The Application Potential of Data Mining in Higher Education Management

A Study Based on German Universities

zur Erlangung des akademischen Grades eines
Doktors der Wirtschaftswissenschaften
(Dr. rer. pol.)

von der KIT-Fakultät für Wirtschaftswissenschaften
des Karlsruher Instituts für Technologie (KIT)

genehmigte

Dissertation

von

M. A. Karin Hartl

Tag der mündlichen Prüfung: 09.07.2019

Erster Gutachter: Prof. Dr. Gholamreza Nakhaeizadeh
Zweiter Gutachter: Prof. Dr. Melanie Schienle

Karlsruhe, Juli 2019

Acknowledgments

First and foremost, I would like to express my gratitude to my thesis supervisor Prof. Dr. Gholamreza Nakhaeizadeh for the chance to write my thesis under his supervision and for the extraordinary support he gave me from our first encounter to the finish line. I benefited greatly from his extensive experience as a DM specialist and researcher, which he used in his advice on the continuous development and improvement of this thesis.

I would also like to thank Prof. Dr. Melanie Schienle, who acted as my second thesis supervisor. Even though I have not been physically present as a member on her professorship, she always took the time to communicate with me and advise me on my thesis. Her comments have given me new energy and helped me to continually improve my work.

Additionally, I would like to thank Prof. Dr. Olaf Jacob who encouraged me to write this thesis and helped me to obtain the necessary resources to succeed in my project. He always had an open ear for my ideas and questions and helped me with his expertise whenever he could.

Furthermore, I would like to thank my colleagues who accompanied me during the preparation of this research. In many chats and discussions, they have allowed and helped me to reflect on my work, which often gave me the strength to move on. This gratitude goes out to my non-scientific colleagues as well, who especially supported the success of the case studies presented.

Finally, I would like to thank my family and friends who have always encouraged me that I can achieve every goal I set for myself. My gratitude goes especially to my husband Robin, who has accompanied and helped me through all the difficulties I encountered in writing this thesis over the past three years.

Abstract

German universities are facing an intense, competitive environment caused by globalization, digitalization, and public sector reforms. The latter also gave the universities more decision-making autonomy, which goes hand in hand with more responsibilities, but also with the possibility of individualizing their strategy. This thesis examines how German universities can use Data Mining techniques to extract useful information from their available data resources to address these current challenges by supporting management decisions. The use of Data Mining methods in education is called Educational Data Mining. Research in this area has so far focused mainly on supporting students and lecturers. This thesis focuses on researching the benefits of Educational Data Mining for university management, which has been mentioned several times in various Educational Data Mining studies but has not been studied in detail so far.

After discussing the most important challenges faced by German universities, their current tasks and objectives were examined. A framework model was then developed that illustrates how the results of two specific Data Mining projects can help universities tackle the challenges and accomplish their tasks. The selected Data Mining projects are dropout analysis and enrollment prediction because the student and applicant data are available to all the German universities. The proposed framework model was verified with two case studies in which the specified analyses were carried out at a German university of applied sciences. To build well-performing models, several Data Mining methods were used and compared. Subsequently, the results were discussed with representatives from the case university, and suggestions were made how the information generated could be included in the decisions of the university administration.

It has been shown that German universities can use their data resources to support their management activities. An overview of this support was presented in the form of a framework model that is not only a first attempt to close the existing research gap in the field of EDM but should also motivate university decision-makers to use their existing data resources. Therefore, the presented thesis can stimulate further research that combines the results of EDM projects with managerial decisions to increase the efficiency of educational institutions. In addition, university administrators can be inspired to use all available resources to ensure their long-term success.

Contents

List of Figures

List of Tables

List of Abbreviations

AHR	Allgemeine Hochschulreife
ANN	Artificial Neural Network
BA	Business Administration
BAH	Business Administration in Health
CAR	Class Association Rule
CHE	Gemeinnütziges Centrum für Hochschulentwicklung (German center for university development)
CRISP	Cross-Industry Standard Process
DM	Data Mining
EA	Evolutionary Algorithm
ECTS	European Credit Transfer System
EDM	Educational Data Mining
EHEA	European Higher Education Area
EHR	Europäischer Hochschulraum (European Education Area)
fgHR	Fachgebundene Hochschulreife
FHR	Fachhochschulreife
FOIL	First-Order Inductive Learner
FP	Frequent Pattern
HAW	Hochschule für angewandte Wissenschaften (University of Applied Sciences)
HEEQ	Higher Education Entrance Qualification
HEI	Higher Education Institution
HRG	Hochschulrahmengesetz
IE	Industrial Engineering
IEL	Industrial Engineering Logistics
IM	Information Management
IMA	Information Management Automotive
IT	Information Technology
k-NN	k-nearest Neighbors
LA	Learning Analytics
LL	Log Likelihood-Function
LMS	Learning Management Systems
LOF	Local Outlier Factor
LOM	Leistungsorientierte Mittelvergabe (performance-oriented fund allocation)
MOOC	Massive Open Online Course

NN	Nearest Neighbor
NPM	New Public Management
OTH	Ostbayerische Technische Hochschule (Technical University of Applied Sciences of eastern Bavaria)
RIPPER	Repeated Incremental Pruning to Produce Error Reduction
ROC	Receiver Operator Characteristic
SMOTE	Synthetic Minority Oversampling Technique
SS	Summer Semester
TH	Technische Hochschule (Technical University of Applied Sciences)
WS	Winter Semester

1 Introduction

This first chapter aims on clarifying the motivation for writing this dissertation. Therefore, urgent challenges faced by German universities are summarized in Section 1.1, which are discussed in more detail in Chapter 2. It is assumed that German universities can address these current challenges by extracting information from their data resources with Data Mining (DM) techniques. Therefore, the research area of Educational Data Mining (EDM), which is the application of DM in the educational sector, is examined in Section 1.2 with two literature reviews. In the first review, the current state of EDM research is surveyed to identify if previous researches investigated the usefulness of EDM for the management of universities. In the second literature review, the status of EDM research in Germany is recorded because the presented research is focused on German universities. Section 1.3 presents the research objectives before describing the approach to achieving these goals in Section 1.4. The last section of this chapter, Section 1.5, gives an overview of the research contributions of this dissertation.

1.1 Relevance of the Research

The recent reforms at German universities changed the tasks of and the demands on the university management, which now has more decision-making autonomy. As a result, university management is more flexible in decision-making but also has the responsibility of defining its own organizational goals. This increases the need for an entrepreneurially oriented strategic university management (Scherm 2013, 2014; Fangmann 2014; Blümel 2016). The growth of universities in terms of number of students, number of employees, and financial resources supports this need and make universities more comparable to small and medium-sized enterprises than to public authorities (Heinrichs 2010: Introduction). Accordingly, in recent years German universities have become service providers exposed to national and international competition (Marettek & Ákos 2010).

This has sharpened the competitive situation of German universities. Privately held universities, international Higher Education Institutions (HEIs), and independent research institutes offer qualified students and researchers alternatives to public universities (Erhardt, D. 2011: 43), which are the focus of this study. In addition, online universities and online degree programs make studying more flexible and allow students to choose from a broad variety of courses, regardless of where they live. Furthermore, the diversity of study programs has been increased, but the number of first-year students is slowly declining, driving universities into a steadily growing competitive environment (Hochschulrektorenkonferenz 2015; Wehrlin 2011). This is supported by globalization, international agreements and advances in information and communication technology that are opening up even more opportunities for the diminishing student body (Erhardt, D. 2011: 2). In addition,

students are more demanding, due to the many choices they nowadays have (Hochschulrektorenkonferenz 2015; Heinrichs 2010: 214; Gerhard 2004: 133). Because satisfied and successful students and graduates are the showpiece of universities that influence the reputation and perception of others, meeting these demands becomes an increasingly important task for universities to ensure their future and long-term success (Kamm 2014: 183).

In summary, two major challenges have been identified for the universities in Germany:

- More decision-making autonomy: During the deregulation[1] of the German public sector and the reform of the New Public Management (NPM), the administrative regulations of public institutions were changed to make them more economical, effective, and efficient (Schmücker 2011: 13). For the management of public universities this meant more decision-making rights, enabling need-based and flexible decisions that increase competitiveness. The accelerated decision-making process, in turn, increased the strategic and operational responsibility of universities and thus the need for professionalized university management (Berthold 2011: 2,33,47; Zechlin 2012: 54).
- Intense competitive environment: In Germany, it is endorsed that universities advocate and compete for highly qualified students, researchers, and resources (Enders 2008: 234; 2009: 7; Haberecht 2009: 31). This has been promoted by the country for several years, because it is considered that the university landscape is mediocre without competition (Haberecht 2009: 31; Kamm 2014: 351). In addition, global networking influences the environment in which HEIs in Germany have to work. This increases the number of competitors, which goes hand in hand with the wide range of degree program choices that students have today. In the course of the Bologna reform and the establishment of university rankings, the comparability of study programs has also increased as these developments support the standardization of the university profiles (Erhardt, D. 2011: 45). In addition, German universities are financially poorly positioned in international comparison (Krull 2008: 243); therefore, they are expected to seek additional financial resources (Kamm 2014: 253).

These challenges are accompanied by new tasks that university decision-makers must face, and new goals must be achieved. For example, universities need to become service providers, reduce their dropout rate, and ensure the quality of education. Accordingly, the need for a professionalized strategic management,[2] based on objective decisions, is more present than ever (Mühlenbein 2006: 30; Wehrlin 2011; Berthold 2011: 33; Scherm 2014). As shown in Figure 1, this is associated with

[1] Deregulation describes the division of responsibility between the state and social institutions, which strengthens their self-responsibility (Erhardt, M. et al. 2008: 4).

[2] Strategic management can be understood as a process of developing and implementing strategies to successfully position an organization in the environment and to build potential and competencies (Bea & Haas 2017).

the growing importance of information[3] as information and knowledge support decision-making, planning security, transparency, and monitoring of results (Wehrlin 2011: 26; Mühlenbein 2006: 30; Scherm 2014; Fangmann 2014). With the help of DM methods, information and knowledge can be extracted from data, which is considered one of the most important assets in the so-called "data age" (Han et al. 2012: 2; Reinsel, Gantz & Rydning 2018). Data are also being generated and stored at German HEIs but have only been used to a limited extend (Scholz 2014).

This study focuses on student- and applicant-related data resources because the Higher Education Statistics Act[4] requires the capture of certain attributes (BMJV 1990). As a result, all public universities in Germany have student data available and can use them to inform their management about their 'clients', including their needs and claims as well as their resource and service requirements.

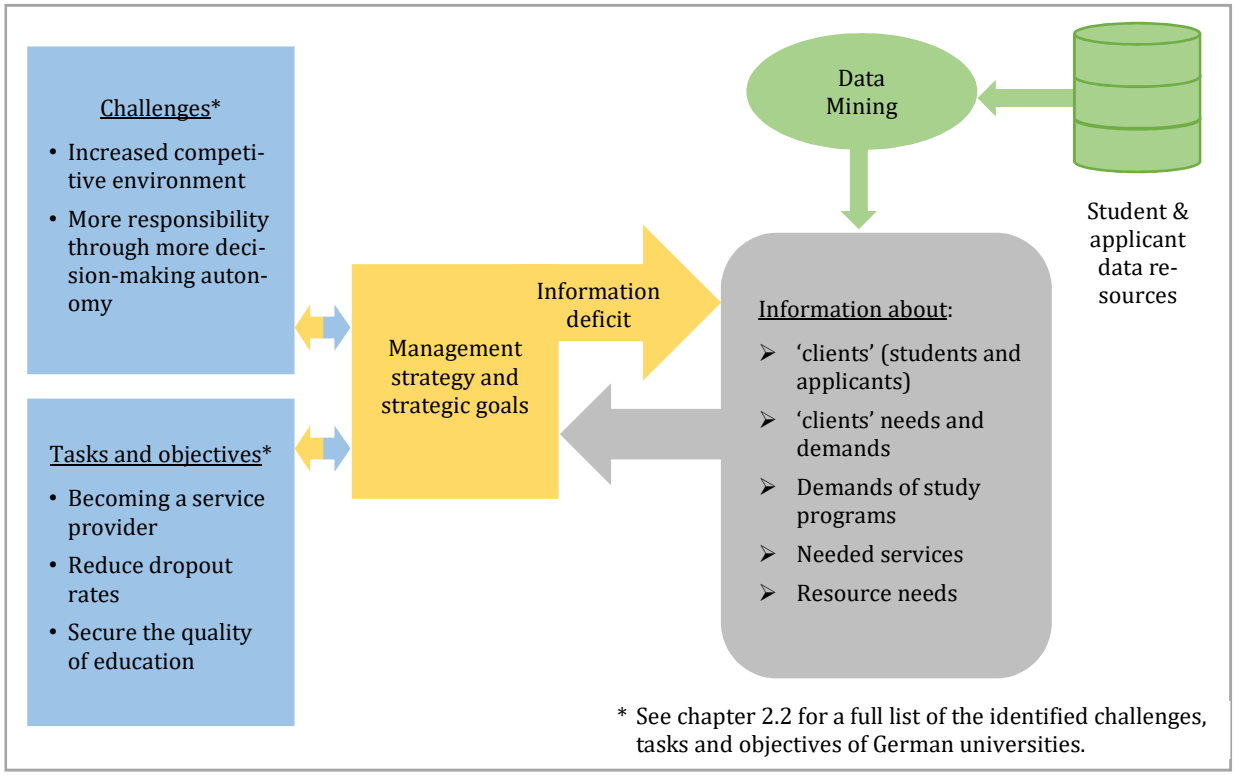


Figure 1. Illustration of the motivation for using DM.

1.2 Current State of the Research

Data Mining is defined by Han et al. (2012) as "… the process of discovering interesting patterns and knowledge from large amounts of data (Han et al. 2012: 5)." Consequently, DM can be understood as an action that leads to meaningful insights that are retrieved from data relative to a predefined target. When used in an educational setting, DM is referred to as EDM, which is understood as "…the development of methods to explore the unique types of data in educational settings and

[3] Information are facts or details on a particular topic that are of value and knowledge for the information recipient.
[4] In German: *Hochschulstatistikgesetz* (HStatG).

using these methods to better understand students and the settings they learn in (Romero & Ventura 2010)." With the development of the term Learning Analytics (LA), which focuses on optimizing the learning environment by collecting and analyzing student data (Papamitsiou & Economides 2014; Romero & Ventura 2013), the understanding of EDM has broadened. Nowadays, it is widely understood as "…the field of using DM techniques in educational environments (Bakhshinategh et al. 2017)." Accordingly, EDM encompasses all applications of DM techniques in educational institutions aimed at identifying knowledge from data generated in an educational environment (Kumar & Chadha 2011). The purpose of EDM, therefore, ranges from understanding and encouraging students and lecturers to assist the management of educational institutions in addressing their various daily challenges (Barahate 2012; Huebner 2013).

1.2.1 State of the art of EDM research

Research efforts in the field of EDM have matured over the last decade, with more publications available each year. To examine whether EDM has already been used to support managerial decisions at universities, a literature review was carried out. This comprehensive review focused on identifying the topics addressed so far in the EDM research community, which in turn provides a good overview of the state of the art. The focus is on survey articles, which have been identified as the optimal starting point because they present an overview of studies already undertaken in the EDM area. In order to find relevant literature, academic libraries and databases were screened for corresponding books and book chapters, journal publications, and conference papers. The following databases were searched: *IGI Global InfoSci Journals Archive*, *IEEE Xplore Digital Library*, *ACM Digital Library*, *AIS e-Library*, *Business Source Premier*, *Springer Link*, and *Google Scholar*. In addition to the keyword "Educational Data Mining", the search terms "Educational Data Mining - Overview", "Educational Data Mining - Review", "Educational Data Mining - State of the Art", "Data Mining in Universities", "Data Mining for University Management", "Data Mining and University Management", and "Knowledge Discovery in University Databases" were used. After surveying the titles and abstracts of the articles found, 39 articles were chosen for further investigation. These are summarized in Table 1 and described below.

Even before the application of DM techniques in an educational context became known as EDM, Luan (2002) discusses possible uses of DM in educational institutions. The author suggests that the enrollment process, marketing measures, and institutional effectiveness can be supported by DM applications. In a second study, conducted two years later, Luan (2004) discusses the possibilities of DM applications for academic planning and intervention, such as the creation of student profiles or the prediction of alumni pledges. Therefore, both studies suggest that the management of educational institutions can benefit from analyzing their data resources with DM methods.

Table 1. Overview of the objectives focused by EDM research, identified from the reviewed articles.

References	Student modeling				Feedback for lecturers	Visual analytics	Support planning & scheduling activities	Construct and evaluate courseware & e-learning	Social Network Analysis	Decision support & DSS[5]
	Performance & dropout prediction	Generate profiles & groups	Support & recommendation	Enrollment prediction						
Luan (2002)				X			X			X*
Beikzadeh et al. (2005)	X*	X*	X*		X*		X*			
Luan (2004)				X		X				X*
Delavari et al. (2005)	X*	X	X*	X*	X*		X*	X*		X*
Romero & Ventura (2007)			X		X		X			
Delavari et al. (2008)	X	X	X				X	X		
Baker & Yacef (2009)	X	X						X		
Romero et al. (2010)	X	X	X		X	X	X	X	X	
Kumar et al. (2011)	X	X	X	X	X		X	X		
Scheuer & McLaren (2012)		X	X		X		X			
Bala & Ojha (2012)	X		X	X				X		
Barahate (2012)	X	X	X					X		
Calders & Pechenizkiy (2012)			X		X		X	X	X	
Goyal & Vohra (2012)	X	X		X		X	X			X
AlHammadi & Aksoy (2013)	X	X	X		X		X			
Huebner (2013)	X	X	X					X		
Mohamad & Tasir (2013)		X						X	X	
Romero et al. (2013)	X	X			X		X			
Bousbia & Belamri (2014)	X	X	X		X		X	X		
Papamitsiou et al. (2014)	X	X	X	X	X					
Peña-Ayala (2014)	X	X	X		X		X	X		
Prakash et al. (2014)	X	X					X	X		
Suhirman, Zain & Chiroma (2014)	X	X	X		X	X	X	X	X	
Fatima, Fatima & Prasad (2015)	X	X	X					X		
Ganesh & Christy (2015)	X	X			X			X		X
Jacob et al. (2015)	X	X	X		X		X			
Sen (2015)	X	X		X				X		
Sukhija, Jindal & Aggarwal (2015)	X	X					X	X		X
Thakar, Mehta & Manisha (2015)	X	X		X	X		X	X		
Mehta & Buch (2016)	X									
Tandale (2016)	X	X	X	X						
Bakhshinategh et al. (2017)	X	X				X	X	X	X	X
Dhingra & Sardana (2017)	X	X	X		X			X		
Dutt, Ismail & Herawan (2017)		X	X					X	X	
Silva & Fonseca (2017)	X	X	X				X			
Thilagaraj & Sengottaiyan (2017)	X	X		X		X	X			
Lenin (2018)	X	X			X					X
Manjarres, Sandoval & Suárez (2018)	X	X	X		X			X		
Aldowah, Al-Samarraie & Fauzy (2019)	X	X	X			X	X	X		
Count	**32**	**33**	**24**	**11**	**19**	**7**	**23**	**24**	**6**	**8**

* Only suggested by the authors.

In the same year, Beikzadeh & Delavari (2005) published a study identifying six processes within an HEI that could benefit from DM applications. The authors suggest useful DM methods and knowledge that can be generated to increase the motivation of HEIs to use DM. The enhanced analytics model was released in 2005, which the authors believe should serve as a roadmap for the use

[5] Decision Support Systems.

of DM in HEIs (Delavari, Beikzadeh & Phon-Amnuaisuk 2005). This roadmap clearly indicates the usefulness of the various DM applications for university administration and management, and the attached brief case study shows that modeling student data can help universities to understand the requirements students have. In 2008, Delavari, Phon-Amnuaisuk & Beikzadeh (2008) re-published the roadmap, combined with analytical guidelines aimed at improving the current decision-making process of HEIs. However, a thorough review of the roadmap has not been published, so the identified decision support remains a proposal from the authors.

The first comprehensive review of literature published in the domain, which coined the term EDM, examines the research conducted between 1995 and 2005 (Romero et al. 2007). The authors note that the focus of DM applications in the educational sector was on support and recommendations for students and lecturers, and on the improvement of planning and scheduling of educational activities. A further literature review published by Baker et al. (2009) categorizes the EDM's main areas of research until 2009. The authors point out that the main interest of EDM research has been in student models generated based on web data and student interaction data. In 2010, Romero et al. (2010) published their second comprehensive EDM literature review, which identifies the key stakeholders of EDM applications and categorizes the tasks that so far have been addressed in the research community. In addition to students and lecturers, course developers, educational researchers, managers, and administration are identified as main target groups of EDM.

In 2011, Scheuer et al. (2012) provide an overview of the topics discussed in EDM research and the DM methods used. The overview presented shows that reports and alerts that support students and lecturers are a major topic in EDM. It is suggested that these reports and alerts can also help other university stakeholders, which leaves room for the assumption that EDM applications are also profitable for university management. In the same year, Kumar et al. (2011) provide another summary of the tasks addressed in EDM research, namely, student modeling, feedback for lecturers and researchers, planning and scheduling, and the evaluation and construction of courseware.

The Bala et al. (2012) study discusses the DM methods used in the EDM area and the main applications for which they were used. The authors state that EDM can provide decision support for university management, but the article does not provide details on how to achieve this managerial support. In comparison, Barahate (2012) identifies the learners and lecturers as the main target groups of EDM applications. The review conducted by Goyal et al. (2012) combines the aforementioned articles by discussing the uses of EDM in HEIs. Accordingly, it is proposed that learners' and lecturers' profit from EDM, which in turn can improve educational practice.

In 2013, Romero et al. (2013) provide another overview of the current state of the EDM domain, which aims to categorize the research area. Therefore, the authors discuss related topics, stakeholders, and the main applications of DM in the educational sector. They also suggest a need for

educational institutions to develop a data-driven culture that uses data to improve decision-making. This indicates that, although EDM has become a developed research discipline, EDM applications thus far are mainly isolated and not integrated into an institution-wide strategy. The survey conducted by Huebner (2013) in the same year points to research gaps that have not been addressed by the EDM community so far. The authors point out that several studies have discussed the benefits of EDM for improving institutional efficiency, but the question of how institutions can use EDM to achieve improvements is unanswered. In addition, they note that many of the studies to date are not generalizable case studies. This is underpinned by the EDM review by Mohamad et al. (2013), which also points out that the research results available so far are difficult to apply in a different context than the one in which they are generated. In the short paper published by AlHammadi et al. (2013), an impact cycle is presented that suggests that knowledge about the student created with the help of DM can influence the work of lecturers, administrators, and managers. This cycle is a first attempt to link the results of EDM projects with institutional efficiency.

Another comprehensive literature review aimed at organizing recent work in EDM research according to its goals and the methods used was presented by Peña-Ayala (2014). The author states that most research focuses on student modeling to characterize the learner, their performance, and their behavior. In the same year, Papamitsiou et al. (2014) published a literature review focusing on the practical work achieved in EDM and LA. The authors emphasize that most case studies conducted so far are exploratory or experimental in nature and primarily support learners and lecturers. The comprehensive review conducted by Suhirman et al. (2014) focuses on the research carried out for educational decision support. An overview of the tasks addressed by the EDM community is presented, and the key benefiters are identified that are analogous to those discussed in previous research. The comprehensive description of the tasks discussed so far mentions the possibilities EDM provides for the management of educational institutions. In addition, the authors suggest that standardizing DM applications and tools for the needs of educational institutions can increase the use of EDM. However, the research does not discuss details of how the management of educational institutions is supported.

In 2015, Thakar et al. (2015) presented another overview of the EDM domain, which subdivides the past research into five main areas: surveys of articles published in EDM, prediction of academic achievement, comparison of DM techniques to predict the academic performance, the correlation between pre- and post-enrollment factors and others, including the determination of student satisfaction or the assessment of the faculty. The authors note that the approaches used in different researches are separate. Accordingly, a need for models and regulatory frameworks is identified that will ensure the sustainable growth of EDM applications at all levels of a university. A second study to summarize EDM submissions by 2015 has been published by Jacob et al. (2015). The au-

thors point out that EDM has so far mainly aimed at studying, predicting, and improving the academic performance of learners, but the knowledge generated can be helpful to all academic stakeholders, including management. The short papers published by Ganesh et al. (2015) and Sukhija et al. (2015) discuss a limited number of researches according to their contributions. Both articles summarize that the main focus of the research investigated is on supporting the students and lecturers of educational institutions. In 2016, Mehta et al. (2016) published a study discussing the utility of EDM for the educational sector, which states that the use of DM techniques supports its general improvement. This conclusion is corroborated by the work of Tandale (2016), which presents a survey of the EDM domain, concluding that all the educational stakeholders can benefit from DM applications. Nevertheless, the study does also not discuss details of how the management of educational institutions is supported.

Another survey, which covers the applications and tasks being investigated by the EDM community, was published by Bakhshinategh et al. (2017). The research focuses on the papers published between 2006 and 2016. The previously addressed EDM applications are compared to the stakeholders they are targeting, with the result that the primary benefiters are the lecturers. In addition, the study indicates that the EDM applications performed examine student characteristics, predict student performance, detect unwanted student behavior, analyze social networks, generate reports and alerts, and improve the planning and scheduling for administrators. Again, no details are provided on how this support is achieved. The short overview provided by Silva et al. (2017) on the EDM domain comes to the same conclusion, but once more it is not specified how the decision-makers of an educational institution can benefit from the insights DM applications provide. The three-decade literature review presented by Dutt et al. (2017) focuses on the use of clustering methods in the EDM domain, which shows that clustering has been mainly used for e-learning-related tasks. The authors conclude their research with a general statement that reiterates that the question of how institutions use DM to improve their institutional efficiency is still unanswered.

In 2018, another comprehensive literature review on the use of DM techniques between 1993 and 2015 was published by Manjarres et al. (2018). The 127 studies investigated in detail were classified according to addressed tasks and applied DM techniques. The authors conclude that the studies published so far in EDM are the description of specific case studies conducted in educational institutions worldwide that are not transferable to other institutions. Lenin (2018) focuses his research on DSS in education and the implementation of DM methods in these systems. The author concludes that little research has been done in this field, which leaves room for the assumption that in general little research has been done with the focus on the decision support that can be provided by EDM applications.

At the beginning of 2019, the comprehensive review of Aldowah et al. (2019) was published, discussing the usefulness of EDM and LA for higher education in the 21st century. In total, 402 articles

were examined. The authors emphasize that both EDM and LA can help HEIs to develop strategies that focus on students. In addition, they state in several phases that decision support is a major advantage of EDM in HEIs. However, a framework or other overview on how the findings generated in EDM projects can support university management is still not provided. Accordingly, the previously identified research question of Huebner (2013) and Dutt et al. (2017), as to how EDM can improve institutional efficiency, still remains unaddressed.

As previously mentioned, Table 1 summarizes the EDM topics that were identified by the articles examined as being investigated by research in the EDM community. The main field of application of EDM research is, therefore, student modeling, which summarizes the prediction of student dropouts and performance, the creation of student profiles, the grouping of students according to their characteristics and their learning behavior, the provision of support and recommendations to the student as well as the prediction of course enrollments. Providing feedback to lecturers is another well-researched application of EDM, with the aim of leveraging lessons learned to tailor teaching and services. The EDM applications, which focus on planning and scheduling, aim to enhance the traditional educational processes with facts. This includes the need-based development of curricula, the optimization of student timetables, and the prediction of alumni donations. Furthermore, EDM researches use data derived from Learning Management Systems (LMS) and e-learning applications to evaluate existing courses and course materials and to create new ones.

In addition, visual analytics has been identified as a scope of EDM research to support the decision-making process by highlighting useful information and abnormalities. The detailed study of research using visual analytics shows that the focus is on visualizing the behavior of students in online environments, often only based on descriptive analytics methods. The use of DM in DSS and social networks are further areas of interest. So far, however, relatively little research has focused on these topics.

The results of the literature review show that EDM has become an established field of research in recent years. So far, EDM research has focused on analyzing students and their learning environment (Romero et al. 2007, 2010; Huebner 2013; Ganesh et al. 2015; Mehta et al. 2016). Nevertheless, several review articles and literature studies identify EDM as an appropriate decision-support approach. As one of the first researchers in the field, Luan (2002, 2004) argues that the DM models used in business are also applicable to the educational environment. This is supported by a framework developed by Beikzadeh et al. (2005) that identifies opportunities for EDM applications in the areas of planning, evaluation, and marketing. The trend is continuous. More and more researchers are emphasizing the need for HEIs to use data-driven facts in their administrative processes and argue that DM applications have great potential (Thakar et al. 2015; Silva et al. 2017; Bakhshinategh et al. 2017; Aldowah et al. 2019). Therefore, in addition to students and lecturers, the administration and management of educational institutions are referred to as the benefiters of

EDM applications. Consequently, the supportive power of DM applications for objective decision support to managers is already recognized.

Nevertheless, these assumptions and statements are solely conclusions of the authors of the articles examined and have not yet been investigated in details. Only the work of AlHammadi et al. (2013) attempts to combine the insights provided by EDM applications with the efficiency of educational institutions. Accordingly, there is a research gap, first pointed out by Huebner (2013). As mentioned earlier, the authors note that the question of how educational institutions use EDM to improve institutional efficiency remains unanswered. This statement was repeated by Dutt et al. (2017) and Aldowah et al. (2019). Consequently, the research gap still seems to exist, along with the need for a framework that ensures the sustainable application of EDM at all levels of educational institutions (Thakar et al. 2015; Aldowah et al. 2019). This thesis aims to fill this gap by developing a framework model that combines the results of EDM projects with the administration and management to increase the effectiveness and efficiency of management decisions.

1.2.2 Educational Data Mining in Germany

In a second literature review, the current state of the art of EDM research in Germany was examined because the focus of this thesis are German universities. Again, academic libraries and databases[6] were searched for relevant books and book chapters, journal publications, and conference papers. The search terms used were "Educational Data Mining in Germany OR Deutschland", "Data Mining at German Universities OR Hochschulen", "Data Mining for German University Management", and "Predict Student Performance in Germany OR Deutschland". This search revealed only a very small number of studies conducted by German researchers based on data resources from German educational institutions. As a result, a secondary study was conducted in which the literature cited and examined in the survey articles analyzed in Section 1.2.1 was reviewed for German participation. In addition, the proceedings of the *International Conference on Educational Data Mining* from its beginning in 2008 until now have been studied as well.

Table 2 shows an overview of the studies found, which were further investigated for the main topics addressed. It can be seen that most publications are so far case studies that use LMS data to model learner behavior and predict student success. Only the study by Schönbrunn & Hilbert (2007) was identified as obviously targeting the improvement of planning and scheduling, but no continuation of the research efforts was found. Kemper, Vorhoff & Wigger (2018) and Berens, Oster, et al. (2018) use DM techniques to predict the dropout of students at German universities to decrease dropout rates. This is also a part of the presented dissertation; the focus, however, in this research is on the

[6] The following databases were searched: *IGI Global InfoSci Journals Archive*, *IEEE Xplore Digital Library*, *ACM Digital Library*, *AIS e-Library*, *Business Source Premier*, *Springer Link*, and *Google Scholar*.

support these models provide for the managerial decision-making process of universities. The visual analytics suggested by Askinadze, Liebeck & Conrad (2019) and Askinadze & Conrad (2018a) may also support the university management, but the authors have not yet discussed this analysis purpose.

Table 2. Overview of the EDM studies conducted in Germany.

Reference	Content overview	Main topics of the investigated research
Askinadze et al. (2019)	Presentation of alternative visualization techniques that can illustrate complex relationships in educational databases	Data visualization
Kemper et al. (2018)	Application of different machine learning approaches to predict student dropout for industrial engineering study	Student performance prediction
Askinadze et al. (2018a)	A short paper about integrating a dashboard visualizing student data	Data visualization
Berens & Schneider (2018); Berens, Oster, et al. (2018)	Discussion and presentation of an early intervention system used to reduce dropouts	Student performance prediction
Askinadze & Conrad (2018b)	Proposal of a concept that provides transparency for the use of student data by having the students decide on the data used for analysis	Data security
Backenköhler et al. (2018)	Proposal of an approach that allows the recommendation of personalized curricula based on a generated course dependency model	Student success, planning, and scheduling
An, Krauss & Merceron (2017)	Investigation into whether the research conducted in MOOCs[7] can be generalized to other LMS-supported courses	LA, LMS
Albrecht (2017)	Estimation of students' knowledge of programming by analyzing the responses to open-ended programming tasks	Student performance prediction, LA
Ifenthaler, Mah & Yau (2017)	General discussion on how LA can support the success of students in German HEIs	Student success, LA
Dyckhoff (2018); Dyckhoff et al. (2012)	Proposal for an explorative learning analytics tool to improve technology-based learning	LA, LMS
Stapel, Zheng & Pinkwart (2016)	A case study that uses an ensemble method to improve the predictive accuracy of student performance prediction models in online learning environments for math	LA, student performance prediction, DM model improvement
Paaßen, Jensen & Hammer (2016)	Proposal for representation of computer programs via execution traces for the analysis of intelligent tutoring systems	LA, LMS
Klüsener & Fortenbacher (2015)	Prediction of student success based on their interaction in MOOCs	Student performance prediction
Voß et al. (2015)	Use of matrix visualization to predict performance in intelligent tutoring systems	Student performance prediction, LA
Zheng, Zhilin, Stapel & Pinkwart (2016)	Discussion of whether perfect scores in math learning systems are actually predictors for good student performance	Student performance prediction
Zheng, Zhilin, Vogelsang & Pinkwart (2015)	Discussion of organizing students into groups in MOOCs using a grouping algorithm to improve performance	Student performance prediction, LA
Bengs & Brefeld (2014)	Proposal of a novel approach for computer-based adaptive speed tests evaluating skill levels	LA, LMS
Koch et al. (2014)	Discussion of the use of text mining to improve software engineering courses	Course evaluation and improvement
Bergner et al. (2012)	Prediction of student responses from student's online interaction (authors use a dataset from the Massachusetts Institute of Technology)	Student performance prediction, LMS
Merceron et al. (2012)	A case study examining the use of e-learning platforms by students	LA
Holzhüter, Frosch-Wilke & Klein (2012)	Exploration of learner models with rules based on log data from an e-learning system	LA, LMS
Thai-Nghe, Horváth & Schmidt-Thieme (2011)	Application of factorization techniques to improve the accuracy of performance prediction models based on log files	Student performance prediction, LMS
Goguadze et al. (2011)	Assessment of how Bayesian models use log files to predict students' misconception	Student performance prediction, LMS

[7] Massive Open Online Courses.

Reference	Content overview	Main topics of the investigated research
Lemmerich, Ifland & Puppe (2011)	A short paper discussing how to identify students' success factors with subgroup discovery	Student performance prediction
Merceron (2011)	Short paper examining the use of the data available in LMS using association rules	LMS
Krüger, Merceron & Wolf (2010)	Presentation of a data model that structures data storage in LMS to enable and support data analysis	LMS
Merceron & Yacef (2008)	Investigation into the interestingness measures that are considered appropriate for the application of association rules to educational data	DM model improvement
Schönbrunn et al. (2007)	Presentation of an approach to demand-oriented planning of Bachelor's and Master's degree programs at German universities	Planning and scheduling

In addition to the research already published, two BMBF[8] projects have been identified that address the problem of student dropout using DM methods. The project *FragSte*,[9] represented by the work of Berens & Schneider (2018) and Berens, Oster, et al. (2018), will investigate whether early intervention methods based on predictive models actually decrease the students' dropout rates at universities (BMBF 2017b). The aim of the *DMPS*[10] project is to generate models that predict students' dropout rates based on the data of the *National Educational Panel* study, which "...collects longitudinal data on the development of competencies, educational processes, educational decisions, and returns to education in formal, non-formal, and informal contexts throughout the life span (NEPS 2018)." At the time of writing this thesis, no results were available on the *DMPS* project. It has to be noted, however, that both projects focus on the prediction of student dropout to reduce dropout rates. As already mentioned, this is also one of the topics of this study. Nevertheless, in addition to creating predictive models that forecast student dropout, this study discusses the positive effects of these models on the success of a university and the positive effects of other DM-based analyses for the university management.

In summary, it should be noted that EDM research in Germany is still in its infancy because relatively few contributions are made by German researchers or on the basis of data from German education institutions. It should also be noted that most of the studies examined are based on data from LMS or other online courses. Therefore, the analysis of further data resources from German educational institutions seems to only have been sparsely researched. In addition, studies investigating the usefulness of DM methods to support management decisions at German universities were not identified in the literature reviews presented, suggesting that this topic has not yet been discussed.

8 In German: *Bundesministerium für Bildung und Forschung*.

9 In German: *Früherkennung abbruchgefährdeter Studierender und experimentelle Studien zur Wirksamkeit der Maßnahmen* (BMBF 2017b).

10 In German: *Determinanten und Modelle zur Prognose von Studienabbrüchen* (BMBF 2017a).

1.3 Goals of the Research

To support German universities in addressing their current situation and to contribute to the identified research gap in the EDM community, this thesis explores the possibilities that EDM offers for the sustainable development of HEIs. Therefore, the main objective of this thesis is to explore and illustrate how HEIs can use their available data resources to objectively support their management decisions with DM techniques. The focus is on the student and applicant data that are available to all universities. It is therefore assumed that every single German university can benefit from EDM.

Specifically, the following objectives are pursued:

- A framework model, hereinafter simply referred to as framework, will be created that will close the identified EDM research gap by linking the results achieved in EDM projects to university management decisions. It, therefore, illustrates how the efficiency of HEIs can be increased through DM. More specifically, the potentials of analyzing available student and applicant data for German university administration and management, including tackling challenges and fulfilling tasks, will be explored in this context. In addition to clarifying the relationship between EDM outcomes and managerial decision support, the framework aims to illustrate the possibilities contained in data and to motivate universities to use that resource to their advantage.
- Two case studies will be carried out to validate the proposed decision support by analyzing the student and applicant data of a case university. The first case study tackles the challenge of overbooking study programs with predictive DM models that can forecast the enrollment of applicants. In the second case, the student data are analyzed to estimate dropout and identify reasons for student failure. The results of both case studies are validated and discussed with the responsible decision-makers of the case university. The findings and proposed actions that will be extracted from the generated models will show once again how EDM can be used to support the decision-making of university administrators and managers.
- An 'EDM-process box' will be generated containing all analytical processes performed during the case studies. This allows the management of HEIs to reconstruct the applied analysis. In addition, EDM researchers can use the processes to further develop the efforts made in this research.
- A contribution will be made to further develop EDM research in Germany. Like other educational institutions worldwide, German universities cannot turn away from digitalization and globalization. Therefore, it is important that German universities and researchers engage in and use the advantages of EDM so as not to loose touch with international competitors. Ideally, the results presented should encourage further research in this area.

1.4 Research Outline

To achieve the identified goals, the following research steps were performed. First, the current situation of German universities is analyzed in Chapter 2 to get an overview of the challenges they face. This includes a brief introduction to the structure of German public universities and the foundations of the reforms that have influenced their current situation. Afterwards, the present challenges of German universities will be discussed and their main tasks and goals described.

Chapter 3 introduces and explains the DM methods used in this study. This includes the introduction of the CRISP-DM, which is used by the case studies in Chapter 5. Also explained are the steps required to preprocess the data, including the treatment of missing values, outlier detection, feature selection, and the balancing of datasets. Following the introduction of applied DM methods, the final section of this chapter focuses on the performance assessment measures used to validate and select classification models.

Chapter 4 discusses the main assumption that has motivated this research, which is that the current challenges of the universities can be tackled with DM methods. Therefore, the incentive to use DM methods to solve existing problems is explained. Subsequently, a framework is proposed which illustrates how the results of two selected EDM projects can contribute to the achievement of the current tasks and objectives of German universities. In the final section of this chapter, the specificities encountered during the DM-based analysis of student and applicant data are addressed and proposals are made that aim on helping other German HEIs analyze their data resources.

The two case studies, which were conducted with data from a German university, are presented in Chapter 5. The first case study analyzes the applicant data to predict the enrollment of applicants. The results are intended to help university decision-makers to cope with the ever-present challenge of overbooked study programs at German universities. The second case study analyzes the student data with the aim of predicting the dropout of students and identifying the requirements of the study programs that may lead to student resignations or transfers. Both case studies are structured according to the CRISP-DM described in Chapter 3. After several models have been created and compared, each case study ends with a discussion. In this discussion, conclusions and suggestions are made on how the case university can use the analysis results to support its management decisions, which may also be helpful for other universities.

In Chapter 6, the thesis is concluded with a final discussion and a summary, containing proposals for further research.

1.5 Research Contribution

In the course of this dissertation, the following contributions have been made:

1. The current state of EDM research has been examined through a comprehensive literature review that clarifies the need for this study. Two major research gaps have been identified, namely the need to clarify the support that EDM provides to increase the effectiveness and efficiency of educational institutions and the lack of EDM research contributions that have emerged in Germany. The literature reviews and the identified research gaps can be used as the starting point for further relevant research.
2. The situation of German universities has been analyzed and the main challenges they face are elaborated. In addition, the core tasks and objectives with the focus on the universities of applied sciences in southern Germany were summarized and prioritized. Accordingly, an overview of the existing challenges, opportunities, tasks, and goals of German universities is given.
3. A framework model has been created that provides an overview of how the current challenges and tasks of German universities can be supported with EDM. This has been done for the first time, to the knowledge of this study. The focus of the framework is on student and applicant data as all German universities have access to these resources. In addition to raising awareness and motivating university decision-makers to use their data resources, the framework is helping to close the identified EDM research gap. This is achieved by linking the EDM project outcomes with the decision support they explicitly provide to the university decision-makers. Furthermore, the framework presented can serve as a guideline for universities and motivate them to use all available resources to ensure their long-term success and existence. In addition, it may serve as a starting point to encourage further EDM research focusing on the management support that EDM provides for HEIs.
4. The specifics encountered during the application of the DM process at the case university are discussed. This may especially help German EDM researchers to reconstruct the results in other institutions and settings and to conduct further research.
5. Two case studies have been carried out, showing that the data resources mentioned, although they are not as extensive as in other countries,[11] contain interesting information for the university decision-makers. The models generated to predict the enrollment of applicants show that universities can gain a deeper understanding of their 'clients', which helps them act purposefully and to support the need-based planning of resources. They also provide information that helps to optimize administrative processes. The knowledge gained in the second

[11] In Germany, only a few demographic characteristics of the students are recorded, including gender, age, place of birth, and final grade of the Higher Education Entrance Qualification (HEEQ). As a rule, there is no information available about the parents, the health, or the financial situation of the student or applicant. If these data would be collected by German universities as well, it is assumed that even better predictive models can be generated.

case study with the dropout analysis shows that among other things, the university management can use EDM to maintain and increase the quality of their study programs and reduce the dropout rate. In addition, the case studies contain ideas and measures to achieve the proposed management support, which can be further explored by research and practice.

6. A RapidMiner[12] 'EDM-process box' was created containing the DM analysis performed in this study in the form of processes. These processes can be downloaded and implemented in RapidMiner, which may help practitioners and researchers to reconstruct the research results presented, adapt them to their own environments, and to develop the proposed ideas further.

[12] RapidMiner is a software platform for DM that "…unites data prep, machine learning and predictive model deployment."(RapidMiner 2019).

2 The Current Situation of German Universities

This chapter examines the current situation of German universities. The first section briefly introduces the German higher education sector before discussing recent developments in this sector. The presented developments influence the current situation as well as the tasks and objectives of German universities, which are discussed in the second section of this chapter. In this context, tasks are understood as the responsibilities a university has to fulfill according to state regulations, especially the ones defined in the State University Law.[13] With objectives, the goals are described that the universities define individually, for example in the target agreement or the development plan.[14]

2.1 German Universities and Recent Developments in the Higher Education Sector

The education system in Germany is divided into five main steps, which are shown in Figure 2. Usually, children start their education in kindergarten and preschool, before attending elementary school up to the 4th grade. Afterwards, pupils start their secondary education, and the educational path is individualized. Depending on their performance in the first couple of school years, they then either attend the *Hauptschule,*[15] the *Realschule,*[16] the *Gesamtschule,*[17] or the *Gymnasium*.[18] In the *Hauptschule* and the *Gesamtschule*, the students learn until the 9th grade before graduating and continuing with an apprenticeship or the *Berufsschule.*[19] The *Realschule* is more demanding, and graduation takes place after successfully completing the 10th grade. Afterwards, the pupils attend an apprenticeship or continue their education at a *Fachoberschule.*[20] Students that attend the *Gymnasium* graduate after successfully completing the 12th or 13th grade.[21] The students choosing to attend the *Gymnasium* or the *Fachoberschule* or any other school qualifying for a Higher Education Entrance Qualification (HEEQ) may continue their education at the university.

13 In German: *Landeshochschulgesetz*.

14 It has to be noted that these goals are usually in line with the state tasks. Nevertheless, they can be individualized and complimented by the university.

15 Secondary modern school from the 5th until the 9th grade.

16 Junior high.

17 Comprehensive school.

18 Grammar school.

19 Vocational school.

20 Specialized secondary school.

21 The number of school years for *Gymnasium* depends on the German federal state the school belongs to. For example, in Bavaria a student has to complete 13 school years in total, whilst in Saxony the Gymnasium ends with the 12th grade.

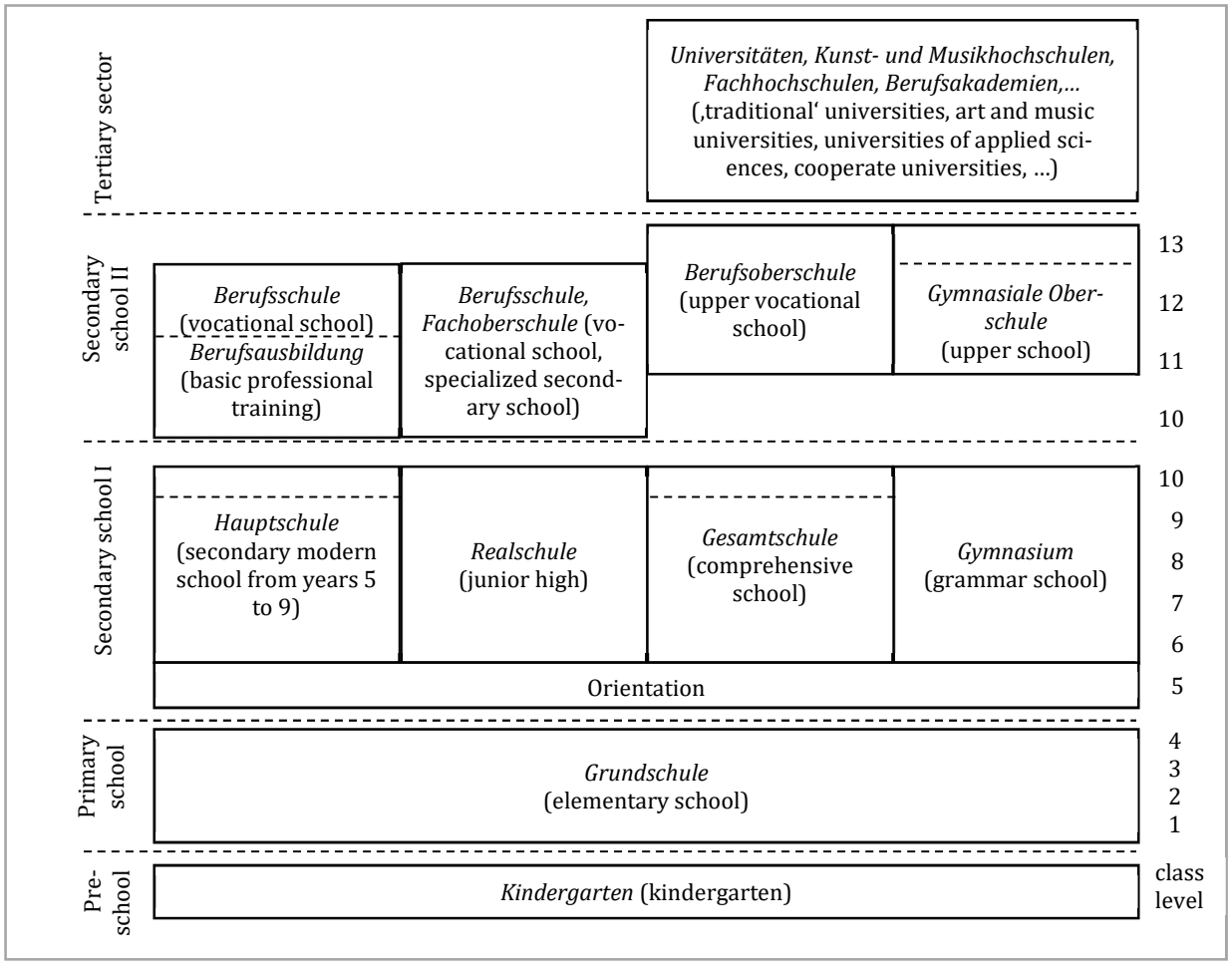


Figure 2. Structure of the German educational sector, based on Mühlenbein (2006: 28).

Alternatively, they choose a profession and continue with an apprenticeship. The successful completion of an apprenticeship in combination with professional work experience can also pave the way to university education.

Universities are at the highest academic level of the German education system. Legally, universities are understood as corporations under public law and as state facilities simultaneously (Kohmann 2012: 68). They are well-established institutions with a long history dating back until the middle ages. With this long history, the university landscape is nowadays very complex, and many forms of HEI can be differentiated. One main distinction is to be made by their sponsorship and management, which may be either private or public. The legal organization, as well as the funding situation between those two types of universities, differ significantly. Private universities are privately managed and funded whereas public universities are non-profit institutions[22] that are financially supported by the federal states with the respective state budgets (Warnecke 2016: 36).

[22] Non-profit organizations do not produce output to maximize profits, they produce a product or service to address an existing demand (Erhardt, D. 2011: 30).

This research focuses on public universities, which in turn can be grouped into the following six main types, which can be distinguished according to their focus, their academic rights, and their size (Heinrichs 2010: 28):

- The 'traditional' universities, or scientific universities, focus on research activities and the education of academics (Warnecke 2016: 2). Generally, these are long-established institutions offering a wide variety of academic programs (Warnecke 2016: 40). Traditionally, they have the largest student body of up to 54.000 students.[23] In addition, they have the sole right to grant the academic titles of a Ph.D. and a habilitation.[24]
- Universities of applied sciences[25] exist since the 1960's and focus on applied teaching and research (Warnecke 2016: 41). Their activities are more practice-oriented and often seek to provide qualified personnel to regional companies (Gerhard 2004: 59). Predominately, the academic programs are clearly structured and aim to prepare the student for a specific profession.
- Pedagogical universities have the same rights and responsibilities as 'traditional' universities but focus on educating school lecturers in their degree programs.
- Theological universities are also comparable to 'traditional' universities, but their academic programs and research activities are clearly focused on religious topics and professions.
- Art and music universities focus on educating and researching art-related topics, including visual arts, performance arts, and music.
- Administration universities are public sector universities that train professionals for the public sector. Accordingly, they are closely associated with the federal government and the German state.

The most common type of universities in Germany are the universities of applied sciences (Hachmeister et al. 2013: 6), as illustrated in Figure 3. Accordingly, 51% of the entire university body in Germany are universities of applied sciences. These play an important role in Germany not only because of their sheer numbers. They are innovators and trainers for regionally needed professionals since they aim to educate students with practical knowledge and they usually have a strong connection to their region and its businesses (Erhardt, D. 2011: 8). For this reason, their existence is very important for the German economy.

Apart from the mentioned differences, German universities are similar in their main rights and obligations. They are legal entities of their own in statutory law, with the right of self-administration,

[23] For example: University of Cologne and Ludwig-Maximilian-University in Munich.
[24] Postdoctoral qualification as a university lecturer.
[25] In German: *Fachhochschulen* oder *Hochschulen der angewandten Wissenschaften* (HAW).

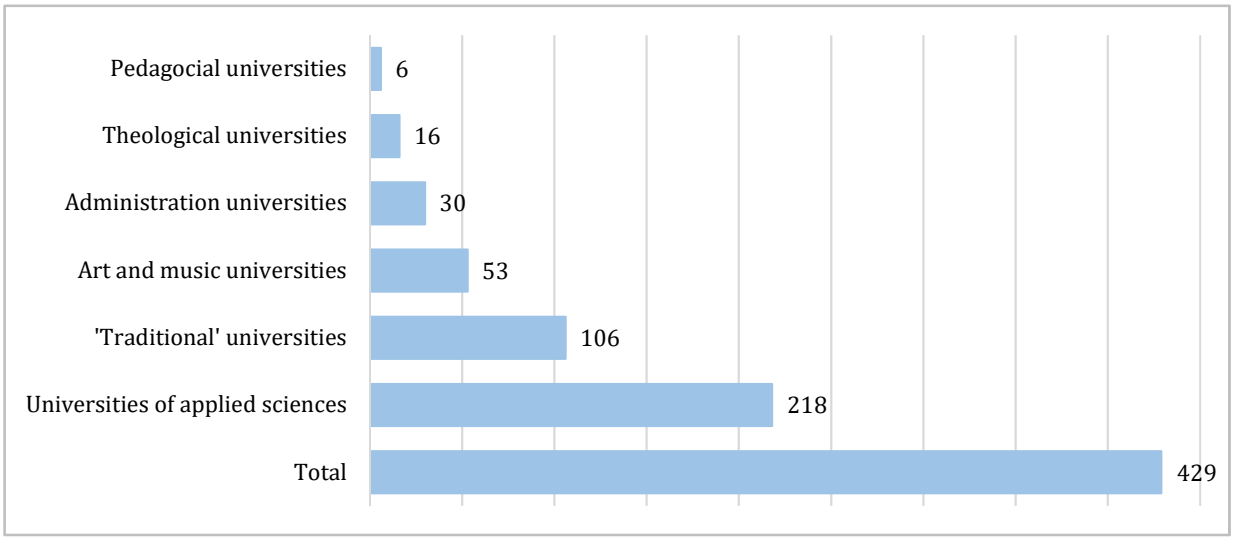


Figure 3. Number of German universities in WS 2017/2018 differentiated according to the university type (DESTATIS 2018a).

freedom in science, freedom in research and teaching as well as the right of granting academic degrees[26] (Heinrichs 2010: 24). The organizational structure of universities is democratic, and committees make the important decisions and are therefore responsible for the management, which is understood as the systematic planning, steering, and control of institutional activities (Erhardt, D. 2011: 79). The main committees are the *Hochschulleitung,*[27] the *Senat,*[28] the *Hochschulrat,*[29] and the *Fakultätsrat.*[30] The main tasks of each committee are described in Table 3.

The basic guidelines for the university management in Germany are set out in the constitution[31] of each university. In this constitution, each university defines separately (Heinrichs 2010: 42):

- their organizational structure,
- the composition and election procedure for the Senate,
- the composition and election procedure for the University Council,
- the leadership and composition of their faculties and institutes,
- the methodical procedures in their committees,
- the university services,
- the role and responsibilities of the equal opportunities commissioner and the student body, and
- the appeals procedure for academic staff and lecturers.

[26] Except Ph.D. and habilitation, which only can be granted by 'traditional' universities with the right to award these titles.
[27] University Management Board.
[28] University Senate.
[29] University Council.
[30] Faculty Council.
[31] In German: *Grundordnung*.

The overarching tasks of universities are summarized in the *Hochschulrahmengesetz (HRG),*[32] which specifies that universities are responsible for educating their students with the necessary knowledge that best prepares them for their future profession, which can be in academia or the economy (Kamm 2014: 122). Furthermore, they have the obligation to nurture and develop science and art through research, teaching, studies, and further education (Warnecke 2016: 37).

Table 3. Main tasks of the committees in German universities, derived from Heinrichs (2010: 45-47) and Geis (2017: 388-415).

Committee	Main tasks
University Management Board	The main management body of German universities consists of the rector, who is also known as the president, the co-rectors (or vice-presidents), and the chancellor. The president is the chairman of the university, who is supported through the vice-presidents. These positions are elected and filled for a certain tenure, usually between 2 and 6 years. The chancellor is the only permanent member of the university management. He or she is responsible for the economic management and human resource management. Together, all bodies of the university management are responsible for supporting the efficiency, the flexibility, and the competitiveness of the university. In addition, the University Management Board decides on the main objectives of the university, which are formulated in a structural and development plan.[33] These main goals are discussed and agreed on with the federal state. Furthermore, the university management is responsible for ensuring the quality of its offered services, for distributing the available resources, for the external representation, and for the management of the day-to-day business of the university.
University Senate	The Senate consists of representatives of the various internal interest groups of a university, namely the management, the research staff, the administrative staff, and the student body. Furthermore, a women's representative and an equal rights representative are part of the Senate. The members of this council are elected through an internal procedure, which is usually set out in the constitution of the university. In general, the Senate is responsible for all tasks and decisions related to research and teaching that have not explicitly been assigned to the faculties.
University Council	The university council has the tasks of a directorate and consist of members belonging to the university as well as external representatives, who are members of other universities, companies, and the general environment of the university. The main responsibilities of the council are the strategic development of the university taking into account regional, countrywide, and international developments.
Faculty Council	Faculties are grouped areas of knowledge and structure a university. In an organizational context, they are seen as units, managed by an elected deanery and by a faculty board composed of representatives of the different groups within a faculty. As a rule, faculties are responsible for exam plans, curricula, faculty-related structure and development plans as well as overseeing and securing the educational quality of the study programs offered.

In addition to research and teaching, universities also have several secondary targets. They are service providers[34] that offer education and professional training, which supports students in their professional life. They are also service providers to employees by offering family-friendly working conditions, a healthy work environment, and the possibility to participate in further education measures that support individual development and good living conditions (Kohmann 2012: 77). Last but not least, they are service providers for the society because they support the progress and the sustainable as well as responsible development of society (Kohmann 2012: 77). In addition,

[32] Higher Education Law.

[33] In German: *Hochschulentwicklungsplan*.

[34] A service provider is understood as an institution that develops and offers services.

universities have to be economically viable. Accordingly, universities need to balance their income and expenditures and deliver the best possible performance in research and teaching using the available resources (Kohmann 2012: 78; Bolsenkötter 1976: 42-45).

The main processes of a German university that aim on achieving their main tasks and objectives (Becker 2011: 9-10) are shown in the process map in Figure 4. Comparable to businesses, universities have core processes, management processes and support processes. The management processes define the strategic direction of the university (Kocian 2007). The core processes are the interface to the external university stakeholders and are directly associated with the value creation and the target achievement (Kocian 2007; Becker 2011: 10). Furthermore, they support the brand and market development of the university (Njenga et al. 2017). The success of the core processes depends on the support processes, which therefore contribute significantly to value creation and target achievement. As illustrated in Figure 4, the two central processes of a university are *studies, teaching & further education* and *research & transfer* (Kocian 2007; Hochschule Karlsruhe 2018; Hochschule Ansbach 2017; Appelfeller & Boentert 2014; Altvater, Hamschmidt & Sehl 2010).

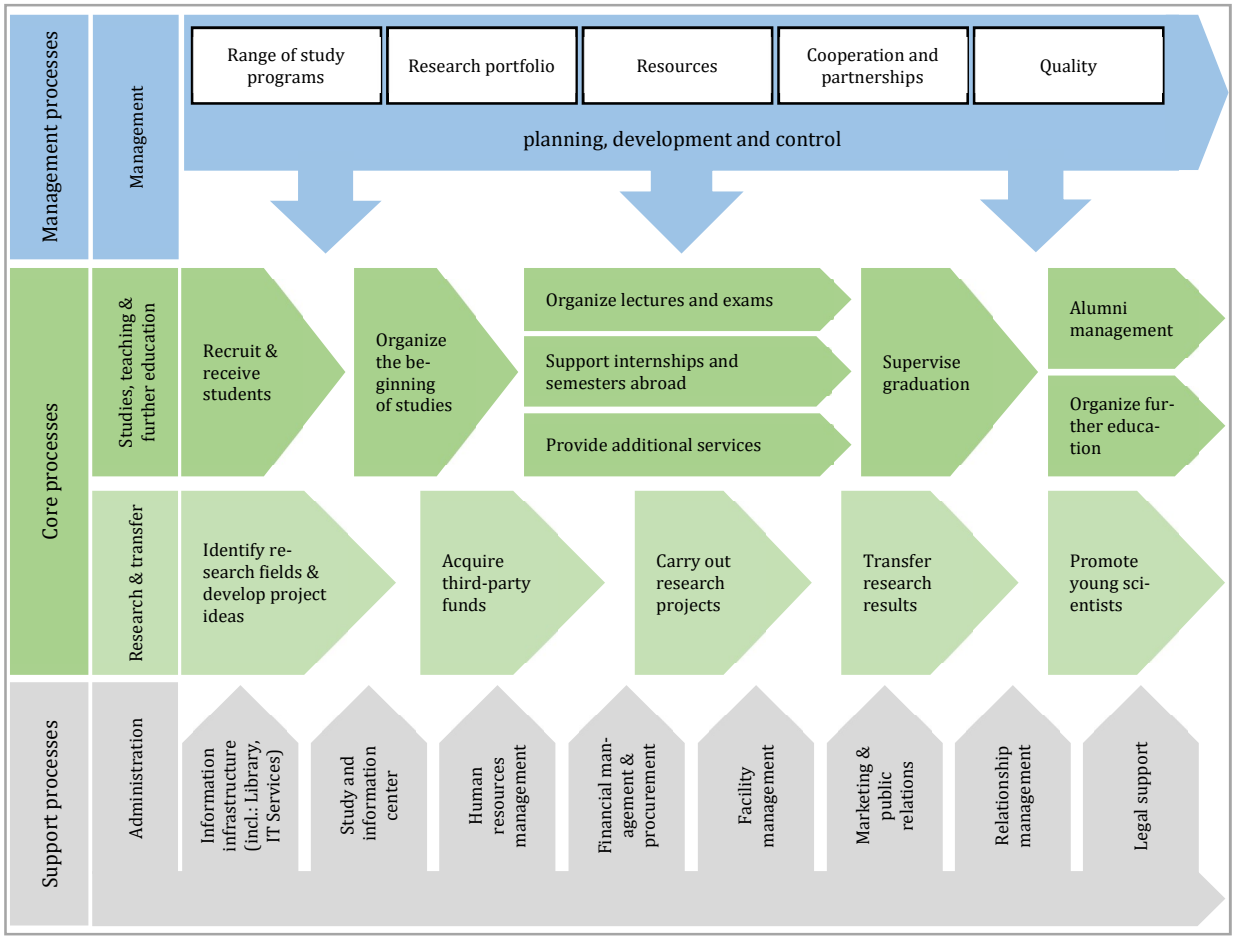


Figure 4. Process map of a German university.

The process *studies, teaching & further education* is oriented on the ideal student lifecycle, beginning with the acquisition of students up to the support of graduates, the maintenance of alumni

relationships and the provision of further education options. The process *research & transfer* is oriented on the ideal lifecycle of a research project, starting by the development of a research idea up to the communication and transfer of the research results and the support of young scientists.

Notwithstanding the fact that it has long been a goal of universities to be a service provider for students, employees, and society and that they have long been urged to work economically, at the end of the 20th century, the organization of the university has been repeatedly criticized (Kehm 2015; Kohmann 2012: 45). This gave the impression that German universities in their present structure are not able to face the challenges of the global knowledge society (Kohmann 2012: 45). Furthermore, the productivity and efficiency of universities were criticized. Consequently, several reform movements and statutory changes have been initiated that changed the tasks of the university management. The ones relevant to this research are described in the following.

2.1.1 New Public Management reform

New Public Management (NPM) was a reform aimed at modernizing public organizations (Schedler & Proeller 2009: 5; Kehm 2012: 17). It came from the criticism on public institutions that they were not productive, powerful, and efficient (Martinez 2009: 25). Therefore, the NPM reform aimed to transform the management of public organizations from a state-controlled administrative body into competitive and economically managed service providers (Kamm 2014: 53; Schedler et al. 2009: 18-19; Heß 2005: 151-153). As a result, the role of the state as the main control mechanism that dictated the tasks and objectives in detail changed for universities (Lanzendorf & Paternack 2009). Nowadays, the state is responsible for setting the general objectives, whilst the universities can decide for themselves on their detailed courses of action (Berthold 2011: 31). The main content and objectives of the NPM are summarized in the following (Hood: 1991:4-5):

- Establishment of a practice-oriented and professionalized management with a high degree of autonomy
- Development of objectives and performance indicators that enable performance measurement
- Establishment of performance-oriented steering mechanisms that reward productivity
- Formation of manageable subdivisions that are self-responsible
- Increasing the competition through temporary contracts and public calls for proposals
- Orientation to economic management styles that provide more flexibility in changing environmental conditions
- Establishment of economic resource management

In order to improve the performance of the public sector, the NPM reform, therefore, relies on entrepreneurial structures. In particular, the strategic goals outlined in Table 4 have influenced the

university landscape and thus the strategic objectives of the HEI management.

Table 4. Strategic changes in the NPM reform affecting the management of German universities (Schedler et al. 2009: 67-84; Bogumil & Heinze 2009: 7).

Strategic goal of NPM	Affects for universities
Customer orientation	The public institutions are encouraged to consider the needs of their citizens and understand them as their clients.[35] At universities, students are the main clients. Accordingly, universities need to consider their demands and requirements in their management decisions and understand themselves as a service provider.
Competitiveness	The creation of structures comparable to the free market economy creates a national and international competitive environment. This means that universities have to compete. Accordingly, they have to find a way to set themselves apart from the competition in order to attract qualified students, employees and funds.
Performance orientation	Traditionally, public organizations and authorities have been controlled through their assigned inputs. As a result, resource-efficient work, which aims at reducing costs, often lead to less financial outlay in the next funding period. Consequently, so not to lose funds, the waste of resources was not unlikely. Therefore, a shift from an input to an output orientation for the resource allocation in public institutions is desired. In the case of universities, the German federal states increasingly use performance-oriented reimbursement allocation,[36] which is linked to predefined strategic goals between the federal state and the universities. Ideally, these goals are measurable and regularly reviewed for their efficiency, effectiveness, and implementation. The funds are then distributed according to the fulfillment of the predefined performance indicators (Kamm 2014: 216-217).
Quality assurance	Quality assurance in public institutions is an important and challenging task as measurable goals and indicators are needed to ensure clarity. The definition of objective achievement metrics is particularly difficult for universities since their main products are intangible services, e.g. study programs or research results. Nevertheless, quality assurance is a major task of universities, and many HEIs are beginning to introduce quality management systems to enhance the value of their research, teaching, and processes (Nickel 2014: 3). Such quality assurance procedures are the *Lehrbericht,*[37] course evaluations, accreditation of degree programs, university rankings, target agreements,[38] and the performance-oriented allocation of funds (Michalk & Richter 2007).

2.1.2 Bologna process

The Bologna process is a European-wide reform with the aim to establish and strengthen the development of an European Higher Education Area (EHEA) and to improve the quality of study programs, to increase the international competitiveness of universities, and to promote mobility and employability of the university members (Berg & Dahm 2009: 46; BMBF 2015; Jubara et al. 2006: 55). In detail, the following main goals have been defined (BMBF 2015; Heinrichs 2010: 58; Maassen 2004: 12; Berg et al. 2009: 48-51):

[35] In the public context, a client is understood as a person that consumes the services of a public entity (Schedler et al. 2009: 69).

[36] In German: *Leistungsorientierte Mittelvergabe* (LOM).

[37] Teaching Report.

[38] In German: *Zielvereinbarungen*.

- Establishment of a staged study structure, which is separated in undergraduate studies and graduate studies with comparable degrees across Europe: For this reason, the traditional German HEI degrees *Diplom* and *Magister* have been replaced with the Bachelor's and Master's degrees. With the Bachelor's degree, the students gain a first professional qualification that enables them to pursue a professional career. The Master's degree complements the Bachelor's degree and serves to develop the scientific abilities of the students as well as to deepen their knowledge. Both degrees are assessed through an accreditation process that aims to ensure consistent quality at European universities (Kamm 2014: 237).
- Introduction of comparable grading structures through ECTS[39] credit points and the diploma supplement:[40] Both the ECTS and the diploma supplement have been introduced to support easy recognition of study achievements at European universities. Specifically, ECTS credit points make the workload of courses and programs comprehensible. The diploma supplement illustrates the course content and indicates the skills that a student acquired by successfully completing a specific course.
- Structuring of study programs into modules: Modules[41] have been set up to structure the study programs in terms of content and chronology. This was done to simplify the recognition process of study achievements, improve the efficiency of the study programs, and to make study programs more flexible to change.
- Establishment of binding structures and cooperation across Europe: The introduction of common pan-European structures and guidelines has been undertaken in order to develop a quality control mechanism and to improve the quality assurance process. A well-established measure is the accreditation process, which ensures the maintenance of guidelines and structures through objectively evaluating the quality of the study programs.
- Enhancing the attractiveness of the EHEA: All the above measures were introduced to increase the attractiveness and worldwide recognition of European education.
- Introducing lifelong learning opportunities: The Bologna process recognized the importance of lifelong learning and supported the development of flexible learning opportunities, e.g. further education programs or flexible university entry requirements for career changers.
- Supporting mobility: In order for all the members of the university to benefit from the establishment of the EHEA as a pillar of modern society, the mobility between universities is promoted. This allows students, lecturers, and staff members to share knowledge and experiences with other institutions.

[39] Abbreviation for the European Credit Transfer System (Maassen 2004: 43).

[40] The diploma supplement is a uniform and internationally understandable document that describes university degrees and the qualifications obtained in those degrees. It is an additional document, provided to each graduate in combination with the degree certificates (Maassen 2004: 226).

[41] A module is a combination of lectures that are focused thematically on one subject area (Maassen 2004: 37; Graumann et al. 2004: 133).

The Bologna process was constantly criticized during its implementation. Nevertheless, the most important changes have so far been adopted and established in over 48 countries, including Germany (European Higher Education Area 2018). The introduced changes do support the interconnectedness of universities, but they increased the competitive environment as well.

2.1.3 University rankings and the Excellence Initiative

University rankings were created by independent institutions to compare universities, their achievements, and services (Gerhard 2004: 171; Kamm 2014: 214). In these comparisons, the scarce good high-ranking places are awarded to a limited number of institutions, which can profit from a respectable reputation that goes with this distinction (Kamm 2014: 215). Internationally renowned rankings are the *Shanghai Academic Ranking of World Universities* and the *Times Higher Education World University Ranking* (Kamm 2014: 289). In both, the ranking places of German universities still offer room for improvement with only 4 universities being in the Top 100 of the *Shanghai Academic Ranking of World Universities* (Shanghai Ranking 2018).

A well-known ranking in Germany is provided by the *Gemeinnütziges Centrum für Hochschulentwicklung* (CHE). Through a survey, which is based on openly available measures, the ranking analyzes indicators in the areas students, student results, internationalization, research, studies and teaching, equipment, practice orientation, and learning environment (Berghoff et al. 2009). The individual results that a university achieves in these areas can be compared online (CHE 2018). These are designed to be informative and can be used in order to obtain an overview of the various study offers and services of the individual Germany universities. Furthermore, the individual strengths and weaknesses of the study programs can be assessed and compared. Accordingly, the CHE ranking assigns no ranking place. Instead, it allows students and other stakeholders to benchmark universities, which has become increasingly important through the growing choice of study opportunities and the growing desire of students to secure their choice of university and degree program with information (Erhardt, D. 2011: 54).

The Excellence Initiative[42] is a project funded by the German research association[43] and the German science council,[44] which was launched in 2006 in response to the poor rankings of German universities in the above-mentioned international rankings (Kamm 2014: 154; Neundorf 2009: 109). The aim of the initiative is to support cutting-edge university research, to improve international competitiveness, and to strengthen the reputation of German higher education (Neundorf 2009: 110; Roessler 2012: 4; DFG 2013: 13; Münch & Pechmann 2009: 71). In the last funding period from

[42] In German: *Exzellenzinitiative*.
[43] In German: *Deutsche Forschungsgemeinschaft*.
[44] In German: *Wissenschaftsrat*.

2012 until 2017, 99 individual projects were funded at 44 universities (DFG 2013: 15). These universities benefit from financial support over a period of 5 years with total funding of 2.4 billion euros. In addition, they are awarded the title of *Excellence University* (DFG 2013: 17). The final decision on the universities and initiatives receiving funding in the upcoming period was not communicated at the time of writing, but the program is definitively continuing. This will give a selected body of universities again the opportunity to benefit from additional funding and added prominence.

Even though the initiative did not catapult German universities into the top ranks internationally, the image growth in Germany is noticeable for the participating institutions (Krull 2008: 247). So far, including the upcoming funding period, the program is open only to 'traditional' universities. Accordingly, universities of applied sciences cannot benefit from this initiative and the corresponding support (Kamm 2014: 155). Therefore, they must find different ways to succeed in today's competitive environment.

2.2 Current Challenges of German Universities

Universities can no longer escape the social, political, and economic changes of recent years (Kohmann 2012: 1). In Figure 5 an overview is given of these main influences as well as their consequences for universities.

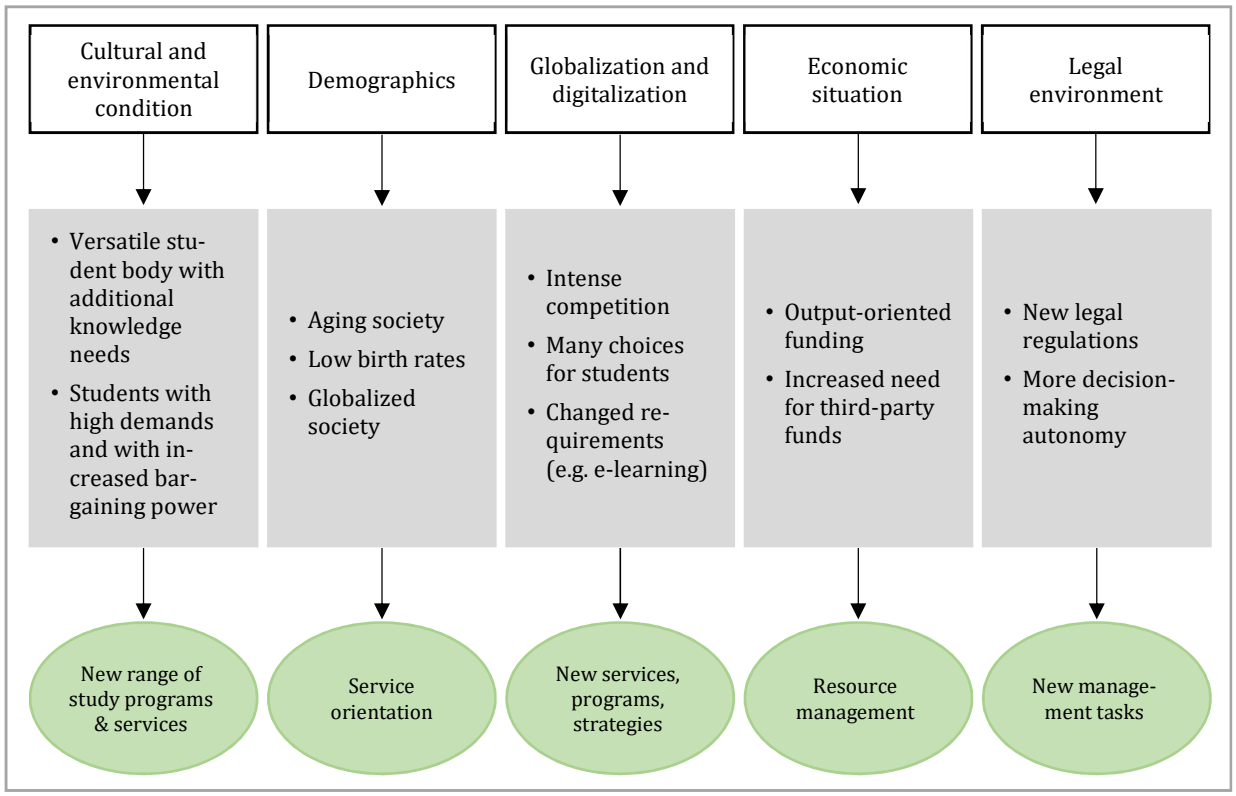


Figure 5. Main external influences on German universities based on Schmücker (2011: 37).

As became clear, in the course of the internationalization and the establishment of modern information and communication technologies, the demands of the labor market, as well as the cultural and social values of the students, changed. The current student body is versatile and has a wide multitude of choices and new demands on its education. Furthermore, students have more influence than ever because demographic changes, such as the declining birth rate in Germany and the willingness of German students to study abroad, lead to a decrease in student numbers. This increases the need for German universities to actively secure their student numbers. Therefore, universities need to adapt their study programs and services to the demands of their stakeholders. They must become service providers who respond to the needs of their students.

Due to the technological developments of the communication and digitalization age, the universities are confronted with additional competitors. In order for German universities to secure their position, they must invent new strategies that include the development of their services and programs to set themselves apart from their competition. Furthermore, the universities are facing changes in legislation, which present them with new tasks. In addition, the economic situation of universities changed significantly. Basic funding is still provided to the universities by the state, but a portion of the financial allowance of universities is coupled with the achievement of pre-defined targets. Moreover, universities are increasingly dependent on third-party funds to have sufficient financial resources. As a result, the universities are in need of effective and efficient resource management.

The following sections of this chapter discuss the mentioned challenges in more detail and aim at raising awareness for the current situation of German universities, which brings with it new tasks and objectives. The focus has been laid on universities of applied sciences because they are the majority of German universities (Erhardt, D. 2011: 76; DESTATIS 2018c). In addition, according to their often smaller size and their shorter existence, they on the one hand have a worse position in the competition than ‘traditional’ universities that can spread their risks and leverage economies of scale that often support them in a strong market position (Erhardt, D. 2011: 76). On the other hand, universities of applied sciences are closer to their ‘clients’ and can respond more flexibly to current and future requirements. Therefore, it is assumed that especially universities of applied sciences have the need but also the ability to plan on-demand and provide services that respond to current trends and requirements. It has to be noticed, however, that the identified challenges, tasks and objectives do not exclusively apply to universities of applied sciences and, therefore, in most cases apply to all public universities in Germany.

In Section 2.2.1, the consequences of greater decision-making autonomy are discussed before the competitive situation of German universities is illuminated in Section 2.2.2. Afterwards, the current tasks of German universities are extracted from the State University Laws. After that, the main objectives are examined, which were taken from the target agreements and mission statements of the

applied sciences universities in southern Germany. This focus has been chosen because these applied sciences universities are the closest competitors to the case university of this research. The challenges discussed are just a selection that has been identified as being urgent and ubiquitous by this research. Accordingly, this research does not claim to provide a complete picture of the challenges that universities in Germany currently face.

2.2.1 Decision-making autonomy and output-orientation

The university landscape in Germany has changed substantially in the last two decades. Previously, the strategies and goals of state universities were determined by statutory and financial regulations (Krücken & Wild 2010; Erhardt, D. 2011: 30). These limited the university in its decision and its possibilities of action. With the deregulation of the public sector in connection with the NPM management reform, the management of public authorities has been changed to make these state institutions more economical, efficient, and effective. For universities, this meant an adjustment of the funding-related governance mechanism from input-oriented to output-oriented and granting of decision-making rights to the university governing bodies (Erhardt, D. 2011: 33; Bogumil et al. 2009: 8). This resulted in a strengthened autonomy of universities, which is desirable in all German federal states (Kohmann 2012: 3; Lanzendorf et al. 2009: 17).

Autonomy in the university context is a complex system with many potentials (Haberecht 2009: 33). The following potentials, which have been summarized by Erhardt, M., Meyer-Guckel & Winde (2008), have opened up to the university management by the increase in decision-making autonomy:

- Universities have been given the flexibility to respond to environmental changes and requirements because decisions can now be made and implemented immediately, which accelerates the overall management process.
- Universities can be competitive because they have the ability to develop their own strategies and profiles.
- Universities can make demand-based decisions that are close to their tasks and objectives. A university knows its requirements best and these can now be involved in the decision-making processes. Furthermore, the individual market situation can be considered in the strategic orientation of a university.

Consequently, autonomy provides universities with the abilities to be competitive, to be flexible, and to make demand-based decisions that are close to the actual challenges they individually have to face. However, the autonomy of universities is not without restrictions. Tasks and targets that are influenced and formulated by the federal state and that are therefore prescribed by law, limit

the strategic freedom of the universities. This makes it all the more important for the management to use the scope of their freedom choice to set themselves apart from the competitors.

In addition, autonomy increases the responsibilities of universities and the need for them to work more efficiently (Erhardt, M. et al. 2008: 8; Kohmann 2012: VII). Nowadays, these are economic entities that are in many ways more comparable to companies than most other public authorities (Heinrichs 2010: 17). Surely, they are not profit-oriented, but their management requirements, their market situation, their public relations and their marketing needs are very close to those of commercial service providers (Heinrichs 2010: 215). Accordingly, universities have strategic and operational responsibilities, including defining strategic goals and deriving operational measures to achieve these goals (Berthold 2011: 47; Marettek et al. 2010). In addition, universities are accountable for their outputs and have to justify their actions (Zechlin 2012: 54). Consequently, university management needs to be professionalized to ensure the appropriate use of the benefits that autonomy offers as well as to ensure its future success (Berthold 2011: 33). This research assumes that data and the information contained in this data can help universities considerably in identifying their individual capabilities, which will help them to seize the opportunities for autonomy.

2.2.2 Competitive situation

A situation is described as a competition when more than one market participant tries to appropriate a limited commodity (Kamm 2014: 18). As a result, not all market participants are able to achieve their full interests (Kamm 2014: 68). The increase in the numbers of universities and educational service providers, the elimination of barriers by the recent developments in the information and communication technology, the NPM reform, and the Bologna process have immensely increased the competitive situation of German universities (Erhardt, D. 2011: 45,64). In addition, the globalization of the economy and the development of English as the scientific language of the world is effacing world's borders and causing universities to face many national and international competitors.

In the German education sector, competition has arisen not only because of the changing environmental conditions. It was actively sought by the relevant state institutions to reorient the universities and encourage them to think economically and to be innovative (Kamm 2014: 27,351; Pasternack 2008: 197; Neundorf et al. 2009: 7). But in the course of the above-mentioned reforms, which introduced, for example, the Bachelor's and Master's degree structure and the target agreements, the homogenization of the German higher education sector was supported. Accordingly, the programs and services offered by the universities are often interchangeable because identical targets promote comparable strategies, structures, and processes (Erhardt, D. 2011: 73). This homogenization is further supported by university rankings and the Excellence Initiative because universities are guided by the defined evaluation criteria when aiming for a high-ranking place (Erhardt,

D. 2011: 73). Consequently, the products and services of German universities are highly comparable and, as a rule, several institutions offer similar study programs, research services, and career opportunities.

Therefore, universities compete for students, researchers, assignments, results, cooperation, and resources (Kamm 2014: 144). Moreover, universities contend for reputation. As a result, the competitive situation of German universities is comparable to that of service providers and is influenced by the competitive forces of additional service providers, customers with high bargaining power, comparability, and scarce resources (Erhardt, D. 2011: 30).

2.2.2.1 Additional service providers

The main competitors of universities are other national and international universities, especially those with comparable profiles and therefore, most likely a comparable position in the competitive environment (Schmücker 2011: 40). In addition, the group of institutions active in the education market extends to private and online universities, public and private research institutions, contract research institutions, and research institutions in companies. These increase the offers in the education market, which intensifies the competition in the education and the research sector (Erhardt, D. 2011: 3; Kamm 2014: 21; Schmücker 2011: 40).

In particular, private universities[45] developed into a major competitor for universities of applied sciences because they are comparable in size and in their focus on practice-oriented education. In addition, many private universities are state-recognized, bringing them even more to the same level. Otherwise, their private ownership allows them to develop and implement strategies with the support of commercial services, e.g. professional marketing companies that are not open to public universities. Accordingly, they have competitive advantages.

Table 5 shows the development of the number of students in 'traditional' universities, universities of applied sciences, and private universities.

Table 5. Development of student numbers from WS 2015/16 until WS 2017/18 (DESTATIS 2018d, 2018b).

Type of University	WS 2015/16	WS 2016/17			WS 2017/18		
	Number of students	Number of students	Change from the previous year	Increase in %	Number	Change from the previous year	Increase in %
Private universities	196,450	211,569	15,119	7.7%	230,197	18,628	8.8%
'Traditional' universities	1,729,503	1,747,515	18,012	1.0%	1,754,634	7,119	0.4%
Universities of applied sciences	929,241	956,717	27,476	3.0%	982,188	25,471	2.7%

[45] Private universities are all universities that are privately owned.

The comparison indicates that private universities do perform well in today's competitive environment (Lanzendorf et al. 2009: 15). Whilst the student numbers of the 'traditional' universities only increased by around 1% in recent years, the applied sciences universities recorded an increase of almost 3%. In comparison, the student numbers at private universities grew by 7.7% in the WS 2016/17 and 8.8 % in the WS 2017/18. Accordingly, they have the highest growth in student numbers. Consequently, private universities seem to be able to identify and fulfill the demands of students in a way that they are able to offer services that set themselves apart from their public sector competition and that increase their desirability (Engelke, Müller & Röwert 2017: 7,8). Surely they have a different initial situation than the universities of applied science, but they are nevertheless competitors.

2.2.2.2 Customers with high bargaining power

Students can be understood as the consumers of higher education services and therefore an important stakeholder for a university (Lomas 2007: 32). As the main client, they demand the best possible preparation for their professional lives and have high expectations on the universities, their services, their support, and their infrastructure (Börgmann & Bick 2011: 75). The global and digitalized environment of the education sector offers students and researchers many opportunities. They can choose between many different educational programs and institutions from around the world. Furthermore, students are an important asset for universities because they are drivers and multipliers of the university image and reputation (Erhardt, D. 2011: 15).

Reputation is an asset to universities that can influence long-term success and existence. It cannot be built and introduced solely by the management of a university. It depends much more on all university members, including lecturers and students (Gerhard 2004: 5). Accordingly, when a student experiences deficits of knowledge in his or her professional life, the feeling towards the university will be negatively influenced, which will reflect their opinion of the institution (Erhardt, D. 2011: 17). Otherwise, positive experiences develop loyalty and ensure a long-lasting relationship between the graduate and the universities. Therefore, graduates strengthen the connection between universities and the economy and expand their influence (Erhardt, D. 2011: 18). As a result, a good relationship with their graduates helps universities in connecting with the regional economy, which then may have a positive impact on additional funding.[46]

The rivalry of universities for students is also influenced by the wide range of study programs offered at state-owned and state-recognized universities, combined with the predicted decline in student numbers (Erhardt, D. 2011: 56-59; Gerhard 2004: 4). Since 2013, universities in Germany do have a stable number of freshmen, which is more than 504,000 new students every year[47]

[46] Project funding and third-party funds are often dependent on the recommendation and support through companies.
[47] 2013 = 508,621; 2014 = 504,882; 2015 = 506,580; 2016 = 509,760 (Hochschulrektorenkonferenz 2017: 23).

(Hochschulrektorenkonferenz 2017: 23). Since 2014, the freshman numbers have increased slightly by 0.3% – 0.6%. Nonetheless, von Stuckrad, Berthold & Neuvians (2017) predict a decline in freshman numbers over the years to around 490,000 new students in 2027, which means that German universities will have to deal with a decrease in freshman numbers by at least 5% (von Stuckrad et al. 2017: 30). This is not a massive downtrend, but in parallel, the number of study programs offered has risen by an average of 3% per year in recent years (see Table 6). Accordingly, the ratio between freshman numbers and number of study programs decreases each year, which is also supported by the affinity of German students to study abroad (Erhardt, D. 2011: 47). As a result, an increasing number of knowledge and research providers face a diminishing number of customers that they need to compete for, which in turn strengthens the student's influence (Erhardt, D. 2011: 56).

Table 6. Development of the study program numbers from WS 2012/13 to WS 2017/18 at state and state-recognized universities (Hochschulrektorenkonferenz 2017: 9).

Semester	Number of study programs	From which		
		Bachelor	Master	Other
WS 2012/13	16,082	7,199	6,735	2,148
WS 2013/14	16,634	7,477	7,067	2,090
WS 2014/15	17,437	7,685	7,689	2,063
WS 2015/16	18,044	8,298	8,099	1,647
WS 2016/17	18,467	8,471	8,358	1,638
WS 2017/18	19,011	8,677	8,703	1,631

In order to differentiate themselves from the competition and to attract enough as well as successful students, the universities have to offer demand-oriented services and study programs. These services and study programs should support the students in their future careers, which in turn helps the university to build and strengthen its reputation. Accordingly, customer orientation has become a major topic for universities in Germany.

2.2.2.3 Comparability of study programs and rivalry for high-ranking places

Some of the developments, which are accompanied by the latest reforms and changes in the higher education market, lead to high comparability of universities in Germany. In the context of the Bologna reform, the establishment of the Bachelor's and Master's degree structure made the education comparable between 'traditional' universities and universities of applied sciences, blurring their traditional clear distinctness in research institutions and institutions with a practical orientation (Maassen 2004: 86; Kamm 2014: 128). One reason is the fact that both universities now have to provide a professional qualification with the Bachelor's degree that allows the student to take up employment after successful completion (Erhardt, D. 2011: 51). This development has been accompanied by widely available study program reviews and university rankings that provide students with easy access to the comparison of universities and their study programs. But these rankings and reviews do have some major downsides. First, only a limited number of institutions can

occupy the highest-ranking places. Second, to secure a high rank, universities focus on improving the specific criteria that are measured in the ranking (Erhardt, D. 2011: 73). Therefore, universities tend to develop standardized profiles, which supports the feeling of interchangeability of study programs and universities (Münch et al. 2009: 80; Erhardt, D. 2011: 73,81). As a result, the university stakeholders cannot recognize the uniqueness of an institution, leading to students placing more emphasis on the ranking in order to find the program that best supports their needs (Maassen 2004: 163).

The German Excellence Initiative has a comparable effect, besides its good intentions to catapult German universities in the ranks of international elite universities. In addition, the universities labeled as excellent are only about 10% of the total university body in Germany, leaving the remaining 90% in a highly competitive environment (Hartmann 2006: 449). Accordingly, only a limited number of universities can benefit from the additional funds and use the initiative to set themselves apart from the competition. The main body of universities, which does not belong to the excellent ranked, is perceived as being less attractive, which increases the competition between those (Hartmann 2006: 449). Furthermore, an application to the initiative is only open to 'traditional' German universities.

The establishment of university rankings and the German Excellence Initiative, therefore, supported the homogenization of the university landscape and the development of two classes of scientific institutions, where there are very few winners and a large number of the 'rest' that find themselves in a highly competitive environment (Münch et al. 2009: 90; Hartmann 2006: 448; Marettek et al. 2010; Neundorf 2009: 118).

2.2.2.4 Limited financial resources and increased competition for third-party funding

The financial resources of German universities are made up of public funds, third-party funds, donations, and revenue from services and further education programs. Most of the funds that are made available publicly secure the basic operation of the universities and are mainly a part of the individual budget of the federal states. Accordingly, these financial resources are provided to secure human resources, administrative expenses of a material nature, maintenance needs, construction projects, and other necessary materials.

The second part of the publicly provided funds is linked to the achievement of pre-defined goals[48] in the areas of studies and teaching, research, equality status, and internationalization (Sieweke 2010: 52; Dohmen 2015: 6). In most of the German federal states, predefined indicators measure the target achievement. The funds are then divided between the universities according to their de-

[48] For example: article 5, paragraph 2 in the Higher Education Act of the federal state Bavaria (*Art. 5 Abs. 2 BayHSchG*).

gree in the achievement of the objectives and their position in the comparison with the other universities of the federal state (Dohmen 2015: 6).[49] In most of the German federal states, the evaluation of the universities in terms of achieving their targets is done separately for 'traditional' universities and universities of applied sciences (Dohmen 2015: 6).

Research activities and the development of new services are often only possible through third-party funding. Therefore, the allocation of these resources is an important objective of universities (Hachmeister et al. 2013: 19). In the university context, third-party funds are target-appropriated resources that are usually awarded after a public call for proposals. The institution or institutions with the most promising concepts and thus the greatest chance for success receive the funding (Kamm 2014: 225). Accordingly, third-party funds are generally allocated through competitive procedures and are therefore only available to a very limited number of institutions (Jaeger 2009: 45). Nevertheless, universities more increasingly depend on these additional financial resources because the core funding for universities is limited and stagnating (Hornbostel 2008: 256; Schimank 2009: 130). Furthermore, the availability of third-party funding has several positive effects on universities. First, they have additional financial resources that give them the scope for general expansion. Second, universities can expand their research activities, which produces research results and progress. Third, as a recipient of funds granted in a competitive environment, the university can built an image as a committed education and research institution.

Nevertheless, many of the resources available to universities in Germany are earmarked to defined purposes and tasks (Heinrichs 2010: 40; Erhardt, M. et al. 2008: 119). This limits public universities in their decisions, their strategies, and their competitiveness. Furthermore, it has been noted that the financial resources available to higher education in Germany are comparably low internationally (Erhardt, D. 2011: 50). As a result, HEIs in Germany have a comparatively poor starting position in the global competition (Erhardt, D. 2011: 50; Hartmann 2006: 463).

2.2.3 Tasks of German universities

In Germany, the main tasks of universities are anchored in the State University Law. Each of the 16 German federal states defines this law separately since each federal state has the right of control over the field of education and culture under the Basic Law.[50] In these legal acts, a section focuses solely on the definition of the tasks of the universities, which were analyzed individually in the course of this thesis. Subsequently, these tasks have been summarized in Figure 6 and are described below.

[49] Assignment of the funds via the distribution model (in German: *Verteilmodell*).

[50] In German: *Grundgesetz*.

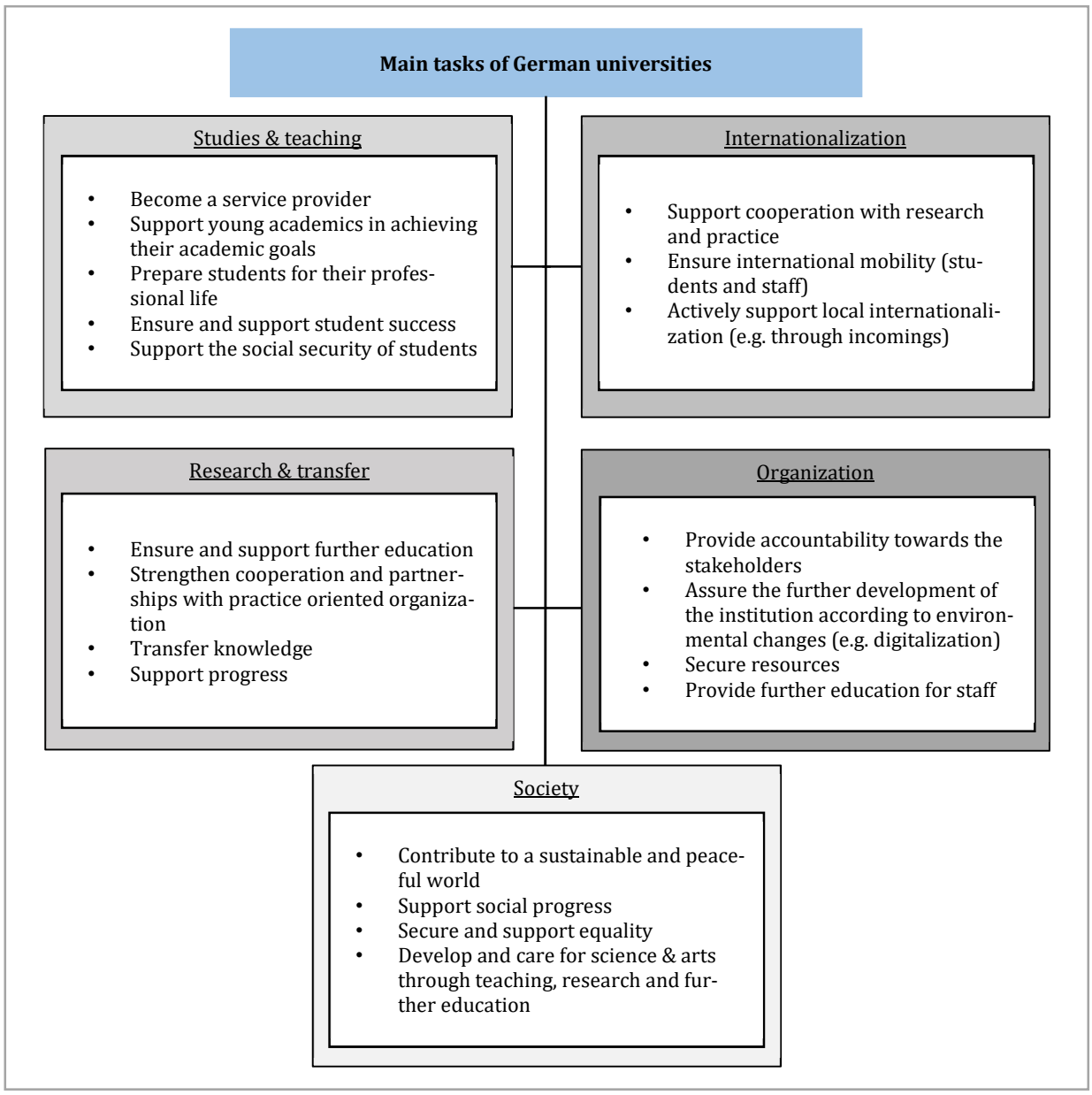


Figure 6. Main tasks of German universities according to the State University Laws.

Tasks in studies and teaching

- Universities must become service providers that advise and support students in their daily challenges.
- Universities have the task to ensure and support the success of their students during and after graduation, which goes hand in hand with their role as a service provider. They must ensure that the majority of the student body successfully completes their studies and continues the desired career.
- Universities are responsible to prepare students for their professional life and future occupation. Besides providing fundamental knowledge in the disciplines relevant to their study program, this inter alia includes the development of social skills, the provision of networking possibilities and the option to gain practical and international experiences. In addition, they

have to help their graduates by providing contacts and networks that support their professional success.

- Universities should provide social assistance to their students by providing support for students with disabilities, children, or relatives in need of special support as well as students with a different educational background.[51] In addition, universities should support the cultural and athletic interests of their students.
- Universities should help young scientists to achieve their goals by giving them guidance and the opportunity to conduct research. Furthermore, they should support Ph.D. students to succeed in their academic careers.

Tasks in research and transfer

- Universities are responsible for promoting and providing continuing education and lifelong learning by offering relevant programs and helping students of these programs succeed. This, in turn, ensures the transfer of knowledge from the educational institution into practice.
- Universities must cooperate nationally and internationally with academic and practice-oriented companies, research institutes and universities. Through this cooperation and partnerships, the development and transfer of knowledge are ensured, which promotes progress.
- Universities are responsible for actively supporting progress by making their research results accessible, reproducible, and usable. Practical institutions may use such results to improve their existing processes, and researchers can further develop ideas and theories.

Tasks for internationalization

- Universities have the task of actively supporting internationalization by enabling and actively seeking to collaborate with research and practice. This includes supporting and enabling staff and student mobility.
- Universities are responsible for local internationalization by offering international programs and easy access to incoming students and staff.

Tasks regarding the organization

- Universities must ensure accountability to the state and other stakeholders. This includes informing the public about the objectives of the institution, the existing courses and subjects, and the research results. In some federal states – e.g. Brandenburg and the region of Hamburg – this includes the presentation of a development plan that defines the goals of the universities and the plan to achieve these goals. Furthermore, some federal states – e.g. Saarland – support the introduction and development of a professional information system, containing data about the students, staff, facilities, and resources.

[51] For example: Students with a foreign background or without a classical HEEQ.

- Universities must ensure the quality of their research and teaching activities by regularly evaluating their programs and adapting them to the needs of society. In addition, they must ensure that the guidelines of good scientific practice are respected and that their processes are developed in line with environmental changes and needs – e.g. digitalization.
- Universities are tasked with securing their needed resources by managing, acting, and planning in an economically responsible way and striving for third-party funding.

Tasks in society

- Universities have the responsibility to contribute to a sustainable and peaceful world by following the guidelines of sustainability and acting ethically in all sectors within their reach.
- Universities are responsible for supporting Germany's social progress. This is achieved for example, by giving the citizens of the country the opportunity to improve their quality of life through education.
- Universities need to secure and support equality, which adds to the task of supporting social progress by ensuring that all citizens have equal access to the education and knowledge they offer. This includes ensuring gender equality, equality of people with disabilities, integration of ethnic minorities, and the implementation of a diversity management.
- Universities must cultivate and develop science and art. This is done through research, teaching, studies, and provision of further education opportunities that support lifelong learning.

2.2.4 Objectives of German universities

In addition to the tasks universities need to fulfill in accordance with the law, they define goals and 'business' objectives to orient themselves, their processes and their strategies (Gerhard 2004: 134), especially at a time when the management of universities gains more and more responsibility (Martinez 2009: 26). Therefore, management by objectives was introduced to universities during the NPM reform, which supported the development of goals among the various areas and committees of a university (Geis 2017: 116; Berthold 2011: 68; Kehm 2012: 18). As a result, target agreements were implemented in all the German federal states as a steering mechanism for university management (Lanzendorf et al. 2009: 18-23; König 2009: 33; Berthold 2011: 93). In these, the university formulates and agrees upon goals with the state. The responsibility in the implementation of these goals lies with the management of the universities, which independently plan actions and formulate strategies for achieving their goals (König 2009: 29). Furthermore, some universities have a university development plan or a structure development plan.[52] In both cases, the goals pursued by a university are usually presented in great detail.

[52] In German: *Strukturentwicklungsplan*.

In this section, the main objectives of universities have been extracted from these plans or the target agreements. If none was available, the goals were derived from the mission statements on the websites of the universities. Due to the fact that there are currently 125 state-funded universities of applied sciences in Germany (DESTATIS 2018a), the focus in analyzing the goals was on those located in southern Germany with a student body of a minimum of 3000 students. This distinction has been made because these applied sciences universities are the most comparable to the university selected for the case studies and are therefore its immediate competition. In total, 31 universities of applied sciences meet these criteria, of which 16 are located in the federal state of Bavaria and 15 in the federal state of Baden-Württemberg.

The main body of the Bavarian universities has publicly available target agreements with the state on the basis of which their current objectives have been extracted. In some cases – e.g. the Universities of Applied Sciences in Augsburg and Hof – the university even had a publicly available university development plan. According to the legal regulations, target agreements should also exist in the federal state of Baden-Württemberg. Unfortunately, these were not publicly available at the time of writing this study. Therefore, the objectives were extracted from information available on the websites or from the structural development or university development plans, if available. An overview of the surveyed universities, sorted to the source from which the required information has been extracted is shown in Table 7.

Table 7. Overview of the sources used to identify the main objectives of German universities.

Source of university objectives analyzed	Names of the universities
University development plan or structural development plan	HAW[53] München, FH[54] Augsburg, HAW Coburg, HAW Landshut, HAW Hof, HAW Aschaffenburg, HAW Heilbronn, HAW Pforzheim, HAW Reutlingen
Target agreement	TH[55] Nürnberg, OTH[56] Regensburg, HAW Würzburg-Schweinfurt, HAW Weihenstephan-Triesdorf, HAW Kempten, TH Deggendorf, HAW Rosenheim, TH Ingolstadt, HAW Neu-Ulm, OTH Amberg-Weiden
Website (e.g. mission statement)	HAW Karlsruhe, HAW Furtwangen, HAW Aalen, HAW Nürtingen-Geislingen, HAW Mannheim, HAW Offenburg, HAW Ulm, TH Stuttgart, HAW Ravensburg-Weingarten, HAW Albstadt-Sigmaringen
No information available	HAW Esslingen, HAW Konstanz

53 This is the abbreviation for *Hochschule für angewandte Wissenschaften*, which means university of applied sciences.

54 This is the abbreviation for *Fachhochschule*, which means university of applied sciences as well. Nowadays, more and more applied sciences universities are called HAW, because it does more directly indicate their applied sciences orientation. Formally, there is no difference between FH's and HAW's.

55 This is the abbreviation for *Technische Hochschule*, which can be used by universities of applied sciences that are clearly specialized on technological professions.

56 This is the abbreviation for *Ostbayersiche Technische Hochschule*, which are applied sciences universities specialized on technological study programs located in eastern Bavaria.

After the objectives were extracted from the information available at each university, these were grouped into the following four areas: (1) *studies, teaching, and further education*; (2) *research and transfer*; (3) *internationalization*, and (4) *organization*. In comparison to the previous section, Section 2.2.3, only these four instead of five areas have been identified because the objectives stated by the universities do not mention goals that are solely aimed on supporting the society. Nevertheless, it is obvious that the identified objectives set by the universities do support the society as well and therefore aim on achieving the tasks defined in the previous section regarding the society.

After the objectives of the universities were extracted, the naming of these goals by the universities was analyzed and counted to identify the goals most important to a variety of universities. The objectives that have been mentioned by at least 10 of the 29 universities surveyed are summarized in Table 8. The relevant matrix for the area *studies, teaching, and further education* is presented in Table 9. The analysis of the remaining operational areas can be found in Appendix B.

Table 8. Main objectives of the state universities in southern Germany.

Operational area	Main objectives
Studies, teaching, and further education	(1) Become a service provider which offers additional student support
	(2) Assure and improve the quality of studies, teaching and further education
	(3) Increase the student success rate
	(4) Extend further education programs
	(5) Extend existing study program and clarify their focus
Research and transfer	(6) Strengthen and increase practical cooperation and knowledge transfer
	(7) Strengthen and increase research cooperation
	(8) Gain more third-party funding and strengthen the research infrastructure
	(9) Support the development of research networks and synergies
Internationalization	(10) Increase international mobility of staff and students
	(11) Increase cooperation with all international partners
	(12) Increase the number of incomings
	(13) Extend existing international study programs and develop new ones
Organization	(14) Provide good working conditions (including: being family friendly and socially inclusive)
	(15) Optimize the administrative processes
	(16) Develop the diversity management further
	(17) Extend the administrative services
	(18) Improve communication in order to create transparency and individualize profiles

In total, 18 goals have been identified that are in the focus of more than 10 applied sciences universities in southern Germany. The goals with the highest representation are:

- Becoming a service provider that offers additional student support (named by 79%),
- Ensure and improve the quality of studies, teaching, and further education (named by 69%), and
- Strengthen and increase practical cooperation and knowledge transfer (named by 66%).

Table 9. Objectives in the area of *studies, teaching, and further education* of the applied sciences universities in southern Germany.

Studies, teaching, and further education											
Universities of applied sciences with more than 3000 students	Secure and increase student numbers	Increase the student success rate	Extend and focus study programs	Assurance and improvement of quality	Support diversity and equality	Extent further education study programs	Become a service provider	Develop new study methods	Secure good scientific practice	Educate and develop students personally and internationally	Support applicants with apprenticeships and alumnus
HAW Munich	●	●	●	●	●	●				●	
TH Nürnberg		●					●	●			
OTH Regensburg		●		●			●			●	
HAW Würzburg-Schweinfurt	●			●	●		●			●	
HAW Weihenstephan-Triesdorf	●	●				●			●		
HAW Kempten		●		●			●			●	
FH Augsburg			●	●			●				
TH Deggendorf	●	●		●							●
HAW Rosenheim		●		●			●				
HAW Coburg						●	●	●		●	
HAW Landshut			●							●	●
TH Ingolstadt		●				●	●				
HAW Neu-Ulm	●	●		●			●	●			●
HAW Hof				●		●	●	●		●	●
HAW Aschaffenburg			●	●			●				
OTH Amberg-Weiden		●		●	●					●	●
HAW Karlsruhe			●	●		●	●				
HAW Heilbronn	●		●	●		●	●			●	●
HAW Furtwangen			●	●		●	●	●		●	
HAW Pforzheim											
HAW Aalen		●	●		●	●	●			●	

Studies, teaching, and further education											
Universities of applied sciences with more than 3000 students	Secure and increase student numbers	Increase the student success rate	Extend and focus study programs	Assurance and improvement of quality	Support diversity and equality	Extent further education study programs	Become a service provider	Develop new study methods	Secure good scientific practice	Educate and develop students personally and internationally	Support applicants with apprenticeships and alumnus
HAW Reutlingen				●		●	●	●		●	
HAW Nürtingen-Geislingen			●			●	●		●		●
HAW Mannheim				●		●	●				
HAW Offenburg			●	●	●		●			●	
HAW Ulm			●	●			●			●	
TH Stuttgart			●	●		●	●				
HAW Ravensburg-Weingarten		●		●			●				
HAW Albstadt-Sigmaringen		●	●			●					
Count (29)	6	13	13	20	5	14	22	6	2	8	6

3 Data Mining

The following section defines the methods used in the case studies in Chapter 5. First, the Cross-Industry Standard Process for Data Mining (CRISP-DM), developed by Chapman et al. in 2000, is explained because this is the process by which the case studies are conducted. It then describes the DM methods used in the case studies, including the data preparation steps and the methods and measures required to evaluate the generated DM models.

3.1 Data Mining Process

The CRISP-DM, shown in Figure 7 is an iterative process consisting of 6 phases (Chapman et al. 2000). The first phase of the process is *Business Understanding*, which aims to give direction to the DM project. Therefore, a business goal is clarified, which should be addressed with the knowledge generated from the DM project. From this a DM problem definition is derived. Subsequently, a preliminary strategy is presented to be followed in order to accomplish the pre-defined tasks, including the initial selection of tools and techniques suitable for the task achievement. The overall business objective of the research presented is to assist the management of German universities in addressing current environmental challenges and to help them secure their long-term success and existence. It is assumed that German universities are able to support their decision-making process with objective facts by analyzing their existing data resources. These additional insights can help universities stand out from their competition. The detailed DM objectives to which the two case studies relate are listed separately in Chapter 5.

In the second CRISP-DM phase – *Data Understanding* – the available and useful data resources are identified and collected. These data resources are explored and analyzed to gain a first understanding and identify limitation, such as data quality issues or lack of detail. In addition, this exploratory analysis gives an overview of the dataset structure and possible relationships and dependencies. If desired, interesting subsets can be selected for further analysis (Larose & Larose 2015: 7). As illustrated through the arrows between the first and the second phase in Figure 7, it may be necessary to redefine the analysis objectives in order to be able to reach them with the available data resources. Circumstances that could cause such a need are the lack of sufficient data or too many missing or inconsistent values in the available resources. Therefore, it may be necessary to gather more information before further analysis can be performed or even to redefine the goal of the DM project.

Once realistic analytical goals are formulated and the needed data resources are available to meet the analysis objectives, the third phase of the CRISP-DM, the *Data Preparation,* begins. In this phase,

all aspects of preparing the final and clean dataset from the original raw data are covered (Larose et al. 2015: 8). Accordingly, the records and attributes to be included in the model generation are selected, new attributes are generated, and erroneous, inconsistent, and missing values are treated. This phase also includes the detection and removal of outliers,[57] which could adversely affect the analysis results. In addition, attributes are transformed as needed to suit the DM method used. Many versions of artificial neural network (ANN) algorithms are examples of data mining methods that require a specific data format because they can only work with attributes that are numerically coded and have a value range between 0 and 1.

After the data resources have been prepared, the actual DM analysis begins in the phase *Modeling*. In this phase, the DM methods identified as suitable to address the task at hand are applied to a training dataset.[58] In order to find the best possible solution, several models are generated by calibrating the parameters of the applied methods or by using several different DM methods that are suitable for solving the problem. It might be required to return to the data preparation phase to

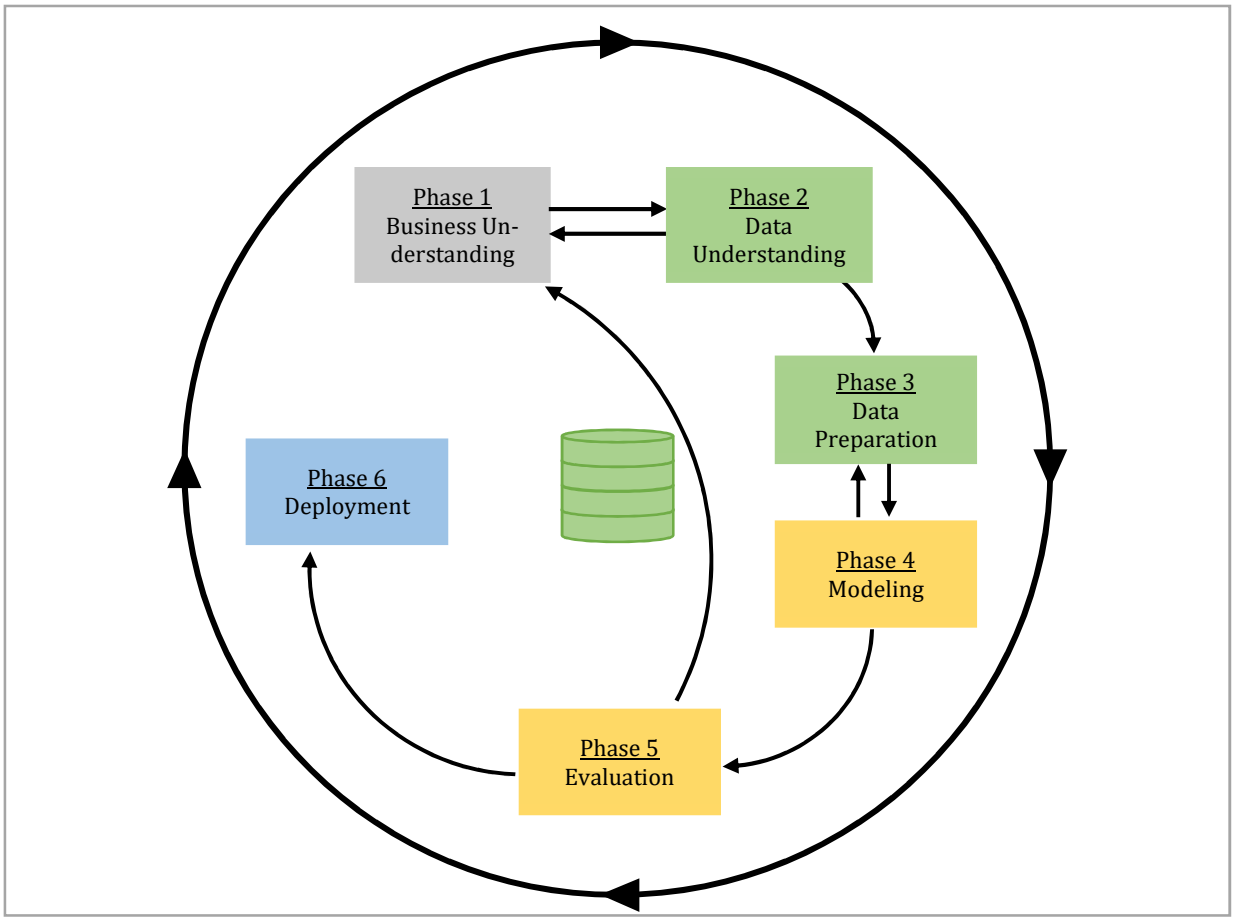


Figure 7. CRISP-DM after Chapman et al. (2000).

[57] According to Han et al. (2012) an outlier can be defined as "…a data object that deviates significantly from the rest of the objects, as if it were generated by a different mechanism."

[58] It is common to split the historic dataset, which is described here as training dataset, into two parts. The bigger portion of the dataset is used for model training and model validation. The remaining smaller portion is used afterwards to test the well-performing models in a real-life scenario.

tailor the form of the data to the requirements of each DM method (Larose et al. 2015: 8). The generated models are compared and evaluated with pre-defined performance measures. If it was not possible to generate a model with satisfactory performance results, it may be necessary to consider whether another preprocessing of the data can improve the model performance or if more data resources are available. If this is not the case, the necessary data must be collected before the desired insights can be generated.

The 5th phase of the CRISP-DM – *Evaluation* – identifies the model or models that best solve the pre-defined tasks. Accordingly, in addition to evaluating the mathematical performance measures, the degree of achievement of the task defined in the *Business Understanding* phase must be assessed. If it has not been possible to create a model with satisfactory performance metrics and insights that support the achievement of the pre-defined task, the CRISP-DM must be restarted. In addition, it is possible that the project plan needs to be adjusted or that the results indicate more DM analysis potentials that can or need to be addressed (Cleve & Lämmel 2016: 9). Once one or more models have been generated that perform well and support the achievement of the pre-defined task, the process continues with Phase 6 – *Deployment*. In this phase, the results are prepared and presented to the decision-makers and integrated into the institutional processes.

3.2 Data Preparation

As mentioned in the previous section, the datasets available for analysis can be incomplete, noisy, or inappropriate for the applicable DM methods. Therefore, the datasets must be processed and transformed by identifying and handling data inconsistencies and outliers. Missing values, data entry errors or unreliable attributes are typically identified in the *Data Understanding* phase and then treated by data cleansing that is, correction or deletion of inaccurate records or attributes. Afterwards, some attributes may need to be transformed to adapt to the needs of certain techniques of DM analysis. In addition, the balancing of datasets with an uneven distribution of the records between the categories of the target variable may be required.

The data preparation steps undertaken in the case studies in Chapter 5 are introduced in the following sections.

3.2.1 Handling missing values

The two main ways to fix missing values in a dataset are either deleting the affected attributes or records or filling in those missing values. The preferable method depends on the number of missing values in an attribute or a data record,[59] the total number of records in a dataset, and the type of DM issue. If circumstances permit, it is recommended to replace the missing values because this

[59] In this research, a data record is understood as one data case, example or object in the dataset.

way no data must be deleted (Larose et al. 2015: 22-25). Missing values can be filled simply by replacing them with the mean or median values of the attribute, by selecting a random value from the observed distribution of an attribute as the replacement, or by generating the missing values using an algorithm. The data imputation methods look for the most likely value, which could actually be the missing value, while the mean or median replacement fills any missing value with the mean or median of the attribute. In the presented thesis, the imputation of missing values is preferred because it is assumed that the heterogeneity of the dataset is obtained this way. It should be noted, however, that replacing missing values creates data points to close the empty areas in the dataset; these are no real-life observations (Larose et al. 2015: 23-25).

Many data imputation methods are available for mixed measures metrics, for example, the prediction of missing values with classification trees or k-nearest neighbor (k-NN) approaches. The imputation of missing values with the classification tree follows the steps described in Section 3.4.3. Accordingly, the attribute with the missing value is the target variable to be forecasted with a classification tree model based on the data records for which there are no missing values in the target variable. The k-NN approach replaces the missing value with the value of the data record that most closely resembles the data record with the missing value. Accordingly, the data record containing the missing value is compared with all the data records in the dataset to identify the one data record that is most similar in all available attributes. The missing value in one attribute is then replaced by the value that the attribute has in the most similar data record. This is the nearest neighbor (NN) data record. A popular distance measure for identifying the NN of a dataset is the *Euclidean distance*. The *Euclidean distance* between two records i and j, which are described by k numeric attributes is (Han et al. 2012: 72; Larose et al. 2015: 305):

$$d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ik} - x_{jk})^2}$$

where $i = (x_{i1}, x_{i2}, \ldots, x_{ik})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jk})$.

The following mathematical properties are satisfied by the *Euclidean distance* (Han et al. 2012: 72-73):

$d(i,j) \geq 0$: The distance between two data records is a non-negative number.

$d(i,i) = 0$: The distance of a data record to itself is 0.

$d(i,j) = d(j,i)$: The distance between two data records is a symmetric function.

$d(i,j) \leq d(i,h) + d(h,j)$: Going directly from record *i* to record *j* in space is no more distant than making a detour over any other record *h*.

For nominal attributes, the distance between an attribute in two data records is encoded with *0*, if the values of the attribute are the same. If the values are different, the distance between the data records is encoded with *1*.

3.2.2 Outlier detection

Outliers in a dataset usually represent extremes that are significantly different from the rest of the data records in the dataset. They can be valid data points or errors during data entry. For some DM tasks, including these extreme values in the analysis means that the outcome can be unreliable (Larose et al. 2015: 26). Outliers should therefore be identified and possibly excluded from modeling (Larose et al. 2015: 26). For the detection of outliers, many different approaches are available. The datasets that are important for this thesis contain numeric and nominal data types. Accordingly, an outlier detection method that can handle mixed data types is needed.

The RapidMiner program provides *density-based* and *distance-based outlier detection* methods that perform unsupervised outlier detection based on proximity. The *distance-based outlier detection* approaches identify outliers according to the distance between a data record i and its k-NN (Han et al. 2012: 552). Therefore, a data record is considered an outlier, based on the assumption that the distance of the outliers to their k-NN is significantly greater than the distance between the other data points and records in the dataset (Aggarwal 2013: 108). Consequently, if the distance is large and a data record is far from its k-NN, the record can be considered as an outlier.

Density-based outlier detection approaches take into account outliers according to their local neighborhood and therefore overcome difficulties that distance-based outlier detection methods may have on datasets containing clusters of different density (Han et al. 2012: 564). Han et al. (2012) discusses this problem using an image that is rebuilt in Figure 8, showing a dense cluster C_1 and a sparse cluster C_2. If the data records o_1 and o_2 were categorized as distance-based outliers, all the records in cluster C_2 would also be classified as outliers because the distance between the records o_1 and o_2 to their neighboring records in cluster C_1 is less than the average distance between the records in cluster C_2. Therefore, o_1 and o_2 cannot be identified as outliers by distance. When viewed locally with respect to the cluster C_1, the data records can be identified as outliers because they are different from the other records in that cluster and are also far away from the records in cluster C_2.

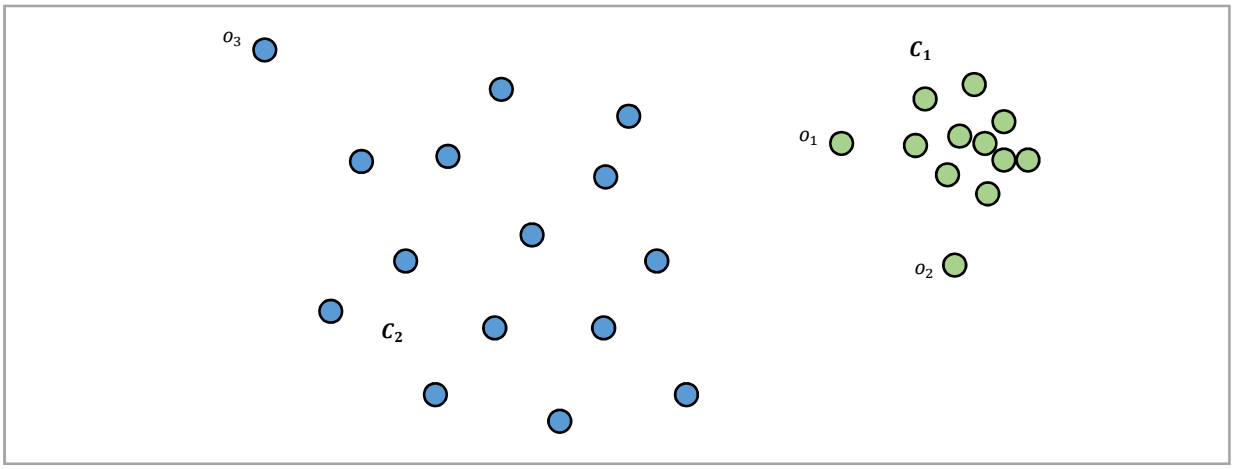


Figure 8. Difference in cluster densities based on Han et al. (2012: 564).

Density-based outlier detection methods regard a data record as an outlier when it is in a low-density region and is therefore a distance d from at least a n-part of all data records in the dataset. In addition, the density around the data record is compared with the density around the neighboring records (Han et al. 2012: 565). When the density around one particular data record is similar to the density around the neighboring records, then the data record is considered a non-outlier. Accordingly, when the density around a data record i is significantly lower than the density around its k-NN records, the data record i is considered as an outlier (Han et al. 2012: 565).

This study uses the *density-based outlier detection* approach *local outlier factor* (LOF), which takes into account the local neighborhood of a data record to identify outliers. Instead of the record's binary tag as an outlier *Outlier(Yes)* or a non-outlier *Outlier(No)*, this outlier detection method assigns a degree of being an outlier to each data record, which may be of particular importance for complex situations (Breunig et al. 2000). Hence, the LOF captures the degree of isolation of a data record with respect to its NN (Breunig et al. 2000). If a data record is deep in a cluster, the LOF value is approximately 1 (Breunig et al. 2000). Accordingly, LOF values that are significantly higher or lower than the LOF values of the other data records in the dataset are considered outliers.

The LOF is estimated by following the below-described steps (Han et al. 2012: 565-567; Breunig et al. 2000):

1. Definition of the *k-distance* of record i:
   The *k-distance* around a data record i ($i \in D$), denoted as $dist_k(i)$, is defined as the distance between the data record i and its k-NN, for each k. Accordingly, the $dist_k(i)$ is the distance between i and another record j in the dataset D, $dist(i,j)$, $j \in D$, such that:
   - There are at least k records $i' \in D\backslash\{i\}$ such that $dist(i,i') \leq dist(i,j)$.
   - There are at most $k-1$ records $i'' \in D\backslash\{i\}$ such that $dist(i,i'') < dist(i,j)$.

2. Definition of the *k-distance neighborhood* of the data record i:

   Given the $dist_k(i)$, the *k-distance neighborhood* of i, $N_{dist_k}(i)$, contains every record in the dataset D whose distance from i is not greater than the $dist_k(i)$:

   $$N_{dist_k}(i) = \{i' | i' \in D, dist(i,i') \leq dist_k(i)\}$$

   where the records i' are called the k-NN of i.

3. Definition of the *reachability distance*:
   The reachability distance between two records is defined in order to generate stable results and to overcome an undesirably high fluctuation of the distance measure. This smoothing effect is achieved by defining the number of k-neighbors that are considered the minimum neighborhood of the data record i.

The *reachability distance* for two records, i and j is denoted by:

$$reachdist_k(i,j) = max\{dist_k(i), dist(i,j)\}$$

Accordingly, the *reachability distance* is either the exact distance between i and j when $dist(i,j) > dist_k(i)$ or otherwise $dist_k(i)$. This concept is shown in Figure 9 assuming of $k = 4$. Accordingly, the distance between i and j_1 is smaller than the *k-distance* of i. Therefore, the $dist(i,j_1)$ is replaced by the $dist_k(i)$. The distance between i and j_2 is greater than the *k-distance* of i. Therefore, the *reachability distance* is the actual distance between these two records. Accordingly, the distance between a record i and the records near i is replaced by the $dist_k(i)$, thereby significantly reducing the statistical variation of the distance for all records j close to i. By specifying the parameter k, this smoothing effect can be controlled. Accordingly, a higher value of k increases the similarity of *reachability distance* of records in the same neighborhood.

It has to be noticed that the *reachability distance* function is not a mathematical distance function, because it is not symmetric and therefore $reachdist_k(i \leftarrow j) \neq reachdist_k(j \leftarrow i)$.

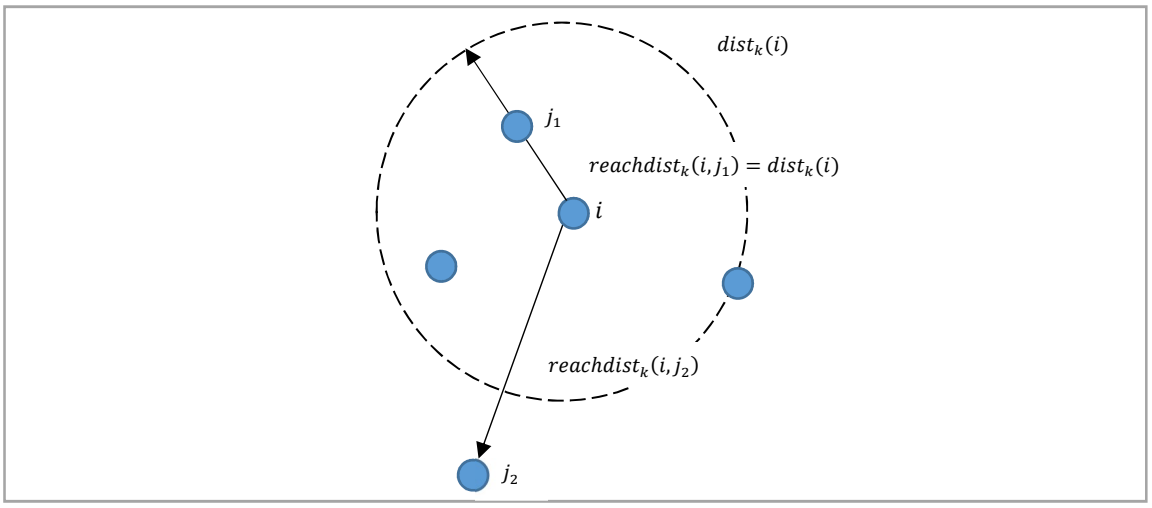


Figure 9. Concept of *reachability distance* for *k = 4* based on Breunig et al. (2000).

4. Definition of the *local reachability density* of record i:

Normally, in density-based clustering, two parameters specify the notion of density: a parameter *MinPts* that specifies a minimum number of records to be considered and a parameter volume. To identify density-based outliers, the density between sets of records is compared. Therefore, only the *MinPts* parameter is included. The volume of the density in the neighborhood of an record i is determined by $reachdist_{MinPts}(i,j)$, for $j \in N_{MinPts}(i)$. Therefore, the *local reachability density* (*lrd*) of a record i is defined as:

$$lrd_{MinPts}(i) = 1/\left(\frac{\sum_{j \in N_{MinPts}(i)} reachdist_{MinPts}(i,j)}{|N_{MinPts}(i)|}\right)$$

Thus, Breunig et al. (2000) indicates the local *reachability distance* of an record i as the inverse of the average *reachability distance* based on the *MinPts*-NN of i.

5. The *local outlier factor* is calculated:

   The LOF is the average of the ratio of the *local reachability density* of i and that of the *MinPts*-NN (or k-NN) of i (Breunig et al. 2000). The notation of the LOF below shows that the lower the $lrd_{MinPts}(i)$ and the higher the *lrd* of i's k-NN, the higher the *LOF* value of i.

$$LOF_{MinPts}(i) = \frac{\sum_{j \in N_{MinPts}(i)} \frac{lrd_{MinPts}(j)}{lrd_{MinPts}(i)}}{|N_{MinPts}(i)|}$$

   In other words, if the local density around a record i is comparably low to the local density around the k-NN records of i, i is considered an outlier. In general, a LOF near 1 shows a great similarity between the density of a record and the density of its NN.

3.2.3 Feature selection

To address a particular DM problem, not all the attributes available in a dataset may be relevant. Accordingly, the application of attribute subset selection methods may be useful to identify the relevant attributes for a particular analysis objective. In addition, it may be necessary to reduce the size of the dataset to reduces the computation time and complexity of the analysis results (Han et al. 2012: 104). The attributes that could most benefit the analysis are usually determined by testing the significance of the attribute against the target variable. Therefore, each attribute is considered as independent and tested accordingly. There are various methods for feature selection. This research uses a *stepwise forward selection* and a *Chi-squared* based attribute filtering approach.

In the *stepwise forward selection*, a model is trained with a learning algorithm, e.g. decision tree or logistic regression. Starting from an empty selection of attributes, the attributes of a given dataset are included individually in the model (Han et al. 2012: 105). Accordingly, each attribute is used to build a model and then the performance of the model is evaluated, e.g. by *cross-validation*.[60] The attribute leading to the highest increase in model performance is then added to the attribute selection. A new analysis round is started in which the remaining attributes are successively added to the previously defined selection to identify the attribute that achieves the highest performance enhancement (Krzysztof et al. 2007: 225). This attribute is then added to the attribute selection as well before starting a new round of analysis with the remaining attributes. The described procedure is repeated until a predefined stop criterion is met, which indicates how much performance improvement is desired by adding an additional attribute to the model each round. The result is an

[60] The principles of *cross-validation* are explained in Chapter 3.5.

attribute selection that contains those attributes that significantly enhance the performance of the DM model.

The *Chi-square* based filtering approach to feature selection was chosen as the second feature selection method because the datasets available for the case studies mainly contain nominal attributes. The *Chi-square* feature selection approach can identify associations between these nominal attributes (Han et al. 2012: 95). Nevertheless, when using this attribute selection method, the numeric attributes in the datasets must also be converted into nominal attributes. This can be done by *discretization* (Krzysztof et al. 2007: 235-237). By *discretizing*, the numerical expressions of an attribute can be combined into a certain number of categorical attribute expressions. In RapidMiner, these bins can either be built by frequency so that the number of unique values in all bins is close to equal, user-specific size, minimizing the entropy in the classes, or by user-specific classes.

After the dataset has been preprocessed, the *Chi-square* test counts the occurrence of an attribute *A* with *c* distinct categories, namely $\{a_1, a_2, a_m, \dots, a_c\}$, with respect to the attribute *B* with *f* distinct categories, namely $\{b_1, b_2, b_n, \dots b_f\}$. This is done by creating a contingency table between the attribute *A* and the attribute *B*. With the information regarding the occurrence, the expected frequencies e_{mn} of the joint event (A_m,B_n) are calculated (Han et al. 2012: 95):

$$e_{mn} = \frac{count(A = a_m) * count(B = b_n)}{r}$$

where r is the number of data records, $count(A = a_m)$ is the number of records with the category a_m for the attribute *A*, and $count(B = b_n)$ is the number of records with the category b_n for the attribute *B*. These expected frequencies are then used to calculate the *Chi-square* χ^2(Han et al. 2012: 95):

$$\chi^2 = \sum_{m=1}^{c} \sum_{n=1}^{f} \frac{(o_{mn} - e_{mn})^2}{e_{mn}}$$

where o_{mn} is the observed frequency (i.e., the actual count) of the joint event $A_m = a_m$ and $B_n = b_n$. The χ^2 tests the hypothesis that there is no correlation between the attributes *A* and *B*. If this hypothesis can be approved, then the two attributes are independent. Conversely, if the hypothesis can be rejected, *A* and *B* are statistically correlated (Han et al. 2012: 95). In this case, if *B* is the target variable the attribute *A* would be included as a feature in the model generated to predict the values of attribute *B*.

3.2.4 Balancing datasets

The main objective of classification problems is, to predict the class to which a particular data record belongs. The starting point is a training dataset, in which the target classes for the data records

are already known. On this dataset, a model is trained that can predict the class attribute for new and unseen data records. In order for a classification algorithm to learn a powerful predictive model, it is helpful to have a dataset of training records evenly distributed among the target classes (Han et al. 2012: 384). A dataset is considered as unbalanced if one of the target classes is under-represented. This imbalance can adversely affect the classification model as only the over-represented class can be well predicted (Han et al. 2012: 384). To avoid this negative effect, the training dataset can be adjusted so that the target classes are evenly distributed. The adjustment of a dataset can be achieved by *undersampling* or *oversampling* (Han et al. 2012: 384).

Undersampling randomly selects a certain number of data records belonging to the over-represented class to match the number of records in the under-represented class (Han et al. 2012: 384). The new dataset, which has an even number of records in both classes, is smaller than the original dataset because some records of the over-represented class are excluded from the downsampled dataset. The goal of the *oversampling* process is to create new data records for the under-represented class to match the number of records to the over-represented class (Han et al. 2012: 384). This procedure results in a dataset that contains more than the original records. Normally, more data records are expected to result in more accurate classification models. However, it should be noted that oversampling creates data records that are either a copy of the original records (*random oversampling*) or 'synthetic' records that are not real-life observations. Accordingly, the models created with the oversampled dataset must be thoroughly evaluated and tested against unseen data records to prove their applicability to the real-life problem.

In the presented research, *oversampling* is performed with the *synthetic minority oversampling technique* (SMOTE), which creates data records in a feature space and not in the data space (Chawla et al. 2002). According to Chawla et al. (2002), this *oversampling* of the under-represented class is performed by "…taking each minority class sample and introducing synthetic examples along the line segments joining one/all *k* minority class NN." Accordingly, the new 'synthetic' data record i_{new} is computed by calculating the difference between the data record i from the minority class of the dataset D and the nearest neighbor j or k-NN j' of this data record, multiplied by a random number δ, which lies between [0,1]. The results of this multiplication are then added to the existing data record i and a new data record i_{new} is created. This method of creating 'synthetic' data records can be referred to as (Zheng, Zhuoyuan & Ye Li 2015):

$$i_{new} = i + (j - i) * \delta$$

The SMOTE oversampling process avoids the possibility of overfitting, which can be the result of *random oversampling*, whereby new data records are created by copying existing data records

(Zheng, Zhuoyuan et al. 2015). Therefore, some records in the under-represented class are represented several times in the training dataset, which may facilitate overfitting (Zheng, Zhuoyuan et al. 2015).

3.3 Frequent Itemsets and Association Rules

Association analysis looks for recurring relationships and connections between the elements in a given dataset (Han et al. 2012: 243-244). These elements are the attributes that describe each data record in a dataset. In the example of a supermarket, the data records are transactions and the items are products that are bought in the individual transactions. Accordingly, each transaction T consists of an itemset. When a combination or set of items occurs in a large number of transactions, these are considered to be frequent itemsets (Larose et al. 2015: 606). From these frequent itemsets, rules can be extracted that describe 'universal' relationships between the items in a given dataset. Such a rule, denoted by $A \Rightarrow B$, could mean that when a product *A* is purchased in a transaction, product *B* is also frequently bought, leaving room to assume that the purchase of the product *A* is likely to result in increases in the purchase of product *B* (Cleve et al. 2016: 63-64).

In DM, association rules are generated in two main steps, which are described below (Han et al. 2012: 246-247):

1. Find all frequent itemsets:
   Let D be a transaction dataset that contains an itemset $I = \{I_1, I_2, \dots I_m\}$ and T transactions that are non-empty itemsets, that $T \subseteq I$. Each transaction T in the dataset has an individual transaction-ID *T-ID*. If *A* is a set of items, then a transaction T will contain *A* if $A \subseteq T$. Accordingly, if *B* is a set of items, a transaction T contains *B* if $B \subseteq T$. With the measure of interestingness *Support*, the percentage of transactions in D is estimated, which contain *A* and *B*, where $A \subset I, B \subset I, A \neq \emptyset,$ and $B \neq \emptyset$. This can be denoted as:

   $$Support\ (A \Rightarrow B) = \frac{Numbers\ of\ transactions\ containing\ (A \cup B)}{Total\ number\ of\ transactions}$$

   A low *Support* count for itemsets indicates that a connection appearing between the items in the dataset is rare. Accordingly, there are most likely no dependencies between the attributes. Whether an itemset appears frequently in a dataset D, this is defined by a predefined minimum *Support* measure (*minsup*). This threshold must be set by the analyst, under consideration of the analysis problem and the available dataset.

2. Generate strong association rules from the frequent itemsets:
   After recognizing the frequent itemsets, rules are created, taking into account the second interestingness measure *Confidence*. The *Confidence* indicates the percentage of transactions in

the dataset D that contains A, which is the rule premises that also contain B, which is the conclusion of the rule:

$$Confidence\ (A \Rightarrow B) = \frac{Support\ (A \Rightarrow B)}{Support\ (A)}$$

A strong association rule will exceed both a predefined *minsup* and a predefined minimum *Confidence* measure (*minconf*). Both thresholds must be set by the analyst.

Several algorithms are available that generate association rules after the minimum thresholds of the interestingness measures have been set. In the presented studies, the *Frequent Pattern Growth* (FP-Growth) approach is used for mining frequent itemsets. The FP-Growth creates a frequent pattern tree to identify frequent itemsets. The first step is to scan the database and generate the number of *support count* of each item. Accordingly, 1-itemsets are created, which are then sorted descending according to their *support count* in the dataset *D*, resulting in a list *L* (Han et al. 2012: 257). Thereafter, the database *D* is scanned a second time and the items in each transaction *T* are sorted by the order in *L* (Han et al. 2012: 257). This is further illustrated by the below example.

Suppose the list *L* extracted from the first scan of the database is as shown in Table 10 and the dataset *D* contains the following transaction $T - ID(10)$: I_1, I_2, I_4. Accordingly, during the second scan of the dataset, the items in the example transaction $ID(10)$ are rearranged in the following order: I_2, I_1, I_4.

Table 10. Example of List *L* created after generating *1-itemsets* from a database *D*.

Item ID	Support count
I_2	7
I_1	6
I_3	6
I_4	2

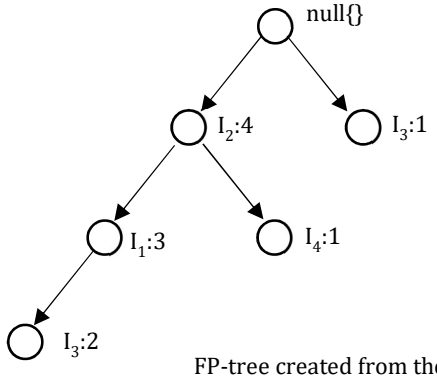
After all transactions *T* have been sorted to *L*, the tree branches are created, starting from the tree root, marked "null". Each transaction builds a branch of the tree, but if transactions use a common prefix, e.g. the combination of I_2 and I_1, then the number of the existing nodes is increased instead of creating a completely new tree root (Han et al. 2012: 257-259), as shown in Figure 10. After all transactions in the dataset are scanned and added to the tree, the FP tree is mined.

Starting with the last item in *L*, the conditional pattern base of each item is created. This is a subdataset that contains all the paths in the tree that lead to the examined item. The conditional pattern base is then used as a "new" transaction dataset, from which a conditional FP tree is generated. From this conditional tree, all combinations of frequent patterns in which the item is displayed are then extracted and further considered if they exceed the predefined *minsup*.[61]

[61] For an even more detailed presentation of the FP-Growth process see Han et al. (2012).

ID	Itemset
1	I_2, I_1
2	I_2, I_1, I_3
3	I_3
4	I_2, I_1, I_3
5	I_2, I_4

Transactions *T* in a dataset *D* sorted to the order in *L*.



FP-tree created from the table on the left.

Figure 10. Example FP-tree.

The identified frequent patterns are then converted into association rules. Accordingly, each frequent itemset is divided into a rule premises and a rule conclusion, also known as rule body and rule head. This can be illustrated in the following example: From a dataset *D*, the frequent itemset A={I_2, I_3, I_6, I_9} was extracted, which exceeds the value for *minsup*. Three rules ($I_2 \Rightarrow I_3, I_6, I_9$), ($I_2, I_3 \Rightarrow I_6, I_9$), or ($I_2, I_3, I_6 \Rightarrow I_9$) could be formed from this itemset. Which rules are considered depends on the *Confidence* of each of the three rule. If the percentage of transactions in *D* that contain the first part of the rule and also contain the second part of the rule exceeds the predefined *minconf*, then the rule is considered as interesting. Therefore, according to the *minconf*, either all of the three rules could be considered interesting, just one, or none at all.

After interesting rules have been drawn up, the results should be further investigated because not every strong rule may be objectively interesting, and the interestingness measures *Support* and *Confidence* do not measure the correlation or implication between the rule premises and rule conclusion (Han et al. 2012: 265-266). A widely used measure of the correlation in association rules is the *Lift*, which evaluates whether the occurrence of the rule premises is independent of the occurrence of the rule conclusion (Han et al. 2012: 266). When both are dependent, the events are correlated, meaning that the rule premises influences the presence or absence of the rule conclusion. The *Lift* is calculated as (Han et al. 2012: 266):

$$Lift\ (A \Rightarrow B) = \frac{Confidence\ (A \Rightarrow B)}{Support\ (A) * Support\ (B)}$$

The *Lift* can be interpreted as follows (Han et al. 2012: 266):

- *Lift = 1*: *A* and *B* are uncorrelated and therefore independent of each other.
- *Lift > 1*: *A* and *B* are positively correlated and the occurrence of one implies the occurrence of the other.
- *Lift < 1*: *A* and *B* are negatively correlated and the occurrence of one probably leads to the absence of the other.

3.4 Classification Analysis

In general, prediction algorithms aim to forecast a particular target variable. For this purpose, they work in two main steps as shown in Figure 11. First, a model is calculated, with which a defined target variable is to be predicted. Therefore, the model attempts to mimic the reality of a historical dataset. Accordingly, the historical dataset must contain the target variable and attributes that can be used to train a model. After training a well-performing predictive model, which is validated by comparing the actual value of the target variable with the predicted value, the analyst can estimate how well the model can depict the reality. In addition, the model should be tested on a testing dataset that also contains the target variable to assess the model performance in a real-life scenario. If the performance results are convincing, the model learned from the historical dataset, also referred to as training dataset, can be applied to the new dataset, and the target variable for the records in the new dataset predicted.

There are several methods available to create classification models. The ones used in the case studies in Chapter 5 are described in the following sections.

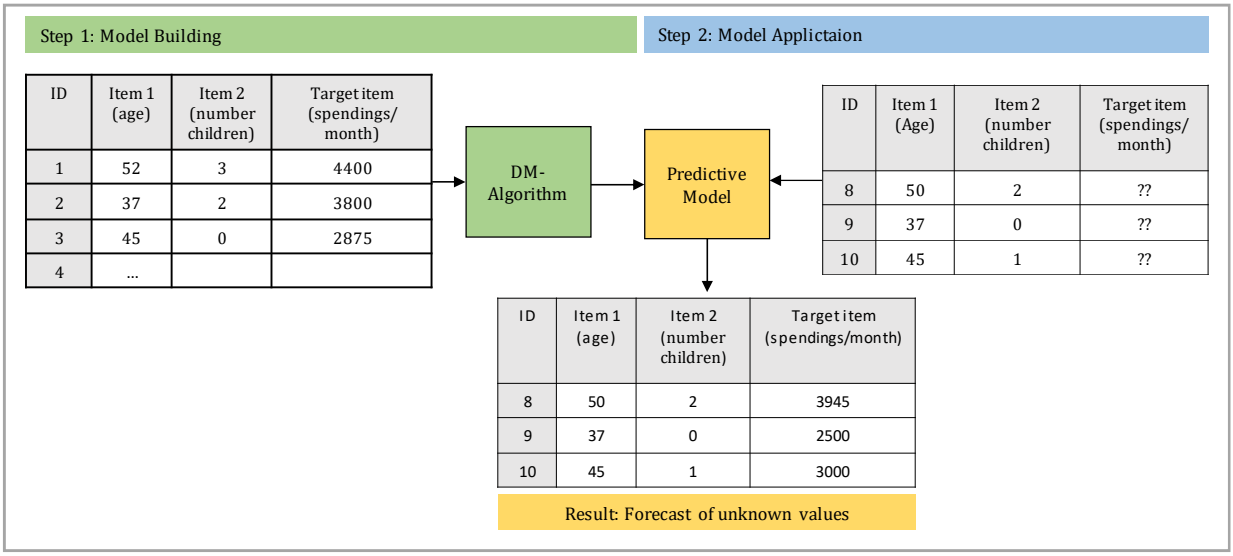


Figure 11. Steps for building and applying a predictive DM model, illustrated on a fictitious example.

3.4.1 Rule Induction

Rule-based classification generates IF-THEN rules that represent the learned model. The IF part of the rule is known as a precondition or antecedent to the rule followed by the THEN part, which is the rule consequent (Han et al. 2012: 355). This consequent of a rule is the class prediction for the target variable. A rule model can either be extracted from a decision tree or from a sequential covering algorithm. In the first approach, a decision tree must be computed and the branches of the tree are then converted into IF-THEN rules, while the second method directly enables the extraction of rules from a training dataset (Han et al. 2012: 359; Kotu & Deshpande 2015: 88).

As the name of the method implies, the rules are generated sequentially, which means that the rules are created one at a time. This process of rule creation is shown in Figure 12.

Sequential Covering Algorithm. Learn a set of IF-THEN rules for classification.

Input:

- *D*, a dataset with class-labelled records;
- *c*, class labels of the records in *D*;
- *Att_vals*, the set of all attributes and their possible values.

Output: A set of IF-THEN rules.

Method:

```
(1)   Rule_set = {}; // initial empty rule set
(2)   for each class c do
(3)     repeat
(4)       Rule = Learn_One_Rule (D, Att-vals, c);
(5)       remove tuples covered by Rule from D;
(6)     Rule_set = Rule_set + Rule; //add new rule to rule set
(7)     until terminating condition;
(8)   endfor
return Rule_set.
```

Figure 12. Sequential covering algorithm for rule learning (Han, Kamber & Pei 2012: 359-360).

The input of the algorithm is a dataset *D* containing class-labeled data records and several attributes describing the labeled records. One class of the target variable *Y* is selected, which can be a binominal target variable with the distinction *Y(yes)* and *Y(no)*, for which the rules are generated first. Usually, this is the least frequently available target class label in the dataset (Kotu et al. 2015: 91). Then the rules R_m are created with the aim of covering all records in a dataset of this first target class, without including any or as few as possible data records of the second target class.

The rules are learned starting with an empty rule set to which conjunctions are added one after another to increase the *accuracy* of the rule (Han et al. 2012: 356):

$$Rule\ accuracy\ (R_m) = \frac{correct\ records\ covered\ by\ the\ rule}{all\ records\ covered\ by\ the\ rule}$$

By itself, *accuracy* is not considered a reliable estimate for the rule quality as high *accuracy* does not mean that many examples in a dataset are covered and the rule is therefore transferable to new datasets (Han et al. 2012: 361). Therefore, measures that integrate aspects of *accuracy* and coverage are preferred for the evaluation of rule quality (Han et al. 2012: 361). The RapidMiner *Rule Induction* operator implements the *Repeated Incremental Pruning to Produce Error Reduction* (RIPPER) algorithm developed by Cohen (1995), which uses the *information gain* measure proposed in the *First Oder Inductive Learner* (FOIL) to evaluate the quality of rules (Cohen 1995; Han et al. 2012: 363).

The information gain of the FOIL is denoted by (Han et al. 2012: 362):

$$FOIL\_Gain = pos' \left[log_2 \left(\frac{pos'}{pos' + neg')} \right) - log_2 \left(\frac{pos}{pos + neg)} \right) \right]$$

where *pos* (*neg*) is the number of positive (negative) tuples a rule *R* holds, pos' (neg') is the number of positive (negative) tuples *R'* holds, and *R'* is the extension of rule *R* (Han et al. 2012: 362). The tuples designated as positive are those belonging to the target class for which the rules are learned while the tuples that belong to the remaining class or classes are said to be negative (Han et al. 2012: 362).

The conjunctions to a rule are added one by one until the *information gain* of a rule reaches its maximum (Cohen 1995). If a new conjunction does not increase the informational content of a rule, the algorithm looks for other conjunctions or stops and starts the iteration of the next rule (Kotu et al. 2015: 93). After a rule is created, the records covered by the rule are removed from dataset *D*, before a new rule is created. After a ruleset is generated that describes all the data records that belong to the selected target class, the rule model is evaluated to determine the need for pruning. To do this, the *FOILprune* measure given for a rule *R* is calculated (Kotu et al. 2015):

$$FOILprune(R) = \frac{pos - neg}{pos + neg}$$

where *pos* is the number of positive objects covered by a rule and *neg* is the number of negative objects covered by the rule. Pruning is done by removing the conjunctions one at a time, taking into account the *FOILprune* value, starting with the most recently added conjunction (Han et al. 2012: 363). Thus, if the *FOILprune* for the pruned version of *R* is higher, pruning is performed. This procedure is necessary because it can counteract overfitting, which can cause the rule to work well in the training dataset but less well on new and unseen data records. Once pruning does nothing further, the rules generated for the first class in the target variable are grouped into a ruleset. For DM problems with two classes, all data objects that are not covered by this rule set are predicted to belong to the second class (Kotu et al. 2015: 94). For multi-class problems, the above rule generation steps are repeated for the next target class (Kotu et al. 2015: 94).

3.4.2 Binary Logistic Regression

A binominal logistic regression model describes the relationship between a set of predictor variables and a binominal target variable (Larose et al. 2015: 359). This relationship is not linear, and the probability of a specific target class being true for a data record – e.g. the probability that an applicant enrolling $p(y = 1)$ – is estimated at known values of the predictor attributes X_m. To predict the probability of the event $y = 1$, the logistic regression function assumes the existence of a latent variable *Z* estimated for a data record *k* as (Backhaus et al. 2011: 254-275):

$$z_k = b_0 + \sum_{m=1}^{M} b_m * x_{mk} + u_k$$

where $k = 1,2, \dots K$, b_0 is the regression constant, b_m is the logistic regression coefficient for x_m, u_k is the error term, and z_k is the latent variable for the case *k*, also referred to as *Logit.* Estimating the latent variable *Z* models the connection between the target variable and the predictor attributes. It is implied that *Z* is generated by a linear combination of the predictor variables X_m and therefore is the aggregated impact strength of the independent variables on the event $y = 1$. To estimate the probability *p* of an event, a logistic probability function is required and the logistic regression function for a case *k* can be defined as (Backhaus et al. 2011: 255):

$$p_k(y = 1) = \frac{1}{1 + e^{-z_k}}$$

where *e* is the *Euler number*. Accordingly, the goal of the logistic regression function is to correctly predict the class of the target variable. This is done by predicting the probability *p*, which is a value between *0* and *1*. Normally, an observed record is assigned to the target class $y = 1$ if $p_k > 0.5$ and the target class $y = 0$ if $p_k < 0.5$, because a logistic regression function is always symmetric about the turning point 0.5 as shown in Figure 13. To predict p_k for each data record *k*, the following relation is taken into account (Backhaus et al. 2011: 259):

$$p_k(y) = \begin{cases} \left(\frac{1}{1 + e^{-z_k}}\right) & for\ y_k = 1 \\ \left(1 - \frac{1}{1 + e^{-z_k}}\right) & for\ y_k = 0 \end{cases}$$

This expression can be combined in the following equation (Backhaus et al. 2011: 259):

$$p_k(y) = \underbrace{\left(\frac{1}{1 + e^{-z_k}}\right)^{y_k}}_{Factor\ A} * \underbrace{\left(1 - \frac{1}{1 + e^{-z_k}}\right)^{1-y_k}}_{Factor\ B}$$

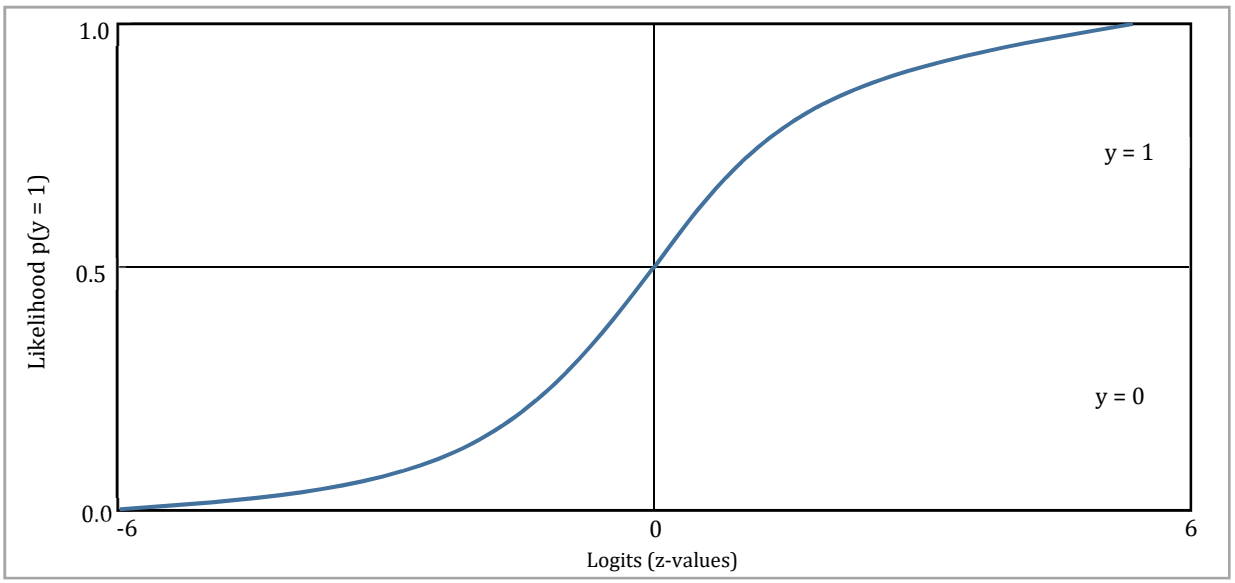


Figure 13. Logistic regression function.

According to the manifestation of the observation y_k for a particular record k, either *Factor A* or *Factor B* is equal to *1* in the above equation. Thus, the main task of the logistic regression function is to estimate the regression coefficients b, describing each predictor attribute in the logistic regression equation so that each data record k in the training dataset is correctly estimated.

The parameters b are usually estimated using the *Maximum Likelihood Estimation*, which aims to determine the coefficients in a way that the likelihood of the model correctly guessing the target class of the data records in the training dataset is maximized (Backhaus et al. 2011: 258-259). The regarding *Likelihood* function is based on the assumption that the probability that independent events, which are the observed values of the target variable, will occur together is estimated by multiplying the individual events (Backhaus et al. 2011: 256). As a result, the *Likelihood* function L maximizes the product ($\prod$) of the probabilities $p_k(y)$ for all cases $k = 1, \dots, K$ in a dataset (Backhaus et al. 2011: 259):

$$L = \prod_{k=1}^{K} \left(\frac{1}{1+e^{-z_k}}\right)^{y_k} * \left(1 - \frac{1}{1+e^{-z_k}}\right)^{1-y_k} \rightarrow max!$$

To simplify the maximization problem, the natural logarithm (ln) of the *Likelihood* function is considered instead of the product ($\prod$) expressed in the *Log-Likelihood-Function* (LL) (Backhaus et al. 2011: 259):

$$LL = \sum_{k=1}^{K} \left[y_k * ln\left(\frac{1}{1+e^{-z_k}}\right)\right] + \left[(1-y_k) * ln\left(1 - \frac{1}{1+e^{-z_k}}\right)\right]$$

The process to determine the best possible logistic regression model aims on minimizing LL statistics by conducting the following steps (Backhaus et al. 2011: 260; Field 2013: 763):

1. Estimates are taken for the regression coefficients b, which can be determined, for example, by a least squares estimation.
2. On the basis of the estimated regression coefficients, the *Logit z* is calculated for a random case k.
3. After the *Logit* has been estimated, the likelihood of $p_k(y = 1)$ is defined.
4. For this case k, the LL-value is estimated.
5. Steps 2 – 4 are repeated for all cases in the dataset to define the total LL statistic.
6. Steps 2 – 5 are repeated with different values for the regression coefficients.
7. The total LL statistics created using the different coefficients are compared. The regression coefficients are changed until the total LL statistics do not significantly improve.

After estimating the logistic regression model and testing the performance of the model,[62] it is desirable to interpret the coefficients in the model. Due to the nonlinear nature of the analysis, the

[62] Please see Chapter 3.5.

relationship between the target variable and the predictor variables is difficult to interpret. Accordingly, it is not obvious how much a predictor contributes to the target variable being $y = 1$ (Backhaus et al. 2011: 263). In addition, the logistic regression coefficients are not comparable with each other, e.g. a high b for a given predictor attribute x_m does not mean that this attribute has the highest impact on the target variable (Backhaus et al. 2011: 263). Only the direction of the influence a predictor variable has on the target variable can be derived from the logistic regression coefficients directly (Backhaus et al. 2011: 263). Accordingly, a positive regression coefficient ($b > 0$) indicates an increase of the probability of the target variable being $y = 1$ and a negative coefficient ($b < 0$) indicates a decrease of the probability of the target variable being $y = 1$ (Backhaus et al. 2011: 263-264).

In addition, the RapidMiner implementation of logistic regression estimates whether the regression coefficient of a given predictor deviates significantly from 0. If this is the case, it can be assumed that the predictor makes a significant contribution to the prediction of the outcome. This test is performed for logistic regression analysis using the *Wald statistic* (Backhaus et al. 2011: 280):

$$W = \left(\frac{b_m}{SE_{b_m}}\right)^2$$

where SE_{bm} is the *standard error SE* of b_m.

The strength of influence that a predictor attribute has on the target variable is more difficult to estimate but can be done using *Odds* and the *Odds Ratios*. The *Odds* of the target variable are the likelihood ratio that reflects the probability of the event occurring ($y = 1$) with respect to the non-occurrence of the event (Backhaus et al. 2011: 264-265):

$$Odds = \frac{p(y = 1)}{1 - p(y = 1)} = e^z$$

Afterwards, the *Odds Ratio* can be calculated. The *Odds Ratio* is an indicator of the change in odds that appears after a unit change in the predictor variable (Field 2013: 767):

$$OddsRatio = \frac{Odds\ after\ a\ unit\ change\ in\ the\ predictor}{Original\ Odds} = e^{b_m}$$

where the *Original Odds* represent the *Odds* of an event occurring before the unit change in the predictor variable. According to Field (2013), the *Odds Ratio* can be interpreted as follows: If the value of the *Odds Ratio* is greater than 1, this means that increasing the predictor variable will increase the odds of the event $y = 1$ happening as well. Backhaus et al. (2011) specifies the interpretation in more detail. So, if an independent variable x_m increases by one unit ($x_m + 1$), the *Odds Ratio* for the event $y = 1$ increases by the factor e^{b_m} (Backhaus et al. 2011: 265).

3.4.3 Classification Trees

Decision trees are inverted tree-like models that extend from a root node over a collection of decisions or internal nodes until they terminate in end nodes, also known as leaf nodes (Larose et al. 2015: 317). Each node represents an attribute that is tested and split with respect to the target variable. The goal of the split is to maximize the purity in the end nodes that contain the target variables and thus to maximize the *accuracy* of the predictive model. Ideally, the cases in the leaf nodes all belong to one target class and are homogeneous (Han et al. 2012: 330-331).

The basic steps to creating a decision tree with a top-down approach are as follows (Cleve et al. 2016: 92-93):

1. A root attribute *A* is selected, which is part of the set of all attributes *Attr_List* in the training dataset *D*, such that $A \in Attr\_List$ and v_A is the set of all values that the attribute *A* can take.
2. For each value $v \in v_A$ the attribute *A* can take, the subset D^{v_m} is formed that contains the data records that have the specific value v_m for the attribute *A*.
3. A branch is established on the node marked with the specific value v_m of the attribute *A*. If all the objects in D^{v_m} belong to the same class, the root terminates in a leaf node that is labeled with the respective class of the target variable. If the subset contains data records that belong to more than one class, a subtree is generated.
4. Steps 1 - 3 are repeated for the subtree but with the reduced attribute list $Attr\_List - A_{used}$ and the reduced example set $D - d_{classified}$, where $d_{classified}$ are the data records in the training dataset *D* already represented by the previous node.

To create a decision tree that is both compact and has high *accuracy* in predicting the target variable, the selection of the attributes that form the nodes in the decision tree is important. Accordingly, a procedure is needed that selects the attribute that "best" partitions the data records in respect to each class of the target variable (Han et al. 2012: 333). The method used for selecting the attributes depends on the algorithm used for modeling.

In the research presented, the well-known *C4.5* algorithm is used to generate decision trees. This method was developed by Quinlan (1992) as the successor to the *ID3* algorithm (Cleve et al. 2016: 107; Han et al. 2012: 332). For attribute selection, the *C4.5* algorithm uses *Shannon Entropy*, which is based on the concept of *information entropy* (Shannon 1948). Accordingly, this method uses knowledge from information theory and selects those attributes for a partition that provide the highest *information gain*. Accordingly, the first step is to define the information content of an attribute that can be measured as (Cleve et al. 2016: 99):

$$I(B) = \sum_{b=1}^{k} -p(v_b) * log_2(p(v_b))$$

where $I(B)$ is the information content, also called entropy, of the attribute B, v_b are the possible values of B, k is the number of categories of the attribute B, and $p(v_b)$ is the relative frequency for the occurrence of v_b in the training dataset in respect to the target variable. The information content is always specified for an attribute and a set of given data records. It is true that the higher the information content in a dataset, the higher the diversity of the data records in the attribute in respect to the target variable (Cleve et al. 2016: 99). So, if in a subset all records belong to one class label, the information content for that subset is *0* (Cleve et al. 2016: 99).

The attributes that are selected as the nodes of the decision tree are those that reduce the *entropy* in the given dataset the most and therefore provide the highest *information gain*. When a training dataset D is split on an attribute B that has v manifestations, the dataset D is divided into v subsets $\{D_1, \dots, D_v\}$. Subset D_i has the value i in attribute B. The remaining information entropy in dataset D after it has been partitioned is calculated as follows (Cleve et al. 2016: 100):

$$I_B(D) = \sum_{i=1}^{v} \frac{|D_i|}{|D|} * I(D_i)$$

where $I_B(D)$ represents the remaining *information entropy* in the training dataset after the dataset has been partitioned. The smaller this measure, the greater the purity of the partitions (Han et al. 2012: 337). The *information gain*, which indicates how much information would be obtained by splitting the training dataset on a specific attribute B with respect to the target variable is calculated as follows (Han et al. 2012: 337):

$$Gain(B) = I(D) - I_B(D)$$

Accordingly, the *information gain* is the expected reduction in information entropy by knowing the value of the attribute B (Han et al. 2012: 337). It is calculated for each attribute in the dataset to identify the attribute, which brings the highest *information gain*. This attribute is selected as the node of the tree structure in each step until it terminates in the leaf nodes containing the class labels of the target variable. This *information gain* measure used by the *ID3* algorithm is biased toward selecting attributes as nodes that have a large number of values, which can result in partitions containing only a very small number of data records (Han et al. 2012: 340). This carries the risk that the resulting tree model may be over-adapted as each partition contains only a very limited number of data records (Han et al. 2012: 340). Therefore, the *C4.5* algorithm contains a kind of normalization, done by including a *split information value* (Han et al. 2012: 340):

$$I\_Split_B(D) = -\sum_{i=1}^{v} \frac{|D_i|}{|D|} * log_2\left(\frac{|D_i|}{|D|}\right)$$

This split *information measure* computes the *entropy* of the data with respect to the number of categories of the attribute (Liu 2011: 74). If an attribute has many categories, the *split info*[63] is generally higher (Liu 2011: 75). Afterwards, the *gain ratio* is calculated, which overcomes the tendency of the *information gain* toward tests with many outcomes and is referred to as (Han et al. 2012: 341):

$$GainRatio(B) = \frac{Gain\ (B)}{I\_Split_B(D)}$$

A decision tree algorithm partitions a dataset with respect to the target variable with the goal of reducing the impurity in the leaf nodes. This could lead to leaf nodes that contain only a very limited number of data records. Accordingly, the tree has a high probability of overfitting, because it memorizes the training records well but may not be able to classify unseen data records. To reduce the overfitting, the decision tree is pruned, i.e. branches or subtrees are deleted and replaced by majority-class leaf nodes to generalize the tree model (Liu 2011: 76-77). There are two common methods of pruning – *pre-pruning* and *post-pruning*. In *pre-pruning*, the trimming of the decision tree is achieved by defining termination criteria that stop splitting the tree when the termination criteria are met. Accordingly, instead of forming another node, a leaf is returned instead. In *post-pruning*, the decision tree is fully grown before subtrees are removed and replaced by a leaf marked with the most frequent subtree class. To decide if a subtree should be pruned, the error rates are estimated. If the error rate is reduced by changing a subtree to a leaf node, the tree is pruned (Han et al. 2012: 345).

3.4.4 Artificial Neural Networks

Artificial Neural Network (ANN) are further learning methods that can be used to estimate a predictive model for a training dataset. This dataset must be encoded in a standardized format, with values between *0* and *1* (Larose et al. 2015: 339). Accordingly, most datasets must be preprocessed, before the analysis can begin. In this thesis, the numeric attributes that have not only values between *0* and *1* are transformed with a *minimum-maximum normalization* (Han et al. 2012: 113):

$$X_i{}' = \frac{X_i - X_{min}}{X_{max} - X_{min}} * (X'_{max} - X'_{min}) + X'_{min}$$

where x_i is the original value for a data record i of the attribute X, X_{min} is the minimum value of the attribute X, X_{max} is the maximum value in the attribute X, and $x_i{}'$ is the normalized value for the attribute X. The variables X'_{min} and X'_{max} are the user-specific minimum and maximum values for the transformed attribute X'.

[63] $split\ info = log_2\left(\frac{|D_i|}{|D|}\right)$.

The categorical variables can be coded by dummy coding. In dummy coding, each category *v* of a nominal attribute B is converted into a separate attribute. If the attribute B has 4 categories $\{v_1, v_2, v_3, v_4\}$, then *4* new attributes are added to the dataset: B_1, B_2, B_3, B_4. If a data record in the dataset has the category v_2, then the attribute B_2 has the expression *1*, while the three remaining attributes B_1, B_3, B_4 are coded with *0*. This procedure can greatly increase the numbers of attributes, as each category must be converted to a new attribute to be binominal-coded.

The ANN used in this thesis are *feedforward networks* that restrict the connections between the layers of the network in a single direction (Larose et al. 2015: 342). The typical layers of a simple neural network are shown in Figure 14. The nodes, also called neurons, in the input layer represent the attributes in the training dataset and the nodes in the output layer represent the target variable. The number of neurons in the input layers usually depends on the number of attributes in the dataset, while the size of the output layer depends on the classification tasks (Larose et al. 2015: 342-343). In the example illustrated in Figure 14, the target variable is binominal and each class in the target variable is represented by a node. The relationship between the input and output layer is modeled by one or more hidden layers. The size and number of hidden layers can be changed to fit the task at hand (Larose et al. 2015: 343). Nevertheless, according to Larose et al. (2015), overly

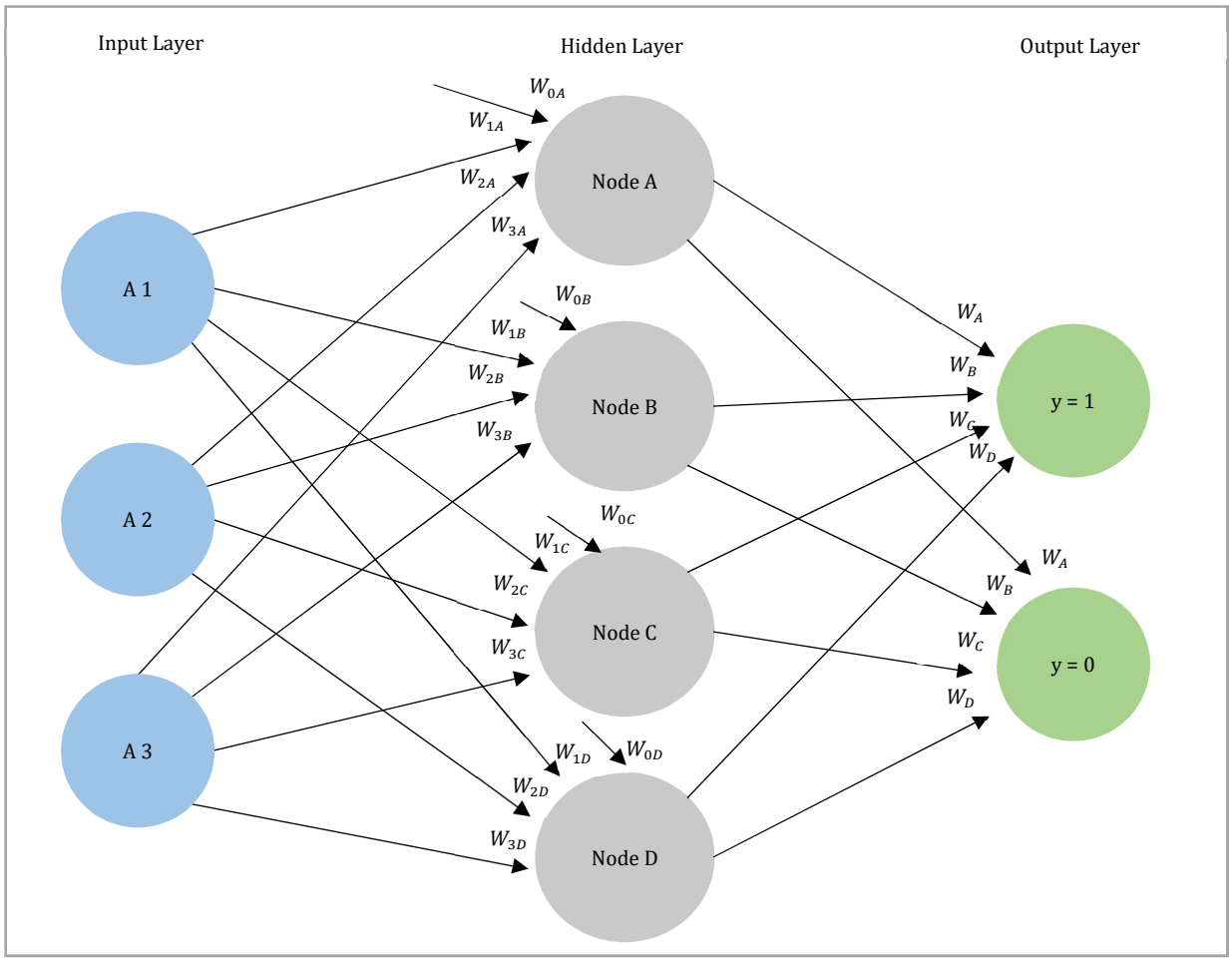


Figure 14. Neural Network model with one hidden layer, based on Larose et al. (2015: 341).

large hidden layers might lead to overfitting, which has negative effects on the predictive power of the model for unseen data records.

The procedure for creating a well-functioning ANN model is summarized in Figure 15 and described below. First, each neuron in a layer is connected to each neuron in the following layer, while the nodes in one layer are not connected to each other. Each connection in the network is assigned a weight between *0* and *1*. The initial weights for each connection and each node W_{nm} are randomly assigned in the initialization phase of the ANN. Subsequently, the attribute values of the input layer are forwarded along the connection to the neurons of the hidden layer. With a combination function, a single scalar value, called *Net*, is estimated for each node in the hidden layer, which is a combination of input values and weights (Larose et al. 2015: 343). For a given node m, this value is calculated as follows (Larose et al. 2015: 343):

$$Net_m = \sum_{n=0}^{N} W_{nm}x_{nm} = W_{0m}x_{0m} + W_{1m}x_{1m} + \cdots + W_{Nm}x_{Nm}$$

where x_{nm} is the n-th input to node m, W_{nm} is the weight associated with the n-th input for node m and a node m has $N + 1$ inputs. The variable x_0 represents a constant input, analogous to the constant factor of regression models, which uniquely assumes the value $x_{0m} = 1$ (Larose et al. 2015: 343). Accordingly, each hidden or output layer contains a certain weighting as additional input $W_{0m}x_{0m} = W_{0m}$ (Larose et al. 2015: 343).

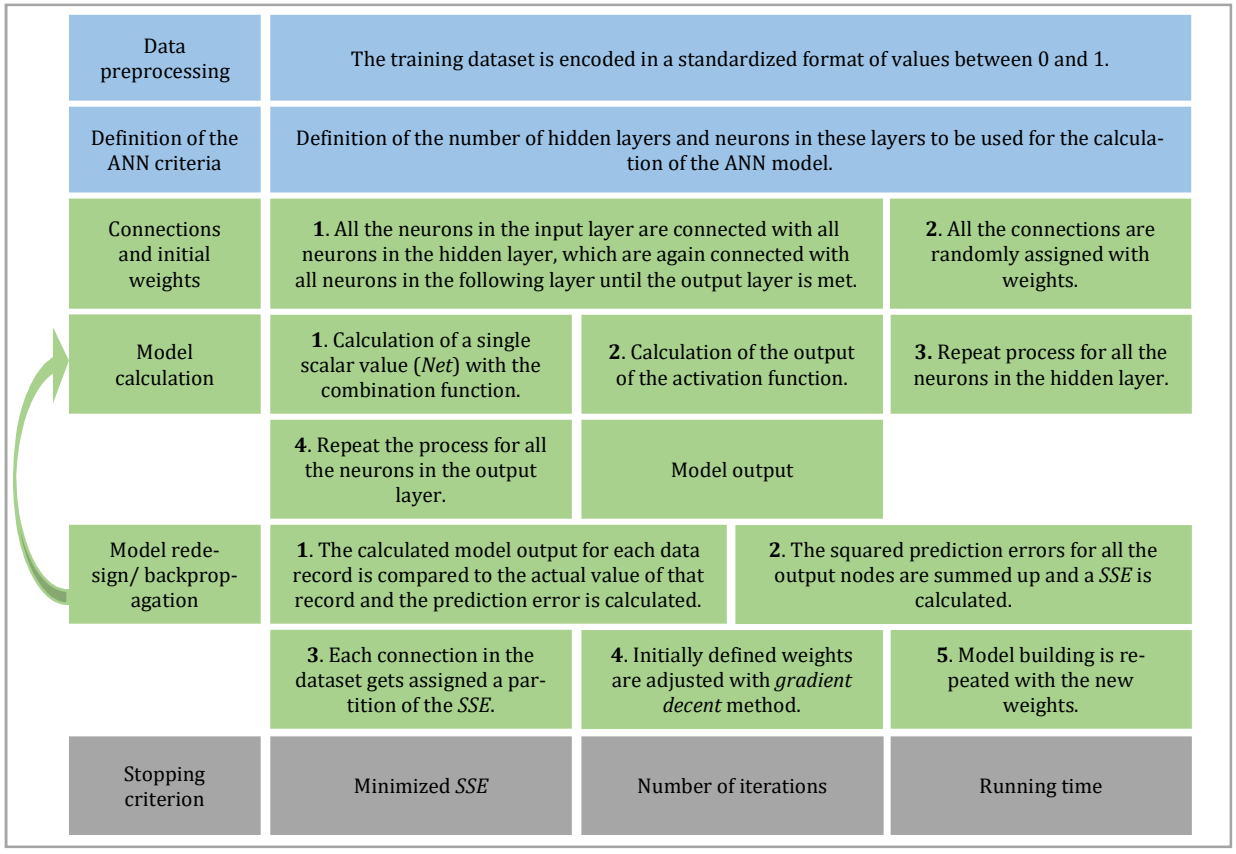


Figure 15. Steps to create an Artificial Neural Network.

Once the Net value of a node has been generated, it is used as an input for a nonlinear activation function. This function models the behavior of biological neurons, which send signals to each other when the combination of inputs for a particular neuron exceeds a certain threshold (Larose et al. 2015: 343). The sigmoid function is the most commonly used activation function (Larose et al. 2015: 344):

$$y = \frac{1}{1 + e^{-Net_m}}$$

where y is the output value of the activation function, e is the base of the natural algorithm, and Net_m is the output of the combination function for node m. The output of the activation function is then used either as an input to the next hidden layer or as an input to the output node (Larose et al. 2015: 344). When the output node is reached, the Net value and the value of the activation function are calculated, and a prediction is made. In a two-output neuron classification problem, ideally, only one is stimulated by the model in order to allow a clear prediction. After the prediction has been made for each data example in the dataset, the predicted value $\hat{y}$ is compared with the actual value of the target variable y. The result of the comparison is a prediction error,[64] which estimates how well the prediction made by the model does fit the actual value in the training dataset. The prediction error is estimated for each data object in the training example and then combined in the *sum of squared errors* (*SSE*) measure (Larose et al. 2015: 345):

$$SSE = \sum_{training\ records} \sum_{output\ nodes} (actual\ value - predicted\ output)^2$$

The *SSE* equation sums the squared prediction errors over all records in the training dataset and over all the output nodes in the model. It indicates how well the initial model worked in predicting the target variable. The goal of the next modeling step is to minimize the *SSE*, which is done by *back-propagation* (Larose et al. 2015: 347). Therefore, the weights of the connections between the neurons in the ANN must be adjusted to best reflect the 'real' connections between the neurons in the training dataset. Due to the nonlinear nature of the sigmoid activation function, *gradient descent* methods are commonly used to adjust the weights (Larose et al. 2015: 346-347). Accordingly, the new weights are estimated as follows (Larose et al. 2015: 346):

$$W_{nm,new} = W_{nm,current} + \Delta W_{nm}$$

where $\Delta W_{nm} = \eta * \delta_m * x_{nm}$. The learning rate η defines the strength with which the error alters the weights, which is typically a selected value between $0.1 \leq \eta \leq 0.8$ (Cleve et al. 2016: 122). The expression δ_m represents the responsibility for a particular error of the node m and is calculated using the partial derivative of the sigmoid function with respect to Net_m.

64 *Prediction error = actual – result =* $(y - \hat{y})$.

The form of the function used depends on whether the node belongs to a hidden or output layer (Larose et al. 2015: 347):

$$\delta_m = \begin{cases} output_m * (1 - output_m) * (actual_m - output_m) & for\ output\ layer\ node \\ output_m * (1 - output_m * \sum_{downstream} W_{mk} * \delta_m & for\ hidden\ layer\ nodes \end{cases}$$

where $\sum_{downstream} W_{mk} * \delta_m$ refers to the weighted sum of error responsibilities for the nodes downstream from the particular hidden layer node. Further details on the mathematical derivation of the *back-propagation* procedure can be found in (Mitchell 1997: 97-103).

After the weights have been adjusted, the described process of model building, shown in Figure 15, continues repeatedly until the stopping criterion is met, which could be a predefined limit on the number of repeats, a real-time limit to training, or the stopping criterion set by the algorithm (Larose et al. 2015: 349). As a rule, the model training is terminated by the algorithm if a set of weights has a significantly higher *SSE* value than the best set of weights used so far (Larose et al. 2015: 350).

3.5 Performance Evaluation for Predictive Models

DM models are created to extract useful information and knowledge from data that can then be used for decision-making. Therefore, it is important to evaluate that the models actually represent the reality found in a particular training dataset and are able to predict the target variable before being used in the field (Larose et al. 2015: 451). When a historical dataset is given, that contains values for the target variable; it is usually possible to generate a model. However, not every predictive model can map the reality well and predict the target variable better than the *default accuracy* rate.[65] These differences between the real world and the model can be caused, among other things, by missing information, incorrect data, incorrect model specifications, or structural changes. Accordingly, not every dataset available for a model building has all the attributes available that affect the manifestation of the target variable. In addition, non-identified data quality problems or the choice of incorrect model parameters adversely affects the predictive accuracy of a model. In addition, structural changes that have not yet been captured in the training database can result in a bed model performance. Therefore, assessing the quality and effectiveness of DM models by their performance is one of the most important steps in the DM process to prevent decisions on unreliable information and insights.

[65] The *default accuracy* rate denotes the *accuracy rate* that would be achieved if the target class labels would be forecasted based on the distribution of the target classes in a historic dataset, without using a trained DM model. For example, if a historic dataset with 100 data records is given that has 60 records belonging to the binominal target class *Yes* and 40 records that belong to the target class *No*, then the *default accuracy* rate would be 60%.

The performance of classification models can be estimated by computing performance measures from a *performance matrix*, also referred to as *contingency table* or *confusion matrix*, illustrated in Table 11 for a two-class classification. As shown in the table, one of the two classes in the target variable is understood as the positive class, while the second class is the negative class. In the example shown, the positive class for the target variable Y is $y = 1$, and the negative class is $y = 0$. The performance matrix captures the number of records in a test dataset that were correctly predicted by the classification model (Han et al. 2012: 365). Therefore, the predicted class for the target variable is compared to the real class of the target variable, to obtain the number of *true-positive* (TP) and *true-negative* (TN) predictions as well as *false-positive* (FP) and *false-negative* (FN) predictions. The number can then be used to generate measurements that indicate the predictive *accuracy* of the models (Han et al. 2012: 365).

Table 11. Confusion matrix for a two-class classification problem.

		Actual class	
		True(y=1)	**True(y=0)**
Predicted class	**Pred(y=1)**	TP (true positives)	FP (false positives)
	Pred(y=0)	FN (false negatives)	TN (true negative)

To generate the confusion matrix, the historical dataset that contains values for the target variable, is normally divided into two subsets. This division can be done by randomly separating the data examples. The *holdout validation* method uses one of the two subsets for training and the other for validation (Larose et al. 2015: 161). The training portion of the dataset should be as large as possible to create a well-trained model. A second method of partitioning a dataset for validation is the *cross-validation* method used in this study. *Cross-validation* splits the training dataset into roughly equal *k* subsets (Larose et al. 2015: 161). The training and validation of the model are then performed *k* times so that each *k* subset is used once as a test set, while the remaining subsets are used to train the model. Compared to the *holdout* method, *cross-validation* ensures that each sample is included in model training, which allows the model to be trained on all available data records in the historical dataset (Larose et al. 2015: 161). Accordingly, no data records are 'lost' during testing.

The metrics extracted from the generated confusion matrix and used in this study are *accuracy*,[66] the *error rate*, the *class recall*, the *class precision*, and the *F-score*. The *accuracy* of a classification model indicates the proportion of correct classifications made by the model, which is calculated as follows (Larose et al. 2015: 456):

$$Accuracy = \frac{TP + TN}{N}$$

[66] In this study, *accuracy* is understood as the overall *recognition rate* that reflects how well the classification model recognized the data records belonging to the various target variable classes (Han et al. 2012: 366).

where N is the total number of data records used for the model testing, which can be calculated as $N = TP + TN + FP + FN$. In comparison, the *error rate* measures the proportion of incorrect classifications (Larose et al. 2015: 456):

$$Error\ rate = 1 - accuracy = \frac{FP + FN}{N}$$

These two measures should always be examined with respect to *class recall* and the *class precision*, since *accuracy* may be high and the *error rate* low even if a classification model can predict well only one of the classes of the target variable (Han et al. 2012: 367). This may be the case, for example, if the balance between the target classes in the training dataset is distorted. The *class recall* specifies the proportion of data records that have been correctly identified as belonging to a particular target class. Accordingly, the proportion of correctly classified data records belonging to the positive class of the target variable is estimated as follows (Larose et al. 2015: 457):

$$Class\ recall(y = 1) = \frac{TP}{TP + FN}$$

The proportion of data records in the test dataset correctly classified as belonging to the negative class can be estimated by:

$$Class\ recall(y = 0) = \frac{TN}{TN + FP}$$

A good classification model should have a high *class recall* for all target classes, but at least for the class that is considered more interesting. A perfect classification model would have a *class recall = 1.0* for all target classes, which means that 100% of the cases in the training dataset were correctly predicted by the classification model (Han et al. 2012: 368).

The *class precision* measures the exactness of the model. It indicates what percentage of the data records in the test dataset that were predicted to belong to a particular class are actually records that belong to that class. The *class precision* for the example performance matrix in Table 11 can be estimated as follows (Han et al. 2012: 368):

$$Class\ Precision(y = 1) = \frac{TP}{TP + FP}$$

$$Class\ Precision(y = 0) = \frac{TN}{TN + FN}$$

Accordingly, the *class precision* specifies the likelihood that a data record in the test dataset will actually be part of a particular class, provided that the model has classified that record as belonging to that class (Larose et al. 2015: 458).

The *F-score*, also known as *F-measure* or *F-value,* is an alternative to the measure of *accuracy*. It is the harmonious mean between *class recall* and *class precision*, taking the differences in the predic-

tion of the individual classes into account (Han et al. 2012: 369). Accordingly, it may be a particularly important measure to consider if the classes of the target variables are unbalanced. The measure is calculated as follows (Han et al. 2012: 369):

$$F - score = \frac{2 * class\ precision(y = 1) * class\ recall(y = 1)}{class\ precision(y = 1) + class\ recall(y = 1)}$$

A *Receiver Operator Characteristic* (*ROC-curve*), illustrated in Figure 16, can be used to visualizes the risks and benefits associated with the classification (Han et al. 2012: 374 - 377). This is illustrated by the representation of trade-off between the data records in a training dataset correctly classified to the positive class and the proportion of examples incorrectly classified as belonging to the positive class. If the *ROC curve* for a classification model is steep at the beginning and significantly higher than the random guessing line, which means that there are many true positive predictions, the model performs well (Han et al. 2012: 376-377). This is also represented by an *area under the curve value AUC* value close to 1.0, which is the value for perfect model *accuracy* (Han et al. 2012: 377).

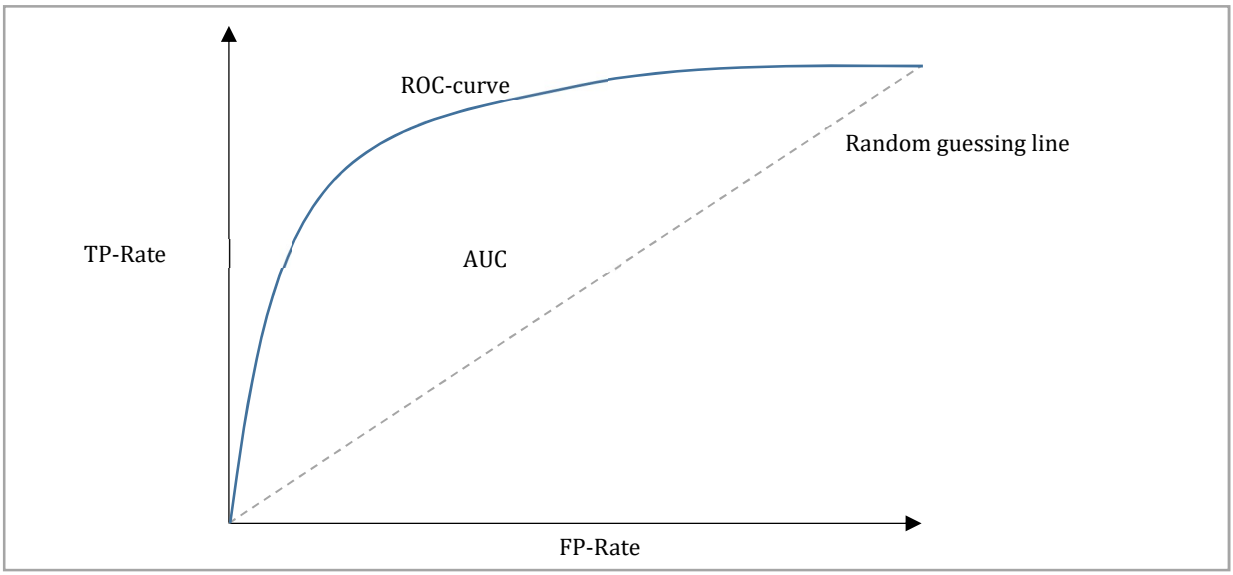


Figure 16. ROC-curve.

4 Decision Support for the University Management through Data Mining

This chapter discusses the incentive for using a DM approach to address the challenges of German universities. The first section emphasizes the main motivation for the DM approach, which among other things is that the resource data, in particular, student and applicant data, are available to all German universities. By analyzing these existing resources with DM methods, universities can gain information that helps them master their challenges, identify their opportunities, and accomplish their tasks and objectives. The second section presents a framework that aims to provide a consolidated view of the possibilities that the DM-based analyses of student and applicant data offer for university management. In addition, the framework aims to identify untapped potential that could be used by university decision-makers. The third section draws attention to the peculiarities that have arisen in the different phases of the DM process in the case studies presented in Chapter 5. In this context, steps are introduced aimed at supporting future DM projects at German universities.

4.1 Motivation for the Data Mining Approach

Chapter 2 describes the environmental conditions that confront universities in Germany with a strong competitive situation for qualified students, lecturers, researchers, staff, and funds. It was found that these and the changed legal situation of the education sector influence the tasks and goals of the universities. Accordingly, increased administration autonomy gives universities flexibility, decision-making power, and the ability to influence their strategic position. As a result, they can respond to the needs of their stakeholders and increase their competitiveness. Otherwise, university management has more responsibilities, and every university must ensure that the predefined government tasks are met. In order to meet these obligations and seize the opportunities of organizational change, a strategy is widely accepted by universities as a necessity (Berthold 2011: 7).

As shown in Figure 17, it is suggested that DM can assist the university management in finding answers to some important questions arising from the opportunities and responsibilities associated with the increased decision-making autonomy. In order to individualize their strategy, universities must adapt their course of action and orientate themselves toward individualized goals (Gerhard 2004: 4). It is believed that identifying students' needs and demands can provide helpful guidance for the university's strategic direction. In addition, if decisions or organizational changes are based on demonstrable information, the university can transparently account for its actions, which is necessary because they are state-supported entities (Kamm 2014: 115).

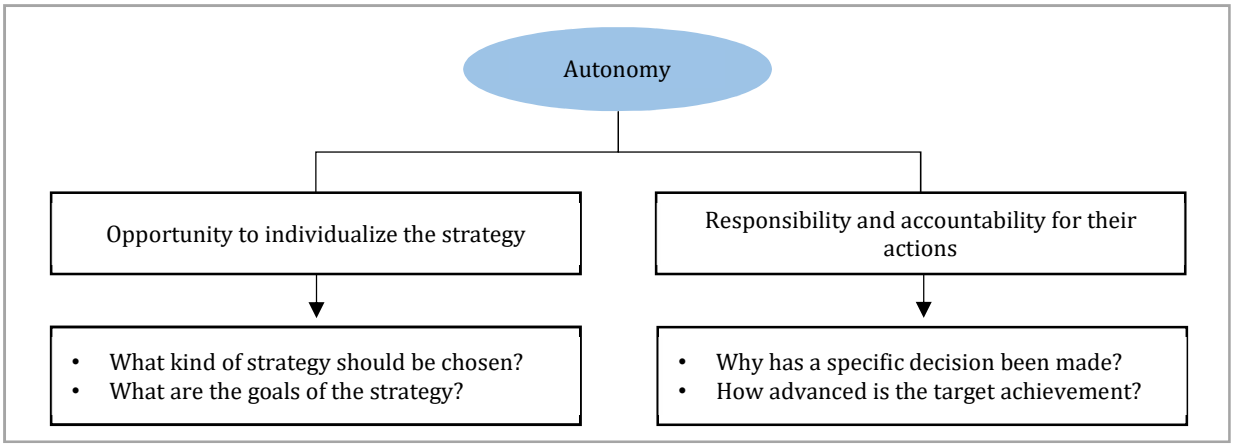


Figure 17. Managerial support that can be provided to universities by analyzing available data.

The above-mentioned need of universities to create an individual strategy is supported by the intense, competitive environment in the education sector, as they have to distinguish themselves from their competitors. Figure 18 summarizes the main university-level competitive forces identified in Section 2.2.2, along with the competitive support assumed by analyzing available student data.

German universities must compete with national and international universities, as well as additional educational service providers, for students who have a very strong negotiation position and are increasingly understanding themselves as clients of the university (Lomas 2007: 34). To make

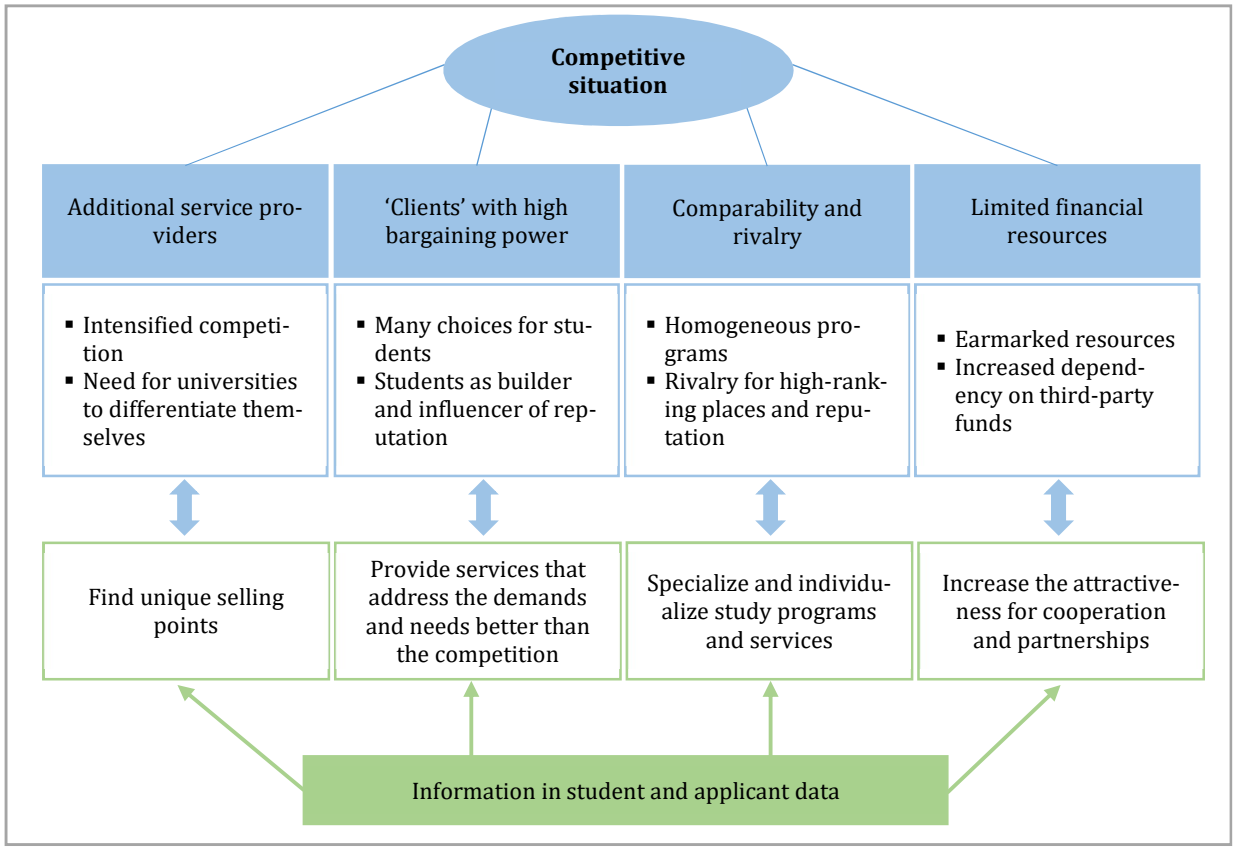


Figure 18. Assumed support that is provided by the analysis of student and applicant data for addressing the main competitive challenges of German universities.

things even more difficult, the universities' programs and services often appear to be homogeneous and therefore interchangeable. In addition, the establishment of new study programs, as well as the declining birth rate, are leading to a downward trend in student numbers, which intensifies the rivalry between the HEIs. In addition, public universities are in a unique financial situation, characterized by limited and earmarked resources as well as dependence on project-based third-party funding (Jaeger 2009: 45).

With information about the real needs and demands of its students, a university can individualize its services, develop specializations, and establish priorities that set it apart from the competition. As a result, universities can develop profiles that are a demand-driven extension of their services (Kamm 2014: 127; Erhardt, M. et al. 2008: 21). This helps them to attract qualified researchers and students who support the establishment of a positive reputation, which in turn supports the attractiveness of the facility for collaborations and partnerships. Partnerships and practical recommendations that consider the relevance of a research idea are often an important part applying for funding.

Accordingly, the analysis of student and applicant data can help universities achieve their current tasks and objectives. It is understood that among the tasks and objectives identified in Sections 2.2.3 and 2.2.4, in particular, these data resources can support the ones summarized in Figure 19. Thus, when the demands of study programs are identified, universities are supported in becoming service providers and improving the success of their students. This can be achieved by developing and providing services that are actually required by students, which in turn improve the student performance. In addition, predictive models can be generated that allow the forecasting of student dropouts and indicate intervention needs, which also has a positive impact on student success. In this way, dropout rates can be minimized, which is one of the mandatory government goals.

Figure 19. Tasks and objectives of German universities, which are assumed to be supported by the analysis of student and applicant data.

The success rate of students can be further improved by ensuring transparency by making the requirements of degree programs visible. The students include their personal interests and their abilities in the decision for a study program (Horstmann & Hachmeister 2016: 4). By creating transparency, the university can prevent students from applying for a program that does not live up to their expectations and thus from becoming dissatisfied. In addition, knowledge of the challenges of study programs can help the university management to ensure the quality of their education programs by further investigating and restructuring courses that are considered to be at greater risk of causing student dropout. Once quality issues are identified and resolved, students' success can be further enhanced.

If universities succeed in becoming service providers for their students and increasing the success rate of their students, the remaining scope of the university's tasks and goals can also be positively influenced. Successful graduates become employees of the economy and as such transfer their knowledge from the university into practice, which optimally supports the success of businesses. Well-educated and successful graduates are therefore valuable employees for companies and research institutions. They help them achieve their goals, face challenges, and ensure their success. These positive experiences with graduates from a particular university support the establishment of a good reputation, which helps to form partnerships between the university and the external institutions. These partnerships can be used to ensure the topicality of study programs by incorporating practical expertise into the curriculum preparation and revision, or by incorporating practical knowledge into the courses through field trips or external lectures. The positive reputation of a university, in turn, can increase the international visibility and reputation of a university, which has a positive impact on its attractiveness for national and international students and researchers. In addition, some of the general objectives and tasks of universities are also expected to be supported by the proposed analyses. As mentioned above, the acquisition of third-party funding may depend on reputation and cooperation that can be increased by using student data. In addition, decisions based on objective information increase transparency.

However, university data resources do not support demand-driven decisions by themselves. They need to be analyzed to extract useful information. Traditional data analysis methods provide an overview of existing data by examining and describing the variance of each attribute. This is done with the support of tables and charts as well as descriptive parameters such as mean value, spread, or standard deviation. Crosstabs and correlation analyses provide first insights into the relationships between the attributes in a dataset. Nevertheless, these methods allow only the comparison of two attributes simultaneously and are therefore not suitable for the recognition of complex connections and patterns. Therefore, descriptive analysis is of limited use for datasets with many attributes and records.

DM methods are able to recognize 'knowledge' in the form of patterns, rules, and relationships in datasets (Witten & Eibe 2001: 2-3). Compared to the traditional data analysis methods, DM methods can automatically examine the connection between multiple attributes, which can detect both suspected and unsuspected relationships (Han et al. 2012: xxi, 1-5). Accordingly, all the details available in a dataset can be included in the analysis at once. This allows for the creation of rules and models representing the relationships between all attributes in a target dataset, or the identification of factors influencing a particular target variable from a large number of attributes. In addition, these models allow the prediction of a target variable that can decisively support decision-makers. For example, by applying classification methods to student and exam data, universities can create predictive models that recognize challenges in their programs and forecast student dropout, which allows for early intervention.

As a result of the above discussion, this study suggests that university management can detect complex patterns and relationships in their existing data resources using DM methods that cannot be detected with traditional data analysis methods. In addition, predictive models can be generated to assist the university management to predict future events and influence them according to their needs, thereby enabling universities to secure their position in a more competitive environment.

4.2 Overview of the Support DM can Provide to the University Management

From the previous discussion of potentials, the analysis of student and applicant data resources can provide to the university management a framework was constructed, which is presented in Figure 20. It aims to present an overview of the possibilities that two specific DM projects represent for university management. The selected issues are the prediction and study of student dropouts and the prediction of applicants' enrollment. It has been decided to focus on these two DM issues as it is assumed that their results have a long reach within the university and because each university in Germany has access to the relevant data resources.

The student-related data resources in Germany usually consist of demographic data and information about the previous education of the student. This information has to be collected by all universities because they are obligated to transfer certain information to the government for statistical purposes (BMJV 1990). The data resource may also be extended to include information about the study career of the student. Moreover, the status of a student in the study program is included, which indicates whether he or she has completed the program successfully. As illustrated in the framework, dropout analysis is expected to be able to identify a student's failure or transfer, student needs, and the demands and needs of study programs. When courses are identified that appear to increase the likelihood of dropping out, they can be further investigated to identify and remedy quality issues. In addition, support measures can be derived to help the students master

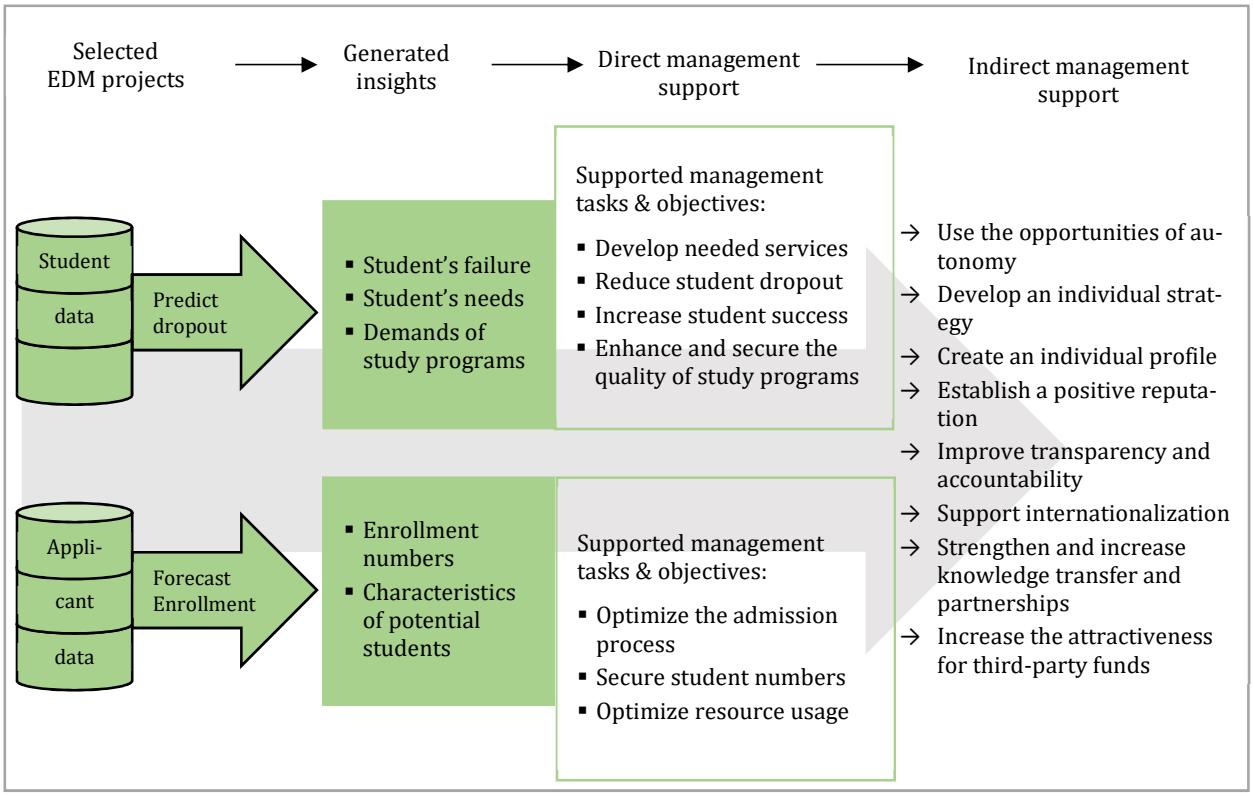


Figure 20. Framework that combines the results of EDM projects with the management support they provide, focused on dropout analysis and enrollment forecast.

the challenges and successfully complete their studies. These support measures can be services, such as tutorials or summer schools, which are available exclusively to the students of the university or even mandatory for the students with difficulties. In addition, the provision of individualized student support can additionally be a service that allows the university to set itself apart from the competition. Furthermore, by knowing their students' difficulties and setting up countermeasures, university management can increase the students' success rate, which can have a positive impact on the university's reputation.

Available applicant-related data resources typically consist of demographic data, information about previous education as well as information about the enrollment status. By analyzing this data, it is assumed that the characteristics of the applicants can be identified which on the one hand support the enrollment at the beginning of the semester positively and on the other hand negatively influence the likelihood of enrollment. In addition, enrollment numbers can be predicted that ideally help avoid overbooking and support demand-oriented resource planning. Furthermore, the knowledge and connections in the model can be used to get to know the applicants of the university better. It is believed that the generated information indicates potential for optimizing the decision-making process in order to attract enough and successful applicants to the university in general and to each study program in particular. Accordingly, it is assumed that the enrollment forecast models directly assist in optimizing the enrollment process and the resource usage.

The indirect management support that can be achieved through the options described above is manifold. As mentioned in the previous section of this chapter, the knowledge generated can help university management to individualize their strategy and to exploit the possibilities of autonomy. Accordingly, uniqueness can be created that can secure the position of a university in today's competitive environment. If this individualization can be justified by insights, the university management can substantiate its actions, which in turn increases transparency. In addition, when defining goals on the basis of facts, the university administrators can monitor the achievement and communicate this to its stakeholders.

As discussed in Section 4.1, new services and on-demand student support can help universities increase their students' success, which can positively impact the university's reputation. This can help the university build national and international partnerships with research institutes and companies. In these partnerships, universities can intensify the transfer of knowledge through joint projects and cooperation. This transfer is particularly useful for universities of applied sciences because they have a practical orientation in their degree programs and their research. A regular exchange with the economy helps the university management to stay up-to-date and provide services and study programs that are current and needed. As a result, students are trained in careers desired by the business community to ensure their employment after the successful completion of the program. This, again, can have a positive impact on the university's reputation, which in turn increases the universities' ability to raise additional funding, and further enhance the national and international visibility and long-term success of the institution in an intense, competitive environment.

4.3 DM Process for Analyzing Data from German Universities

Before the case studies in Chapter 5 are carried out, the specific features of DM projects in the German university environment are discussed. This is done based on the phases of the CRISP-DM described in Section 3.1. The identified specifics are based on the experiences made during this research. Accordingly, there is no claim to a comprehensive picture for all DM projects and German universities.

In the first phase of CRISP-DM, *Business Understanding*, the planned DM project gets its direction. Therefore, the information needs of the management, which motivate the use of DM methods, must be identified to formulate the business objective. In companies, a DM project is usually initiated by decision-makers who seek improvement and expect this to be achieved with additional information and, consequently, data. From an external perspective, the need for universities to gain and use additional information is clearly recognized by this research, but for most of the university departments, the analysis of data for decision support is not yet commonplace. Therefore, a challenge for the implementation of a DM project at a German university is recognized, which is to convince the

decision-makers of their information needs, which can be addressed with DM methods. In addition, the organizational structure of universities may make it difficult for individuals or individual departments to introduce change, as they often need the support of a committee before any real action can be taken. Therefore, it may be necessary to convince a wide range of actors of the need and benefit of the DM project.

Once a realistic business goal has been formulated, such as lowering the student dropout rate by 20%, the DM goal must be deduced. At this stage, DM know-how is required to identify the DM methods that can support the achievement of the identified goal. Although universities are places of progress where new techniques and methods are tested and developed, using these developments to support the institution is not commonplace. As a result, the required know-how for DM in the governing bodies of German universities is probably sparse, especially at a time when all kinds of institutions around the world are looking for well-trained data analysts (Davenport & Patil 2012; Shum et al. 2013). Therefore, formulating a realistic DM objective is another challenge for universities in the *Business Understanding* CRISP-DM phase.

In the second phase of the CRISP-DM, *Data Understanding*, several particularities have been noticed, starting with accessing the data resources. In order to access the data, several actors must be involved and convinced. It is, therefore, proposed to develop and provide a detailed and understandable strategy to describe the planned DM project, including data security compliance measures. Once access to the data has been granted, the next hurdle is to understand the structure of the available database or databases. Without the support of the department responsible, this can be a time-consuming or even impossible process. Accordingly, it is very important that the daily users of the data and the benefiters of the DM project recognize the benefits and support that the project brings. This also applies to the understanding of the available data, as educational information such as course and module names are often encrypted (Prakash et al. 2014; Chau & Phung 2012). In addition, data quality issues have been identified, such as course and module codes not being consistent over time, e.g. they change with new study and examination regulations. Furthermore, different spellings of places or double-filled attributes were identified. This can complicate the generation of valid models, which in turn is aided by the fact that the amount of educational data is often rather low (Chau et al. 2012). In addition, the resources available are usually only part of the reality, and it can be difficult to obtain attributes with impact (Baker 2014; Chau et al. 2012), e.g. data on a student's motivation for participating in a particular study program are typically not collected.

The structure of university databases is based on the tasks that have to be fulfilled, e.g. update the student status, change address information, or update course grades. It was found that the retrieval and further use of the data for the analyses were not yet targeted, which entailed various data preparation steps before the data can be analyzed descriptively. For example, the dataset retrieved from

the student data management system of the case university contained most student records countless times because each exam could only be displayed on a separate line. Therefore, each exam and the related exam information had to be converted to a separate column so that each student would only be represented by one record in the dataset. In addition, peculiarities in the structure of certain attributes must be addressed. For example, there are many attributes available that represent exams, and because not every student attends each exam, the dataset contains an enormous number of missing values. The described missing values for the exam results, however, are not classic missing values due to missing information or data entry errors. The information may be missing for several reasons, e.g. the student did not attend the course, the student did not take the exam, or the student dropped out before the course takes place according to the curriculum.

The third CRISP-DM step, *Data Preparation*, processes the raw data for the analysis. As mentioned above, the preparation steps had to be done before the *Data Understanding* phase can be completed. This may include the anonymization of the data, especially if personal data are being processed. The anonymization of student data requires the generation of a new personal ID. As a rule, the students already have an individual ID assigned to them at the beginning of their studies, but based on this 'matriculation number', the person that it represents can be easily identified. To avoid this, a new ID was created. In addition, unique personal data such as the date of birth or the home address must be anonymized, which may include the generation of new attributes. Beyond this, if possible, the creation of new attributes is generally recommended because the data available at German universities is not extensive. The additional attributes that may be of interest to the decision-makers and for the analysis should be identified together with the department responsible. For the following case studies, for example, the new attribute *TimeHEEQDegree-Application* was considered relevant, which can be extracted from the date of the Higher Education Entrance Qualification (HEEQ) and the date of the enrollment at the case university. In addition, existing data quality issues have to be addressed, caused mainly due to errors in data entry, such as different spellings of city names or duplication of information in one cell.

No specifics were encountered in the remaining CRISP-DM phases *Modeling*, *Evaluation*, and *Deployment*. It has rather been shown that common DM methods can be applied. However, it cannot be ruled out that certain DM methods perform better than others for certain EDM challenges, but this has not been the focus of this research.

To help the establishment of EDM projects across German universities, the following DM process steps are proposed, which were drawn from the experience of this study.

1. Convince the relevant department, decision-makers, and users: In this step, the motivation and willingness of the decision-makers to participate in the planned DM project should be aroused. The planned DM project should be discussed and presented to the direct benefiters

of the analysis results, e.g. the studies and examination office. It can be helpful to present best practices to make the project tangible.

2. Establish a business objective: Once the decision-makers are convinced, a business goal should be jointly identified and formulated. Ideally, the identified goal should be achievable with the available data resources in a relatively short time and with DM methods that provide interpretable results. This helps to ensure that the department receives understandable results within a reasonable period of time, which should positively influence their willingness to participate in further DM projects.
3. Define the DM goal and an analysis plan: It is suggested that this step is already considered in the previous one, at least if first experiences with DM are gained since the aim is then to run a successful DM project to increase the willingness throughout the university for further projects. Nevertheless, this step should ensure again that a clear DM goal is formulated and that an understandable analysis plan is developed.
4. Examine the available data resources: In this step, it is especially necessary that the responsible department supports the DM project. As mentioned earlier, the data available from educational institutions are usually encrypted and require the expertise of the department to be properly understood. In addition, it is possible for data to be stored in multiple databases, and it may not be clear where the data are best retrieved or how it should be combined. In addition, it is essential to engage with the topic of data security and to discuss how all parties involved can ensure the responsible handling of the data. This includes the anonymization of personal data and the assurance that the data are only stored in secure places.
5. Recognize and handle data quality issues: The presented study identified data quality issues, in particular those related to data entry errors, such as the different spelling of city names. The remedy of those is time-consuming and a permanent solution should be found to avoid such problems in the future. In addition, it is suggested that all data preparation steps performed be properly recorded as they must be redone in order to apply and further evolve the DM models.
6. Create comprehensible models and evaluate: It is suggested that the first DM project conducted aims to create interpretable models that can be comprehended by the decision-makers. This includes considering the use of a DM tool that can also be understood and handled by the departments. In addition, it should be ensured that the models are properly evaluated prior to interpretation. The DM analyst should first explain both, the evaluation and the interpretation. It has been found that offering first interpretations and conclusions to the decision-makers ensures that they understand the outcome correctly, which helps them to make their own interpretations.

7. Plan the implementation and further development of the models: This step must be done together with the university departments. It may be worth discussing whether additional information is available or can be collected. In addition, decision-makers should be informed about the functioning of the models and equipped with a tool and the knowledge that enables the application. In addition, a plan should be drawn up as to when and how the models can be improved and evaluated in the future.
8. Keep in mind that the analysis results are only as good as the willingness of the decision-makers to use them: Therefore, it should be ensured that the responsible decision-makers are informed about the DM project and its current status in a contemporary manner. In addition, their opinions and suggestions should be regularly included as only they can ultimately turn the results of the DM project into benefits and support a thorough implementation of EDM.

5 Case Studies

This chapter analyzes existing applicant and student data of a case university of applied sciences with the DM methods described in Chapter 3 to examine whether the management support proposed in Chapter 4 can actually be achieved. The case university is located in southern Germany and currently has around 4000 students. The analyses are performed with RapidMiner, a software platform that combines data preparation, descriptive analysis, predictive analysis, and predictive modeling (RapidMiner 2019). In the user interface of the program, the analysis is created with operators connected to a workflow. The operators each have a specific predefined task that allows comparatively straightforward modeling. In addition, the workflows and analyses can be shared with other RapidMiner users who only need to adapt the modeling steps to their specific data resources. The analysis processes used below can therefore be downloaded to give the target group universities the ability to implement the analysis steps and to advance and use these steps to make the most of their own data resources (see Appendix D).

5.1 Enrollment Forecast

A big challenge for universities which significantly impacts their success and their reputation is the overbooking or underbooking of study programs. Overbooking is a phenomenon that occurs when more units with limited capacity are sold than are actually available (Zenkert 2017). This phenomenon can be observed daily in the aviation or hotel industry, where overbooking is done to avoid loss of revenue due to customers' no-show (Hueglin & Vannotti 2009; Klindokmai et al. 2014). Without the sale of more flights or hotel beds than available, any cancellation would result in an empty seat or empty room and consequently a loss of income (Zenkert 2017; Phumchusri & Meneesophon 2014). However, when capacity is overbooked and every booking appears, there are additional costs, as customers have to be provided with lucrative alternatives to the reservation. Nevertheless, if a customer does not receive the booked service, dissatisfaction is high, most likely resulting in poor ratings and low appreciation for the institution.

Universities in Germany are facing a similar problem. They have a certain number of study places available in each program, which must be optimally utilized. As a rule, a potential student in Germany applies for more than one program and several universities to get a good selection of study offers. From these offers, the applicant chooses only one at the beginning of the semester. Therefore, only a relatively small number of applicants for a degree program are ready to begin the program once accepted. The degree programs in Germany can be divided into open-admissions and admissions-restricted programs, all of which have a limited number of study places. For admissions-restricted degree programs, this fixed number of study places is mandatory, while for open-admissions programs, the number can be understood as an ideal value or guideline. Accordingly,

individuals who apply to an open-admission degree program within the application deadline and, if applicable, meet all the defined requirements for the program (e.g. internship, minimum HEEQ grade) have to be admitted. If the study program is admission-restricted, the applicants with valid applications will be classified according to predefined criteria (e.g. HEEQ grade, number of waiting semesters). If an applicant has a higher-ranking place than available places, he or she receives an admission offer. If an applicant fulfills all the mandatory requirements for the program but does not have a high-ranking place, he or she may take a place if one of the candidates higher up the ranking list does not accept the admission offer.

In the conventional sense, overbooking does not take place for open-admissions programs because the pre-defined study place capacities are officially only an ideal value. Nevertheless, if this plan value is significantly over-utilized, the university faces the same challenges as if admissions-restricted programs are overbooked. Consequently, if significantly more applicants enroll to a program with open-admission than the ideal value of available study places, the capacities are overloaded as well, which is also referred to as overbooking in the following.

As universities in Germany receive part of their government funding for the number of students in the first semester, capacity utilization is very important. Accordingly, unless all available study places are filled, existing resources – human resources, assets, equipment – will not be fully utilized; however, they must still be available and subsidized. As a result, study places in Germany are usually overbooked, in order to have enough applicants who enroll at the beginning of the semester. Conversely, overbooking capacity leads to congested resources, crowded lecture halls, overburdened faculty, and over-utilized supplementary services. Such a situation is visible to students and staff and will most likely lead to dissatisfaction on both sides.

Estimating the exact number of applicants enrolling in a program is largely based in practice on the experience of professionals in the admissions department; it is therefore often instinctive and based on the experiences of past semesters. Although experience is a very valuable asset in the decision-making process, overbooked study programs are not uncommon. Consequently, universities should use all available resources to support their decision-making processes. During the application process, each university collects data about the applicants, their education and career. DM provides a way for universities to extract information from these existing data resources and create models that help predict applicants' enrollment or no-show. It is believed that these predictive models and the information they contain objectively support the enrollment process and positively impact the long-term success and existence of a university.

5.1.1 Introduction of the data resources

To forecast the no-show of applicants two datasets were collected in the application period of the last two semesters – the summer semester SS2018 dataset and the winter semester WS2018/2019 dataset, both of which cover the applications to six Bachelor programs offered each semester. These two datasets have been grouped together in the *Applications2018* dataset. This grouping was done to include the possible structural difference between the two application periods in the model generation and to have a sufficient number of data examples. For that reason, models were created that are applicable to all Bachelor programs at the same time. As soon as more data examples are collected in the upcoming application periods, it may be worthwhile to generate predictive models separately for each Bachelor program.

The attributes available in the *Applications2018* dataset are described in Table 12. In addition to the attributes extracted directly from the case university's application management system, new attributes have been generated to anonymize and compress the information. The created attributes are marked with the *new* expression.

The target variable in the dataset is *Enrollment*, which indicates whether an applicant enrolled at the beginning of the semester *Enrollment(Yes)* or did not enroll *Enrollment(No)*. The distribution of the examples in these two classes is shown in Figure 21. Applicants who have been excluded from the application process due to non-compliance with mandatory requirements or incomplete information are presented separately – *Status(Excluded)*. These candidates, 413 in total, are excluded from the subsequent analysis because they are not considered for admission to one of the study programs. Of the remaining 2611 examples in the dataset, 774 enrolled at the beginning of the semester, and 1837 did not enroll. Accordingly, the case university had an enrollment rate of 29.6% in 2018.

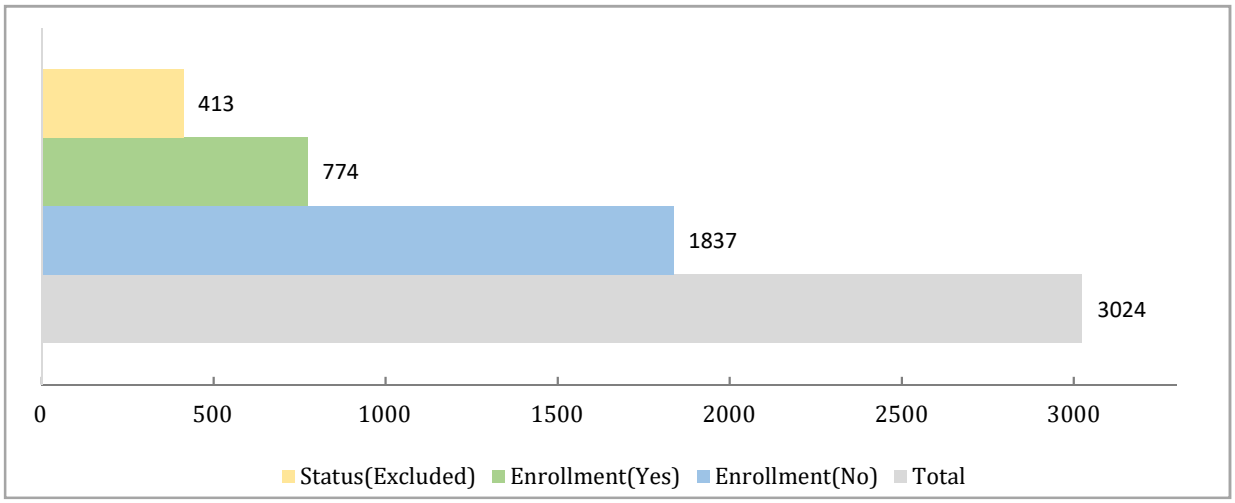


Figure 21. Distribution of applicants in the categories *Enrollment(Yes)*, *Enrollment(No)* and *Status(Excluded)*.

Table 12. Description of the applicant attributes in the available dataset.

		Attribute	Description	Values
Demographic attributes	Personal information	*Gender*	The gender of the applicant.	*male*, *female*
		Age (new)	The age of the applicant at the time of application.	Age in years
		PlaceofBirth	The name of the applicant's place of birth at the time of application, derived from the date of birth.	Name of the city
		Distance_PlaceofBirth (new)	The distance of the applicant's birthplace from the location of the case university. This is the compacted version of the attribute *PlaceofBirth*.	*Radius100* (< 100km), *Radius200* (100km < Radius200 ≤ 200km), *Radius300* (200km < Radius300 ≤ 300km), *Radius400* (300km < Radius400 ≤ 400km), *Radius500* (400km < Radius500 ≤ 500km), *Over500* (> 500km), *Abroad*
		BirthCountry	The name of the country where the applicant has been born.	Country name
		Nationality	The nationality of the applicant.	*German*, *Foreign*
		Residence	The name of the place of residence at the time of application.	Name of a German town or city, else *Foreign*
		Distance_Residence (new)	The distance of the place of residence from the university location.	See attribute *Distance_PlaceofBirth*
		TownSize_Residence (new)	The size of the town or city where the applicant lives at the time of application.	*Community* = below 5.000 inhabitants, *SmallTown* = between 5.000 and 20.000 inhabitants, *Town* = between 20.000 and 50.000 inhabitants, *BigTown* = between 50.000 and 100.000 inhabitants, *City*= between 100.000 and 500.000 inhabitants, *BigCity* = above 500.000 inhabitants, *Foreign* = Place of residence is not in Germany
	Previous education	*HEEQDegree*	The type of Higher Education Entrance qualification (HEEQ) of the applicant.	*AHR* = 'traditional' university entrance degree, *FHR* = university of applied sciences entrance degree, *fgHR* = subject-specific entrance degree, *Abroad* = foreign entrance degree
		HEEQDistrict	The district or country where the HEEQ was earned.	Name of the city, district or country
		Distance_HEEQDistrict (new)	The distance of the HEEQ district from the university.	See attribute *Distance_PlaceofBirth*
		HEEQCountry	This attribute defines whether the HEEQ degree has been obtained in Germany or abroad.	*German*, *Abroad*
		HEEQGrade	The grade of the HEEQ degree.	*1.0* (best grade) – *4.0* (least passing grade)
		HEEQGradeComp (new)	The compact version of the attribute *HEEQGrade.*	*very good* = 1.0 – 1.5, *good* = 1.6 – 2.5, *satisfactory* = 2.6 – 3.5, *sufficient* = 3.6 – 4.0
		TimeHEEQDegree-Application (new)	The time between the successful completion of the HEEQ and the application at the case university.	*< 6 Months, between 6-12 Months, between 13-18 Months, between 19-24 Months, 2-3 Years, 3-4 Years, 4-5 Years, 5-6 Years, > 6 Years*
		PreviousSemesters	The number of semesters in which the applicant has already been enrolled in.	*0* to max
		FirstSemester (new)	This attribute indicates if the student starts the university career at the case university or has been enrolled previously.	*Yes* = Student has not been enrolled in previous study programs, *No* = Student has been enrolled previously

		Attribute	Description	Values
		Apprenticeship	This attribute defines whether the applicant has completed an apprenticeship at the time of application.	*Yes* = student has completed an apprenticeship, *No* = student has not completed an apprenticeship
Studies-relevant attributes	**Application-related information**	*Priority*	The priority of the application for a particular study program at the case university.	*1 – 6*, with *1* being the highest priority, and *6* being the lowest priority
		MultipleAppli (new)	This attribute indicates whether the applicant has applied to more than one study program at the case university.	*Yes* = the applicant applied to more than one program, *No* = the applicant applied only to one program
		AppliNumber	This attribute indicates how many programs the applicant applied for at the case university.	*1 – 6*, with *1* representing a single application
		Status_SevenWeeks *Status_FiveWeeks* *Status_ThreeWeeks* *Status_OneWeek*	The *Status* attribute has been collected seven weeks, five weeks, three weeks and one week prior to the application deadline and changes according to the status of the applicant at any time in the collection. If an applicant has not yet applied to the case university at the time of data collection and therefore has no *Status*, he or she will be given the status *NotApplied*.	*Received* = The online application has been received, but the obligatory printed application has not yet been received. *Valid* = All documents have been received and will now be verified as required. *Excluded* = The applicant was excluded because he/she did not fulfill the formal requirements. *Admission* = The applicant received a study place offer. *Later* = The applicant does not meet all the mandatory requirements but may be offered a place if there are still places available or the required documents are submitted. *OfferRejected* = The applicant has rejected the admission offer. *EnrollmentRequested* = The applicant accepted the admission offer and applied for enrollment at the beginning of the semester. *NotApplied* = The applicant has not yet submitted his/her application at the time the data was retrieved. Applicants marked with this status are currently not represented in the dataset.
		Enrollment (new)	This attribute was collected at the beginning of the semester, as soon as it is clear that an applicant started as a new student.	*Yes* = Applicant is enrolled and became a new student. *No* = Applicant has not enrolled at the case university.

The attribute *Status* has been determined four times throughout the application process and therefore consists of 4 individual attributes – *Status(SevenWeeks)*, *Status(FiveWeeks)*, *Status(ThreeWeeks)*, and *Status(OneWeek)*. As a rule, the application process takes place in several steps. First, an applicant sends an online application and therefore receives the status *Received*. After submitting the required printed application, the submitted documents will be examined and, if applicable, declared *Valid*. If the admission requirements are not fulfilled, the applicant is *Excluded*. If the applicant is considered for a later approval because of missing documents that can be submitted later in the application process, the status changes to *Later*. Once a valid application has been submitted and the documents are thoroughly examined, the applicant receives an admission offer, which is represented by the status *Admission*. The applicant can then actively accepted the

admission offer, which is marked with the status *EnrollmentRequested*, or the offer can be rejected, which is marked with the status *OfferRejected*.

The *Status* attribute has been collected in a two-week period and tracks the progress of the application, from submission to enrollment request. Table 13 shows the distribution of the data records over the four *Status* attributes. Therefore, the majority of the applicants have not yet applied[67] until seven weeks prior to the admission deadline – *Status_SevenWeeks*. This changes the closer the application deadline comes, and almost all the applications have been received one week before the admission closes. Only 47 individuals seem to have not yet applied, which can only be assumed because the information in question is missing in the dataset. For the remaining examples, more than 50% of all the applicants received an admission offer one week prior to the application deadline, and of those 470 have already requested the enrollment. Of these 470 applicants, 434 were enrolled, which corresponds to a 92.3% enrollment rate. Of the applicants with the *Status_OneWeek*(*Admission*), only 253 were enrolled, that is 24.2%. As already mentioned, 413 applicants were excluded by the study and examination office, which is why they were excluded from the following analysis as well because they are not considered as valid.

Table 13. Distribution of the four *Status* attributes.

Status	SevenWeeks		FiveWeeks		ThreeWeeks		OneWeek	
	Number	%	Number	%	Number	%	Number	%
Received	558	18.5%	509	16.8%	780	25.8%	533	17.6%
NotApplied	1828	60.4%	1358	44.9%	842	27.8%		
Valid	475	15.7%	285	9.4%	405	13.4%	15	0.5%
Admission	16	0.5%	440	14.6%	289	9.6%	1045	34.6%
Later	100	3.3%	74	2.4%	79	2.6%	421	13.9%
OfferRejected			17	0.6%	68	2.2%	80	2.6%
EnrollmentRequested			188	6.2%	261	8.6%	470	15.5%
Excluded			106	3.5%	253	8.4%	413	13.7%
Missing	47	1.6%	47	1.6%	47	1.6%	47	1.6%
Total	**3024**	**100.0%**	**3024**	**100.0%**	**3024**	**100.0%**	**3024**	**100.0%**

The dataset contains the application data of six Bachelor programs, which are Business Administration (BA), Business Administration in Health (BAH), Information Management (IM), Information Management Automotive (IMA), Industrial Engineering (IE), and Industrial Engineering in Logistics (IEL). The distribution of the applicants into the individual study programs is shown in Table 14. This table presents the number of applicants who enroll and the number of non-enrolled applicants to indicate the enrollment rate.[68] The BA program receives the majority of valid applications, with 37.3% of the total applicants applying for this program. The second most popular program is

[67] The status *NotApplied* has been concluded at the beginning of the semester because there have not been any records of the new student at any time of data collection.

[68] $Enrollment\ rate = \left(\frac{amount\ of\ new\ students}{total\ amount\ of\ valid\ applicants\ for\ the\ program}\right) * 100.$

BAH, which received 17.3% of the valid applications. The IMA and IEL programs have the least number of applications.

Table 14. Distribution of the applications between the investigated study programs.

Study Program	Valid applications in 2018		Enrollment(Yes)	Enrollment(No)	Enrollment rate
BA	973	37.3%	241	732	24.8%
BAH	451	17.3%	133	318	29.5%
IE	446	17.1%	101	345	22.6%
IEL	162	6.2%	60	102	37.0%
IM	417	16.0%	182	235	43.6%
IMA	162	6.2%	57	105	35.2%
Total	**2611**	**100.0%**	**774**	**1837**	**29.6%**

The BA study program also has the most enrollments. However, compared to the total number of valid applications for this program, these 241 students only account for 24.8% of the applicants. In contrast, the program IM has an enrollment rate of 43.6%, and the IEL program of 37.0%. It appears that a large number of applicants are consciously opting for these study programs, which could also suggest that these programs have unique features that set them apart from the programs of competing universities. The IE course has the lowest enrollment rate at just 22.6%. This is an admission-restricted program and therefore only so many applicants can be admitted as available study places, which could explain the low enrollment rate. Nevertheless, the BA program with an enrollment rate of only 24.8% is not an admission-restricted program. However, the subject is offered by many competing universities. Accordingly, applicants interested in studying business administration have a wide choice. In order to increase the rate of enrollment, the case universities BA program could benefit from identifying or establishing unique features that make it more attractive than competitor programs.

The attribute *MultipleAppli* registers those applicants who have applied for more than one program at the case university. A total of 563 individuals, that is 21.6% of the valid applications, applied for more than one program and 214 of those enrolled at the beginning of the semester. This is an enrollment rate of almost 38.0%. In addition, the percentage distribution of the attribute shown in Figure 22 indicates that 27.6% of the applicants who enrolled at the beginning of the semester had

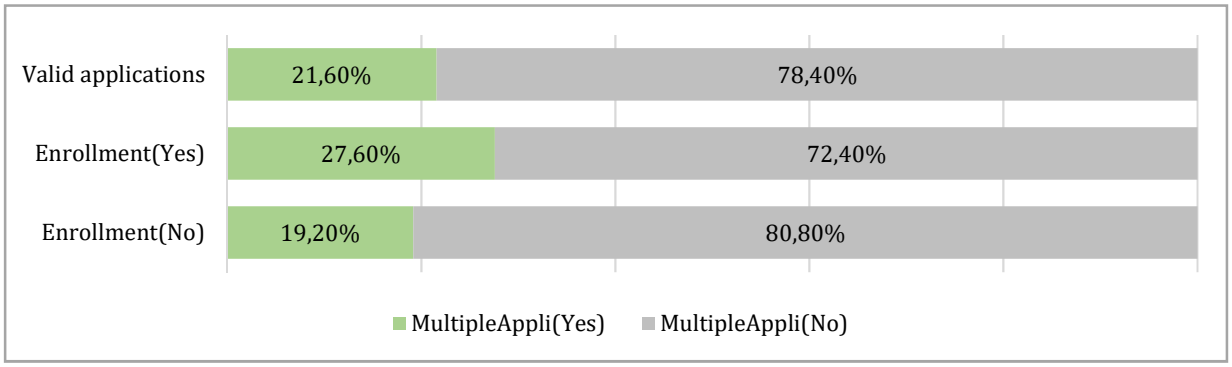


Figure 22. Distribution of the attribute *MultipleAppli*.

applied multiple times to the case university, while only 19.2% of the applicants who did not enroll had applied multiple times. Consequently, it seems that those applicants who have submitted multiple applications are more willing to enroll at the beginning of the semester. This leaves room for the assumption that they prefer the location of their study place to the content of the study program.

The *AppliNumber* attribute is the basis of the *MultipleAppli* consolidated attribute and reflects the number of applications submitted by each applicant. The distribution in the attribute given in Table 15 shows that 78.4% of those who submitted a valid application applied only to one program, 15.6% applied for two, and 4.3% applied for three programs. Accordingly, only 1.6% of the examples in the dataset applied to more than three programs. However, the enrollment rate increases with the number of applications an individual submits to the case university. Accordingly, of the records in the dataset applying for more than 3 programs, 57.1%[69] enroll at the beginning of the semester, while only 29.2%[70] of the applicants with 1 to 3 applications enroll. Consequently, in particular, applicants who submit several applications to the case university seem to make their study choice dependent on the university and not on the content of the program.

Table 15. Distribution of the *AppliNumber* attribute.

AppliNumber	Valid applications in 2018		Enrollment(Yes)	Enrollment(No)	Enrollment rate
1	2048	78.4%	562	1486	27.4%
2	408	15.6%	146	262	35.8%
3	113	4.3%	42	71	37.2%
4	29	1.1%	14	15	48.3%
5	11	0.4%	8	3	72.7%
6	2	0.1%	2		100.0%
Total	**2611**	**100.0%**	**774**	**1837**	**29.6%**

The priority an applicant has for a given study program at the case university is specified in the *Priority* attribute. If the applicant is applying for only one program, this attribute is generated automatically and has the distinction 1. If any person applies for more than one program, they must mark the applications according to their preferences. The courses are marked in descending order, starting from 1, which is the highest priority. To create a consistent database in which each applicant is represented only once, for individuals with multiple valid applications, the individual record that remains in the dataset is randomly selected. If the applicant with multiple applications enrolled at the beginning of the semester, this data record has been stored in the dataset. Therefore, the distribution in the *Priority* attribute is only relevant for the applicants who enroll, also referred to as new students. This distribution is shown in Table 16.

[69] 24 out of 42 individuals that submitted between 4 and 6 applications.

[70] 750 individuals of the 2569 that submitted between 1 and 3 applications.

Table 16. Distribution of the new students in the attribute *Priority*.

Priority	Number	Fraction
1	661	85.4%
2	53	6.8%
3	10	1.3%
4 - 6	3	0.4%
Missing	47	6.1%
Total	**774**	**100%**

Accordingly, 85.4% of the new students enrolled in a program for which they have applied for with a *Priority(1)*, 6.8% for a program with the *Priority(2),* and 1.3% for a program they applied for at the *Priority(3)*. Therefore, most of the applicants who enrolled at the beginning of the semester received an admission offer for the program of their first choice.

Table 17 shows the descriptive analysis of the attributes *Age*, *HEEQGrade*, and *PreviousSemesters*. The *Age* attribute and the *PreviousSemesters* attribute have a broad distribution, and their maximum value is high on average. However, only a 3.4% portion of applicants is older than 26 years (88 examples)[71] and 2% (52 examples) have more than 8 semesters of previous study experience.[72] Nevertheless, these examples are legitimate data points and are included in the analysis. The comparison of the attributes between the two target classes shows that the applicants enrolling at the beginning of the semester are with an average age of 21 years, slightly older than those who decide against the case university. The average HEEQ grade of 2.8 and the average study experience of 1.1 semester differ only minimally in the two target groups as well. It is therefore assumed that none of the three attributes has an impact on the likelihood of an applicant enrolling.

Table 17. Descriptive analysis results of the *Age*, *HEEQGrade* and *PreviousSemesters* attributes.

Attribute	Valid applications in 2018				Enrollment(Yes)				Enrollment(No)			
	Minimum	Maximum	Average	Deviation	Minimum	Maximum	Average	Deviation	Minimum	Maximum	Average	Deviation
Age	17.0	49.0	20.9	2.6	17.0	38.0	21.0	2.7	17.0	49.0	20.8	2.6
HEEQGrade	1.0	4.0	2.8	0.6	1.0	3.9	2.8	0.5	1.0	4.0	2.8	0.6
Previous Semesters	0.0	24.0	1.1	2.4	0.0	17.0	1.2	2.3	0.0	24.0	1.1	2.4

The information whether a student has previous study experience is summarized in the *FirstSemester* attribute, which indicates whether an applicant has – *FirstSemester(No)* – or has not – *FirstSemester(Yes)* – previous study experience at that or any other university. The results of the descriptive analysis are shown in Figure 23 and indicate that the portion of individuals with previous study experience is higher for new students than for those who are not enrolled. Accordingly, 32.6% of the new students have experience of studying, while only 28.0% of those who have not

[71] Two standard deviations larger than the mean.

[72] Three standard deviations larger than the mean.

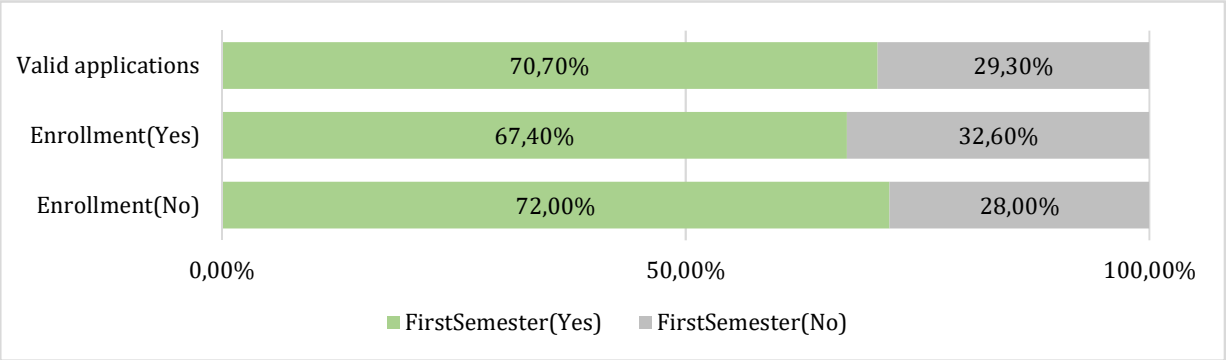


Figure 23. Distribution of the *FirstSemester* attribute.

enrolled have previous study experience. This is also reflected in the enrollment rate, which is 28.3%[73] for those without study experience and 32.9%[74] for those with study experience. Consequently, it seems that the study experience at this or any other university slightly increases the likelihood of enrollment.

Table 18 evaluates the distribution of the *Gender* attribute. The portion of female and male applicants in the valid applications is almost balanced, which can also be noticed for the enrollment rate. Accordingly, 29.5% of the female applicants are enrolled, which represents a slightly lower enrollment rate than that of the male applicants. However, it is assumed that the *Gender* attribute has no impact on the enrollment probability.

Table 18. Distribution of the *Gender* attribute.

Gender	Valid applications in 2018		Enrollment(Yes)	Enrollment(No)	Enrollment rate
Female	1307	50.1%	385	922	29.5%
Male	1304	49.9%	389	915	29.8%
Total	**2611**	**100.0%**	**774**	**2247**	**25.7%**

The attribute *Nationality* is descriptively evaluated in Table 19. The distribution of the attribute shows that the case university has significantly more national than international applicants.

Table 19. Distribution of the *Nationality* attribute.

Nationality	Valid applications in 2018		Enrollment(Yes)	Enrollment(No)	Enrollment rate
German	2353	90.1%	704	1649	29.9%
Foreign	258	9.9%	70	188	27.1%
Total	**2611**	**100.0%**	**774**	**1837**	**29.6%**

Nevertheless, the enrollment rate shows only a small difference with 27.1% of the 258 international applicants enrolled. However, this attribute does not take into account the place of residence

[73] Of the 1845 applicants without previous study experience 522 enrolled.
[74] Of the 766 applicants with previous study experience, 252 are enrolled.

at the time of application, and applicants with a foreign nationality may already live in Germany. This assumption is supported by the fact that the applications in the dataset are intended for programs that are mainly taught in German.

The deviation of the *BirthCountry* attribute illustrated in Table 20 shows that 92.3% of the applicants who submitted a valid application are born in Germany. The remaining 202 applicants are born in 68 different countries. Most foreign applicants come from Russia with 21 individuals, followed by Kazakhstan with 13 applicants, Ukraine with 11 applicants, Turkey with 10 applicants, and Syria with 8 applicants. Accordingly, many foreign birth countries are represented in the dataset by fewer than 8 individuals and 28 birth countries only have one representative. When filtering the dataset for applicants who became new students at the beginning of the semester, the *BirthCountry* attribute consists of 33 distinctions, but of the 774 new students, 718 (92.8%) were born in Germany. Of the 56 foreign-born students, 7 are from Ukraine and 5 from Kazakhstan, while only 3 new students were born in Russia, and 2 new students were born in Turkey. The enrollment rate for applicants born in Turkey and Russia are correspondingly low.

Table 20. Distribution of the *BirthCountry* attribute.

BirthCountry	Valid applications in 2018		Enrollment(Yes)	Enrollment(No)	Enrollment rate
Germany	2409	92.3%	718	1691	29.8%
Russia	21	0.8%	3	18	14.3%
Kazakhstan	13	0.5%	5	8	38.5%
Ukraine	11	0.4%	7	4	63.6%
Turkey	10	0.4%	2	8	20.0%
Syria	8	0.3%	2	6	25.0%
Other	139	5.3%	37	102	26.6%
Total	**2611**	**100.0%**	**774**	**1837**	**29.6%**

The *Distance_PlaceofBirth* attribute shown in Table 21 defines how far away the applicant was born from the place of the case university. In the dataset, 70.7% are born within 100km from the case university, which are 1846 applicants. Of these, 629 were enrolled at the beginning of the semester, which is 81.3% of all new students. Thus, it seems that students born in the region of the case university are enrolled more frequently than students with a foreign birthplace or a birthplace more than 100km away. This is also reflected in the enrollment rates, which is highest for applicants born within 100km of the case university.

Table 21. Distribution of the *Distance_PlaceofBirth* attribute.

Distance_PlaceofBirth	Valid applications in 2018		Enrollment(Yes)	Enrollment(No)	Enrollment rate
Radius100	1846	70.7%	629	1217	34.1%
Radius200	337	12.9%	55	282	16.3%
Radius300	55	2.1%	10	45	18.2%
Radius > 300	172	6.6%	27	145	15.7%
Abroad	201	7.7%	53	148	26.4%
Total	**2611**	**100.0%**	**774**	**1837**	**29.6%**

The *Residence* attribute indicates the place where the applicant lives at the time of application. In total, this attribute includes 893 distinctions, and 571 of these residences are represented by one person only. To compress the information, the *Distance_Residence* attribute was generated, which indicates how far the place of residence is located around the case university. The distribution in the attribute between the two target classes is shown in Figure 24. Of the total 2611 applicants with a valid application, 2100 (80.4%) live no further than 100km from the case university, 325 (12.4%) live no further than 200km, and 47 (1.8%) live no further than 300km at the time of application. Accordingly, only 120 applicants (4.6%) live more than 300km away at the time of application, and only 19 (0.7%) come from abroad. If the analysis of the attribute focuses on the 774 applicants enrolling at the beginning of the semester, the numbers show that 91.1% of the new students (705 individuals) are from a radius within 100km from the case university, and 58 (7.5%) live no further away than 200km at the time of application. Only 11 new students applied from further than 300km away. This supports the assumption made in the analysis of the *Nationality* attribute, in which it has been found that applicants with foreign nationality already reside in Germany at the time of application. Accordingly, the case university has mainly applicants and students from the region, and it seems that a place of residence near the site of the case university increases the likelihood of enrollment. Therefore, the case university seems to have a high degree of regional recognition.

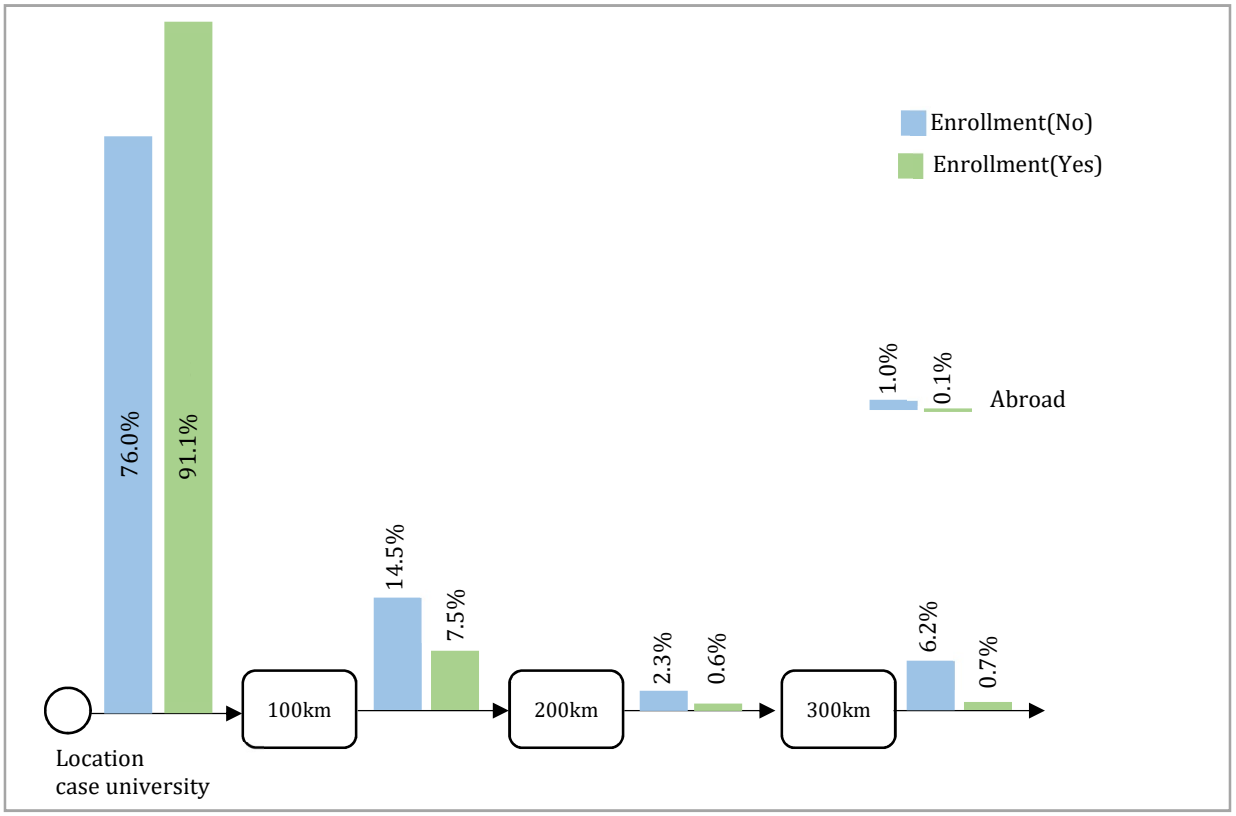


Figure 24. Distribution of the *Distance_Residence* attribute.

The size of the city or town where the applicant resides at the time of application has been recorded in the *TownSize_Residence* attribute to see if the level of the urbanization in which the applicant lives impacts on enrollment behavior. This attribute was developed by manually collecting the population of the residences. Subsequently, they were summarized in the categories described in Table 12. The distribution of the *TownSize_Residence* attribute is shown in Table 22. Enrollment rates for applicants from cities and large towns are highest, while the majority of new students live in small towns at the time of application. Of the applicants who live in a big city, only 17.1% enrolled.

Table 22. Distribution of the *TownSize_Residence* attribute.

TownSize_Residence	Valid applications in 2018		Enrollment(Yes)	Enrollment(No)	Enrollment rate
Community	514	19.7%	158	356	30.7%
SmallTown	846	32.4%	246	600	29.1%
Town	428	16.4%	111	317	25.9%
BigTown	302	11.6%	98	204	32.5%
City	379	14.5%	139	240	36.7%
BigCity	123	4.7%	21	102	17.1%
Abroad	19	0.7%	1	18	5.3%
Total	**2611**	**100.0%**	**774**	**1837**	**29.6%**

The evaluation of the *HEEQDegree* attribute in Figure 25 shows that 60.2% of the applicants who have submitted a valid application have an *AHR* university entrance degree and 36% have an *FHR* university of applied sciences entrance degree, while only 2.8% have an entrance qualification from a foreign country. This is also reflected in the new students and the applicants who did not enroll. Nevertheless, the proportion of individuals with an *AHR* is slightly higher for the new students, while the proportion of applicants with an *fgHR* and a HEEQ from abroad is higher for the non-enrolled applicants. This observation corroborates the assumptions made earlier that applicants of foreign origin are less likely to become enrolled at the case university. However, the nature of the HEEQ is not expected to influence the likelihood of an applicant enrolling.

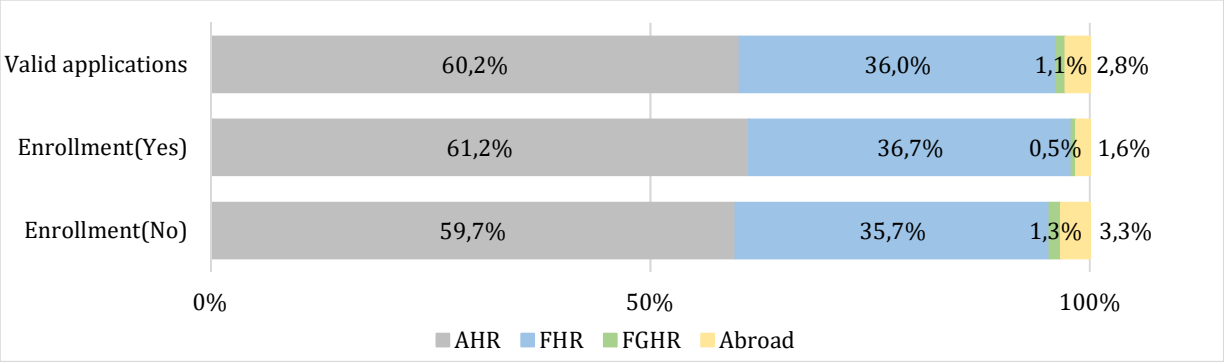


Figure 25. Distribution of the *HEEQDegree* attribute.

Another attribute in our dataset is *HEEQCountry*, which indicates whether an applicant received the HEEQ in Germany or abroad, which is evaluated in Table 23. Of the 2611 applicants with a valid

application, 2535 received their HEEQ in Germany. The detailed investigation of the applicants enrolled at the beginning of the semester shows that 98.4% of the new students have a German HEEQ degree, which is also reflected in the enrollment rate of 30.1%. Given the previous analysis of the *Distance_Residence* and *Nationality* attributes, this was assumed.

Table 23. Distribution of the *HEEQCountry* attribute.

HEEQCountry	Valid applications in 2018		Enrollment(Yes)	Enrollment(No)	Enrollment rate
German	2535	97.2%	762	1773	30.1%
Foreign	76	2.8%	12	64	15.8%
Total	**2611**	**100.0%**	**774**	**1837**	**29.6%**

The *HEEQDistrict* attribute indicates the district where the HEEQ was acquired. These are in total 231 different districts. Therefore, the attribute has been condensed in the *Distance_HEEQDistrict* attribute according to the distance between the case university site and the district where the HEEQ degree was acquired. The results are shown in Table 24. Of all the applicants in the dataset, 77.2% received their HEEQ degree within the region, and 33.8% of these applicants enrolled at the beginning of the semester. Of the 337 applicants who completed their HEEQ degree within 200km of the case university, only 19% enrolled, and for the applicants who acquired a HEEQ degree further away than 200km, the enrollment rate is even lower. This finding again supports the assumption that a regional connection of the applicant increases the likelihood of enrollment.

Table 24. Distribution of the *Distance_HEEQDistrict* attribute.

Distance_HEEQDistrict	Valid applications in 2018		Enrollment(Yes)	Enrollment(No)	Enrollment rate
Radius100	2015	77.2%	682	1333	33.8%
Radius200	337	12.9%	64	273	19.0%
Radius300	45	1.7%	5	40	11.1%
Radius > 300	138	5.3%	12	126	8.7%
Abroad	76	2.9%	11	65	14.5%
Total	**2611**	**100.0%**	**774**	**1837**	**29.6%**

The attribute *Apprenticeship* indicates whether an applicant has completed vocational training before applying at the case university. An apprenticeship in Germany is a necessary step for almost every profession and is usually completed within 3 years. Table 25 shows the descriptive analysis of the attribute, and the results display that the largest proportion of applicants (78.7%) did not complete any apprenticeship prior to application to the case university.

Table 25. Distribution of the *Apprenticeship* attribute.

Apprenticeship	Valid applications in 2018		Enrollment(Yes)	Enrollment(No)	Enrollment rate
Yes	557	21.3%	167	390	30.0%
No	2054	78.7%	607	1447	29.6%
Total	**2611**	**100.0%**	**774**	**1837**	**29.6%**

The enrollment rate for both applicants with an apprenticeship and applicants without an apprenticeship is about 30%. Accordingly, having or not having a completed apprenticeship does not seem to influence the probability of enrollment.

The *TimeHEEQDegree-Application* attribute was generated from the existing attributes to determine how much time passed between the case university application and the time the HEEQ was completed. Therefore, the months between the application date and the date of the HEEQ were counted and condensed. Both dates are available in the original applicant data. The distribution of the generated attribute is shown in Table 26. It is noticeable that, in particular, those who completed their HEEQ degree between 6 and 12 months and those who graduated between 19 and 24 months prior to their application to the case university have a high enrollment rate. On the contrary, the enrollment rate for applicants that finished their HEEQ less than 6 months before the application deadline is relatively low at 25%. It appears that a good proportion of the applicants who become new students take some time off between completing their HEEQ and applying for a case university study program. Unfortunately, there is not much information about the reason for the late application. This can be anything from traveling over completing an internship or gaining first study experience at different programs and universities.

Table 26. Distribution of the *TimeHEEQDegree-Application* attribute.

TimeHEEQDegree-Application	Valid applications in 2018		Enrollment(Yes)	Enrollment(No)	Enrollment rate
< 6 Months	717	27.5%	179	538	25.0%
between 6-12 Months	568	21.8%	209	359	36.8%
between 13-18 Months	316	12.1%	54	262	17.1%
between 19-24 Months	232	8.9%	109	123	47.0%
2-3 Years	232	8.9%	67	165	28.9%
3-4 Years	241	9.2%	74	167	30.7%
4-5 Years	131	5.0%	39	92	29.8%
5-6 Years	87	3.3%	20	67	23.0%
> 6 Years	87	3.3%	23	64	26.4%
Total	**2611**	**100.0%**	**774**	**1837**	**29.6%**

Only the comparison of the *TimeHEEQDegree-Application* attribute with the *FirstSemester* or *Apprenticeship* attributes in crosstabs can give an indication of what the new students did before their studies, which was done in Table 27. Accordingly, only a small portion of new students who finish their HEEQ between 0 and 18 months before the application have prior study experiences or an apprenticeship. Therefore, the available data does not indicate what they have done in this gap. Applicants who apply later than 19 months after completing their HEEQ appear to be delayed in the application because they have already studied in a previous program at this or another university.

Table 27. Crosstab of the *TimeHEEQDegree-Application* attribute and the *Apprenticeship* and *FirstSemester* attributes, with respect to new students.

TimeHEEQDegree-Application	Enrollment(Yes)	Apprenticeship				FirstSemester			
		Yes	*%*	*No*	*%*	*Yes*	*%*	*No*	*%*
< 6 Months	179	**42**	23.5%	**137**	76.5%	**175**	97.8%	**4**	2.2%
between 6-12 Months	209	**36**	17.2%	**173**	82.8%	**171**	81.8%	**38**	18.2%
between 13-18 Months	54	**5**	9.3%	**49**	90.7%	**51**	94.4%	**3**	5.6%
between 19-24 Months	109	**19**	17.4%	**90**	82.6%	**37**	33.9%	**72**	66.1%
2-3 Years	67	**6**	9.0%	**61**	91.0%	**23**	34.3%	**44**	65.7%
3-4 Years	74	**27**	36.5%	**47**	63.5%	**33**	44.6%	**41**	55.4%
4-5 Years	39	**16**	41.0%	**23**	59.0%	**19**	48.7%	**20**	51.3%
5-6 Years	20	**11**	55.0%	**9**	45.0%	**6**	30.0%	**14**	70.0%
> 6 Years	23	**5**	21.7%	**18**	78.3%	**7**	30.4%	**16**	69.6%
Total	**774**				**774**				**774**

5.1.2 Data preparation and analysis plan

After the dataset has been analyzed descriptively, the deviation of the numerical attributes in the dataset has been examined more closely as some of the records are not in the normal distribution of the attributes *Age* and *PreviousSemesters*. These records were studied individually to detect data entry or calculation errors. Such errors were not noticed. Accordingly, the data points are legitimate and remain in the dataset.

The search for outliers taking into account all attributes in the dataset was continued using the LOF method described in Section 3.2.2. Therefore, the dataset has been normalized, which is recommended for distance and density-based outlier detection methods (Kotu et al. 2015: 336). The numeric attributes in the dataset are normalized with a *min-max normalization*, and the nominal attributes with dummy encoding. Both approaches are described in Section 3.4.4. After transformation, the LOF method is executed. The result is a measure indicating whether a data record can be considered an outlier. Values near 1 are non-outliers, while values that differ significantly from 1 can be considered as outliers (Breunig et al. 2000). The results obtained show a range in the LOF value between 0.92 and 1.51, with an average of 1.10 and a standard deviation of 0.08. The visual analysis of the calculated LOF factor is shown in Figure 26, where both the *x-axes* and the *y-axes* represent the calculated LOF values of the data records. The data records with values above a LOF of 1.45 are obviously in a low-density region, far from the adjacent records. This concerns only 3 data examples that are excluded from the following analysis.

In addition, the dataset was examined for missing values. Unfortunately, 47 records do have missing values for the *Priority* attribute and for all four *Status* attributes. Since all of these attributes are missing for the 47 data records, it is assumed that these individuals applied to the case university within one week of the application deadline, since after this time period, no more data collection took place.

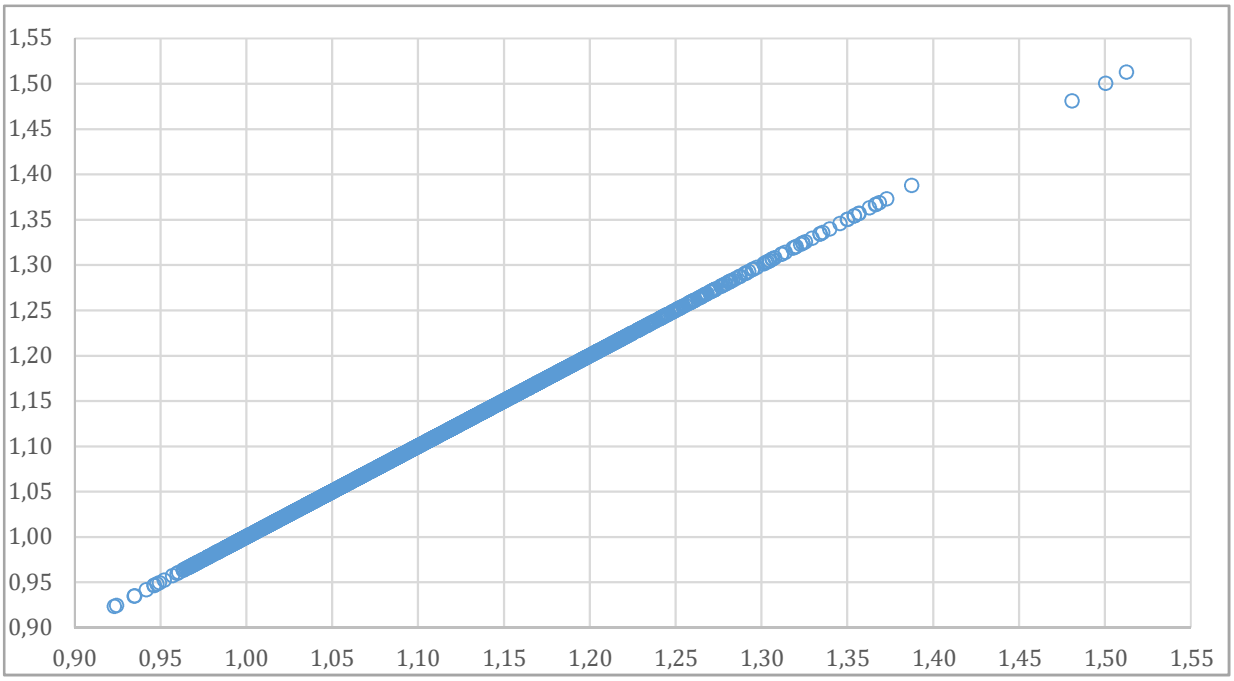


Figure 26. Visual display of the LOF calculated.

Accordingly, these records, along with the target attribute *Enrollment*, were collected at the beginning of the semester when checking who of the applicants enrolled in any of the study programs. Following consultation with a representative of the case university's study and examination office and the fact that all the values of the four *Status* attributes are missing for the 47 examples, it was decided to exclude these individuals from the analysis, as it is assumed that the *Status* attributes are an important predictor of applicant enrollment. Thus, 2561 examples remain in the dataset *Applications2018*.

Of these 2561 data examples, 727 belong to the target class *Enrollment(Yes)* and 1834 to the target class *Enrollment(No)*. Accordingly, the dataset is unbalanced, which could affect the *accuracy* of the analysis results, and it is therefore recommended to balance the dataset (Larose et al. 2015: 166). A dataset can be balanced either by oversampling or by undersampling, as described in Section 3.2.4. Before sampling is performed, the dataset *Applications2018* is split into a training and a testing set. The training part of the dataset is used for model generation and validation, while the testing part is used to test the best performing models to examine their suitability for the decision-makers in the enrollment department in a real-life scenario. If the splitting of the dataset is performed after the original dataset is balanced, the testing dataset will be balanced as well, which does not represent the realities. Accordingly, the dataset *Applications2018* is split into the *TrainingSet2018* and the *UnseenTestSet*. The *TrainingSet2018* contains 85% of the data records, i.e. 2177 records.[75] The remaining 15% of the dataset *Application2018* built the *UnseenTestSet*. Of the 384

[75] Of these, 618 records belong to the target class *Enrollment(Yes)* and 1559 to the target class *Enrollment(No)*.

records in this dataset, 109 belong to the target class *Enrollment(Yes)*, and 275 to the target class *Enrollment(No)*.

After the split, two additional datasets were created from the *TrainingSet2018*, a *DataSetDS2018* and a *DataSetUS2018*. For the down-sampled dataset *DataSetDS2018*, 618 records from the overrepresented target class *Enrollment(No)* were randomly chosen to fit the size of the underrepresented class. Thus, the dataset consists in total of 1236 data records, each target class being represented by 618 records. In the oversampled dataset *DataSetUS2018*, 941 synthetic data records were created for the underrepresented target class *Enrollment(Yes)* to match the 1559 examples that belong to the overrepresented target class *Enrollment(No)*. Accordingly, the *DataSetUS2018* includes a total of 3118 records.

All three datasets – *TrainingSet2018*, *DataSetDS2018*, and *DataSetUS2018* – contain 26 attributes that are considered for analysis. Therefore, it does not seem necessary to use feature selection to minimize the number of attributes in order to minimize the computational time and complexity of the DM models. Nevertheless, feature selection may be helpful in determining the attributes that strongly correlate with the target variable or those that are strongly correlated with each other and may not introduce additional information into the predictive model. Therefore, a *forward selection* was used in some of the following analyses as a step before the model generation. In addition, the expertise of the applications and enrollment department of the case university has been included in the selection of attributes that are believed to influence the target variable. Therefore, members of this department have been asked to propose attributes that might influence the target variable. The following attributes have been suggested: *Status_OneWeek*, *Status_ThreeWeeks*, *Status_FiveWeeks*, *Status_SevenWeeks*, *Priority*, *HEEQDegree*, *AppliNumber*, *HEEQGradeComp*, *TimeHEEQDegree-Application*, *Distance_Residence*, and *Distance_HEEQDistrict*.

Therefore, the analysis was performed considering the following three scenarios:

- Scenario A contains all available attributes. For the attributes that are represented in a compacted attribute – e.g. *Residence* – the latter is included in the analysis – e.g. *Distance_Residence*.
- Scenario B includes those attributes into the model building that have been defined by the study and examination department as factors influencing the target variable.
- Scenario C includes the attributes into the model building that have been identified as influential predictors by feature selection.

5.1.3 Decision Tree models

The decision tree models are calculated using the classification tree approach based on the *C4.5* algorithm described in Section 3.4.3 of this thesis. Several models have been generated for each of

the three datasets and for each of the three scenarios. The attributes contained in Scenario C were identified by stepwise *forward selection* based on a decision tree. Accordingly, Scenario C is generated for each of the training datasets individually and includes those attributes in the model building that are identified to enhance the performance of a classification tree model generated within the stepwise *forward selection*.

To generate well-performing tree models that are able to classify new and unseen datasets, pruning was applied to the decision trees. Consequently, the *minimal gain*, the *minimal size for split*, and the *minimum leaf size* have been adjusted to minimize the *error rates* of the models and to identify classification trees of an appropriate size. Therefore, each generated model is evaluated with *cross-validation*, which is introduced in Section 3.5. The performance validation results for the tree models with the lowest *error rates* are presented in Table 28. These models are described by Larose et al. (2015: 163-164) as those with optimal complexity. The table displays the *accuracy*, *error rate*, *recall*, and *precision*. In addition, the *F-score* is given, which is an important performance measure for the unbalanced *TrainingSet2018*, as it is a harmonic mean that accounts for the fluctuations between the *recall* and the *precision* for the prediction of the positive class. If only the *accuracy* was considered, Model 1, Model 2, and Model 3 would be preferred as they have the highest *accuracy* with at least 86.7% of correct classifications. However, the *recall* for the *Enrollment(Yes)* class is only at 59% on average, which is lower than the *recall* for this class of the other models.

Table 28. Performance results of validating the decision tree models generated.

Model Name	Dataset	Scenario	Accuracy	Recall		Precision		Error rate	F-score
				True (No)	True (Yes)	Pred (No)	Pred (Yes)		
1	*TrainingSet 2018*	*A*	86.73%	97.75%	58.90%	85.71%	91.23%	13.27%	71.49%
2		*B*	86.91%	97.56%	60.03%	86.03%	90.71%	13.09%	72.18%
3		*C*	86.91%	97.75%	59.55%	85.91%	91.32%	13.09%	72.02%
4	*DataSetDS 2018*	*A*	74.51%	88.67%	60.36%	69.10%	84.20%	25.49%	70.21%
5		*B*	77.82%	88.67%	66.99%	72.87%	85.54%	22.18%	74.90%
6		*C*	79.28%	96.28%	62.30%	71.86%	94.34%	20.72%	74.93%
7	*DataSetUS 2018*	*A*	77.78%	94.61%	60.94%	70.78%	91.88%	22.22%	72.89%
8		*B*	81.14%	75.11%	87.17%	85.41%	77.79%	18.86%	82.25%
9		*C*	78.54%	71.20%	85.89%	83.46%	74.89%	21.46%	79.47%

According to the *F-score*, Model 8 is the best predictive model with a relatively low *error rate* of 18.9%. Therefore, the model validation performance metrics suggest that 81.1% of all cases in the *DataSetUS2018* can be properly classified. It is also suggested that Model 8 correctly recognizes 87.2% of the cases that belong to the target class *Enrollment(Yes)*, which is the highest recognition rate for this class. The Model 9 performance validation measures show the second-best performance with an *error rate* of 21.5%, an *F-score* of 79.5%, and a *recall* for the class *Enrollment(Yes)* of 85.9%. In addition, Model 5 is further investigated because it is the model with the highest *recall* rate for the target class *Enrollment(Yes)* of those models trained exclusively on real data records.

Figure 27 shows Model 9, which has the attribute *Status_OneWeek* as the tree root. Accordingly, if an applicant requested admission to a study program one week before the application deadline, it is very likely that he or she enrolls at the beginning of the semester. In addition, the model indicates that enrollment is likely if an applicant has received an admission offer one week prior to the application deadline, whose HEEQ grade is below 1.3 and whose place of residence is near the university – *Distance_Residence(Radius100)*. However, the distribution of the cases in the relevant tree leaf shows that 47.7% of the examples belong to the class *Enrollment(No)*. Therefore, the tree rule is not conclusive and should be further investigated as soon as more data records are available.

```
Status_OneWeek(Admission)
|  HEEQGrade > 1.3
|  |  Distance_Residence(Abroad): No {No=3, Yes=0}
|  |  Distance_Residence(Over500): No {No=8, Yes=2}
|  |  Distance_Residence(Radius100): Yes {No=514, Yes=564}
|  |  Distance_Residence(Radius200): No {No=92, Yes=24}
|  |  Distance_Residence(Radius300): No {No=20, Yes=1}
|  |  Distance_Residence(Radius400): No {No=11, Yes=0}
|  |  Distance_Residence(Radius500): No {No=7, Yes=0}
|  HEEQGrade ≤ 1.3: No {No=4, Yes=0}
Status_OneWeek(EnrollmentRequested): Yes {No=35, Yes=899}
Status_OneWeek(Later): No {No=339, Yes=64}
Status_OneWeek(OfferRejected): No {No=65, Yes=0}
Status_OneWeek(Received): No {No=449, Yes=5}
Status_OneWeek(Valid): No {No=12, Yes=0}
```

Figure 27. Decision tree Model 9.

Figure 28 shows Model 8, which also contains the *Status_OneWeek* attribute as the root of the tree. The model shows that applicants requesting enrollment one week prior to the application deadline are likely to enroll at the beginning of the semester. In addition, the tree indicates that applicants are likely to enroll if they have received an admission offer at least one week prior to the application deadline and have the following additional characteristics:

(1) *Distance_HEEQDistrict(Radius100)* AND *HEEQGradeComp(good)* AND *TownSize_Residence(City)*

(2) *Distance_HEEQDistrict(Radius100)* AND *HEEQGradeComp(good)* AND *TownSize_Residence(Community)* AND *Status_FiveWeeks(NotApplied)* AND *HEEQDegree(FHR)*

(3) *Distance_HEEQDistrict(Radius100)* AND *HEEQGradeComp(good)* AND *TownSize_Residence(Town)* AND *AppliNumber(2)*

(4) *Distance_HEEQDistrict(Radius100)* AND *HEEQGradeComp(sufficient)* OR *HEEQGradeComp(satisfactory)*

```
Status_OneWeek(Admission)
|  Distance_HEEQDistrict(Abroad): No {No=15, Yes=2}
|  Distance_HEEQDistrict(Over500): No {No=6, Yes=0}
|  Distance_HEEQDistrict(Radius100)
|  |  HEEQGradeComp(good)
|  |  |  TownSize_Residence(BigCity): No {No=6, Yes=0}
|  |  |  TownSize_Residence(BigTown): No {No=15, Yes=13}
|  |  |  TownSize_Residence(City): Yes {No=17, Yes=33}
|  |  |  TownSize_Residence(Community)
|  |  |  |  Status_FiveWeeks(Admission): No {No=5, Yes=0}
|  |  |  |  Status_FiveWeeks(NotApplied)
|  |  |  |  |  HEEQDegree(AHR): No {No=16, Yes=11}
|  |  |  |  |  HEEQDegree(FHR): Yes {No=8, Yes=10}
|  |  |  |  Status_FiveWeeks(Received): No {No=5, Yes=2}
|  |  |  |  Status_FiveWeeks(Valid): No {No=11, Yes=3}
|  |  |  TownSize_Residence(SmallTown): No {No=63, Yes=48}
|  |  |  TownSize_Residence(Town)
|  |  |  |  AppliNumber(1): No {No=34, Yes=7}
|  |  |  |  AppliNumber(2): Yes {No=2, Yes=3}
|  |  HEEQGradeComp(satisfactory): Yes {No=300, Yes=388}
|  |  HEEQGradeComp(sufficient): Yes {No=23, Yes=35}
|  |  HEEQGradeComp(very good): No {No=6, Yes=1}
|  Distance_HEEQDistrict(Radius200): No {No=90, Yes=29}
|  Distance_HEEQDistrict(Radius300): No {No=15, Yes=2}
|  Distance_HEEQDistrict(Radius400): No {No=11, Yes=1}
|  Distance_HEEQDistrict(Radius500): No {No=11, Yes=3}
Status_OneWeek(EnrollmentRequested): Yes {No=35, Yes=899}
Status_OneWeek(Later): No {No=339, Yes=64}
Status_OneWeek(OfferRejected): No {No=65, Yes=0}
Status_OneWeek(Received): No {No=449, Yes=5}
Status_OneWeek(Valid): No {No=12, Yes=0}
```

Figure 28. Decision tree Model 8.

Hence, applicants who have received an admission offer will probably enroll if they received their HEEQ within the region of the case university with a grade equal to 2 and if they live in a city at the time of application. If they live in a smaller village and have not yet applied five weeks prior to the application deadline and have a HEEQ for the applied sciences university (FHR), then the tree indicates that they are probably enrolling as well. Because 44.4% of the cases in this leaf belong to the class *Enrollment(No)* the applicability of the tree rule should be re-assessed. The same must be considered for the third tree rule since the leaf in question contains only 5 examples in total, and 2 out of these 5 do belong to the class *Enrollment(No)*. In addition, the model indicates that applicants from the case university region with a HEEQ grade of 3 or worse have a higher probability of university enrollment.

To assess whether Model 8 and Model 9 can assist university decision-makers during the enrollment process, they were tested with the *UnseenTestSet*, although some of the generated rules are not yet conclusive. The performance results of the model testing are shown in Table 29. With a predictive *accuracy* of 76.0% overall, Model 8 is slightly better at classifying the unseen data records correctly. Accordingly, 73.5% of the cases of the class *Enrollment(No)* and 82.6% of the cases of the class *Enrollment(Yes)* were correctly identified. Of all the cases in the test dataset classified as belonging to the class *Enrollment(No)* 91.4% were correctly predicted, while only 55.2% of the

cases predicted to belong to the class *Enrollment(Yes)* are actually cases of this class. Thus, the model can identify a majority of applicants who actually enroll at the beginning of the semester, but for this purpose, 73 applicants who did not enroll at the beginning of the semester were wrongly classified as belonging to the class *Enrollment(Yes)*. The Model 9 is even able to correctly identify 89% of the cases belonging to the target class *Enrollment(Yes)*, but 95 of the examples classified as belonging to this class are actual applicants who do not enroll at the beginning of the semester. Nevertheless, both models are useful to the university decision-makers when they want to identify as many applicants as possible who actually enroll at the beginning of the semester. If universities would be able to offer admissions only to the applicants who are expected to be new students, they could increase their enrollment rate to about 55.2% using Model 8, which is much lower for the *UnseenTestSet* at 28.3%.[76]

Table 29. Performance results of testing Model 8 and Model 9 with the *UnseenTestSet*.

Model 8 | accuracy = 76.04%	**True(No)**	**True(Yes)**	**Class precision**
Pred(No)	202	19	91.40%
Pred(Yes)	73	90	55.21%
Class Recall	73.45%	82.57%	
Model 9 | accuracy = 72.14%	**True(No)**	**True(Yes)**	**Class precision**
Pred(No)	180	12	93.75%
Pred(Yes)	95	97	50.52%
Class Recall	65.45%	88.99%	

The decision tree Model 5 is based on the *DataSetDS2018*, i.e. no synthetic data records were used to generate the tree model. Therefore, this model is also tested with the *UnseenTestSet* to investigate whether the sole use of real-world records increases predictive performance. The performance test results are presented in Table 30, indicating that the model can correctly classify 84.9% of all cases in the *UnseenTestSet*. In detail, the model is able to identify 90.9% of the cases that do not enroll. Of the applicants that enrolled, almost 70% were identified. These 76 examples are 75.3% of all records associated with the target class *Enrollment(Yes)*. As a result, Model 5 identifies fewer of the applicants enrolling at the beginning of the semester but has fewer misclassifications. Therefore, the model is useful for university decision-makers, especially if they want to estimate the exact number of enrollments since the number of unidentified cases belonging to the target class *Enrollment(Yes)* and the number of misclassifications is only different by 7 records.

Table 30. Performance results of testing Model 5 with the *UnseenTestSet*.

Model 5 | accuracy = 84.90%	**True(No)**	**True(Yes)**	**Class precision**
Pred(No)	250	33	88.34%
Pred(Yes)	25	76	75.25%
Class Recall	90.91%	69.72%	

[76] Of the 384 cases in the *UnseenTestSet* 109 enrolled at the beginning of the semester.

The relevant tree model is shown in Figure 29. Again, the *Status_OneWeek* attribute is identified as the best predictor. Furthermore, the model reiterates that applicants who have requested the enrollment one week prior to the application deadline are likely to actually enroll at the beginning of the semester. Applicants who have received an admission offer one week prior to the application deadline and have applied to more than two degree programs at the case university – *AppliNumber ≥ 3* – are also likely to enroll.

If an application for only one program has been approved by the case university, the probability of enrollment for applicants with the following characteristics is high:

(1) HEEQGradeComp(good) AND *TownSize_Residence(BigTown)* OR *TownSize_Residence(City)*

(2) *HEEQGradeComp(satisfactory)* AND *Status_SevenWeeks(Admission)* OR *Status_SevenWeeks(Later)*

```
Status_OneWeek(Admission)
|  AppliNumber(1)
|  |  HEEQGradeComp(good)
|  |  |  TownSize_Residence(BigCity): No {No=2, Yes=1}
|  |  |  TownSize_Residence(BigTown): Yes {No=4, Yes=7}
|  |  |  TownSize_Residence(City): Yes {No=7, Yes=9}
|  |  |  TownSize_Residence(Community): No {No=23, Yes=13}
|  |  |  TownSize_Residence(SmallTown): No {No=31, Yes=16}
|  |  |  TownSize_Residence(Town): No {No=18, Yes=4}
|  |  HEEQGradeComp(satisfactory)
|  |  |  Status_SevenWeeks(Admission): Yes {No=1, Yes=2}
|  |  |  Status_SevenWeeks(Later): Yes {No=4, Yes=5}
|  |  |  Status_SevenWeeks(NotApplied): No {No=80, Yes=76}
|  |  |  Status_SevenWeeks(Received): No {No=10, Yes=3}
|  |  |  Status_SevenWeeks(Valid): No {No=8, Yes=8}
|  |  HEEQGradeComp(sufficient): No {No=11, Yes=11}
|  |  HEEQGradeComp(very good): No {No=5, Yes=1}
|  AppliNumber(2)
|  |  TownSize_Residence(BigCity): No {No=4, Yes=0}
|  |  TownSize_Residence(BigTown): Yes {No=3, Yes=8}
|  |  TownSize_Residence(City): Yes {No=2, Yes=8}
|  |  TownSize_Residence(Community)
|  |  |  Priority(1): No {No=6, Yes=4}
|  |  |  Priority(2): Yes {No=0, Yes=3}
|  |  TownSize_Residence(SmallTown)
|  |  |  Priority(1): No {No=15, Yes=9}
|  |  |  Priority(2): Yes {No=0, Yes=8}
|  |  TownSize_Residence(Town)
|  |  |  Priority(1)
|  |  |  |  HEEQDegree(AHR): Yes {No=1, Yes=3}
|  |  |  |  HEEQDegree(FHR): No {No=4, Yes=0}
|  |  |  Priority(2): Yes {No=0, Yes=4}
|  AppliNumber(3): Yes {No=8, Yes=11}
|  AppliNumber(4): Yes {No=1, Yes=7}
|  AppliNumber(5): Yes {No=1, Yes=2}
Status_OneWeek(EnrollmentRequested): Yes {No=16, Yes=365}
Status_OneWeek(Later): No {No=140, Yes=27}
Status_OneWeek(OfferRejected): No {No=24, Yes=0}
Status_OneWeek(Received): No {No=184, Yes=3}
Status_OneWeek(Valid): No {No=5, Yes=0}
```

Figure 29. Decision tree Model 5.

If an individual applied for two case university study programs and an admission offer has been granted one week before the application deadline, the enrollment is likely if:

(3) TownSize_Residence(BigTown) OR *TownSize_Residence(City)*

(4) TownSize_Residence(Community) AND *Priority(2)*

(5) TownSize_Residence(Town) AND *Priority(1)* AND *HEEQGrade(AHR)*

(6) TownSize_Residence(SmallTown) AND *Priority(2)*

Thus, the model finds that applicants who have submitted applications to several degree programs at the case university will probably enroll if they have received an admission offer at least one week prior to the application deadline. These findings support the assumption made in the descriptive analysis of the data in Section 5.1.1 that the university is locally popular. The tree rules also suggest that the size of the town or city of residence at the time of application affects the likelihood of enrollment. It appears that applicants who live in a location with between 50,000 and 500,000 inhabitants are likely to enroll, especially if they have applied to at least two case university degree programs or have a HEEQ grade of 2. If they reside in a smaller town or community, the likelihood of enrollment seems high if they receive an admission offer for the program they applied for with the second priority. In addition, the tree suggests that applicants with a HEEQ grade of 3 are likely to enroll if they have already been admitted seven weeks before the application deadline. Nevertheless, the tree leaf concerned is only represented by a very limited number of data records, making a binding statement impossible.

In order to improve the generality of the model, the *minimum leaf size* of Model 5 has been further increased, resulting in lower *accuracy* and thus a higher *error rate* in the model validation as well as in the model testing shown in Table 31. The recognition rate in the model testing for the target class *Enrollment(Yes)* dropped only slightly to 66.0%, but the *precision* of the prediction decreased significantly to 49.7%.

Table 31. Performance results of Model 5a.

Model Name	Performance measures of	Accuracy	Error rate	Recall		Precision	
				True(No)	True(Yes)	Pred(No)	Pred(Yes)
5a	Model validation on *DataSetDS2018*	69.82%	30.18%	80.10%	59.55%	66.44%	74.95%
	Model testing on *UnseenTestSet*	71.35%	28.65%	73.45%	66.06%	84.52%	49.66%

In addition, the rules in the tree model shown in Figure 30 are based only on the *Status_ThreeWeeks* attribute. They show that applicants who already requested enrollment three weeks before the application deadline, as well as applicants already admitted at that time, will probably enroll. It should also be noted that applicants who have submitted a valid application three weeks before the application deadline are also likely to enroll. Nevertheless, the distribution of the records in the tree leaf

Status_ThreeWeeks(Admission): Yes {No=40, Yes=89}
Status_ThreeWeeks(EnrollmentRequested): Yes {No=6, Yes=207}
Status_ThreeWeeks(Later): No {No=23, Yes=8}
Status_ThreeWeeks(NotApplied): No {No=214, Yes=122}
Status_ThreeWeeks(OfferRejected): No {No=22, Yes=0}
Status_ThreeWeeks(Received): No {No=225, Yes=100}
Status_ThreeWeeks(Valid): Yes {No=88, Yes=92}

Figure 30. Decision tree Model 5a.

suggests that this rule needs further investigation. Thus, it seems that even though some of the rules in Model 5 are only represented by a relatively small number of examples, they not only increase the model's detail but also its predictive ability.

The *default accuracy* of the unseen test dataset is 71.6%. Thus, if all of the cases in the dataset were classified as belonging to the overrepresented class *Enrollment(No)*, 71.6% of the examples, which are the 275 cases in the *UnseenTestSet* that do not enroll at the beginning of the semester, would be classified correctly. In particular, Model 5 exceeds this default in the model testing with an *accuracy* of 84.9% and can therefore assist the university administration in the enrollment process, especially to estimate the exact number of new students.

5.1.4 Binominal Logistic Regression models

Another suitable approach to forecast the applicant enrollment is binominal logistic regression analysis, which creates models that predict the likelihood of an event occurring – *Enrollment(Yes)* – or an event not occurring – *Enrollment(No)*. The starting point of a logistic regression analysis is usually the assumption of a model based on the logical understanding of the problem. As described in Section 5.1.2, the attributes that are logically assumed to affect the target variable were defined in collaboration with the case university's enrollment department and included in Scenario B. The attributes contained in Scenario C are determined by calculation of a logistic regression model in the pre-processing step of the *forward selection*. Accordingly, the attributes that enhance the performance of the logistic regression model identified in this step are then included in the generation of the actual model.

The resulting overview of the best-possible models created for each of the three scenarios and datasets is shown in Table 32. According to the *F-score*, the Model 18 is the best performing model that can correctly classify 86.6% of the cases in the *DataSetUS2018*. Of the models generated with the *DataSetDS2018*, Model 15 has the highest performance validation metrics and can, therefore, classify 81.2% of the cases in the dataset correctly. In comparison, Model 12 has the highest *accuracy*. Nevertheless, the recognition rate of the class *Enrollment(Yes)* is comparatively low at 64.2%, and it is shown that the model can, in particular, predict the non-enrollment of applicants. Overall, Model 18 has the highest performance validation measures with a *recall* for both target classes above 86% and an *accuracy* of 86.6%.

Table 32. Performance results of validating the logistic regression models generated.

Model Name	Dataset	Scenario	Accuracy	Recall		Precision		Error rate	F-score
				True (No)	True (Yes)	Pred (No)	Pred (Yes)		
10	*TrainingSet 2018*	A	86.72%	95.96%	63.43%	86.88%	86.15%	13.28%	73.01%
11		B	87.69%	97.24%	63.59%	87.08%	90.14%	12.31%	74.49%
12		C	88.15%	97.63%	64.24%	87.32%	91.47%	11.85%	75.41%
13	*DataSetDS 2018*	A	78.15%	81.07%	75.24%	76.61%	79.90%	21.85%	77.45%
14		B	79.36%	83.33%	75.40%	77.21%	81.90%	20.64%	78.48%
15		C	81.15%	91.91%	70.39%	75.63%	89.69%	18.85%	78.82%
16	*DataSetUS 2018*	A	82.52%	82.55%	82.49%	82.50%	82.54%	17.48%	82.52%
17		B	82.33%	83.19%	81.46%	81.78%	82.90%	17.67%	82.18%
18		C	86.59%	86.47%	86.72%	86.69%	86.50%	13.41%	86.60%

In order to assess which of the highlighted models perform best on the data from the field, they were tested with the *UnseenTestSet*. The results of these performance tests are shown in Table 33. According to the *accuracy*, Model 12 still outperforms the other two models as nearly 90% of all cases in the dataset were correctly predicted. However, only 66.1% of the students enrolling at the beginning of the semester were identified, while 99.3% of those who did not enroll were identified. Model 15 is able to correctly classify as many as 86.2% of the data records in the *UnseenTestSet* and 69.7% of the applicants who enroll at the beginning of the semester were recognized by the model. This percentage is even higher for Model 18, where up to 78.9% of the cases that belong to the target class *Enrollment(Yes)* have been correctly classified. However, the *class precision* for the target class *Enrollment(Yes)* of 50.9% is significantly lower than that of Model 15 which is 79.2%. The one of Model 12 is with 97.3% even higher. Consequently, all three models will be further investigated.

Table 33. Performance results of testing Model 12, Model 15 and Model 18 with the *UnseenTestSet*.

Model 12 | accuracy = 89.84%	**True(No)**	**True(Yes)**	**Class precision**
Pred(No)	273	37	88.06%
Pred(Yes)	2	72	97.30%
Class Recall	99.27%	66.06%	
Model 15 | accuracy = 86.20%	**True(No)**	**True(Yes)**	**Class precision**
Pred(No)	255	33	88.54%
Pred(Yes)	20	79	79.17%
Class Recall	92.73%	69.72%	
Model 18 | accuracy = 72.40%	**True(No)**	**True(Yes)**	**Class precision**
Pred(No)	192	23	89.30%
Pred(Yes)	83	86	50.89%
Class Recall	69.82%	78.90%	

The generated logistic regression models specify a regression coefficient that represents the change in the logarithm of the *odds* of the outcome for an increase of one unit in the predictor attribute (Backhaus et al. 2011: 265-266). Consequently, the coefficient only indicates clearly whether the change of one of the attributes has a positive or negative influence on the probability of an example belonging to the positive target class *Enrollment(Yes)*. Furthermore, the *p-values* are

given, which are the result of the statistical test of the *0-hypothesis* and indicate whether an attribute has a statistically significant influence on the target variable. Normally, attributes with a *p-value* of $p < 0.05$ are considered statistically significant, which means that the probability that the connection between the attribute and the target variable being random is less than 5% (Field 2013: 71). Accordingly, a high *p-value* indicates that the influence of the target variable is random with a high probability. In the following, the attributes with a $p < 0.10$ are also considered statistically significant since the probability that the connection between the attribute and the target variable being random is less than 10%.

To generate the logistic regression models, the categorical attributes were transformed using *dummy coding*, which is executed in RapidMiner directly by the logistic regression operator. Consequently, the influence of each attribute class is estimated separately. Table 34 presents Model 12 and indicates that the characteristics *Status_OneWeek(EnrollmentRequested)*, *Priority(2)*, *Priority(3)*, and *Distance_HEEQDistrict(Radius100)* have a positive influence on the probability of an applicant enrolling. It can, therefore, be assumed that persons applying for more than one program and that are admitted to their second or third choice of a degree program at the case university are likely to enroll as well as applicants who have acquired their HEEQ in the region of the case university. In addition, it appears that the HEEQ from a distance of more than 500km increases the likelihood of enrollment. This would not have been assumed from the descriptive analysis of the data as at the beginning of the semester, only 3 applicants of the 38 that applied from more than 500km away had enrolled. Accordingly, the result is based on a limited number of observations and should be investigated further as soon as more data records are available. The characteristics *Status_OneWeek(Received)*, *Status_OneWeek(Later)*, *Distance_PlaceofBirth(Over500)*, and *Distance_PlaceofBirth(Radius500)* have a negative effect on the probability that an applicant will enroll, indicating that applicants whose roots are in different areas of Germany than the case university are likely not to enroll.

Table 34. Logistic regression Model 12.

Attribute	Regression coefficient	p-value
Status_OneWeek(EnrollmentRequested)	3.740	0.000
Status_OneWeek(Received)	-4.120	0.000
Status_OneWeek(Later)	-1.326	0.000
Priority(2)	2.466	0.000
Priority(3)	3.386	0.003
Distance_PlaceofBirth(Over500)	-2.414	0.068
Distance_PlaceofBirth(Radius500)	-1.367	0.096
Distance_HEEQDistrict(Radius100)	1.867	0.099
Distance_HEEQDistrict(Over500)	3.162	0.102
Distance_HEEQDistrict(Radius200)	1.742	0.130
HEEQDegree(fgHR)	-1.394	0.141
…		

Table 35 presents Model 15, which shares the same characteristics as Model 12 that increase the likelihood that an applicant will enroll. In addition, other characteristics are identified that reduce the likelihood of enrollment. Thus, the model shows that if an applicant either lives in a town with 5,000 to 50,000 inhabitants or in a large city with more than 500,000 inhabitants at the time of application, the likelihood of enrollment decreases.

Table 35. Logistic regression Model 15.

Attribute	Regression coefficient	p-value
Status_OneWeek(EnrollmentRequested)	3.327	0.000
Status_OneWeek(Received)	-4.565	0.000
Status_OneWeek(Later)	-1.445	0.000
Priority(2)	2.964	0.000
TownSize_Residence(Town)	-0.768	0.011
TownSize_Residence(SmallTown)	-0.636	0.015
Priority(3)	2.843	0.020
TownSize_Residence(BigCity)	-0.957	0.047
TownSize_Residence(Community)	-0.373	0.182
TownSize_Residence(Abroad)	-2.051	0.317
TownSize_Residence(BigTown)	-0.227	0.480
Status_OneWeek(OfferRejected)	-12.575	0.808
Status_OneWeek(Valid)	-11.793	0.922
Priority(6)	12.612	0.963
Priority(5)	9.286	0.973
Priority(4)	9.023	0.973

Table 36 presents Model 18, which has the highest *recall* for the target class *Enrollment(Yes)* in the model testing, and therefore, can correctly identify most cases in the *UnseenTestSet* that belong to that class. The model suggests that in addition to the above attributes, the *Status_OneWeek(Admission), Status_OneWeek(Later)*, and *Status_ThreeWeeks(EnrollmentRequested)* have a positive effect on the likelihood of the applicant enrolling at the beginning of the semester. Accordingly, applicants who have been admitted one week before the application deadline will probably enroll as well as those who requested the enrollment three weeks before the application deadline.

Table 36. Logistic regression Model 18.

Attribute	Regression coefficient	p-value
Status_OneWeek(EnrollmentRequested)	9.051	0.000
Status_OneWeek(Admission)	5.202	0.000
Status_OneWeek(Later)	3.527	0.000
Priority(2)	2.731	0.000
Status_ThreeWeeks(Later)	-2.692	0.000
Status_ThreeWeeks(Admission)	-1.043	0.000
HEEQ_Country(Foreign)	-2.557	0.001
Priority(3)	3.057	0.038
Status_ThreeWeeks(EnrollmentRequested)	1.087	0.057
TownSize_Residence(BigCity)	-2.121	0.065
Status_ThreeWeeks(Valid)	0.363	0.189

The positive impact of the *Status_OneWeek(Later)* on the likelihood of an applicant enrolling is in contrast to the statements in the above models, which indicate that this status has a negative impact. This difference can be traced back to the nature of the *DataSetUS2018*, which consist of many synthetic data records that could cause this contradiction. The descriptive analysis of the *Status* attribute in Table 13 shows that 421 individuals with a valid application have the *Status_OneWeek(Later)*. Of these, 37 were enrolled at the beginning of the semester, which corresponds to an enrollment rate of only 8.8%. Accordingly, this attribute is considered to have a negative rather than a positive effect on enrollment.

Which of the models should be preferred by the university administration depends on the decisions that are to be supported. If the target is to have the highest *accuracy* in predicting the target class *Enrollment(Yes)*, then Model 12 should be preferred because the *precision* for the target class is 97.3% in the model testing. However, 33.9% of the applicants enrolling at the beginning of the semester are not identified and would, therefore, be lost if only those students are admitted whose enrollment is considered likely by the model. If it is more urgent for decision-makers to identify as many applicants as possible that actually enroll at the beginning of the semester, Model 18 should be preferred as nearly 80% of the applicants who enroll are identified correctly. If the goal of the prediction is to estimate the actual number of applicants enrolling, the use of Model 15 is suggested. According to the model testing results shown in Table 33, the *precision* and the *recall* for the target class *Enrollment(Yes)* only differ by 10%, and therefore, the number of misclassifications and the number of unidentified applicants who enroll are relatively balanced.

5.1.5 Artificial Neural Networks models

In the third round of analysis, ANNs were used to forecast the enrollment of applicants. Before the ANN models were generated, the dataset had to be transformed according to the process described in Section 3.4.4. Most attributes in the given dataset are categorical and are therefore transformed with *dummy coding*. Especially the variables *HEEQDistrict* with 242 categories, *BirthCountry* with 68 categories, and *Residence* with 893 categories increase the size of the datasets immensely. As a result, the transformation of the available attributes results in a training dataset with 1416 attributes. This huge number of attributes has a negative impact on the model calculation time. A first experiment showed that the model training with the *TrainingSet2018* already took 1 hour and 45 minutes for only 10 training cycles and 1 hidden layer. This small number of training cycles, where in each one the goal is to minimize the *SSE* for the model by identifying the optimal weights for each neuron connection, results in an undertrained model with limited predictive power.

This issue has been resolved by applying feature selection to the newly generated dataset. The *Chi-squared*-based feature selection used is described in detail in Section 3.2.3. Accordingly, only those attributes of the 1416 are included in the analysis that have a proven link to the target variable

Enrollment. In addition, all models are generated based on one hidden layer[77] and 500 training cycles. The first ANN model was generated with *TrainingSet2018* and those attributes whose *weight ≥ 0.1* in the direction of the target variable. A total of 6 attributes were included in the model building – *Status_OneWeek(EnrollmentRequested)*, *Status_OneWeek(Received)*, *Status_OneWeek(Later)*, *Status_ThreeWeeks(EnrollmentRequested)*, *Status_FiveWeeks(EnrollmentRequested)*, and *Status_SevenWeeks(Valid)*. The performance measures of this Model 19 are shown in Table 37. They show that the model can correctly identify 97.8% of the cases that belong to the class *Enrollment(No)* but only 59.1% of the cases that belong to the target class *Enrollment(Yes)*.

The second ANN model evaluated in Table 37 contains the attributes in the analysis whose *weight ≥ 0.01* toward the target variable. Accordingly, 43 attributes were included. Model 20 has a slightly lower *accuracy* than Model 19, but the *recall* for the target class *Enrollment(Yes)* at 66.9% is higher and therefore this model seems to be able to classify more applicants who enroll at the beginning of the semester correctly.

Table 37. Performance results of validating the ANN models Model 19 and Model 20.

Model Name	Dataset	Inclusion criteria	Number of included attrib.	Calculation time	Accuracy	Recall		Precision	
						True (No)	True (Yes)	Pred (No)	Pred (Yes)
19	*TrainingSet 2018*	Weight ≥ 0.1	6	< 1 min	86.77%	97.75%	59.06%	85.76%	91.25%
20		Weight ≥ 0.01	43	6:52 min	85.16%	92.37%	66.99%	87.59%	77.67%

In addition, the undersampled dataset and the oversampled dataset were used to generate models that predict both target classes well. Again, each dataset has been transformed to fit the requirements of the ANN algorithm, and then the *Chi-squared* based feature selection was applied. Again, two scenarios were analyzed with both balanced datasets. First, all the attributes that have a *weight ≥ 0.1* with respect to the target variable are included in the model creation, and secondly, all attributes with an importance *weight ≥ 0.01* are included.

The performance results of the four ANN models, shown in Table 38, display that the inclusion of additional attributes and data records in the model calculation increases the model performance but also adversely affects the model calculation time. These calculation times can be reduced by using the *split-validation* operator for model validation, which uses a fixed number of records as training and a fixed number of records for validation purposes. In the study presented, it was decided to continue using *cross-validation* to have more training data records available, which generally results in models with better performance. As shown in Table 38, Model 24 has the highest *accuracy* and can correctly classify 84.8% of all records in the training datasets. In this model, the

[77] Larose et al. (2015: 342) specifies that the usage of one hidden layer is sufficient for most predictive problems.

recall and the *precision* are high for both target classes, and the model appears to classify 85.1% of the cases that belong to the target class *Enrollment(Yes)* correctly. To evaluate the real-life performance of the models, however, model testing is required.

Table 38. Performance results of validating ANN models Model 21, Model 22, Model 23, and Model 24.

Model Name	Dataset	Inclusion criteria	Number of included attrib.	Calculation time	Accuracy	Recall		Precision	
						True (No)	True (Yes)	Pred (No)	Pred (Yes)
21	*DataSetDS 2018*	Weight ≥ 0.1	11	< 1 min	77.26%	76.38%	78.16%	77.76%	76.79%
22		Weight ≥ 0.01	55	3.25 min	77.75%	78.32%	77.18%	77.44%	78.07%
23	*DataSetUS 2018*	Weight ≥ 0.1	11	< 1 min	79.80%	71.58%	88.01%	85.65%	75.59%
24		Weight ≥ 0.01	77	16.13 min	84.77%	84.41%	85.12%	85.01%	84.52%

The performance results of testing the models on the *UnseenTestSet* are shown in Table 39. These suggest that Model 24 performs best in forecasting the target class *Enrollment(Yes)*. Of the 109 cases in the test set that enroll, 72.5% were identified, which is 69.3% of all the cases predicted by the model as belonging to the class *Enrollment(Yes)*. Accordingly, the number of misclassifications, which are 35 cases, is close to the number of unidentified cases belonging to the target class *Enrollment(Yes)*.

Table 39. Performance results of testing the ANN models with the *UnseenTestSet*.

Model 19 | accuracy = 89.06%	**True(No)**	**True(Yes)**	**Class precision**
Pred(No)	273	40	87.22%
Pred(Yes)	2	69	97.18%
Class Recall	99.27%	63.30%	
Model 20 | accuracy = 84.11%	**True(No)**	**True(Yes)**	**Class precision**
Pred(No)	253	39	86.64%
Pred(Yes)	22	70	76.09%
Class Recall	92.00%	64.22%	
Model 21 | accuracy = 72.92%	**True(No)**	**True(Yes)**	**Class precision**
Pred(No)	188	17	91.71%
Pred(Yes)	87	92	51.40%
Class Recall	68.36%	84.40%	
Model 22 | accuracy = 78.39%	**True(No)**	**True(Yes)**	**Class precision**
Pred(No)	218	26	89.34%
Pred(Yes)	57	83	59.29%
Class Recall	79.27%	76.15%	
Model 23 | accuracy = 75.00%	**True(No)**	**True(Yes)**	**Class precision**
Pred(No)	179	18	91.63%
Pred(Yes)	78	91	53.85%
Class Recall	71.64%	83.49%	
Model 24 | accuracy = 83.07%	**True(No)**	**True(Yes)**	**Class precision**
Pred(No)	240	30	88.89%
Pred(Yes)	35	79	69.30%
Class Recall	87.27%	72.48%	

Therefore, Model 24 allows the estimation of the number of applicants that enroll at the beginning of the semester, as it is anticipated that 114 applicants from the *UnseenTestSet* will enroll, and 109 actually enroll. If the goal of the prediction was to identify applicants who actually enroll, regardless of the wrong classifications, Model 21 should be brought forward, as the model correctly identifies 84.4% of all applicants who enroll.

5.1.6 Discussion of the analysis results

The overarching objective of analyzing the existing applicant data was to demonstrate that they can assist universities in the relevant decision-making process and to help optimize the enrollment process. All of the models highlighted in the presented case study can predict the enrollment of applicants in the model testing with greater *accuracy* than the *default accuracy* of the presented dataset, which is 71.6%. In addition, the interpretable decision tree and logistic regression models indicate features that increase the likelihood of an applicant enrolling. In particular, the attribute *Status_OneWeek* and its categories are an important predictor for the enrollment. If an applicant has received an admission offer or has actively requested enrollment at least one week prior to the application deadline, he or she will most likely become a new student at the beginning of the semester. Furthermore, the models suggest that applicants with a HEEQ grade from the area of the university site are also likely to enroll, indicating a sense of regional affiliation of the new students. This assumption is further supported by the positive impact the submission of multiple applications has for the enrollment. It seems that individuals applying for multiple study programs have a particular interest in studying at the case university and are less focused on a particular subject. In addition to the *Status_OneWeek* attribute, Model 5a suggests that the *Status_ThreeWeeks* attribute can also be helpful in predicting an applicant's enrollment, but only with an *accuracy* of 71.4% in the model test. If the generated model would have a higher accuracy, it could allow the university management to predict the enrollment of applicants already three weeks before the application deadline.

Nevertheless, in view of 4 out of the 6 programs examined being admission-free, the university may only use the knowledge generated to plan and adapt its capacities and resources according to the predicted requirements. This is due to the fact that anyone who applies for an admission-free program when it is due and, if applicable, meets the minimum admission requirements, is entitled to a place in the program. The study and examination office, therefore, have no opportunity to intervene if overbooking is predicted for an admission-free program, as they cannot stop the admission process prematurely or only admit a certain number of applicants.

Nevertheless, the presented models generate decision support and enable universities to prepare their resources for the beginning of the semester. This is possible because the models permit for

the forecasting of the actual enrollment numbers at the beginning of the semester, allowing demand-oriented capacity and resource planning. When the overbooking of a study program is predicted, the decision-makers have the possibility to plan ahead. For example, they can prepare the division of a course into groups, which may be necessary to adapt to the spatial conditions of the university. Furthermore, additional lecturers could be recruited if part of the funds are used according to the predicted requirements. As a result, the academic and administrative members of the university can prepare themselves, which can reduce dissatisfaction that can occur with overcrowded programs and too many students. If one week before the application deadline, it is predicted that the available study places are not fully utilized, applicants with applications having the status *Later* can be encouraged to submit the missing documents. In addition, to promote more applications, it is conceivable that marketing for a study program can be intensified or the application deadline extended.

Based on the current structure of the enrollment process, this study suggests that Model 24 could be most useful for demand-based capacity and resource planning. As highlighted in Section 5.1.5, Model 24 can correctly predict 72.5% of the applicants that enroll at the beginning of the semester in the model test and 69.3% of the applicants who are considered new students are actually new students. Accordingly, the model wrongly classifies 30.7% of the cases in the testing dataset as new students, while 27.5% of the actual new students are not identified by the model. The number of misclassifications and the number of not identified new students only differs minimally, so the model is able to predict nearly the right number of applicants that enroll at the beginning of the semester. Since the current structure of the admission procedure does not allow admissions to be granted only to those applicants that are forecasted to be likely to enroll, it is not considered dramatic that 27.5% of the applicants most likely to enroll are not identified by Model 24 because every student who submits a valid application must receive an admission offer.

If the goal of the study and examination office is to correctly identify as many applicants as possible that enroll, Model 9 should be preferred, which is able to correctly identify 89.0% of the cases that belong to the target class *Enrollment(Yes)* in the *UnseenTestSet*. If the decision-makers are looking for an interpretable model and correctly predicting the number of enrollments, it is suggested to apply Model 5 or Model 15. Both models were generated with the *DataSetDS2018*, i.e. no synthetic data records are included in the modeling. In addition, features are identified that affect the likelihood of enrollment, such as *Status_OneWeek(EnrollmentRequested)* and *Status_OneWeek(Admission)* in combination with *AppliNumber(2)*, *TownSize_Residence(BigTown)*, *TownSize_Residence(City)*, or *Priority ≥ 2*. In addition, both models can correctly identify 69.7% of the applicants who enroll at the beginning of the semester and at 84.9% and 86.2% have significantly higher *accuracy* than the *default accuracy* in the *UnseenTestSet*, which is 71.6%.

For admissions-restricted programs, predictive models present additional options for optimizing the enrollment process. The study places for admission-restricted programs are usually awarded in several rounds until all available study places are filled. With the help of predictive models, the number of rounds required to fill the available study places can be minimized by predicting how many applicants eligible for admission will accept an admission offer. If this number is below the available minimum places, the invitation for admission could be extended to lower-ranking places until a satisfactory number of enrollment is predicted. This procedure could help the university to award exactly as many admissions as needed in the first round of admissions to fill all available study places. If this procedure is successfully established, the enrollment process becomes more time-saving. Furthermore, applicants that would have been granted admission in the second or third admission round will receive an immediate offer to enroll in the program, which may secure the university students who would otherwise have migrated to a competing university.

5.1.7 Proposal for optimization

In order for German universities to face the challenge of overcrowded study programs, the current structure of the admission process must change. One discussed solution at the case university is the subdivision of the application process for admissions-free programs into several 4-week slots. In each slot, applications will be received by a pre-defined deadline. Subsequently, the applicants who qualify for admission will be invited to a program. Until a certain period, they must respond to the admission offers, otherwise, their claim to a study space expires. This procedure will continue until all available study places have been filled. Nevertheless, overbooking would still be necessary as candidates who have already accepted an admission offer may still apply to more universities and opt for another program, even though they have already enrolled at the case university.

Therefore, a further modification of the admission process is proposed, based on the conversion of admission-free to admission-restricted programs, which allows intervention by the university management. This proposal does not mean that only individuals with a very good HEEQ should be given the opportunity to study. Rather, it is suggested that the universities have the right to not have to admit every candidate to their programs, even if they do not want to restrict their admissions to ranking places that are based on grades or wait semesters. Consequently, a new admissions procedure is proposed, based on the proven possibilities that models predicting the enrollment numbers offer university decision-makers. This suggestion of an optimized enrollment process is shown in Figure 31.

As shown, it is proposed that the university collects applications until a fixed application deadline. After the end of the deadline, all persons who have submitted a valid application, which might include the fulfillment of some admission requirements, will be considered for further procedure. Therefore, no admissions will be granted before the end of the application deadline. Then, from all

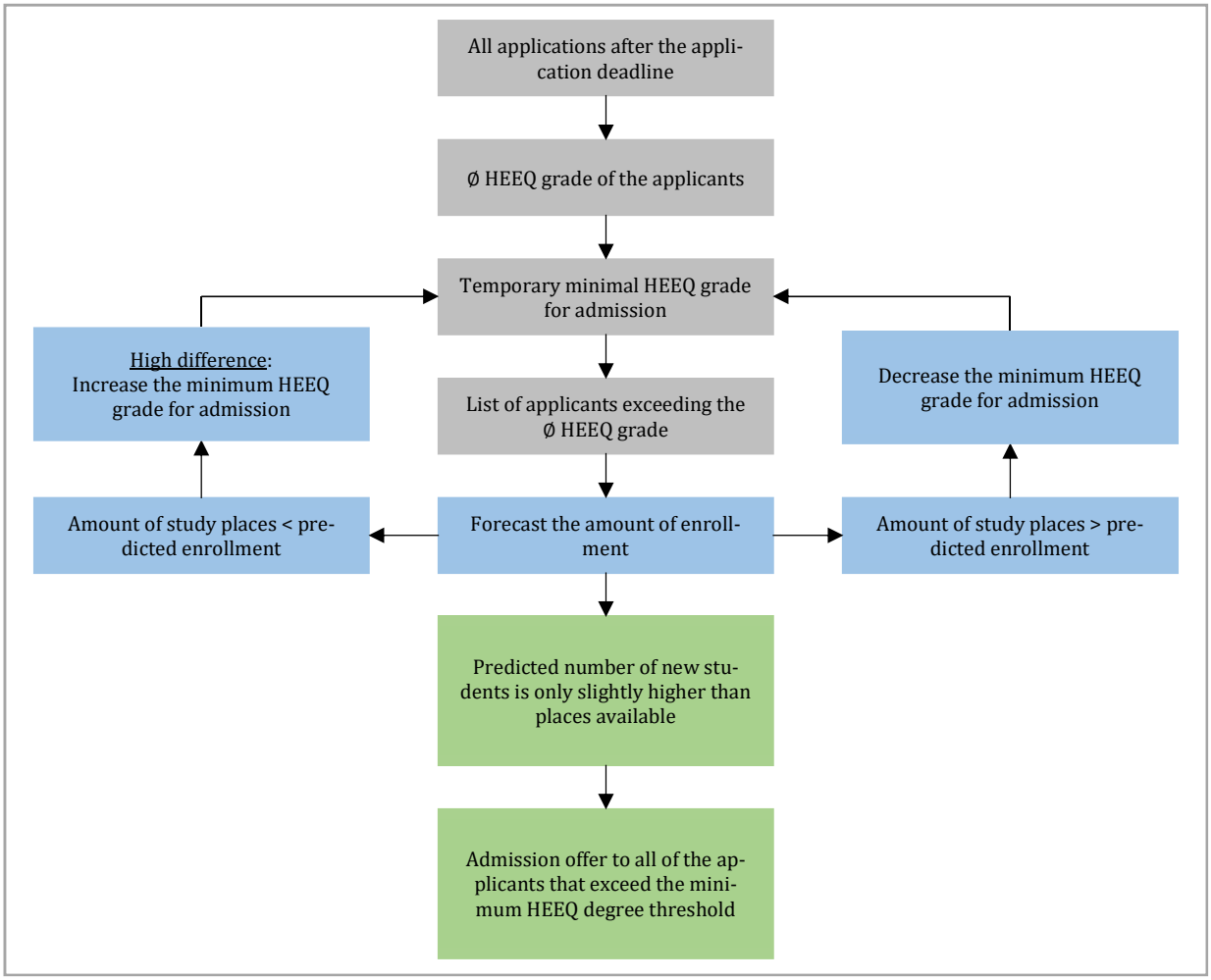


Figure 31. Proposal of an optimized enrollment process based on predictive models for the allocation of 50 - 80% of the available study places.

valid applications to a program, the average HEEQ grade is calculated, which is used as a threshold for admissions. For all the applicants that exceed the calculated minimum threshold, enrollment is predicted. The forecast enrollment numbers are then compared with the number of available study places. If the predicted enrollment numbers far exceed the available study places, then the minimum threshold of the HEEQ grade could be increased.[78] Accordingly, fewer applicants will be considered for enrollment. For these, an enrollment prediction is performed again, and the new predicted enrollment numbers are compared with the actual study places available. This procedure is repeated until a satisfactory number of new students is predicted.

If, after the first round of predictions, the candidate numbers are lower than the available study places, the threshold for the HEEQ grade could be lowered.[79] Accordingly, more applicants will be considered for enrollment. Also for these, the enrollment number is predicted and compared with the number of available places. This is repeated until the predicted result is satisfactory. Once a satisfactory number of predicted enrollments corresponds to the number of places available, all the

[78] For example: If the original threshold is a HEEQ grade of 2.7 than this could be increased to the HEEQ grade of 2.5.
[79] For example: If the original threshold is a HEEQ grade of 2.7, it could be decreased to a HEEQ grade of 3.0.

applicants who exceed the specified minimum HEEQ grade threshold are invited to the program. If the forecast has been made on a well-performing prediction model, the study places should be fully occupied and the overbooking minimized.

In order not to exclude from the program those who do not have a high HEEQ grade, it is proposed that not all the available study places be awarded by the procedure described above. It is suggested that a drawing process will award a certain number of available places, which could be between 20% and 50% of the total available places. Accordingly, it is proposed that a certain number of candidates be chosen randomly from any remaining candidates who have not yet received a study place offer and therefore do not reach the minimum threshold for the HEEQ degree. For these applicants, the enrollment numbers are predicted. In line with the procedure illustrated in Figure 31, more or fewer applicants will be considered for an admission offer, depending on how well the forecasted number actually fills the available study spaces. Once the number of applicants predicted to enroll at the beginning of the semester is satisfactory, and therefore the available study places are filled, the admission offers are granted to those applicants randomly selected for admission. To ensure that not only overbooking is addressed, but also all the available spaces are occupied, a slight overbooking of the available places is still conceivable.

The predictive models presented in this chapter are mainly based on the *Status* attribute. Therefore, in order for the method described above to work, new predictive models need to be generated because fewer status features are available if no admissions are granted before the application deadline. The presented models show that the demographic data currently available to German universities does not significantly influence the enrollment probability of an applicant. Therefore, new attributes need to be collected in order to generate well-performing predictive models without the *Status* attribute being available. Therefore, it is proposed to gather more information about a student's intention to apply for a program. Therefore, in addition to prioritizing the application, the applicants must provide additional information about their intentions to enroll in a program. These intentions could be queried during the application process by asking standardized questions, for example:

- How high has been your interest in the main topic of the study program before? (very high/high/medium/low/very low)
- Did you apply for other programs that are similar in focus to this study program? (Yes/No)
- Is the location of the university an important factor in your application for the study program? (Yes/No)

The answer to these and other questions, which, however, should not get out of hand to keep the application process clear and efficient, are then implemented in the creation of forecasting models that are believed to help optimize the process.

The above-proposed changes and optimizations of the enrollment process surely cannot be implemented immediately and require several steps and measures ranging from the provision of the required resources to the adaption of admission regulations to the development of well-performing predictive models. Consequently, such changes in Germany are not alone the responsibility of the university itself and depend on the support and cooperation of the federal states. However, without a consistent change in the admission procedure, in particular, the problem of overcrowded admissions-free programs cannot be tackled by German universities.

Although overbooking cannot be prevented all together with the models presented above, they have shown to provide important information that support the university decision-makers, especially in preparing their capacities in the case of overbooked programs.

5.2 Dropout Analysis

Ensuring student success, reducing student dropout rates, and improving the quality of study programs are other important tasks of universities, which are believed to be supported by the analysis of student-related data resources with DM methods. Such analyses are expected to identify the requirements of study programs and the students' challenges. These can then be addressed by providing services that are needed and supporting the students' success. This, in turn, can support the university to create a unique profile that could also increase its attractiveness.

The degree program with the highest number of students and the longest history in the case university is the BA Bachelor, which is also offered by many competitors. As a result, the case university could benefit greatly from analyzing their students data resources, which could help them to optimize their program and to differentiate it from the programs of the competition. Therefore, the students' data are examined for patterns that might be related to student dropout. In addition, predictive models will be generated to enable the university to forecast dropout. With this knowledge, the university administration and management has the opportunity to intervene and to increase the success of its students through individual support.

5.2.1 Introduction of the available student data resources

The dataset available for the model generation consists of 1813 records of students who have been enrolled in the BA Bachelor's program since 2008 and either successfully completed their studies or terminated them before graduation. The dataset has been extracted from the case university's data management system and consist of 245 attributes, of which 221 attributes represent exams with their individual *exam-ID* and *exam grade*. The structure of the BA program is shown in Figure 32. The boxes marked in green represent the required courses, whilst the gray boxes stand for the elective courses. Accordingly, the students attend required courses in the first three semesters,

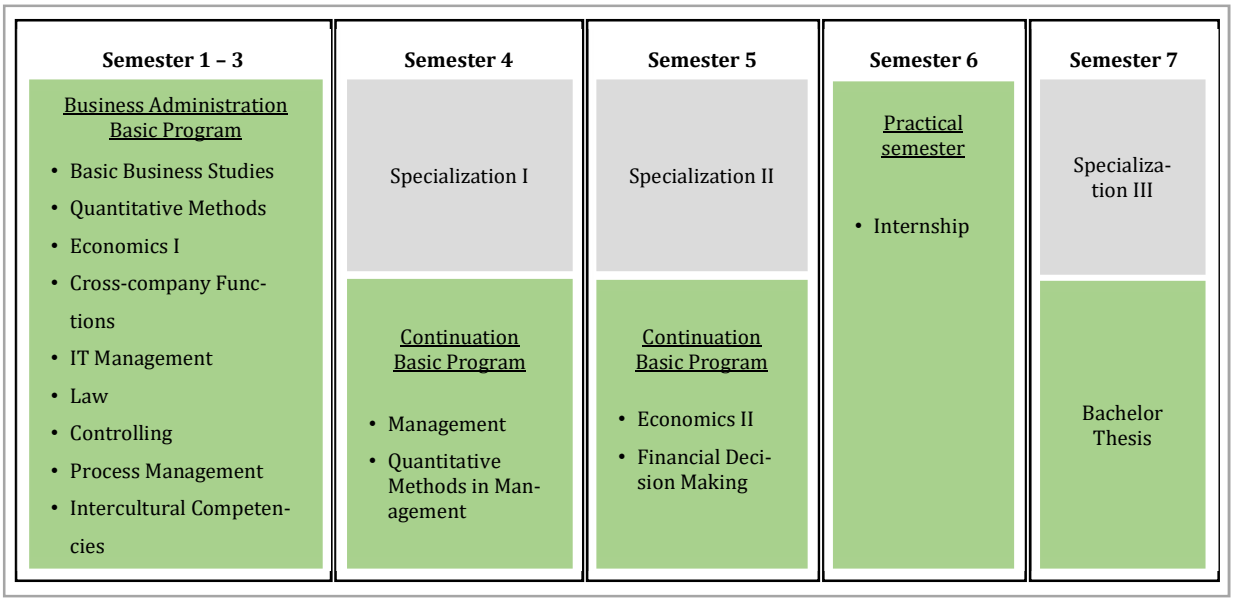


Figure 32. Structure of the BA study program at the case university.

which serve as the basis for the following semesters. After that, they can choose between specializations and different electives to earn the required number of credits to successfully complete their studies. The variety and range of courses offered at the case university are large, and the electives that may be chosen to earn a sufficient number of credit points changes each semester. Hence, many elective courses have a very limited number of participants and are not offered regularly. These courses, which obviously also have only a very limited number of values in the dataset, were excluded from the analysis. This involved a total of 94 courses.

In addition to the examination names, the examination grades, and general study program-relevant data, the dataset contains demographic attributes and information about the previous education of the students. The attributes available for analysis are presented in Table 40.

Table 40. Attributes in the student dataset of the case university.

		Attribute	Description	Values
Demographic attributes	**Personal information**	*Age* (new)	The age of the student at the time of enrollment for the study program.	Age in years
		Gender	The gender of the applicant.	*male, female*
		PlaceofBirth	The name of the student's place of birth.	Name of the Germany city or community, else *Abroad*
		Distance_ PlaceofBirth (new)	The distance of the student's birthplace from the location of the case university.	*Radius100* (≤ 100km), *Radius200* (100km < Radius200 ≤ 200km), *Radius300* (200km < Radius300 ≤ 300km), *Radius400* (300km < Radius400 ≤ 400km), *Radius500* (400km < Radius500 ≤ 500km), *Over500* (> 500km), *Abroad*
		Nationality	The nationality of the student.	*German, Foreign*
		Residence	The name of the place of residence of the student during the studies at the case university.	Name of a German town or city
		Distance_Residence (new)	The distance of the place of residence from the university site.	See attribute *Distance_PlaceofBirth*

		Attribute	Description	Values
		TownSize_Residence (new)	The size of the town the student lives in during the studies at the case university.	*Community* = below 5.000 inhabitants, *SmallTown* = 5.000 - 20.000 inhabitants, *Town* = 20.000 - 50.000 inhabitants, *BigTown* = 50.000 - 100.000 inhabitants, *City* = 100.000 - 500.000 inhabitants, *BigCity* = above 500.000 inhabitants
	Previous education	*HEEQDegree*	The type of HEEQ the student has.	*AHR* = 'traditional' university entrance degree, *FHR* = university of applied sciences entrance degree, *fgHR* = subject-specific entrance degree, *Abroad* = foreign entrance degree
		HEEQDistrict	The district or country where the HEEQ was earned	Name of the city, district or country
		Distance_HEEQDistrict (new)	The distance of the HEEQ district from the location of the case university.	See attribute *Distance_PlaceofBirth*
		HEEQGrade	The grade of the HEEQ degree.	*1.00* (best grade) – *4.0* (least passing grade)
		HEEQGradeComp (new)	The compact version of the attribute *HEEQGrade*.	*1* (1.0 – 1.5) = very good, *2* (1.6 – 2.5) = good, *3* (2.6 – 3.5) = satisfactory, *4* (3.6 – 4.0) = sufficient
		TimeHEEQDegree-Application (new)	The time between the successful completion of the HEEQ and the application at the case university.	*< 6 Months*, *between 6-12 Months*, *between 13-18 Months*, *between 19-24 Months*, *between 25 and 36 Months*, *> 37 Months*
		FirstSemester (new)	This attribute indicates if the student collected first study experiences before starting the investigated study program.	*Yes* = student has not been enrolled in previous study programs, *No* = student has been enrolled previously
		Apprenticeship	This attribute defines whether the student has completed an apprenticeship before starting the investigated study program.	*Yes* = student has a finished apprenticeship, *No* = student does not have a finished apprenticeship
Studies-specific attributes	**General information**	*Course of Study*	The name of the study program in which the student was enrolled.	*Business Administration (BA)*
		Enrollment date	The date of the enrollment to the program.	Date
		Deregistration date	The date of termination from the program.	Date
		DeregReason	The reason for the termination of the studies.	*Dropout after ultimate failed exam*, *Willful dropout by the student*, *No re-registration*, *University change*, *Successful completion*, *Other*
		Completion (new)	This attribute is the target variable and indicates if a student has successfully completed the study program. It was derived from the attribute *DeregReason*.	*Yes* = student successfully completed the study program, *No* = student did not successfully complete the program and dropped out
		SemestersStudied (new)	The number of semesters the student spend studying in the investigated program.	0 to max
	Exam-specific data	*Course/Module/ Exam name*	The individual name of each course or module.	A list of the courses and modules in the BA program can be found in Appendix C.
		Exam grade	The grade for the course or module.	*1.0* until *5.0*, where *1.0* is the highest possible grade, *4.0* is the minimum passing grade and *5.0* is the grade a student gets for a non-passed exam. Further coding for exams without grades: *1001* = passed without a grade; *5005* = failed without a grade

In addition to the attributes directly available from the university's data management system, supplementary attributes have been generated either to develop new information or to compress attributes. The latter has been found to be particularly necessary for attributes with many categories, some of which are represented by only one record, as generalizable rules and predictive models require enough data records for each category. The attributes created for the analysis are marked with the expression *new*.

The target variable of the presented dataset shown in Figure 33 is *Completion*, which describes whether the student has successfully completed the study program – *Completion(Yes)* – or has dropped out – *Completion(No)*. This target variable is almost balanced in the available dataset, with 988 records (54.5%) successfully completing their studies and 825 records (45.5%) dropping out without successfully completing the program.

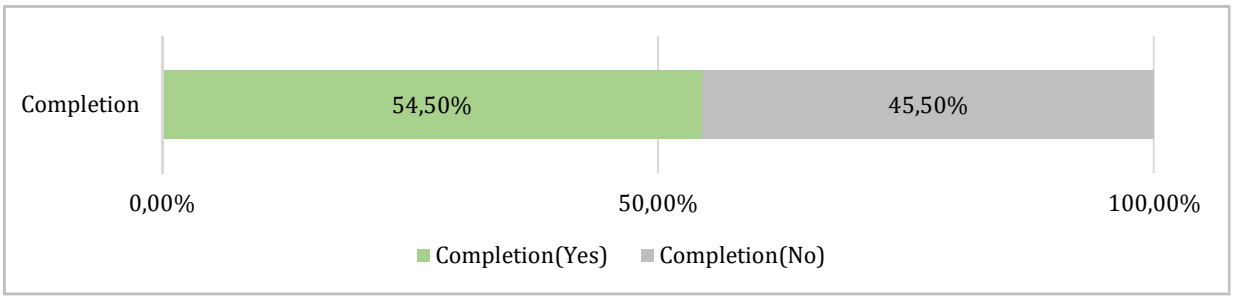


Figure 33. Distribution of the *Completion* target variable.

The descriptive evaluation of the *DeregReason* attribute is shown in Table 41. It specifies the reasons for the students deregistration from the program. Accordingly, the most common reason for students to drop out is the forced dropout after an ultimately failed exam, which is applicable for 419 of the 825 records in the dataset that dropped out, which are 50.8%.

Table 41. Distribution of the *DeregReason* attribute.

DeregReason	Total[80]		Graduates[81]	Dropouts[82]		Dropout rate
	Number	Fraction		Number	Fraction	
Dropout after ultimate failed exam	419	23.1%	0	419	50.8%	100.0%
Willful dropout by the student	218	12.0%	0	218	26.4%	100.0%
No re-registration	104	5.7%	0	104	12.6%	100.0%
University change	70	3.9%	0	70	8.5%	100.0%
Other	14	0.8%	0	14	1.7%	100.0%
Successful completion	988	54.5%	988	0	0.0%	0.0%
Total	**1813**	**100.0%**	**988**	**825**	**100%**	**45.5%**

[80] Total number of students/records in the dataset.

[81] Students in the dataset who successfully completed their studies and therefore belong to the *Completion(Yes)* target class.

[82] Students in the dataset that dropped out before successfully completing the study program and therefore belong to the *Completion(No)* target class.

The second most frequent reason is the willful termination by the student, which applies to 218 individuals and, therefore 26.4% of the dropouts. Other reasons are lack of re-registration, which applies to 104 individuals, or the transfer to another university, which is valid for 70 individuals. For the remaining 14 students that dropped out, the reason for termination is not stated.

The descriptive analysis of the *Age* attribute is shown in Table 42. Accordingly, the average age of the students in the dataset is 21.4 years, with a standard deviation of 2.8 years. In total, 274 students (15.1%) of 1813 are older than 23 years of age, and only two are younger than 18 years. The average age of the graduates is 21 years, with a standard deviation of only 2.4, while the average age of the dropouts is 21.8 years with a standard deviation of 3.1 years. These differences make room for the assumption that younger students are more likely to successfully complete their studies. Nevertheless, 112 graduates are older than 23, which is still a proportion of 11.3%. As a result, it is assumed that the age of the student does not significantly influence the successful completion of the study program.

Table 42. Distribution of the *Age* attribute.

Attribute	Total				Graduates				Dropouts			
	Minimum	Maximum	Average	Deviation	Minimum	Maximum	Average	Deviation	Minimum	Maximum	Average	Deviation
Age	17.0	48.0	21.4	2.8	17.0	48.0	21.0	2.4	18.0	37.0	21.8	3.1

The distribution of the *Gender* attribute is illustrated in Table 43. The evaluation of this attribute shows that the BA study program of the case university has slightly more female than male students. Furthermore, female students seem more likely to successfully complete the study program, with 59.9% of the graduates being female. Yet, it is assumed that gender alone is not a predictor attribute that influences the drop out of a student because less male than female students are enrolled in the study program, which could be the reason for the difference in the distribution.

Table 43. Distribution of the *Gender* attribute.

Gender	Total		Graduates		Dropouts	Dropout rate
	Number	Fraction	Number	Fraction		
Female	969	53.4%	592	59.9%	377	38.9%
Male	844	46.6%	396	40.1%	448	53.1%
Total	**1813**	**100.0%**	**988**	**100%**	**825**	**45.5%**

The distribution of the attribute *Nationality*, shown in Table 44 displays that 61.7% of the students of a foreign nationality dropped out, while only 44.6% of the students of German nationality dropped out. Accordingly, the dropout rate is much higher for students of a foreign background, which might be due to the language barrier as the BA program is taught mainly in German.

Table 44. Distribution of the *Nationality* attribute.

Nationality	Total		Graduates	Dropouts	Dropout rate
	Number	Fraction			
German	1693	93.4%	942	751	44.6%
Foreign	120	6.6%	46	74	61.7%
Total	**1813**	**100%**	**988**	**852**	**45.5%**

The *Distance_PlaceofBirth*, *Distance_Residence*, and *TownSize_Residence* attributes were generated from the demographic *PlaceofBirth* and *Residence* attributes to compress the detailed information in these attributes. For example, the *Residence* attribute has 504 distinctions, and 296 of these places are only represented by one individual in the dataset. The separate analysis of these distinctions will most probably not generate universally applicable rules that can be used to predict the student's dropout. Therefore, aggregation is a necessary step. The distribution of the data records in the *Distance_PlaceofBirth* and *Distance_Residence* attributes is shown in Table 45.

Table 45. Distribution of the *Distance_PlaceofBirth* and the *Distance_Residence* attributes.

Distance_PlaceofBirth	Total		Graduates	Dropouts	Dropout rate
	Number	Fraction			
Radius100	1372	75.7%	774	598	43.6%
Radius200	163	9.0%	82	81	49.7%
Radius300	37	2.0%	18	19	51.4%
Radius400	24	1.3%	14	10	41.7%
Over500	29	1.6%	13	16	55.2%
Abroad	188	10.4%	87	101	53.7%
Total	**1813**	**100.0%**	**988**	**825**	**45.5%**
Distance_Residence	**Total**		**Graduates**	**Dropouts**	**Dropout rate**
	Number	Fraction			
Radius100	1569	86.5%	849	720	45.9%
Radius200	194	10.7%	108	86	44.3%
Radius300	21	1.2%	13	8	38.1%
Radius400	12	0.7%	6	6	50.0%
Radius500	9	0.5%	7	2	22.2%
Over500	4	0.2%	2	2	50.0%
Abroad	4	0.2%	3	1	25.0%
Total	**1813**	**100.0%**	**988**	**825**	**45.5%**

This distribution indicates that students born in the region of the university are more likely to complete their studies than students with a foreign birthplace or a birthplace more than 500km away from the case university. The attribute *Distance_Residence* is not that conclusive, possibly because many students have moved their place of residence to nearby the case university. This cannot be reconstructed because the dataset does not include any information about when the place of residence was last changed. Furthermore, it is not apparent whether the attribute describes the primary or secondary place of residence or at what point in the student lifecycle it was collected or

updated. Therefore, the attribute is not included in the following DM analyses. The *TownSize_Residence* attribute is also based on the *Residence* attribute. For the same reasons as mentioned above, this attribute is not further investigated.

The distribution of the *HEEQDegree* attribute is illustrated in Figure 34. Of the total cases in the dataset, 929 individuals, or 51.2%, have a HEEQ for the 'traditional' university (*AHR*), and 46.4% have a university of applied sciences entrance degree (*FHR*). The proportion of graduates with an *AHR* HEEQ degree is even higher at 58.5%. It seems that an *AHR* HEEQ degree increases the likelihood of the student graduating.

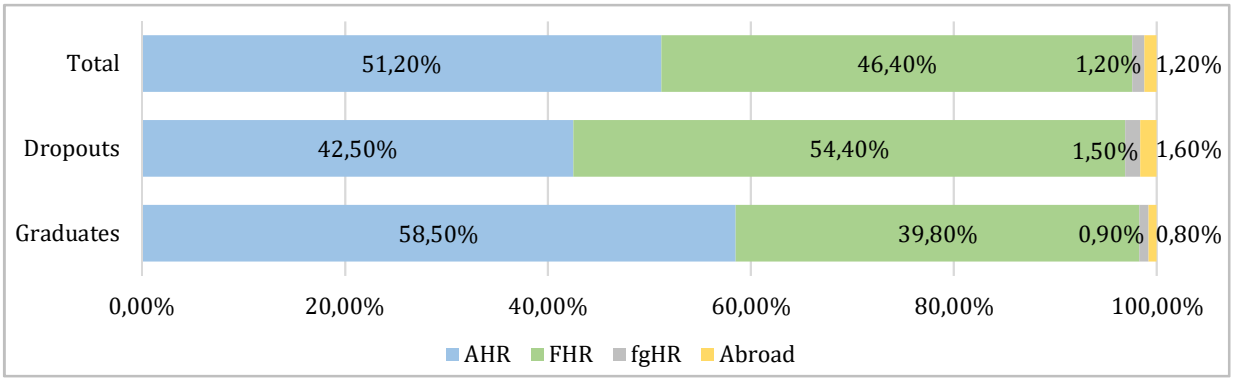


Figure 34. Distribution of the *HEEQDegree* attribute.

The *HEEQDistrict* attribute specifies the area in which the HEEQ degree has been obtained. Because the attribute has 139 characteristics, it has been compressed into the *Distance_HEEQDistrict* attribute to define whether the HEEQ degree was obtained in the region and, therefore, determine the reach of the university. The distribution of the attribute is shown in Table 46. The majority of the students received their HEEQ degrees from the region of the case university, 150 received it within a radius of 200km, and only 94 individuals have a HEEQ degree from further away than 200km. In the case of the graduates, 882 individuals, or 89.3%, have a HEEQ degree from the region. Furthermore, the students with a HEEQ from the region have the lowest dropout rate. As a result, it seems that the case university mainly has a regional reach, and the students coming from its region are also more likely to graduate.

Table 46. Distribution of the *Distance_HEEQDistrict* attribute.

Distance_HEEQDistrict	Total		Graduates	Dropouts	Dropout rate
	Number	Fraction			
Radius100	1569	86.5%	882	687	43.8%
Radius200	150	8.3%	71	79	52.7%
Radius > 200	94	5.2%	35	59	62.8%
Total	**1813**	**100.0%**	**988**	**825**	**45.5%**

The *HEEQGrade* and *HEEQGradeComp* attributes contain information about the final grade the students achieved in their HEEQ degree. The second attribute is based on the first and is only the compact portrayal of the *HEEQGrade* attribute. The average HEEQ grade of all the students in the dataset is 2.7. The average HEEQ grade of the graduates is slightly better at 2.6, and the average HEEQ grade of the non-graduates is slightly worse at 2.8. The distribution of the *HEEQGradeComp* attribute between all cases in the dataset is shown in Table 47. It can be seen that 49.7% of the students with a *HEEQGrade* of 3 dropped out, while for a HEEQ grade of 2, this was only 33.9% of the students, and for a HEEQ grade of 1, the student dropout rate was 36.7%. It is obvious that the fraction of students with a good HEEQ is higher for the graduates, and the proportion of students with a satisfactory or sufficient HEEQ grade is higher for the dropouts. As a result, this attribute reflects the general assumption that students with a better HEEQ grade are more successful in completing their studies.

Table 47. Distribution of the *HEEQGradeComp* attribute.

HEEQGradeComp	Total		Graduates		Dropouts		Dropout rate
	Number	Fraction	Number	Fraction	Number	Fraction	
1	30	1.7%	19	1.9%	11	1.3%	36.7%
2	511	28.2%	338	34.3%	173	21.0%	33.9%
3	1198	66.1%	603	61.0%	595	72.1%	49.7%
4	74	4.1%	28	2.8%	46	5.6%	62.2%
Total	**1813**	**100.0%**	**988**	**100.0%**	**825**	**100.0%**	**45.5%**

The *TimeHEEQDegree-Application* attribute contains information about the time elapsed between the completion of the HEEQ and beginning of the study at the case university. This attribute has been generated from the difference between the *Enrollment date* and the *Date of HEEQDegree* attributes available in the original dataset. An overview of the distribution of this attribute is shown in Table 48, which displays that only half of the students in the dataset begin their studies within one year of completing their HEEQ. Accordingly, 47.9% of the students in the BA program have more than a one-year gap between successfully completing their HEEQ degree and enrolling at the case university. The dropout rate is highest for the students enrolled at the case university 19-24 months after successfully completing their HEEQ degree.

Table 48. Distribution of the *TimeHEEQDegree-Application* attribute.

TimeHEEQDegree-Application	Total		Graduates	Dropouts	Dropout rate
	Number	Fraction			
< 6 Months	546	30.12%	322	224	41.0%
7-12 Months	398	21.95%	192	206	51.8%
13-18 Months	236	13.02%	160	76	32.2%
19-24 Months	143	7.89%	57	86	60.1%
25-36 Months	187	10.31%	96	91	48.7%
> 37 Months	303	16.71%	161	142	46.9%
Total	**1813**	**100.00%**	**988**	**825**	**45.5%**

One reason for the one year gap between the completion of the HEEQ degree and the beginning of studies could be the fact that 565 students (31.2%) in the dataset have previous study experience – *FirstSemester(No)* – and this means that they spend several months after successfully completing the HEEQ in a different study program. In addition, 35.6% (645) of the students in the dataset do have a completed apprenticeship – *Apprenticeship(Yes)* – possibly completed after the HEEQ and before the beginning of studies. Therefore, in Table 49 a crosstab has been created that confronts the *TimeHEEQDegree-Application* attribute with the *Apprenticeship* and *FirstSemester* attributes. The confrontation of the attributes shows that with increasing time between the HEEQ and the beginning of study, the probability increases that a student has a completed apprenticeship or previous study experience. Nevertheless, 43.9% of the students with an apprenticeship completed their HEEQ within one year before their application. It seems that these students opted for an academic career after they successfully completed their apprenticeship.

Table 49. Crosstab of the *TimeHEEQDegree-Application* attribute and the *Apprenticeship* and *FirstSemester* attributes.

TimeHEEQDegree-Application	Apprenticeship				FirstSemester			
	Yes		*No*		*Yes*		*No*	
	Number	Fraction	Number	Fraction	Number	Fraction	Number	Fraction
< 6 Months	193	29.9%	354	30.3%	455	36.5%	91	16.1%
7-12 Months	90	14.0%	308	26.4%	280	22.4%	118	20.9%
13-18 Months	42	6.5%	194	16.6%	164	13.1%	72	12.8%
19-24 Months	29	4.5%	114	9.8%	82	6.6%	61	10.8%
25-36 Months	78	12.1%	108	9.2%	93	7.5%	94	16.6%
> 37 Months	213	33.0%	90	7.7%	174	13.9%	129	22.8%
Total	**645**	**100%**	**1168**	**100%**	**1248**	**100%**	**565**	**100%**

The attribute *SemestersStudied* was calculated from the difference between the *Deregistration date* and the *Enrollment date*. A descriptive evaluation is presented in Table 50. Accordingly, the average number of semesters spend in the study program across all students in the dataset is 5.1 semesters. This number changes significantly between the target groups graduates and dropouts. Since the regular period of study in the BA program investigated is 7 semesters, the average number of *SemestersStudied* for the graduates is 7.7 semesters. Accordingly, many students are able to finish their studies successfully within 8 semesters. The average *SemestersStudied* for the dropouts is 2.0 semesters with a deviation of 1.5 semesters. It is therefore assumed that a large portion of students that drop out will do so before the completion of the third semester.

Table 50. Distribution of the *SemestersStudied* attribute.

Attribute	Total		Graduates		Dropouts	
	Average	Deviation	Average	Deviation	Average	Deviation
SemestersStudied	5.1	3.1	7.7	1.0	2.0	1.5

A major portion of the attributes in the dataset is the course and module exams and their responding grades. Table 51 descriptively evaluates these with the least missing values. The missing values in the dataset can also be interpreted as 0-values because they indicate that a student did not attend a module or at least in the exam of the module. This non-participation could, for example, be caused by the student deregistering from the exam, by the student skipping the course, or by the student simply not showing up for the exam. The exams with most participants take place between the first and third semester. These exams logically have the least missing values because they are mandatory and most dropouts take place before the third semester.

The module with the most attendees is *Economics*, followed by the *Law* module. The students usually attend these modules in the first semester and, therefore, most of the dropouts do attend the regarding courses and exams. The low average grade of 3.7 in the module *Economics* gives reason to believe that this is a challenging subject for the students. The same applies to the course *Business Ethics* and the module *Quantitative Methods*, which have an average grade of 3.7 and 4.0. This means that many students, whether they graduate or drop out, do not pass the regarding module exam on their first attempt. For the *Intercultural Competencies* module, it is noticeable that the minimum grade is 4.0, while all other modules and courses have a minimum grade of 5.0. This is due to the structure of the module, which extends over two semesters. Therefore, the examinations for the subjects of these modules are not combined and take place at different times of the studies. Therefore, the module only receives a grade when all courses belonging to the module have been successfully completed. Otherwise, the student has no grade for the module.

Table 51. Descriptive analysis of the *Exam* attributes with the least missing or 0-values.

Exam (first attempt)	Missing	Maximum	Minimum	Average	Deviation
Economics	156	1.0	5.0	3.7	1.1
Law	255	1.0	5.0	3.4	1.2
Business Ethics	414	1.0	5.0	3.7	1.2
Cross-company Functions	432	1.0	5.0	3.6	1.2
Quantitative Methods	533	1.0	5.0	4.0	1.2
Controlling[83]	612	1.0	5.0	3.4	1.1
Process Management	637	1.0	5.0	3.4	1.2
Intercultural Competencies	678	1.0	4.0	2.3	0.6
Management	812	1.0	5.0	2.9	0.9

Table 52 compares the number of graduates and dropouts who passed or failed a module or exam based on the first attempt. Accordingly, the totals do not cover all students who graduated or dropped out because some students may have chosen to postpone the exam. Grades 1 to 4 have been condensed into the category *passed*, while grade 5 is condensed into the category *failed*. The comparison clearly shows that a higher percentage of graduates successfully complete the exams at the first try. For example, 85.2% of the graduates who participated in the examination of the

[83] Financial Controlling.

module *Economics* have done so successfully, while only 33.1% of the dropouts successfully completed the module in their first attempt. Merely the module *Quantitative Methods* was passed by only 52.4% of the graduates, which underlines the above assessment that this module is a challenge for all students.

Table 52. Overview of the students successfully completing an exam on the first attempt, distinguished by graduates and dropouts.

Exams (first attempt)	Graduates					Dropouts				
	Passed (Grades 1-4)		Failed (Grade 5)		Total	Passed (Grades 1-4)		Failed (Grade 5)		Total
	Number	Fraction	Number	Fraction	Number	Number	Fraction	Number	Fraction	Number
Economics	830	85.2%	144	14.8%	974	226	33.1%	457	66.9%	683
Law	875	89.8%	99	10.2%	974	173	29.6%	411	70.4%	584
Business Ethics	929	94.6%	53	5.4%	982	223	53.5%	194	46.5%	417
Cross-company Functions	752	77.8%	214	22.2%	966	125	30.1%	290	69.9%	415
Quantitative Methods	381	52.4%	346	47.6%	727	148	26.8%	405	73.2%	553
Controlling	827	85.3%	143	14.7%	970	66	28.6%	165	71.4%	231
Process Management	777	80.8%	185	19.2%	962	49	22.9%	165	77.1%	214
Business Administration	560	90.3%	60	9.7%	620	361	70.0%	155	30.0%	516

5.2.2 Data preparation and analysis plan

The goal of the planned analysis is twofold. First, a predictive model is to be generated that can predict the dropout of the students in the BA program. Second, the modules that may contribute to the dropout of students are to be identified. To achieve the first goal, classification models are created. It was chosen to perform rule induction and decision tree modeling as these methods provide interpretable models that can be used to identify modules that increase the likelihood of dropping out of the study program, which also provides information for the second goal. In addition, association rule modeling has been applied to the dataset in order to find rules or patterns that influence the success or failure of students.

Prior to the model generation, the dataset was analyzed for outliers. Neither the LOF approach nor distance-based outlier detection identified actual outliers. Therefore, all the records were included in the analysis. In addition, the dataset was examined for missing values. The demographic data and the general study information do not contain any missing values, but all the module and course related exam attributes do. This is normal due to the nature of the dataset because not all the students in the dataset completed or attended each course in the program. Some students even dropped out before the first semester ended, or they decided to postpone exams and attempt them in a later semester. Therefore, the missing values also contain information and can be understood as 0-values, and these are kept as such in the dataset. The attributes that specify only descriptive facts, such as the *Course of Study* or the *Status*, which is *Deregistered* for all of the students in the

dataset, are excluded from the analysis because they do not contain values influencing the target variable. Furthermore, the *DeregReason* attribute is excluded because it clearly indicates the class of the target variable, which also applies to the attribute *SemestersStudied*. The remaining attributes are taken into account during modeling. In order to determine the reason for dropping out and because most of the students drop out before the beginning of the 4th semester, the mandatory courses that are usually attended in the first three semesters of the BA study program are considered the most interesting. Therefore, the following analyses will focus on these courses.

In addition, the original dataset has been divided into a training and a test portion. The *TrainingSet* contains 85% of the original dataset, which are 1541 records. The *TestSet* contains the remaining 15% of the records. The target variable *Completion* has a small imbalance between the two target classes. In order to assess whether this inequality has a negative effect on the model performance, a down-sampled dataset was generated from the *TrainingSet*. This *TrainingSetDS* contains 701 records belonging to the target class *Completion(Yes)* and 701 records belonging to the target class *Completion(No)*. To generate and identify well-performing models, the following analyses were done with the *TrainingSet* and the *TrainingSetDS*, and each satisfactory model was tested on the same *TestSet*.

5.2.3 Rule models

The rules were built using the rule induction approach described in Section 3.4.1 of this thesis. In the first step of the rule induction, all attributes of the student and the modules, which take place in the first 3 semesters of the BA study program, were included in the analysis. The Model A shown below is calculated with the *TrainingSet* and the Model B with the *TrainingSetDS*. Both models are based on the same attribute and indicate that the student is likely to complete the studies when the module *Controlling* is successfully completed:[84]

(Model A) IF *Controlling* < *4.5* THEN *Yes* (709 / 60)
ELSE *No* (117 / 576)

(Model B) IF *Controlling* ≤ *4.5* THEN *Yes* (591 / 60)
ELSE *No* (99 / 576)

The performance measures of the models are shown in Table 53 and indicate that both models are able to correctly identify the cases in the training datasets with an *F-score* and an *accuracy* of > 87%. In detail, both models are able to correctly classify more than 91% of the students who drop out before completing their studies. Of the cases predicted to belong to the target class *Completion(No)*,

[84] If it is not stated otherwise, the rules always refer to the first attempt of a student completing the module or course.

83.0% in Model A and 84.1% in Model B are classified correctly. Therefore, both models seem to be able to identify students at risk of dropping out.

Table 53. Performance results of validating Model A and Model B.

Model Name	Accuracy	Recall		Precision		Error rate	F-score
		True(Yes)	True(No)	Pred(Yes)	Pred(No)		
A	88.19%	84.40%	91.44%	92.20%	83.03%	11.81%	87.67%
B	87.23%	82.60%	91.87%	91.04%	84.07%	12.77%	87.82%

In order to find more rules indicating the dropout of students, the *Controlling* module was excluded from modeling and the following rule models were identified:

(Model C) IF *Business Ethics* ≤ *3.5* THEN *Yes* (733 / 136)
ELSE *No* (96 / 508)

(Model D) IF *Process Management* ≤ *4.5* THEN *Yes* (546 / 42)
ELSE *No* (139 / 593)

The performance results of the generated rules are shown in Table 54. Accordingly, Model C can correctly identify 92.4% of the cases belonging to the target class *Completion(No)*, while 79.3% of all the students predicted as belonging to this class actually drop out before the successful completion of the studies.

Table 54. Performance results of validating Model C and Model D.

Model Name	Accuracy	Recall		Precision		Error rate	F-score
		True(Yes)	True(No)	Pred(Yes)	Pred(No)		
C	85.59%	79.88%	92.44%	92.68%	79.31%	14.41%	85.40%
D	85.88%	77.18%	94.58%	93.44%	80.56%	14.12%	87.04%

The Model C indicates that if a student does not have a grade better or equal to 3 in the course *Business Ethics*, which is part of the module *Intercultural Competencies*, the student is likely to drop out, while the successful completion of the course will increase the likelihood of the student graduating. Model D performs slightly better and is able to correctly detect 94.6% of the cases in the dataset that drop out, and 80.6% of the records predicted to be the target class *Completion(No)* actually belong to this class. The model indicates that if students successfully complete the *Process Management* module with a grade of at least 4, the probability of successful completion of their studies increases. The course *Business Ethics* is usually completed between the first and the second semester, whereas the *Process Management* module is attended in the third semester. Therefore, given the current structure of the BA program, Model C offers the possibility of predicting the dropout of students after the second semester, while Model D can predict the dropout at the earliest after the third semester.

In another analysis round based on the *TrainingSet*, a Model E was generated, excluding the *Business Ethics* course from the model calculation. With the *TrainingSetDS* a Model F was created that excludes the *Process Management* module from the model generation. Both models are presented below:

(Model E) IF *Process Management* ≤ 4.5 THEN *Yes* (662 / 42)
IF *Law* > 3.5 THEN *No* (61 / 377)
IF *Economics* > 4.5 AND *Cross-company Functions (second attempt)* > 4.5 AND *Intercultural Competencies* > 3.0 THEN *No* (0 / 2)
ELSE *No* (93 / 224)

(Model F) IF *Business Ethics* ≤ 3.5 THEN *Yes* (609 / 136)
ELSE *No* (73 / 452)

The models indicate even more courses that seem to influence the dropping out of students. Accordingly, if a student has a grade in the *Law* module that is worse than 3, he or she will likely drop out. In addition, Model E states that if a student has not successfully completed the *Economics* module as well as the second attempt on the *Cross-company Functions* module and does not score better than 3 in the *Intercultural Competencies* module, then he or she will not graduate. Since this rule is based on only two records, it should be further investigated as more data records become available. The rule generated in Model F states that students who manage to pass the *Business Ethics* course with a minimum grade of 3 are more likely to successfully complete the program than the ones with a grade worse than 3. The performance results of the two models are presented in Table 55 and show that Model E is able to identify as many as 93.3% of the cases belonging to the target class *Completion(No)*, and Model F can identify 80.6% of these cases.

Table 55. Performance results of validating Model E and Model F.

Model Name	Accuracy	Recall		Precision		Error rate	F-score
		True(Yes)	True(No)	Pred(Yes)	Pred(No)		
E	85.40%	78.81%	93.30%	93.37%	78.61%	14.60%	85.35%
F	83.53%	86.45%	80.60%	81.67%	85.61%	16.47%	83.09%

Due to the nature of the modules and courses on which the above rules are based, they can only predict dropout after the second semester at the earliest. Therefore, a further scenario was modeled, which only includes the modules *Quantitative Methods* and *Economics*, as these are scheduled in the first semester.[85]

[85] In Appendix C the detailed current structure of the investigated BA program is presented.

The rules that have been built are:

(Model G) IF *Economics* ≤ 4.5 THEN *Yes* (697/189)
ELSE *No* (100/358)

(Model H) IF *Economics* ≤ 4.5 THEN *Yes* (580 / 189)
ELSE *No* (72 / 307)

Accordingly, students who successfully complete the *Economics* module are also likely to successfully complete the study program. The Model G performance validation metrics presented in Table 56 show that 75.0% of the students who drop out of the program can be identified correctly. Model H, which points to the same, has a *recall(No)* of 72.3%, while the *precision* of the prediction with 78.2% is slightly higher than that of Model G. Nevertheless, both models seem to be able to recognize more than 70% of the cases in the training datasets that leave the program as early as after the first semester.

Table 56. Performance results of validating Model G and Model H.

Model Name	Accuracy	Recall		Precision		Error rate	F-score
		True(Yes)	True(No)	Pred(Yes)	Pred(No)		
G	78.20%	80.83%	75.04%	79.51%	76.56%	21.80%	75.68%
H	76.89%	79.89%	72.33%	74.27%	78.24%	23.88%	74.79%

Because all generated rule models provide promising results, a model test was performed on the unseen *TestSet* to identify the models that are able to classify most of the students at risk of dropping out. The results of the model test are presented in Table 57. They show that all generated models have significantly higher *accuracy* than the *default accuracy* in the *TestSet*, which is 54.4%.[86] Accordingly, they all support the management of universities in predicting the dropout of students. When examining the performance metrics in more detail, it is noticeable that Models A and B, Models D and E, Models C and F, as well as Models G and H, have identical performance results, which can be traced back to all these rulesets having a commonality. Accordingly, the slight imbalance of the target variable in the dataset does not appear to influence the model performance.

Table 57. Performance results of testing the generated rule models on the unseen *TestSet*.

Model Name	Accuracy	Recall		Precision		Error rate
		True(Yes)	True(No)	Pred(Yes)	Pred(No)	
A	86.76%	79.73%	95.16%	95.16%	79.73%	13.24%
B	86.76%	79.73%	95.16%	95.16%	79.73%	13.24%
C	81.62%	83.78%	79.03%	82.67%	80.33%	18.38%
D	85.29%	77.70%	94.35%	94.26%	78.00%	14.71%
E	85.29%	77.70%	94.35%	94.26%	78.00%	14.71%
F	81.62%	83.78%	79.03%	82.67%	80.33%	18.38%
G	80.88%	89.86%	70.16%	78.24%	85.29%	19.12%
H	80.88%	89.86%	70.16%	78.24%	85.29%	19.12%

[86] Of the 272 data records in the *TestSet*, 148 belong to the *Completion(Yes)* class and the remaining 124 belong to the *Completion(No)* class.

According to the *recall*, Models A and B are able to identify 95.2% of the cases in the testing dataset correctly as belonging to the target class *Completion(No)*. Models G and H are the ones with the highest *precision* for the *Completion(No)* target class, with 85.3% of the cases predicted to belong to that class actually being students who drop out of the study program.

The models that should be preferred for decision support depend on the main tasks, which for the presented case, are to identify the students who are at risk for dropping out. This allows the university decision-makers to intervene and to provide additional support that may increase the likelihood that students will succeed in the study program. In terms of performance metrics, Models A and B should be preferred for decision support because they are able to identify 95.2% of all students who drop out of the program before successfully completing it in the unseen *TestSet*. However, the rule models are based on a module currently scheduled in the third semester of the program; therefore, the capabilities of the model to support early intervention are limited. Model C and Model F are based on the *Business Ethics* course, which is scheduled in the second semester. According to these models, the decision-makers can identify nearly 80% of students who drop out after completing the second semester. Models G and H can identify up to 70% of students that drop out after completing the first semester, allowing for the earliest possible intervention.

The models not only indicate the possibility of early intervention but also demonstrate that the successful completion of the *Business Ethics* course and the *Controlling*, *Process Management* and *Economics* modules positively influence the likelihood that a student will successfully complete the BA program, while failure in the *Law* module increases the risk of dropping out.

5.2.4 Decision tree models

The following decision tree models are generated based on the *C4.5* algorithm following the steps described in Section 3.4.3. Furthermore, it was decided to only use the *TrainingSet* for the model generation because it best reflects the actual structure of the student dataset, and the slight imbalance in the target variable does not appear to affect the performance of the models, which has been noticed during the previous rule induction analysis. In order to find the best possible outcome and a model that will assist the study and examination office in early intervention and in supporting student success, the following scenarios have been used:

- Scenario 1: In conjunction with the general information about the student, exams taking place between the first and the third semester are included in the model generation.
- Scenario 2: In conjunction with the general student information, only the first and second semester exams are included in the analysis.
- Scenario 3: In conjunction with the general student information, only the exams taking place in the first semester are included in the analysis.

- Scenario 4: Only the general information about the student has been included in the analysis.

In addition, the models were pruned to identify models of optimal complexity. The performance results for the best models in all scenarios are shown in Table 58. The performance results show that student dropout cannot be predicted solely on demographic student data. The Model L in question states that only 48.2% of the records in the training dataset associated with the *Completion(No)* target class can be identified by the model, meaning that 51.8% have not been identified. In addition, the performance results show that only 57.9% of the cases identified as *Completion(No)* actually belong to this class. Accordingly, 42.1% of the cases designated by the model as students who drop out are actually students who successfully complete their studies. Therefore, the chance of identifying a student's correct class only on their demographic characteristics is very close to the *default accuracy* of 54.5%.[87]

Table 58. Performance results of validating the generated decision tree models.

Model Name	Scenario	Accuracy	Recall		Precision		Error rate	F-score
			True(Yes)	True(No)	Pred(Yes)	Pred(No)		
I	1	92.47%	96.67%	87.45%	90.22%	95.63%	7.53%	91.33%
J	2	91.43%	98.33%	83.17%	87.50%	97.65%	8.57%	89.74%
K	3	76.44%	76.90%	75.89%	79.26%	73.28%	23.56%	74.44%
L	4	60.48%	70.71%	48.22%	62.07%	57.88%	39.52%	50.66%

Model K, calculated only taking into account the modules that the student generally attends and completes in the first semester, can identify 75.9% of the students who drop out of the program before graduating, Model J is able to identify 83.2%, and Model I can identify 87.5%. Accordingly, all three models perform better than the *default accuracy* in the *TrainingSet* and are therefore tested on the unseen *TestSet*. The results of the model test are shown in Table 59.

Table 59. Performance results of testing Model I, Model J and Model K on the unseen *TestSet*.

Model I | accuracy = 93.38%	**True(Yes)**	**True(No)**	**Class precision**
Pred(Yes)	143	13	91.67%
Pred(No)	5	111	95.69%
Class recall	96.62%	89.52%	
Model J | accuracy = 93.38%	**True(Yes)**	**True(No)**	**Class precision**
Pred(Yes)	145	15	90.62%
Pred(No)	3	109	97.32%
Class recall	97.97%	87.90%	
Model K | accuracy = 81.99%	**True(Yes)**	**True(No)**	**Class precision**
Pred(Yes)	133	37	78.24%
Pred(No)	15	87	85.29%
Class recall	89.86%	70.16%	

[87] 840 of the 1541 examples in the *TrainingSet* belong to the *Completion(Yes)* class and 701 to the *Completion(No)* class.

As the number of attributes increases, the percentage of cases recognized correctly by the models as belonging to the target class *Completion(No)* increases. Accordingly, Model I can correctly identify 89.5% of these cases, while Model K can only detect 70.2%. Nevertheless, Model K can predict the dropout of students as early as the end of the first semester and allows the university to provide additional support to the 87 students identified as potentially at risk for dropping out. Therefore, the generated Model K, shown in Figure 35, gives the management of universities the opportunity to intervene as early as after the first semester to reduce the number of student dropouts. The root of the presented tree model is the *Economics* module, which according to the curriculum is usually attended in the first semester. Hence, if a student does not pass the *Economics* module or does not try to pass it at the first attempt, he or she will probably drop out. If the student has passed the module with at least a grade 4, it is likely that he or she will successfully complete the study program.

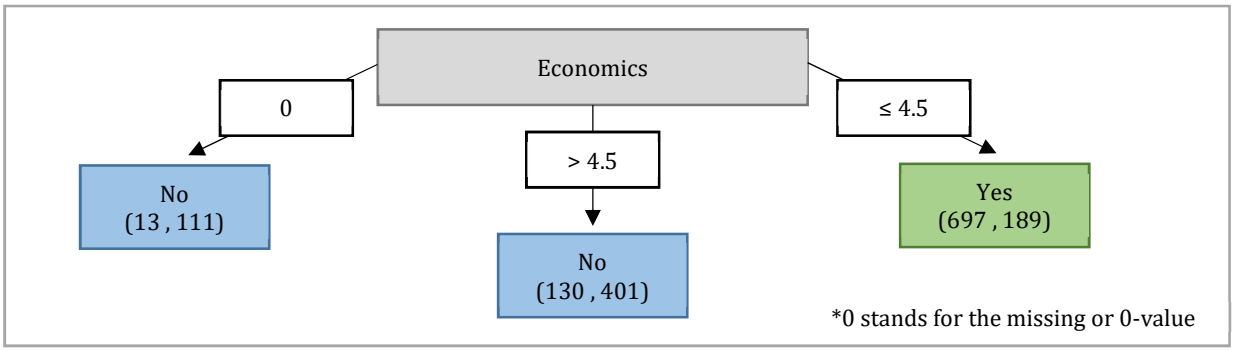


Figure 35. Decision tree Model K.

Figure 36 depicts Model J, which allows university decision-makers to predict the dropout of students after the second semester. In this model, the module *Intercultural Competencies* (first attempt) is the tree root, which indicates that the probability of dropping out is high if a student does not have a grade for this module. If a student does have a grade equal to or better than 2 in this

```
Intercultural Competencies = 0: No {Yes=1, No=570}
Intercultural Competencies > 2.6
|  Cross-company Functions = 0: No {Yes=8, No=9}
|  Cross-company Functions > 3.5: Yes {Yes=117, No=74}
|  Cross-company Functions ≤ 3.5
|  |  Quantitative Methods = 0: Yes {Yes=35, No=0}
|  |  Quantitative Methods > 4.5
|  |  |  Quantitative Methods (second attempt) = 0: Yes {Yes=7, No=5}
|  |  |  Quantitative Methods (second attempt) > 3.8
|  |  |  |  Business Simulation (1001): No {Yes=2, No=6}
|  |  |  |  Business Simulation = 0: Yes {Yes=6, No=0}
|  |  |  Quantitative Methods (second attempt) ≤ 3.8: Yes {Yes=47, No=2}
|  |  Quantitative Methods ≤ 4.5: Yes {Yes=112, No=4}
Intercultural Competencies ≤ 2.6
|  Economics (second attempt) = 0: Yes {Yes=442, No=17}
|  Economics (second attempt) > 4.5
|  |  Intercultural Competencies > 2.1: Yes {Yes=7, No=3}
|  |  Intercultural Competencies ≤ 2.1: No {Yes=0, No=7}
|  Economics (second attempt) ≤ 4.5: Yes {Yes=56, No=4}
```

Figure 36. Decision tree Model J.

module, he or she will probably still drop out if the second attempt to successfully complete the *Economics* module has failed. In addition, Model J indicates that a student is likely to drop out if he or she passes the *Intercultural Competencies* module with a grade worse than 2, the module *Cross-company Functions* has not been attempted at the first try or has been successfully completed with at least a grade 3, the first attempt at completing the *Quantitative Methods* module was failed, the second attempt at completing this module was failed or was passed with a grade 4, but the *Business Simulation* course was successfully completed.

Model I is not as well-suited as the other two models to give universities the ability to provide early intervention to its students, as most of the students drop out before the third-semester exams. However, the model is shown in Figure 37 to identify the potentially challenging third-semester modules that get the students to drop out. The model again shows that the *Intercultural Competencies* module is a predictor of student dropout. Accordingly, if a student does not have a grade for the first attempt on completing the *Intercultural Competencies* module, he or she is likely to drop out. The same can be said for students who did not pass the *Intercultural Competencies* module or did pass it with a grade equal to or worse than 3 and do not have a grade for the *Controlling* module. If they did pass the module *Controlling* but do not have a grade for the *Process Management* module, a dropout appears likely, but only 4 out of 7 cases in the tree leaf actually belong to the *Completion(No)* target class. If they fail to successfully complete the *Controlling* module on the first attempt, the results are also not conclusive because only 52 records of 99 in this tree leaf complete the studies successfully, and the remaining 47 belong to the *Completion(No)* target class.

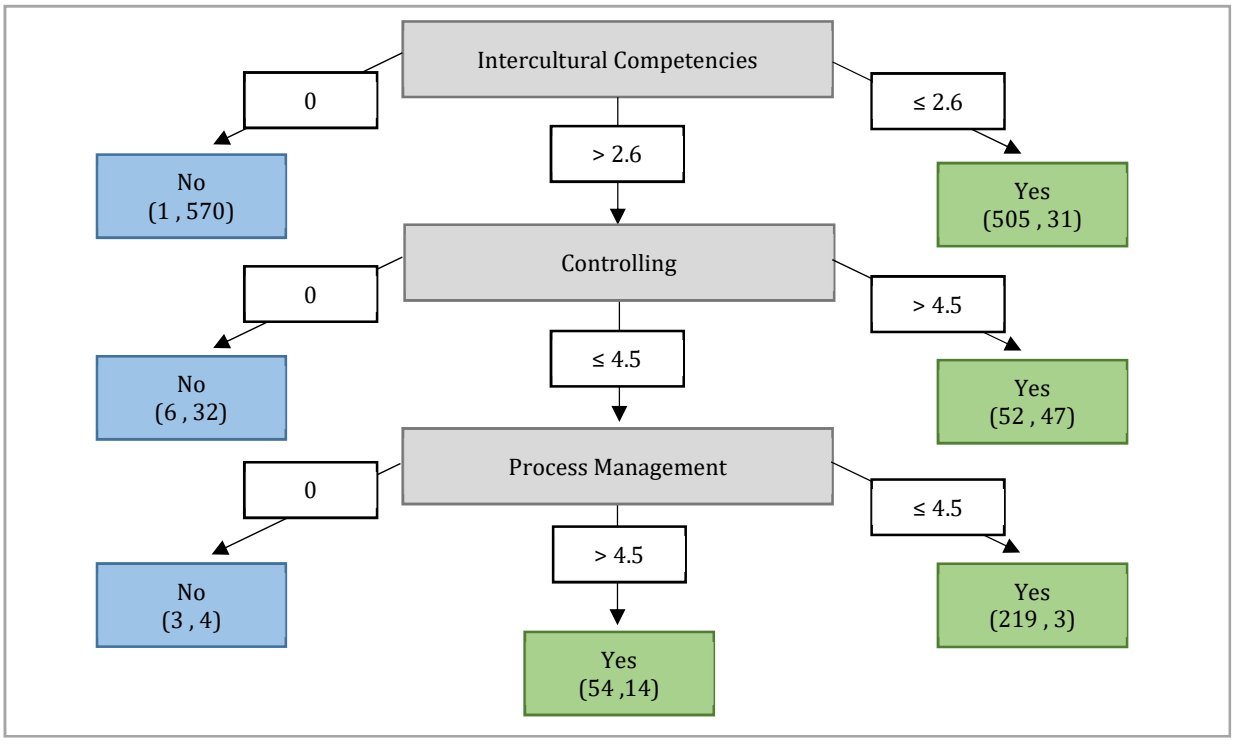


Figure 37. Decision tree Model I.

The models presented and tested above are all informative and convince with the *accuracy* measures, which are significantly higher than the *default accuracy* of 54.4% of the *TestSet*. In particular, Model I can predict the dropout of students well. However, Model J and Model K should be preferred by the university administration if they want to increase the success of their students and reduce their student dropout rate. Model K can identify 70% of the students who are at risk of dropping out after the first semester. These can be provided with additional tutoring, which increases their chances and motivation to continue to study successfully. By additionally applying Model J after the second semester, it is assumed that the university management can identify even more students at risk for dropping out because the model is able to identify as many as 87% of the students that do not graduate. As already mentioned, Model I can identify almost 90% of the students dropping out, but it can only be applied after the third semester, and intervention to increase the student success may be too late at that time for many of the dropout candidates.

All three models do provide interesting information to identify modules and courses that are challenging for students. It is emphasized that the *Economics* and *Intercultural Competencies* modules are demanding for students as they are likely to drop out if they did not succeed to complete the modules on their first attempt. In addition, dropping out is likely if the students do not have a grade for the module. This can be the case if they do not participate in all or one of the course exams that must be completed as part of the modules. This, in turn, may indicate that the student had difficulty with the module's courses during the semester. In addition to the *Economics* and *Intercultural Competencies* modules, Models I and J indicate that the *Quantitative Methods* and *Controlling* modules can also influence the success of students.

5.2.5 Association rules

With frequent pattern mining, associations and correlations between the attributes in a dataset can be identified. In the case presented, this DM method is used to identify relationships between the target variable *Completion* and the other attributes in the dataset. Accordingly, the focus is on association rules that include the target variable *Completion* and characteristics and modules related to student success and dropout.

Before applying the FP-Growth algorithm described in Section 3.3, the dataset was converted to a transactional dataset. Accordingly, all the numerical and nominal variables are converted into binominal variables that either apply (*true*) or do not apply (*false*) for a particular data record. If each numerical attribute in the dataset were first converted into a nominal attribute and secondly to a binominal attribute via *dummy encoding*, an enormous number of attributes would result since each individual number would be converted to a separate attribute. In addition, the likelihood of finding frequent itemsets would be minimized because each new attribute applies only to a small

number of data records.[88] The generation of models for such datasets can be inefficient and difficult (Krzysztof et al. 2007: 235). Therefore, it was decided to discretize the values of the numerical attributes. For the attributes that represent a module, user-specific binning has been performed because of the substantive meaning of the numerical expressions in these attributes. Accordingly, grades between 1.0 and 4.0 indicate that a student has passed a certain exam, and the grade of 5.0 is given if a student did not pass. Hence, grades between 1.0 and 4.0 were assigned to the *passed* category, and grade 5.0 to the *failed* category. For the sake of completeness, *equal-frequency binning*[89] and *entropy-based binning*[90] were also performed with the result that the association analysis did not identify any interesting rules that include the target variable *Completion* and modules. The numerical attributes *Age* and *SemestersStudied* were discretized into three partitions using *equal-frequency binning*.

The performed association rule analysis included the demographic attributes and modules that students usually attend in the first three semesters of their studies. In addition, minimum performance measures have been established. The defined minimum support threshold *minsup* = 0.45 indicates that the items or itemsets must occur in at least 45% of the *DataSetOriginal* to be considered during rule generation. In total, 134 frequent itemsets were found that exceed the *minsup*, and 21 of these frequent itemsets contain the target variable *Completion*. These items and itemsets are listed in Table 60.

Table 60. Frequent itemsets, including the *Completion* target variable, with a *minsup* = 0.45.

Sup.[91]	Item 1	Item 2	Item 3	Item 4
0.545	Completion(Yes)			
0.455	Completion(No)			
0.520	Nationality(German)	Completion(Yes)		
0.486	Distance_HEEQDistrict (Radius100)=true	Completion(Yes)		
0.496	BasicBusinessStudies (passed)	Completion(Yes)		
0.541	Intercultural Competencies(passed)	Completion(Yes)		
0.458	Economics(passed)	Completion(Yes)		
0.483	Law(passed)	Completion(Yes)		
0.456	Completion(Yes)	Controlling(passed)		
0.468	Nationality(German)	Distance_HEEQDistrict (Radius100)=true	Completion(Yes)	
0.475	Nationality(German)	BasicBusinessStudies (passed)	Completion(Yes)	

[88] For example, of the 1813 records in the dataset, only 7.2% (130 data records) are 18 years of age, or of the 1657 cases that have a grade for the *Economics* module, only 5.1% (85 data records) have the grade 2.0.

[89] The numerical expressions of the attribute are subdivided into a predefined number of bins based on the distribution of the data records in the dataset so that each bin ideally contains the same number of data records (Witten et al. 2001: 261; Han et al. 2012: 115-116). Hence, "…the numerical predictor is partitioned into *v* categories, each having *v/n* records, where *n* is the total number of records" (Larose et al. 2015: 41).

[90] For more details, see Witten et al. (2001: 262-265).

[91] *Support* measure.

Sup.[91]	Item 1	Item 2	Item 3	Item 4
0.516	Nationality(German)	Intercultural Competencies(passed)	Completion(Yes)	
0.463	Nationality(German)	Law(passed)	Completion(Yes)	
0.484	Distance_HEEQDistrict (Radius100)=true	Intercultural Competencies (passed)	Completion(Yes)	
0.494	BasicBusinessStudies (passed)	Intercultural Competencies(passed)	Completion(Yes)	
0.456	Intercultural Competencies(passed)	Economics(passed)	Completion(Yes)	
0.480	Intercultural Competencies(passed)	Law(passed)	Completion(Yes)	
0.453	Intercultural Competencies(passed)	Completion(Yes)	Controlling(passed	
0.466	Nationality(German)	Distance_HEEQDistrict (Radius100)=true	Intercultural Competencies(passed)	Completion(Yes)
0.473	Nationality(German)	BasicBusinessStudies (passed)	Intercultural Competencies(passed)	Completion(Yes)
0.461	Nationality(German)	Intercultural Competencies(passed)	Law(passed)	Completion(Yes)

According to the list of frequent itemsets, the variable *Completion* is displayed together with the modules *Controlling*, *Law*, *Economics*, *Intercultural Competencies*, and *Basic Business Studies* in the dataset. From these frequent itemsets, association rules were generated with a *minconf* = 0.6. Accordingly, it was decided to consider only rules as interesting in which at least 60% of the data records displaying the first part of the rule, display the second part as well. Accordingly, at least 490[92] records of the *DataSetOriginal* must be presented by the rule.

The full set of rules that exceed the *minsup*, the *minconf*, and include the variable *Completion* in the rule conclusion is shown in Table 61 in descending order according to the *support* measure. The generated rules display that the successful completion of the modules *Intercultural Competencies*, *Basic Business Studies*, *Economics*, *Controlling*, and *Law* positively influences the successful completion of the study program. The performance metrics show that all rules have a *Support* > 0.45 and, therefore, the frequent itemsets are found in at least 816 records in our dataset. In addition, the *confidence* measures indicate that at least 72% of the records that contain the rule premises also show the rule conclusion. In addition to *support* and *confidence*, the table includes the *Lift* measure, which indicates whether there is a correlation between the premise and the conclusion of the rule. Rule 41 has the highest *Lift* value of 1.75, which indicates that the successful completion of the study program is 1.75 times more likely if the student passes the *Intercultural Competencies* and *Controlling* modules. The positive *Lift* measure for the other rules, which is at least *Lift* > 1.39, shows that the rule body and the rule head are positively correlated. It can therefore be concluded that a student who has passed the *Intercultural Competencies*, *Basic Business Studies*, *Law*, *Economics* and *Controlling* modules at the first attempt is more likely to successfully complete the program. In addition, the rules indicate that both the German nationality and a HEEQ from the region of the

[92] 45% of 1813 cases are 816 examples. 60% of 816 examples are 490.

case university have a positive effect on the successful completion of the study program. However, it is assumed that these rules only exist because the attributes are true for at least 86% of all records in the dataset.

Table 61. Rules that exceed the minimum performance measures and have the *Completion* target variable in the rule conclusion.

No	Rule body	Conclusion	Sup.	Conf.[93]	Lift
1	Intercultural Competencies(passed)	Completion(Yes)	0.541	0.870	1.596
2	Intercultural Competencies(passed)	Nationality(German), Completion(Yes)	0.516	0.830	1.597
3	Nationality(German), Intercultural Competencies(passed)	Completion(Yes)	0.516	0.875	1.605
4	BasicBusinessStudies(passed)	Completion(Yes)	0.496	0.761	1.397
5	BasicBusinessStudies(passed)	Intercultural Competencies(passed), Completion(Yes)	0.494	0.757	1.399
6	Intercultural Competencies(passed)	BasicBusinessStudies(passed), Completion(Yes)	0.494	0.793	1.598
7	BasicBusinessStudies(passed), Intercultural Competencies(passed)	Completion(Yes)	0.494	0.894	1.641
8	Intercultural Competencies(passed)	Distance_HEEQDistrict(Radius100)=true, Completion(Yes)	0.484	0.778	1.600
9	Distance_HEEQDistrict(Radius100)=true, Intercultural Competencies(passed)	Completion(Yes)	0.484	0.877	1.610
10	Law(passed)	Completion(Yes)	0.483	0.835	1.532
11	Intercultural Competencies(passed)	Law(passed), Completion(Yes)	0.480	0.771	1.600
12	Law(passed)	Intercultural Competencies(passed), Completion(Yes)	0.480	0.830	1.534
13	Intercultural Competencies(passed), Law(passed)	Completion(Yes)	0.480	0.925	1.697
14	BasicBusinessStudies(passed)	Nationality(German), Completion(Yes)	0.475	0.729	1.404
15	Nationality(German), BasicBusinessStudies(passed)	Completion(Yes)	0.475	0.769	1.411
16	BasicBusinessStudies(passed)	Nationality(German), Intercultural Competencies(passed), Completion(Yes)	0.473	0.726	1.406
17	Intercultural Competencies(passed)	Nationality(German), BasicBusinessStudies(passed), Completion(Yes)	0.473	0.761	1.600
18	Nationality(German), BasicBusinessStudies(passed)	Intercultural Competencies(passed), Completion(Yes)	0.473	0.765	1.415
19	Nationality(German), Intercultural Competencies(passed)	BasicBusinessStudies(passed), Completion(Yes)	0.473	0.802	1.615
20	BasicBusinessStudies(passed), Intercultural Competencies(passed)	Nationality(German), Completion(Yes)	0.473	0.857	1.650
21	Nationality(German), BasicBusinessStudies(passed), Intercultural Competencies(passed)	Completion(Yes)	0.473	0.899	1.650
22	Intercultural Competencies(passed)	Nationality(German), Distance_HEEQDistrict(Radius100)=true, Completion(Yes)	0.466	0.748	1.600
23	Nationality(German), Intercultural Competencies(passed)	Distance_HEEQDistrict(Radius100)=true, Completion(Yes)	0.466	0.789	1.621
24	Distance_HEEQDistrict(Radius100)=true, Intercultural Competencies(passed)	Nationality(German), Completion(Yes)	0.466	0.843	1.623
25	Nationality(German), Distance_HEEQDistrict(Radius100)=true, Intercultural Competencies(passed)	Completion(Yes)	0.466	0.881	1.617
26	Law(passed)	Nationality(German), Completion(Yes)	0.463	0.801	1.541

[93] *Confidence* measure.

No	Rule body	Conclusion	Sup.	Conf.[93]	Lift
27	Nationality(German), Law(passed)	Completion(Yes)	0.463	0.842	1.544
28	Intercultural Competencies(passed)	Nationality(German), Law(passed), Completion(Yes)	0.461	0.740	1.600
29	Nationality(German), Intercultural Competencies(passed)	Law(passed), Completion(Yes)	0.461	0.780	1.617
30	Law(passed)	Nationality(German), Intercultural Competencies(passed), Completion(Yes)	0.461	0.797	1.543
31	Nationality(German), Law(passed)	Intercultural Competencies(passed), Completion(Yes)	0.461	0.838	1.548
32	Intercultural Competencies(passed), Law(passed)	Nationality(German), Completion(Yes)	0.461	0.887	1.708
33	Nationality(German), Law(passed), Intercultural Competencies(passed)	Completion(Yes)	0.461	0.928	1.702
34	Economics(passed)	Completion(Yes)	0.458	0.786	1.442
35	Intercultural Competencies(passed)	Economics(passed), Completion(Yes)	0.456	0.733	1.601
36	Economics(passed)	Intercultural Competencies(passed), Completion(Yes)	0.456	0.783	1.450
37	Intercultural Competencies(passed), Economics(passed)	Completion(Yes)	0.456	0.916	1.681
38	Controlling(passed)	Completion(Yes)	0.456	0.926	1.700
39	Intercultural Competencies(passed)	Completion(Yes), Controlling(passed)	0.453	0.729	1.598
40	Controlling(passed)	Intercultural Competencies(passed), Completion(Yes)	0.453	0.920	1.701
41	Intercultural Competencies(passed), Controlling(passed)	Completion(Yes)	0.453	0.954	1.750

The first set of association rules in Table 61 does not contain any rules with the target variable class *Completion(No)*. This is due to the fact that only 825 cases in the *DataSetOriginal* belong to this class, which accounts for 45.5% of the total dataset. Accordingly, combinations with the target class *Completion(No)* for the set *minsup* = 0.45 are not sufficiently available. As a result, a second association analysis was performed with a *minsup* = 0.20. Therefore, at least 20% of the dataset (362 records) must contain the item or itemset. The *minconf* measure for rule generation has been increased to 0.7. In total, 65 frequent itemsets that contain the target variable *Completion(No)* were identified, from which the rules containing only the target variable *Completion(No)* in their rule conclusion are presented in Table 62.

The rules detected indicate that a student who fails the *Law* module is nearly 1.8 times more likely to drop out. The regarding confidence measure of 0.806 indicates furthermore a kind of rule accuracy (Han et al. 2012: 416), indicating that 80.6% of the records in the dataset that fail the *Law* module belong to the *Completion(No)* class.

In addition, the rules show that students who fail the *Economics* module are almost 1.7 times more likely to drop out, and 76% of all cases in the dataset who fail the *Economics* module are in the *Completion(No)* class. Accordingly, it can be concluded that the failure of a student in the *Economics* and *Law* modules increases the risk of dropping out. In addition, the attributes *Nationality(German)* and *Distance_HEEQDistrict(Radius100)* are displayed in the rules. This supports the above assumption that they are only included in the rules since they apply to at least 86% of all the cases in the

DataSetOriginal. In addition, the rules indicate that a study duration between 0 and 2.5 semesters increases the risk of students dropping out. This is a logical connection since the students that drop out stay on average 2.0 semesters in the study program, which is shown in Table 50 in Section 5.2.1. In combination with the attribute *SemestersStudied(Range1)* being *true*, the attributes *Apprenticeship(No)*, *FirstSemester(Yes)*, and *HEEQGradeComp(3)=true* are shown as increasing a student's risk to drop out. Since all students in the dataset who only spend up to 2.5 semesters in the study program have dropped out, these are not considered informative.

Table 62. Association rules with *Completion(No)* as rule conclusion and *minsup* = 0.2.

No	Rule body	Conclusion	Sup.	Conf.	Lift
1	Nationality(German), Economics(failed)	Completion(No)	0.228	0.756	1.662
2	Distance_HEEQDistrict(Radius100)=true, Economics = failed	Completion(No)	0.210	0.757	1.665
3	Economics(failed)	Completion(No)	0.252	0.760	1.671
4	Law(failed)	Completion(No)	0.227	0.806	1.771
5	Nationality(German), Law(failed)	Completion(No)	0.205	0.807	1.772
6	Nationality(German), Apprenticeship(No), SemestersStudied(Range1)[94]=true	Completion(No)	0.205	0.997	2.192
7	Apprenticeship(No), SemestersStudied(Range1)	Completion(No)	0.228	0.998	2.192
8	Nationality(German), FirstSemester(Yes), SemestersStudied(Range1)=true	Completion(No)	0.231	0.998	2.192
9	Nationality(German), HEEQGradeComp(3), SemestersStudied(Range1)=true	Completion(No)	0.238	0.998	2.192
10	FirstSemester(Yes), SemestersStudied(Range1)=true	Completion(No)	0.249	0.998	2.192
11	HEEQGradeComp(3), SemestersStudied(Range1)=true	Completion(No)	0.257	0.998	2.193
12	Nationality(German), SemestersStudied(Range1)=true	Completion(No)	0.328	0.998	2.194
13	SemestersStudied(Range1)=true	Completion(No)	0.356	0.998	2.194
14	Distance_HEEQDistrict(Radius100)=true, SemestersStudied(Range1)=true	Completion(No)	0.297	1.000	2.198
15	Distance_PlaceofBirth(Radius100)=true, SemestersStudied(Range1)=true	Completion(No)	0.257	1.000	2.198
16	SemestersStudied(Range1)=true, Economics(failed)	Completion(No)	0.214	1.000	2.198
17	Nationality(German), Distance_HEEQDistrict(Radius100)=true, SemestersStudied(Range1)=true	Completion(No)	0.277	1.000	2.198
18	Nationality(German), Distance_PlaceofBirth(Radius100)=true, SemestersStudied(Range1)=true	Completion(No)	0.245	1.000	2.1978
19	Distance_HEEQDistrict(Radius100)=true, Distance_PlaceofBirth(Radius100)=true, SemestersStudied(Range1)=true	Completion(No)	0.246	1.000	2.198
20	Distance_HEEQDistrict(Radius100)=true, FirstSemester(Yes), SemestersStudied(Range1)=true	Completion(No)	0.210	1.000	2.198
21	Distance_HEEQDistrict(Radius100)=true, HEEQGradeComp(3), SemestersStudied(Range1)=true	Completion(No)	0.216	1.000	2.198
22	Nationality(German), Distance_HEEQDistrict(Radius100)=true, Distance_PlaceofBirth(Radius100)=true, SemestersStudied(Range1)=true	Completion(No)	0.234	1.000	2.198
23	Nationality(German), Distance_HEEQDistrict(Radius100)=true, HEEQGradeComp(3), SemestersStudied(Range1)=true	Completion(No)	0.201	1.000	2.198

[94] *Range1* compacts 0 - 2.5 semesters.

Finding association rules that have the target variable in the consequent of the rule is also the first step in the *association rule-based classification* set out by Liu, Yiming & Wong (2001), who propose the generation of a classification model based on class association rules (CARs). Consequently, in this DM method, association rules of the form *condset* → *y* are identified, where *condset* is a set of items, and $y \in Y$, where Y is a set of class labels (Palanisamy 2006: 13). Rules of this form that exceed *minconf* and *minsup* are considered CARs. From this set of CARs, one or several are selected to generate a classifier that can then be used to predict the target variable (Liu et al. 2001). At the time of writing this dissertation, the RapidMiner program did not provide an application to create or apply CARs. Since RapidMiner was chosen as the DM tool for this thesis, and a RapidMiner EDM-process box is generated for the purpose of easy reconstruction of all performed analyses, it has been decided that the DM method *association rule-based classification* will not be further investigated. Nevertheless, the above-generated association rules provide interesting insights for the university decision-makers, which can be used to reduce the number of students dropping out and increase the quality of the study programs offered.

5.2.6 Discussion of the generated models

The models generated on the basis of the student data show that many informative insights can be extracted with DM methods that support the management of universities. The descriptive analysis of the *DataSetOriginal* showed that on average, students drop out of the BA program after 2 semesters. With the generated rules and decision trees, the decision-makers of universities can predict the student dropout already after completion of the first semester. According to the test of Model G and Model K, it is possible to identify 70% of the students that drop out during their studies after the first semester based on their performance in the *Economics* module.

After the second semester, which includes the modules *Basic Business Studies*, *Intercultural Competencies*, *Cross-company Functions* and *Law*, the rule models Model C and Model F can identify 79.0% of the students that drop out in the *TestSet*, and the decision tree Model J can even identify 87.9% of those students. Because of this significant difference in performance, when testing these models, the decision-makers of the case university should prefer Model J. By applying this model after the end of the second semester, the university can identify more students who might benefit from additional support, especially in the modules *Intercultural Competencies* and *Quantitative Methods*.

The models, which additionally include the modules *IT Management*, *Controlling*, and *Process Management* allow the prediction of dropout after the end of the third semester. Although many students consider to or actually do drop out before that date, it is believed that the university can still identify more dropout candidates who could be retained. According to the performance test results,

Model A can identify most of the students in the test dataset who drop out, and this model is therefore recommended for the task. It indicates that the successful completion of the module *Controlling* has a positive impact on the likelihood of the student continuing to study.

Aside from enabling universities to identify students who are likely to drop out, the generated predictive models and the association rules indicate modules that challenge the students. The generated prediction rules state that students successfully passing the modules *Economics*, *Process Management*, and *Controlling* on the first attempt, as well as the *Business Ethics* course, are likely to successfully complete the study program. In addition, the association rules indicate that the successful completion of the module *Basic Business Studies* leads to the successful completion of the study program. In addition, the generated models indicate that a student is at particular risk for dropping out if the *Economics* and *Law* modules are not successfully completed on the first attempt. The decision tree also states that if the student does not have a grade for the first attempt to finish the *Economics*, *Intercultural Competencies*, and *Controlling* modules, he or she is most likely a dropout candidate. This 0-value in the module attribute can be caused by several occasions, e.g. the student has decided not to take the exam or has already dropped out. Consequently, it can be seen that all three of the applied DM methods provide interesting results for the university decision-makers, which are generally comparable.

Because of the current structure of the study program, which is reflected in the available dataset, it is not clear whether all the topics associated with a module, or just a specific course, pose a challenge to the student and lead to failure. Only for the *Intercultural Competencies* module it is indicated that the successful completion of the *Business Ethics* course increases the likelihood that students successfully complete the study program. This, in turn, indicates that a failure of the course increases the likelihood of dropping out. With this information, the university decision-makers can examine the course to determine the cause of the students struggle. For the *Business Ethics* course, which aims to help students understand and reflect on the economy in a moral context, the hurdle could be the course language English. Another reason for the failure could also be that the importance of the topic is not understood by the student, which could be addressed by making the content more practical by involving company representatives in the class. However, to define the true reasons why this topic is challenging for some students requires a detailed study that could be conducted in the form of interviews with students and faculty.

The regular evaluation of courses is obligatory in Germany. At the case university, this is done with a software-based standardized evaluation questionnaire. Accordingly, the evaluation of the courses is already available for several periods, and therefore, it might not even be necessary to conduct further interviews to identify and address the main challenges in the subjects. However, these evaluations are currently only provided to the lecturers themselves. Therefore, the opportunity of using

them to improve the quality of the study program and reduce the student dropout rate should be discussed with all parties concerned.

5.2.7 Proposal for the usage of the analysis results

The information collected with DM can help universities take targeted action to increase student success, reduce the number of dropouts, improve and secure the quality of study programs, and develop required services. Nevertheless, the orientation of universities on the needs of their students should not mean that the demands set on the student by the university should become lower, because lower demands can lead to reduced chances of the student on the job market (Erhardt, D. 2011: 82). It is much more needed of universities to provide students with the necessary support to reach the demands of the study programs on offer.

As shown in Figure 38, the measures proposed by this research to increase student success and, therefore, to help universities to achieve their tasks and objectives can be provided (A) prior to their studies; (B) during the semester, and (C) after the semester.

Prior to the beginning of studies, the insights generated by DM analyses can empower the university to make the needs and challenges of their study programs more visible. This may result in potential students rethinking their choice of study, which on the downside reduces the number of applications and enrollments for a program. On the upside, student retention could be reduced because students usually choose a program that fits their abilities consciously. The information on the content of a study program is normally presented by the universities on their websites in the form of curriculum and content overviews. In addition, applicants can find out about the detailed contents of the programs by reading the module descriptions, which are usually also available for download. These are often very detailed documents that do not appeal to the applicant. It is therefore recommended that the information on the demands, needs, and opportunities of a study program are made visible in another way. It is conceivable that short video clips are provided that convey the experiences of students from higher semesters or those already graduated. These clips should contain the challenging and rewarding moments they have experienced. In addition, information events could be offered, which include a contribution of current students and graduates. These events are not a new invention, but in addition to the possibilities of the program, the requirements should also be the focus.

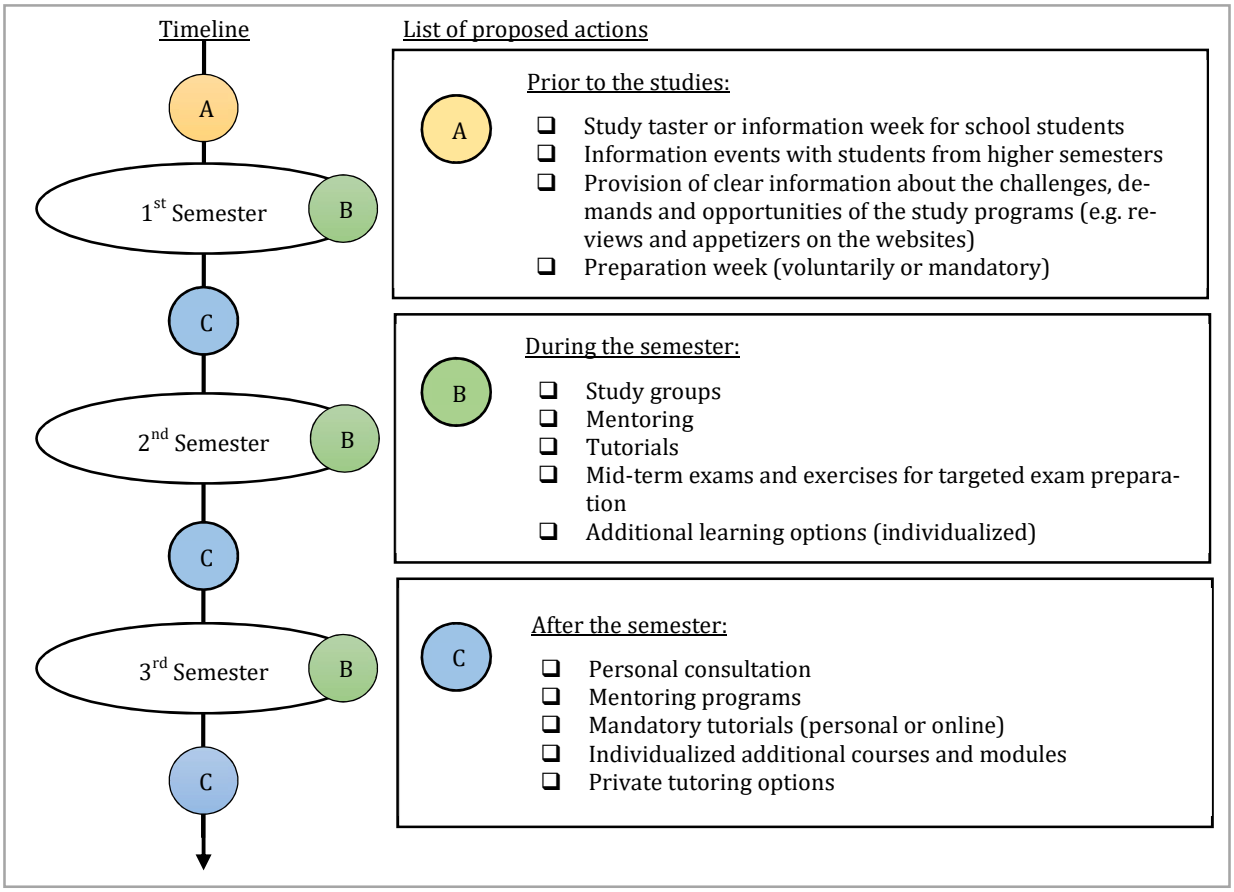


Figure 38. Proposed measures to provide targeted student support throughout the student life, supported by the information generated with DM analysis.

Another measure already established at some universities is a taster program that allows students to attend courses of their choice while they are still at school. In this way, the students can get an idea of the structure and content of the university program, which helps them to decide if this meets their expectations. It is conceivable that such an information period will be offered nationwide and will be compulsory for all pupils aspiring a HEEQ. This could help students to understand that the transfer of university education is not so different from learning at school since the introduction of the Bachelor's and Master's degrees, which could help to correct their expectations. Also possible is a training or preparation phase for already enrolled applicants before the official start of the study program. In these one or two weeks, the new students can become familiar with the requirements of the program, which may help them to succeed in the actual modules. In addition, they will have the opportunity to connect with their new classmates or mentors who could oversee the program, which can positively impact student motivation and perseverance.

During the semester, the students can be offered tutorials that are either available physically or via e-learning. In these tutorials, the students can be given basic and background knowledge about the challenging topics. It is conceivable that online tutorials are individualized and made compulsory. Tests can be used to identify the level of knowledge and knowledge deficits, which are then ad-

dressed through extra learning options. In addition, mentoring programs with higher-semester students may support success, as it can be assumed that they can explain content more tangible and that the inhibition of questions is less because these mentors come from the same target group. Another option to provide support during the semester is regular mid-term exams or exercises that allow the students to assess their current skill level, possibly alerting them that they need to consider further learning options.

After the semester, once the students at risk for dropping out have been identified, the measures and support to improve a student's success can be further individualized and made available to those students who need it most. It is conceivable that students identified as being at risk of dropping out should attend personal consultation, mentoring programs, or tutorials. These services can be tailored precisely to the students who have problems, so they can learn in smaller groups and thus in a more protected environment. In addition, simply informing a student at risk for dropping out about his or her current performance could be a wake-up call that may motivate at least some students to up their game to reach their goal of succeeding in their desired career.

In conjunction with the developments of the digitalization age, the analysis of the available data resources opens up universities even more possibilities to individualize their services. The following measures are proposed:

- Students who are classified as being at risk for dropping out can be offered e-learning opportunities from which they can profit anonymously. Accordingly, students who are unsuccessful in a course or have low scores in a preparatory test may be offered additional e-learning materials or tasks that they need to complete without their classmates' knowledge, which may help them to improve their performance without the feeling of embarrassment. This is possible, because they can decide for themselves if they want to share their additional tasks or not.
- Study groups that are built to perform a specific task can be combined on the basis of the students level of competence. These skill levels can be defined after an online placement test. According to the results of this test, which could be analyzed with DM methods, students with different skill levels can be combined to study groups, and therefore, the stronger students may be able to help the weaker students overcome their problems.
- In a prediction that identifies the risk of a student dropping out or a test defining the student's skill level, the students may be ranked. This can either be a clear rank or an indication of the group the student belongs to – e.g. the best 10% or the worst 10%. Based on their placement, students can receive further tasks and study recommendations. It is also assumed that the motivation of the students may increase because they can visibly monitor their own performance.

- Another suggestion is the integration of a personalized module into the curriculum of a study program. The content of this module should be based on the student's skill level and aimed at meeting the individual challenges a student faces. For example, if a student has problems with the *Quantitative Methods* module, which requires mathematical knowledge, the student may need to attend an additional course that is related to these subjects and that prepares him or her for the actual module exam. In this way, all students must complete additional training in challenging topics, which could be an important measure to increase student success and reduce student dropout rates.

The proposed measures show that student success can be improved by the consistent implementation of DM at German universities, as the management can develop necessary services and support. It was also stated that additional data resources, e.g. e-learning results and log data can provide the university management with even more opportunities to increase student engagement and thus retention. However, the successful implementation of EDM at German universities requires the cooperation and commitment of the stakeholders. Without their support, implementing DM analyses is an insurmountable challenge. Therefore, it is necessary to convince all stakeholders of a university of the benefits of introducing EDM. It is hoped that the analysis presented and the resulting proposed actions will encourage the university management to look for the opportunities and benefits EDM offers and to support the further development and consistent establishment of EDM, which will be beneficial to all university stakeholders.

6 Conclusion

This chapter concludes this thesis with a final discussion and a summary. The first section compares the results of the case studies with the assumptions of the framework that combines the insights gained from the presented DM projects with the management support for German universities. In the second section, the thesis is summarized and pointed to possible future research interests.

6.1 Final Discussion

The framework presented and described in Section 4.2 suggests that predictive models that forecast the enrollment of applicants or the dropout of students help universities solve their management tasks, master their challenges and support their processes. It has been proposed that models predicting the dropout of students provide insights into the reasons for the failure of students, the demands and needs of the study programs, and the needs of the students. While the models generated in Section 5.2 clearly indicate modules that increase student success or the risk of dropping out, and thus identify the demands of the study programs, student needs can only be derived as the support required by the student to meet those demands. The generated models, therefore, do not allow identifying the individual needs and requirements of the student that go beyond special measures to increase their success in the study program investigated, such as additional learning materials or tutorials. For this purpose, additional data resources would be required, e.g. further background information on the financial and social situation of the student and/or further information about the motivation of the student. In addition, it is suggested that the analysis of data from e-learning systems can help identify individual student needs.

Nevertheless, the presented models offer insights that support and inform the decision-makers of the university. For example, required services, such as further learning options, can be offered especially for those courses that increase the risk of student dropout. This helps the university become a service provider, which is one of the tasks specified in the State University Law. It can be argued that the provision of support services to students does not necessarily require DM applications and for most general services, e.g. student counseling, this is true. Nonetheless, the presented study shows that the DM models generated do identify challenging modules and topics that have not been recognized as such. An example in the case study presented here is the module *Intercultural Competencies*, which mainly teaches soft skills on the specifics of working with and in international teams and in a globalized environment as well as doing business in an ethical way. In contrast, the *IT Management* module, which includes the *Databases* course that requires technical know-how, does not seem to challenge the dropout portion of the students in the same way. As a

result, by using DM applications, universities gain additional information that enables them to design and deliver support services that are actually needed and which are unlikely to be otherwise established. This can help the university to develop services that competitors do not offer, which, in turn, increases its uniqueness.

By helping the students, in particular, those students who are considered vulnerable and at risk for dropping out, and exactly in those courses that have been identified as challenging, the management of universities can increase the student success and reduce the dropout rate. Both are important tasks and objectives of German universities. These can also be achieved by increasing the transparency of the study programs so that the future student can assess whether the study program is what he or she expects. As mentioned in the discussion in Section 5.2.7, if the demands of study programs are transparently communicated, this may help students decide before the first semester whether their interests and expectations coincide with the actual content of the chosen study program. In addition, the generated models can contribute to securing and improving the quality of the study program, which is another important task that German universities have to fulfill. Once the courses discerned as demanding are further investigated, issues in structure and content can be identified and improved. At this point, it should be emphasized that this measure requires the support of the university lecturers, which therefore have to be included in the DM project.

The framework in Section 4.2 then proposed that in addition to the direct management support the findings of the DM projects provide to the university decision-makers, they positively support the university in addressing its environmental challenges and achieving further management tasks. A big advantage offered by the use of DM methods for the management of universities, e.g. by increasing the student success and improving the quality of their degree programs, is establishing a positive reputation. A student who has successfully graduated from university becomes an employee, an independent services provider, or an employer. As the university prepares its students for these successes, they positively reflect back on the training they have received. Satisfied graduates, who disseminate their experiences positively, additionally support this and therefore additionally support the universities positive reputation. As mentioned before throughout Chapter 4, a positive reputation supports collaboration with national and international research organizations and companies. These are important for the long-term success and existence of a university and increase the attractiveness of the institution for national and international partnerships as well as qualified researchers and students. In turn, external partnerships may increase the university's ability to attract third-party funding since support from external partners is often an integral part of applications for funds.

It has also been suggested in the framework that the results of the case studies may support the management of universities in developing and directing their strategy. As discussed in Section 2.2,

German universities today have more decision-making autonomy and, therefore, the right and the obligation to formulate, at least in part, their own strategies. As Berthold (2011) states, strategic management for universities means goal-orientation, organized pursuit of the achievement of goals and review of the achievement of goals. The identification of realistic goals and the monitoring of the achievement of goals can be supported by the conducted DM projects, which identify current requirements and challenges. For example, the models that predict the enrollment of applicants in Section 5.1 indicate an increased likelihood of applicants enrolling who apply multiple times to the case university and are rooted in the region. Therefore, the case university may benefit from a profile that emphasizes regional connectivity. This does not mean that internationalization should become less important for the institution, but the university could also have many regional potentials (de Wit 2008: 379-380). Nevertheless, it should be noted that only a few characteristics have been identified that increase the likelihood of enrollment, due to the lack of attributes in the dataset. Accordingly, this research suggests that further data resources on applicants and students should be collected at German universities, which would allow even more individualized support.

In addition, it was proposed in the framework in Section 4.2 that predictive models forecasting the enrollment of applicants can estimate the enrollment numbers, which will help the university management to optimize their admissions process, secure their freshman numbers and optimize their resource allocation. As shown in Section 5.1, with the estimate of enrollment numbers decision-makers can plan demand-oriented. Therefore, it can be ensured that sufficient resources are available to meet and satisfy the needs of all new students. This ensures that students, lecturers, and staff are satisfied even when overbooking. In addition, the need for an optimized admissions procedure was recognized so that German universities with admissions-free study programs can tackle the challenge of overcrowded programs. The solution proposed in Section 5.1.7 is based on the opportunities that DM offers to university decision-makers and thoroughly integrates enrollment prediction models. This suggestion is so far only a vision of the future, how an optimized admissions procedure in a data-driven and digitalized university can look-like. Therefore, the preconditions and the applicability for such a modified admission process have not yet been examined.

The presented case studies confirm that the assumptions made in the framework in Section 4.2 apply. It has thus been shown that the insights gained from two specific DM projects based on relatively small amounts of data already have a very long reach within the university and provide valuable information that supports management decisions. Nevertheless, it should be noted that this proposed support can only be achieved if the university stakeholders are willing to participate, trust, and make use of the lessons learned.

6.2 Summary

The presented study was motivated by the assumption that the current challenges and tasks confronting German universities can be solved and overcome by additional information that can be extracted from the available data resources of the universities with DM methods. In particular, it was assumed that EDM supports management decisions at German universities, as DM enables the detection of complex and unexpected patterns in given datasets. In order to substantiate these assumptions, two literature reviews were carried out in Chapter 1. The first review provided an overview of the state of the art in EDM research to identify work that focused on implementing EDM projects to support the management of educational institutions. It was noted that the benefits of EDM to support the management of educational institutions were recognized by several researchers. However, no details have yet been provided on how the results of the EDM project can be linked to managerial decision-making. This was also noted by Huebner (2013), Dutt et al. (2017), and Aldowah et al. (2019) as a research gap. In addition, Aldowah et al. (2019) and Thakar et al. (2015) identified the need for a framework that ensures the sustainable use of EDM at all levels of educational institutions. The second literature review examined the current state of EDM research in the German higher education sector. It was found that there are only a few contributions from German researchers using data from Germany.

By demonstrating that German universities can benefit from the analysis of their student and applicant data resources with DM, a link has been established between DM outcomes and the management support they provide for universities. This helps to close the research gap found in the first literature review. In addition, current tasks and challenges of German universities were tackled with DM methods, and a contribution to the currently small number of German EDM researches was made.

The above contributions were achieved in several steps. First, the current situation of German universities was examined in Chapter 2, and their core tasks and objectives were identified and prioritized. This was achieved by an analysis of the German State University Laws as well as the university development plans or the university mission statements. It was decided to focus on the German universities of applied sciences because they face additional disadvantages due to their smaller size and shorter existence but, on the contrary, they can adapt more flexibly to innovations. Furthermore, they are the largest body of universities in Germany.

Chapter 3 introduced the CRISP-DM and the DM methods used in the case studies. Afterwards in Chapter 4, the motivation for using the DM approach to assist universities in addressing their challenges and achieving their goals was discussed. This chapter also developed the framework that illustrates the relationship between the insights that can be generated with two specific DM projects and the direct and indirect support they provide to the university decision-makers. This

framework provides an overview of how the results of the proposed EDM projects can be linked to supporting the efficiency of educational institutions. In addition, it can serve as a guideline that shows how DM applications can be used sustainably at German universities. In addition, the framework aims to motivate national and international universities to make use of their available data resources to ensure their long-term success and existence.

The two EDM problems that the framework focuses on are the forecast of enrollment numbers and the prediction of student dropouts. Both analyses were conducted at a case university and are described in Chapter 5. These case studies have shown that the insights proposed in the framework can indeed be extracted from the available student and applicant data. Accordingly, it has been proven that by forecasting the enrollment of their applicants, universities can gain a deeper understanding of their 'clients' and plan their resources according to their needs. The models that have been developed to identify the reasons for and predict the dropout of students enable universities, among other things, to increase students' success, which has a long-term positive impact on their reputation. In addition, the results of both case studies may deliver helpful information for the university's strategic direction and for creating a profile that supports the uniqueness of the university vis-à-vis its competitors. The ideas and measures outlined in Section 5.1.7 and 5.2.7 aim on illustrating how the findings of the DM projects can be transformed into management support and are considered a reference point for universities wishing to implement DM projects and for further research activities. This is also supported by the discussion of the specifics of a DM project at German universities in Section 4.3, which have been encountered during the case studies and the RapidMiner 'EDM-process box' (see Appendix D). Both support the simple implementation and further development of the presented analyses for researchers and the practice.

It is therefore assumed that the contributions in this thesis are useful for both research and practice. For the **research community**, the presented framework, in combination with the case studies, shows how the results of DM projects can increase the efficiency of universities. Accordingly, a contribution was made to close the research gap identified by Huebner (2013), Dutt et al. (2017) and Aldowah et al. (2019) with focus on German universities. In addition, the studies conducted are research contributions based on data from the German education sector, further developing EDM research in Germany. Furthermore, it is believed that the framework presented may serve as a starting point for developing similar models according to the specific needs of other HEIs. In addition, the findings can motivate other research activities that explore the value of more and different data resources for decision-making by leaders in education institutions using different DM techniques or for further challenges.

In practice, the presented research results may especially be useful to German universities, their students, companies, and the German state:

- **Universities** can tailor the contributions made to their individual challenges and leverage knowledge from their stored data to objectively support their decision-making processes. This can help identify current and future student's needs, optimize resource planning and provide administrative support. By providing necessary services, universities can increase the success of their students, which has a positive impact on the reputation of the university and thus ensures its attractiveness in the long run and, therefore, positions the university as a valuable education site.
- Analyses of student and applicant data essentially support potential and actual **students**. As soon as challenges are identified, intervention and support measures can be developed for the students. This ensures that they receive the best possible training and are prepared for their future careers and qualified for the best possible start in their working life.
- By increasing the students' success and providing them with up-to-date education, universities provide **companies** with well-trained young professionals. These professionals become valuable employees who can use modern technologies and techniques that support the progress of their employers and their success in a changing environment.
- The German **state** can also benefit from the successful use of DM techniques in higher education, as high-quality education leads to well-educated graduates who support a thriving economy. In addition, the information gained can show possibilities for change and improvement that apply not only to one institution but also to all universities in Germany.

It should be noted, however, that this study did not provide a solution generally applicable to every university in Germany. Different mission statements, dissimilar organizational structures, and variations of available data resources may require adjustments. Therefore, the analysis processes that are performed during the case studies are provided so that the management of other universities can tailor them to their own conditions and issues. Furthermore, it should be noted that the presented evaluation of the proposals made in the framework is qualitative, which reduces the representativeness of the results. Nevertheless, public universities in Germany are very similar in terms of structure and management requirements as they depend on government regulations. In addition, most of the student-relevant data resources used in this research must be collected by each university in accordance with the *HStatG* (BMJV 1990). Consequently, it is assumed that all German public universities can predict the dropout of students and the enrollment of applicants.

For the continuation of the presented study, it is suggested that further case studies be carried out to additionally investigate the assumptions and suggestions made. In addition, it is proposed to examine the usefulness of analyzing further data resources available to German universities using DM methods in order to support administrative decision-making. Furthermore, researchers from other countries should investigate the propositions made for the applicability to the corresponding

educational environment or should work on developing similar models for the particular requirements of their countries. As the research activities in the EDM communities of recent years show a great potential in using data generated in e-learning environments, another interesting topic for exploration is how the results of the corresponding DM projects can be useful for the university management. Last but not least, the development of a CRISP-DM is conceivable, which is especially geared to the needs and requirements of EDM projects.

In summary, this thesis contributes to closing two existing research gaps in the field of EDM by illustrating how the current challenges of German universities can be met with DM. In addition, accessible results have been provided so that other researchers and practitioners can understand and comprehend this research. The proposed framework and the availability of an 'EDM-process box' ensure this in particular. In addition, concrete recommendations were made on how the DM outcomes can be used to help the university management meet current challenges and achieve tasks and objectives. These additionally increase the applicability of the research for practical use and further research activities.

List of References

Aggarwal, Charu C. (2013): *Outlier Analysis*, New York: Springer International Publishing.

Albrecht, Ella (2017): A Framework for the Estimation of Students' Programming Abilities, in: *Proceedings of the 10th International Conference on Educational Data Mining*, 424-426.

Aldowah, Hanan; Al-Samarraie, Hosman & Fauzy, Wan M. (2019): Educational Data Mining and Learning Analytics for the 21st Century Higher Education: A Review and Synthesis, in: *Telematics and Informatics,* 37: 13-49.

AlHammadi, Dina A. A. & Aksoy, Mehmet S. (2013): Data Mining in Higher Education, in: *Periodicals of Engineering and Natural Sciences,* 1(2): 1-4.

Altvater, Peter; Hamschmidt, Martin & Sehl, Ilka (2010): Prozessorientierte Hochschule, in: *Wissenschaftsmanagement,* 4: 42-47.

An, Truong-Sinh; Krauss, Christopher & Merceron, Agathe (2017): Can Typical Behaviors Identified in MOOCs be Discovered in Other Courses, in: *Proceedings of the 10th International Conference on Educational Data Mining*, 220-225.

Appelfeller, Wieland & Boentert, Annika (2014): Systematische Entwicklung von Prozesslandkarten, in: *Fachhochschule Münster*, https://www.fh-muenster.de/ipl/downloads/Appelfeller_Boentert_Landkarte-1.pdf [05.03.2019].

Askinadze, Alexander & Conrad, Stefan (2018a): Development of an Educational Dashboard for the Integration of German State Universities Data, in: *Proceedings of the 11th International Conference on Educational Data Mining*, 508-509.

Askinadze, Alexander & Conrad, Stefan (2018b): Respecting Data Privacy in Educational Data Mining: An Approach to the Transparent Handling of Student Data and Dealing with the Resulting Missing Value Problem, in: *Proceedings of the 27th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises*, IEEE, 160-164.

Askinadze, Alexander; Liebeck, Matthias & Conrad, Stefan (2019): Using Venn, Sankey, and UpSet Diagrams to Visualize Students Study Progress Based on Exam Combinations, *9th International Conference on Learning Analytics and Knowledge*, 1-5.

Backenköhler, Michael; Scherzinger, Felix; Singla, Adish & Wolf, Verena (2018): Data-driven Approach towards a Personalized Curriculum, in: *Proceedings of the 11th International Conference on Educational Data Mining*, 246-251.

Backhaus, Klaus; Erichson, Bernd; Plinke, Wulff & Weiber, Rolf (2011): *Multivariate Analysemethoden: Eine anwendungsorientierte Einführung*, Vol. 13, Berlin Heidelberg: Springer-Verlag.

Baker, Ryan S. (2014): Educational Data Mining: An Advance for Intelligent Systems in Education, in: *IEEE Intelligent Systems,* 29(3).

Baker, Ryan S. & Yacef, Kalina (2009): The State of Educational Data Mining in 2009: A Review and Future Vision, in: *Educational Data Mining,* 1(1): 3-16.

Bakhshinategh, Behdad; Zaiane, Osmar R.; ElAtia, Samira & Ipperciel, Donald (2017): Educational Data Mining Applications and Tasks: A Survey of the last 10 Years, in: *Education and Information Technologies,* 23(1).

Bala, Manoj & Ojha, Deo B. (2012): Study of Applications of Data Mining Techniques in Education, in: *International Journal of Research in Science and Technology,* 1(4): 1-10.

Barahate, Sachin R. (2012): Educational Data Mining as a Trend of Data Mining in Educational Systems, *International Conference & Workshop on Recent Trends in Technology, Proceedings published in International Journal of Computer*, 11-16.

Bea, Franz X. & Haas, Jürgen (2017): *Strategisches Management*, Vol. 9, Konstanz, München: UVK Verlagsgesellschaft.

Becker, Jörg (2011): Was ist Geschäftsprozessmanagement und was bedeutet prozessorientierte Hochschule, in: Degkwitz, Andreas & Klapper, Frank (Eds.), *Prozessorientierte Hochschule: Allgemeine Aspekte und Praxisbeispiele*, Bad Honnef: BOCK + HERCHEN Verlag, 8-22.

Beikzadeh, Mohammad R. & Delavari, Naeimeh (2005): A New Analysis Model for Data Mining Processes in Higher Education Systems, *6th Information Technology Based Higher Education and Training*, 7-9.

Bengs, Daniel & Brefeld, Ulf (2014): Computer-based Adaptive Speed Tests, in: *Proceedings of the 7th International Conference on Educational Data Mining*, 221-224.

Berens, Johannes; Oster, Simon; Schneider, Kerstin & Burghoff, Julian (2018): *Early Detection of Students at Risk - Predicting Student Dropouts Using Administrative Student Data and Machine Learning Methods*, Wuppertal: Bergische Universität Wuppertal, http://elpub.bib.uni-wuppertal.de/edocs/dokumente/fbb/wirtschaftswissenschaft/sdp/sdp18/sdp18006.pdf [08.02.2018].

Berens, Johannes & Schneider, Kerstin (2018): How to prevent dropouts? Experimental analysis of prevention programs at public and private universities, *11th International Conference on Educational Data Mining*.

Berg, Christian & Dahm, Jochen (2009): Der Bologna-Prozess im Überblick - Stationen und Ziele, AkteurInnen, Strukturen und Umsetzung, in: Neundorf, Anja; Zado, Julian & Zeller, Joela (Eds.), *Hochschulen im Wettbewerb - Innenansicht über die Herausforderungen des deutschen Hochschulsystems*, Bonn: Verlag J. H. W. Dietz Nachf. GmbH, 44 - 60.

Berghoff, Sonja; Federkeil, Gero; Giebisch, Petra; Hachmeister, Cort-Denis; Hennings, Mareike; Roesseler, Isabel & Ziegele, Frank (2009): *CHE Hochschulranking - Vorgehensweise und Indikatoren*, No. 119, Gütersloh: CHE gemeinnütziges Centrum für Hochschulentwicklung, https://www.che.de/downloads/CHE_AP119_Methode_Hochschulranking_2009.pdf [09.03.2019].

Bergner, Yoav; Dröschler, Stefan; Kortemeyer, Gerd; Rayyan, Saif; Seaton, Daniel & Pritchard, David E. (2012): Model-based Collaborative Filtering Analysis of Student Response Data: Machine-Learning Item Response Theory, *5th International Conference on Educational Data Mining*, 95-102.

Berthold, Christian (2011): *"Als ob es einen Sinn machen würde...", Strategisches Management an Hochschulen*, No. 140, Gütersloh: CHE gemeinnütziges Centrum für Hochschulentwicklung, https://www.che-consult.de/fileadmin/pdf/publikationen/CHE_AP140_Strategie.pdf [09.03.2019].

Blümel, Albrecht (2016): *Von der Hoschulverwaltung zum Hochschulmanagement*, Vol. 1, Wiesbaden: Springer Fachmedien.

BMBF (2015): *Bericht der Bundesregierung über die Umsetzung des Bologna-Prozesses 2012 - 2015 in Deutschland*: Bundesministerium für Bildung und Forschung, https://www.bmbf.de/files/Bericht_der_Bundesregierung_zur_Umsetzung_des_Bologna-Prozesses_2012-2015.pdf [28.02.2017].

BMBF (2017a): *Determinaten und Modelle zur Progose von Studienabbrüchen*, Bundesministerium für Bildung und Forschung, https://www.wihoforschung.de/de/dmps-1466.php [13.02.2019].

BMBF (2017b): *Früherkennung abbruchgefährdeter Studierender und experimentelle Studien zur Wirksamkeit von Maßnahmen (FragSte)*, Bundesministerium für Bildung und Forschung, https://www.wihoforschung.de/de/fragste-1329.php [13.02.2019].

BMJV (1990): *Gesetz über die Statistik für das Hochschulwesen sowie für die Berufsakademien (HStatG)*, Bundesministerium der Justiz und für Verbraucherschutz, https://www.gesetze-im-internet.de/hstatg_1990/HStatG.pdf [16.02.2019].

Bogumil, Jörg & Heinze, Rolf G. (2009): Einleitung, in: Bogumil, Jörg & Heinze, Rolf G. (Eds.), *Neue Steuerung von Hochschulen: Eine Zwischenbilanz*, Berlin: edition sigma.

Bolsenkötter, Heinz (1976): *Ökonomie der Hochschule - Eine betriebswirtschaftliche Untersuchung*, Baden-Baden: Nomos Verlagsgesellschaft mbH & Co.

Börgmann, Kathrin & Bick, Markus (2011): IT-Governance in deutschen Hochschulen - eine qualitative Untersuchung, in: *Hochschulmanagement,* 6(2).

Bousbia, Nabila & Belamri, Idriss (2014): Which Contribution Does EDM Provide to Computer-based Learning Environments, in: Peña-Ayala, Alejandro (Ed.) *Educational Data Mining: Applications and Trends*, Switzerland: Springer International Publishing.

Breunig, Markus M.; Kriegel, Hans-Peter; Raymund, Ng & Sander, Jörg (2000): LOF: Identifying Density-based Local Outliers, in: *Proceedings of the International Conference on Management of Data (MOD)*, ACM, 93-104.

Calders, Toon & Pechenizkiy, Mykola (2012): Introduction to the special section on Educational Data Mining, in: *ACM SIGKDD Explorations Newsletter,* 13: 3-6, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.401.9500&rep=rep1&type=pdf.

Chapman, Pete; Clinton, Julian; Kerber, Randy; Khabaza, Thomas; Reinartz, Thomas; Shearer, Colin & Wirth, Rüdiger (2000): *CRISP-DM 1.0*: SPSS Inc.

Chau, Vo Thi Ngoc & Phung, Nguyen Hua (2012): A Knowledge-driven Educational Decision Support System, *International Conference on Computing and Communication Technologies, Research, Innovation and Vision for the Future*, IEEE.

Chawla, Nitesh V.; Bowyer, Kevin W.; Hall, Lawrence O. & Philip, Kegelmeyer W. (2002): SMOTE: Synthetic Minority Over-sampling Technique, in: *Journal of Artificial Intelligence Research,* 16: 321-357.

CHE (2018): *CHE Hochschulranking - Deutschlands größtes Hochschulranking*, CHE gemeinnütziges Centrum für Hochschulentwicklung, https://ranking.zeit.de/che/de/?wt_zmc=fix.ext.zonpmr.che.ranking.che-startseite.bildtext.indikatoren.x&utm_medium=fix&utm_source=che_zonpmr_ext&utm_campaign=ranking&utm_content=che-startseite_bildtext_indikatoren_x [23.11.2018].

Cleve, Jürgen & Lämmel, Uwe (2016): *Data Mining*, Vol. 2, Berlin: Walter de Gruyter GmbH.

Cohen, William W. (1995): Fast Effective Rule Induction, in: *Proceedings of the 12th International Conference on Machine Learning*, 115-123.

Davenport, Thomas H. & Patil, D. J. (2012): Data Scientist: The Sexiest Job of the 21st Century, in: *Harvard Business Review*, https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century [27.02.2019].

de Wit, Hans (2008): Internationalization of Higher Education: Issues and Challenges, in: Kehm, Barbara M. (Ed.) *Hochschulen im Wandel*, Frankfurt am Main: Campus Verlag GmbH, 379-391.

Delavari, Naeimeh; Beikzadeh, Mohammad & Phon-Amnuaisuk, Somnuk (2005): Application of Enhanced Analysis Model for Data Mining Processes in Higher Education Systems, *6th International Conference on Information Technology Based Higher Education and Training*, IEEE.

Delavari, Naeimeh; Phon-Amnuaisuk, Somnuk & Beikzadeh, Mohammad (2008): Data Mining Application in Higher Learning Institutions, in: *Informatics in Education,* 7(1): 31-54.

DESTATIS (2018a): *Hochschulen insgesamt*, Statistisches Bundesamt, https://www.destatis.de/DE/ZahlenFakten/GesellschaftStaat/BildungForschungKultur/Hochschulen/Tabellen/HochschulenHochschularten.html [21.12.2018].

DESTATIS (2018b): *Studierende an privaten Hochschulen insgesamt nach Hochschularten*, https://www.destatis.de/DE/ZahlenFakten/GesellschaftStaat/BildungForschungKultur/Hochschulen/Tabellen/PrivateHochschulenStudierendeInsgesamtHochschulart.html [10.11.2018].

DESTATIS (2018c): *Studierende in Deutschland nach Hochschule*, Statistisches Bundesamt, https://www-genesis.destatis.de/genesis/online/logon?sequenz=tabelleErgebnis&selectionname=21311-0002 [16.02.2019].

DESTATIS (2018d): *Studierende insgesamt nach Hochschularten*, https://www.destatis.de/DE/ZahlenFakten/GesellschaftStaat/BildungForschungKultur/Hochschulen/Tabellen/StudierendeInsgesamtHochschulart.html [10.11.2018].

DFG (2013): *Exzellenzinitiative auf einen Blick - Der Wettbewerb des Bundes und der Länder zur Stärkung der universitären Spitzenforschung*, Bonn: Deutsche Forschungsgemeinschaft (DFG), https://www.dfg.de/download/pdf/dfg_im_profil/geschaeftsstelle/publikationen/exin_broschuere_de.pdf [09.03.2019].

Dhingra, Kriti & Sardana, Kanika Dingra (2017): Educational Data Mining: A Review to its Future Vision, in: *International Journal of Technology Transfer and Commercialisation,* 15(3).

Dohmen, Dieter (2015): *Anreize und Steuerung in Hochschulen - Welche Rolle spielt die leistungsbezogene Mittelzuweisung?*, Berlin: Forschungsinstitut für Bildungs- und Sozialökonomie, https://www.researchgate.net/publication/272678986_Anreize_und_Steuerung_in_Hochschulen_-_Welche_Rolle_spielt_die_leistungsbezogene_Mittelzuweisung [11.12.2018].

Dutt, Ashish; Ismail, Maizatul A. & Herawan, Tutut (2017): A Systematic Review on Educational Data Mining, in: *IEEE Access,* 5.

Dyckhoff, Anna L. (2018): *Action Research and Learning Analytics in Higher Education*, Dissertation, RWTH Aachen.

Dyckhoff, Anna L.; Zielke, Dennis; Bültmann, Mareike; Chatti, Mohamed A. & Schroeder, Ulrik (2012): Design and Implementation of Learning Analytics Toolkit for Teachers, in: *Educational Technology & Society,* 15(3).

Enders, Jürgen (2008): Hochschulreform als Organisationsreform, in: Kehm, Barbara M. (Ed.) *Hochschule im Wandel*, Frankfurt am Main: Campus Verlag GmbH, 232-241.

Engelke, Jens; Müller, Ulrich & Röwert, Ronny (2017): *Erfolgsgeheimnisse privater Hochschulen: Wie Hochschulen atypische Studierende gewinnen und neue Zielgruppen erschließen können*: CHE gemeinnütziges Centrum für Hochschulentwicklung, https://www.che.de/downloads/Im_Blickpunkt_Erfolgsgeheimnisse_privater_Hochschulen.pdf [06.08.2018].

Erhardt, Dominik (2011): *Hochschulen im strategischen Wettbewerb - Empirische Analyse der horizontalen Differenzierung deutscher Hochschulen*, Vol. 1, Wiesbaden: Gabler Verlag.

Erhardt, Manfred; Meyer-Guckel, Volker & Winde, Mathias (2008): *Leitlinien für die deregulierte Hochschule*, Essen: Edition Stifterverband für die deutsche Wissenschaft.

European Higher Education Area (2018): *Full Members of the EHEA*, European Higher Education Area (EHEA) and Bologna Process, http://www.ehea.info/pid34249/members.html [07.12.2018].

Fangmann, Helmut (2014): Hochschulmanagement als politisches Projekt, in: Scherm, Ewald (Ed.) *Management unternehmerischer Universitäten: Realität, Vision oder Utopie?*, München, Mehring: Rainer Hampp Verlag, 35-42.

Fatima, D.; Fatima, Sameen & Prasad, A. V. K. (2015): A Survey on Research Work in Educational Data Mining, in: *Journal of Computer Engineering,* 17(2).

Field, Andy (2013): *Discovering Statistics using IBM SPSS*, Vol. 3, London: SAGE Publications Ltd.

Ganesh, Hari S. & Christy, Joy A. (2015): Application of Educational Data Mining: A Survey, *2nd International Conference in Information Embedded and Communication Systems*, IEEE.

Geis, Max-Emanuel (Ed.) (2017): *Hochschulrecht im Freistaat Bayern - Ein Handbuch für Wissenschaft und Praxis*, Vol. 2, Heidelberg: C.F. Müller GmbH.

Gerhard, Julia (2004): *Die Hochschulmarke*, Vol. 1, Cologne: Josef Eul Verlag GmbH.

Goguadze, Geroge; Sosnovsky, Sergey; Isotani, Seiji & McLaren, Bruce M. (2011): Evaluating a Bayesian Student Model of Decimal Misconceptions, in: *Proceedings of the 4th International Conference of Educational Data Mining*, 301-306.

Goyal, Monika & Vohra, Rajan (2012): Applications of Data Mining in Higher Education, in: *International Journal of Computer Science,* 9(2).

Graumann, Olga; Keck, Rudolf W.; Pewsner, Machail & Rakhkochkine, Antoli (2004): *Schul- und Hochschulmanagement: 100 aktuelle Begriffe: Ein vergleichendes Wörterbuch in deutscher und russischer Sprache*, Hildesheim: Hildesheimer Universitätsschriften.

Haberecht, Christian (2009): Hochschulen zwischen Demokratie und Wettbewerb, in: Neundorf, Anja; Zado, Julian & Zeller, Joela (Eds.), *Hochschulen im Wettbewerb - Innenansicht über die Herausforderungen des deutschen Hochschulsystems*, Bonn: Verlag J. H. W. Dietz Nachf. GmbH, 31 - 43.

Hachmeister, Cort-Denis; Herdin, Gunvald; Roessler, Isabel & Berthold, Christian (2013): *Forschung an deutschen Fachhochschulen/HAW - Gesetzliche Regelungen, Zielvereinbarungen und Förderprogramme im Jahr 2013*, No. 171, Gütersloh: CHE gemeinnütziges Centrum für Hochschulentwicklung, https://www.che.de/downloads/CHE_AP_171_FH_Forschung.pdf [10.07.2018].

Han, Jiawei; Kamber, Micheline & Pei, Jian (2012): *Data Mining: Concepts and Techniques*, Vol. 3, Waltham, USA: Morgan Kaufmann Publishers.

Hartmann, Michael (2006): Die Exzellenzinitiative - ein Paradigmawechsel in der deutschen Hochschulpolitik, in: *Leviathan,* 34(4).

Heinrichs, Werner (2010): *Hochschulmanagement*, Vol. 1, München: Oldenburg Verlag.

Heß, Jürgen (2005): Sind die neuen Steuerungsinstrumente wissenschaftsadäquat? Hochschulen zwischen Ökonomie, Effizienzdruck und Wissenschaftsfreiheit, in: Fisch, Rudolf & Koch, Stefan (Eds.), *Neue Steuerung von Bildung und Wissenschaft*, Bonn: Lemmens Medien.

Hochschule Ansbach (2017): *Prozesslandkarte*, https://www.hs-ansbach.de/hochschule/organisation/qualitaetsmanagement/prozesslandkarte.html [05.03.2019].

Hochschule Karlsruhe (2018): *Erfassen und entwickeln: Prozesslandkarte der HsKA*, Hochschule Karlsruhe, https://www.hs-karlsruhe.de/home/hochschule/einrichtungen/qualitaetsmanagement/prozesse/ [05.03.2019].

Hochschulrektorenkonferenz (2015): *Statistische Daten zu Studienangeboten an Hochschulen in Deutschland: Wintersemester 2015/2016*, Bonn: Hochschulrektorenkonferenz, https://www.hrk.de/fileadmin/_migrated/content_uploads/HRK_Statistik_WiSe_2015_16_webseite_01.pdf [15.06.2018].

Hochschulrektorenkonferenz (2017): *Statistische Daten zu den Studienangeboten an Hochschulen in Deutschland*, Bonn: Hochschulrektorenkonferenz, https://www.hrk.de/fileadmin/redaktion/hrk/02-Dokumente/02-02-PM/HRK_Statistik_BA_MA_UEbrige_WiSe_2017_18_Internet.pdf [26.11.2018].

Holzhüter, Marianne; Frosch-Wilke, Dirk & Klein, Ulrich (2012): Exploiting Learner Models Using Data Mining for E-Learning: A Rule based Approach, in: Peña-Ayala, Alejandro (Ed.) *Intelligent and Adaptive Educational-Learning Systems*, Vol. 17, Berlin, Heidelberg: Springer-Verlag.

Hood, Christopher (1991): A Public Management for all Seasons?, in: *Public Administration,* 69(1): 3-19.

Hornbostel, Stefan (2008): Exzellenz und Differenzierung, in: Kehm, Barbara M. (Ed.) *Hochschule im Wandel*, Frankfurt am Main: Campus Verlag GmbH, 253-266.

Horstmann, Nina & Hachmeister, Cort-Denis (2016): *Anforderungsprofile für die Fächer im CHE Hochschulranking aus Professor(inn)ensich*, No. 194: Centrum für Hochschulentwicklung, https://www.che.de/downloads/CHE_AP_194_Anforderungsprofile_Studienfaecher.pdf [12.03.2019].

Huebner, Richard A. (2013): A survey of educational data-mining research, in: *Research in Higher Education,* 19: 1-13.

Hueglin, Christoph & Vannotti, Francesco (2009): Data Mining Techniques to Improve Forecast Accuracy in Airline Businesses, in: *Proceedings of the SIGKDD 7th International Conference in Knowledge Discovery and Data Mining*, ACM, 438-442.

Ifenthaler, Dirk; Mah, Dana-Kristin & Yau, Jane Yin-Kim (2017): *Studienerfolg mittels Learning Analytics*: e-teaching.org, https://www.e-teaching.org/praxis/erfahrungsberichte/studienerfolg-mittels-learning-analytics [12.03.2019].

Jacob, John; Jha, Kavya; Kotak, Paarth & Puthran, Shubha (2015): Educational Data Mining Techniques and their Applications, *1st International Conference on Green Computing and The Internet of Things*, IEEE, 1344-1348.

Jaeger, Michael (2009): Steuerung durch Anreizsysteme an Hochschulen: Wie wirken formelgebundene Mittelvergabe und Zielvereinbarungen, in: Bogumil, Jörg & Heinze, Rolf G. (Eds.), *Neue Steuerung von Hochschulen: Eine Zwischenbilanz*, Berlin: edition sigma, 45-65.

Jubara, Annett; Kaschlun, Grunhild; Kiessler, Oliver & Smolarczyk, Rudolf (2006): *Glossary on the Bologna Process*, Bonn: Hochschulrektorenkonferenz, https://www.hrk.de/fileadmin/redaktion/hrk/02-Dokumente/02-10-Publikationsdatenbank/Beitr-2006-07_Glossary_Bologna.pdf [12.03.2019].

Kamm, Ruth (2014): *Hochschulreformen in Deutschland: Hochschulen zwischen staatlicher Steuerung und Wettbewerb*, Vol. 18, Bamberg: University of Bamberg Press.

Kehm, Barbara M. (2012): Hochschulen als besondere und unvollständige Organisationen? - Neue Theorien zur 'Organisation Hochschule', in: Wilkesmann, Uwe & Schmid, Christian J. (Eds.), *Hochschule als Organisation*, Wiesbaden: Springer Fachmedien, 17-25.

Kehm, Barbara M. (2015): Deutsche Hocchschulen: Entwicklung, Probleme, Perpektiven, in: *Bundeszentrale für politische Bildung*, http://www.bpb.de/gesellschaft/bildung/zukunft-bildung/205721/hochschulen-in-deutschland?p=all [05.03.2019].

Kemper, Lorenz; Vorhoff, Gerrit & Wigger, Berthold U. (2018): Predicting Student Dropout: a Machine Learning Approach, in: *Karlsruhe Institute of Technology*, https://www.researchgate.net/publication/322919234_Predicting_Student_Dropout_a_Machine_Learning_Approach [12.03.2019].

Klindokmai, Sirikhorn; Neech, Peter; Wu, Yu; Ojiako, Udechukwu; Chipulu, Max & Marshall, Alasdeir (2014): Evaluation of Forecasting Models for Air Cargo, in: *The International Journal of Logistics Management,* 25(3): 635-655.

Klüsener, Marcus & Fortenbacher, Albrecht (2015): Predicting Students' Success Based on Forum Activities in MOOCs, *8th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications*, 925-928.

Koch, Michael; Ring, Markus; Otto, Florian & Landes, Dieter (2014): Combining Statistical and Semantic Data Sources for the Improvement of Software Engineering Courses, in: *Proceedings of the 7th International Conference on Educational Data Mining*, 341-342.

Kocian, Claudia (2007): Prozesslandkarte für Hochschulen, in: *Die neue Hochschule,* 2: 32-36.

Kohmann, Oliver (2012): *Strategisches Management von Universitäten und Fakultäten*, Vol. 1, Wiesbaden: Gabler Verlag.

König, Karsten (2009): Hierarchie und Kooperation: Die zwei Seelen der Zielvereinbarung zwischen Staat und Hochschule, in: Bogumil, Jörg & Heinze, Rolf G. (Eds.), *Neue Steuerung von Hochschulen: Eine Zwischenbilanz*, Berlin: edition sigma, 29-44.

Kotu, Vijay & Deshpande, Bala (2015): *Predictive Analytics and Data Mining*, Waltham (USA): Morgan Kaufmann by Elsevier Inc.

Krücken, Georg & Wild, Elke (2010): Zielkonflikte - Herausforderung für Hochschulforschung und Hochschulmanagement, in: *Hochschulmanagement,* 5(2).

Krüger, André; Merceron, Agathe & Wolf, Benjamin (2010): A Data Model to Ease Analysis and Mining of Educational Data, in: *Proceedings of the 3rd International Conference on Educational Data Mining*, 131-140.

Krull, Wilhelm (2008): Die Exzellenzinitiative und ihre Folgen: Die Hochschule vor neuen Herausforderungen, in: Kehm, Barbara M. (Ed.) *Hochschulen im Wandel: Die Universität als Forschungsgegenstand*, Frankfurt am Main: Campus Verlag, 243-251.

Krzysztof, Cios J.; Witold, Pedrycz; Swiniarski, Roman W. & Kurgan, Lukasz A. (2007): *Data Mining: A Knowlegde Discovery Approach*, New York: Sringer Science+Business Media.

Kumar, Varun & Chadha, Anupama (2011): An Empirical Study of the Applications of Data Mining Techniques in Higher Education, in: *International Journal of Advanced Computer Science and Applications,* 2(3): 80-84.

Lanzendorf, Ute & Paternack, Peer (2009): Hochschulpolitik im Ländervergleich, in: Bogumil, Jörg & Heinze, Rolf G. (Eds.), *Neue Steuerung von Hochschulen: Eine Zwischenbilanz*, Berlin: edition sigma, 13-28.

Larose, Daniel T. & Larose, Chantal D. (2015): *Data Mining and Predictive Analytics*, Vol. 1, New Jersey: John Wiley & Sons Inc.

Lemmerich, Florian; Ifland, Marianus & Puppe, Frank (2011): Identifying Influence Factors of Students Success by Subgroup Discovery, in: *Proceedings of the 4th International Conference on Educational Data Mining*, 345-346.

Lenin, Thingbaijam (2018): A review on Data Mining Algorithms and Attributes for Decision Support Systems in Educational Domain, in: *International Journal of Engineering Research in Computer Science and Engineering,* 5(8): 1-7.

Liu, Bing (2011): *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*, Vol. 2, Berlin, Heidelberg: Springer-Verlag.

Liu, Bing; Yiming, Ma & Wong, Ching Kian (2001): Classification using Association Rules: Weaknesses and Enhancements, in: Grossman, Robert L.; Kamath, Chandrika; Kegelmeyer, Philip; Kumar, Vipin & Namburu, Raju R. (Eds.), *Data Mining for Scientific and Engineering Applications*, Boston (USA): Springer Science+Business Media Dordrecht, 591-605.

Lomas, Laurie (2007): Are Students Customers? Perceptions of Academic Staff, in: *Quality in Higher Education,* 13(1): 31-44.

Luan, Jing (2002): Data Mining and Knowledge Management in Higher Education - Potential Applications, *42nd Annual Forum for the Association for Institutional Research*, 1-16.

Luan, Jing (2004): *Data Mining Applications in Higher Education*, SPSS Executive Report: SPSS Inc., http://www.spss.ch/upload/1122641492_Data%20mining%20applications%20in%20higher%20education.pdf [06.03.2019].

Maassen, Oliver T. (2004): *Die Bologna-Revolution: Auswirkungen der Hochschulreform in Deutschland*, Vol. 1, Frankfurt am Main: Bankakademie-Verlag GmbH.

Manjarres, Andrés V.; Sandoval, Luis Gabriel M. & Suárez, Martha S. (2018): Data Mining Techniques Applied in Educational Environments: Literature Review, in: *Digital Education Review,* 33: 235-266.

Marettek, Christian & Ákos, Barna (2010): Aktuelle Probleme des Hochschulmanagements im Namen der "deregulierten Hochschule", in: *Hochschulmanagement,* 5(1): 2-13.

Martinez, Madeleine S. (2009): Der Wandel des deutschen Hochschulwesens: Von der Ordinarien- zur Wettbewerbshochschule, in: Neundorf, Anja; Zado, Julian & Zeller, Joela (Eds.), *Hochschulen im Wettbewerb - Innenansicht über die Herausforderungen des deutschen Hochschulsystems*, Bonn: Verlag J.H. W. Dietz Nachf. GmbH, 16 - 30.

Mehta, Apurva A. & Buch, Niati J. (2016): Depth and Breadth of Educational Data Mining: Researchers' point of view, *10th International Conference on Intelligence Systems and Control*, IEEE.

Merceron, Agathe (2011): Investigating Usage of Resources in LMS with Specific Association Rules, in: *Proceedings of the 4th International Conference on Educational Data Mining*, 361-362.

Merceron, Agathe; Schwarzrock, Sebastian; Elkina, Margarita; Pursian, Andreas; Beuster, Liane; Fortenbacher, Albrecht; Kappe, Leonard & Wenzlaff, Boris (2012): Learning Paths in a Non-Personalizing e-Learning Environment, in: *Proceedings of the 5th International Conference on Educational Data Mining*, 228-229.

Merceron, Agathe & Yacef, Kalina (2008): Interestingness Measures for Association Rules in Educational Data, *1st International Conference of Educational Data Mining*.

Michalk, Barbara & Richter, Heike (Eds.) (2007): *Verfahren der Qualitätssicherung und Qualitätsentwicklung*, Bonn: Hochschulrektorenkonferenz, https://www.hrk.de/fileadmin/redaktion/hrk/02-Dokumente/02-10-Publikationsdatenbank/Beitr-2007-08_Verfahren_der_QS.pdf [12.03.2019].

Mitchell, Tom M. (1997): *Machine Learning*, Singapore: McGraw-Hill Education Ltd.

Mohamad, Siti Khadijah & Tasir, Zaidatun (2013): Educational Data Mining: A Review, in: *Procedia-Social and Behavioral Sciences,* 97: 320-324.

Mühlenbein, Karen (2006): *Fehlsteuerung von Hochschulreform in Deutschland: Eine Untersuchung der Informationssystem über das Hochschulwesen*, Vol. 1, Bern, Stuttgart, Wien: Haupt Verlag.

Münch, Richard & Pechmann, Max (2009): Der Kampf um Sichtbarkeit: Die Kolonisierung des wissenschaftsinternen Wettbewerbs durch wissenschaftsexterne Evaluationsverfahren, in: Bogumil, Jörg & Heinze, Rolf G. (Eds.), *Neue Steuerung von Hochschulen: Eine Zwischenbilanz*, Berlin: edition sigma, 67-92.

NEPS (2018): *Project National Educational Panel Study (NEPS)*, Leibnitz Institute for Educational Trajectories, https://www.neps-data.de/en-us/projectoverview.aspx [13.02.2019].

Neundorf, Anja (2009): Auf die Plätze, fertig, los! Der Wettbewerb um die Elite und Exzellenzgelder, in: Neundorf, Anja; Zado, Julian & Zeller, Joela (Eds.), *Hochschulen im*

Wettbewerb - Innenansicht über die Herausforderungen des deutschen Hochschulsystems, Bonn: Verlag J. H. W. Dietz Nachf. GmbH, 109 - 124.

Neundorf, Anja; Zado, Julian & Zeller, Joela (Eds.) (2009): *Hochschulen im Wettbewerb - Innenansicht über die Herausforderungen des deutschen Hochschulsystems*, Bonn: J. H. W. Dietz Nachf. GmbH.

Nickel, Sigrun (Ed.) (2014): *Implementierung von Qualitätsmanagementsystemen: Erfahrungen aus der Hochschulpraxis*, No. 163, Gütersloh: CHE gemeinnütziges Centrum für Hochschulentwicklung, http://www.che.de/downloads/CHE_AP_163_Qualitaetsmanagementsysteme_2014.pdf [12.03.2019].

Njenga, James; Rodello, Ildeberto A.; Hartl, Karin & Jacob, Olaf (2017): Identifying Opportunities and Challenges for Adding Value to Decision-Making in Higher Education Through Academic Analytics, in: Rocha, Álvaro; Correia, Anna M.; Adeli, Hojjat; Reis, Luís P. & Costanzo, Sandra (Eds.), *Recent Advances in Information Systems and Technologies*, Vol. 2, Cham: Springer, 474-480.

Paaßen, Benjamin; Jensen, Joris & Hammer, Barbara (2016): Execution Traces as a Powerful Data Representation for Intelligent Tutoring Systems for Programming, in: *Proceedings of the 9th International Conference of Educational Data Mining*, 183-190.

Palanisamy, Senthil K. (2006): *Association Rule Based Classification*, Master Thesis, Worcester Polytechnic Institute.

Papamitsiou, Zacharoula & Economides, Anastasios A. (2014): Learning Analytics and Educational Data Mining in Practice: A Systematic Literature Review of Empirical Evidence, in: *Educational Technology & Society,* 17(4): 49-64.

Pasternack, Peer (2008): Teilweise neblig, überwiegend bewölkt: Ein Wetterbericht zur deutschen Hochschulsteuerung, in: Kehm, Barbara M. (Ed.) *Hochschule im Wandel: Die Universität als Forschungsgegenstand*, Frankfurt am Main: Campus Verlag GmbH, 195-206.

Peña-Ayala, Alejandro (2014): Educational data mining: A survey and a data mining-based analysis of recent works, in: *Expert Systems with Applications,* 41: 1432-1462.

Phumchusri, Naragain & Meneesophon, Panaratch (2014): Optimal Overbooking Decision for Hotel Rooms Revenue Management, in: *Journal of Hospitality and Tourism Technology,* 5(3): 261-277.

Prakash, B. R.; Hanumanthappa, M. & Kavitha, V. (2014): Big Data in Educational Data Mining and Learning Analytics, in: *International Journal of Innovative Research in Computer and Communication Engineering,* 2(12): 7515-7520.

Quinlan, J. Ross (1992): *C4.5: Programs for Machine Learning*, San Francisco (USA): Morgan Kaufmann.

RapidMiner (2019): *One Platform. Does Everything*, https://rapidminer.com/products/ [26.02.2019].

Reinsel, David; Gantz, John & Rydning, John (2018): *Data Age 2025: The Digitization of the World from Edge to Core*, IDC White Paper: Seagate, https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf [12.03.2019].

Roessler, Isabel (2012): *Auswirkungen der Zukunftskonzepte auf die Studienbedingungen*, No. 156: CHE gemeinnütziges Centrum für Hochschulentwicklung, https://www.che.de/downloads/CHE_AP156_Auswirkungen_der_Exzellenzinitiative_auf_die_Studienbedingungen_2012.pdf [12.03.2019].

Romero, Cristóbal & Ventura, Sebastián (2007): Educational Data Mining: A Survey from 1995 - 2005, in: *Expert Systems with Applications,* 33(1): 135-146.

Romero, Cristóbal & Ventura, Sebastián (2010): Educational Data Mining: A Review of the State-of-the-Art, in: *IEEE Transactions on Systems, Man, and Cybernetics - Part C: Applications and Reviews,* 40(6): 601-618.

Romero, Cristóbal & Ventura, Sebastián (2013): Data Mining in Education, in: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* 3(1): 12-27.

Schedler, Kuno & Proeller, Isabella (2009): *New Public Management*, Vol. 4, Stuttgart, Wien: Haupt Verlag.

Scherm, Ewald (2013): Universitäten zwischen Zukunftskonzept und Orientierungslosigkeit: Hochschulautonomie als Chance und Risiko, in: *wissenschaftsmanagement.de*, https://www.wissenschaftsmanagement.de/dateien/downloads-open-access/wissenschaftsmanagement_openaccess_scherm.pdf [07.08.2017].

Scherm, Ewald (2014): Management unternehmerischer Universitäten: (k)eine konfliktäre Beziehung, in: Scherm, Ewald (Ed.) *Management unternehmerischer Universitäten: Realität, Vision oder Utopie?*, München, Mehring: Rainer Hampp Verlag, 1-34.

Scheuer, Oliver & McLaren, Bruce M. (2012): Educational Data Mining, in: Seel, Norbert M. (Ed.) *Encyclopedia of the Science of Learning*, Vol. 3: Springer, 1075-1079.

Schimank, Uwe (2009): Governance-Reformen nationaler Hochschulsysteme: Deutschland in internationaler Perspektive, in: Bogumil, Jörg & Heinze, Rolf G. (Eds.), *Neue Steuerung von Hochschulen: Eine Zwischenbilanz*, Berlin: edition sigma, 123-137.

Schmücker, Stefanie (2011): *Universitätsprofile - Konzeption, Komponenten sowie empirische Umsetzung an deutschen Universitäten*, Vol. 79, München: Bayerisches Staatsinstitut für Hochschulforschung und Hochschulplanung.

Scholz, Tobias M. (2014): Big Data in Faculty Performance Measurement: The Dean's Role in the Brave New (Data) World, in: Scholz, Christian & Stein, Volker (Eds.), *The Dean in the University of the Future*, Vol. 1, München, Mering: Rainer Hampp Verlag, 155-161.

Schönbrunn, Karoline & Hilbert, Andreas (2007): Data Mining in Higher Education, in: Decker, Reinhold & Lenz, Hans-Joachim (Eds.), *Advances in Data Analysis*: Springer Verlag Berlin Heidelberg, 489-496.

Sen, Umesh Kumar (2015): A Brief Review Status of Educational Data Mining, in: *International Journal of Advanced Research in Computer Science & Technology,* 3(1).

Shanghai Ranking (2018): *Academic Ranking of World Universities 2017*, http://www.shanghairanking.com/arwu2017.html [23.11.2018].

Shannon, Claude E. (1948): A Mathematical Theory of Communication, in: *The Bell System Technical Journal,* 27: 379 - 423.

Shum, Simon B.; Baker, Ryan S.; Behrens, John T.; Hawskey, Martin; Jeffery, Naomi & Pea, Roy (2013): Educational Data Scientists: A Scarce Breed, in: *Proceedings of the 3rd International Learning Analytics and Knowledge Conference*, ACM, 278-281.

Sieweke, Simon (2010): Leistungsbewertung im Hochschulbereich durch Peer-Review-Verfahren, in: *Hochschulmanagement,* 5(2): 52-57.

Silva, Carla & Fonseca, José (2017): Educational Data Mining: A Literature Review, in: Rocha, Álvaro; Serrhini, Mohammed & Felgueiras, Carlos (Eds.), *Europe and MENA Cooperation Advances in Information and Communication Technologies*: Springer Verlag International Publishing, 87-95.

Stapel, Martin; Zheng, Zhilin & Pinkwart, Niels (2016): An Ensemble Method to Predict Student Performance in an Online Math Learning Environment, in: *Proceedings of the 9th International Conference of Educational Data Mining*, 231-238.

Suhirman, S.; Zain, Jasni M. & Chiroma (2014): Data Mining for Educational Decision Support: A Review, in: *International Journal of Emerging Technologies in Learning,* 9(6): 4-19.

Sukhija, Karan; Jindal, Manish & Aggarwal, Naveen (2015): The Recent State of Educational Data Mining: A Survey and Future Vision, *3rd International Conference on MOOCs, Innovation and Technology in Education*, IEEE, 354-359.

Tandale, Prashant G. (2016): A Review on Applications of Data Mining Techniques in Higher Education, in: *International Journal of Current Trends in Engineering and Research,* 2(5): 102-107.

Thai-Nghe, Nguyen; Horváth, Thomás & Schmidt-Thieme, Lars (2011): Factorization Models for Forecasting Student Performance, in: *Proceedings of the 4th International Conference on Educational Data Mining*, 11-20.

Thakar, Pooja; Mehta, Anil & Manisha (2015): Performance Analysis and Prediction in Educational Data Mining: A Research Travelogue, in: *International Journal of Computer Applications,* 110(15): 60-68.

Thilagaraj, T. & Sengottaiyan, N. (2017): A Review of Educational Data Mining in Higher Education Systems, in: *Proceedings of the 2nd International Conference on Research in Intelligent and Computing in Engineering*, RICE, 349-358.

von Stuckrad, Thimo; Berthold, Christian & Neuvians, Tim (2017): *Auf dem Hochschulplateau der Studiennachfrage: Kein Tal in Sicht!: Modellrechnung zur Entwicklung der Studienanfängerzahlen bis zum Jahr 2050*, Gütersloh: CHE gemeinnütziges Centrum für Hochschulentwicklung, https://www.che.de/downloads/Laenderbericht_Niedersachsen_2105.pdf [12.03.2019].

Voß, Lydia; Schatten, Carlotta; Mazziotti, Claudia & Schmidt-Thieme, Lars (2015): A Transfer Learning approach for applying Matrix Factorization to small ITS Datasets, in: *Proceedings of the 8th International Conference on Educational Data Mining*, 372-375.

Warnecke, Christian (2016): *Universitäten und Fachhochschulen im regionalen Innovationssystem - eine deutschlandweite Betrachtung*, Vol. 1, Bochum: Universitätsverlag Brockmeyer

Wehrlin, Ulrich (2011): *Universitäten und Hochschulen im Wandel*, Vol. 2, Göttingen: Optimus Verlag.

Witten, Ian H. & Eibe, Frank (2001): *Data Mining: Praktische Werkzeuge und Techniken für das maschinelle Lernen*, München: Carl Hanser Verlag

Zechlin, Lothar (2012): Zwischen Interessensorganisation und Arbeitsorganisation? Wissenschaftsfreiheit, Hierarchie und Partizipation in der 'unternehmerischen Hochschule', in: Wilkesmann, Uwe & Schmid, Christian J. (Eds.), *Hochschule als Organisation*, Wiesbaden: Springer Fachmedien, 41-59.

Zenkert, David (2017): *No-show Forecast Using Passenger Booking Data*, Bachelor Thesis, Lund University.

Zheng, Zhilin; Stapel, Martin & Pinkwart, Niels (2016): Perfect Scores Indicate Good Students!? The Case of One Hundred Percenters in a Math Learning System, in: *Proceedings of the 9th International Conference on Educational Data Mining*, 660-661.

Zheng, Zhilin; Vogelsang, Tim & Pinkwart, Niels (2015): The Impact of Small Learning Group Composition on Student Engagement and Success in a MOOC, in: *Proceedings of the 8th International Conference of Educational Data Mining*, 500-503.

Zheng, Zhuoyuan & Ye Li, Yungpeng C. (2015): Oversampling Method for Imbalanced Classification, in: *Computing and Informatics,* 34: 1017 - 1037.

Appendices

Appendix A. List of Thesis Relevant Publications

Published in 2018:

(1) Hartl, Karin (2018), A first Attempt to Address the Problem of Overbooking Study Programs, in: *Proceedings of the 11th International Conference on Educational Data Mining*, 541-544, http://educationaldatamining.org/files/conferences/EDM2018/EDM2018_Preface_TOC_Proceedings.pdf.

(2) Hartl, Karin & Nakhaeizadeh, Gholamreza (2018), The Potentials of Educational Data Mining for Decision-making at German Universities, Doctorial Consortium Paper, *11th International Conference on Educational Data Mining*, http://educationaldatamining.org/files/conferences/EDM2018/papers/EDM2018_paper_249.pdf.

(3) Hartl, Karin (2018), How can Universities Profit from their Existing Data Resources, in: *Proceedings of the 13th DESRIST Conference*, https://www.researchgate.net/publication/326551345_How_can_universities_benefit_from_their_existing_data_resources-A_Data_Mining_approach.

Appendix B. Further Objectives of the Southern German Universities of Applied Sciences

Identified objectives in the area of *research and transfer*

Research and transfer										
Universities of applied sciences with over 3000 students	Increasing third-party funds and strengthen the research infrastructure	Strengthen and increase research co-operation's	Secure good scientific practice and the quality of research	Support spin-offs and start-ups	Increase patent applications	Strengthen and increase practical cooperation's and knowledge transfer	Increase personal resources and support for researchers	Support research networks and synergies	Increase applied research activities	Increase degree of recognition
HAW Munich	•	•								
TH Nürnberg	•	•	•							
OTH Regensburg	•	•		•	•	•				
HAW Würzburg-Schweinfurt	•	•	•	•						
HAW Weihenstephan-Triesdorf	•			•		•	•			
HAW Kempten		•	•	•						
HAW Augsburg	•			•		•	•	•		
TH Deggendorf	•			•	•	•		•	•	
HAW Rosenheim	•		•	•			•			
HAW Coburg			•				•			
HAW Landshut		•				•		•		•
TH Ingolstadt	•					•		•		•
HAW Neu-Ulm		•		•	•	•		•	•	•
HAW Hof	•					•		•		
HAW Aschaffenburg	•	•							•	
OTH Amberg-Weiden	•	•	•	•				•		
HAW Karlsruhe		•				•				
HAW Heilbronn	•	•	•	•		•		•	•	
HAW Furtwangen		•							•	
HAW Pforzheim										
HAW Aalen				•		•		•		
HAW Reutlingen		•		•		•		•		

Research and transfer										
Universities of applied sciences with over 3000 students	Increasing third-party funds and strengthen the research infrastructure	Strengthen and increase research co-operation's	Secure good scientific practice and the quality of research	Support spin-offs and start-ups	Increase patent applications	Strengthen and increase practical cooperation's and knowledge transfer	Increase personal resources and support for researchers	Support research networks and synergies	Increase applied research activities	Increase degree of recognition
HAW Nürtingen-Geislingen		•				•	•		•	
HAW Mannheim		•				•	•		•	
HAW Offenburg		•				•		•		
HAW Ulm						•			•	
TH Stuttgart		•				•	•	•	•	
HAW Ravensburg-Weingarten						•				
HAW Albstadt-Sigmaringen						•		•		
Count (29)	13	17	7	12	3	19	7	13	9	3

Identified objectives in the area of *internationalization*

Internationalization						
Universities of applied sciences with above 3000 students	Support and increase cooperation with international partners	Support and increase mobility	Increase the amount of international students and incomings	Increase international reputation	Intensify the internationalization strategy	Develop and extent international study programs
HAW Munich	●	●				
TH Nürnberg			●	●		
OTH Regensburg					●	
HAW Würzburg-Schweinfurt	●	●	●			
HAW Weihenstephan-Triesdorf			●			
HAW Kempten		●	●		●	
HAW Augsburg	●	●			●	●
TH Deggendorf	●					
HAW Rosenheim	●	●	●			●
HAW Coburg	●	●				
HAW Landshut						●
TH Ingolstadt						●
HAW Neu-Ulm	●	●	●			●
HAW Hof	●					●
HAW Aschaffenburg		●	●			
OTH Amberg-Weiden		●	●			●
HAW Karlsruhe					●	
HAW Heilbronn	●	●	●		●	●
HAW Furtwangen						●
HAW Pforzheim						
HAW Aalen			●		●	●
HAW Reutlingen		●	●			●

Internationalization						
Universities of applied sciences with above 3000 students	Support and increase cooperation with international partners	Support and increase mobility	Increase the amount of international students and incomings	Increase international reputation	Intensify the internationalization strategy	Develop and extent international study programs
HAW Nürtingen-Geislingen	●					
HAW Mannheim	●	●	●			●
HAW Offenburg					●	
HAW Ulm		●				
TH Stuttgart	●	●	●			
HAW Ravensburg-Weingarten	●			●		
HAW Albstadt-Sigmaringen	●	●	●			
Count (29)	14	15	14	2	7	12

Identified objectives in the area *human resources, infrastructure, organization and social responsibility*

Human resources, infrastructure, organization and social responsibility										
Universities of applied sciences with above 3000 students	Optimize administration and administrative processes	Adjust and extent the service range	Secure sustainability (in general and in staff policy)	Secure good working conditions (family friendliness, social inclusion)	Strengthen diversity management	Extent and support digitalization	Secure and strengthen quality management	Optimize data storage and usage	Improve internal and external communication (establish transparency, profiles)	Secure resource efficiency
HAW Munich	●	●	●	●						
TH Nürnberg	●			●	●	●	●	●		
OTH Regensburg	●				●	●	●	●		
HAW Würzburg-Schweinfurt	●			●	●			●		
HAW Weihenstephan-Triesdorf	●				●	●		●		
HAW Kempten	●			●	●	●		●		
HAW Augsburg				●					●	
TH Deggendorf					●		●			
HAW Rosenheim	●			●	●			●	●	●
HAW Coburg	●	●		●						
HAW Landshut		●							●	
TH Ingolstadt				●	●					
HAW Neu-Ulm	●			●	●				●	
HAW Hof	●		●							●
HAW Aschaffenburg	●			●					●	
OTH Amberg-Weiden	●				●			●		
HAW Karlsruhe				●	●					
HAW Heilbronn			●	●	●		●		●	●
HAW Furtwangen		●	●						●	
HAW Pforzheim										
HAW Aalen	●	●							●	

Human resources, infrastructure, organization and social responsibility										
Universities of applied sciences with above 3000 students	Optimize administration and administrative processes	Adjust and extent the service range	Secure sustainability (in general and in staff policy)	Secure good working conditions (family friendliness, social inclusion)	Strengthen diversity management	Extent and support digitalization	Secure and strengthen quality management	Optimize data storage and usage	Improve internal and external communication (establish transparency, profiles)	Secure resource efficiency
HAW Reutlingen	●	●		●	●	●	●	●	●	
HAW Nürtingen-Geislingen		●								
HAW Mannheim				●	●				●	
HAW Offenburg	●								●	
HAW Ulm		●		●						
TH Stuttgart		●								
HAW Ravensburg-Weingarten		●		●						
HAW Albstadt-Sigmaringen		●								
Count (29)	15	11	4	16	14	5	5	8	11	3

Appendix C. Detailed List of the Courses and Modules in the BA Program

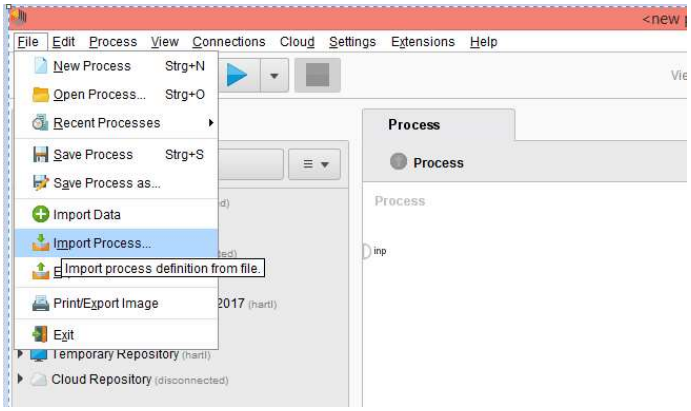
Semester	Modules and courses	Kind of module
1	Quantitative Methods • Business Mathematics • Statistics	Mandatory
	Economics	Mandatory
1 and 2	Basic Business Studies • Introduction to Business Administration • Business English • Business Simulation Game	Mandatory
	Intercultural Competences • Intercultural Competences Seminar • Business Ethics • Second Foreign Language (Spanish, French)	Mandatory
2	Cross-company Functions • Organization Studies • Human Resource Management	Mandatory
	Law • Civil and Public Law • Commercial and Corporate Law • Principles of Taxation	Mandatory
3	IT Management • Databases • Information Systems	Mandatory
	Controlling and Financial Management • Accounting and Bookkeeping • Cost and Activity Calculation • Stock Control and Manufacturing	Mandatory
	Process Management • Marketing • Transport Economics • Material and Production Management	Mandatory
4	Management	Mandatory
	Quantitative Methods in Management	Mandatory
	Start of Specialization (4th until 7th Semester) Examples: • Corporate Finance • Business Information Systems • Accounting • Marketing • …	Elective courses / Specializations
5	Economics II	Mandatory
	Financial Decision Making	Mandatory
	Continuation of Specialization	Elective courses
6	Internship • Seminar to Internship • Internship Project	Mandatory
	Continuation of Specification	Elective courses
7	Bachelor Thesis • Thesis • Thesis Seminar	Mandatory

Appendix D. Download Instructions for the RapidMiner 'EDM-process box'

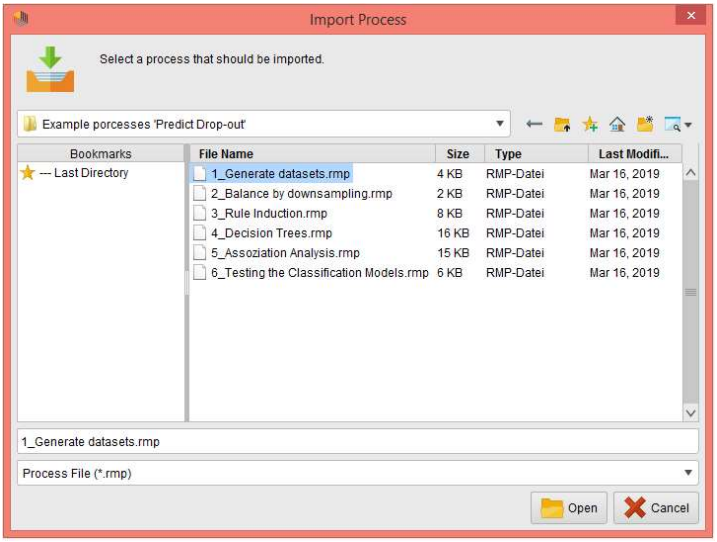
The RapidMiner 'EDM-process box' contains those processes that have been used in the case studies presented in Chapter 5. In the following, the download procedure is described so that the processes can be used as a reference point for practitioners and research.

1. Download the Zip-folder 'EDM-process box' from KITopen:[1]
   https://doi.org/10.5445/IR/1000092542
2. Unpack the Zip-folder.
3. Open the RapidMiner program. If you do not have the program yet, it can be downloaded from: https://rapidminer.com/get-started/
4. Open the downloaded processes in RapidMiner:

   Step 1: Import the process

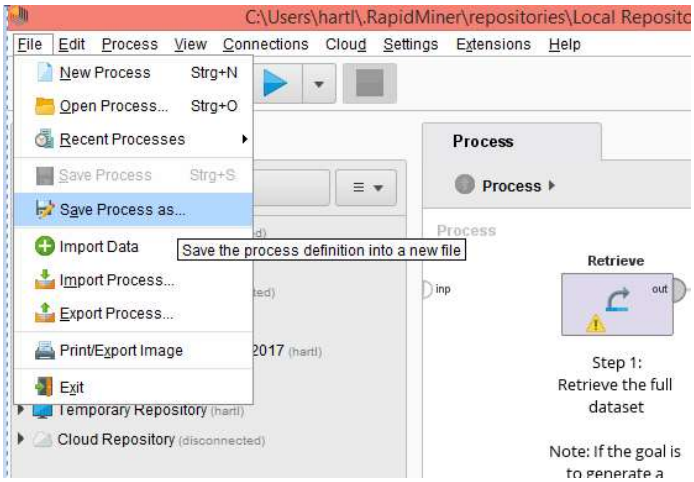


   Step 2: Locate and open the process



[1] Please note that the provided Link is a temporary link. Once the thesis is finalized, the folder is assigned a permanent DOI and will be available long-term.

Step 3: Save the process in your own RapidMiner repository



5. Upload your data and start analyzing