MDSAA

Master's Degree Program in
Data Science and Advanced Analytics

Knowledge extraction from courses and online learning activities

Catarina dos Reis Urbano

Dissertation

presented as partial requirement for obtaining the Master's Degree Program in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

KNOWLEDGE EXTRACTION FROM COURSES AND ONLINE LEARNING ACTIVITIES

by

Catarina dos Reis Urbano

Dissertation presented as partial requirement for obtaining the Master's degree in Advanced Analytics, with a Specialization in Business Analytics

Supervisor / Co Supervisor: Roberto Henriques

November 2022

Acknowledgements

I want to express my gratitude to my advisor, Roberto Henriques, who supported and guided me along the way.

I also want to thank Professor Jorge Mendes for his kindness and all the extra help.

Finally, to my family, friends and dear boyfriend, who have always believed and encouraged me during my academic path and beyond. I am deeply grateful to all those who have given me strength throughout the process and aided me with their endless contributions.

I dedicate my work to you.

“There are only two ways to live your life. One is as though nothing is a miracle. The other is as though everything is a miracle.” (Albert Einstein)

Abstract

Technological advancement has led to the increasing use of all types of electronic devices, which causes large volumes of data to be constantly generated and stored in repositories. This growth in data through Information Technology (IT) systems makes it necessary to continue its exploration and analysis to support institutions in the decision-making process. Due to the importance of education in society, this field has been the target of several studies over the years.

Taking that into account, and knowing that association rules and regression analysis are among the most popular data mining algorithms for finding the hidden patterns in data, the purpose of this paper is to find exciting trends across courses considering the students' grades, as well as study if, and to what extent, the student's learning performance is related to their interaction in moodle. The data used were collected through the netp@ and moodle systems, consisting of all student learning data and activities/logs history. This data belongs to students of all masters who attended the academic years between 2012-2013 and 2020-2021.

We chose Sample, Explore, Modify, Model, and Assess (SEMMA) methodology for the applicability of its steps to accomplish the study's goals.

Through the Partial Least Squares Regression (PLSR) algorithm, it was shown that *Gestão do Conhecimento*, *Metodologias de Investigação* and *Métodos Descritivos de Data Mining* are the most importants courses that affect the grades of Dissertation/Work Project/Intership Report in the Business Intelligence specialization. In addition, according to the predictive model, *Metodologias de Investigação* was the most important variable for predicting the performance of the Dissertation/Work Project/Internship Report of Information Systems and Technologies Management specialization.

Finally, the association rules algorithms used were the Apriori, FP-Growth and Eclat. From their results, it was found that courses with continuous assessment methods achieve better academic performance compared to others. Furthermore, higher levels of online interaction are associated with better achievement.

Keywords: Learning Analytics; Educational Data Mining; Learning Management System; Association Rule Mining; Partial Least Squares Regression; E-learning; Machine Learning.

Contents

Appendices

List of Figures

List of Tables

Acronyms/Abbreviations

CBA	Classification Based on Associations
CS	Computer Science
DT	Decision Tree
e.g.	Exempli gratia (for example)
EDM	Educational Data Mining
GPA	Grade Point Average
i.e.	Id est (that is)
IQR	Inter Quartile Range
IT	Information Technology
KDD	Knowledge Discovery in Databases
LA	Learning Analytics
LCMS	Learning Content Management System
LMS	Learning Management System
MAE	Mean Absolute Error
MLR	Multiple Linear Pegression
MSE	Mean Squared Error
PCA	Principal Component Analysis
PLSR	Partial Least Squares Regression
RMSE	Root Mean Squared Error
SEMMA	Sample, Explore, Modify, Model, and Assess

VIP	Variable Importance in the Projection

1 Introduction

Since the beginning of the Digital Revolution, technology has become increasing part of our daily lives, transforming society and its habits (Cowan *et al.* [18]). In education, this has brought a variety of learning alternatives that go beyond the traditional classroom environment. With the adaptation to this new reality, the digital era, an increase in the use of electronic devices directly leads to a significant rise in data. As new data is regularly created and stored in databases, organizations can gain valuable insights through analytics that help them in decision-making. In the case of higher education institutions, the data collected on students increases from year to year, making it necessary to use data mining techniques to better understand their learning behaviours.

Data Mining or Knowledge Discovery in Databases (KDD) can be defined as the process of discovering new helpful information from vast repositories of data (Algarni *et al.* [6]). It has been applied in several areas, one of them being education, called educational data mining, which has become a topic of great interest.

Educational Data Mining has emerged as an interdisciplinary research field in recent years that aims at the "development of methods for making discoveries within the unique kinds of data that come from educational settings and using those methods to better understand students and the settings in which they learn in" (Baker *et al.* [11]).

According to the *First International Conference on Learning Analytics and Knowledge in 2011*, the Learning Analytics definition is as follows: "Learning analytics is the measurement, collection, analysis, and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs" [33].

Although both subjects seek to find significant knowledge to support teaching and learning aiming to solve educational issues, improve learning experience and institutional effectiveness, their purposes and definitions differ, being that "Learning analytics has a relatively greater focus on human interpretation of data and visualization while EDM has a relatively greater focus on automated methods" (Bajpai *et al.* [10]).

Currently, the most common practice in learning environments is using e-learning or web-based education platforms, such as Learning Content Management System (LCMS) or Learning Management System (LMS), which refers to the "use of information and communication technologies to allow access to online teaching/learning resources" (Arkorful *et al.* [7]. This new context emerged from the widespread use of the internet in education.

Therefore, the usage of a LMS like moodle allows students to learn on their own, providing them with learning materials (like pdf files and video lectures), communication (like announcements, chats, and discussion forums), reporting and instruction tools (Algarni *et al.* [6]). An LMS is defined as an e-learning software application that can create virtual learning environments (Coates *et al.* [15]). In other words, this type of platform comprises digital resources related to the course.

Since one of the biggest challenges that educational institutions face is the exponential growth of educational data that constantly needs to be explored and analyzed, the present research project aims to extract knowledge from it to better manage the associations found between courses and online learning activities on the final grades, using to that end descriptive and predictive analytics techniques.

According to the survey carried out by Peña-Ayala *et al.* [44], which analyzed 240 works on Educational Data Mining (EDM) published between 2010 and the first quarter of 2013, most of them use classification and clustering techniques. This emphasizes that studies that apply association rules and regression models in the field of education are on a much smaller scale.

Although few works study the learning activity patterns, some described in Chapter 2, none of them consider a comprehensive set of engagement metrics and their impact on academic performance, taking into account the structure and evaluation methods of the courses. Besides that, since limited research is available concerning the relationships between master's degrees, and almost always only two or three courses are included in the analysis, none of them studies their impact on the final course project.

To address these gaps, we put into practice association rules and a regression model. Specifically, we employ Apriori, Eclat, and F-growth algorithms to discover how online learning activities influence the final grades of two courses. Furthermore, we used the PLSR model to see which courses most affect the final grades of the Dissertation/Work Project/Internship Report.

Compared to others, the main advantage of these methods is that both are the most suited to discover and study relationships between variables, given the nature of the data. It can be helpful to use association rules in e-learning environments, as they can catch meaningful correlations between the different features of the dataset. In particular, they allow us to understand the link between student performance and the different factors that can positively or negatively affect their learning experience. In addition, they intuitively present the results, making their interpretation very easy for teachers, as they require less knowledge in Data Mining than other methods. On the other hand, the Partial Least Squares Regression technique handles the multicollinearity problem while modelling the shared underlying information of the predictors and response variable.

The remainder of the document is organized as follows: Chapter 2 presents some of the related works from the literature, Chapter 3 describes all the steps performed until the modelling phase, giving an overview of the methods considered in this work, Chapter 4 discusses the experiments conducted and the results obtained based on the evaluation metrics, Chapter 5 presents the main conclusions of the project, and finally, Chapter 6 concludes the thesis by naming the limitations faced and the future works to be pursued.

2

Literature Review

Several studies have been carried out in the last few decades by researchers in the fields of EDM and Learning Analytics (LA). All these studies vary in both their research purpose and the methods that they use. While in some of them the objective is to encounter the most suitable predictive model to predict students' academic performance, in others the aim is to find the factors that most contribute to students' success along with new strategies to improve their learning achievement.

Different machine learning algorithms have been addressed over the years in the most varied areas, divided into unsupervised and supervised learning techniques. Whereas supervised learning focuses on predictive analysis, unsupervised learning aims at a more descriptive analysis.

The first objective of this dissertation lies in understanding the relationships between courses. That said, and knowing that there are two types of supervised learning algorithms, classification and regression, respectively, this study will focus on the last one. Hence, some examples of relevant studies in this field will be displayed.

Gouzouasis *et al.* [24] examined the relationship between participation and performance in music courses and performance in other academic courses, using a sample of 130217 British Columbia students. The methods employed were linear regression and t-tests. The results showed a positive correlation between music courses and other courses regarding student performance in the different disciplines. In addition, the t-tests indicated a consistent pattern of differences in academic performance between students who attended 11th-grade music courses and those who did not participate in any 11th-grade music courses. To sum up, the course Band 11 was associated with better overall performance. In comparison, participation in the music course was associated with better performance in mathematics and biology but not in English.

Haverila *et al.* [25] analyzed how the e-learning experience of students affects perceived learning outcomes. This research was carried out at Tamk University of Applied Sciences in Tampere and the data was collected by a questionnaire, resulting in 57 observations. The Modified 3P Learning Model analysis was performed on the JMP 1-2-3 software program by SAS using multiple regression. Through the results obtained by the model, it can be established that the proposition "The students' prior E-learning experience as presage variable correlates significantly with the learning process" was not satisfied. On the contrary, the model supported the second proposition: "The students' prior E-learning experience as presage variable correlates significantly with the learning outcome variables". In addition, PLSR was used to test the model's validity and relative importance of the variables, where the critical collaborations and meetings were used as explanatory variables and the effectiveness, productivity, and amount of learning as response variables. The results showed the importance of prior experience for successful e-learning outcomes. This study extended the first part (presage) of a work already done by incorporating e-learning experience into the 3P model.

Gamulin *et al.* [21] aimed at understanding the relationships between scores on written midterm exams, scores on web-based formative assessment during seminar teaching, scores on web-based formative assessment during laboratory teaching, scores and time used for online self-assessment tests, number of Moodle logins, number of approaches to specific Moodle resources, and final exam grades. The data gathered from 302 students included the assessment scores, the time spent on Moodle, and logs during the course. The models employed were Principal Component Analysis (PCA) and PLSR, with ten independent variables (counting-based and duration-based features, scores, among others) and four dependent variables (final grades features). The software used to conduct the analysis was Matlab and PLS toolbox. Two metrics were used to evaluate the performance of models, the first one was Root Mean Squared Error (RMSE) and the second was the correlation coefficient (R^2). The results showed that the PLSR is preferable, and the outcome variable that produces better results was y2 (final oral exam grade at first three terms calculated as the average of all three terms).

Oliveira *et al.* [43] used data from 1977 to 2012 to identify the most relevant determinants of aggregate demand for higher education in Portugal. For this end, the PLSR model was employed. In this study, the dependent variable is the aggregate demand. The independent variables are: personal disposable income, government spending in higher education as a percentage of gross domestic product, the wage premium of a tertiary education degree, the ratio of tertiary education fees to the minimum wage, the percentage of females in higher education, the portion of the population with a higher education degree, the participation rate in primary education lagged seven years, the participation rate in secondary education lagged two years, the percentage of enrolled students completing the last year of secondary education, the number of higher education institutions in the Portuguese system, the number of years of compulsory education, the duration of undergraduate study programs, and the percentage of programs adapted to the Bologna guidelines. As a feature selection technique, the VIP was employed to identify the most relevant factors. The efficiency and reliability of the PLSR model were evaluated by the percentage of explained variance (R^2Y), predictive capacity (Q^2), and goodness of the fit (R^2). Two-component were extracted, and the results suggested that Portuguese aggregate demand for public higher education is more affected by policy and social context factors than by economic variables.

Wang *et al.* [51] conducted statistical analyses at different types of Chinese universities to explore students' experiences and attitudes toward contract cheating and the contextual characteristics that relate to these behaviours. Data were collected from four kinds of Chinese universities through questionnaires, one for teachers, and another for students, resulting in a total of 509 observations. First, Pearson's linear correlation tests were conducted to find potential factors in contract cheating, and then, PLSR was employed. The results indicate that both personal and institutional reasons significantly influence cheating intentions. The internal motivations for students to cheat were the desire to obtain a good final grade in the course, dissatisfaction with the learning outcomes, and a low sense of accomplishment. External motivations that led to a high cheating rate were inadequate teacher feedback about cheating assignments, lack of institutional regulations and cheat detection software tools.

The types of unsupervised tasks can be categorized in clustering, visualization, and association rules. As stated earlier, one of our goals is to detect correlations between student behaviours on moodle platform with their performance to determine what is positively or negatively impacting their learning experience. Therefore, the second part of this study will focus on association rules. Taking that into account, some papers that address this topic are described below.

Morris *et al.* [40] explored students' behaviour and performance in online courses, using data from the University System of Georgia. The results revealed statistically significant differences in the behaviours between withdrawers and completers and between successful and non-successful completers. Furthermore, the students' activities showed to be significantly related to their achievement, being the variables that most explain the final grades: the number of discussion posts viewed, number of content pages viewed, and seconds of viewing discussions.

Hung *et al.* [28] aimed to identify and analyze patterns of online learning behaviours and predict student performance. The results showed that the most common online learning activities were logging into the LMS and accessing course materials. In addition, students were classified into three groups. The first and second groups corresponded to above-average performance students, whereas the third group is relative to below-average performance students. In summary, the most crucial variable to predict student performance is the frequency of accessing course materials.

Romero *et al.* [48] intended to identify rare associations from e-learning data, Id est (that is) (i.e.), the discovery of unusual student behaviours in the moodle system that may be significant to consider. Data were collected from 230 students in five Moodle courses on computer science at the University of Córdoba, where the algorithms used and compared were: Apriori-Frequent, Apriori-Infrequent, Apriori-Inverse, and Apriori-Rare. The reason behind applying rare association rule mining algorithms is due to the imbalanced nature of educational data. With this study, the authors showed the importance of using those algorithms in the academic field. The results showed that the number and the time spent on doing assignments and quizzes are directly related to students' performance and can be used to help the students identified at risk of failing the final exam. Also, the time spent on the forum and the number of submitted and read messages on the forum influence the final mark.

Badr *et al.* [9]'s objective was to identify the relationships between courses and then use those associations to predict student's grades, using for that purpose records of mathematics graduate students of Kansas State University. A peculiarity of this study is that it built an algorithm (Classification Based on Association Rule Mining) to predict students' achievement before enrolling in the course in question. Primarily, only the courses related to the programming course have been selected. The results revealed that mathematics courses don't influence students' performance in programming courses, contrary to English courses.

Yoo *et al.* [53] aimed to discover course sequences and understand which ones may influence students' dropout likelihood. This study was made at a public university using sequential pattern mining. The data gathered includes student characteristics and enrollment information for computer science courses. The findings show that students, especially females, often follow course trajectories recommended by their academic advisors. Also, the introductory courses in computer programming proved to be the most important ones, followed by the fundamental math courses.

Moubayed *et al.* [41] focused on the association between engagement and academic performance in a second-year undergraduate science course offered at a North American university. The results indicate that student engagement is highly correlated with academic achievement.

Dahdouh *et al.* [19] built a recommender system using association rules. The objective of this study was to understand the relationships between students' activities in each course and then advise the pupils on the appropriate learning materials and the most suitable courses to follow. Fourteen rules were extracted from the dataset of 1218 students from the High School of Technology of Fez. The FP-growth algorithm showed to be faster than Apriori.

Ahmed *et al.* [4] sought to understand the progress of students' academic performance, degradation of their merit, dropout, and retention in the Computer Science and Engineering department. The study was conducted at the Bangladesh University of Engineering and Technology. The Apriori algorithm was implemented using the WEKA tool to get all this knowledge. The rules showed that most students were male and lived in university dormitories. Also, male students had a higher probability of having a poor Grade Point Average (GPA) as did students who lived on the university campus. In addition, prerequisite courses may influence others; for example, if a student had an excellent grade in the structured programming language course, the same student would also have an excellent grade in the object-oriented programming course. Furthermore, students who did not achieve well tended to struggle with course grades. Besides that, the students who would usually drop out were males and lived on campus. However, the abandonment rate is meagre. The results also showed that good performance depends on the course activities, like quizzes, and the grades of departmental courses affect the final cumulative GPA. Overall, the Apriori algorithm found interesting and useful rules that answered the questions presented in this study.

Abdullah *et al.* [2] employed a study at the University of Malaysia Terengganu where the objective was to identify the uncommon relationships among the programs chosen by students of the computer science program. The algorithms used with the definite factors measure were the LP-Tree and LP-Growth. A total of 4177 rules were extracted, and the results showed that 32% of the students weren't enrolled in any computer science program, and 36% of them applied to four computer science programs. Also, it was seen that if a student chose the Forestry program, the same student would also select the following programs: banking, nursing, art design, pure science, physiotherapy, management, and radiotherapy. In addition, if a student chose the IT program, they would also choose nursing or/and physiology. The nursing program would also be selected when the IT and forest programs were chosen. This proved that students have many different interests when applying to their favourite university programs, and most of them chose the forest program.

Holanda *et al.* [26] intended to understand the lack of interest of female students in computer-related courses at the University of Brasilia. The data was obtained through a questionnaire aimed at female high school students. The algorithm used was the Apriori complemented by a graphical statistic analysis, which extracted a total of 32 rules. The results revealed that family approval had a significant influence on the choice of a major in university. Also, girls believed most students in the Computer Science (CS) major are male. In addition, employability did not appear to be an important factor, although girls were not sure about the computer science job market. However, they know the importance of mathematics in a CS major. Other results showed that girls that like computer games are more likely to pursue a major in computer science.

Yuliansyah *et al.* [54] tried to perceive the relationship between the length of study duration, length of thesis duration, GPA, and English proficiency score. The algorithm used was the Apriori, where eight rules were extracted. The results indicate that when GPA is 3.0, 3.1, and 3.2, the English proficiency score is 400 and vice-versa. Also, when English proficiency is 400, the length of study is 4.1 years and vice-versa. Since the factor length of the thesis duration doesn't appear on the eight rules, this variable proved to be irrelevant.

Mamcenko *et al.* [35] focused on analyzing students' C++ programming language exam data through clustering and association rules techniques. This study was developed at Vilnius Gediminas Technical university. The algorithms used were Kohonen for clustering and Simultaneous Depth-first Expansion for association rules. During clustering, three clusters were obtained. The first one represented the students who pass the first exam, the second one the students who pass on the second exam, and the third one the students who retake the second exam. The results highlighted that the course development policies are the most challenging part to be absorbed by students. The association rules analysis showed that students are more prepared to know the correct answers when they retake an exam. Also, when students spent between 5 to 40 seconds answering a question, the questions were usually right. On the contrary, if students took longer than 180 seconds, the answers were often incorrect.

Silva *et al.* [50] had the objective of discovering which factors influenced the students to remain in their courses longer than expected or to leave the university before the end of the course, using for this purpose the Apriori algorithm. This research was developed at the federal university of Bahia for the second and third semesters, where twenty-four and sixty-nine rules were extracted accordingly. The second semester results demonstrated that when twenty-two students failed on Discrete Math for the second time, all of them flunked this course. Students who do Discrete Math for the second time are not allowed to enrol in the Logic Maths Introduction course or Data Structures. In addition, out of thirty-seven students enrolled in Digital Circuits and Computer Architecture, thirty-two are retained. Of sixteen students approved on Data Structures, thirteen are also approved in Digital Circuits and ComputerArchitectures courses. For the third semester, it was concluded that the eight students who failed in Formal Language and Theory failed the discipline. Also, Discrete Math is a prerequisite course for Formal Language and Theory. Finally, fourteen of the fifteen students took Calculus A for the second time and failed the course.

The summarization of all the studies presented in the literature is displayed in Table 2.1. The Table contains the following information: author, study objective, sample size, software, algorithm, and results.

Table 2.1: Summary of reviewed articles.

Author	Study Objective	Sample Size	Software	Algorithm	Results
Morris et al. (2005)	Explore students' engagement in asynchronous online courses	300000	SPSS	Multiple Linear Pegression (MLR) and T-tests	Completers engaged in online learning activities more often and for a more extended amount of time than unsuccessful and withdrawing students. The most important variables that affect the final grade are the number of discussion posts viewed, the number of content pages viewed, and the seconds on viewing discussion pages.
Gouzouasis et al. (2007)	Study the relationship between participation and performance in music disciplines and performance in other disciplines	130217	Not referenced	Linear Regression and T-tests	Music involvement is correlated with overall higher achievement. 11th Grade of music courses predicted the grade 12 of music courses. The time a student spends in music courses does not diminish their performance in other disciplines; in fact, it even makes them better.
Hung et al. (2008)	Analyze online learning behaviours and predict student performance	17934	Weka, SPSS, and Knime	K-means, Association Rules, and Decision Tree (DT)	Students who participate more actively, i.e., who frequently access course materials, post and read messages, and participate in synchronous discussions, will perform better.
Romero et al. (2010)	Identify unusual students behaviours	230	Not referenced	Apriori-Frequent, Apriori-Infrequent, Apriori-Inverse and Apriori-Rare	The number and the time spent doing assignments and quizzes are directly related to students performance. The time spent on the forum and the number of submitted and read messages on the forum influence the final mark.
Mamcenko et al. (2011)	Explore students' behaviours when taking online exams of C++ Programming language	3167	Not referenced	Kohonen and Simultaneous Depth-first Expansion	Three clusters were formed, representing the first exam pass, the first exam repass, and the second exam retake. Students who repeat the exam are more prepared for the tougher questions. When an answer is correct, it depends on the execution time.
Haverila et al. (2011)	Determine how e-learning experience influences learning outcomes	57	JMP	MLR and PLSR	The variable "Prior E-learning experience" proved to be the most significant contributor to the learning outcomes efficiency, amount, and productivity.
Gamulin et al. (2013)	Understand the relationship between learning and log features on final grades	302	Matlab and PLS toolbox	PCA and PLSR	PLSR outperforms PCA. The outcome variable that produced better results was y2.

Continued on next page

Author	Study Objective	Sample Size	Software	Algorithm	Results
Silva et al. (2013)	Determine which factors and which classes cause students retention or abandonment	135	Not referenced	Apriori	For the second and third semester, the results point out that Discrete Math is a prerequisite course for Formal Language and Theory, Logic Math Introduction and Data Structures courses. It is recommended to teach calculus A for the first semester, being not a prerequisite for any course in the second or third semester.
Abdullah et al. (2014)	Find rare associations among the chosen university programs by students	160	C+	LP-Tree, and LP-Growth with Definite Factors measure	Most of the students chose the forestry program. Students had mixed interests. 32% of the students aren't enrolled in any computer science program, and 36% of the students applied to four computer science programs.
Ahmed et al. (2014)	Get knowledge about students' academic progress, retention, success, and dropout	9210	WEKA	Apriori	Most students are male and live in the university dormitories. The students with tendencies to have poor GPA are male and live on the university campus, being these the ones who usually drop out. The prerequisite courses influence the grades of the other ones. The final grade of a course is influenced by the grades obtained in the course activities. The grades of departmental courses affect the final GPA.
Oliveira et al. (2015)	Identify the most relevant determinants of aggregate demand	Not referenced	Not referenced	PLSR	Portuguese aggregate demand for public higher education is more affected by factors related to the policy and social contexts.
Badr et al. (2016)	Identify the relations between courses and make predictions of students' grades	203	Java	Classification Based on Associations (CBA)	Mathematical courses don't influence the student's performance in the programming courses, contrary to the English courses.
Yoo et al. (2017)	Detect course sequences and find which ones may influence students' dropout likelihood	665	Not referenced	PrefixSpan	Students follow the courses trajectories recommended by the academic advisors, especially female students. Introductory courses in computer programming proved to be the most important ones, followed by fundamental math courses.
Dahdouh et al. (2018)	Understand the relationships between students' activities in the different courses	1218	R	FP-growth	With the construction of this recommender system, courses are advised to students based on other students' enrollments in historical data. FP-growth algorithm showed to be faster than Apriori and produced a set of 14 rules.
Moubayed *et al.* (2018)	Understand the relationship between student engagement with their academic performance	305933	Matlab and WEKA	Apriori	Students with higher levels of engagement tend to achieve better grades in the course.
Yuliansyah et al. (2019)	Identify the relationships among study duration, thesis duration, GPA, and English proficiency score	1437	Python	Apriori	The variables English proficiency, GPA, and length of study duration are strongly related to students' data, contrary to the size of the thesis duration.

Continued on next page

Author	Study Objective	Sample Size	Software	Algorithm	Results
Holanda et al. (2019)	Discover why female students are not inclined to choose a major in computer science	3707	R	Apriori	Girls who like computer games are more likely to apply to a computer science major. Family approval plays an essential role in choosing the major to pursue. Female students believe most students enrolled in CS majors are boys. Employability was not an important factor, and girls weren't sure about the computer science job market.
Wang et al. (2021)	Understand the reasons behind students' contract cheating	509	Not referenced	PLSR	Both personal and institutional reasons significantly influence students cheating intentions.

3 Methodology

The methodology used in this research is based on SEMMA, which was developed by the SAS institute. The use of the SEMMA model allows for the understanding, organization, development, and maintenance of data mining projects (Shafique *et al.* [49]). SEMMA is also driven by a highly iterative cycle.

According to Azevedo *et al.* [8], this methodology can be seen as a practical implementation of the five phases of the KDD process. Figures 3.1 and 3.2 presents the steps to be carried out for each experiment in this study, where the first experiment consists of using the PLSR to see which courses of two specializations of the Master's Degree in Information Management, have the most significant influence on the Dissertation/Work Project/Internship Report. Concerning the second experiment, association rules algorithms were employed in two courses with different assessment methods to understand how online behaviour patterns influence students' final grades

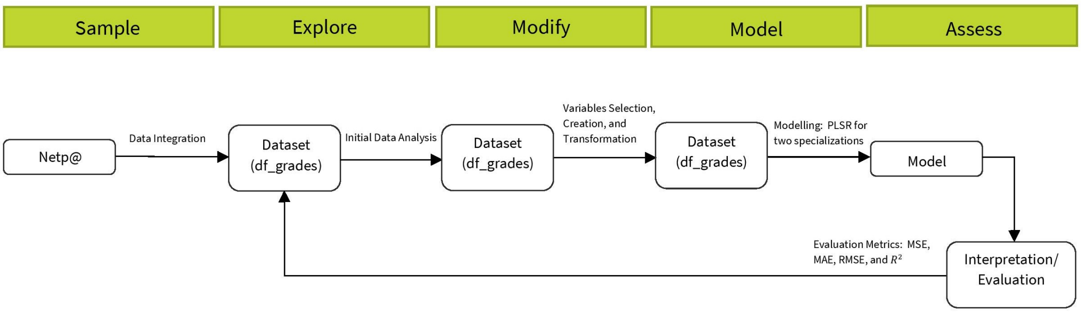

Sample
Explore
Modify
Model
Assess
Netp@
Data Integration
Dataset (df_grades)
Initial Data Analysis
Dataset (df_grades)
Variables Selection, Creation, and Transformation
Dataset (df_grades)
Modelling: PLSR for two specializations
Model
Evaluation Metrics: MSE, MAE, RMSE, and R^2
Interpretation/ Evaluation

Figure 3.1: SEMMA methodology steps for the first experiment.

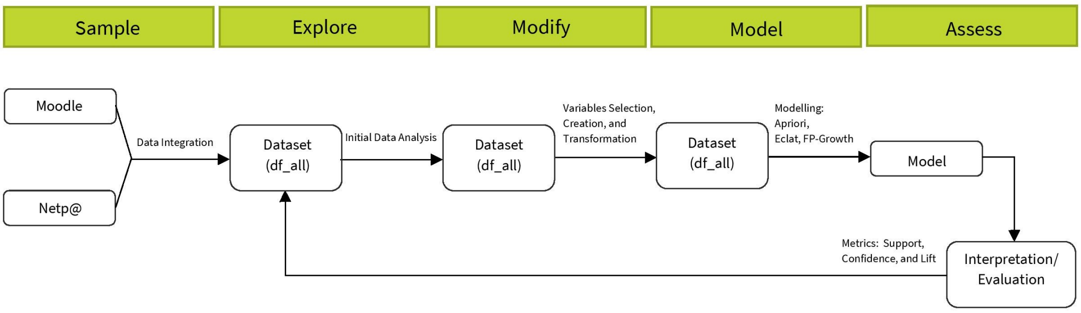

Sample
Explore
Modify
Model
Assess
Moodle
Netp@
Data Integration
Dataset (df_all)
Initial Data Analysis
Dataset (df_all)
Variables Selection, Creation, and Transformation
Dataset (df_all)
Modelling: Apriori, Eclat, FP-Growth
Model
Metrics: Support, Confidence, and Lift
Interpretation/ Evaluation

Figure 3.2: SEMMA methodology steps for the second experiment.

3.1 Data Extraction and Description

The data received from moodle and netp@ systems were anonymized, where from moodle, log data were gathered, and from netp@, the students' learning data.

To obtain a significant set of data to analyze and compare, the records of nine academic years for all the master's degrees were considered, from 2012-2013 to 2020-2021, respectively. However, as Moodle logs only started to be recorded from the 2017-2018 academic year, only this and the subsequent academic years can be used when analyzing the logs data. Also, it's important to mention that since the master's degree comprises two academic years, it is only possible to assess the student's success after completing these years, assuming that the students finish their studies on time.

That being said, nine datasets were provided, each concerning the respective academic year. They came in excel format, with the data obtained by each system separated into sheets. The data acquired are described in Tables 3.2 and 3.1.

Table 3.1: Structure of Learning data.

Field	Description	Example
year_code	Academic year code	201718
master_code	Course code	4281
master_nm_pt	Course name in Portuguese	Mestrado em Estatística e Gestão de Informação
branch_nm	Master's degree specialization name	Análise e Gestão de Informação
course_code	code	200086
course_nm_pt	Course name in Portuguese	Metodologias de Investigação
year	Year	1
semester	Semester	S1
nr_ects	ECTS	7.5
mandatory	Type of course	Yes
userId	Student identifier	1496
evaluationType	Assessment type	Nota Parcial 1
evaluationStatus	Assessment status	Avaliado
evaluationGrade	Evaluation result	10.6
final_code	Whether it's the final grade or not	Yes
finalStatus	Final assessment status	Aprovado
finalGrade	Final course grade	15

Table 3.2: Structure of Log data.

Field	Description	Example
userid	Student identifier	762
masterid	Course code	189
fullname	Full name of the course. Contains the courseid and codDisciplina	201718 - Data Mining I (200026)
courseCode	Course code	200026
eventname	Full action description	\mod_forum\event\course_module_viewed
action	Moodle action. Piece of the eventname	viewed
target	Moodle target. Piece of the eventname	course_module
fileName	course contents	Exploratory Network Visualization
timecreated	Time of the event	02/09/2017 16:18:47

3.2 Data Integration

After receiving the data from the two learning management systems, all the datasets were loaded into the Jupiter notebook using the python programming language. As not all students have interactions in moodle, to avoid losing valuable information about their grades in the courses, two datasets will be used, one that only contains the learning data (df_grades) from Table 3.1 and the other that holds all the data provided (df_all), which is the junction of Tables 3.1 and 3.2. For the first task, which stands for understanding the influence of the courses on the Dissertation/Work Project/Internship Report grades, the learning data of the nine academic years were concatenated together, resulting in a dataset of 143667 records and 17 attributes.

Then, for the second task, which aims at understanding the relationship between moodle activities and the final grade of two courses, for each of the four datasets (2017-2018 to 2020-2021 academic years), it was necessary to merge all the sheets (log and learning data) to have the data altogether. This operation was performed using as keys, i.e., common identifiers, the students' ids, and the courses codes. Consequently, all datasets were concatenated into one, resulting in a total of 18137046 records and 26 attributes.

3.3 Data Exploration

After loading the data, it's important to get familiar with it to gain useful insights and detect possible problems. Therefore, this is a crucial step to take in, as it is where we will understand the meaning and importance of each variable to the study at hand. By performing initial data analysis, it will be possible to discover which records to maintain, which variables should be chosen to address the established goals, and how to modify them appropriately.

First, we need to verify if the variables are well defined; if they are not, we will need to make the necessary modifications to correct them (Exempli gratia (for example) (e.g.), the variable *year* was defined as a floating number when it should be an integer).

Relatively to the missing values presented in the df_all, the first thing that caught our attention was their high number in the *finalGrade* variable, as shown in Table 3.3. The missing data in this feature correspond to dropouts, i.e., to students that, for an unspecified reason, didn't complete the course. For the other variables with incomplete data, we noticed that for all students who did not have information in the final grade, the remaining ones were also missing, except the variable *filename*. As for the df_grades dataframe, the only variables that present missing values are the *finalGrade* and *evaluationGrade*, with 3227 and 1191, respectively. Students who do not have information about the *evaluationGrade* also do not have it in *finalGrade*, which also happens in the other dataframe, as explained.

Table 3.3: Number of missing values of initial variables.

Field	Missing values
userid	0
courseid	764358
fullname	804364
courseCode	804364
eventname	0
action	0
target	0
fileName	18121939
timecreated	0
year_code	1954658
course_code	1954658
course_nm_pt	1954658
branch_nm	1954658
course_code	1954658
course_nm_pt	1954658
year	1954658
semester	1954658
nr_ects	1954658
mandatory	1954658
userId	1954658
evaluationType	1954658
evaluationStatus	1954658
evaluationGrade	1954658
final_code	1954658
finalStatus	1954658
finalGrade	2259222

At the beginning of the analysis, the only interval variables are *finalGrade* and *evaluationGrade*, where their descriptive statistics are displayed in Table 3.4. When looking at it, we can observe that both variables present an average of 15, which indicates a certain consistency between the students' grades and shows that overall they tend to perform well. When looking at the minimum value of *finalGrade*, it seems suspicious since it's improbable for a student to have zero as the final grade of a course. On the contrary, it's already possible that a student obtains zero as a partial grade, as it's enough that the pupil doesn't deliver a project or a quiz for this to happen.

Table 3.4: Descriptive statistics of initial interval variables.

	finalGrade	evaluationGrade
mean	15	15
std	4	4
min	0	0
25%	14	13
50%	16	15
75%	17	17
max	20	20

Then, some characteristics of the data were examined using data visualization techniques. For that end, the hour, day, and month were extracted from the variable *timecreated*.

From the analysis of Figure 3.3, it is possible to observe that students tend to be more active in the afternoon, especially at 6 pm, when the highest activity peak occurs.

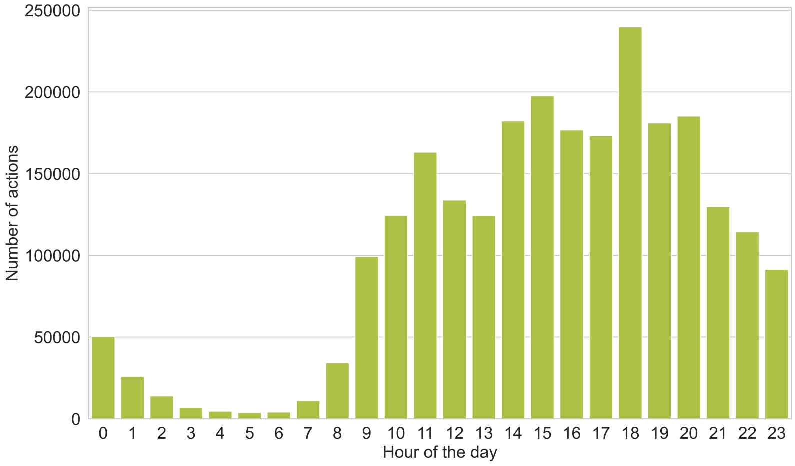


Figure 3.3: Number of actions by the hour of the day.

The second barplot, Figure 3.4, shows that October is the month with more moodle platform interactions, followed by November. A plausible reason for this to happen is because it is a period when students start working on projects and/or studying for quizzes, for example.

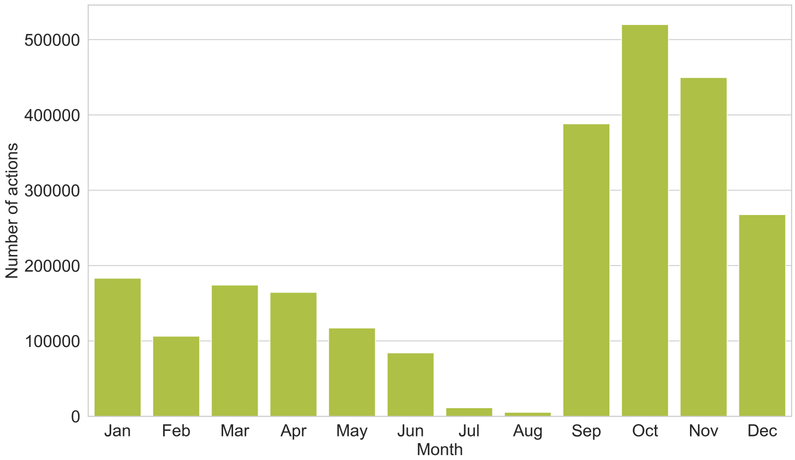


Figure 3.4: Number of actions by month.

In relation to Figure 3.5, it was already expected to find more students attending the Normal assessment period. From then on, there is an exponential decrease in the number of students who attend the other types of assessment, which suggests that most students get a sufficient grade to pass the courses in the first assessment phase.

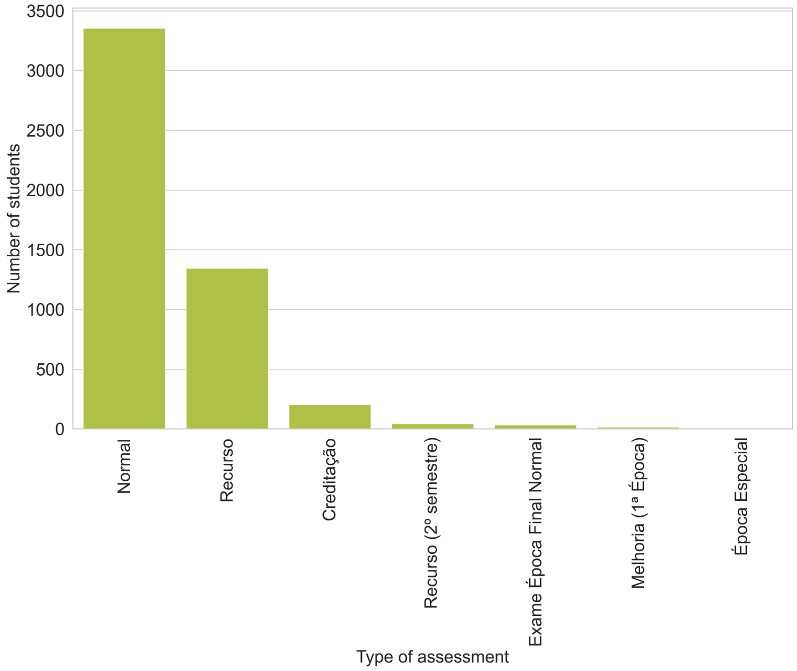


Figure 3.5: Number of students by type of assessment performed.

These conclusions are supported by Figure 3.6 which proves that students obtain their final grade in the normal evaluation.

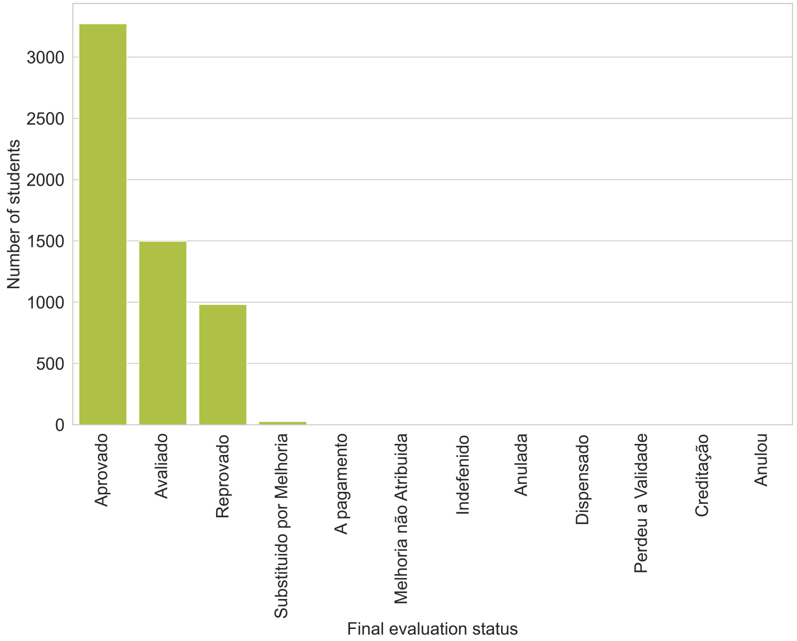


Figure 3.6: Number of students by evaluation outcome.

The displayed barplots, Figures 3.7 and 3.8, reveals the most and least common actions performed by students. That said, we can see that course_viewed is the predominant action, followed by course_module_viewed. This is natural since students usually access the pages of the courses as well as their materials frequently.

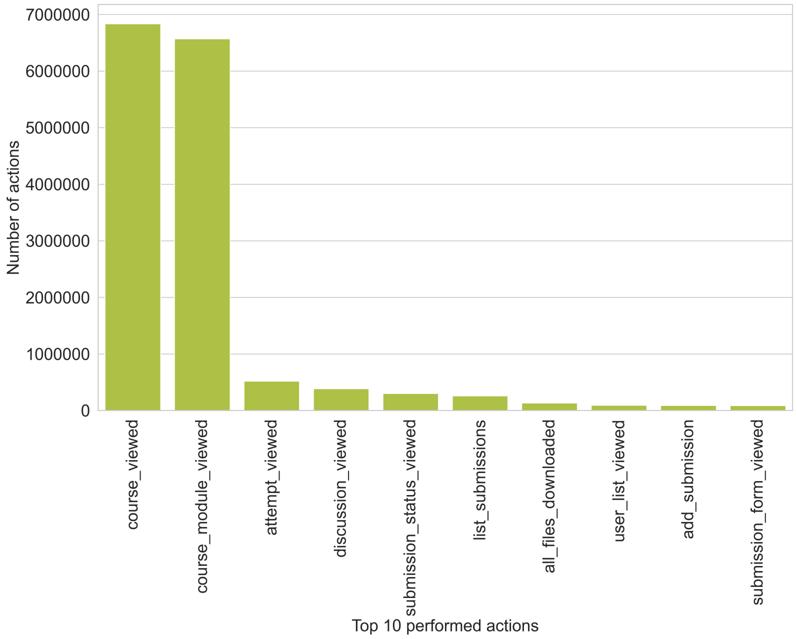


Figure 3.7: Most performed actions.

On the other hand, the less representative action is the subscription_deleted, which also makes sense as students don't usually interact much in discussion forums. These visualizations required the creation of an additional variable called *target_action*, which is the combination of the variable *target* with the variable *action*, and which will serve as an auxiliary variable in the construction of the new features, that will be explained in the next section.

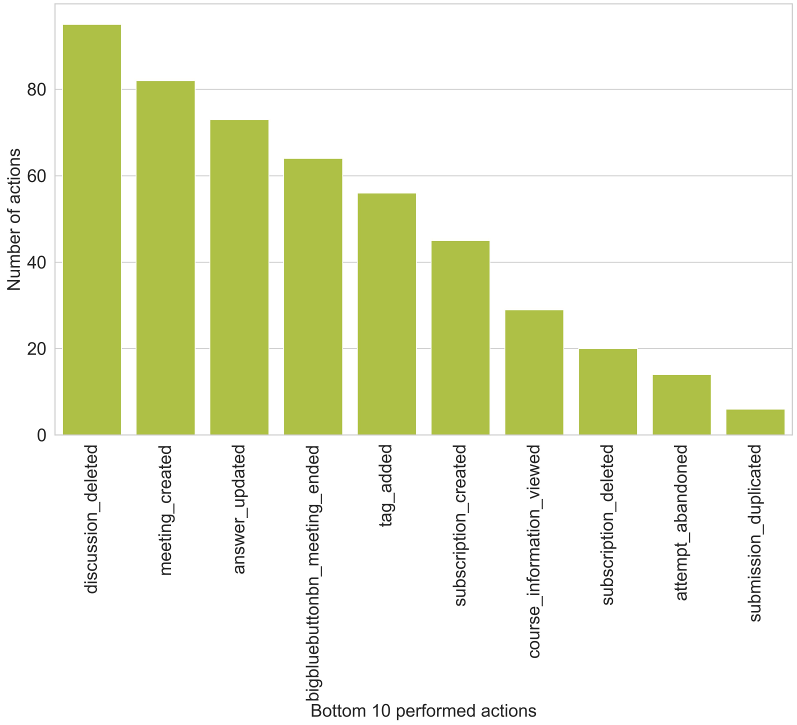


Figure 3.8: Least performed actions

From the interpretation of Figure 3.9, we can conclude that the master's degree with more enrollments in the nine academic years is always the Master's Degree in Information Management. It is also observable that more master's degrees were introduced in the last two years.

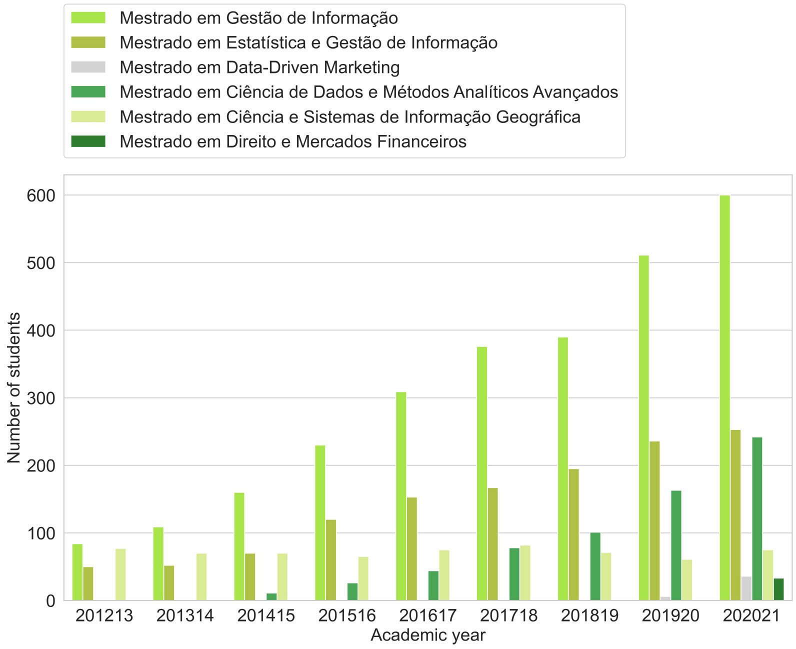


Figure 3.9: Number of students per academic year.

When analyzing Figure 3.10, it is clear that the master's degrees average grades tend to be good and quite similar to each other, usually between 14 and 16 values, being the Master in Data Science and Advanced Analysis the one that obtains better results in more academic years.

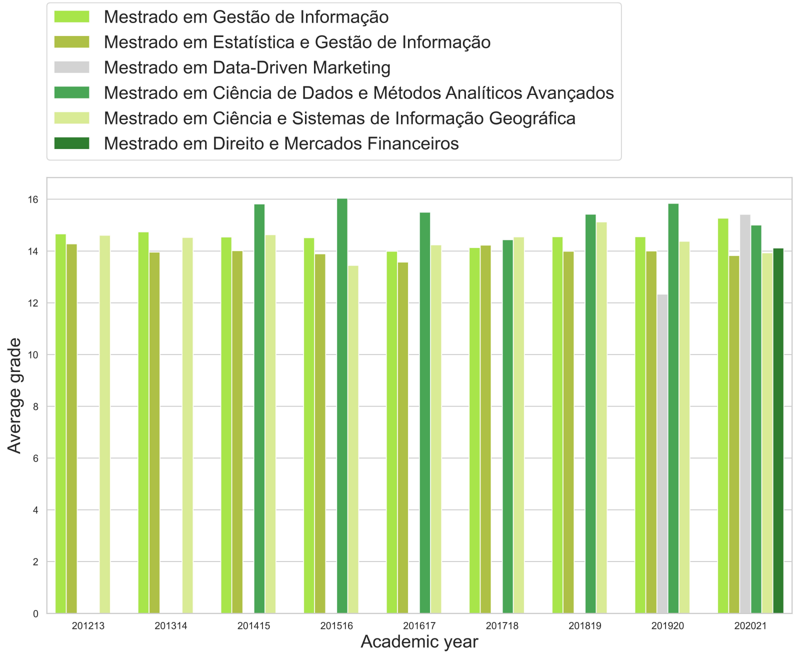


Figure 3.10: Master's degrees average grade by academic year.

It is important to mention that for the creation of the Figure 3.10 it was necessary to create an auxiliary variable, referring to the student's average grade (weighted average), which follows the formula:

$$\text{Average grade } = \frac{\sum \text{nr\_ects} \times \text{ finalGrade}}{\sum \text{nr\_ects}}$$

3.4 Data Preparation

Data Preparation is a crucial step in arranging the data for further analysis. In this step, we perform record and feature selection, feature transformation and creation, and data cleaning to prepare the information to ease the model implementation and enhance its results.

It's important to clarify from the beginning that since the dataset df_all contains the df_grades, any changes made on the initial variables that belong to the learning data will be done on both datasets, except for the *finalGrade* feature, as the df_all dataset will use the values of *finalGrade* in its original form, which will not happen in the df_grades dataframe since the datasets will be used to fulfil different objectives. Furthermore, different variables will be created for each dataset according to their purposes.

3.4.1 Data Construction

As the name suggests, this task refers to creating new features, either by creating derived attributes or simply creating new ones based on existing variables, to retrieve as much information as possible from the provided datasets.

That said, all explanations about the meaning and construction of these additional attributes in the df_all dataset are displayed in Table 3.5.

Table 3.5: New interval features.

New variable	Description	Construction
dedication_time	Time (seconds) that a student spent in a session	From the variable *timecreated* and *action*, for each student get the time between logins, *loggedin* category. After that, sum up all the session times. A maximum session length of 2h has been set as established to NOVA IMS configuration.
inactive_time	Time (seconds) elapsed between sessions	From the variables *timecreated* and *action*, for each student obtain the time elapsed between a login and its previous action. Then add up those durations.
messages_sent	Number of times the student sent chat messages in moodle	From the variable *action*, get the frequency of the category *sent* using the column *userid* as the index.
log_freq	Number of times the student logged in to moodle	From the variable *action*, get the frequency of the category *loggedin* using the column *userid* as the index.
quizes_done	Number of quizzes taken by a student	From the variable *target_action*, get the frequency of the category *attempt_submitted* using the column *userid* as the index.
		Continued on next page

New variable	Description	Construction
submissions	Number of times the student submitted course materials	From the variable *target_action*, get the frequency of all the categories related to students submissions, being them: *add_submission*, *answer_submitted*, *response_submitted*, and *submission_created*, using the column *userid* as the index.
discussion_forum_posts	Number of times the student posts on the discussion forum	From the variable *target_action*, get the frequency of all the categories related to students posts, being them: *post_created*, and *discussion_created*, using the column *userid* as the index.
downloads	Number of times the student downloads course materials	From the variable *target_action*, get the frequency of the category *all_files_downloaded* using the column *userid* as the index.
contents_updated	Number of times the student made updates	From the variable *target_action*, get the frequency of all categories related to student updates, being them: *submission_updated*, *post_updated*, *answer_updated*, and *choice_updated*, using the column *userid* as the index.
web_links_viewed	Number of times the student accessed external URLs (e.g. mp4 videos) and HTML pages	From the variable *target_action*, get the frequency of all categories related to the links accessed by students, being them: *course_module_viewed*, for both *mod_url* and *mod_page* components, and *content_page_viewed*, using the column *userid* as the index.
discussion_forum_views	Number of times the student accessed forum discussions	From the variable *target_action*, get the frequency of *discussion_viewed* category, using the column *userid* as the index.
quizes_reviwed	Number of times the student reviewed his/her quiz solution before submitting it	From the variable *target_action*, get the frequency of *attempt_reviewed* category, using the column *userid* as the index.
files_viewed	Number of times the student read supplementary materials and additional learning resources (pdf, ppt, word, and excel files)	From the variable *target_action*, get the frequency of all categories related to the files accessed by students, being them: *course_module_viewed*, for both *mod_resource* and *mod_book* components, and *chapter_viewed*, using the column *userid* as the index.

From the Table displayed above, as well as the reviewed articles, we can see that all the variables belong to activity engagement indicators, including counting-based and duration-based features derived from the trace data (Yang *et al.* [52], Cerezo *et al.* [13], Hu *et al.* [27], Kovanovic *et al.* [31], Macfadyen *et al.* [34], and Conijn *et al.* [16]).

As for the df_grades dataset, from the variables *course_nm_pt* and *finalGrade*, we extract a new set of features that aggregate these two, where the columns correspond to the courses, and their values to the final grades, using the *userId* as the index. The names of these new variables correspond to the values of *course_nm_pt*, i.e., to the courses' names.

3.4.1.1 Exploration of Constructed Data

Beyond the initial data analysis described in 3.3, and for the same reasons mentioned earlier, it is always relevant to repeat this procedure for new data.

The missing values in the Table 3.6 mean that the student did not perform the action in question. From its analysis, we can infer that the action that students do the least is to post in the discussion forum.

Table 3.6: Number of missing values of the constructed variables.

Field	Missing values
dedication_time	0
inactive_time	0
messages_sent	5234952
log_freq	0
quizes_done	10116823
quizes_reviewed	11269329
submissions	783620
discussion_forum_posts	11275191
downloads	2756466
contents_updated	2294821
web_links_viewed	323707
discussion_posts_views	592209
files_viewed	5432

Now considering the main statistics of the created attributes exhibited in Table 3.7, we can extract valuable knowledge about students' interactivity in moodle platform. From the analysis, we can see that the maximum values in *quizes_done* and *quizes_reviewed* do not make much sense as they are out of the ordinary, which could have been a system error when collecting the actions taken by students, because even if a student fails in multiple courses more than one time, where it is necessary to carry out quizzes regularly, this value would still be too high to be considered plausible. Additionally, details on the variables' distributions can be consulted in Appendix B.

Table 3.7: Descriptive statistics of constructed interval variables.

	mean	std	min	25%	50%	75%	max
dedication_time	12725830	13667540	4441	4978486	10313300	16140136	170378197
inactive_time	8505248	7747943	37	3582428	7412742	11303107	64334543
messages_sent	22	83	0	0	1	6	2174
log_freq	277	274	1	110	228	359	2604
quizes_done	3	12	0	0	0	1	203
quizes_reviewed	4	20	0	0	0	0	400
submissions	8	8	0	2	6	11	54
discussion_forum_posts	36	1	0	0	0	0	14
downloads	8	15	0	0	2	10	234
contents_updated	2	2	0	0	2	4	24
web_links_viewed	41	44	0	8	29	59	301
discussion_posts_views	24	31	0	3	15	36	585
files_viewed	277	254	0	90	210	404	2741

Finally, to identify the existence of associations between courses of the Master's Degree Program in Information Management, Pearson correlations were calculated, and pair plots were obtained for the two Specializations, which are displayed in Figures 3.11, and 3.12. The Pearson's correlation coefficient checks if a linear correlation between variables exists and quantifies it. The correlation matrices help us measure the direction (Positive/Negative) and the intensity of the variables' interrelationship (Low/Medium/High).

When looking at Pearson's correlations for both specializations, we observe that all the courses are correlated with each other, indicating the presence of multicollinearity.

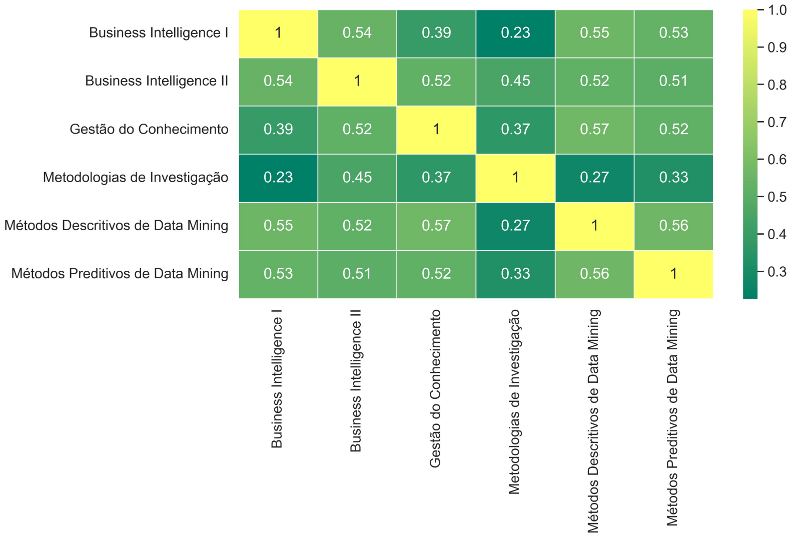


Figure 3.11: Pearson Correlation Matrix of courses of Business Intelligence Specialization.

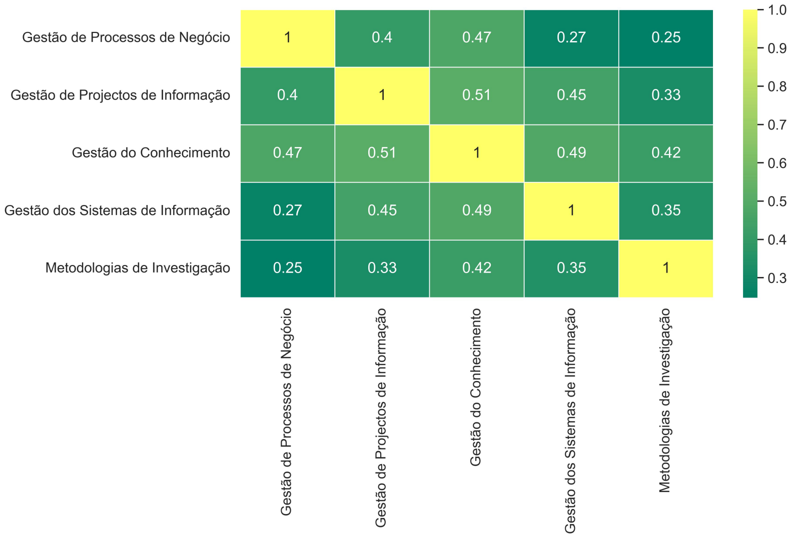


Figure 3.12: Pearson Correlation Matrix of courses of Information Systems and Technologies Management Specialization.

After that, we will also analyze Pearson's correlations between the predictors with the target variable for the two specializations, which are presented in Figures 3.13, and 3.14.

Regarding the Business Intelligence specialization, the variable *Gestão do Conhecimento* obtained the highest Pearson correlation coefficient, with a value of 0.51. So it can be said that there is a moderate positive relationship between this variable with *Dissertação/Trabalho de Projecto/Relatório de Estágio*.

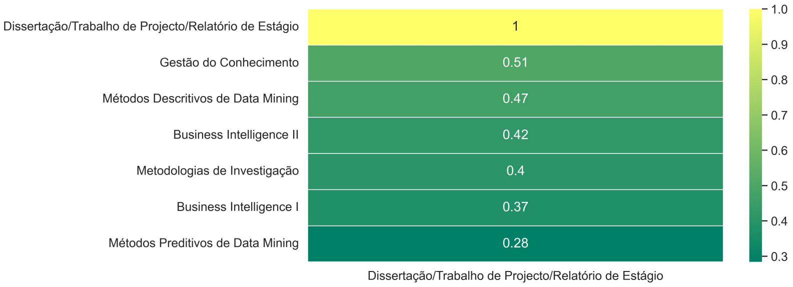


Figure 3.13: Pearson Correlation Matrix of courses with target variable of Business Intelligence Specialization.

For the specialization in Management of Information Systems and Technologies, the course with the highest Pearson correlation coefficient is *Metodologias de Investigação* with a value of 0.55, respectively. Therefore, this variable also has a moderate positive relationship with the variable *Dissertação/Trabalho de Projecto/Relatório de Estágio*.

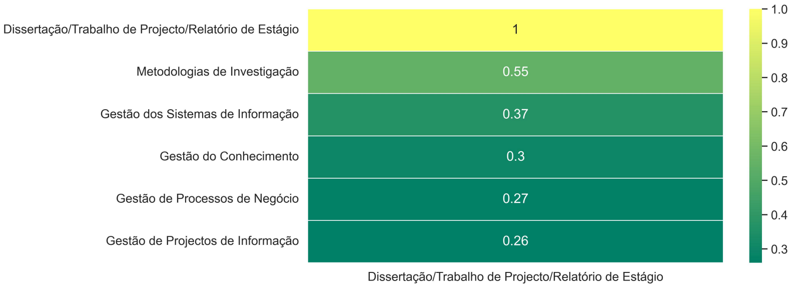


Figure 3.14: Pearson Correlation Matrix of courses with target variable of Information Systems and Technologies Management Specialization.

In addition, Appendix A explores the pairwise relationships between the selected courses of each specialization with the Dissertation/ Work Project/ Internship Report.

3.4.2 Data Transformation

Data transformation is the process of changing the format, structure, or values of data.

The same data can be arranged in different ways, without losing information or its content, like in the case of students' grades, either by expressing them in a binary form (pass/fail), by levels (insufficient to excellent), or numerically (Minaei-Bidgoli *et al.* [37], Cortez *et al.* [17]).

So, in the first instance, for the df_grades dataframe, the values of the variable *finalGrade* will be grouped according to the Portuguese Decree-Law, where each grade will be assigned a qualitative mark by degree, made as follows: Insufficient (0 to 9), Sufficient (10 to 13), Good (14 to 15), Very Good (16 to 17) and Excellent (18 to 20) [46].

In addition, for the df_all dataframe, the values of the created variables defined in Table 3.5 will be discretized into bins using the Inter Quartile Range (IQR) method (Islam *et al.* [29]).

This technique is called data binning, discrete binning, or bucketing, and it was used as it can improve the quality of the models by strengthening the relationship between attributes. Apart from that, it can also increase the interpretation and comprehensibility of the results (Dougherty *et al.* [20]).

Subsequently, since some master's degrees and courses changed their names over the years, we decided to replace the oldest terms of the variables *ds_discip_pt* and *nm_curso_pt* with the most recent ones.

3.4.3 Data Cleaning

Data cleaning is the process of modifying and correcting data. This task aims to make the data uniform and prepared for analysis. This includes handling missing values, resolving inconsistencies, and detecting and removing redundant data.

As already stated in the previous Section, Table 3.3, there are a lot of missing values in the variable *finalGrade*. Since we are only interested in the courses with grades associated, these records can be deleted. After performing this operation, the missing data observed in the other columns disappeared, except in the *fileName*. Still, as this feature does not provide any additional relevant information for the analysis, it can also be removed.

After the creation of more consistent and complete variables, the ones used to build them can be discarded, as they are no longer needed, such as the variables *action*, *target*, *timecreated* and *target_action*. The latter's purpose was only to assist in the creation of the others, thus having no additional utility.

It is also important to point out that some of these new features have missing values, like *discussion_forum_posts*, *discussion_forum_views*, *quizes_done*, *quizes_reviwed*, *files_viewed*, *submissions*, *messages_sent*, *web_links_viewed*, *contents_updated*, and *downloads*, respectively. The missing data in this specific case means that the student did not perform any of these actions, so they will be replaced by the value zero. Additionally, as the variables *quizes_done*, and *quizes_reviwed* showed absurd maximum values, we defined a reasonable limit of quizzes taken and reviewed to be 60 and 120, respectively.

Besides that, some variables share the same information like the keys of the Tables and the variable *fullname*, that embodies the variables *courseid*, *courseCode* and *course_nm_pt*. Moreover, features related to codes and ids have no use. Thereby all of these variables can also be dropped, except one of the student's identifiers that we will use as an index.

Another aspect to consider is that for each course, there are several final grades for the same students (depending on the evaluation period); for example, a student can pass the exam in the first period and go to the second period to improve their grade, ending up with a higher grade, in these cases two final grades are recorded. In this way, only the highest grade was maintained to have only one final grade in each course.

The observations of students who obtained a final grade of zero will be ignored from the analysis due to their lack of meaning and possible transcription errors.

3.5 Data Selection

This section will specify which data and variables are needed to carry out each of the proposed goals.

For the first task of this dissertation, it's necessary to properly stratify the data for modelling purposes. For this, the dataset was divided into two subsets, with 70% of the data being used to train the model and the remaining 30% to test and evaluate it. Furthermore, only the records corresponding to the Information Management Master's will be considered since it is the master's degree with the highest number of observations registered, as seen in Figure 3.9. As the specialization in Marketing Intelligence has little data available, it was not considered for analysis. Therefore, we will focus on only two specializations of this master's degree. Regarding the needed variables, only the student's identifier and their grades in each mandatory course will be used. It is important to mention that, as some of these disciplines have few records or were introduced in the last academic years under study, they will not be part of this analysis. This way, we end up considering only the mandatory courses common to all or almost all academic years, as detailed in Table 3.8.

Table 3.8: Courses for each specialization.

Knowledge Management and Business Intelligence	**Information Systems and Technologies Management**
• Dissertação/Trabalho de Projecto/Relatório de Estágio • Gestão do Conhecimento • Métodos Descritivos de Data Mining • Métodos Preditivos de Data Mining • Business Intelligence II • Business Intelligence I • Metodologias de Investigação	• Dissertação/Trabalho de Projecto/Relatório de Estágio • Gestão de Processos de Negócio • Gestão de Projectos de Informação • Gestão dos Sistemas de Informação • Metodologias de Investigação • Gestão do Conhecimento

For the second task, we will only take into account the grades of two courses, *Programação para a Ciência de Dados* and *Aprendizagem Profunda*, namely. The first corresponds to a course where students have many interactions in moodle (e.g. lots of quizzes and homework) and the other with very few interactions in the system, which only serves as a repository, i.e., contains only lectures. Additionally to the student's grades in these two courses, which will be seen as target variables, only the constructed variables presented in Table 3.5 will be necessary for future analysis. Regarding the course *Aprendizagem Profunda*, as taking quizzes is not part of the disciplinary assessment methods, the variables quizes_done and quizes_reviewed will be omitted since they would always take the value zero.

3.6 Modelling

This section will explain the most relevant theoretical notions behind the learning algorithms employed in this study.

3.6.1 Regression

Regression analysis is a widely used statistical technique to make inferences about the relationships between a dependent variable and one or more independent variables (Montgomery *et al.* [38]). This type of analysis helps to understand the explanatory variables' influence on the response variable and make predictions. In short, it shows how the dependent (target) variable changes when varying one of the independent (predictor) variables, keeping the others constants.

3.6.1.1 Partial Least Squares Regression

PLSR or Projection to Latent Structures is a combination between multiple linear regression and principal component analysis that serves as a dimensionality reduction technique used to analyze or predict a set of response variables from a set of explanatory variables (Abdi *et al.* [1]). This prediction is performed by grouping the correlated independent features into sets of orthogonal factors called latent variables, where the coefficients (loadings) are determined in order to maximize the covariance between these new variables and the dependent variable, maximizing this way the predictive ability.

To sum up, the objective of this model is to describe the relationship between one or more response variables and predictors through the latent variables.

Furthermore, this technique is commonly used when there are many correlated independent variables or missing values (Pirouz *et al.* [45]). That said, the reason for choosing this model over Principal Components Regression and Multiple Linear Regression is that in the first, the components are created without taking the dependent variable into consideration and in the latter multicollinearity cannot be dealt with.

As with any regression model, some assumptions must always be met; being them:

1. Independence: The residuals are independent of each other;

2. Homoscedasticity: The residuals have constant variance;

3. Zero Conditional Mean: The residuals are always centred.

For the first assumption mentioned, the Durbin-Watson test is the most commonly used. The premise is not satisfied if its score is less than 1.5 (positive autocorrelation) or greater than 2.5 (negative autocorrelation). Otherwise, if its score is between 1.5 and 2.5, there is no autocorrelation, and the assumption is satisfied (King *et al.* [30]).

In order to identify and select the most important features, the measures used were the VIP scores and the regression coefficients (β) obtained by the PLSR (Chong *et al.* [14]). Both belong to filter methods, where the primary purpose is variable identification.

The VIP score measures the explicative power of the independent variables in relation to the dependent variable.

The value of the VIP score, which is greater or equal to 1, is the typical cutoff point for selecting relevant variables (Akarachantachote *et al.* [5]). Thus, variables that satisfy this criterion must be selected. In addition, from the vector of the regression coefficients, which is a measure of association between each explanatory variable and the response, the ones with low absolute values can be ignored from the model (Mehmood *et al.* [36]).

The VIP measure v_j is defined as follows:

$$v_j = \sqrt{P \sum_{a=1}^{A} SS_a \left(w_{\mathrm{aj}}/\|w_a\|\right)^2 / \mathrm{SST}}$$

Where P is the total number of variables, A the total number of components, SS_a the sum of squares explained by the a^{th} component, SS_t the total variance explained by all the components and $\frac{w_a}{\|w_a\|^2}$ represents the importance of the j^{th} variable. Hence, the percentage of the variance explained by each PLS component is reflected in the weight of each v_j, which represents the contribution of every single variable.

3.6.2 Association Rules Mining

In Data Mining, association rules are a type of rule-based machine learning algorithm that can be described as a way of finding frequent patterns in large sets of data items (García *et al.* [22]). Essentially, these rules provide information about things that frequently occur together, being a classic example the Market Basket Analysis. The rules are expressed in the form of X → Y, where the rule is interpreted as If X Then Y (i.e., based on if/then statements), being X the antecedent and Y the consequent. In a transaction, X and Y are considered independent item sets.

3.6.2.1 Apriori Algorithm

Apriori is a level-wise algorithm introduced by R. Agrawal and R. Srikant in 1994 that proceeds by mining frequent sets of items for boolean association rules. The algorithm's name comes from the fact that it uses prior knowledge of the properties of the frequent itemsets.

The biggest bottleneck of the algorithm is that it needs to scan the database repeatedly, generating a huge number of candidate sets, which implies a high cost of time and memory.

The algorithm can be divided into two main steps:

1. Candidate generation: Generate (K+1) itemset by joining each item with oneself;

2. Candidate Pruning: Count the frequency of each item in the database, and in case the items don't meet the requirements of minimum support, they will be removed since they are considered infrequent.

3.6.2.2 Frequent Pattern Growth Algorithm

FP-Growth Algorithm (or frequent-pattern growth) is an enhancement of the Apriori as it compresses the dataset, being used to detect frequent itemsets in a database without candidate itemset generation. Although this algorithm is much faster than Apriori, it has limitations as the FP-tree may not fit in memory and is difficult to build due to its complex data structure.

This algorithm consists of two steps:

1. Construct an FP-Tree;

2. From the FP-Tree extracts the frequent itemsets recursively.

3.6.2.3 ECLAT Algorithm

Eclat Algorithm stands for Equivalence Class Clustering and Bottom-up Lattice Transversal, which uses prefix-based equivalence relation along with a depth-first search for discovering frequent elements. Eclat only needs to scan the dataset once. Compared to Apriori and FP-Growth algorithms, this algorithm is much faster as it only uses vertical datasets, therefore more scalable and efficient. One of its downsides is that, due to how the algorithm runs, it lacks other interesting measures such as lift and confidence.

The two major steps of this algorithm are:

1. Invert the data structure → item: TID_set.

2. Obtain (K+1) itemsets by finding the intersection of frequent K itemsets.

In short, the comparison of the three algorithms is explained in the Table 3.9.

Table 3.9: Comparative analysis between algorithms.

	Apriori	**FP-Growth**	**Eclat**
Storage Structure	Array-based	Tree based	Array-based
Data Format	Horizontal	Horizontal	Vertical
Search Type	Breadth first	Depth-first	Depth-first
Technique	Join and prune	Divide and conquer	Intersection based approach
Speed	The slowest	Faster than Apriori	The faster
Memory	Requires more memory space	Requires less memory space than Apriori	Requires less memory space
Candidates	Use self-joining for candidate generation	No candidate generation	Use intersection of transaction ids for candidate generation
Frequent Patterns	Candidates whose support is higher than minimum support	Mining conditional FP Trees	Transactions whose support is higher than the minimum support
Scans	Scan the database repeatedly	Only requires two scans	Only requires one scan

3.6.3 Evaluation measures

When assessing model performance within the data analysis pipeline, choosing a good evaluation metric is fundamental. The quality of this metric will reflect how good the generalization of the model results is to real-world data.

3.6.3.1 Regression

The most widely used regression metrics for evaluating model performance are the Mean Squared Error (MSE), Mean Absolute Error (MAE), RMSE, and Coefficient of Determination. The first three focus on measuring the difference between predicted and observed values, while the last one allows us to measure the accuracy of our model in predicting the dependent variable. The closer the coefficient of determination is to one, and are to zero, the better the model will be.

Mean Absolute Error It represents the average of the absolute differences between the model's actual and predicted values. Its formula is as follows:

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} \left| y_i - y'i \right|}{n} \quad \text{,where n is the number of instances.}$$

Mean Squared Error It represents the average of the squared differences between the model's actual and predicted values. Its formula is as follows:

$$\mathrm{MSE} = \frac{\sum_{i=1}^{n} (y_i - y'i)^2}{n} \quad \text{,where n is the number of instances.}$$

Root Mean Squared Error It represents the average root-squared difference between the model's actual and predicted values. Its formula is as follows:

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \quad \text{,where n is the number of instances.}$$

Coefficient of Determination It's a statistical measure that indicates the model's strength, i.e., how well the data fits the model (the goodness of fit). It shows the proportion of variance in the dependent variable that can be explained by the independent variable (Rencher *et al.* [47]). It can be written as:

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \quad \text{,where n is the number of instances.}$$

3.6.3.2 Association Rules Mining

Knowing that the main obstacles to using association rules are the discovery of a considerable number of rules as well as obtaining non-interesting or poorly understandable ones, it is important to define the threshold values of the parameters (minimum support and minimum confidence) wisely (Garcia *et al.* [23], Moreno *et al.* [39], Wang *et al.* [51]).

In this way, the most commonly used measures to understand the strength and interestingness of an association are support, confidence, and lift. The criteria used to compare and select the most appropriate association rules were the followings: the minimum value of support defined was 0.07, i.e., the rule needs to have occurred at least 7% of the time to be considered, and the level of confidence was at least 75%, i.e., we would like to have more than 75% confidence that the rule applies. Nevertheless, it is important to emphasize that a credible rule must have a good confidence factor, a high level of support, and a lift higher than 1.

Support The support of a rule shows how frequently an itemset appears in the database (a measure of popularity). In other words, support is the proportion of transactions in which both X and Y occur. The support formula, introduced by Agrawal *et al.* [3], is given as follows:

$$\operatorname{supp}(X) = \frac{|\{t \in T; X \subseteq t\}|}{|T|}$$

Confidence The confidence of a rule indicates the strength of an association, being seen as a conditional probability of the rule. Hence, it shows how often the rule has been found to be true. Its formula, introduced by Agrawal *et al.* [3], is given as follows:

$$\operatorname{conf}(X \Rightarrow Y) = \frac{\operatorname{supp}(X \cup Y)}{\operatorname{supp}(Y)}$$

Lift The lift of a rule expresses how interesting the rule is, i.e., it measures the importance of a rule. A lift value larger (less) than 1 indicates a positive (negative) dependence or complementary (substitution) effect. Its formula, introduced by Brin *et al.* [12], is given as follows:

$$\operatorname{lift}(X \Rightarrow Y) = \frac{P(Y \mid X)}{P(Y)} = \frac{P(X \wedge Y)}{P(X)P(Y)}$$

4

Results and discussion

This section presents the results of the learning algorithms discussed in 3.6.1 and 3.6.2, which will be assessed according to the evaluation metrics referenced in 3.6.3.

4.1 Relationships between courses

First, it's important to highlight that the optimal number of PLSR components to be retained in each model was determined through the lowest RMSE value (Olav *et al.* [32], Nocita *et al.* [42]).

4.1.1 Specialization in Business Intelligence

Of the selected courses, a total of 122 students had grades recorded in all of them.

From Figure 4.1, which shows the RMSE value for each component, we can conclude that we must retain only one principal component, as it is where a smaller RMSE value is reached.

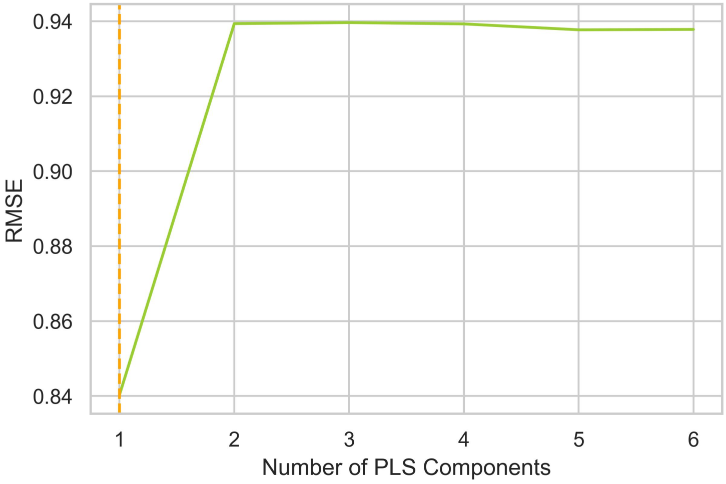


Figure 4.1: Number of components to retain.

Looking at Table 4.1, we notice that the important variables to be kept in the model are *Business Intelligence II*, *Gestão do Conhecimento*, *Metodologias de Investigação* and *Métodos Descritivos de Data Mining*, as they have VIP scores greater than one and the highest regression coefficient values.

Table 4.1: VIP scores and regression coefficients for Business Intelligence Specialization.

Course	VIP score	β
Business Intelligence I	0.79	0.21
Business Intelligence II	1.02	0.27
Gestão do Conhecimento	1.27	0.33
Metodologias de Investigação	1.08	0.28
Métodos Descritivos de Data Mining	1.12	0.29
Métodos Preditivos de Data Mining	0.56	0.15

From Table 4.2, it can be said that 57% of the variability of the grades in Dissertation/Work Project/Internship Report is explained by the courses *Business Intelligence II*, *Gestão do Conhecimento*, *Metodologias de Investigação* and *Métodos Descritivos de Data Mining*.

Table 4.2: Evaluation metrics for Business Intelligence Specialization.

MSE	RMSE	MAE	R-Squared
0.69	0.79	0.66	0.57

By analyzing the residual scatterplot, Figure 4.2, the points are randomly dispersed with very few fluctuations, i.e., without any pattern around the axis y = 0, so due to this lack of pattern and structure, it is concluded that the errors are independent and show constant variance. The Durbin-Watson test also supports this statement since our model got a score of about 1.89, which is between 1.5 and 2.5, so there is no autocorrelation in our residuals. Finally, since the average of the residuals is about 0.04, we can assume they are centred.

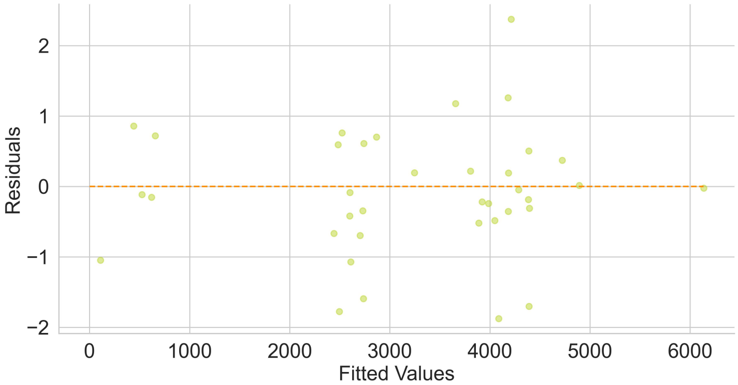


Figure 4.2: Scatterplot to test the homoscedasticity of the residuals.

Therefore, the model does not violate any of the required assumptions, so we can assume that our model can perform well to predict future grades of Dissertation/Work Project/Internship report by using the four independent variables, *Business Intelligence II*, *Gestão do Conhecimento*, *Metodologias de Investigação* and *Métodos Descritivos de Data Mining*. However, since our model has an R-squared score of 57%, there are still about 43% unknown factors affecting the grades of Dissertation/Work Project/Internship Report.

4.1.2 Specialization in Information Systems and Technologies Management

A total of 86 students had grades recorded in the selected courses for the academic years under study.

Figure 4.3 displays the RMSE value for each component, and from it, we can infer that one principal component must be selected as this is where the minimum RMSE value is reached.

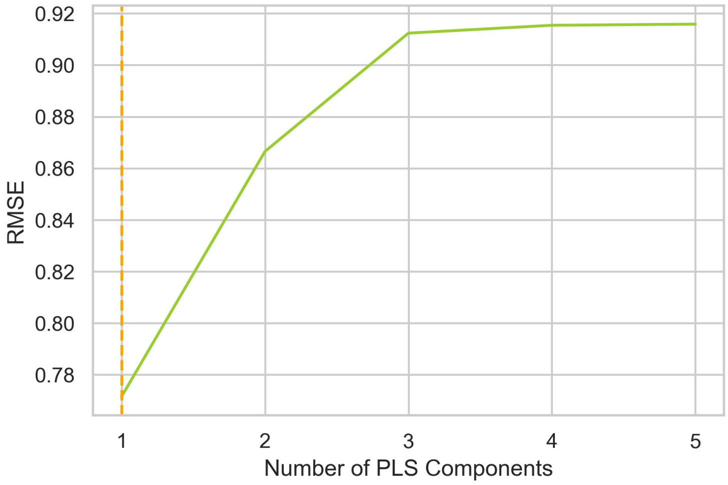


Figure 4.3: Number of components to retain.

When observing Table 4.3, we see that the only significant variable to be maintained in the model are *Gestão dos Sistemas de Informação* and *Metodologias de Investigação* since they are the only ones with a VIP score greater than one. Apart from that, they are the ones with the highest regression coefficient values.

Table 4.3: VIP scores and regression coefficients for Information Systems and Technologies Management Specialization.

Course	VIP score	β
Gestão de Processos de Negócio I	0.62	0.20
Gestão de Projectos de Informação II	0.78	0.25
Gestão do Conhecimento	0.68	0.22
Gestão dos Sistemas de Informação	1.10	0.35
Metodologias de Investigação	1.53	0.48

Concerning Table 4.4 it can be said that 53% of the variability in Dissertation/Work Project/Internship Report was accounted for in the courses *Gestão dos Sistemas de Informação* and *Metodologias de Investigação*.

Table 4.4: Evaluation metrics for Information Systems and Technologies Management Specialization.

MSE	RMSE	MAE	R-Squared
0.65	0.80	0.60	0.53

Through the scatterplot presented in Figure 4.4, our residuals seem to have a constant and uniform variance. In other words, the variation of our residuals (or amount of error in the model) is similar at each point across the model. Thus, the errors are independent of each other. From the Durbin-Watson test, a score of about 1.86 was obtained, showing no correlation in our residuals, which lies within the permissible limit of 1.5 to 2.5. Finally, since the average of the residuals is about 0.03, we can assume they are centred.

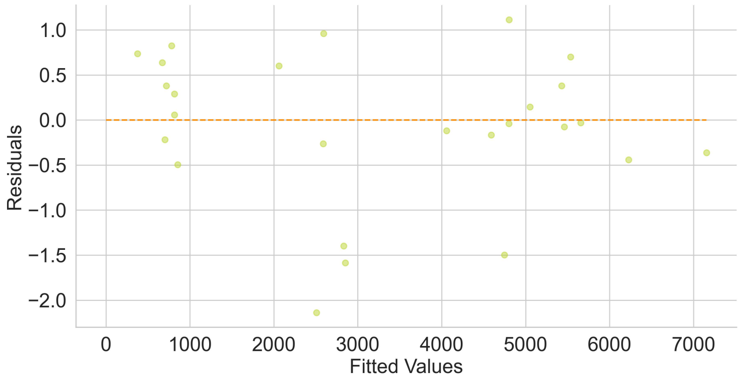


Figure 4.4: Scatterplot to test the homoscedasticity of the residuals.

Therefore, our model successfully passed all the assumptions in the model validation steps, so we can conclude that our model can perform well to predict future grades of Dissertation/Work Project/Internship report by only using two variables, *Gestão dos Sistemas de Informação* and *Metodologias de Investigação*. However, our model only has an R-squared score of 53%, which means there are still about 47% unknown factors affecting the grades of Dissertation/Work Project/Internship Report.

4.2 Relationships between online learning activities in the final grade

As our objective is to understand the influence of online learning behaviour on the final grade, we defined the consequent (target) to be the final grade.

Beyond that, for both courses, after obtaining all the rules for the three algorithms, we looked for the thirty rules that had the highest confidence and the highest level of complementarity, i.e., with the highest lift value (assuming that confidence and support were of at least 75% and 7%, respectively).

Ultimately, the networks of the top 50 confidence or support rules are presented in Appendix C.

4.2.1 Programação para a Ciência de Dados

To contextualize, three hundred and twenty-one students took this discipline in the four academic years under study, and most (two hundred and twelve) obtained an Excellent. Sixty-six obtained a Very Good, twenty-five a Good, seventeen a Sufficient, and only one was not approved in this course, getting a final grade of Insufficient.

After setting the minimum support threshold, we kept 3106 rules for Apriori and FP-Growth algorithms. As for the Eclat algorithm, 3099 rules were retained.

4.2.1.1 Apriori

From the top thirty association rules presented in Table 4.5, we only got rules where the grade was marked as Excellent.

Table 4.5: Top 30 rules.

Antecedents	Consequents	Support	Confidence	Lift
(contents_updated_[5:10], dedication_time_[16140137:32882610])	(notaFinal_Excellent)	0.071651	0.884615	1.339441
(submissions_[5:7], quizes_done_[1:2], web_links_viewed_[66:139], quizes_reviewed_[1:2], discussion_posts_viewed_[40:87])	(notaFinal_Excellent)	0.071651	0.884615	1.339441
(web_links_viewed_[66:139], quizes_reviewed_[1:2], discussion_posts_viewed_[40:87], submissions_[5:7])	(notaFinal_Excellent)	0.071651	0.884615	1.339441
(dedication_time_>32882610, log_freq_>733)	(notaFinal_Excellent)	0.090343	0.878788	1.330617
(quizes_reviewed_[1:2], discussion_posts_viewed_[40:87], submissions_[5:7])	(notaFinal_Excellent)	0.084112	0.870968	1.318777
(quizes_done_[1:2], discussion_posts_viewed_[40:87], quizes_reviewed_[1:2], submissions_[5:7])	(notaFinal_Excellent)	0.084112	0.870968	1.318777
(discussion_posts_viewed_[40:87], files_viewed_[100:219])	(notaFinal_Excellent)	0.084112	0.870968	1.318777
(log_freq_[360:733], contents_updated_[5:10])	(notaFinal_Excellent)	0.080997	0.866667	1.312264
(quizes_done_[1:2], discussion_posts_viewed_[40:87], submissions_[5:7])	(notaFinal_Excellent)	0.118380	0.863636	1.307676
(log_freq_[360:733], inactive_time_[11303108:22884126], contents_updated_[5:10])	(notaFinal_Excellent)	0.077882	0.862069	1.305303
(dedication_time_>32882610, discussion_forum_posts_0)	(notaFinal_Excellent)	0.077882	0.862069	1.305303
(dedication_time_>32882610, inactive_time_>22884126, log_freq_>733)	(notaFinal_Excellent)	0.077882	0.862069	1.305303
(dedication_time_>32882610, inactive_time_>22884126)	(notaFinal_Excellent)	0.077882	0.862069	1.305303
(log_freq_>733)	(notaFinal_Excellent)	0.096573	0.861111	1.303852
(dedication_time_>32882610)	(notaFinal_Excellent)	0.096573	0.861111	1.303852
(web_links_viewed_[66:139], quizes_done_[1:2], discussion_posts_viewed_[40:87], submissions_[5:7])	(notaFinal_Excellent)	0.096573	0.861111	1.303852
(discussion_forum_posts_0, quizes_done_[1:2], discussion_posts_viewed_[40:87], quizes_reviewed_[1:2])	(notaFinal_Excellent)	0.074766	0.857143	1.297844
(discussion_forum_posts_0, quizes_reviewed_[1:2], discussion_posts_viewed_[40:87])	(notaFinal_Excellent)	0.074766	0.857143	1.297844
(discussion_forum_posts_0, quizes_done_[1:2], downloads_[3:7])	(notaFinal_Excellent)	0.074766	0.857143	1.297844
(inactive_time_[11303108:22884126], contents_updated_[5:10])	(notaFinal_Excellent)	0.090343	0.852941	1.291482
(quizes_done_[1:2], discussion_posts_viewed_[40:87], dedication_time_[16140137:32882610])	(notaFinal_Excellent)	0.090343	0.852941	1.291482
(quizes_reviewed_[1:2], contents_updated_[3:5], discussion_posts_viewed_[40:87])	(notaFinal_Excellent)	0.071651	0.851852	1.289832
(dedication_time_>32882610, discussion_forum_posts_0, log_freq_>733)	(notaFinal_Excellent)	0.071651	0.851852	1.289832
(quizes_done_[1:2], contents_updated_[3:5], discussion_posts_viewed_[40:87], quizes_reviewed_[1:2])	(notaFinal_Excellent)	0.071651	0.851852	1.289832
(files_viewed_[100:219], quizes_reviewed_[1:2], submissions_[5:7])	(notaFinal_Excellent)	0.071651	0.851852	1.289832
(discussion_forum_posts_0, discussion_posts_viewed_>87)	(notaFinal_Excellent)	0.071651	0.851852	1.289832
(submissions_[5:7], quizes_done_[1:2], web_links_viewed_[66:139], inactive_time_[11303108:22884126], discussion_posts_viewed_[40:87])	(notaFinal_Excellent)	0.071651	0.851852	1.289832

Continued on next page

Antecedents	Consequents	Support	Confidence	Lift
(files_viewed_[100:219], quizes_done_[1:2], quizes_reviewed_[1:2], submissions_[5:7])	(notaFinal_Excellent)	0.071651	0.851852	1.289832
(quizes_reviewed_[1:2], discussion_posts_viewed_[40:87], quizes_done_[1:2])	(notaFinal_Excellent)	0.105919	0.850000	1.287028
(quizes_reviewed_[1:2], discussion_posts_viewed_[40:87])	(notaFinal_Excellent)	0.105919	0.850000	1.287028

From the analysis, we can draw some meaningful and insightful conclusions:

- 88% of the students who have made 1-2 quizzes and reviewed them the same amount of times, made 5-7 submissions, read 40-87 forum posts, and opened 66-139 web links, also achieved a final grade of Excellent. In addition, this rule has the highest lift value (lift = 1.34), so it can be specified as the most impressive and valuable rule among the rules;

- 88% of the students that had a dedication time of more than 32882610 seconds (approximately one year and fifteen days) and logged in into the system more than 733 times also obtained a final grade of Excellent;

- 87% of the students who read between 40-87 forum posts and 100-219 moodle files also obtained a final grade of Excellent;

- 86% of the students who made 1-2 quizzes, downloaded between 3-7 course materials, but didn't post anything on the discussion forum, also had a final grade of Excellent;

- 85% of the students who read more than 87 messages on the discussion forum but didn't post anything also achieved a final grade of Excellent.

- 85% of the students who were between 11303108-22884126 seconds (approximately 131 to 265 days) out of the system and took between 5-10 updates also finished the course with a final grade of Excellent.

4.2.1.2 FP-Growth

As in the case of the Apriori algorithm, we only obtained rules presented in Table 4.6, where the grade was marked as Excellent.

Table 4.6: Top 30 rules.

Antecedents	Consequents	Support	Confidence	Lift
(web_links_viewed_[66:139], quizes_reviewed_[1:2], discussion_posts_viewed_[40:87], submissions_[5:7])	(notaFinal_Excellent)	0.071651	0.884615	1.339441
(contents_updated_[5:10], dedication_time_[16140137:32882610])	(notaFinal_Excellent)	0.071651	0.884615	1.339441
(submissions_[5:7], quizes_done_[1:2], web_links_viewed_[66:139], quizes_reviewed_[1:2], discussion_posts_viewed_[40:87])	(notaFinal_Excellent)	0.071651	0.884615	1.339441
(dedication_time_>32882610, log_freq_>733)	(notaFinal_Excellent)	0.090343	0.878788	1.330617
(quizes_done_[1:2], discussion_posts_viewed_[40:87], quizes_reviewed_[1:2], submissions_[5:7])	(notaFinal_Excellent)	0.084112	0.870968	1.318777
(quizes_reviewed_[1:2], discussion_posts_viewed_[40:87], submissions_[5:7])	(notaFinal_Excellent)	0.084112	0.870968	1.318777
(discussion_posts_viewed_[40:87], files_viewed_[100:219])	(notaFinal_Excellent)	0.084112	0.870968	1.318777
(log_freq_[360:733], contents_updated_[5:10])	(notaFinal_Excellent)	0.080997	0.866667	1.312264
(quizes_done_[1:2], discussion_posts_viewed_[40:87], submissions_[5:7])	(notaFinal_Excellent)	0.118380	0.863636	1.307676
(dedication_time_>32882610, discussion_forum_posts_0)	(notaFinal_Excellent)	0.077882	0.862069	1.305303
(dedication_time_>32882610, inactive_time_>22884126, log_freq_>733)	(notaFinal_Excellent)	0.077882	0.862069	1.305303
(dedication_time_>32882610, inactive_time_>22884126)	(notaFinal_Excellent)	0.077882	0.862069	1.305303
(log_freq_[360:733], inactive_time_[11303108:22884126], contents_updated_[5:10])	(notaFinal_Excellent)	0.077882	0.862069	1.305303
(log_freq_>733)	(notaFinal_Excellent)	0.096573	0.861111	1.303852
(dedication_time_>32882610)	(notaFinal_Excellent)	0.096573	0.861111	1.303852
(web_links_viewed_[66:139], quizes_done_[1:2], discussion_posts_viewed_[40:87], submissions_[5:7])	(notaFinal_Excellent)	0.096573	0.861111	1.303852
(discussion_forum_posts_0, quizes_done_[1:2], downloads_[3:7])	(notaFinal_Excellent)	0.074766	0.857143	1.297844
(discussion_forum_posts_0, quizes_done_[1:2], discussion_posts_viewed_[40:87], quizes_reviewed_[1:2])	(notaFinal_Excellent)	0.074766	0.857143	1.297844
(discussion_forum_posts_0, quizes_reviewed_[1:2], discussion_posts_viewed_[40:87])	(notaFinal_Excellent)	0.074766	0.857143	1.297844
(quizes_done_[1:2], discussion_posts_viewed_[40:87], dedication_time_[16140137:32882610])	(notaFinal_Excellent)	0.090343	0.852941	1.291482
(inactive_time_[11303108:22884126], contents_updated_[5:10])	(notaFinal_Excellent)	0.090343	0.852941	1.291482
(quizes_done_[1:2], contents_updated_[3:5], discussion_posts_viewed_[40:87], quizes_reviewed_[1:2])	(notaFinal_Excellent)	0.071651	0.851852	1.289832
(quizes_reviewed_[1:2], contents_updated_[3:5], discussion_posts_viewed_[40:87])	(notaFinal_Excellent)	0.071651	0.851852	1.289832
(discussion_forum_posts_0, discussion_posts_viewed_>87)	(notaFinal_Excellent)	0.071651	0.851852	1.289832
(files_viewed_[100:219], quizes_reviewed_[1:2], submissions_[5:7])	(notaFinal_Excellent)	0.071651	0.851852	1.289832
(dedication_time_>32882610, discussion_forum_posts_0, log_freq_>733)	(notaFinal_Excellent)	0.071651	0.851852	1.289832
(submissions_[5:7], quizes_done_[1:2], web_links_viewed_[66:139], inactive_time_[11303108:22884126], discussion_posts_viewed_[40:87])	(notaFinal_Excellent)	0.071651	0.851852	1.289832
(files_viewed_[100:219], quizes_done_[1:2], quizes_reviewed_[1:2], submissions_[5:7])	(notaFinal_Excellent)	0.071651	0.851852	1.289832
(quizes_reviewed_[1:2], discussion_posts_viewed_[40:87], quizes_done_[1:2])	(notaFinal_Excellent)	0.105919	0.850000	1.287028
(quizes_reviewed_[1:2], discussion_posts_viewed_[40:87])	(notaFinal_Excellent)	0.105919	0.850000	1.287028

Some relevant information can be extracted from the displayed combinations, such as:

- 88% of the students who have made 5-10 updates, and spent between 16140137-32882610 seconds (between 187 to 381 days) on moodle, also achieved a final grade of Excellent. This is the rule with the highest lift value (lift = 1.34), so it can be specified as the most impressive and valuable rule among the rules;
- 86% of the students who logged in into the system more than 733 times also ended up with a final grade of Excellent;
- 86% of the students who logged in more than 733, spent more than 32882610 seconds (>381 days) on the system and had more than 22884126 seconds (>265 days) out of the system also acquired a final grade of Excellent.

4.2.1.3 Eclat

The Tables 4.7 and 4.8 present the rules obtained where the final grade was marked as Very Good and Excellent.

Table 4.7: All the rules - Final grade of Very Good.

Antecedents	Consequents	Support
(discussion_forum_posts_0)	(notaFinal_Very good)	0.161994
(quizes_reviewed_0, discussion_forum_posts_0)	(notaFinal_Very good)	0.133956
(quizes_reviewed_0)	(notaFinal_Very good)	0.121495
(quizes_reviewed_0, quizes_done_0)	(notaFinal_Very good)	0.112150
(quizes_reviewed_0, discussion_forum_posts_0, quizes_done_0)	(notaFinal_Very good)	0.109034
(discussion_forum_posts_0, quizes_done_0)	(notaFinal_Very good)	0.105919
(discussion_forum_posts_0, contents_updated_[3:5])	(notaFinal_Very good)	0.096573
(submissions_[14:27], quizes_done_0)	(notaFinal_Very good)	0.096573
(chat_messages_0)	(notaFinal_Very good)	0.096573
(discussion_posts_viewed_[40:87])	(notaFinal_Very good)	0.093458
(quizes_done_0)	(notaFinal_Very good)	0.093458
(quizes_reviewed_0, quizes_done_0, contents_updated_[3:5])	(notaFinal_Very good)	0.087227
(submissions_[14:27], quizes_reviewed_0)	(notaFinal_Very good)	0.087227
(files_viewed_[220:412])	(notaFinal_Very good)	0.087227
(quizes_done_0, contents_updated_[3:5])	(notaFinal_Very good)	0.087227
(quizes_reviewed_0, contents_updated_[3:5])	(notaFinal_Very good)	0.087227
(contents_updated_[3:5])	(notaFinal_Very good)	0.074766
(web_links_viewed_[66:139])	(notaFinal_Very good)	0.074766
(submissions_[14:27])	(notaFinal_Very good)	0.074766

Continued on next page

Antecedents	Consequents	Support
(submissions_[14:27], quizes_reviewed_0, quizes_done_0)	(notaFinal_Very good)	0.071651
(web_links_viewed_[36:65])	(notaFinal_Very good)	0.071651

Table 4.8: Top 30 rules - Final grade of Excellent

Antecedents	Consequents	Support
(discussion_forum_posts_0)	(notaFinal_Excelent)	0.660436
(files_viewed_[220:412], contents_updated_[3:5])	(notaFinal_Excelent)	0.504673
(web_links_viewed_[66:139], log_freq_[360:733])	(notaFinal_Excelent)	0.386293
(files_viewed_[220:412])	(notaFinal_Excelent)	0.380062
(quizes_done_[1:2], discussion_posts_viewed_[40:87])	(notaFinal_Excelent)	0.314642
(chat_messages_0, discussion_forum_posts_0)	(notaFinal_Excelent)	0.308411
(discussion_posts_viewed_[40:87], contents_updated_[3:5])	(notaFinal_Excelent)	0.302181
(web_links_viewed_[36:65])	(notaFinal_Excelent)	0.295950
(inactive_time_[7412743:11303107], dedication_time_[10313281:16140136])	(notaFinal_Excelent)	0.295950
(web_links_viewed_[66:139], dedication_time_[10313281:16140136])	(notaFinal_Excelent)	0.292835
(quizes_done_[1:2], inactive_time_[11303108:22884126])	(notaFinal_Excelent)	0.274143
(quizes_reviewed_0, files_viewed_[413:882])	(notaFinal_Excelent)	0.271028
(submissions_[14:27], quizes_done_0)	(notaFinal_Excelent)	0.264798
(inactive_time_[11303108:22884126], dedication_time_[16140137:32882610], contents_updated_[3:5])	(notaFinal_Excelent)	0.264798
(files_viewed_[220:412], quizes_done_0)	(notaFinal_Excelent)	0.264798
(submissions_[5:7], web_links_viewed_[66:139], discussion_forum_posts_0, contents_updated_[3:5])	(notaFinal_Excelent)	0.258567
(log_freq_[360:733], inactive_time_[11303108:22884126])	(notaFinal_Excelent)	0.258567
(inactive_time_[11303108:22884126])	(notaFinal_Excelent)	0.258567
(log_freq_[360:733], contents_updated_[3:5], discussion_forum_posts_0, inactive_time_[11303108:22884126], dedication_time_[16140137:32882610])	(notaFinal_Excelent)	0.249221
(submissions_[5:7], quizes_done_[1:2], inactive_time_[11303108:22884126])	(notaFinal_Excelent)	0.239875
(discussion_forum_posts_[1:2])	(notaFinal_Excelent)	0.236760
(web_links_viewed_[66:139])	(notaFinal_Excelent)	0.227414
(inactive_time_[11303108:22884126], log_freq_[360:733], files_viewed_[220:412], dedication_time_[16140137:32882610])	(notaFinal_Excelent)	0.224299
(web_links_viewed_[66:139], files_viewed_[220:412])	(notaFinal_Excelent)	0.221184
(log_freq_[111:228], quizes_reviewed_0, discussion_forum_posts_0)	(notaFinal_Excelent)	0.221184
(web_links_viewed_[66:139], dedication_time_[16140137:32882610])	(notaFinal_Excelent)	0.221184
(quizes_reviewed_0, dedication_time_[10313281:16140136], quizes_done_0)	(notaFinal_Excelent)	0.221184
(dedication_time_[16140137:32882610])	(notaFinal_Excelent)	0.218069
(log_freq_[360:733], dedication_time_[16140137:32882610], contents_updated_[3:5])	(notaFinal_Excelent)	0.218069
(discussion_posts_viewed_[40:87], quizes_reviewed_[1:2])	(notaFinal_Excelent)	0.214953

From their analysis, we can point out that :

- 39% of all students logged in into the system between 360-733 times, opened 66-139 web links and had a final grade of Excellent;
- 31% of all students read between 40-87 forum discussions, made 1-2 quizzes and achieved a final grade of Excellent;

- 21% of all students viewed 40-87 forum discussions, made 1-2 quizzes reviews and ended up with a final grade of Excellent;
- 11% of all students didn't make any quizzes, didn't make any quiz reviews and acquired a final grade of Very Good;
- 7% of all students opened 36-65 web links and got a final grade of Very Good;

To sum things up, we can observe that more students achieved a final grade of Excellent. These students opened more web links and did and reviewed more quizzes than those who obtained a final grade of Very Good.

4.2.2 Aprendizagem Profunda

Of one hundred and twenty-one students enrolled, most obtained a grade marked as Very Good (forty-nine), and none failed this course. Of the remaining, thirty-eight achieved an Excellent, twenty-one a Good, and thirteen a Sufficient as final grades.

After setting the minimum support threshold, we kept 1357 rules for Apriori, FP-Growth and ECLAT algorithms.

4.2.2.1 Apriori

The Table 4.9 shows all the rules obtained, and as can be seen, we only found rules where the final grade was marked as Very Good.

Table 4.9: All the rules with a minimum confidence of 0.75

Antecedents	Consequents	Support	Confidence	Lift
(chat_messages_0, dedication_time_[4978487:10313280], discussion_forum_posts_0, log_freq_[111:228])	(notaFinal_Very good)	0.074380	0.818182	2.020408
(chat_messages_0, discussion_forum_posts_0, inactive_time_[3582429:7412742], contents_updated_[3:5])	(notaFinal_Very good)	0.074380	0.818182	2.020408
(dedication_time_[4978487:10313280], discussion_forum_posts_0, inactive_time_[3582429:7412742], chat_messages_0, log_freq_[111:228])	(notaFinal_Very good)	0.074380	0.818182	2.020408
(chat_messages_0, dedication_time_[4978487:10313280], discussion_forum_posts_0, inactive_time_[3582429:7412742])	(notaFinal_Very good)	0.082645	0.769231	1.899529
(chat_messages_0, discussion_forum_posts_0, inactive_time_[3582429:7412742], log_freq_[111:228])	(notaFinal_Very good)	0.082645	0.769231	1.899529
(chat_messages_0, discussion_forum_posts_0, log_freq_[111:228])	(notaFinal_Very good)	0.082645	0.769231	1.899529
(chat_messages_0, discussion_forum_posts_0, inactive_time_[3582429:7412742])	(notaFinal_Very good)	0.099174	0.750000	1.852041
(downloads_[8:15], discussion_forum_posts_0)	(notaFinal_Very good)	0.074380	0.750000	1.852041
(chat_messages_0, discussion_forum_posts_0, submissions_[14:27], log_freq_[111:228])	(notaFinal_Very good)	0.074380	0.750000	1.852041
(discussion_forum_posts_0, inactive_time_[3582429:7412742], chat_messages_0, submissions_[14:27], log_freq_[111:228])	(notaFinal_Very good)	0.074380	0.750000	1.852041

Some of the most interesting findings are listed below:

- 82% of the students who didn't send any messages, didn't post on the discussion forum, spent between 4978487-10313280 seconds (58-119 days approximately) on moodle platform, had a total of idle time of between 3582429-7412742 seconds (42-86 days approximately), and logged into the system between 111-228, also ended up with a final grade of Very Good. As this rule has the highest lift value (lift = 2.02), it can be specified as the most impressive and valuable rule among the rules;
- 82% of the students who didn't send any messages didn't post on the discussion forum, made 3-5 updates, and spent between 3582429-7412742 seconds (42-86 days approximately) out of the system, also got a final grade of Very Good;
- 75% of the students who didn't post on the discussion forum, and made 8-15 downloads, also achieved a final grade of Very Good;
- 75% of the students who didn't send any message, didn't post on the discussion forum, made between 14-27 content submissions, and logged in between 111-228 times, also got a final grade of Very Good.

Since the rules obtained through the specified criteria were too few for this discipline, we proceeded to lower the confidence threshold to 50%, keeping the minimum support at 7%, in an attempt to find unexpected and surprising rules. By decreasing the confidence level, more rules were generated, thus allowing us to analyze more rules where the consequent has the value of Very good, presented in Table 4.10. In addition, rules were found where the consequent assumes the value of Excellent (there were not any with the confidence of 75% or higher), shown in Table 4.11.

Table 4.10: Top 30 rules with a minimum confidence of 0.50 - Final grade of Very Good

Antecedents	Consequents	Support	Confidence	Lift
(chat_messages_0, discussion_forum_posts_0, inactive_time_[3582429:7412742], contents_updated_[3:5])	(notaFinal_Very good)	0.074380	0.818182	2.020408
(dedication_time_[4978487:10313280], discussion_forum_posts_0, inactive_time_[3582429:7412742], chat_messages_0, log_freq_[111:228])	(notaFinal_Very good)	0.074380	0.818182	2.020408
(chat_messages_0, dedication_time_[4978487:10313280], discussion_forum_posts_0, log_freq_[111:228])	(notaFinal_Very good)	0.074380	0.818182	2.020408
(chat_messages_0, discussion_forum_posts_0, log_freq_[111:228])	(notaFinal_Very good)	0.082645	0.769231	1.899529
(chat_messages_0, discussion_forum_posts_0, inactive_time_[3582429:7412742], log_freq_[111:228])	(notaFinal_Very good)	0.082645	0.769231	1.899529
(chat_messages_0, dedication_time_[4978487:10313280], discussion_forum_posts_0, inactive_time_[3582429:7412742])	(notaFinal_Very good)	0.082645	0.769231	1.899529
(chat_messages_0, discussion_forum_posts_0, submissions_[14:27], log_freq_[111:228])	(notaFinal_Very good)	0.074380	0.750000	1.852041
(discussion_forum_posts_0, inactive_time_[3582429:7412742], chat_messages_0, submissions_[14:27], log_freq_[111:228])	(notaFinal_Very good)	0.074380	0.750000	1.852041
(downloads_[8:15], discussion_forum_posts_0)	(notaFinal_Very good)	0.074380	0.750000	1.852041

Continued on next page

Antecedents	Consequents	Support	Confidence	Lift
(chat_messages_0, discussion_forum_posts_0, inactive_time_[3582429:7412742])	(notaFinal_Very good)	0.099174	0.750000	1.852041
(discussion_posts_viewed_[7:19], contents_updated_[3:5])	(notaFinal_Very good)	0.082645	0.714286	1.763848
(discussion_forum_posts_0, discussion_posts_viewed_[7:19], contents_updated_[3:5])	(notaFinal_Very good)	0.082645	0.714286	1.763848
(chat_messages_0, discussion_forum_posts_0, inactive_time_[3582429:7412742], submissions_[14:27])	(notaFinal_Very good)	0.082645	0.714286	1.763848
(submissions_[14:27], discussion_posts_viewed_[7:19], contents_updated_[3:5])	(notaFinal_Very good)	0.074380	0.692308	1.709576
(chat_messages_0, dedication_time_[4978487:10313280], discussion_forum_posts_0, contents_updated_[3:5])	(notaFinal_Very good)	0.074380	0.692308	1.709576
(discussion_forum_posts_0, contents_updated_[3:5], submissions_[14:27], discussion_posts_viewed_[7:19])	(notaFinal_Very good)	0.074380	0.692308	1.709576
(chat_messages_0, dedication_time_[4978487:10313280], discussion_forum_posts_0)	(notaFinal_Very good)	0.090909	0.687500	1.697704
(discussion_forum_posts_0, submissions_[14:27], discussion_posts_viewed_[7:19])	(notaFinal_Very good)	0.090909	0.687500	1.697704
(submissions_[14:27], discussion_posts_viewed_[7:19])	(notaFinal_Very good)	0.090909	0.687500	1.697704
(dedication_time_[4978487:10313280], discussion_forum_posts_0, inactive_time_[3582429:7412742], submissions_[14:27], log_freq_[111:228])	(notaFinal_Very good)	0.082645	0.666667	1.646259
(discussion_forum_posts_0, discussion_posts_viewed_[7:19])	(notaFinal_Very good)	0.099174	0.666667	1.646259
(discussion_posts_viewed_[7:19])	(notaFinal_Very good)	0.099174	0.666667	1.646259
(dedication_time_[4978487:10313280], discussion_forum_posts_0, inactive_time_[3582429:7412742], log_freq_[111:228])	(notaFinal_Very good)	0.090909	0.647059	1.597839
(chat_messages_0, dedication_time_[4978487:10313280], discussion_forum_posts_0, submissions_[14:27])	(notaFinal_Very good)	0.074380	0.642857	1.587464
(discussion_forum_posts_0, inactive_time_[3582429:7412742], log_freq_[111:228], contents_updated_[3:5])	(notaFinal_Very good)	0.074380	0.642857	1.587464
(chat_messages_0, inactive_time_[3582429:7412742], contents_updated_[3:5])	(notaFinal_Very good)	0.074380	0.642857	1.587464
(dedication_time_[4978487:10313280], discussion_forum_posts_0, inactive_time_[3582429:7412742], contents_updated_[3:5])	(notaFinal_Very good)	0.074380	0.642857	1.587464
(discussion_forum_posts_0, discussion_posts_viewed_[7:19], files_viewed_[220:412])	(notaFinal_Very good)	0.074380	0.642857	1.587464
(discussion_posts_viewed_[7:19], files_viewed_[220:412])	(notaFinal_Very good)	0.074380	0.642857	1.587464
(downloads_[3:7], submissions_[14:27])	(notaFinal_Very good)	0.074380	0.642857	1.587464

Table 4.11: All the rules with a minimum confidence of 0.50 - Final grade of Excellent

Antecedents	Consequents	Support	Confidence	Lift
(files_viewed_[413:882], discussion_forum_posts_[1:2], contents_updated_[3:5])	(notaFinal_Excellent)	0.074380	0.642857	2.046992
(files_viewed_[413:882], discussion_forum_posts_[1:2], submissions_[14:27])	(notaFinal_Excellent)	0.082645	0.555556	1.769006
(downloads_[1:2])	(notaFinal_Excellent)	0.074380	0.529412	1.685759
(chat_messages_>70)	(notaFinal_Excellent)	0.082645	0.526316	1.675900
(contents_updated_[5:10], log_freq_[111:228])	(notaFinal_Excellent)	0.074380	0.500000	1.592105
(files_viewed_[413:882], discussion_forum_posts_[1:2], discussion_posts_viewed_[40:87])	(notaFinal_Excellent)	0.074380	0.500000	1.592105
(submissions_[14:27], discussion_forum_posts_[1:2], discussion_posts_viewed_[40:87])	(notaFinal_Excellent)	0.074380	0.500000	1.592105

From their analysis, we can obtain additional information in relation to those pointed out above :

- 77% of the students who didn't send any chat message, didn't make any post on the discussion forum and had a login frequency of 111-228 also achieved a final grade of Very Good;
- 71% of the students who didn't post on the discussion forum but read 7-19 times its content and made 3-5 updates also obtained a final grade of Very Good;
- 64% of the students who read the contents of the discussion forum between 7-19 times and read the course materials between 220-412 times also got a final grade of Very Good;
- 56% of the students who viewed course materials between 413-882 times, posted between 1-2 times on the discussion forum, and made 14-27 submissions, also achieved a final grade of Excellent;
- 53% of the students who sent more than 70 chat messages also achieved a final grade of Excellent;
- 50% of the students who read course materials between 413-882 times, posted 1-2 times on the discussion forum, and read their contents between 40-87 times, also acquired a final grade of Excellent;
- 50% of the students who made 5-10 updates and logged in 111-228 times on the moodle platform also had a final grade of Excellent.

4.2.2.2 FP-Growth

In Table 4.12, all the rules obtained are presented, and once again, only rules where the final grade was classified as Very Good were found.

Table 4.12: All the rules with a minimum confidence of 0.75

Antecedents	Consequents	Support	Confidence	Lift
(chat_messages_0, dedication_time_[4978487:10313280], discussion_forum_posts_0, log_freq_[111:228])	(notaFinal_Very good)	0.074380	0.818182	2.020408
(chat_messages_0, discussion_forum_posts_0, inactive_time_[3582429:7412742], contents_updated_[3:5])	(notaFinal_Very good)	0.074380	0.818182	2.020408
(dedication_time_[4978487:10313280], discussion_forum_posts_0, inactive_time_[3582429:7412742], chat_messages_0, log_freq_[111:228])	(notaFinal_Very good)	0.074380	0.818182	2.020408
(chat_messages_0, discussion_forum_posts_0, log_freq_[111:228])	(notaFinal_Very good)	0.082645	0.769231	1.899529
(chat_messages_0, discussion_forum_posts_0, inactive_time_[3582429:7412742], log_freq_[111:228])	(notaFinal_Very good)	0.082645	0.769231	1.899529
(chat_messages_0, dedication_time_[4978487:10313280], discussion_forum_posts_0, inactive_time_[3582429:7412742])	(notaFinal_Very good)	0.082645	0.769231	1.899529
(downloads_[8:15], discussion_forum_posts_0)	(notaFinal_Very good)	0.074380	0.750000	1.852041

Continued on next page

Antecedents	Consequents	Support	Confidence	Lift
(chat_messages_0, discussion_forum_posts_0, inactive_time_[3582429:7412742])	(notaFinal_Very good)	0.099174	0.750000	1.852041
(discussion_forum_posts_0, inactive_time_[3582429:7412742], chat_messages_0, submissions_[14:27], log_freq_[111:228])	(notaFinal_Very good)	0.074380	0.750000	1.852041
(chat_messages_0, discussion_forum_posts_0, submissions_[14:27], log_freq_[111:228])	(notaFinal_Very good)	0.074380	0.750000	1.852041

From its analysis, we can gather some interesting insights:

- 77% of the students who didn't post on the discussion forum, didn't send any chat message, became offline from 3582429-7412742 seconds (42-86 days approximately), and logged into the system between 111-228, also ended up with a final grade of Very Good;
- 75% of the students who didn't post anything on the discussion forum, and made 8-15 downloads, also got a final grade of Very Good;

Once more, as only ten rules were acquired according to the established criteria, the confidence threshold was lowered to 50%, keeping the minimum support at 7%. By decreasing the confidence level, we had the opportunity to analyze more rules where the consequent has been set to Very good, displayed in Table 4.13. In addition, rules were found where the consequent takes the value of Excellent (there were not any with the confidence of 75% or higher), exhibited in Table 4.14.

Table 4.13: Top 30 rules with a minimum confidence of 0.50 - Final grade of Very Good

Antecedents	Consequents	Support	Confidence	Lift
(dedication_time_[4978487:10313280], discussion_forum_posts_0, inactive_time_[3582429:7412742], chat_messages_0, log_freq_[111:228])	(notaFinal_Very good)	0.074380	0.818182	2.020408
(chat_messages_0, dedication_time_[4978487:10313280], discussion_forum_posts_0, log_freq_[111:228])	(notaFinal_Very good)	0.074380	0.818182	2.020408
(chat_messages_0, discussion_forum_posts_0, inactive_time_[3582429:7412742], contents_updated_[3:5])	(notaFinal_Very good)	0.074380	0.818182	2.020408
(chat_messages_0, discussion_forum_posts_0, inactive_time_[3582429:7412742], log_freq_[111:228])	(notaFinal_Very good)	0.082645	0.769231	1.899529
(chat_messages_0, dedication_time_[4978487:10313280], discussion_forum_posts_0, inactive_time_[3582429:7412742])	(notaFinal_Very good)	0.082645	0.769231	1.899529
(chat_messages_0, discussion_forum_posts_0, log_freq_[111:228])	(notaFinal_Very good)	0.082645	0.769231	1.899529
(chat_messages_0, discussion_forum_posts_0, inactive_time_[3582429:7412742])	(notaFinal_Very good)	0.099174	0.750000	1.852041
(chat_messages_0, discussion_forum_posts_0, submissions_[14:27], log_freq_[111:228])	(notaFinal_Very good)	0.074380	0.750000	1.852041
(discussion_forum_posts_0, inactive_time_[3582429:7412742], chat_messages_0, submissions_[14:27], log_freq_[111:228])	(notaFinal_Very good)	0.074380	0.750000	1.852041
(downloads_[8:15], discussion_forum_posts_0)	(notaFinal_Very good)	0.074380	0.750000	1.852041
(discussion_forum_posts_0, discussion_posts_viewed_[7:19], contents_updated_[3:5])	(notaFinal_Very good)	0.082645	0.714286	1.763848
(discussion_posts_viewed_[7:19], contents_updated_[3:5])	(notaFinal_Very good)	0.082645	0.714286	1.763848

Continued on next page

Antecedents	Consequents	Support	Confidence	Lift
(chat_messages_0, discussion_forum_posts_0, inactive_time_[3582429:7412742], submissions_[14:27])	(notaFinal_Very good)	0.082645	0.714286	1.763848
(discussion_forum_posts_0, contents_updated_[3:5], submissions_[14:27], discussion_posts_viewed_[7:19])	(notaFinal_Very good)	0.074380	0.692308	1.709576
(chat_messages_0, dedication_time_[4978487:10313280], discussion_forum_posts_0, contents_updated_[3:5])	(notaFinal_Very good)	0.074380	0.692308	1.709576
(submissions_[14:27], discussion_posts_viewed_[7:19], contents_updated_[3:5])	(notaFinal_Very good)	0.074380	0.692308	1.709576
(chat_messages_0, dedication_time_[4978487:10313280], discussion_forum_posts_0)	(notaFinal_Very good)	0.090909	0.687500	1.697704
(discussion_forum_posts_0, submissions_[14:27], discussion_posts_viewed_[7:19])	(notaFinal_Very good)	0.090909	0.687500	1.697704
(submissions_[14:27], discussion_posts_viewed_[7:19])	(notaFinal_Very good)	0.090909	0.687500	1.697704
(discussion_posts_viewed_[7:19])	(notaFinal_Very good)	0.099174	0.666667	1.646259
(discussion_forum_posts_0, discussion_posts_viewed_[7:19])	(notaFinal_Very good)	0.099174	0.666667	1.646259
(dedication_time_[4978487:10313280], discussion_forum_posts_0, inactive_time_[3582429:7412742], submissions_[14:27], log_freq_[111:228])	(notaFinal_Very good)	0.082645	0.666667	1.646259
(dedication_time_[4978487:10313280], discussion_forum_posts_0, inactive_time_[3582429:7412742], log_freq_[111:228])	(notaFinal_Very good)	0.090909	0.647059	1.597839
(downloads_[3:7], submissions_[14:27])	(notaFinal_Very good)	0.074380	0.642857	1.587464
(discussion_forum_posts_0, inactive_time_[3582429:7412742], log_freq_[111:228], contents_updated_[3:5])	(notaFinal_Very good)	0.074380	0.642857	1.587464
(chat_messages_0, inactive_time_[3582429:7412742], contents_updated_[3:5])	(notaFinal_Very good)	0.074380	0.642857	1.587464
(discussion_forum_posts_0, discussion_posts_viewed_[7:19], files_viewed_[220:412])	(notaFinal_Very good)	0.074380	0.642857	1.587464
(chat_messages_0, dedication_time_[4978487:10313280], discussion_forum_posts_0, submissions_[14:27])	(notaFinal_Very good)	0.074380	0.642857	1.587464
(dedication_time_[4978487:10313280], discussion_forum_posts_0, inactive_time_[3582429:7412742], contents_updated_[3:5])	(notaFinal_Very good)	0.074380	0.642857	1.587464
(discussion_posts_viewed_[7:19], files_viewed_[220:412])	(notaFinal_Very good)	0.074380	0.642857	1.587464

Table 4.14: All the rules with a minimum confidence of 0.50 - final grade of Excellent

Antecedents	Consequents	Support	Confidence	Lift
(files_viewed_[413:882], discussion_forum_posts_[1:2], contents_updated_[3:5])	(notaFinal_Excellent)	0.074380	0.642857	2.046992
(files_viewed_[413:882], discussion_forum_posts_[1:2], submissions_[14:27])	(notaFinal_Excellent)	0.082645	0.555556	1.769006
(downloads_[1:2])	(notaFinal_Excellent)	0.074380	0.529412	1.685759
(chat_messages_>70)	(notaFinal_Excellent)	0.082645	0.526316	1.675900
(files_viewed_[413:882], discussion_forum_posts_[1:2], discussion_posts_viewed_[40:87])	(notaFinal_Excellent)	0.074380	0.500000	1.592105
(contents_updated_[5:10], log_freq_[111:228])	(notaFinal_Excellent)	0.074380	0.500000	1.592105
(submissions_[14:27], discussion_forum_posts_[1:2], discussion_posts_viewed_[40:87])	(notaFinal_Excellent)	0.074380	0.500000	1.592105

From their analysis, we can extract new information compared to those mentioned above:

- 82% of the students who didn't send any message, didn't post anything on the discussion forum, made between 3-5 content updates, and had an idle time of between 3582429-7412742 seconds, also got a final grade of Very Good. As this rule has the highest lift value (lift = 2.02), it can be specified as the most impressive and valuable rule among the rules;

- 69% of the students who made 14-27 submissions, read between 7-19 forum posts and updated their contents 3-5 times also achieved a final grade of Very Good;
- 67% of the students who didn't post anything on the discussion forum but read their contents 7-19 times also obtained a final grade of Very Good;
- 56% of the students who made 14-27 submissions, posted 1-2 times on the discussion forum, and read the learning materials between 413-882 times, also acquired a final grade of Excellent.
- 53% of the students who sent more than 70 chat messages also achieved a final grade of Excellent;

4.2.2.3 Eclat

The rules extracted where the final grade was set as Good, Very Good, and Excellent are displayed in Tables 4.15, 4.16, and 4.17, respectively.

Table 4.15: All the rules - Final grade of Good

Antecedents	Consequents	Support
(discussion_forum_posts_0)	(notaFinal_Good)	0.157025
(submissions_[14:27], contents_updated_[3:5])	(notaFinal_Good)	0.157025
(submissions_[14:27])	(notaFinal_Good)	0.148760
(web_links_viewed_[36:65])	(notaFinal_Good)	0.123967
(inactive_time_[7412743:11303107])	(notaFinal_Good)	0.107438
(discussion_forum_posts_0, contents_updated_[3:5])	(notaFinal_Good)	0.099174
(inactive_time_[7412743:11303107], log_freq_[229:359], dedication_time_[10313281:16140136])	(notaFinal_Good)	0.082645
(files_viewed_[413:882])	(notaFinal_Good)	0.082645
(inactive_time_[7412743:11303107], log_freq_[229:359])	(notaFinal_Good)	0.082645
(contents_updated_[3:5])	(notaFinal_Good)	0.082645
(inactive_time_[7412743:11303107], dedication_time_[10313281:16140136])	(notaFinal_Good)	0.082645
(log_freq_[229:359], dedication_time_[10313281:16140136])	(notaFinal_Good)	0.074380
(dedication_time_[10313281:16140136])	(notaFinal_Good)	0.074380
(log_freq_[229:359])	(notaFinal_Good)	0.074380
(discussion_posts_viewed_[40:87])	(notaFinal_Good)	0.074380

Table 4.16: Top 30 rules - Final grade of Very Good

Antecedents	Consequents	Support
(discussion_forum_posts_0)	(notaFinal_Very good)	0.330579
(files_viewed_[413:882])	(notaFinal_Very good)	0.280992
(inactive_time_[3582429:7412742], log_freq_[111:228])	(notaFinal_Very good)	0.264463
(discussion_forum_posts_0, submissions_[14:27], files_viewed_[220:412])	(notaFinal_Very good)	0.256198
(web_links_viewed_[66:139])	(notaFinal_Very good)	0.239669
(log_freq_[229:359])	(notaFinal_Very good)	0.231405

Continued on next page

Antecedents	Consequents	Support
(log_freq_[229:359], submissions_[14:27], dedication_time_[10313281:16140136])	(notaFinal_Very good)	0.223140
(dedication_time_[4978487:10313280])	(notaFinal_Very good)	0.214876
(contents_updated_[3:5])	(notaFinal_Very good)	0.198347
(files_viewed_[413:882], discussion_forum_posts_[1:2])	(notaFinal_Very good)	0.198347
(files_viewed_[413:882], submissions_[14:27], contents_updated_[3:5])	(notaFinal_Very good)	0.181818
(chat_messages_0, inactive_time_[3582429:7412742], log_freq_[111:228])	(notaFinal_Very good)	0.173554
(downloads_>35)	(notaFinal_Very good)	0.173554
(log_freq_[360:733], inactive_time_[11303108:22884126], dedication_time_[16140137:32882610])	(notaFinal_Very good)	0.173554
(dedication_time_[10313281:16140136])	(notaFinal_Very good)	0.165289
(chat_messages_0, discussion_forum_posts_0, submissions_[14:27], log_freq_[111:228])	(notaFinal_Very good)	0.165289
(inactive_time_[7412743:11303107], dedication_time_[10313281:16140136])	(notaFinal_Very good)	0.165289
(chat_messages_0, dedication_time_[4978487:10313280], discussion_forum_posts_0)	(notaFinal_Very good)	0.165289
(chat_messages_0, contents_updated_[3:5])	(notaFinal_Very good)	0.165289
(discussion_forum_posts_0, submissions_[14:27])	(notaFinal_Very good)	0.165289
(chat_messages_0, log_freq_[111:228])	(notaFinal_Very good)	0.157025
(discussion_forum_posts_0, submissions_[14:27], log_freq_[111:228])	(notaFinal_Very good)	0.157025
(chat_messages_0, log_freq_[229:359])	(notaFinal_Very good)	0.148760
(discussion_forum_posts_0, discussion_posts_viewed_[7:19])	(notaFinal_Very good)	0.148760
(files_viewed_[220:412])	(notaFinal_Very good)	0.148760
(discussion_forum_posts_0, inactive_time_[3582429:7412742], log_freq_[111:228])	(notaFinal_Very good)	0.148760
(discussion_forum_posts_[1:2])	(notaFinal_Very good)	0.148760
(inactive_time_[7412743:11303107], submissions_[14:27])	(notaFinal_Very good)	0.140496
(web_links_viewed_[66:139], contents_updated_[3:5])	(notaFinal_Very good)	0.140496
(chat_messages_0, files_viewed_[220:412])	(notaFinal_Very good)	0.140496

Table 4.17: Top 30 rules - Final grade of Excellent

Antecedents	Consequents	Support
(submissions_[14:27], contents_updated_[3:5])	(notaFinal_Excelent)	0.413223
(discussion_forum_posts_[1:2])	(notaFinal_Excelent)	0.314050
(discussion_forum_posts_0)	(notaFinal_Excelent)	0.231405
(discussion_posts_viewed_[40:87])	(notaFinal_Excelent)	0.214876
(contents_updated_[5:10])	(notaFinal_Excelent)	0.181818
(web_links_viewed_[36:65])	(notaFinal_Excelent)	0.173554
(dedication_time_[4978487:10313280], inactive_time_[3582429:7412742], submissions_[14:27])	(notaFinal_Excelent)	0.173554
(log_freq_[360:733])	(notaFinal_Excelent)	0.165289
(dedication_time_[4978487:10313280], log_freq_[111:228])	(notaFinal_Excelent)	0.165289
(inactive_time_[11303108:22884126])	(notaFinal_Excelent)	0.157025

Continued on next page

Antecedents	Consequents	Support
(log_freq_[111:228])	(notaFinal_Excelent)	0.157025
(submissions_[14:27])	(notaFinal_Excelent)	0.140496
(discussion_posts_viewed_[20:39])	(notaFinal_Excelent)	0.140496
(web_links_viewed_[36:65], discussion_forum_posts_[1:2])	(notaFinal_Excelent)	0.132231
(contents_updated_[3:5])	(notaFinal_Excelent)	0.132231
(inactive_time_[3582429:7412742])	(notaFinal_Excelent)	0.123967
(web_links_viewed_[66:139], submissions_[14:27])	(notaFinal_Excelent)	0.123967
(files_viewed_[413:882], discussion_posts_viewed_[40:87])	(notaFinal_Excelent)	0.115702
(discussion_posts_viewed_[40:87], submissions_[14:27], contents_updated_[3:5])	(notaFinal_Excelent)	0.115702
(inactive_time_[3582429:7412742], submissions_[14:27])	(notaFinal_Excelent)	0.107438
(dedication_time_[4978487:10313280], inactive_time_[3582429:7412742])	(notaFinal_Excelent)	0.107438
(discussion_posts_viewed_[40:87], files_viewed_[413:882], submissions_[14:27])	(notaFinal_Excelent)	0.107438
(files_viewed_[413:882], contents_updated_[3:5])	(notaFinal_Excelent)	0.107438
(files_viewed_[413:882], inactive_time_[3582429:7412742])	(notaFinal_Excelent)	0.107438
(dedication_time_[4978487:10313280])	(notaFinal_Excelent)	0.107438
(discussion_forum_posts_0, submissions_[14:27])	(notaFinal_Excelent)	0.099174
(inactive_time_[3582429:7412742], submissions_[14:27], log_freq_[111:228])	(notaFinal_Excelent)	0.099174
(discussion_posts_viewed_[40:87], discussion_forum_posts_[1:2], submissions_[14:27])	(notaFinal_Excelent)	0.099174
(discussion_posts_viewed_[40:87], submissions_[14:27])	(notaFinal_Excelent)	0.099174
(files_viewed_[413:882], submissions_[14:27])	(notaFinal_Excelent)	0.099174

From them, we can emphasize:

- 41% of all students made 3-5 updates, 14-27 submissions and had a final grade of Excellent;
- 33% of all students never post on the discussion forum and achieved a final grade of Very Good;
- 31% of all students made 1-2 posts on the discussion forum and got a final grade of Excellent;
- 26% of all students had between 3582429-7412742 seconds out of the system, logged in between 111-228 times and got a final grade of Very Good;
- 21% of all students read between 40-87 learning materials and obtained a final grade of Excellent;
- 13% of all students made 14-27 submissions, viewed 66-139 web links and got a final grade of Excellent;
- 12% of all students opened 36-65 web links and ended up with a final grade of Good;
- 11% of all students had an inactive time between 7412743-11303107 seconds and achieved a final grade of Good.

In short, from the results of the three algorithms, we can conclude that the students that obtained a final grade of Excellent tend to make more updates, read more course materials, post more times in the discussion forum, read more forum contents, open more web links, send more chat messages, made and review more quizzes when compared to those who got a final grade of Very Good.

4.2.3 Comparisons

As seen, for both courses, the Apriori and FP-growth algorithms found identical association rules (for both minimum confidence thresholds of 50% and 75%) when defining the consequent as the final grade. The only difference between them is in the execution times, where for both cases, the FP-growth algorithm proved to be faster to run. Although the Eclat algorithm uses only the support metric, relevant information was extracted from it that coincides with and reinforces the interpretations made by the other algorithms.

To sum up, from this analysis, we can conclude that:

- The login frequency and the time spent in moodle are higher in *Programação para a Ciência de Dados*, which makes sense because it is a course associated with a regular basis of work. Also, when students get final grades marked as Excellent, we saw that they send more chat messages in this course than in *Aprendizagem Profunda*. Furthermore, as most students obtained a final grade of Excellent in this course and Very Good in *Aprendizagem Profunda*, we can infer that the structure of a course directly impacts student performance. Therefore, a discipline in which the student requires daily work tends to generate better achievement, as it can reduce procrastination behaviours and encourage students to stay committed to the course;

- The students prefer to read forum discussions rather than post on the forum;

- The level of student interactivity on the moodle platform is directly related to their academic performance since the more actions a student performs, the more time he/she spends online, and consequently, the better their achievement will be.

We will also analyze the extent to which the patterns of online learning activities influence the courses that stood out in the first experiment The goal is to understand if they are more similar to the course *Programação para a Ciência de Dados* or *Aprendizagem Profunda*.

According to Table 4.18, and comparing with the previous analysis, we can see that the courses *Gestão do Conhecimento*, *Gestão dos Sistemas de Informação* and *Métodos Descritivos de Data Mining* have a high level of online engagement (the average values of the features web_links_viewed, dedication_time, files_viewed and log_freq lie within the boundaries), similar to that found in the *Programação para a Ciência de Dados* course. As for the courses *Business Intelligence II* and *Metodologias de Investigação*, they present a low level of online participation (the average values of the features inactive_time, files_viewed, and web_links_viewed lie within the boundaries), thus identifying themselves more with emphAprendizagem Profunda course.

Table 4.18: Average of online learning activities for the courses characterised as important in the first experiment.

	Metodologias de Investigação	Gestão dos Sistemas de Informação	Business Intelligence II	Gestão do Conhecimento	Métodos Descritivos de Data Mining
dedication_time	11145130	19073870	12801840	17499240	17965960
inactive_time	7288766	11828380	8500136	11290020	11682020
messages_sent	15	9	24	27	37
log_freq	236	402	278	375	390
quizes_done	0	3	0	0	0
quizes_reviewed	0	6	0	0	0
submissions	11	5	7	7	9
discussion_forum_posts	0	0	1	0	1
downloads	8	5	18	12	17
contents_updated	2	2	5	3	5
web_links_viewed	31	50	48	75	72
discussion_posts_views	26	26	35	28	43
files_viewed	381	413	632	435	462

5 Conclusions

The objectives of this research were first to identify which courses have a positive impact on the final grade of the Dissertation/Work Project/Internship Report, and secondly, to discover the associations between online student interactions with final grades of two different courses (*Programação para a Ciência de Dados* and *Aprendizagem Profunda*), whereby this research went from a general to a specific approach.

Nine datasets from 2012-2013 to 2020-2021 were extracted, containing the logs and learning data from moodle and netp@ systems. From them, two datasets were formed since different goals were set. For the first one, all the academic years were taken into account where only the learning data were considered, resulting in a dataframe (df_grades) with 143667 records and 17 attributes. As for the second, we merged all the student historical data available from the last four academic years, resulting in dataframe (df_all) with a total of 18137046 records and 26 features.

The initial analysis and good understanding of the data performed in python programming language led to the transformation and creation of attributes, as well as data filtering resulting in two final datasets containing 10527 records spread over 3 variables and with 13674 records distributed in 15 variables for the dataframes df_grades and df_all, respectively.

After properly preparing the data for modelling, a predictive analysis was employed with Partial Least squares Regression, using a set of courses as predictors and the Dissertation/Work Project/Internship report as the target variable. This analysis was done for the two specializations of the Master in Information Management, using the df_grades dataset.

The results of the PLSR for the Business Intelligence specialization suggest that the most important variables that affect the final grade of the Dissertation/ Work Project/Internship Report are *Business Intelligence II*, *Gestão do Conhecimento*, *Metodologias de Investigação* and *Métodos Descritivos de Data Mining*. Now for the specialization in Information Systems and Technologies Management, the results suggest that the variables that most influence the grades in Dissertation/Work Project/Internship Report are *Gestão dos Sistemas de Informação* and *Metodologias de Investigação*. In short, the course Metodologias de Investigação seems to be the one with more influence on the grades of the Dissertation/Work Project/Internship Report since it appears in both specializations, so its implementation in other master's degrees may prove to be beneficial.

As for the second part, a descriptive analysis was conducted with association rules techniques to depict students' learning patterns for *Programação para a Ciência de Dados* and *Aprendizagem Profunda*. According to the rules' support, lift and confidence values, we first conclude that the study scientifically identified students' behavioural patterns across different courses. Additionally, it was proved that courses with continuous assessment methods achieve better school performance. Also, students' interactions on the moodle platform are associated with academic performance since the more engaged the students are, the better their grades will be.

6

Limitations and recommendations for future works

During the development of this dissertation, some problems arose about the quality of the data coming from moodle and netp@ systems. One of the most significant issues found was that some students presented final grades of zero, which is almost impossible to happen. In addition, some students showed absurd values for the number of quizzes taken and reviewed. For the first case described, as teachers manually insert the grades into the system, we advise that they pay extra attention when entering the data. Apart from that, as both situations directly influence the study in question, we recommend regular monitoring of the data, so these types of incongruences can be identified and corrected in time.

Another aspect is that master's degrees and courses undergo restructuring over time, such as name changes or adding and removing courses from master's degrees. This proved to be a challenge and a limitation that this study faced since, due to the lack of knowledge of such restructuring, it was tough to understand which disciplines kept their content but changed their name, thus letting us end up with fewer observations than expected. So, to facilitate the analysis, we recommend that whenever changes are made, the old names are automatically replaced by the most recent ones so that there is only one name for the same master's degree/course. An alternative to this proposal is that whenever modifications are made, a note is entered into the system describing the changes.

Another constraint is that the number of students completing each specialization of the master's degree in Information Management (courses + final project) is small, so in order to overcome this issue, more data should be collected so that more courses can be included in the analysis, leading to more reliable and consequently better results.

For future works, to better understand the connections between courses and the student's learning behaviours that influence their academic performance, this study should be replicated at other levels of higher education, such as bachelor's and graduate degrees. This way, for the first task, instead of seeing which courses most influence the Dissertation/Work Project/Internship Report grades, we can see the influence that one set of courses (of a semester/year) has on another course (of another semester/year). Furthermore, we could draw more general conclusions, as this will allow us to determine if students of different academic degrees show the same educational patterns regarding courses.

As for the second task, although we have the number of submissions made by students and their partial grades, one way to improve the analysis and perhaps obtain more satisfactory results would be to identify what these partial grades correspond to (e.g. projects, homework, and peer reviews).

Since one common application of association rules mining is in the domain of recommender systems, from the analysis performed and the insights obtained, we can use them to provide a list of recommended activities or study strategies (like tutorials and books). Beyond that, another suggestion is to get more data to do a more detailed and in-depth analysis of the factors that could affect the student's performance, more specifically, extracting more students' data such as age, sex, and nationality (sociodemographic data).

In addition, to complement the work developed, cluster analyzes can be carried out using either the students' grades in the courses, which will allow us to see which students have bad/medium/good grades in a set/or all of the courses or using the engagement metrics created to understand which type of learners students are (e.g., passive and active).

Bibliography

[1] H. Abdi. "Partial least squares regression and projection on latent structure regression (PLS Regression)". In: *Wiley interdisciplinary reviews: computational statistics* 2.1 (2010), pp. 97–106. doi: 10.1002/wics.51.

[2] Z. Abdullah et al. "Mining Least Association Rules of Degree Level Programs Selected by Students". In: *International Journal of Multimedia and Ubiquitous Engineering* 9 (2014-01), pp. 241–254. doi: 10.14257/ijmue.2014.9.1.23.

[3] R. Agrawal, T. Imielinski, and A. Swami. "Mining associations between sets of items in large databases". In: *Proceedings of the ACM SIGMOD International Conference on Management of Data*. 1993, pp. 207–216. isbn: 0897915925. doi: 10.1145/170035.170072.

[4] S. Ahmed, R. Paul, and A. S. M. L. Hoque. "Knowledge discovery from academic data using Association Rule Mining". In: *2014 17th International Conference on Computer and Information Technology (ICCIT)*. 2014, pp. 314–319. doi: 10.1109/ICCITechn.2014.7073107.

[5] N. Akarachantachote, S. Chadcham, and K. Saithanu. "Cutoff threshold of variable importance in projection for variable selection". In: *International Journal of Pure and Apllied Mathematics* 94.3 (2014), pp. 307–322. doi: 10.12732/ijpam.v94i3.2.

[6] A. Algarni. "Data Mining in Education". In: *International Journal of Advanced Computer Science and Applications* 7.6 (2016). doi: 10.14569/IJACSA.2016.070659.

[7] V. Arkorful and N. Abaidoo. "The role of e-learning, advantages and disadvantages of its adoption in higher education". In: *International journal of instructional technology and distance learning* 12.1 (2015), pp. 29–42.

[8] A. Azevedo and M. F. Santos. "KDD, SEMMA and CRISP-DM: a parallel overview". In: *IADIS European Conf. Data Mining*. 2008, pp. 182–185.

[9] G. Badr et al. "Predicting Students' Performance in University Courses: A Case Study and Tool in KSU Mathematics Department". In: *Procedia Computer Science* 82 (2016). 4th Symposium on Data Mining Applications, SDMA2016, 30 March 2016, Riyadh, Saudi Arabia, pp. 80–89. issn: 1877-0509. doi: https://doi.org/10.1016/j.procs.2016.04.012.

[10] S. Bajpai and S. Mani. "Big Data in Education and Learning Analytics". In: *TechnoLearn: An International Journal of Educational Technology* 7 (2017-01), p. 45. DOI: 10.5958/2249-5223.2017.00005.5.

[11] R. S. Baker. "Data mining for education". In: *International Encyclopedia of Education* 7 (2010-01), pp. 112–118.

[12] S. Brin et al. "Dynamic itemset counting and implication rules for market basket data". In: *Proceedings of the 1997 ACM SIGMOD international conference on Management of data*. 1997, pp. 255–264. DOI: 10.1145/253260.253325.

[13] R. Cerezo et al. "Students' LMS interaction patterns and their relationship with achievement: A case study in higher education". In: *Computers & Education* 96 (2016), pp. 42–54. DOI: https://doi.org/10.1016/j.compedu.2016.02.006.

[14] I.-G. Chong and C.-H. Jun. "Performance of some variable selection methods when multicollinearity is present". In: *Chemometrics and Intelligent Laboratory Systems* 78.1 (2005), pp. 103–112. ISSN: 0169-7439. DOI: https://doi.org/10.1016/j.chemolab.2004.12.011.

[15] H. Coates, R. James, and G. Baldwin. "A critical examination of the effects of learning management systems on university teaching and learning". In: *Tertiary Education and Management* 11.1 (2005), pp. 19–36. DOI: 10.1080/13583883.2005.9967137.

[16] R. Conijn et al. "Predicting Student Performance from LMS Data: A Comparison of 17 Blended Courses Using Moodle LMS". In: *IEEE Transactions on Learning Technologies* 10.1 (2017), pp. 17–29. DOI: 10.1109/TLT.2016.2616312.

[17] P. Cortez and A. M. G. Silva. "Using data mining to predict secondary school student performance". In: (2008).

[18] R. Cowan. "The "Industrial Revolution" in the Home: Household Technology and Social Change in the 20th Century". In: *Technology and culture* 17 (1976), pp. 1–23. DOI: 10.2307/3103251.

[19] K. Dahdouh et al. "Smart Courses Recommender System for Online Learning Platform". In: *2018 IEEE 5th International Congress on Information Science and Technology (CiSt)*. 2018, pp. 328–333. DOI: 10.1109/CIST.2018.8596516.

[20] J. Dougherty, R. Kohavi, and M. Sahami. "Supervised and Unsupervised Discretization of Continuous Features". In: *Machine Learning Proceedings 1995*. Ed. by A. Prieditis and S. Russell. Morgan Kaufmann, 1995, pp. 194–202. ISBN: 978-1-55860-377-6. DOI: https://doi.org/10.1016/B978-1-55860-377-6.50032-3.

[21] J. Gamulin, O. Gamulin, and D. Kermek. "Data mining in hybrid learning: Possibility to predict the final exam result". In: *2013 36th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. 2013, pp. 591–596.

[22] E. García et al. "A collaborative educational association rule mining tool". In: *The Internet and Higher Education* 14.2 (2011). Web mining and higher education: Introduction to the special issue, pp. 77–88. ISSN: 1096-7516. DOI: https://doi.org/10.1016/j.iheduc.2010.07.006.

[23] E. García et al. "Drawbacks and solutions of applying association rule mining in learning management systems". In: *CEUR Workshop Proceedings* 305 (2007-01), pp. 13–22.

[24] P. Gouzouasis, M. Guhn, and N. Kishor. "The predictive relationship between achievement and participation in music and achievement in core grade 12 academic subjects". In: *Music Education Research* 9.1 (2007), pp. 81–92. DOI: 10.1080/14613800601127569.

[25] M. Haverila. "Prior E-learning experience and perceived learning outcomes in an undergraduate E-learning course". In: *MERLOT Journal of Online Learning and Teaching* 7.2 (2011), pp. 206–218.

[26] M. Holanda et al. "The Brazilian School Girls' Perspectives on a Computer Science Major". In: *CLEI Electronic Journal* 22.2 (2019-08). DOI: 10.19153/cleiej.22.2.2.

[27] Y.-H. Hu, C.-L. Lo, and S.-P. Shih. "Developing early warning systems to predict students' online learning performance". In: *Computers in Human Behavior* 36 (2014), pp. 469–478. ISSN: 0747-5632. DOI: https://doi.org/10.1016/j.chb.2014.04.002.

[28] J.-L. Hung and K. Zhang. "Revealing online learning behaviors and activity patterns and making predictions with data mining techniques in online teaching". In: *MERLOT Journal of Online Learning and Teaching* (2008).

[29] O. Islam, M. Siddiqui, and N. R. Aljohani. "Identifying Online Profiles of Distance Learning Students Using Data Mining Techniques". In: *Proceedings of the 2019 The 3rd International Conference on Digital Technology in Education*. ICDTE 2019. Association for Computing Machinery, 2019, pp. 115–120. ISBN: 9781450372206. DOI: 10.1145/3369199.3369249.

[30] M. L. King. "The Durbin-Watson test for serial correlation: Bounds for regressions using monthly data". In: *Journal of Econometrics* 21.3 (1983), pp. 357–366. ISSN: 0304-4076. DOI: https://doi.org/10.1016/0304-4076(83)90050-7.

[31] V. Kovanovic et al. "Penetrating the Black Box of Time-on-task Estimation". In: *LAK '15 Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*. 2015, pp. 184–193. DOI: 10.1145/2723576.2723623.

[32] O. M. Kvalheim, B. Grung, and T. Rajalahti. "Number of components and prediction error in partial least squares regression determined by Monte Carlo resampling strategies". In: *Chemometrics and Intelligent Laboratory Systems* 188 (2019), pp. 79–86. ISSN: 0169-7439. DOI: https://doi.org/10.1016/j.chemolab.2019.03.006.

[33] *LAK '11: Proceedings of the 1st International Conference on Learning Analytics and Knowledge*. Association for Computing Machinery, 2011. ISBN: 9781450309448.

[34] L. P. Macfadyen and S. Dawson. “Mining LMS data to develop an “early warning system” for educators: A proof of concept”. In: *Computers & education* 54.2 (2010), pp. 588–599. ISSN: 0360-1315. DOI: https://doi.org/10.1016/j.compedu.2009.09.008.

[35] J. Mamcenko et al. “Analysis of e-exam data using data mining techniques”. In: *Proc of 17th International Conference on Information and Software Technologies (IT 2011), Kaunas, Lithuania*. 2011, pp. 215–219.

[36] T. Mehmood et al. “A review of variable selection methods in Partial Least Squares Regression”. In: *Chemometrics and Intelligent Laboratory Systems* 118 (2012), pp. 62–69. ISSN: 0169-7439. DOI: https://doi.org/10.1016/j.chemolab.2012.07.010.

[37] B. Minaei-Bidgoli et al. “Predicting student performance: an application of data mining methods with an educational Web-based system”. In: *33rd Annual Frontiers in Education, 2003. FIE 2003.* Vol. 1. 2003, T2A–13. DOI: 10.1109/FIE.2003.1263284.

[38] D. C. Montgomery, E. A. Peck, and G. G. Vining. *Introduction to linear regression analysis*. John Wiley & Sons, 2021.

[39] M. Moreno García, S. Segrera, and V. Batista. “Association Rules: Problems, solutions and new applications Abstract”. In: *Actas del III Taller Nacional de Minería de Datos Y Aprendizaje* (2005-01).

[40] L. V. Morris, C. Finnegan, and S.-S. Wu. “Tracking student behavior, persistence, and achievement in online courses”. In: *The Internet and Higher Education* 8.3 (2005), pp. 221–231. ISSN: 1096-7516. DOI: https://doi.org/10.1016/j.iheduc.2005.06.009.

[41] A. Moubayed et al. “Relationship Between Student Engagement and Performance in E-Learning Environment Using Association Rules”. In: *2018 IEEE World Engineering Education Conference (EDUNINE)*. 2018, pp. 1–6. DOI: 10.1109/EDUNINE.2018.8451005.

[42] M. Nocita et al. “Prediction of soil organic carbon content by diffuse reflectance spectroscopy using a local partial least square regression approach”. In: *Soil Biology and Biochemistry* 68 (2014), pp. 337–347. ISSN: 0038-0717. DOI: https://doi.org/10.1016/j.soilbio.2013.10.022.

[43] M. Oliveira, C. Vieira, and I. Vieira. “Modelling demand for higher education: A partial least-squares analysis of Portugal”. In: *European Journal of Higher Education* 5.4 (2015), pp. 388–406. DOI: 10.1080/21568235.2015.1084589.

[44] A. Peña-Ayala. “Educational data mining: A survey and a data mining-based analysis of recent works”. In: *Expert Systems with Applications* 41.4, Part 1 (2014), pp. 1432–1462. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2013.08.042.

[45] D. Pirouz. “An Overview of Partial Least Squares”. In: *SSRN Electronic Journal* (2006-10). DOI: 10.2139/ssrn.1631359.

[46] PORTUGAL. "Decreto-Lei Nº 42/2005, de 22 de fevereiro de 2005. Aprova os princípios reguladores dos instrumentos para a criação do espaço europeu de ensino superior". In: *Diário da República* 37/2005 (2005).

[47] A. C. Rencher and G. B. Schaalje. *Linear Models in Statistics*. John Wiley & Sons, 2008. ISBN: 9780470192603.

[48] C. Romero et al. "Mining Rare Association Rules from e-Learning Data". In: 2010-10, pp. 171–180.

[49] U. Shafique and H. Qaiser. "A comparative study of data mining process models (KDD, CRISP-DM and SEMMA)". In: *International Journal of Innovation and Scientific Research* 12.1 (2014), pp. 2351–8014.

[50] C. V. Silva et al. "Mining retention rules from student transcripts: A case study of the information systems programme at a federal university". In: *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*. Vol. 24. 1. 2013, p. 577. DOI: 10.5753/CBIE.SBIE.2013.577.

[51] Y. Wang and Z. Xu. "Statistical Analysis for Contract Cheating in Chinese Universities". In: *Mathematics* 9.14 (2021). ISSN: 2227-7390. DOI: 10.3390/math9141684.

[52] Y. Yang et al. "Predicting course achievement of university students based on their procrastination behaviour on Moodle". In: *Soft Computing* 24.24 (2020), pp. 18777–18793.

[53] J. S. Yoo, Y.-S. Woo, and S. J. Park. "Mining Course Trajectories of Successful and Failure Students: A Case Study". In: *2017 IEEE International Conference on Big Knowledge (ICBK)*. 2017, pp. 270–275. DOI: 10.1109/ICBK.2017.55.

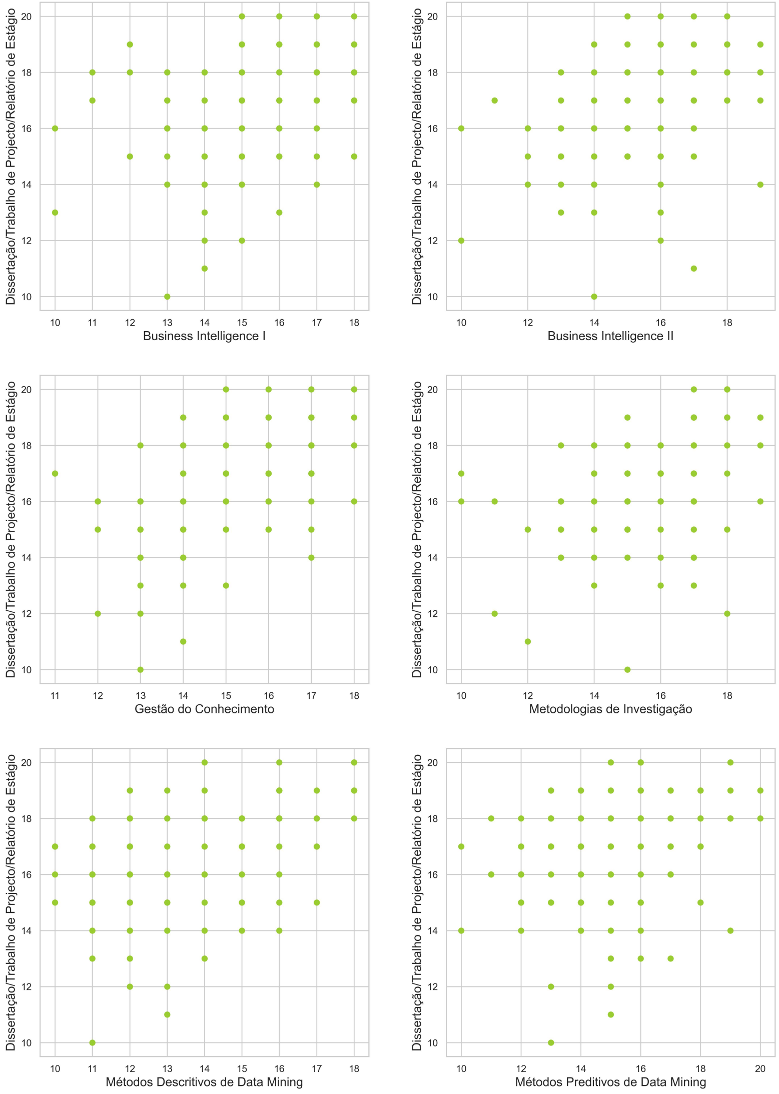
[54] H. Yuliansyah et al. "Discovering Meaningful Pattern of Undergraduate Students Data using Association Rules Mining". In: *Proceedings of the 2019 Ahmad Dahlan International Conference Series on Engineering and Science (ADICS-ES 2019)*. Atlantis Press, 2019, pp. 43–47. ISBN: 978-94-6252-844-4. DOI: https://doi.org/10.2991/adics-es-19.2019.4.

A

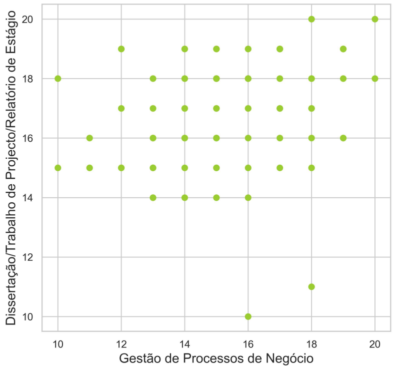
Pairwise Relationships

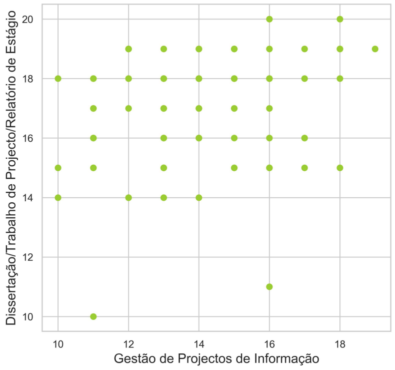
Through the scatterplots, we can see the actual relationship between the predictors with the response variable.

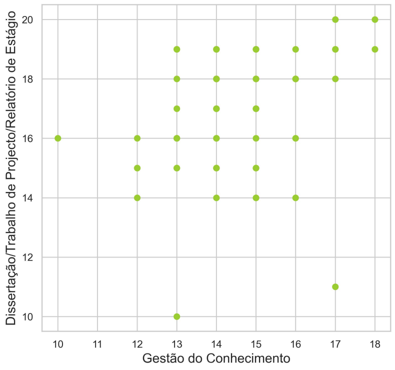
A.1 Specialization in Business Intelligence

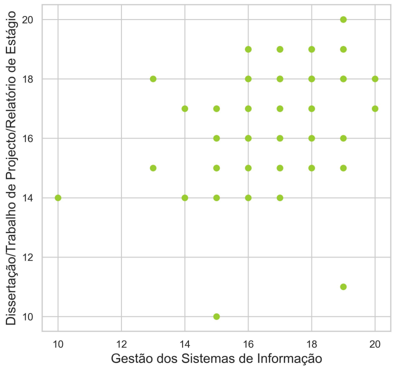


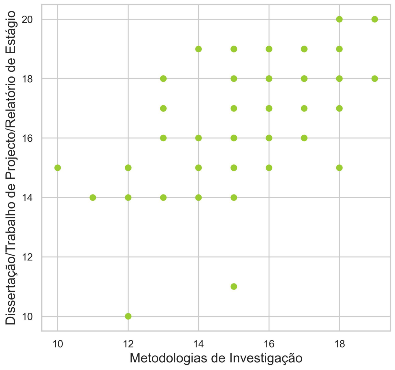
A.2 Specialization in Information Systems and Technologies Management







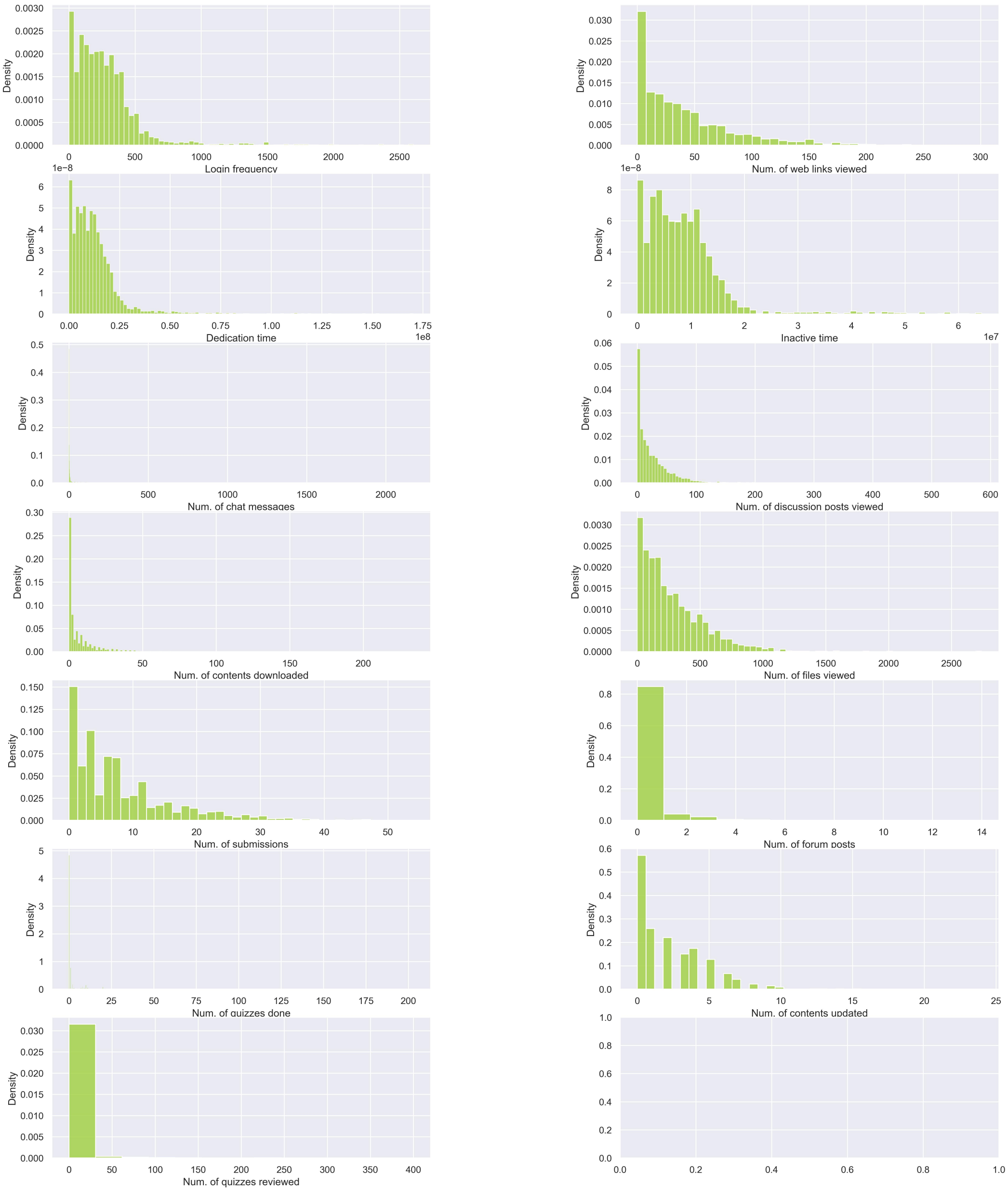




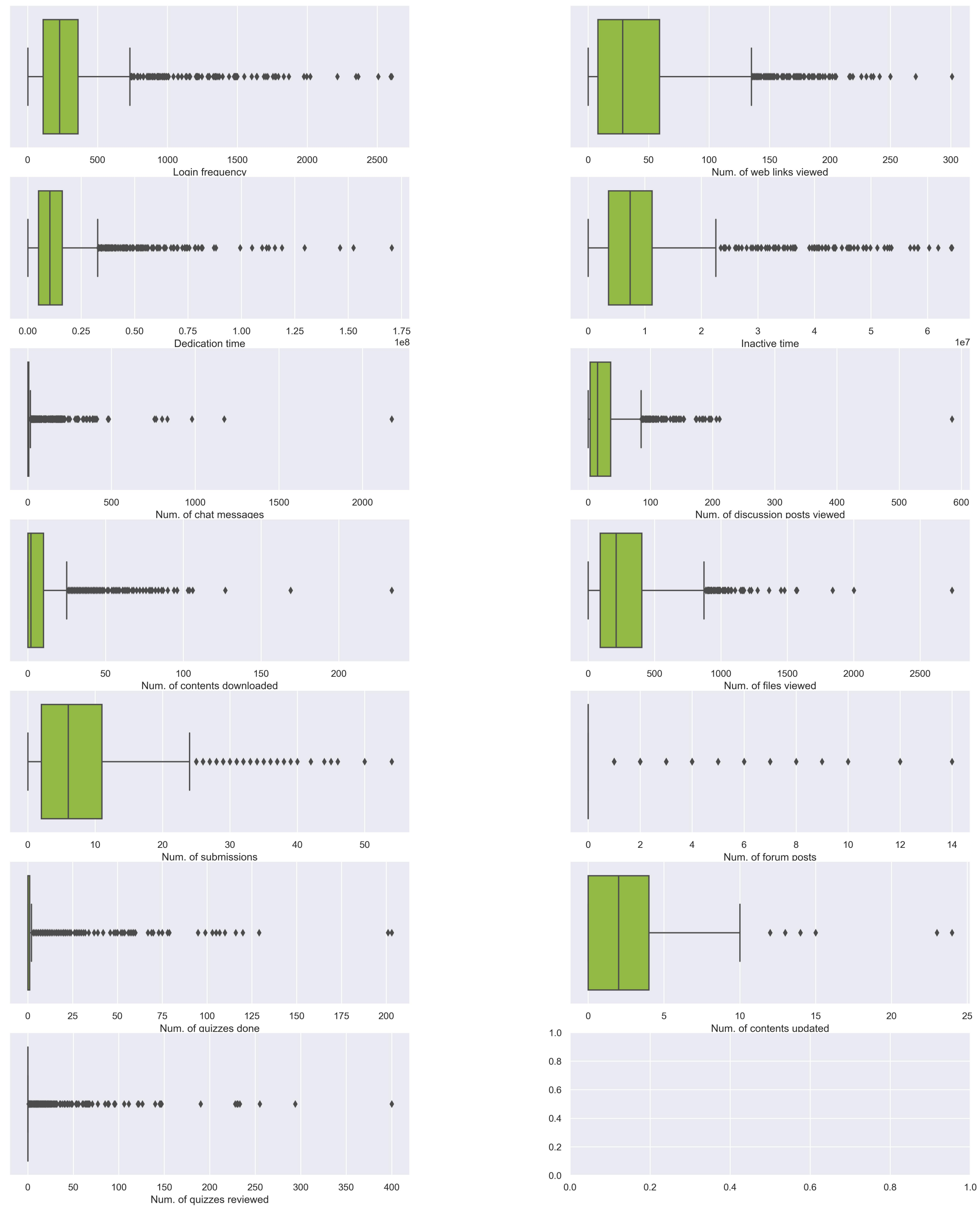
B

Variables' Distribution

B.1 Histograms



B.2 Box Plots

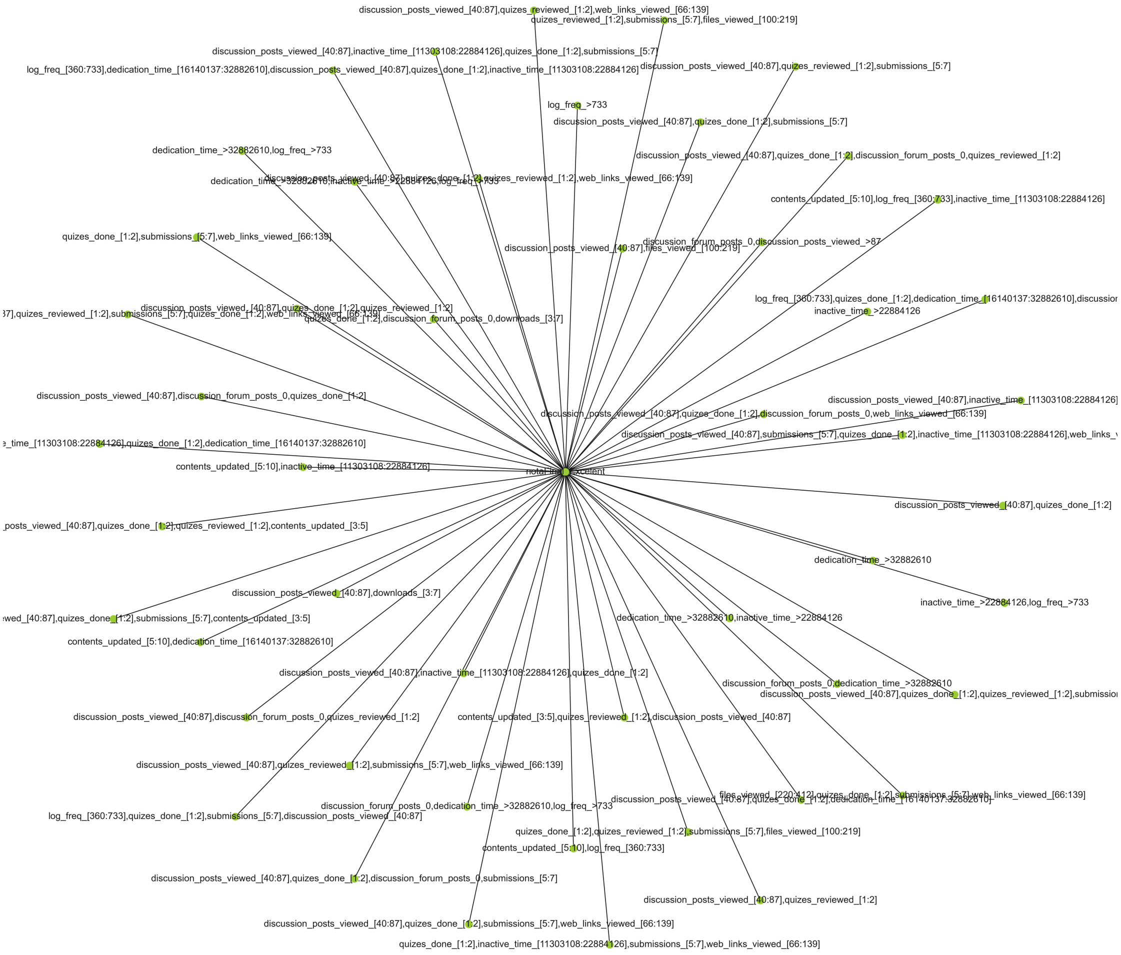

0 500 1000 1500 2000 2500
Login frequency
0 50 100 150 200 250 300
Num. of web links viewed
0.00 0.25 0.50 0.75 1.00 1.25 1.50 1.75
1e8
Dedication time
0 1 2 3 4 5 6
1e7
Inactive time
0 500 1000 1500 2000
Num. of chat messages
0 100 200 300 400 500 600
Num. of discussion posts viewed
0 50 100 150 200
Num. of contents downloaded
0 500 1000 1500 2000 2500
Num. of files viewed
0 10 20 30 40 50
Num. of submissions
0 2 4 6 8 10 12 14
Num. of forum posts
0 25 50 75 100 125 150 175 200
Num. of quizzes done
0 5 10 15 20 25
Num. of contents updated
0 50 100 150 200 250 300 350 400
Num. of quizzes reviewed
0.0 0.2 0.4 0.6 0.8 1.0
0.0 0.2 0.4 0.6 0.8 1.0

C

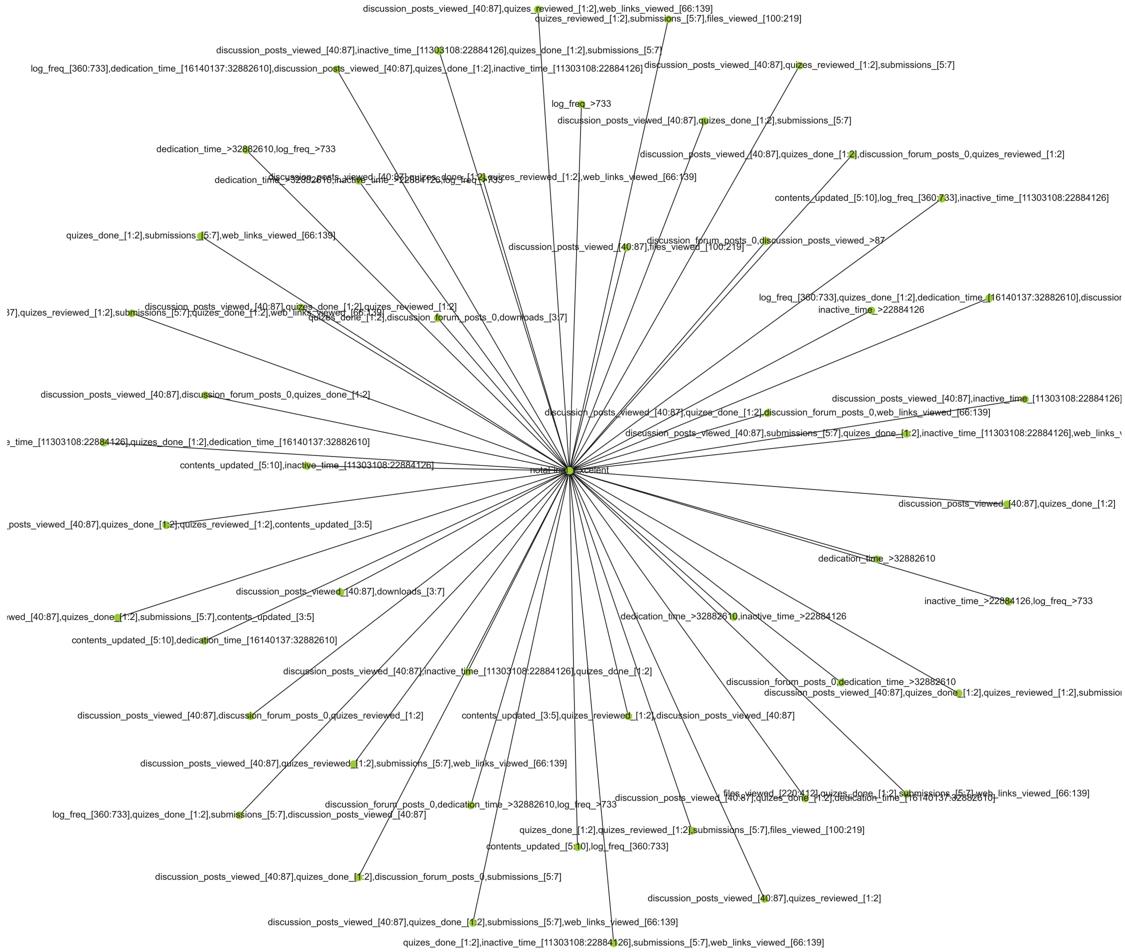
Rules Networks

C.1 Programação para a Ciência de Dados

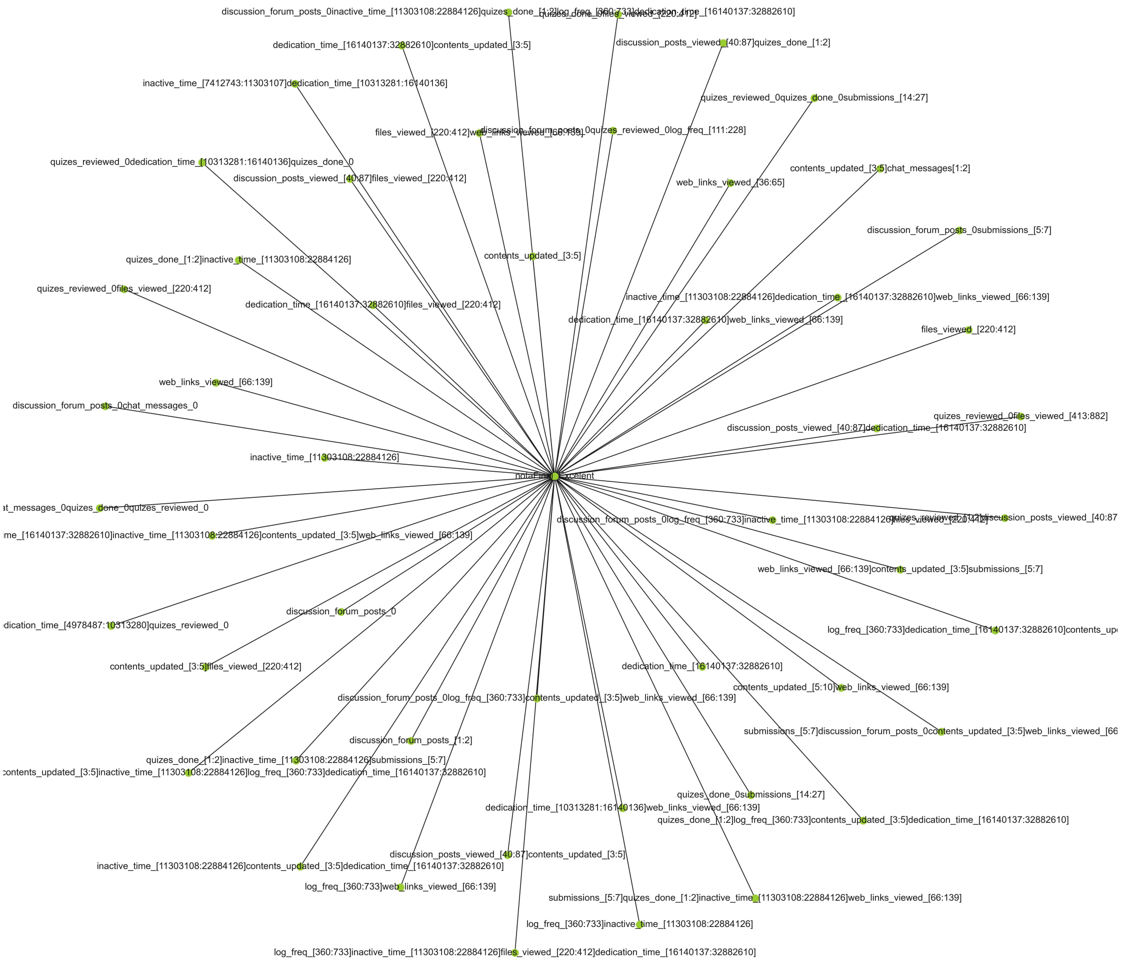
C.1.1 Apriori - Top 50 Confidence Rules


discussion_posts_viewed_[40:87],quizes_reviewed_[1:2],web_links_viewed_[66:139]
quizes_reviewed_[1:2],submissions_[5:7],files_viewed_[100:219]
discussion_posts_viewed_[40:87],inactive_time_[11303108:22884126],quizes_done_[1:2],submissions_[5:7]
log_freq_[360:733],dedication_time_[16140137:32882610],discussion_posts_viewed_[40:87],quizes_done_[1:2],inactive_time_[11303108:22884126]
discussion_posts_viewed_[40:87],quizes_reviewed_[1:2],submissions_[5:7]
log_freq_>733
discussion_posts_viewed_[40:87],quizes_done_[1:2],submissions_[5:7]
dedication_time_>32882610,log_freq_>733
discussion_posts_viewed_[40:87],quizes_done_[1:2],discussion_forum_posts_0,quizes_reviewed_[1:2]
discussion_posts_viewed_[40:87],quizes_done_[1:2],quizes_reviewed_[1:2],web_links_viewed_[66:139]
dedication_time_>32882610,inactive_time_>22884126,log_freq_>733
contents_updated_[5:10],log_freq_[360:733],inactive_time_[11303108:22884126]
quizes_done_[1:2],submissions_[5:7],web_links_viewed_[66:139]
discussion_forum_posts_0,discussion_posts_viewed_>87
discussion_posts_viewed_[40:87],files_viewed_[100:219]
log_freq_[360:733],quizes_done_[1:2],dedication_time_[16140137:32882610],discussior
inactive_time_>22884126
discussion_posts_viewed_[40:87],quizes_done_[1:2],quizes_reviewed_[1:2]
37],quizes_reviewed_[1:2],submissions_[5:7],quizes_done_[1:2],web_links_viewed_[66:139]
quizes_done_[1:2],discussion_forum_posts_0,downloads_[3:7]
discussion_posts_viewed_[40:87],discussion_forum_posts_0,quizes_done_[1:2]
discussion_posts_viewed_[40:87],inactive_time_[11303108:22884126]
discussion_posts_viewed_[40:87],quizes_done_[1:2],discussion_forum_posts_0,web_links_viewed_[66:139]
e_time_[11303108:22884126],quizes_done_[1:2],dedication_time_[16140137:32882610]
discussion_posts_viewed_[40:87],submissions_[5:7],quizes_done_[1:2],inactive_time_[11303108:22884126],web_links_v
contents_updated_[5:10],inactive_time_[11303108:22884126]
notaFinal_excelent
discussion_posts_viewed_[40:87],quizes_done_[1:2]
_posts_viewed_[40:87],quizes_done_[1:2],quizes_reviewed_[1:2],contents_updated_[3:5]
dedication_time_>32882610
discussion_posts_viewed_[40:87],downloads_[3:7]
inactive_time_>22884126,log_freq_>733
ewed_[40:87],quizes_done_[1:2],submissions_[5:7],contents_updated_[3:5]
dedication_time_>32882610,inactive_time_>22884126
contents_updated_[5:10],dedication_time_[16140137:32882610]
discussion_posts_viewed_[40:87],inactive_time_[11303108:22884126],quizes_done_[1:2]
discussion_forum_posts_0,dedication_time_>32882610
discussion_posts_viewed_[40:87],quizes_done_[1:2],quizes_reviewed_[1:2],submissior
discussion_posts_viewed_[40:87],discussion_forum_posts_0,quizes_reviewed_[1:2]
contents_updated_[3:5],quizes_reviewed_[1:2],discussion_posts_viewed_[40:87]
discussion_posts_viewed_[40:87],quizes_reviewed_[1:2],submissions_[5:7],web_links_viewed_[66:139]
files_viewed_[220:412],quizes_done_[1:2],submissions_[5:7],web_links_viewed_[66:139]
discussion_posts_viewed_[40:87],quizes_done_[1:2],dedication_time_[16140137:32882610]
discussion_forum_posts_0,dedication_time_>32882610,log_freq_>733
log_freq_[360:733],quizes_done_[1:2],submissions_[5:7],discussion_posts_viewed_[40:87]
quizes_done_[1:2],quizes_reviewed_[1:2],submissions_[5:7],files_viewed_[100:219]
contents_updated_[5:10],log_freq_[360:733]
discussion_posts_viewed_[40:87],quizes_done_[1:2],discussion_forum_posts_0,submissions_[5:7]
discussion_posts_viewed_[40:87],quizes_reviewed_[1:2]
discussion_posts_viewed_[40:87],quizes_done_[1:2],submissions_[5:7],web_links_viewed_[66:139]
quizes_done_[1:2],inactive_time_[11303108:22884126],submissions_[5:7],web_links_viewed_[66:139]

C.1.2 FP-Growth - Top 50 Confidence Rules

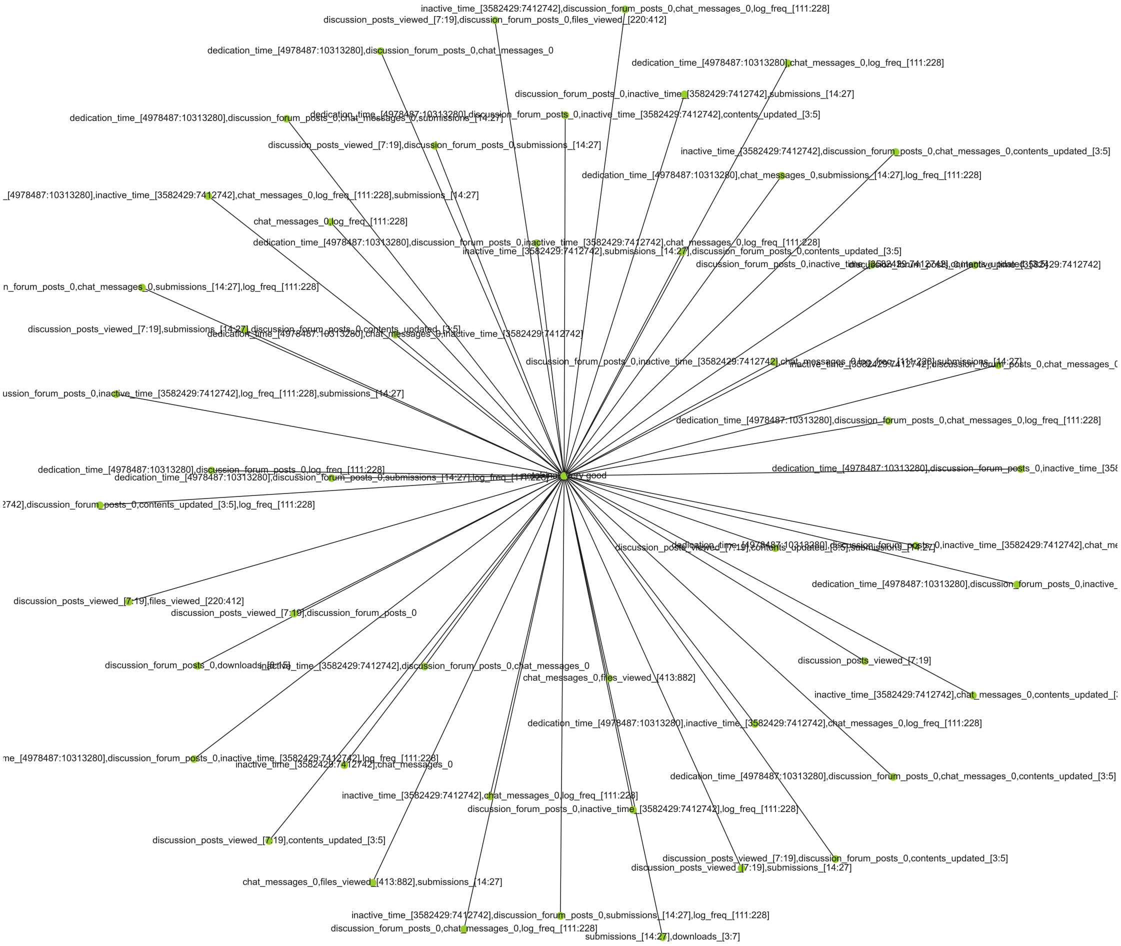

discussion_posts_viewed_[40:87],quizes_reviewed_[1:2],web_links_viewed_[66:139]
quizes_reviewed_[1:2],submissions_[5:7],files_viewed_[100:219]
discussion_posts_viewed_[40:87],inactive_time_[11303108:22884126],quizes_done_[1:2],submissions_[5:7]
log_freq_[360:733],dedication_time_[16140137:32882610],discussion_posts_viewed_[40:87],quizes_done_[1:2],inactive_time_[11303108:22884126]
discussion_posts_viewed_[40:87],quizes_reviewed_[1:2],submissions_[5:7]
log_freq_>733
discussion_posts_viewed_[40:87],quizes_done_[1:2],submissions_[5:7]
dedication_time_>32882610,log_freq_>733
discussion_posts_viewed_[40:87],quizes_done_[1:2],discussion_forum_posts_0,quizes_reviewed_[1:2]
discussion_posts_viewed_[40:87],quizes_done_[1:2],quizes_reviewed_[1:2],web_links_viewed_[66:139]
dedication_time_>32882610,inactive_time_>22884126,log_freq_>733
contents_updated_[5:10],log_freq_[360:733],inactive_time_[11303108:22884126]
quizes_done_[1:2],submissions_[5:7],web_links_viewed_[66:139]
discussion_forum_posts_0,discussion_posts_viewed_>87
discussion_posts_viewed_[40:87],files_viewed_[100:219]
log_freq_[360:733],quizes_done_[1:2],dedication_time_[16140137:32882610],discussion
inactive_time_>22884126
discussion_posts_viewed_[40:87],quizes_done_[1:2],quizes_reviewed_[1:2]
87],quizes_reviewed_[1:2],submissions_[5:7],quizes_done_[1:2],web_links_viewed_[66:139]
quizes_done_[1:2],discussion_forum_posts_0,downloads_[3:7]
discussion_posts_viewed_[40:87],discussion_forum_posts_0,quizes_done_[1:2]
discussion_posts_viewed_[40:87],inactive_time_[11303108:22884126]
discussion_posts_viewed_[40:87],quizes_done_[1:2],discussion_forum_posts_0,web_links_viewed_[66:139]
discussion_posts_viewed_[40:87],submissions_[5:7],quizes_done_[1:2],inactive_time_[11303108:22884126],web_links_
e_time_[11303108:22884126],quizes_done_[1:2],dedication_time_[16140137:32882610]
contents_updated_[5:10],inactive_time_[11303108:22884126]
notaFinal_excelent
discussion_posts_viewed_[40:87],quizes_done_[1:2]
_posts_viewed_[40:87],quizes_done_[1:2],quizes_reviewed_[1:2],contents_updated_[3:5]
dedication_time_>32882610
inactive_time_>22884126,log_freq_>733
discussion_posts_viewed_[40:87],downloads_[3:7]
ewed_[40:87],quizes_done_[1:2],submissions_[5:7],contents_updated_[3:5]
dedication_time_>32882610,inactive_time_>22884126
contents_updated_[5:10],dedication_time_[16140137:32882610]
discussion_posts_viewed_[40:87],inactive_time_[11303108:22884126],quizes_done_[1:2]
discussion_forum_posts_0,dedication_time_>32882610
discussion_posts_viewed_[40:87],quizes_done_[1:2],quizes_reviewed_[1:2],submissio
discussion_posts_viewed_[40:87],discussion_forum_posts_0,quizes_reviewed_[1:2]
contents_updated_[3:5],quizes_reviewed_[1:2],discussion_posts_viewed_[40:87]
discussion_posts_viewed_[40:87],quizes_reviewed_[1:2],submissions_[5:7],web_links_viewed_[66:139]
files_viewed_[220:412],quizes_done_[1:2],submissions_[5:7],web_links_viewed_[66:139]
discussion_forum_posts_0,dedication_time_>32882610,log_freq_>733
discussion_posts_viewed_[40:87],quizes_done_[1:2],dedication_time_[16140137:32882610]
log_freq_[360:733],quizes_done_[1:2],submissions_[5:7],discussion_posts_viewed_[40:87]
quizes_done_[1:2],quizes_reviewed_[1:2],submissions_[5:7],files_viewed_[100:219]
contents_updated_[5:10],log_freq_[360:733]
discussion_posts_viewed_[40:87],quizes_done_[1:2],discussion_forum_posts_0,submissions_[5:7]
discussion_posts_viewed_[40:87],quizes_reviewed_[1:2]
discussion_posts_viewed_[40:87],quizes_done_[1:2],submissions_[5:7],web_links_viewed_[66:139]
quizes_done_[1:2],inactive_time_[11303108:22884126],submissions_[5:7],web_links_viewed_[66:139]

C.1.3 ECLAT - Top 50 Support Rules

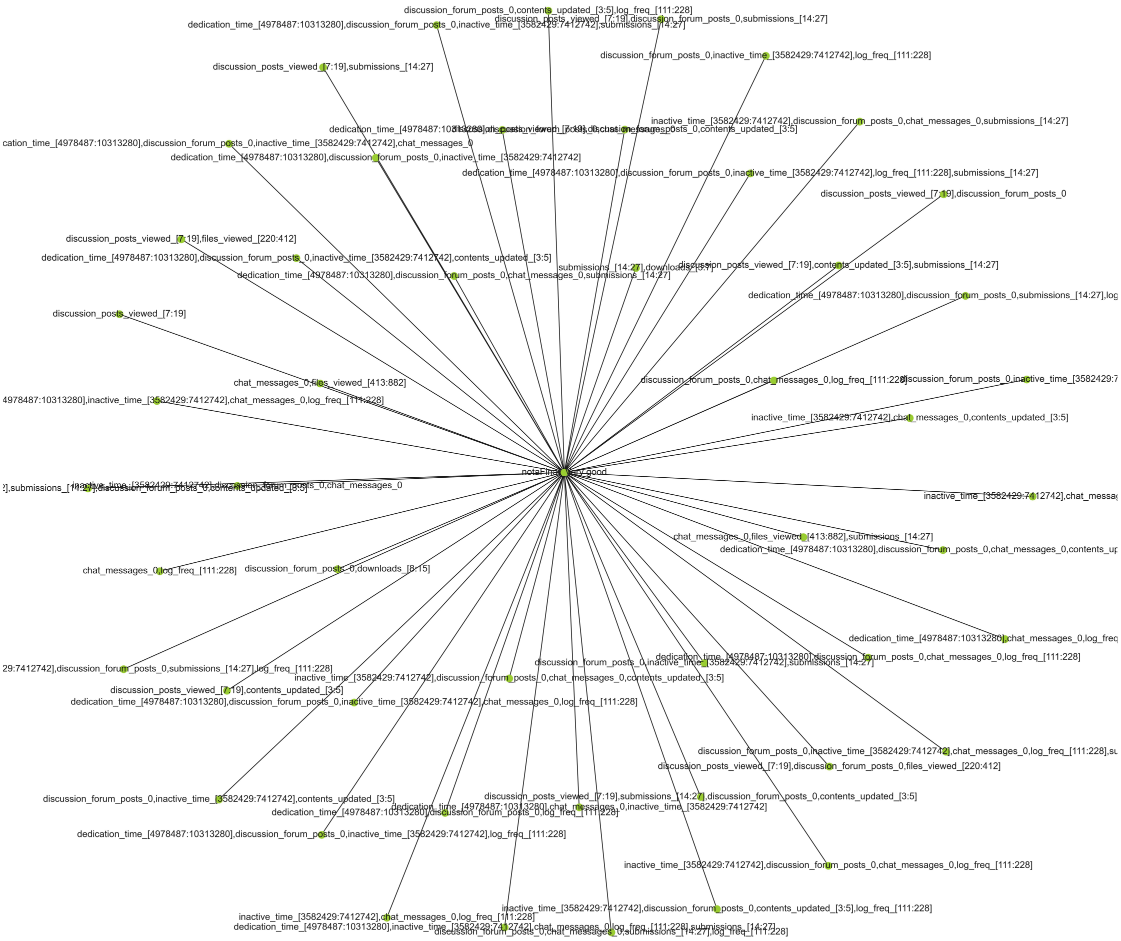

discussion_forum_posts_0inactive_time_[11303108:22884126]quizes_done_[1:2]log_freq_[360:733]dedication_time_[16140137:32882610]
quizes_done_0files_viewed_[220:412]
dedication_time_[16140137:32882610]contents_updated_[3:5]
discussion_posts_viewed_[40:87]quizes_done_[1:2]
inactive_time_[7412743:11303107]dedication_time_[10313281:16140136]
quizes_reviewed_0quizes_done_0submissions_[14:27]
files_viewed_[220:412]web_links_viewed_[66:139]
discussion_forum_posts_0quizes_reviewed_0log_freq_[111:228]
quizes_reviewed_0dedication_time_[10313281:16140136]quizes_done_0
contents_updated_[3:5]chat_messages[1:2]
discussion_posts_viewed_[40:87]files_viewed_[220:412]
web_links_viewed_[36:65]
discussion_forum_posts_0submissions_[5:7]
contents_updated_[3:5]
quizes_done_[1:2]inactive_time_[11303108:22884126]
quizes_reviewed_0files_viewed_[220:412]
inactive_time_[11303108:22884126]dedication_time_[16140137:32882610]web_links_viewed_[66:139]
dedication_time_[16140137:32882610]files_viewed_[220:412]
dedication_time_[16140137:32882610]web_links_viewed_[66:139]
files_viewed_[220:412]
web_links_viewed_[66:139]
discussion_forum_posts_0chat_messages_0
quizes_reviewed_0files_viewed_[413:882]
discussion_posts_viewed_[40:87]dedication_time_[16140137:32882610]
inactive_time_[11303108:22884126]
notaFinal_excelent
at_messages_0quizes_done_0quizes_reviewed_0
discussion_forum_posts_0log_freq_[360:733]inactive_time_[11303108:22884126]files_viewed_[220:412]
quizes_reviewed_[1:2]discussion_posts_viewed_[40:87
ime_[16140137:32882610]inactive_time_[11303108:22884126]contents_updated_[3:5]web_links_viewed_[66:139]
web_links_viewed_[66:139]contents_updated_[3:5]submissions_[5:7]
discussion_forum_posts_0
log_freq_[360:733]dedication_time_[16140137:32882610]contents_up
dication_time_[4978487:10313280]quizes_reviewed_0
contents_updated_[3:5]files_viewed_[220:412]
dedication_time_[16140137:32882610]
contents_updated_[5:10]web_links_viewed_[66:139]
discussion_forum_posts_0log_freq_[360:733]contents_updated_[3:5]web_links_viewed_[66:139]
submissions_[5:7]discussion_forum_posts_0contents_updated_[3:5]web_links_viewed_[66
discussion_forum_posts_[1:2]
quizes_done_[1:2]inactive_time_[11303108:22884126]submissions_[5:7]
contents_updated_[3:5]inactive_time_[11303108:22884126]log_freq_[360:733]dedication_time_[16140137:32882610]
quizes_done_0submissions_[14:27]
dedication_time_[10313281:16140136]web_links_viewed_[66:139]
quizes_done_[1:2]log_freq_[360:733]contents_updated_[3:5]dedication_time_[16140137:32882610]
discussion_posts_viewed_[40:87]contents_updated_[3:5]
inactive_time_[11303108:22884126]contents_updated_[3:5]dedication_time_[16140137:32882610]
log_freq_[360:733]web_links_viewed_[66:139]
submissions_[5:7]quizes_done_[1:2]inactive_time_[11303108:22884126]web_links_viewed_[66:139]
log_freq_[360:733]inactive_time_[11303108:22884126]
log_freq_[360:733]inactive_time_[11303108:22884126]files_viewed_[220:412]dedication_time_[16140137:32882610]

C.2 Aprendizagem Profunda

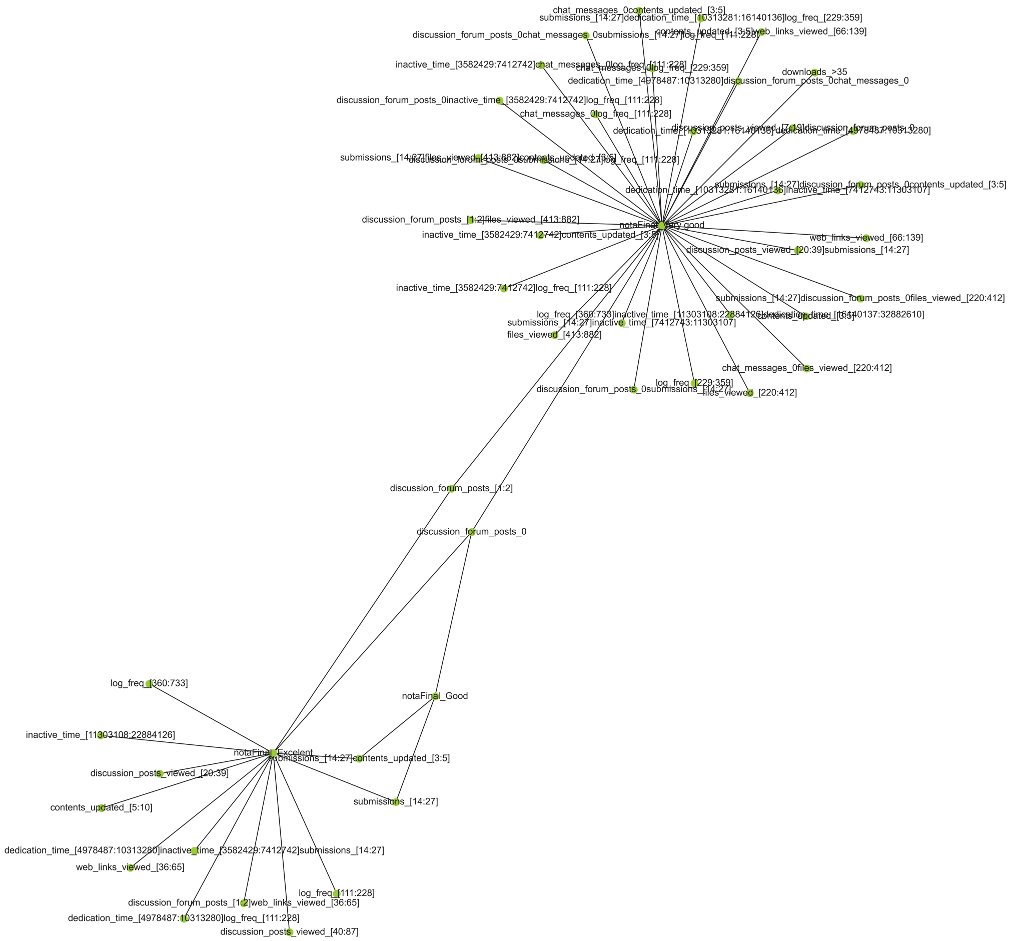
C.2.1 Apriori - Top 50 Confidence Rules


inactive_time_[3582429:7412742],discussion_forum_posts_0,chat_messages_0,log_freq_[111:228]
discussion_posts_viewed_[7:19],discussion_forum_posts_0,files_viewed_[220:412]
dedication_time_[4978487:10313280],discussion_forum_posts_0,chat_messages_0
dedication_time_[4978487:10313280],chat_messages_0,log_freq_[111:228]
discussion_forum_posts_0,inactive_time_[3582429:7412742],submissions_[14:27]
dedication_time_[4978487:10313280],discussion_forum_posts_0,chat_messages_0,submissions_[14:27]
dedication_time_[4978487:10313280],discussion_forum_posts_0,inactive_time_[3582429:7412742],contents_updated_[3:5]
discussion_posts_viewed_[7:19],discussion_forum_posts_0,submissions_[14:27]
inactive_time_[3582429:7412742],discussion_forum_posts_0,chat_messages_0,contents_updated_[3:5]
dedication_time_[4978487:10313280],chat_messages_0,submissions_[14:27],log_freq_[111:228]
_[4978487:10313280],inactive_time_[3582429:7412742],chat_messages_0,log_freq_[111:228],submissions_[14:27]
chat_messages_0,log_freq_[111:228]
dedication_time_[4978487:10313280],discussion_forum_posts_0,inactive_time_[3582429:7412742],chat_messages_0,log_freq_[111:228]
inactive_time_[3582429:7412742],submissions_[14:27],discussion_forum_posts_0,contents_updated_[3:5]
discussion_forum_posts_0,inactive_time_[3582429:7412742],contents_updated_[3:5]
discussion_forum_posts_0,contents_updated_[3:5],inactive_time_[3582429:7412742]
n_forum_posts_0,chat_messages_0,submissions_[14:27],log_freq_[111:228]
discussion_posts_viewed_[7:19],submissions_[14:27],discussion_forum_posts_0,contents_updated_[3:5]
dedication_time_[4978487:10313280],chat_messages_0,inactive_time_[3582429:7412742]
discussion_forum_posts_0,inactive_time_[3582429:7412742],chat_messages_0,log_freq_[111:228],submissions_[14:27]
inactive_time_[3582429:7412742],discussion_forum_posts_0,chat_messages_0
ussion_forum_posts_0,inactive_time_[3582429:7412742],log_freq_[111:228],submissions_[14:27]
dedication_time_[4978487:10313280],discussion_forum_posts_0,chat_messages_0,log_freq_[111:228]
dedication_time_[4978487:10313280],discussion_forum_posts_0,log_freq_[111:228]
dedication_time_[4978487:10313280],discussion_forum_posts_0,submissions_[14:27],log_freq_[111:228]
very good
dedication_time_[4978487:10313280],discussion_forum_posts_0,inactive_time_[358
2742],discussion_forum_posts_0,contents_updated_[3:5],log_freq_[111:228]
dedication_time_[4978487:10313280],discussion_forum_posts_0,inactive_time_[3582429:7412742],chat_me
discussion_posts_viewed_[7:19],contents_updated_[3:5],submissions_[14:27]
dedication_time_[4978487:10313280],discussion_forum_posts_0,inactive_
discussion_posts_viewed_[7:19],files_viewed_[220:412]
discussion_posts_viewed_[7:19],discussion_forum_posts_0
discussion_posts_viewed_[7:19]
discussion_forum_posts_0,downloads_[8:15]
inactive_time_[3582429:7412742],discussion_forum_posts_0,chat_messages_0
chat_messages_0,files_viewed_[413:882]
inactive_time_[3582429:7412742],chat_messages_0,contents_updated_[3
dedication_time_[4978487:10313280],inactive_time_[3582429:7412742],chat_messages_0,log_freq_[111:228]
me_[4978487:10313280],discussion_forum_posts_0,inactive_time_[3582429:7412742],log_freq_[111:228]
inactive_time_[3582429:7412742],chat_messages_0
dedication_time_[4978487:10313280],discussion_forum_posts_0,chat_messages_0,contents_updated_[3:5]
inactive_time_[3582429:7412742],chat_messages_0,log_freq_[111:228]
discussion_forum_posts_0,inactive_time_[3582429:7412742],log_freq_[111:228]
discussion_posts_viewed_[7:19],contents_updated_[3:5]
discussion_posts_viewed_[7:19],discussion_forum_posts_0,contents_updated_[3:5]
discussion_posts_viewed_[7:19],submissions_[14:27]
chat_messages_0,files_viewed_[413:882],submissions_[14:27]
inactive_time_[3582429:7412742],discussion_forum_posts_0,submissions_[14:27],log_freq_[111:228]
discussion_forum_posts_0,chat_messages_0,log_freq_[111:228]
submissions_[14:27],downloads_[3:7]

C.2.2 FP-Growth - Top 50 Confidence Rules


discussion_forum_posts_0,contents_updated_[3:5],log_freq_[111:228]
discussion_posts_viewed_[7:19],discussion_forum_posts_0,submissions_[14:27]
dedication_time_[4978487:10313280],discussion_forum_posts_0,inactive_time_[3582429:7412742],submissions_[14:27]
discussion_forum_posts_0,inactive_time_[3582429:7412742],log_freq_[111:228]
discussion_posts_viewed_[7:19],submissions_[14:27]
inactive_time_[3582429:7412742],discussion_forum_posts_0,chat_messages_0,submissions_[14:27]
dedication_time_[4978487:10313280],discussion_forum_posts_0,chat_messages_0
discussion_posts_viewed_[7:19],discussion_forum_posts_0,contents_updated_[3:5]
cation_time_[4978487:10313280],discussion_forum_posts_0,inactive_time_[3582429:7412742],chat_messages_0
dedication_time_[4978487:10313280],discussion_forum_posts_0,inactive_time_[3582429:7412742]
dedication_time_[4978487:10313280],discussion_forum_posts_0,inactive_time_[3582429:7412742],log_freq_[111:228],submissions_[14:27]
discussion_posts_viewed_[7:19],discussion_forum_posts_0
discussion_posts_viewed_[7:19],files_viewed_[220:412]
dedication_time_[4978487:10313280],discussion_forum_posts_0,inactive_time_[3582429:7412742],contents_updated_[3:5]
submissions_[14:27],downloads_[3:7]
discussion_posts_viewed_[7:19],contents_updated_[3:5],submissions_[14:27]
dedication_time_[4978487:10313280],discussion_forum_posts_0,chat_messages_0,submissions_[14:27]
dedication_time_[4978487:10313280],discussion_forum_posts_0,submissions_[14:27],log
discussion_posts_viewed_[7:19]
chat_messages_0,files_viewed_[413:882]
discussion_forum_posts_0,chat_messages_0,log_freq_[111:228]
discussion_forum_posts_0,inactive_time_[3582429:7
4978487:10313280],inactive_time_[3582429:7412742],chat_messages_0,log_freq_[111:228]
inactive_time_[3582429:7412742],chat_messages_0,contents_updated_[3:5]
notaFinal_very good
2],submissions_[14:27],discussion_forum_posts_0,contents_updated_[3:5]
inactive_time_[3582429:7412742],discussion_forum_posts_0,chat_messages_0
inactive_time_[3582429:7412742],chat_messag
chat_messages_0,files_viewed_[413:882],submissions_[14:27]
dedication_time_[4978487:10313280],discussion_forum_posts_0,chat_messages_0,contents_up
chat_messages_0,log_freq_[111:228]
discussion_forum_posts_0,downloads_[8:15]
dedication_time_[4978487:10313280],chat_messages_0,log_freq
dedication_time_[4978487:10313280],discussion_forum_posts_0,chat_messages_0,log_freq_[111:228]
discussion_forum_posts_0,inactive_time_[3582429:7412742],submissions_[14:27]
29:7412742],discussion_forum_posts_0,submissions_[14:27],log_freq_[111:228]
inactive_time_[3582429:7412742],discussion_forum_posts_0,chat_messages_0,contents_updated_[3:5]
discussion_posts_viewed_[7:19],contents_updated_[3:5]
dedication_time_[4978487:10313280],discussion_forum_posts_0,inactive_time_[3582429:7412742],chat_messages_0,log_freq_[111:228]
discussion_forum_posts_0,inactive_time_[3582429:7412742],chat_messages_0,log_freq_[111:228],su
discussion_posts_viewed_[7:19],discussion_forum_posts_0,files_viewed_[220:412]
discussion_forum_posts_0,inactive_time_[3582429:7412742],contents_updated_[3:5]
discussion_posts_viewed_[7:19],submissions_[14:27],discussion_forum_posts_0,contents_updated_[3:5]
dedication_time_[4978487:10313280],chat_messages_0,inactive_time_[3582429:7412742]
dedication_time_[4978487:10313280],discussion_forum_posts_0,log_freq_[111:228]
dedication_time_[4978487:10313280],discussion_forum_posts_0,inactive_time_[3582429:7412742],log_freq_[111:228]
inactive_time_[3582429:7412742],discussion_forum_posts_0,chat_messages_0,log_freq_[111:228]
inactive_time_[3582429:7412742],discussion_forum_posts_0,contents_updated_[3:5],log_freq_[111:228]
inactive_time_[3582429:7412742],chat_messages_0,log_freq_[111:228]
dedication_time_[4978487:10313280],inactive_time_[3582429:7412742],chat_messages_0,log_freq_[111:228],submissions_[14:27]
discussion_forum_posts_0,chat_messages_0,submissions_[14:27],log_freq_[111:228]

C.2.3 ECLAT - Top 50 Support Rules


chat_messages_0contents_updated_[3:5]
submissions_[14:27]dedication_time_[10313281:16140136]log_freq_[229:359]
discussion_forum_posts_0chat_messages_0submissions_[14:27]log_freq_[111:228]
contents_updated_[3:5]web_links_viewed_[66:139]
inactive_time_[3582429:7412742]chat_messages_0log_freq_[111:228]
chat_messages_0log_freq_[229:359]
downloads_>35
dedication_time_[4978487:10313280]discussion_forum_posts_0chat_messages_0
discussion_forum_posts_0inactive_time_[3582429:7412742]log_freq_[111:228]
chat_messages_0log_freq_[111:228]
discussion_posts_viewed_[7:19]discussion_forum_posts_0
dedication_time_[10313281:16140136]dedication_time_[4978487:10313280]
submissions_[14:27]files_viewed_[413:882]contents_updated_[3:5]
discussion_forum_posts_0submissions_[14:27]log_freq_[111:228]
submissions_[14:27]discussion_forum_posts_0contents_updated_[3:5]
dedication_time_[10313281:16140136]inactive_time_[7412743:11303107]
discussion_forum_posts_[1:2]files_viewed_[413:882]
notaFinal_Very good
inactive_time_[3582429:7412742]contents_updated_[3:5]
web_links_viewed_[66:139]
discussion_posts_viewed_[20:39]submissions_[14:27]
inactive_time_[3582429:7412742]log_freq_[111:228]
submissions_[14:27]discussion_forum_posts_0files_viewed_[220:412]
log_freq_[360:733]inactive_time_[11303108:22884126]dedication_time_[16140137:32882610]
contents_updated_[3:5]
submissions_[14:27]inactive_time_[7412743:11303107]
files_viewed_[413:882]
chat_messages_0files_viewed_[220:412]
log_freq_[229:359]
discussion_forum_posts_0submissions_[14:27]
files_viewed_[220:412]
discussion_forum_posts_[1:2]
discussion_forum_posts_0
log_freq_[360:733]
notaFinal_Good
inactive_time_[11303108:22884126]
notaFinal_Excelent
submissions_[14:27]contents_updated_[3:5]
discussion_posts_viewed_[20:39]
contents_updated_[5:10]
submissions_[14:27]
dedication_time_[4978487:10313280]inactive_time_[3582429:7412742]submissions_[14:27]
web_links_viewed_[36:65]
log_freq_[111:228]
discussion_forum_posts_[1:2]web_links_viewed_[36:65]
dedication_time_[4978487:10313280]log_freq_[111:228]
discussion_posts_viewed_[40:87]

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa