A data-driven approach to predict first-year students' academic success in Higher Education Institutions

Abstract

This study presents a data mining approach to predict academic success of the first-year students. A dataset of 10 academic years for first-year bachelor's degrees from a Portuguese Higher Institution (N = 9652) has been analysed. Features' selection resulted in a characterising set of 68 features, encompassing socio-demographic, social origin, previous education, special statutes and educational path dimensions. We proposed and tested three distinct course stage data models based on entrance date, end of the first and second curricular semesters. A support vector machines (SVM) model achieved the best overall performance and was selected to conduct a data-based sensitivity analysis. The previous evaluation performance, study gaps and age-related features play a major role in explaining failures at entrance stage. For subsequent stages, current evaluation performance features unveil their predictive power. Suggested guidelines include to provide study support groups to risk profiles and to create monitoring frameworks. From a practical standpoint, a data-driven decision-making framework based on these models can be used to promote academic success.

Keywords

Academic success; data mining; higher education; modelling; SVM; sensitivity analysis.

Introduction

Research areas, such as higher education, are expanding their interest in extracting meaningful and more complex knowledge from their data sources (Koedinger et al., 2008). Recently, a research area that combines Data Mining (DM) and education has emerged and consolidated. Educational Data Mining (EDM) is a field that explores DM applied on different types of educational data (Howard et al., 2016). EDM uses data mainly obtained from educational

information systems to unfold knowledge and find answers to questions and problems concerning the education system.

This study aims to apply data mining techniques to an academic data set provided by a Portuguese Higher Institution, and present meaningful information to increase academic success rate. The resulting models' performance is evaluated and its suitability to predict potential success and failure cases are scrutinized. To achieve the predictors for academic success in the first-year we implemented CRoss-Industry Standard Process for Data Mining (CRISP-DM). This methodology defines a project as a cyclic process and applies a non-rigid sequence of six main stages (Chapman et al, 2000). At the end of this process a knowledge extraction process is conducted, and the collected insights used to formulate guidelines and suggestions regarding institutional policies and pedagogical approaches to improve academic success. On an institutional and management level, the suggested guidelines are expected to leverage decision-making, optimize allocation of educational resources and increase overall institutional performance.

Background

The concept of academic success, which is pivotal to an analytical tool for assessing the quality of HEIs, has several problems in its definition and, consequently, in its operationalization (York et al, 2015). This concept has a myriad of meanings and very diverse uses, depending on the various scientific approaches, but also on its recognition in the various systems and public policies of higher education systems, and the practices and cultures prevailing in educational institutions. From the point of view of its observation, our perspective can be placed at several levels of analysis (cf. Costa and Lopes, 2011): at the level of the performance of higher education systems (in a macro or structural perspective), institutional (where the present study is located) or individual paths of success (in a biographical perspective). Success can also be

interpreted from the point of view of learning and acquired competences and skills, the persistence and the achievement of degrees or certifications (this is the measure of success that we can analyse), students' engagement in academic activities, or even the possibility of having a better and more qualified entry into the labour market (among other social opportunities) (York et al., 2015). Academic success concept is being applied as a definition base that aggregates a multiple number of student and institutional outcomes in students in all grade levels (e.g. Guo, Zhou and Feng, 2018; Pace, Alperb, Burchinal, Golinkoff and Hirsh-Pasek, 2019). Success, in conceptual terms, remains relevant in its appeal and motivation for attainment or achievement of a goal (Hannon et al., 2017). The Astin model first proposed in 1991 (Astin, 2012) clearly identifies academic success as an outcome of input factors and the environment. The model also suggests that the environment functions as a mediator. However, the relationship between environment and student outcomes cannot be understood without considering student inputs. According to Tinto (2006), students enter Higher Educational Institutions [HEI] with a variety of abilities, skills, levels of high education preparation, attributes, specifically with differences on social class, age, gender, attitudes, values and knowledge about higher education. At the same time, students participate in external commitments, such as family, work and community. These set of features is being used as root to correlation and patterns studies regarding academic success.

Pascarella and Terenzini (2005) refine Astin's framework by explaining higher education outcomes as functions of three sets of elements: inputs, environment and outputs. The inputs are composed by demographic characteristics, family backgrounds, academic and social experiences. The environment encompasses people, programs, policies, cultures, and experiences that students encounter in HEI.

The first-year student's achievement is predictive for subsequent years, as seen in Brouwer, Jansen, Flash and Hofman's (2016) study, so the students must be supported early. For HEI to

be able to create the most appropriate support for students, it is necessary to understand which factors predict the academic success. An approach with data mining techniques will allow to achieve this goal. According to Romero and Ventura (2007), the introduction of DM techniques in academic domains could improve decision-making processes in higher learning institutions. This improvement is expected to promote student's retention, transition rate and academic success. EDM is a field that explores data-mining approaches and techniques on different types of educational data, aiming at solving problems within the educational context (Baker and Yacef, 2009). It concerns to better understand students and the settings in which they learn (Baker, 2010). Over the years, students' enrolment and practicing in HEI has generated huge sets of student related data that may reflect the efficiency of the learning process (Koedinger et al., 2008). Converting raw data originated by educational systems into useful information can potentially have a great impact on educational research and practice. Regardless of the origin, all DM techniques show one common characteristic: automated discovery of new relations and dependencies between attributes in the observed data.

Through this research effort it was possible to infer that academic success' modelling is significantly affected by diverse factors, such as, higher educational context, educational system and its specificities, available data and its quality. Other aspects such as problem and modelling decisions lead to distinct operationalization of success and how it is measured. Regarding datasets there is a great diversity in terms of source, nature and volume (Khan et al., 2017). The data sources are mostly originated through surveys to students and/or form the HEI database. It is possible to label the reviewed features in five distinct clustering groups: socio-demographic features, social origin features, educational path features, previous education features and special statute features. Regarding student's success operationalizations, the following main definitions have been reviewed: passing grade in a specific module or course, passing grade in a specific exam, passing grade point average, student's graduation and student's

graduation with no failures. A wide spectrum of relevant explanatory features is observed, as there is a large number of distinct features pointed as the most relevant in the literature depending on each study's characteristics. There is no standard in the used datasets, as each study relies on distinct sources. Even so, the following features' groups showed great impact on multiple studies: previous education features, educational path features and socio-demographic features.

Methodology and Methods

We adopted the CRISP-DM (Cross Industry Standard Process for Data Mining) methodology which is particularly suited for data-driven research projects as it was demonstrated by Moro et al. (2011), with several recent data-driven studies adopting it (e.g., Almahadeen et al., 2017). Such approach consists in an iterative sequence of six phases (business understanding, data understanding, data preparation, modeling, evaluation, deployment) involved in a cycle with the goal to tune the final result, i.e., the data model capability to adequately model a problem according to evaluation metrics. The CRISP-DM is not prescriptive, rather it suggests a sequence flow. Therefore, the methodology is flexible, although usually the six phases tend to be followed. In research, usually the deployment phase is replaced by knowledge extraction to understand a given problem (Moro et al., 2011). All the experiments were implemented using the R statistical tool and the "rminer" package (Cortez, 2010). In the next subsections, we detail the main tasks implemented to address each CRISP-DM phase.

Business Understanding

The studied institution is a public HEI located in Portugal, where the socioeconomic status constitutes a high impact on academic success of the students (Mestre and Baptista, 2016). This relationship is also verified in other countries, as reported in Sirin (2005), Ingram (2011) and Brouwer et al. (2016), who have found that the financial and social capital of families have a

high influence on students' academic success, even at the college. All the used data was anonymized and its use for this research was made under a confidentiality agreement signed by the institution's Data Protection Office representative and all the authors.

All data used in our study was extracted from the institution's information system. Therefore, the earlier in the academic path the DM model could predict failures by avoiding a high level of incorrectly predicted cases (false positives), the better. This study adopts student's graduation with no failures as student's success operationalization. Thus, DM's goal and the main analysis' subject are devoted to predicting students that would not complete their degree's programme within the optimal number of curricular years. In other words, students that fails and/or repeats at least one curricular year. This study follows a classification DM approach, as it builds a predictive model that classifies a data record into one of two predefined classes. Predefined classes used for success are "Failure" and "Success".

Data Understanding

The original data collected from the institutional database included bachelor students' records who effectively enrolled a programme provided by the institution between 2006/2007 and 2015/2016 (10 years' timeframe). This ensures success operationalization requirements to be met. An analytical base table (ABT) was created to collect features' candidates, originally spread among distinct tables of the relational database model. No data collected after the first curricular year were considered for feature's gathering purposes because the first-year student's academic success is predictive for later years (Brouwer et al., 2016), as it concentrates in freshmen. Thus, thirty-two directly extracted features were added to the ABT in first instance. Additionally, fifty-four derived features total were designed for student statutes and social services. Social services features were also split in two categories: accepted and requested, adding extra detail to the analysis. Further data understanding effort exposed potential features based on pre-existing data. According to Authors (2019), feature engineering is key for data

mining. Therefore, eighteen new computed features were designed applying non-straightforward logic, requiring distinct transformation, aggregation or/and calculation processes. For instance, five new computed features were designed for candidacy preference considering the relationship between student's preference, HEI and degree student ended up registering and entry exams grades average. On its turn, new computed evaluation related features were designed, comprising overall evaluation by each semester of the first curricular year. Thus, six new computed educational path features were designed for student's evaluation. It is important to detail that weighted average features were calculated relying on the premise that 30 ECTS are the optimum amount of ECTS to be collected per semester. Additional computed features representing, student's age at entry, study gap time between precedent and current educational degree, and student's residence location were also developed. Table 1 depicts the description and resultant classes of each of the one-hundred and four features gathered at this stage.

Table 1.

Features' description and classes

Data Preparation

Data preparation stage requires to take decisions on final features' set, establishing the foundation for modelling. Five distinct approaches were applied at this stage. The first approach was based in data generalization (through replacing low level attributes with high level concepts). A conceptual review process was carried out to design a meaningful higher aggregation level, to deal with several features' low quality, setting bases for appropriate

modelling. For instance, six distinct classes were designed, taking ESCO[1] (European Skills, Competences, Qualifications and Occupations) multilingual classification of occupations, for parent's occupations, one of the indicators for socioeconomic status (Costa et al, 2002 and Smith and Lynch, 2004). The second approach consisted in dealing with missing data' features. Hotdeck imputation algorithm (k-nearest neighbour) was applied to some feature, while for the remaining missing data' features, a 1% threshold was set up for decision taking (input with "unknow" value or exclude). The third approach consisted in reviewing dependencies between the DM goal and each feature. For instance, partial-time students are unable to meet operationalized success conditions, so, all records, which *partialTimeStudentAtEntry* is true were excluded. The fourth approach consisted in removing single class features. A clear example is *degreeType feature*, that due to this proposed scope, is only represented by a single class: bachelor. The fifth and final approach is based on outliers and conflicting data' features. At the end of this process further CRISP-DM iteration based on features' selection tuning decided on the imputed values' features. Table 2 summarizes the final ABT by features' group, data type and collection time, as the result of data preparation stage. Final dataset is composed by a total 9652 records for regular bachelor's degrees and 789 records for 4-year bachelor's degrees. A total of 68 features are represented, 36 special statute features, 12 education path features, 6 previous education features, 10 socio-demographic features and 4 social origin features.

The population represented in the dataset consists in 48.1% male and 51.9% female. The average age of the students is 20.1 years, with a standard deviation of 5.3 years. As expected for these ages, most of the students are single (95.8%), with the remaining consisting in 2.4% married, 0.8% divorced, 0.1% widowed, and 0.9% unknown.

[1] ESCO is a Europe 2020 initiative, the current version is ESCO v1.0.3 (Last update 26/04/2018). DG Employment, Social Affairs and Inclusion of the European Commission developed ESCO in collaboration with stakeholders and with the European Centre for the Development of Vocational Training (Cedefop).

Table 2

Final ABT for DM modelling purposes.

Modelling

Considering nature and structure of the final ABT and the techniques that produced best results in the related works, supported by Shahiri et al. (2015) analysis, we decided to develop models based on the following four techniques: Decision Trees (DT) (Apté and Weiss, 1997), Random Forests (RF) (Breiman, 2001), Support Vector Machines (SVM) (Cortes and Vapnik, 1995) and Artificial Neural Networks (ANN) (Haykin, 1994). Rminer provides the *mining* function which we applied using the following setup: RPART (Recursive Partitioning and Regression Trees), DT and CTREE (Conditional Inference Trees) (distinct DT algorithms), RF, SVM, and MLPE (multilayer perceptron), as ANN representative. Models' training plan is based on k-fold cross-validation method (Trevor, Robert and Friedman, 2009). The k parameter was set to 10 (k=10), as per the most recent related works' guidelines. Each DM model analysis is submitted to 20 runs in order to enhance results' robustness.

Results

Evaluation

Receiver Operating Characteristic (ROC) curve and the corresponding Area Under the ROC Curve (AUC) (Bradley, 1997), based on the confusion matrix (Kohavi and Provost, 1998), were used for measuring purposes. Three main models are evaluated, one for each data collection time, entrance, end of the first curricular semester and end of the second curricular semester. Each model relies on a distinct number of features depending on collection time it is based on. The 4-year degrees' model is additionally evaluated at this stage

The first model being evaluated are composed by 30 features collected at entrance. This model is henceforth referred as DM_Entrance. Table 3 depicts SVM, as the best predictive model (AUC slightly higher than 0.77). RF model also demonstrates a considerable predictive result surpassing 0.76, while MLPE model almost reaches 0.75. CTREE achieves the best result by far within the decision tree model, even performing considerably worse than the previous models.

Table 3

AUC results for DM_Entrance model.

Figure 1 shows the ROC curve for CTREE, as DT's representative, SVM, RF and MLPE. It is possible to observe that SVM curve achieves higher TPR (True Positive Rate) values along the entire FPR-axis (False Positive Rate). SVM model proves its higher discriminatory capacity, outperforming remaining models for the whole cut-off probability's range. The points highlighted in the graphic represent a threshold value of 50%, for each model's curve.

Figure 1. ROC curves for DM_Entrance model.

Table 4 details 50% threshold analysis through confusion matrices and resulting sensitivity and 1-specificity values for DM_Entrance models.

Table 4

Confusion matrices for DM_Entrance model.

DM_EntryYear1Sem model establishes the basis for succeeding collection time model, summing up to 44 features. Table 5 demonstrates a huge predictive performance boost

compared to DM_Entrance model's results (greater or equal than 13%). Once more SVM and RF achieve the best AUC results, surpassing 0.90.

Table 5

AUC results for DM_EntryYear1Sem model

Figure 2 shows the ROC curves for DM_EntryYear1Sem models' performance analysis. RF curve clearly intersects SVM curve for an FPR close to 0.5. SVM slightly achieves better performance for lower values of FPR, while RF is slightly better above that value. Threshold values of 50% and 30%, for each model's curve are highlighted in the figure. 30% threshold represents an optimized TPR/FPR trade-off.

Figure 2. ROC curves for DM_EntryYear1Sem model.

Table 6 details 30% threshold analysis through confusion matrices and resulting sensitivity and 1-specificity values for DM_EntryYear1Sem model.

Table 6

Confusion matrix for DM_EntryYear1Sem model.

DM_EntryYear2Sem model relies on the whole set of features collected by the end of the second curricular semester (68 features). Table 7 shows DM_EntryYear2Sem model's increased predictive performance results. Newly included features allowed SVM and RF models to reach, approximately, 0.94. The discriminatory capacity of the whole features' set, at this point, is so robust that distinct models' performance results tend to converge.

Table 7

AUC results for DM_EntryYear2Sem model.

Figure 3 shows ROC analysis for DM_EntryYear2Sem models. SVM achieves the best performance for an FPR below 0.3. RF intersects SVM around that value, outperforming it for above values. 20% threshold value was scrutinized, following same threshold selection reasoning applied previously.

Figure 3. ROC curves for DM_EntryYear2Sem model.

Table 8 details 20% threshold analysis through confusion matrix for DM_EntryYear2Sem model. At this point the models' sensitivity is so high, that special attention is given to 1-specificity review. So, comparing RF and SVM, RF achieves a slightly better sensitivity while SVM achieves a reduced and considerably better 1-specificity results.

Table 8

Confusion matrices for DM_EntryYear2Sem model.

Knowledge Extraction

Sensitivity analysis (SA) method described by Cortez and Embrecht (2011) was adopted to perform feature's relevance analysis based on SVM resultant models. Specifically, the data-based sensitivity analysis algorithm (DSA) is selected among others, as it induces several features values to be changed simultaneously, allowing interactions between input features to be detected.

Figure 4 shows the relevance for the most impacting features in DM_Entrance model. Feature's relevance is measured through its contribution percentage to the output. Each of the illustrated features, 8 out of 30, demonstrates great relevance, above 5%. Their combined contribution to the model surpasses 63%.

Figure 4. Features' relevance for DM_Entrance model.

Reviewing high impact features on its characteristics, it is noticeable that all features' groups are represented, except social origin.

Even submitted to an imputation process, during data preparation stage, *entryGradeHotdeck* feature keep its prominent importance and shows the highest relevance. The detailed influence of *entryGradeHotDeck* feature is depicted in Figure 5.

Figure 5. Impact of entryGradeHotdeck on DM_Entrance model.

This feature quantifies secondary or high school evaluation performance, so as highlighted in Tinto (1999), high school evaluation performance provides insight into potential academic performance of the freshmen. Previous education features are commonly pointed out as relevant predictors of academic success. Related studies, such as, Osmanbegović and Suljić (2012), Goker et al. (2013), Trstenjak and Donko (2014) and Asif et al. (2017), present previous evaluation related features as the most impacting features on their models. As per Trstenjak and Donko (2014), great part of socio-demographic and social origin features doesn't change over time, having previously influenced secondary school evaluation performance. This helps explaining the leveraged relevance of *entryGradeHotDeck* feature in the model. The initial perception regarding previous student's performance is confirmed, as lower *entryGradeHotdeck* values, presents a much stronger contribution to failure, especially for entry grade values below 13.

The second most impacting feature is *studyGapYears*, it is quite an interesting finding, since no similar feature is found in related works' models. Figure 6 shows no significant impact for *studyGapYears* values below 10.

Figure 6. Impact of studyGapYears on DM_Entrance model.

Even so several years' gap shows slightly inferior impact than gap's absence. For gaps above 10 years, a prominent influence is verified.

The third most impacting feature is *yearOfBirth*, registering a similar contribution percentage. In order to review its impact illustrated in Figure 7, it is important to remind that original dataset was trimmed to enrolments between 2006/2007 and 2015/2016.

Figure 7. Impact of yearOfBirth on DM_Entrance model.

In general terms, yearOfBirth show considerable to high contribution to failure for values below 1990. This impact trend demonstrates that failure is higher among older students, as most of these cases represent students that enrolled in later life stages. These findings follow indications presented in Natek and Zwilling (2014), Authors (2017) and Fernandes et al. (2018).

Following DSA is based on DM_EntryYear1Sem model. Figure 8 shows the relevance of the 8 most impacting features in the model. Several features, collected at the end of first curricular semester, showed great impact, placing 4 features among the 8 most relevant. This impact confirms the directions discussed in model's evaluation. The combined contribution of these 8 most impacting features is close to 65%.

Figure 8. Features' relevance for DM_EntryYear1Sem model.

The two higher relevance features are educational path features' group representatives. Particularly, they represent first curricular semester evaluation' information, achieving a combined relevance greater than 30%. These feature's relevance supports the findings presented in Martins et al. (2018), that relying on the same educational system, observed similar results for these features. Other studies, such as, Mishra et al. (2014), Slim et al. (2014), Zimmermann et al. (2015) and Asif et al. (2017) demonstrate similar level of impact for equivalent features in their models. Asif et al., (2017) points two groups of academic students according to their performance, high-performing students and low-achieving students and claim that many students tend to stay in the same kind of groups for all academic path. This standpoint may provide some insight regarding these features' great impact in the model. Figures 9 and 10 demonstrate that the lower their values (worst evaluation performance), the stronger their contribution to academic failure.

Figure 9. Impact of weightedAverageGradeEntryYear1stSem on DM_EntryYear1Sem model.

Figure 10. Impact of ectsCreditsEntryYear1stSem on DM_EntryYear1Sem model.

Figure 11 shows the relevance of the 8 most importance features in DM_EntryYear2Sem model. The combined contribution of the 8 most impacting features in the model is approximately 63%.

Figure 11. Features' relevance for DM_EntryYear2Sem model.

Following DM_EntryYear1Sem model's trend, most recent evaluation-related features are the most important features. These features' relevance is aligned with Zimmermann et al. (2015) insights regarding the higher impact of most recent evaluation performances over the academic path. Despite showing equivalent trend, compared to most recent evaluation-related features in DM_EntryYear1Sem model, a slightly lower contribution is verified. This can be explained by the fact that second semester evaluation-related features share their importance with first semester evaluation-related features in DM_EntryYear2Sem model, and by the greater number of features in this model.

In contrast to DM_EntryYear1Sem DSA, there is no big percentage gaps between the 8 most relevant features. Although the first semester evaluation-related features still show high importance, they are exceeded by several previously collected features. Time-domain features show leveraged impact on failure at this point (end of first curricular year). Student's decision on retention, transition or dropout are potentially more influenced by individual life cycles at the end of first curricular year, directly impacting success.

Discussion

Figure 12 shows a wrapped-up analysis for the three main reviewed models' performance, considering each DM model per features' collection time.

Figure 12. Shows a wrapped-up analysis for reviewed models' performance.

SVM is clearly the best model for DM_Entrance, as it outperforms other models for all threshold range. DM_Entrance model can be developed to predict student's performance before the beginning of the first curricular semester. This a-priori predictive model shows good evaluation results (AUC=0.77 for SVM). DM_EntryYear1Sem model predicts student's performance by the end of the first curricular semester, achieving improved evaluation results (AUC around 0.91 for SVM). DM_EntryYear2Sem model can be set up by the end of the first curricular year, achieving near perfect performance (AUC around 0.94 for SVM and RF models). SVM results can be partially explained due to improved performance of the SVM training algorithm for small sized datasets. As for the DM_entryYear2Sem model, both SVM and RF achieved similar performance, with RF even surpassing SVM's results in a few of the evaluation metrics. This RF performance boost can be explained by algorithm's improved ability to deal with a mixture of numerical and categorical features, bearing in mind that relevant numerical features amount has increased significantly, with the inclusion of first and second semesters' students' evaluation features. Although relying on slightly later stages, reducing timings for decision-making and actions to be taken, these models provide an enhanced predictive potential, achieving great performances. These results demonstrate that collecting fresh features during the first curricular year, such as, student's evaluation performance features, it is possible to enrich model's ability to predict unsuccessful cases, while reducing false positive detections.

Conclusion

The overall success of an EDM project is very much accounted for providing educational stakeholder, such as deans, coordinators, teachers and managers, with meaningful information when making decisions concerning educational policies, courses offered, etc (Fernandes et al., 2018). It is therefore useful to underline this type of knowledge as a basis for informed intervention. This may point to clues for the institution's various forms of action. Following are

some of these guidelines for the case of the analysed institution, especially regarding the success of its freshman students:

- Providing specific study supporting groups for lower entry grade's students, since the beginning of first curricular semester. Some literature suggests that low performing secondary school students tend to maintain their low performance level on further higher education.

- Monitoring performance evolution of a specific students' group. This group would be gathered using the following criteria: low entry grade (below 13); older students (above 26 years old) and large study gap (above 20 years).

- Identifying students that collect less than 18 ECTS or achieve weighted average grade below 7, at the end of the first curricular semester. Extended institutional support can be provided to these students, such as, helping them defining individual study plan for second curricular semester, clearly identifying effort requirements and work balance for better performance achievement.

- Again, at the end of the second curricular semester, poor performance students could be identified. Proceeding with pedagogical support is important at this stage.

A significant part of this study's effort consisted in data quality tasks. Nevertheless, predictive potential has been lost due to some bad quality data, this is a limitation on this study. Consistent and coherent academic data is easier to analyse and include in further DM models and frameworks. Specifically, in the data preparation phase, some features were removed due to consisting in single class features. As programmes have unique resource needs, contact hours, credits per module, prior-entry qualification requirements and, laboratory/fieldwork demands, the dataset after the preparation tasks does not reflect important aspects necessary in gauging students' academic success in higher education, which consists in an important limitation that must be mentioned. Simple processes, as empty/incomplete fields validation could be applied

to academic forms in order to reduce inadequate data. Creating a segmented list of answers for each field would enhance the quality of collected data. These suggestions would facilitate and promote DM applications as it would potentially reduce the data preparation, cleansing and quality stages' effort as well as increasing the number of data and specially the number of candidates' features to be included in the model.

Ketonen, Haarala-Muhonen, Hirsto, Hänninen, Wähälä and Lonka (2016) characterized the first-year students through a set of profiles: alienated, engaged, disengaged and undecided. They found that the engaged students performed better in academic achievement and the undecideds one received the lowest grades. For future work, it would be interesting to compare the results presented in our study mediated by defined features and Ketonen et al. (2016) to create a more complete model that considers both approaches.

We also propose to designing individual school's DM models based on presented models, in order to capture specific school's characteristics; considering additional data sources, such as, end of semester's student satisfaction surveys; scrutinizing the effect of post-labour feature on academic failure; and extending data quality approaches on social origin, candidacy preference and secondary school related features and revisiting their impact on predicting academic failure. Ultimately, an information system encompassing these models can be used as a data-driven decision-making framework for supporting and optimising institutional policies and actions for academic success, also in other educational systems and social contexts.

This study has important limitations that derive, in part, from its main empirical reference. It deals with institutional data with essentially administrative and educational management functions. However, the proposed model has many advantages for monitoring and defining policies within the framework of this Portuguese university and may be replicable to other institutions.

Despite the limits of the variables available to cover some of the dimensions inscribed in the theoretical models used here - such as those related to learning and skills acquisition processes (among others mentioned by authors such as York et al, 2015); or those related to the engagement and integration of students in the academic environment and their activities (as mentioned by Tinto, 2006) - it enables the use, fulfilling all the ethical requirements of data protection, of an information system on students in order to study a relevant range of attributes and factors involved in academic success. In general, these information systems have, among others, attributes of socio-demographic characterization, family background, previous academic paths and sometimes some indicators on student satisfaction that can be related to variables of educational outcomes (which allow approaching a reading of academic success). This type of exercise enables the adaptation of theoretical models to the conditions of availability of existing information, still allowing to add causal and relational knowledge about the factors of success in higher education, particularly useful, because it can provide relevant knowledge to the higher education institutions. Although this model has only been tested in one university institution, it can be tested and produce interesting results in other institutional contexts, in Portugal or in other countries, reinforcing the possibilities of *monitoring* and *intervening in advance* regarding academic success. Its focus on first-year students allows not only to act in a recognized critical segment, but also to intervene early in the sustainability of successful paths (Brouwer, Jansen, Flash and Hofman, 2016).

References

Almahadeen, L., Akkaya, M. and Sari, A. (2017). Mining student data using CRISP-DM model. *International Journal of Computer Science and Information Security*, 15(2), pp. 305-316.

Apté, C. and Weiss, S. (1997). Data mining with decision trees and decision rules. *Future generation computer systems*, 13(2-3), pp. 197-210.

Asif, R., Merceron, A., Ali, S. A. and Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers & Education*, 113, pp. 177-194.

Astin, A. W. (2012). *Assessment for excellence: The philosophy and practice of assessment and evaluation in higher education*. Rowman & Littlefield Publishers.

Baker, R. S. (2010). Data mining for education. *International Encyclopedia of Education*, 7(3), pp. 112-118.

Baker, R. S. and Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *JEDM Journal of Educational Data Mining*, 1(1), pp. 3-17.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7), pp. 1145-1159.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1), pp. 5-32.

Brouwer, J., Jansen, E., Flache, A. and Hofman, A. (2016). The impact of social capital on self-efficacy and study success among first-year university students. *Learning and Individual Differences*, 52, pp. 109-118.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wirth, R. (2000). *CRISP-DM 1.0 -Step-by-step data mining guide*, CRISP-DM Consortium.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3), pp. 273-297.

Cortez, P. (2010). Data mining with neural networks and support vector machines using the R/rminer tool. In: *Industrial Conference on Data Mining*. Berlin: Springer, pp. 572-583.

Cortez, P. and Embrechts, M. J. (2011). Opening black box data mining models using sensitivity analysis. In: *2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*. IEEE, pp. 341-348.

Costa, A. F. and Lopes, J. T. (2011). The Diverse Pathways of Higher Education Students: A Sociological Analysis on Inequality, Context and Agency. *Portuguese Journal of Social Science*, 10, pp. 43-58. https://doi.org/10.1386/pjss.10.1.43_1

Mestre, C. and Baptista, J. O. (2016). *Desigualdades Socioeconómicas e Resultados Escolares: 3º ciclo do ensino público geral*. [online] Lisbon: Direção-Geral de Estatísticas da Educação e da Ciência. Available at: http://www.dgeec.mec.pt/np4/316/%7B$clientServletPath%7D/?newsId=607&fileName=DesigualdadesResultadosEscolares.pdf

Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R. and Van Erven, G. (2019). Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. *Journal of Business Research*, 94, pp. 335-343.

Guo, Q., Zhou, J. and Feng, L. (2018). Pro-social behavior is predictive of academic success via peer acceptance: A study of Chinese primary school children. *Learning and Individual Differences*, 65, pp. 187-194.

Goker, H., Bulbul, H. I. and Irmak, E. (2013). The estimation of students' academic success by data mining methods. In: *12th International Conference on Machine Learning and Applications*. IEEE, pp. 535-539.

Hannon, O., Smith, L. R. and Lã, G. (2017). Success at University: The Student Perspective. In: L. Wood and Y. Breyer, eds., *Success in Higher Education*. Singapore: Springer, pp. 257-268.

Haykin, S. (1994). *Neural networks: a comprehensive foundation*. Prentice Hall PTR.

Ingram, N (2011). Within School and Beyond the Gate: The Complexities of Being: Educationally Successful and Working Class, *Sociology*, 45(2), pp. 287–302.

Ketonen, E. E., Haarala-Muhonen, A., Hirsto, L., Hänninen, J. J., Wähälä, K. and Lonka, K. (2016). Am I in the right place? Academic engagement and study success during the first years at university. *Learning and Individual Differences,* 51, pp. 141-148.

Khan, S., Liu, X., Shakil, K. A. and Alam, M. (2017). A survey on scholarly data: From big data perspective. *Information Processing & Management,* 53(4), pp. 923-944.

Koedinger, K., Cunningham, K., Skogsholm, A. and Leber, B. (2008). An open repository and analysis tools for fine-grained, longitudinal learner data. In: *1st International Conference on Educational Data Mining*. Montreal: International Working Group on Educational Data Mining, pp. 157-166.

Kohavi, R. and Provost, F. (1998). Glossary of terms. *Machine Learning*, 30(271), pp. 127-132.

Martins, M. P., Migueis, V. L. and Fonseca, D. S. B. (2018). A data mining approach to predict undergraduate students' performance. In: *13th Iberian Conference on Information Systems and Technologies* (CISTI). IEEE, pp. 1-7.

Mishra, T., Kumar, D. and Gupta, S. (2014). Mining students' data for prediction performance. In: *Fourth International Conference on Advanced Computing & Communication Technologies*. IEEE, pp. 255-262.

Moro, S., Laureano, R. and Cortez, P. (2011). Using data mining for bank direct marketing: An application of the crisp-dm methodology. In: *Proceedings of European Simulation and Modelling Conference-ESM'2011*. Guimarães: Eurosis, pp 117-121.

Natek, S. and Zwilling, M. (2014). Student data mining solution–knowledge management system related to higher education institutions. *Expert Systems with Applications*, 41(14), pp. 6400-6407.

Osmanbegović, E. and Suljić, M. (2012). Data mining approach for predicting student performance. *Economic Review*, 10(1), pp. 3-12.

Pace, A., Alper, R., Burchinal, M. R., Golinkoff, R. M. and Hirsh-Pasek, K. (2019). Measuring success: Within and cross-domain predictors of academic and social trajectories in elementary school. *Early Childhood Research Quarterly*, 46, pp. 112-125.

Pascarella, E. T. and Terenzini, P. T. (2005). *How college affects students: A third decade of research*. Vol 2. San Francisco: Jossey-Bass.

Romero, C. and Ventura, S. (2010). Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetic*s, 40(6), pp. 601-618.

Shahiri, A. M. and Husain, W. (2015). A review on predicting student's performance using data mining techniques. *Procedia Computer Science, 72*, pp. 414-422.

Slim, A., Heileman, G. L., Kozlick, J. and Abdallah, C. T. (2014). Predicting student success based on prior performance. In: *Symposium on Computational Intelligence and Data Mining (CIDM)*. IEEE, pp. 410-415.

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of educational research*, 75(3), pp. 417-453.

Tinto, V. (1999). Taking retention seriously: Rethinking the first year of college. *NACADA journal*, 19(2), pp. 5-9.

Tinto, V. (2006). Research and practice of student retention: What next? *Journal of College Student Retention: Research, Theory & Practice*, 8(1), pp. 1-19.

Trevor, H., Robert, T. and Friedman, JH. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Trstenjak, B. and Donko, D. (2014). Determining the impact of demographic features in predicting student success in Croatia. In: *37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. IEEE, pp. 1222-1227.

Zimmermann, J., Brodersen, K. H., Heinimann, H. R. and Buhmann, J. M. (2015). A model-based approach to predicting graduate-level performance using indicators of undergraduate-level performance. *Journal of Educational Data Mining*, 7(3), 151-176.

Tables

Table 1 - Features' description and classes

Feature	Description	Classes
Area	Student's residence code area	"1495";"2795";etc.
areaCode	Student's postal residence	
gender	Gender	"M";"F"
yearOfBirth	Year of birth	"1986";"1967";etc.
fatherOccupation	Father's occupation	"Managers";"Elementary occupations";etc.
motherOccupation	MotherOccupation	"Managers";"Elementary occupations";etc.
fatherOccupationConditionType	Father's condition in labour force	"Unemployed"; "Dep.worker"; "Imp.worker";etc.
motherOccupationConditionType	Mother's condition in labour force	"Unemployed"; "Dep.worker"; "Imp.worker";etc.
occupation	Student's occupation	"Student";"Filled Occupation", "Unknown"
firstExecutionYear	Curricular year of first admission in the institution	"2006/2007";"2007/2008";etc
maritalStatusType	Marital status	"Married";"Single";etc.

nationality	Nationality	"French"; "Spanish";etc.
secondNationality	Second Nationality	"French"; "Spanish";etc.
entryYear	Year of registration in current degree	"2006/2007";"2007/2008";etc
fatherLiteraryHabilitationType	Father's educational level	"Illiterate", "Higher education";etc.
motherLiteraryHabilitationType	Mother's educational level	"Illiterate", "Higher education";etc.
workingStudentAtEntry	Student worker statute (required at admission process)	"True;"False".
partialTimeStudentAtEntry	Partial time statute (required at admission process)	"True;"False".
specialEducationNeedsAtEntry	Special Education statute (required at admission process)	"True;"False".
scholarShipAtEntry	Scholarship (granted at admission process)	"True;"False".
dislocatedAtEntry	Dislocated statute (required at admission process)	"True;"False".
degreeCode	Code that represents each degree	"MNG";"CS";etc.
degreeType	Degree Type	"Bachelor";"Master";etc.
degreeSchool	Degree school	"Business";"IT";etc
entryGrade	Entry Grade for HEI admission	"9.5" to "20"
precedentDegreeDesignation	Field of study in secondary school	"Sciences";"Sports";etc.
precedentConclusionYear	Secondary school conclusion year	"1984";"2005";etc.
secondarySchoolType	Secondary (high) school's sector	"Public";"Private";etc
ingression	Ingression type	"CNA";"CM23";etc
highSchoolDegreeType	Secondary school via (or path)	"Scientific_Humanistic";"Other";etc.
wasFirstChoice	Was the university chosen in 1st place?	"True;"False".
erasmusOutgoing	Student accepted for Erasmus outgoing	"True;"False".
workingStudentEntryYear1stSem	Student worker statute granted during 1st semester	"True;"False".
InternationalStudentEntryYear1stSem	International student statute granted during 1st semester	"True;"False".
partialTimeStudentEntryYear1stSem	Partial time statute granted during 1st semester	"True;"False".
fctgrantOwnerEntryYear1stSem	FCT grant granted during 1st semester	"True;"False".
classSubRepresentativeEntryYear1stSem	Statute granted during 1st semester	"True;"False".
classRepresentativeEntryYear1stSem	Statute granted during 1st semester	"True;"False".
handicappedEntryYear1stSem	Handicapped statute granted during 1st semester	"True;"False".
pregnantOrChildrenUnder3EntryYear1stSem	Pregnant or children under 3 years old statute granted during 1st semester	"True;"False".
professionalAthleteEntryYear1stSem	Professional athlete statute granted during 1st semester	"True;"False".
sasGrantOwnerEntryYear1stSem	Social support statute (SAS) granted during 1st semester	"True;"False".
militaryEntryYear1stSem	Military statute granted during 1st semester	"True;"False".
temporaryDisabilityEntryYear1stSem	Temporary disability statute granted during 1st semester	"True;"False".
religiousEntryYear1stSem	Religious statute granted during 1st semester	"True;"False".
associativeLeaderEntryYear1stSem	Associative leader statute granted during 1st semester	"True;"False".
athleteEntryYear1stSem	Athlete statute granted during 1st semester	"True;"False".
firefighterEntryYear1stSem	Firefighter statute granted during 1st semester	"True;"False".
erasmusGuestEntryYear1stSem	Erasmus guest statute granted during 1st semester	"True;"False".
deathOfSpouseOrFamilyEntryYear1stSem	Death of spouse or family statute granted during 1st semester	"True;"False".
appearancePoliceOrMilitaryAuthorityEntryYear1stSem	Appearance in police or military authority statute granted during 1st semester	"True;"False".
monitorEntryYear1stSem	Monitor statute granted during 1st semester	"True;"False".
previousIBSStudentEntryYear1stSem	Previous IBS student statute granted during 1st semester	"True;"False".
top15IBSEntryYear1stSem	Top 15 IBS statute granted during 1st semester	"True;"False".
workingStudentEntryYear2ndSem	Student workerstatute granted during 1st semester	"True;"False".
InternationalStudentEntryYear2ndSem	International student statute granted during 2nd semester	"True;"False".

partialTimeStudentEntryYear2ndSem	Partial time statute granted during 2nd semester	"True;"False".
fctgrantOwnerEntryYear2ndSem	FCT grant granted during 2nd semester	"True;"False".
classSubRepresentativeEntryYear2ndSem	Class sub-representative statute granted during 2nd semester	"True;"False".
classRepresentativeEntryYear2ndSem	Class representative statute granted during 2nd semester	"True;"False".
handicappedEntryYear2ndSem	Handicapped statute granted during 2nd semester	"True;"False".
pregnantOrChildrenUnder3EntryYear2ndSem	Pregnant or children under 3 years old statutes granted during 2nd semester	"True;"False".
professionalAthleteEntryYear2ndSem	Professional athlete statute granted during 2nd semester	"True;"False".
sasGrantOwnerEntryYear2ndSem	Social support statute (SAS) granted during 2nd semester	"True;"False".
militaryEntryYear2ndSem	Military statute granted during 2nd semester	"True;"False".
temporaryDisabilityEntryYear2ndSem	Temporary disability statute granted during 2nd semester	"True;"False".
religiousEntryYear2ndSem	Religious statute granted during 2nd semester	"True;"False".
associativeLeaderEntryYear2ndSem	Associative leader statute granted during 2nd semester	"True;"False".
athleteEntryYear2ndSem	Athlete statute granted during 2nd semester	"True;"False".
firefighterEntryYear2ndSem	Firefighter statute granted during 2nd semester	"True;"False".
erasmusGuestEntryYear2ndSem	Erasmus guest statute granted during 2nd semester	"True;"False".
deathOfSpouseOrFamilyEntryYear2ndSem	Death of spouse or family statute granted during 2nd semester	"True;"False".
appearancePoliceOrMilitaryAuthorityEntryYear2ndSem	Appearance in police or military authority statute granted during 2nd semester	"True;"False".
monitorEntryYear2ndSem	Monitor statute granted during 2nd semester	"True;"False".
previousIBSStudentEntryYear2ndSem	Previous IBS student statute granted during 2nd semester	"True;"False".
top15IBSEntryYear2ndSem	Top 15 IBS statute granted during 2nd semester	"True;"False".
requestedSocialServiceEntryYear	Requested any social service during 1st year	"True;"False".
acceptedSocialServiceEntryYear	Granted any social service during 1st year	"True;"False".
requestedSStransportSupplementEntryYear	Requested transport supplement during 1st year	"True;"False".
requestedSSaccommodationSupplementEntryYear	Requested accommodation supplement during 1st year	"True;"False".
requestedSSresidenceRequestEntryYear	Requested residence during 1st year	"True;"False".
requestedSSFinantialSupportEntryYear	Requested financial support during 1st year	"True;"False".
acceptedSStransportSupplementEntryYear	Granted transport supplement during 1st year	"True;"False".
acceptedSSaccommodationSupplementEntryYear	Granted accommodation supplement during 1st year	"True;"False".
acceptedSSresidenceRequestEntryYear	Granted residence during 1st year	"True;"False".
acceptedSSFinantialSupportEntryYear	Granted financial support during 1st year	"True;"False".
firstChoice	Was It the first choice (University+degree)?	"True";"False".
firstChoiceUniversity	Was the first choice?	"True";"False".
firstChoiceCourse	Was the enrolled degree the first choice?	"True";"False".
orderPreference	which order of preference did the student registered?	"1";"2";"3";"4";"5";"6".
gapEntryExames	Grade average points for entry exams	"9.5" to "20"
entryAge	Student's age at entry	"16" to "74"
entryAgeRange	Student's age at entry	"[16-18]";"[19-23]"; etc.
municipality	Student's residence municipality	"Lisboa";"Oerias"; etc.
district	Student's residence district	"Lisboa";"Setúbal";etc.
lisbonMetropolitanArea	Does student live within Lisbon metropolitan area?	"True";"False".
studyGap	Any time gap since previous educational programme?	"True;"False".
studyGapYears	Time Gap since previous educational programme	"0";"1";"2"; etc.

ectsCreditsEntryYear1stSem	number of course passed in the entry year 1st semester	"0";"6";"12"; etc.
ectsCreditsEntryYear2ndSem	number of course passed in the entry year 2nd semester	"0";"6";"12"; etc.
averageEntryYear1stSem	average grade of the passed courses in entry year 1st semester	"0" to "20"
weightedAverageEntryYear1stSem	weighted average grade of the passed courses in entry Year 1st semester	"0" to "20"
averageGradeEntryYear2ndSem	average grade of the passed courses in entry year 2nd semester	"0" to "20"
weightedAverageEntryYear2ndSem	weighted average grade of the passed courses in entry year 2nd semester	"0" to "20"

Table 2 - Final ABT for DM modelling purposes.

Feature Name	Features' Group	Data Type	Collection time
gender	Socio-demographic	Cat.	Entrance
yearOfBirth	Socio-demographic	Num.	
fatherOccupationConditionType	Social Origin	Cat.	
motherOccupationConditionType	Social Origin	Cat.	
occupation	Socio-demographic	Cat.	
FirstExecutionYear	Educational Path	Cat.	
maritalStatusType	Socio-demographic	Cat.	
nationality	Socio-demographic	Cat.	
secondNationality	Socio-demographic	Cat.	
entryYear	Educational Path	Cat.	
fatherLiteraryHabilitationType	Social Origin	Cat.	
motherLiteraryHabilitationType	Social Origin	Cat.	
entryAge	Educational Path	Num.	
entryAgeRange	Educational Path	Cat.	
district	Socio-demographic	Cat.	
lisbonMetropolitanArea	Socio-demographic	Cat.	
fatherOccupation	Socio-demographic	Cat.	
motherOccupation	Socio-demographic	Cat.	
degreeCode	Educational Path	Cat.	
degreeSchool	Educational Path	Cat.	
precedentConclusionYear	Previous Education	Cat.	
secondarySchoolType	Previous Education	Cat.	
ingression	Previous Education	Cat.	
entryGradeHotDeck	Previous Education	Num.	
studyGap	Previous Education	Cat.	
studyGapYears	Previous Education	Num.	
workingStudentAtEntry	Special Statute	Cat.	
specialEducationNeedsAtEntry	Special Statute	Cat.	
scholarshipAtEntry	Special Statute	Cat.	
dislocatedAtEntry	Special Statute	Cat.	
workingStudentEntryYear1stSem	Special Statute	Cat.	At the end of first curricular semester
InternationalStudentEntryYear1stSem	Special Statute	Cat.	
classSubRepresentativeEntryYear1stSem	Special Statute	Cat.	
classRepresentativeEntryYear1stSem	Special Statute	Cat.	
handicappedEntryYear1stSem	Special Statute	Cat.	
pregnantOrChildrenUnder3EntryYear1stSem	Special Statute	Cat.	
professionalAthleteEntryYear1stSem	Special Statute	Cat.	
sasGrantOwnerEntryYear1stSem	Special Statute	Cat.	
temporaryDisabilityEntryYear1stSem	Special Statute	Cat.	
associativeLeaderEntryYear1stSem	Special Statute	Cat.	
athleteEntryYear1stSem	Special Statute	Cat.	
ectsCreditsEntryYear1stSem	Educational Path	Num.	
averageEntryYear1stSem	Educational Path	Num.	
weightedAverageEntryYear1stSem	Educational Path	Num.	

workingStudentEntryYear2ndSem	Special Statute	Cat.	
InternationalStudentEntryYear2ndSem	Special Statute	Cat.	
classSubRepresentativeEntryYear2ndSem	Special Statute	Cat.	
classRepresentativeEntryYear2ndSem	Special Statute	Cat.	
handicappedEntryYear2ndSem	Special Statute	Cat.	
pregnantOrChildrenUnder3EntryYear2ndSem	Special Statute	Cat.	
professionalAthleteEntryYear2ndSem	Special Statute	Cat.	
sasGrantOwnerEntryYear2ndSem	Special Statute	Cat.	
temporaryDisabilityEntryYear2ndSem	Special Statute	Cat.	
associativeLeaderEntryYear2ndSem	Special Statute	Cat.	
athleteEntryYear2ndSem	Special Statute	Cat.	At the end of second curricular semester (first curricular year)
requestedSocialServiceEntryYear	Special Statute	Cat.	
acceptedSocialServiceEntryYear	Special Statute	Cat.	
requestedSStransportSupplementEntryYear	Special Statute	Cat.	
requestedSSaccommodationSupplementEntryYear	Special Statute	Cat.	
requestedSSresidenceRequestEntryYear	Special Statute	Cat.	
requestedSSFinantialSupportEntryYear	Special Statute	Cat.	
acceptedSStransportSupplementEntryYear	Special Statute	Cat.	
acceptedSSaccommodationSupplementEntryYear	Special Statute	Cat.	
acceptedSSresidenceRequestEntryYear	Special Statute	Cat.	
acceptedSSFinantialSupportEntryYear	Special Statute	Cat.	
ectsCreditsEntryYear2ndSem	Educational Path	Num.	
averageGradeEntryYear2ndSem	Educational Path	Num.	
weightedAverageEntryYear2ndSem	Educational Path	Num.	

Data type: Num. = Numerical, Cat. = Categorical.

Table 3 - AUC results for DM_Entrance model.

RPART	DT	CTREE	SVM	RF	MLPE	Model Details
0.6764	0.6772	0.7273	0.7732	0.7611	0.7476	DM_Entrance model 30 features 9652 records

AUC mean values after 20 runs of 10-fold for each modelling technique

Table 4 - Confusion matrices for DM_Entrance model.

Threshold = 50%					
SVM		Predicted		Sensitivity	1-specificity
		Failure	Success		
Target	Failure	3469	1462	0.7035	0.2959
	Success	1397	3324		
RF		Predicted		Sensitivity	1-specificity
		Failure	Success		
Target	Failure	2885	2046	0.5851	0.2197
	Success	1037	3684		
MLPE		Predicted		Sensitivity	1-specificity
		Failure	Success		
Target	Failure	3318	1613	0.6729	0.3215
	Success	1518	3203		
CTREE		Predicted		Sensitivity	1-specificity
		Failure	Success		
Target	Failure	3336	1595	0.6765	0.3527
	Success	1665	3056		

Table 5 - AUC results for DM_EntryYear1Sem model

RPART	DT	CTREE	SVM	RF	MLPE	Model Details
0.8463	0.8466	0.8954	0.9097	0.9082	0.8936	DM_EntryYear1Sem model 44 features 9652 records

AUC mean values after 20 runs of 10-fold for each modelling technique

Table 6 - Confusion matrix for DM_EntryYear1Sem model.

Threshold = 30%					
SVM		Predicted		Sensitivity	1-specificity
		Failure	Success		
Target	Failure	4303	628	0.8726	0.2402
	Success	1134	3587		
RF		Predicted		Sensitivity	1-specificity
		Failure	Success		
Target	Failure	4297	634	0.8714	0.2480
	Success	1171	3550		
MLPE		Predicted		Sensitivity	1-specificity
		Failure	Success		
Target	Failure	4293	638	0.8706	0.3093
	Success	1460	3261		
CTREE		Predicted		Sensitivity	1-specificity
		Failure	Success		
Target	Failure	4279	652	0.8678	0.2824
	Success	1333	3388		

Table 7 - AUC results for DM_EntryYear2Sem model.

RPART	DT	CTREE	SVM	RF	MLPE	Model Details
0.8882	0.8886	0.9257	0.9378	0.9380	0.9263	DM_EntryYear2Sem model 68 features 9652 records

AUC mean values after 20 runs of 10-fold for each modelling technique

Table 8 - Confusion matrices for DM_EntryYear2Sem model.

Threshold = 20%					
SVM		Predicted		Sensitivity	1-specificity
		Failure	Success		
Target	Failure	4536	395	0.9199	0.2724
	Success	1286	3435		
RF		Predicted		Sensitivity	1-specificity
		Failure	Success		
Target	Failure	4624	307	0.9377	0.3247
	Success	1533	3188		
MLPE		Predicted		Sensitivity	1-specificity
		Failure	Success		
Target	Failure	4293	638	0.9187	0.3173
	Success	1460	3261		
CTREE		Predicted		Sensitivity	1-specificity
		Failure	Success		
Target	Failure	4600	331	0.9329	0.3739
	Success	1765	2956		

Figures

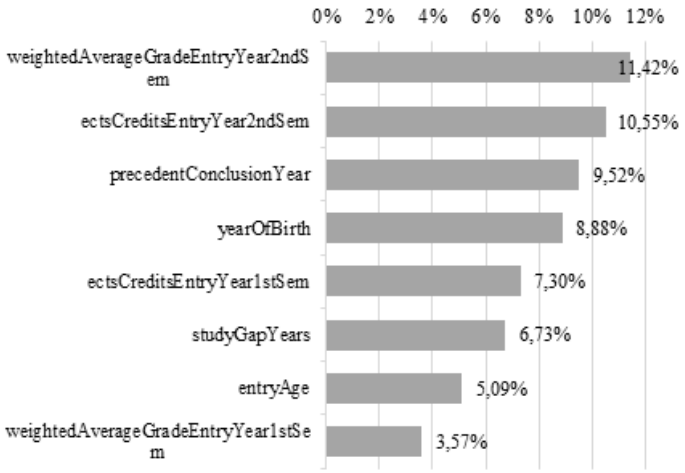

Figure 1

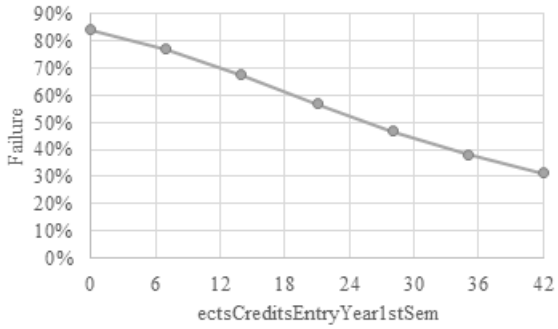

Figure 2

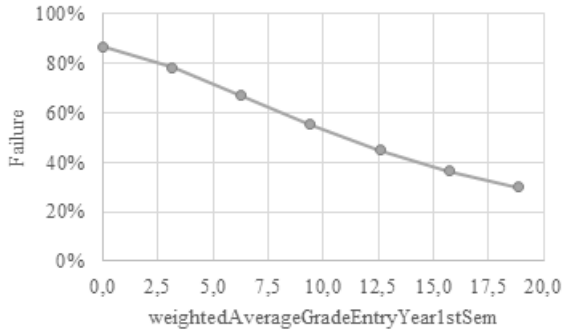

Figure 3

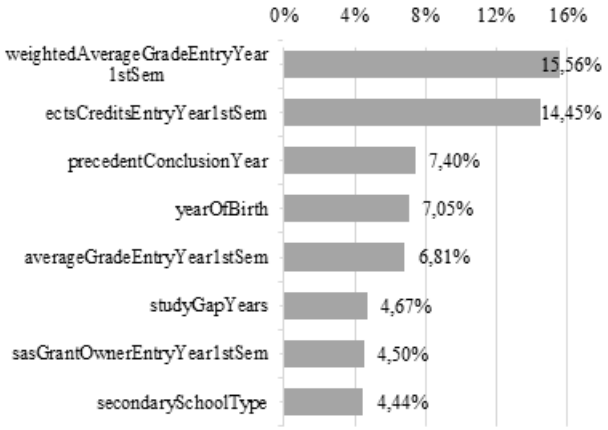

Figure 4

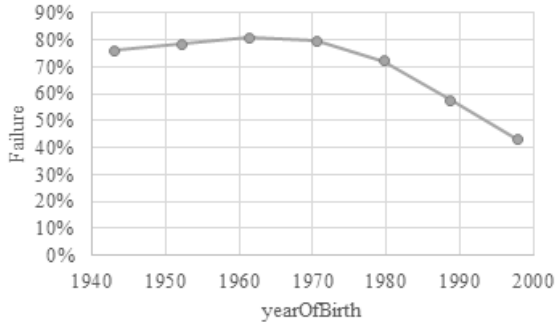

Figure 5

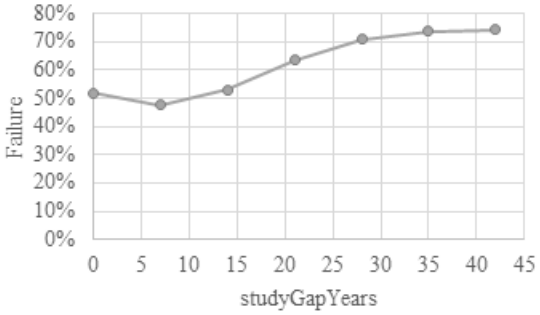

Figure 6

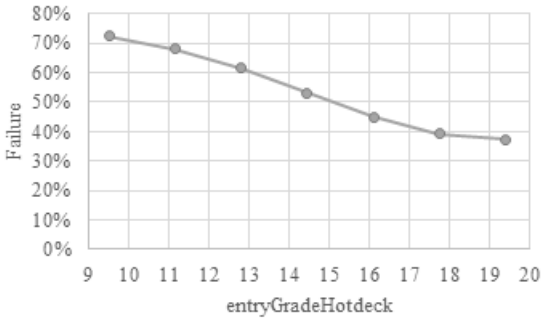

Figure 7

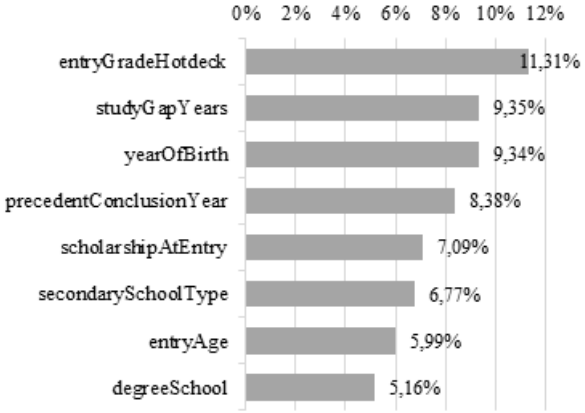

Figure 8

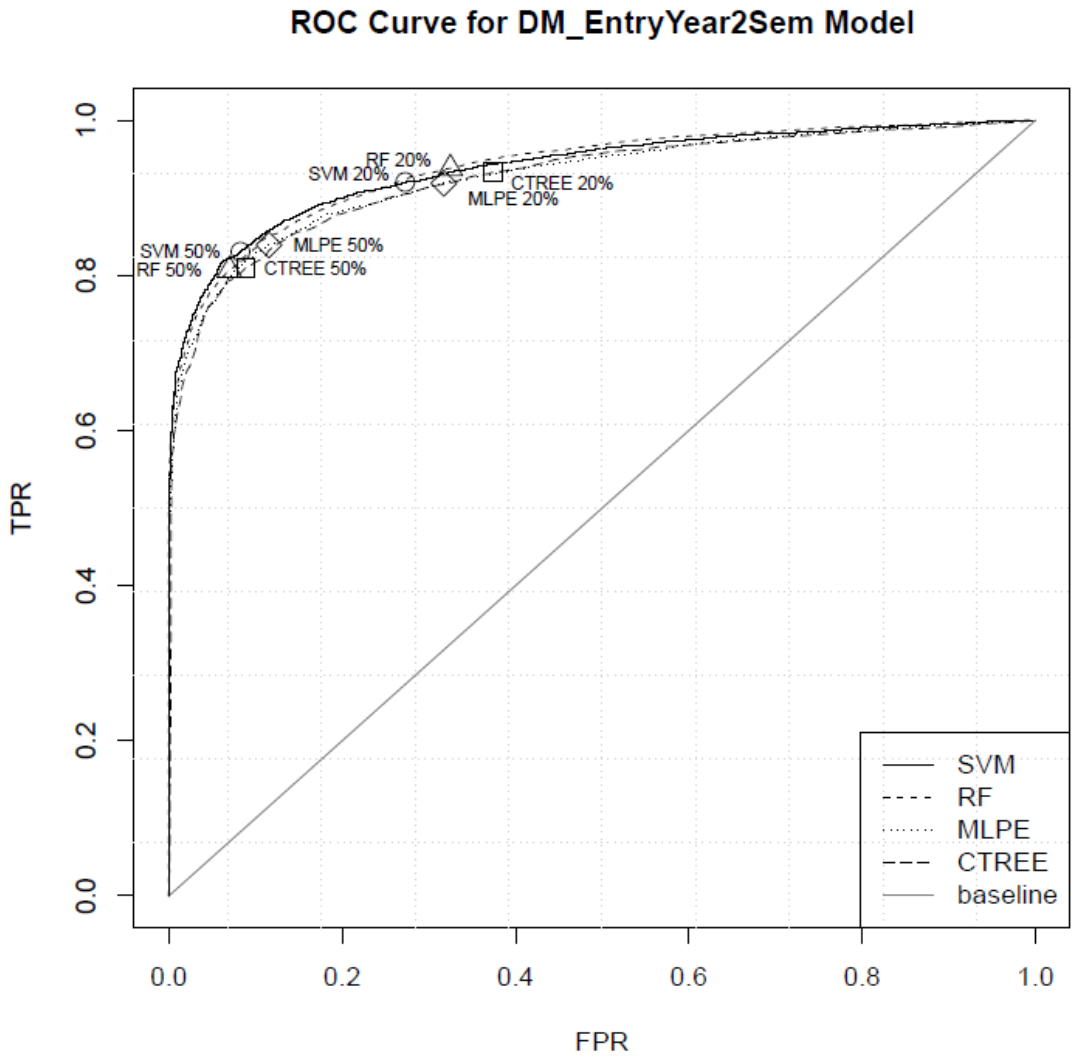
ROC Curve for DM_EntryYear2Sem Model
TPR
FPR
0.0
0.2
0.4
0.6
0.8
1.0
SVM 20%
RF 20%
CTREE 20%
MLPE 20%
SVM 50%
RF 50%
MLPE 50%
CTREE 50%
SVM
RF
MLPE
CTREE
baseline

Figure 9

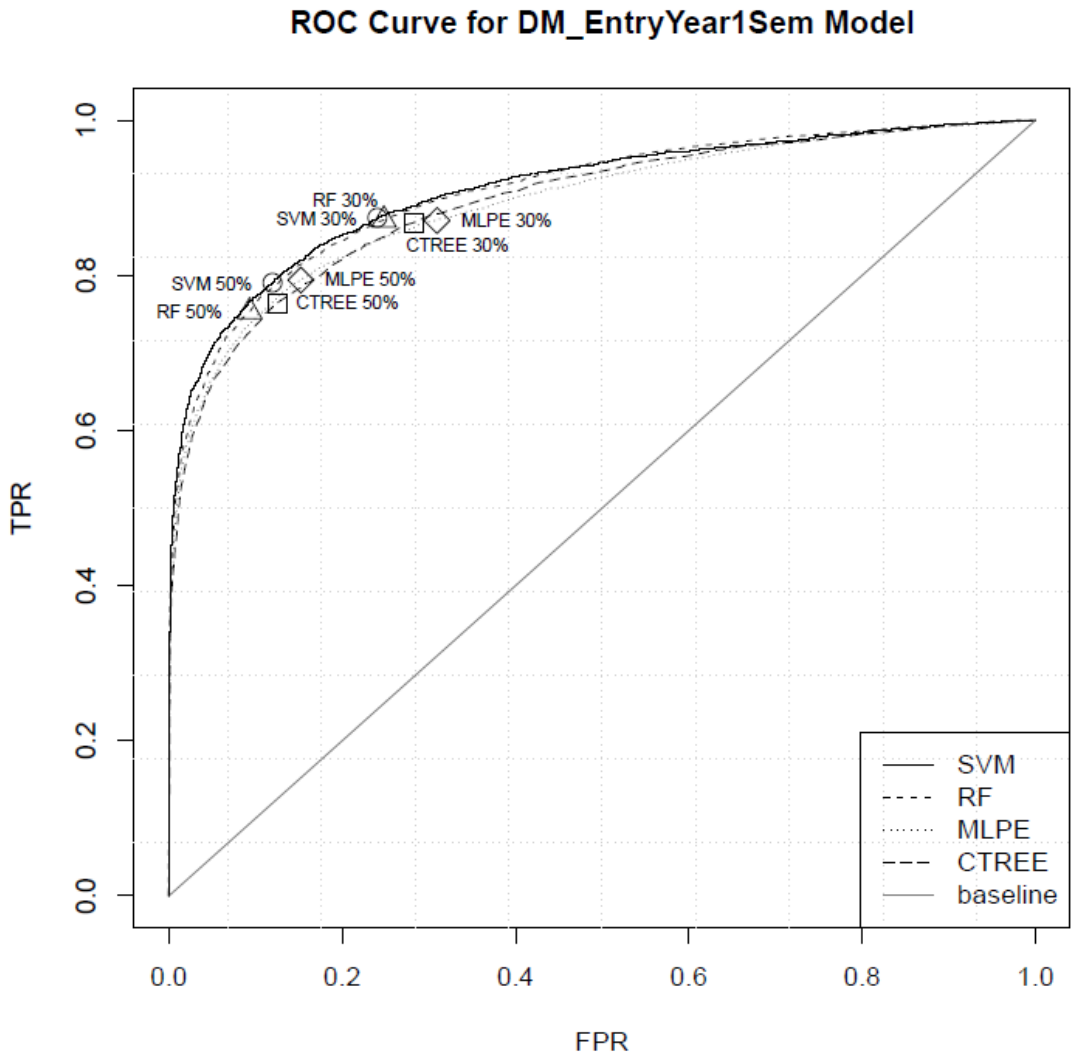
ROC Curve for DM_EntryYear1Sem Model
TPR
FPR
0.0
0.2
0.4
0.6
0.8
1.0
RF 30%
SVM 30%
MLPE 30%
CTREE 30%
SVM 50%
MLPE 50%
RF 50%
CTREE 50%
SVM
RF
MLPE
CTREE
baseline

Figure 10

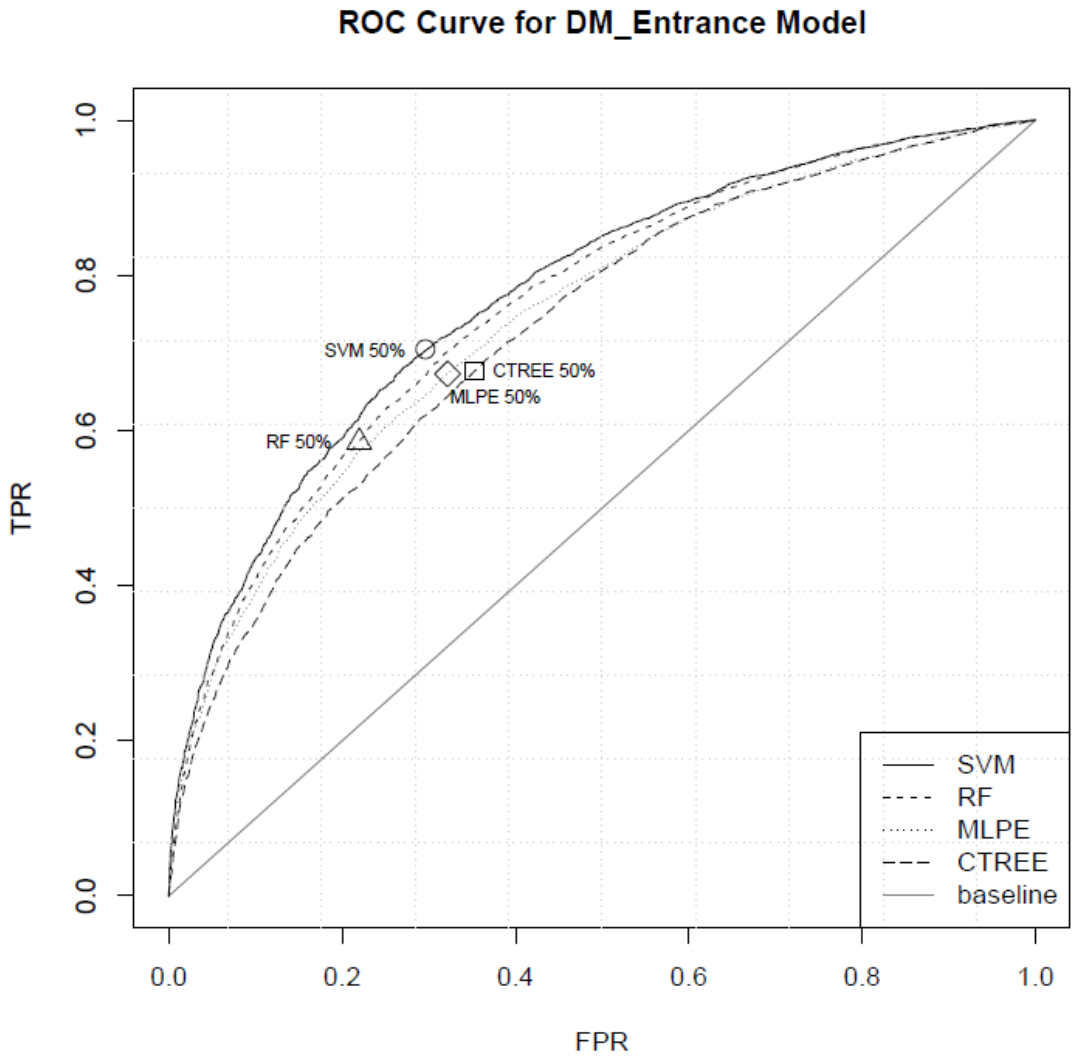

Figure 11

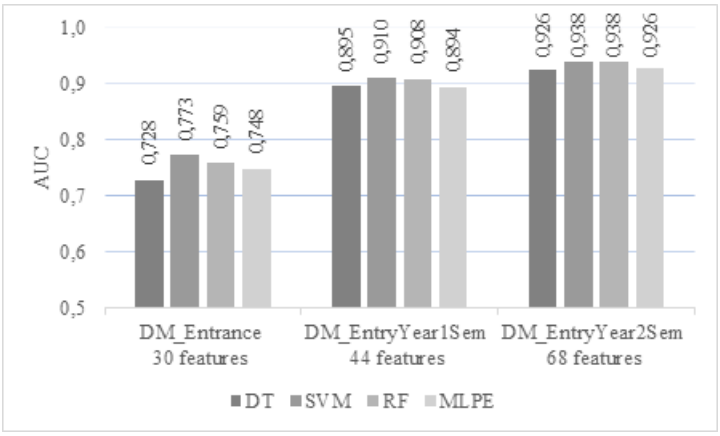

Figure 12

Figure Captions