Educational Data Mining: Analysis of Drop out of Engineering Majors at the UnB - Brazil

Rodrigo da Fonseca Silveira
Centro de Informática
University of Brasilia
Brasilia, Brazil
rodrigofonseca@unb.br

Marcio de Carvalho Victorino
Faculdade de Ciência da Informação
University of Brasilia
Brasilia, Brazil
mcvictorino@unb.br

Maristela Holanda
Depart. de Ciência da Computação
Brasilia, Brazil
University of Brasilia
maholanda@unb.br

Marcelo Ladeira
Depart. de Ciência da Computação
Brasilia, Brazil
University of Brasilia
mladeira@unb.br

***Abstract*—This paper presents an analysis of data about the drop out of undergraduate engineering students at the University of Brasilia(UnB), Brazil. In Brazil, similar to other countries, there is a representative amount of engineering students that enroll in engineering majors, however, they don't get to graduate in those majors. Information about the reason for that phenomenon is important for action on the matter by university decision-makers. This paper aims to answer the research question: What are the main factors that motivate engineering students to drop out of engineering majors at UnB? We have collected the social and performance data of engineering students from 2009 to 2019. Some of the data can be considered rare in similar studies, like students' distance from home to campus and factors like students' leave of absence requests rather than performance factors. We used three data mining techniques: Generalized Linear Model (GLM), Boosting algorithm (GBM) and Random Forest(RF). The results of the study showed that international students deserve some attention from the university and courses like Physics 1 can be challenging for engineering students.**

***Index Terms*—Educational Data Mining; EDM; Machine Learning; Students dropout rate; Sparkling Water Data Mining**

I. Introduction

The topic of the school dropout rate has been the subject of several studies and research in the last decade. In Brazil, it is a common problem in education at practically all levels, from the basic to the higher level. The present study aims to contribute to the work already done at the University of Brasilia, Brazil. We intend to identify patterns of dropout based on data that are still little explored in the context of this university for Engineering majors.

In this context, this paper presents research focused on the dropout rate among UnB engineering students at the Darcy Ribeiro campus over the last ten years. We analyzed not only performance data but also socio-economic data and the distance of the address given by the student to the campus. The analysis applied data mining techniques and supervised algorithms to forecast the probable dropout rate of active students. In addition, the critical profiles of students who may deserve greater attention from engineering departments will be highlighted, so that they are more likely to succeed in their academic lives.

The rest of this work is organized as follows: Section II presents a literature review; Section III provides proposed methodology; Section IV presents the results; and, finally, Section IV outlines the conclusions and indicates possibilities for future work.

II. Literature Review

Educational Data Mining (EDM) can be defined as the application of techniques of data mining to educational data analysis aimed at solving problems in the educational context [1] [2] [3] [4]. EDM uses the databases of educational systems to understand the students to design educational policies that will improve their academic performance. Over the years, several studies have been done to identify patterns or models of prediction of undergraduate student dropout [5] [6] [7] [8] [9]. Some papers are presented below.

Costa et al. [10] cite a study based on supervised models for distance education versus on-campus courses. Roy et al. [11] provide an overview of how the K-means Clustering algorithm can be used to discover students' discipline patterns. Barbu et al. [12] cite K-means as a Drop-out Prediction tool based on admission score and others to predict students' profiles. This article is particularly interesting because it brings a vision of how important the engineering students' skills before the course itself can be.

Bouslimani et al. [13] used logistic regression to predict students' performance in electrical engineering at Université de Moncton, in Canada. Perez et al. [14] bring a study with decision Trees, Logistic Regression and Random Forest to prevent students dropping out in Systems Engineering in Colombia. This is the closest situation found in comparison

978-1-7281-4550-1/19/$31.00 ©2019 IEEE
DOI 10.1109/ICMLA.2019.00048

with UnB. From this article, the insights into using the same algorithms were used.

Bautista et al. [15] used, for example, Decision Trees to analyze if students performances in algebra, calculus and physics student performances can predict the engineering specialization.

Fernandes et al. [16] presented a methodology for analyzing the predictive performance of students in public schools of the Federal District of Brazil, and the proposed methodology was based on CRISP-DM and employed a dataset obtained from a repository of the State Department of Education of the Federal District of Brazil. Martins et al. [17] used H2O software as a data mining tool and employed parameter tuning to train some three classification algorithms and Deep Learning for high school education data. The authors work predicted 71.1% of the cases of dropout given the characteristics of college attrition.

In contrast to the cited studies, the present work adds analysis from rarely used variables, like students' distance from home to campus and factors like students' leave of absence requests or housing assistance applications, rather than performance factors.

III. Methodology

The methodology used in this paper was based on CRISP-DM (Cross Industry Process Model for Data Mining) [18]. The database was composed of social and performance data collected from about 5289 engineering students, graduated or not, but no longer studying in the UnB, in a .csv format file. The database didn't have personal identification since the student enrollment data were fully encrypted. Data were also collected from about 3071 students still active to implement the chosen model.

As a product of the study, we made a predictive and supervised model. We trained the model from records of students disconnected from UnB in the last ten years. The following variables were analyzed: How many times did the student ask for housing assistance or food aid (*qtHelpRequests*); How many times did the student apply for a leave of absence from a course - (*qtGradesTR*); How many times did the student apply for a leave of absence from major (*qtLeaveAbsenceMajor*); How many times did the student transfer Credits (*qtGradesCC*); How many previous university courses (*qtPreviosCoursesUNB*); How many failures in compulsory courses (*qtFailuresOBR*); How many times did calculus 1 (*qtCalculus1*); How many times did calculus 2 (*qtCalculus2*); How many times did physics 1 (*qtPhysics1*); How many times did physics 2 (*qtPhysics2*); Incoming age in the major (*incomingAge*); Distance from home to campus (*campusDistance*); Racial level (*racialLevel*); Father's name informed ? (*dadNameInformed*); Foreign student ? (*foreingStudent*); Student was teacher's assistant sometimes ? (*monitoringStudent*); Student studied in public school ? (*publicSchool*); Student not born in Brasilia ? (*notBornInDf*); Naturalized student in Brazil ? (*naturalizedStudent*) and Bachelor's major ? (*bachelorCourse*). In addition to this information, there is also the variable that represents the student's Major dropout, *droppedOUT*, which was the goal of this work.

Since the distance of the address given by the student to the campus was a desired item of information, geoprocessing software from CODEPLAN, called GEOCODE [19] was used. This software receives as input a list of ZIP codes and returns the latitude and longitude coordinates of the address. With this information and from the coordinates of the UnB campus, it was possible to calculate the Euclidean distance between two points.

The rest of this section is organized as the CRISP-DM steps: Data Understanding, Preparation of Data, Modeling and Evaluation.

A. Data Understanding

Students' data were previously collected with SQL language from the DBMS SQL Server, using the following filters: year from 2009, undergraduate level, students of Darcy Ribeiro campus, non-zero address zip code with eight positions, and formed or dropout students. The zip code filter removed 1240 students from the total. In the data understanding phase, postal Zip numbers that were not found by CODEPLAN software were disregarded, since there was only interest in students whose distance to the campus was known. In addition, since the software only recognizes ZIP code numbers for the Federal District of Brazil, addresses reported outside this federation unit were also discarded. In addition, any form of departure other than graduation was considered to be an evasion. Only students that had relevant and valid social studies validated by the University Student Aid Commission were considered in the help request variable.

The distance of the student to the Campus was observed from Statistical graphs. Figure 1 shows that most students live less than 30 km from the Campus. It also can be analyzed that the longest distance was less than 50 km, because GEOCODE only gets zip codes from city of Brasilia.

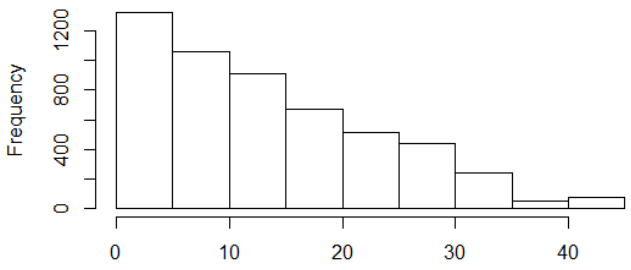

Fig. 1. Histogram of distance (km) to campus

Another interesting analysis could be made by the *incomingAge* versus *droppedOUT* box plot graphic in Figure 2. It can be seen that students who drop out the course are older than those who don't. It may be that this happens due to their work or due to sons care, but the University could take some actions for these students.

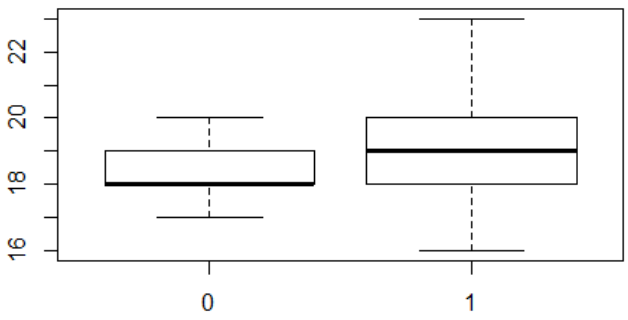

Fig. 2. Boxplot: incoming age x drop out course (0-retention / 1- drop out)

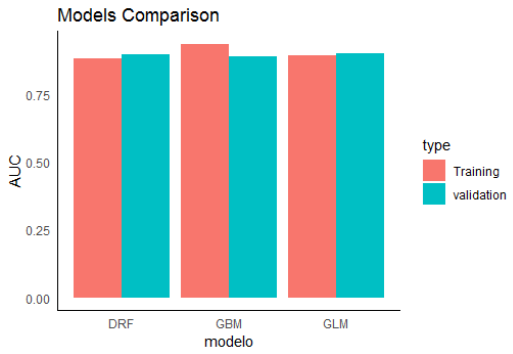

Fig. 3. Models comparison by AUC value

qtLeaveAbsenceMajor.<3	2,1041	POS	monitoringStudent.1	1,4460	NEG
foreingStudent.1	1,7122	POS	rangePreviousCourses.0	1,2711	NEG
rangePhysics1.>2	1,6394	POS	naturalizedStudent.1	0.7779	NEG
bachelorCourse.1	1,5656	POS	rangeAge.<20	0.7411	NEG
monitoringStudent.0	1,3574	POS	rangeGradesSR.0	0.7255	NEG
rangeGradesSR.>5	0.8724	POS	rangeGradesTR.0	0.7179	NEG
rangePreviousCourses.2-5	0.8593	POS	rangeFailuresOBR.1	0.6590	NEG
rangeFailuresOBR.>3	0.7822	POS	rangeAge.21-30	0.6156	NEG
rangeAge.41-50	0.6510	POS	rangeGradesCC.>10	0.5509	NEG
rangeCalculus1.1	0.6037	POS	rangeDistance.41-50	0.5299	NEG
rangeCalculus2.>2	0.5975	POS	rangeFailuresOBR.2	0.3829	NEG
racialLevel.1	0.4159	POS	racialLevel.0	0.3632	NEG
rangePreviousCourses.>5	0.3594	POS	rangePhysics2.1	0.3597	NEG
rangeDistance.21-30	0.2996	POS	rangePhysics1.0	0.3517	NEG
rangeGradesCC.<10	0.2960	POS	rangeHelpRequests.<3	0.3169	NEG
rangeDistance.31-40	0.2679	POS	rangePhysics1.1	0.2801	NEG
rangeGradesTR.>5	0.2164	POS	rangeCalculus2.1	0.2776	NEG
rangeGradesSR.3-5	0.1666	POS	rangeGradesSR.<3	0.2042	NEG
rangeGradesTR.3-5	0.1348	POS	dadNameInformed.0	0.1437	NEG
dadNameInformed.1	0.1239	POS	rangeCalculus1.>2	0.1192	NEG
rangeFailuresOBR.0	0.0920	POS	publicSchool.0	0.0631	NEG
publicSchool.1	0.0665	POS	rangeDistance.<20	0.0241	NEG
notBornInDf.1	0.0276	POS	notBornInDf.0	0.0229	NEG
rangeAge.31-40	0.0085	POS			

Fig. 4. Coeficients from GLM Model

```
Confusion Matrix (vertical: actual; across: predicted)
         0   1    Error        Rate
0      451 131 0.225086   =131/582
1       59 397 0.129386    =59/456
Totals 510 528 0.183044  =190/1038
```

Fig. 5. Confusion Matrix from GLM Model

B. Preparation of Data

In the Preparation of Data step, some variables, such as *droppedOut*, *foreignStudent*, *dadNameInformed*, *publicSchool* and *naturalizedStudent* had to be categorized by R code [20], because as they were considered them as numerical. In addition, some variables were also created in ranges of values, for example *incomingAge* was converted to *rangeAge* and *campusDistance* was converted to *rangeDistance*. By the end of the preparation of data, the data were ready to be used in the H2O algorithms. As will be seen in following section, some H2O models convert the range values in new separated variables of binary values.

C. Modeling

Due to the nature of the objective of the study, that is, prediction of dropout rate from data in which the graduation results are already known, a model for supervised classification was used.As a test plan, data were separated into frames and divided into training, validation and test data, with a ratio of 50%, 30%, and 20%. The test data was only used in the chosen model.

Three algorithms were used from the H2O package [21] on their default parameters: Generalized Linear Model (GLM), Gradient Boosting Machine (GBM) and Random Forest (RF). The GLM was the chosen model due to its accuracy (AUC) for the validation frame. Accuracy is the number of correct predictions divided by the total number of predictions. The AUC values for the three models can be seen in Figure 3.

For the GLM, the coefficients from the model are listed in Figure 4. Figure 4 also shows that the GLM algorithm converts the range values in new separated variables of binary values.

The confusion matrix in Figure 5 shows an accuracy of almost 90% for the GLM model for validation data. This value can be considered satisfactory, but analysis could be made to check if it is an overfitted model .

D. Evaluation

Figure 6 shows that when we used the test frame in GLM, which contains data not yet tested for the model, it had 86.56% of accuracy. That can be considered a good value, since it represents that almost only one from ten predictions is incorrect, and nine are correct.

IV. Results

When the active students frame was applied to GLM, which represents the students that are still studying at UnB, the model predicted a 57% drop out rate in the next years for these engineering course students.

An output csv file was created by RStudio with the probabilities of dropout. Thus, another file it could be generated with the students enrollments and their chances of dropping out or not. When the probability from the GLM model is higher than 50%, it is considered that the student will probably drop out of the Engineering majors.

```
MSE:  0.1478742
RMSE:  0.3845441
LogLoss:  0.4582488
Mean Per-Class Error:  0.20148
AUC:  0.8656126
Gini:  0.7312252
R^2:  0.3992617
Residual Deviance:  945.8255
AIC:  1041.826

Confusion Matrix (vertical: actual; across: predicted)
         0   1    Error       Rate
0      490  90 0.155172    =90/580
1      112 340 0.247788   =112/452
Totals 602 430 0.195736  =202/1032
```

Fig. 6. Validation Performance data resume from GLM Model

This analysis can be used from the University of Brasilia to monitor these students, and adopt some help policies if needed. For example, some help-policy could be created for international students, maybe a Portuguese course would be useful for this issue. On the other hand, this study could be used by other researchers as a baseline for future analysis in this area.

V. Conclusion and Future Works

This paper presented a methodology for analyzing the predictive dropout of undergraduate engineering students at the University of Brasilia. The proposed methodology is based on CRISP-DM and used a dataset of social and performance data about 5289 engineering students, graduated or not, but no longer studying in the University. For training the models and from about 3071 students still active to implement the chosen model. The GLM model was chosen after three models created, trained and compared and a prediction was made from real enrolled students.

As conclusion, the GLM predicted that almost 57% of the engineering students from UnB might drop out of the major in the next years. The model also predicts that the international students need some attention from the university, as they were noted as critical class of students related to low retention.

Another interesting conclusion from GLM is that those who do the Physics 1 course more than two times and that those apply for a leave of absence from major more than three times have greater chance of not graduating in the future. As this work also generated a file that aims to predict real drop out from active students, UnB could use this as a direction to create help policies in order to mitigate this academic issue. pattern values from the algorithms, the grid function to tune the parameters could be used. Data from other campi from UnB could be collected with larger time range.

Future work includes a new study improving H2O models by using not just the pattern values from the algorithms, the grid function to tune the parameters could be used. Data from other campi from UnB could be collected with larger time range. In relation to parallelism, a Hadoop cluster could be analyzed to run Sparkling Water as a Yarn application and be compared to the Spark cluster used in this work.

Acknowledgments

Maristela Holanda thanks the Fundação de Apoio a Pesquisa do Distrito Federal (FAP-DF) Brazil, Edital UnB-DPG 05-2018 Bolsa de Pos-doutorado no Exterior FAP-DF, for supporting this work.

References

[1] R. S. Baker and P. S. Inventado, "Educational data mining and learning analytics," in *Learning analytics*, pp. 61–75, Springer, 2014.
[2] C. Romero and S. Ventura, "Educational data mining: A survey from 1995 to 2005," *Expert systems with applications*, vol. 33, no. 1, pp. 135–146, 2007.
[3] A. Dutt, M. A. Ismail, and T. Herawan, "A systematic review on educational data mining," *IEEE Access*, vol. 5, pp. 15991–16005, 2017.
[4] A. Peña-Ayala, "Educational data mining: A survey and a data mining-based analysis of recent works," *Expert systems with applications*, vol. 41, no. 4, pp. 1432–1462, 2014.
[5] M. Zaffar, S. Iskander, and M. A. Hashmani, "A study of feature selection algorithms for predicting students academic performance," *Int. J. Adv. Comput. Sci. Appl*, vol. 9, no. 5, pp. 541–549, 2018.
[6] L. Zhang and H. Rangwala, "Early identification of at-risk students using iterative logistic regression," in *International Conference on Artificial Intelligence in Education*, pp. 613–626, Springer, 2018.
[7] S. M. M. Rubiano and J. A. D. Garcia, "Formulation of a predictive model for academic performance based on students' academic and demographic data," in *2015 IEEE Frontiers in Education Conference (FIE)*, pp. 1–7, IEEE, 2015.
[8] B. K. Francis and S. S. Babu, "Predicting academic performance of students using a hybrid data mining approach," *Journal of medical systems*, vol. 43, no. 6, p. 162, 2019.
[9] M. Agaoglu, "Predicting instructor performance using data mining techniques in higher education," *IEEE Access*, vol. 4, pp. 2379–2387, 2016.
[10] E. B. Costa, B. Fonseca, M. A. Santana, F. F. de Araújo, and J. Rego, "Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses," *Computers in Human Behavior*, vol. 73, pp. 247–256, 2017.
[11] D. Roy, P. Bermel, K. Douglas, H. Diefes-Dux, M. Richey, K. Madhavan, and S. Shah, "Synthesis of clustering techniques in educational data mining," in *ASEE Annual Conference & Exposition, Columbus, Ohio, USA*, 2017.
[12] M. Barbu, R. Vilanova, J. Vicario, M. J. Pereira, P. Alves, M. Podpora, A. Kawala-Janik, M. Prada, M. Dominguez, A. Spagnolini, *et al.*, "Data mining tool for academic data exploitation: publication report on engineering students profiles," *ERASMUS+ KA2/KA203*, 2019.
[13] Y. Bouslimani, G. Durand, and N. Belacel, "Educational data mining approach for engineering graduate attributes analysis," *Proceedings of the Canadian Engineering Education Association (CEEA)*, 2016.
[14] B. Pérez, C. Castellanos, and D. Correal, "Predicting student drop-out rates using data mining techniques: A case study," in *IEEE Colombian Conference on Applications in Computational Intelligence*, pp. 111–125, Springer, 2018.
[15] R. M. Bautista, M. Dumlao, M. A. Ballera, and V. A. S. B.-Q. City, "Recommendation system for engineering students specialization selection using predictive modeling," in *The Third International Conference on Computer Science, Computer Engineering, and Social Media (CSCESM2016)*, p. 34, 2016.
[16] E. Fernandes, M. Holanda, M. Victorino, V. Borges, R. Carvalho, and G. Van Erven, "Educational data mining: Predictive analysis of academic performance of public school students in the capital of brazil," *Journal of Business Research*, vol. 94, pp. 335–343, 2019.
[17] L. C. B. Martins, R. N. Carvalho, R. S. Carvalho, M. Holanda, *et al.*, "Early prediction of college attrition using data mining," in *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 1075–1078, IEEE, 2017.
[18] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, "Crisp–dm 1.0 step-by-step data mining guide," 2000.
[19] "Geocode." http://geocode.codeplan.df.gov.br/. (Accessed on 05/10/2019).
[20] "R project." https://www.r-project.org. (Accessed on 05/10/2019).
[21] "H2o.ai." http://www.h2o.ai/. (Accessed on 05/10/2019).