2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT)

Supervised Learning in the Context of Educational Data Mining to Avoid University Students Dropout

Kelly J. de O. Santos
Department of Computer Science
Federal University of Sergipe
Aracaju, SE, Brazil
kelly.joany@gmail.com

Angelo G. Menezes
Department of Computer Science
Federal University of Sergipe
Aracaju, SE, Brazil
angelomenezes.eng@gmail.com

Andre B. de Carvalho
Department of Computer Science
Federal University of Sergipe
Aracaju, SE, Brazil
andre@ufs.br

Carlos A. E. Montesco
Department of Computer Science
Federal University of Sergipe
Aracaju, SE, Brazil
estombelo@ufs.br

***Abstract*—Educational data mining is a research field that looks for extracting useful information from large educational datasets. This area provides tools for improving student retention rates around the world. In this paper we propose a computational approach using educational data mining and different supervised learning techniques (Decision Trees, K-nearest Neighbor, Neural Networks, Support Vector Machines, Naive Bayes and Random Forests) to evaluate the behaviour of different prediction models in order to identify the profile of at-risk university students in a Brazilian university environment. The results of this paper indicate that some algorithms can be used as tools for supporting decisions that reduce school dropout.**

***Index Terms*—Educational Data Mining, Big Data, Machine Learning, University Dropout.**

I. Introduction

According to the Census of Higher Education of Brazil 2015, 11% of the students who concluded high school in 2010 dropped out already in the first year. By 2014, almost half (49%) of students left the majors they had chosen in 2010 [1]. Thus, student dropout is one of the major problems that affect Brazilian educational institutions in general.

In Brazil, there is an effort to sustain and support basic education. There are already some contributions that point out the interest of the Brazilian researchers in this area, but such contributions are still in scarcity [2]. Despite the difficulties that are present in this scenario, data mining and machine learning have been gaining more and more relevance in scientific studies and research applied to the educational field.

II. Related Work

Marquez-Vera et al. [3] proposed a study of educational data mining (EDM) techniques to predict school dropout comprising 670 high school students in Zacatecas, Mexico. The results obtained accuracy between 75% and 98% of the ten selected classifiers for this case study when applied in different contexts in the same experiment.

Lam-On and Boongoen [4] pointed out that databases often add many redundant attributes which and proposed the implementation of an ETL framework before applying the selected algorithms. After it was possible to obtain accuracy results around 92% during the research.

For this work, we have selected for evaluation the main EDM algorithms: Decision Tree, K-nearest Neighbor (KNN), Neural Networks (NN), Support Vector Machine (SVM), Naive Bayes (NB) and Random Forests (RF). According to [5], these algorithms are among the most used ones in EDM.

III. Methodology

The processing steps were as follows to obtain the results.

1) Data acquisition from Federal University of Sergipe (UFS) [6];
2) Data pre-processing;
3) Feature selection based on accuracy contribution;
4) Application of the selected algorithms;
5) Evaluation and analysis of results.

About data acquisition: 23,690 students from the Department of Computer Science in the UFS were selected. Among these, Computer Science (CS) courses obtained 12,079 records, followed by Information Systems (IS) with 5,592 students and Computer Engineering (CE) with 5,389 students. These data included records from 2010 until 2018 related to courses from the first until the sixth semester in each major.

In the data pre-processing step, we first checked for missing elements, applied semantic analysis and then normalization.

For the classification step, parameters were chosen based on the default parameter settings of the Scikit-learn library. Also, the features were selected based on how they contributed to the accuracy of all algorithms, and K-Fold cross validation with a five subset split was applied to the model.

IV. Results

After analyzing the accuracy of all algorithms in each semester, the average accuracy could be used as means of comparison. Table I shows the best related results.

TABLE I
Average accuracy for the best algorithms

Course	Best Algorithms	Average Accuracy
Computer Science	- Decision Tree - Random Forest	66%
Information System	- Random Forest - SVM - Decision Tree	70%
Computer Engineering	- Decision Tree	72%

2161-377X/19/$31.00 ©2019 IEEE
DOI 10.1109/ICALT.2019.00068

The accuracy of all algorithms for the three courses (IS, CS and CE) in each semester (1^o to 6^o semester) are shown in graphics of Figures 1, 3 and 2, respectively.

The difference in accuracy is directly related to the amount of presented subjects in each semester. As presented in Figure 1, the third semester for the IS program, which has currently 6 courses, showed the lowest accuracy while the fourth semester, which has only 4 subjects, had a higher accuracy. This points out to the fact that a higher number of courses may lead to a decrease in students' performance and failing a course early in a program could incidentally influence their behavior and expectations with the major.

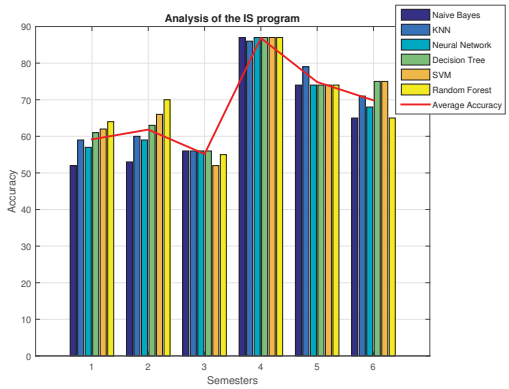

Fig. 1. Accuracy of the students dropout from IS program

As it is presented in Figure 2, the lowest accuracy for the CE is presented in the first 3 semesters. This may be related to the fact this major has different groups of subjects being taken all at once (math, computer science, electronics, ...) in the beginning of the program and not all students may be prepared to that.

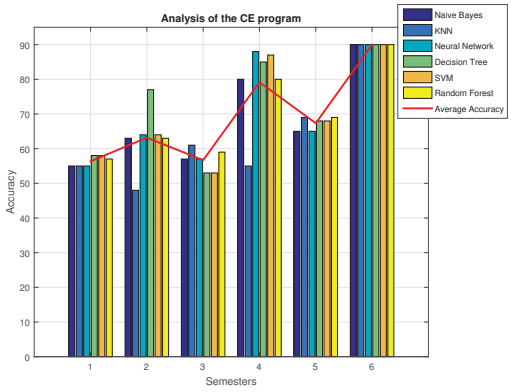

Fig. 2. Accuracy of the students dropout from CE program

In Figure 3, it can be observed that accuracy was the lowest on the second semester for the CS program. This may be explained by the fact that many math and logic subjects are introduced and studied altogether. In the third, fourth and sixth semesters, the accuracy has an up trend which may point out to the fact that students are more comfortable with technical subjects.

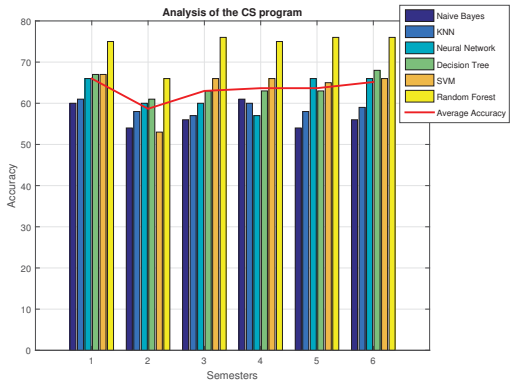

Fig. 3. Accuracy of the students dropout from CS program

V. Conclusion

This paper presented the analysis of different supervised learning techniques in the context of educational data mining for university students dropout avoidance. It was found that students dropped from their respective programs more frequently in the 4^o semester for the CE and IS programs, and in the 6^o semester for the CS one since those were the semesters where the algorithms had their best results.

While some algorithms presented solid results for every semester (RF and Decision Trees), some of them were not able to achieve such high outcomes (i.e., NN) most likely because they need more data or a parameter tuning step in order to be more robust to fit the data. Therefore, as future work, it is intended to expand the parameter tuning session using grid search or an evolution based technique for this work.

References

[1] I. Teixeira, *Instituto Nacional de Estudos e P. E. A. Censo da educao superior 2015*, 2015.

[2] L. Manhaes, S. Cruz, and G. Silva, "Wave: An architecture for predicting dropout in undergraduate courses using edm," *Proceedings of the ACM Symposium on Applied Computing*, 03 2014.

[3] C. Marquez-Vera, C. R. Morales, and S. V. Soto, "Predicting school failure and dropout by using data mining techniques," *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje*, vol. 8, no. 1, pp. 7–14, Feb 2013.

[4] N. Lam-On and T. Boongoen, "Using cluster ensemble to improve classification of student dropout in thai university," *2014 Joint 7th International Conference on Soft Computing and Intelligent Systems (SCIS) and 15th International Symposium on Advanced Intelligent Systems (ISIS)*, pp. 452–457, Dec 2014.

[5] R. Machado, E. Nara, J. Schreiber, and G. Schwingel, "Estudo bibliometrico em mineracao de dados e evasao escolar," *Congresso Nacional de Excelencia em Gestao*, 08 2015.

[6] A. G. Barroca Filho I. and J. Santa Rosa, "Sigaa mobile – o caso de sucesso da ferramenta de gestão acadêmica na era da computação móvel," in *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, vol. 24, no. 1, 2013, p. 92.