Predicting Student Drop-Out Rates Using Data Mining Techniques: A Case Study

Boris Pérez[1,2(✉)], Camilo Castellanos[2], and Darío Correal[2]

[1] Univ. Francisco de Paula Stder., Cúcuta, Colombia
borisperezg@ufps.edu.co
[2] Universidad de los Andes, Bogotá, Colombia
{cc.castellanos87,dcorreal}@uniandes.edu.co

Abstract. The prevention of students dropping out is considered very important in many educational institutions. In this paper we describe the results of an educational data analytics case study focused on detection of dropout of Systems Engineering (SE) undergraduate students after 6 years of enrollment in a Colombian university. Original data is extended and enriched using a feature engineering process. Our experimental results showed that simple algorithms achieve reliable levels of accuracy to identify predictors of dropout. Decision Trees, Logistic Regression, Naive Bayes and Random Forest results were compared in order to propose the best option. Also, Watson Analytics is evaluated to establish the usability of the service for a non expert user. Main results are presented in order to decrease the dropout rate by identifying potential causes. In addition, we present some findings related to data quality to improve the students data collection process.

Keywords: Student drop out · Student desertion prediction
Educational data mining · Prediction models

1 Introduction

In every country, education is synonymous with rapid economic growth [13]. Universities can play a key role as builders of qualified human capital and innovation systems in their countries [5]. Also, they help supporting the economic growth of a country [13]. High level of education along with more students completing their studies are the required conditions to improve level of human capital in the society. The completion of university is correlated to the increase of life expectancy, increase in social status, reduction of the risk of unemployment, among others [10].

There has been increasing interest from universities in understanding the behavior of successful students. The reputation of these institutions is measured by the percentage of students who graduate and by the strategies the university has to retain its students [16]. The early identification of students who are at risk of dropping out is critical for the success of any retention strategy [16]. Then

A. D. Orjuela-Cañón et al. (Eds.): ColCACI 2018, CCIS 833, pp. 111–125, 2018.
https://doi.org/10.1007/978-3-030-03023-0_10

it becomes necessary to detect these students as early as possible, maintaining intensive and continuous intervention to reduce the dropout levels [15,18].

The identification of dropout rates has become a major issue for Universities. In Colombia, there is an initiative from the Ministerio de Educación called SPADIES (System for Prevention and Analysis of Desertion in Institutions of Higher Education) [11]. This initiative was designed by the Center for Economic Studies (SEDE) at the University of the Andes to follow up on the problem of dropout in higher education, to calculate the risk of desertion of each student, and to classify these students by groups. This initiative can support the evaluation of strategies for each of the situations that influence dropout such as student status, academic program and institution; and also promotes the consultation, consolidation, interpretation and use of this information (tables and graphs, each by various criteria).

In the literature, researches have been investigating the finding of the characteristics of both the student and his context influencing his decision to drop out of university. Tinto's model [21] is the most widely accepted model in student retention literature. Tinto concluded that the student dropout is strongly related to their degree of academic (grade performance and intellectual development) and social integration (peer-group interactions and faculty interactions) at university. Bharadwaj and Pal [3,4] identified that attributes of the score in a senior secondary exam, residence, various habits, annual family income, and family status were shown to be important parameters for dropout. Additionally, Kovacic [14] identified that enrollment data could be used to predict dropout. Variables like students attendance in class, hours spent studying after class, family income, mother's age and mother's education are significantly related to student dropout. Specifically, it has been found that the factors like mother's education and family income are highly correlated with student performance [9].

However, there is still no consensus in the literature on the causes of dropout in universities. Despite the existence of multiple studies on university dropout, there is a little research related to the computer science area. Also, the vast majority of studies are focused on static variables, leaving aside the dynamic component of the grades the student obtains during his studies. It is fundamental to investigate the causes that lead students of Systems Engineering program to withdraw the course before its completion.

Thus, considering the factors that influence the dropout rates in universities, the aim of the present study is to answer the following questions: (a) what are the key determinants of undergraduate student dropout rates in a Systems Engineering program of a Colombian Private University? And, (b) what data mining technique is more suitable to find these key determinants? To do so, we model student dropout using data gathered from academic databases from 2004 to 2010. In addition, this paper presents the differences between using a programmatic approach to identify the key determinants versus the automatic approach offered by Watson Analytics from IBM. We do not take into account data related to the enrollment process like demographic information. Our approach considered a significant population of a private university.

This paper is organized as follows: Sect. 2 reviews the related work. Section 3 introduces the data mining methodology. Section 4 describes the data set used. Section 5 presents the data preparation. Section 6 describes the modeling process. Section 7 reports results. Section 8 offers a preliminary proposal to use these results in a real context. Finally, Sect. 9 outlines the conclusions.

2 Related Work

Data mining is the area which analyzes huge repositories of data to extract important patterns, association and relations among all these and is therefore a valuable tool for converting data into usable information [9]. Data mining can discover hidden information in various domains, including marketing, banking, educational research, surveillance, telecommunications fraud detection, and scientific discovery. Education is one of these domains where the primary concern is the evaluation and, in turn, enhancement of educational organizations [20].

Educational Data Mining (EDM) is a discipline engaged with the develop of methods and techniques not only for exploring and analyzing the data that come from educational context but also for extracting hidden information for better understanding students. This information can be used in several educational processes such as predicting course enrollment, estimating student dropout rate, detecting atypical values in students' transcripts, and improvement of student models that would predict student's characteristics or academic performances [20,23].

Tinto's model [21] is the most widely accepted model in student retention literature. Tinto concluded that the decision of students to persist or drop out of their studies is strongly related to their degree of academic and social integration at university. In the academic system, Tinto analyzed the grade performance and intellectual development. In the social system, peer-group interactions and faculty interactions were also analyzed. However, Brunsden et al. in [6] tested the model with path analysis using LISREL8 software and concluded that Tinto's model may not be the most appropriate for dropout research. They

Bharadwaj and Pal [4] used EDM to evaluate student performance among 300 students from five different colleges who were enrolled in an undergraduate computer course. They employed a Bayesian classification scheme of 17 attributes, of which the score in a senior secondary exam, residence, various habits, annual family income, and family status were shown to be important parameters for academic performance. In a second study, Bharadwaj and Pal [3] constructed a new data set which included student attendance, and test, seminar, and assignment grades in order to predict academic performance. A similar study was proposed by Kovacic [14], who applied EDM to identify which enrollment data could be used to predict student academic performance. In this study, he used CHAID and CART algorithms on a dataset of student enrollment.

In another study, Al-Radaideh et al. [1] analyzed student's academic data (student gender, student age, student department, high school grade, lecturer degree, lecturer gender, among others) building a classification model using the

decision tree method to improve the quality of the higher educational system. They found that high school grade was the attribute with the highest gain ratio and was considered the root node of the decision tree. The Holdout method and the K-Cross-Validation method (k-CV) were used to evaluate the model. However, they found that the collected samples and attributes were not sufficient to generate a classification model of high quality.

Gerben et al. [8] conducted a case study in which they used machine learning techniques to predict student success using features extracted from student pre-university academic records. Their experimental results showed that simple and intuitive classifiers (i.e.: decision trees) gave useful results with accuracies between 75% and 80%. One of their findings was that the strongest predictor of success was the grade for the Linear Algebra course, which was not seen as the decisive course.

Despite these studies, it's not clear which data mining algorithms are preferable in this context. For example, Luan in [12] built predictors using clustering as means of data exploration and classification. In [17], Romero and Ventura presented a survey on EDM where one of their findings was that association analysis has become a popular approach. Finally, Herzog in [19] presented the results of a case study where Bayesian networks and neural networks were outperformed by decision tree algorithms but on small educational datasets.

3 Methodology

The analytic task performed in this work is a binary classification task where dropout (0, 1) is the target or dependent variable. To do this, we follow the Cross Industry Standard Process for Data Mining methodology (CRISP–DM) [7]. This methodology is useful for planning, communicating the project team, and documenting. It provides a generic check-lists which advises the steps to be taken and provides practical advice for all steps. The life cycle of a data mining project [22] is broken down into six phases as presented in Fig. 1.

The phases are described below. *Business understanding* allows definition of the business goal, in our case the student dropout phenomenon, covered in the previous sections. *Data understanding* involves data collection, identification of data quality problems and discovering of insights. *Data preparation* covers feature extraction, data wrangling, and can require multiple iterations. *Modeling* consists of technique selection, application and calibrating parameters. There is a close link between Data Preparation and Modeling. Often, one realizes data problems while modeling and gets ideas for constructing new data. *Evaluation* is focused on the performance assessment of the models built in the previous phase. *Deployment* deals with the operationalization of the model within the real context. These phases will be tackled in the following sections.

For data understanding, data preparation and modeling, we use an open source tool in Python as our data science development environment: Jupyter

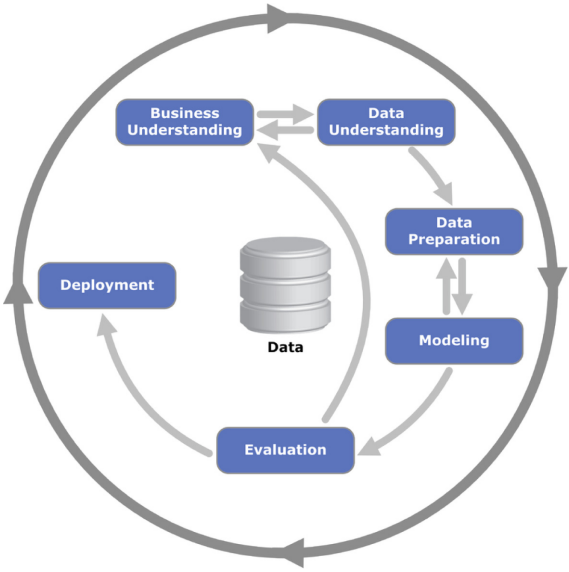

Fig. 1. Cross industry standard process for data mining methodology

notebook[1], Python data analysis library (Pandas[2]) to deal with data structures, Scikit-learn[3] (machine learning library), Seaborn[4] a statistical data visualization tool and Graphviz[5] a graph visualization software to generate the decision tree charts.

4 Dataset

The data set used in this study comes from 762 students enrolled in the Systems Engineering Program at a private university in Bogotá, Colombia. The data is organized in three tables, and together they include the following features:

- Admission information, including minimum demographic information (gender, birth date, marital status).
- Graduation dates, including date of graduation and the academic program.
- Transcript records including the courses taken and the grades for each of them, the academic program and the academic cumulative average.

Our focus was on students entering the university from first term of 2004 through the second term of 2010. Although the institutional databases has the

[1] http://jupyter.org.
[2] http://pandas.pydata.org.
[3] http://scikit-learn.org.
[4] http://seaborn.pydata.org.
[5] http://graphviz.org.

latest student enrollment data, 2010 was chosen as the last year for analyses since student graduation was defined as six years from enrollment. The confidentiality of data was preserved by not using any personal data like Colombian national ID number, date of birth, campus wide identification number, or name. The overall graduation rate in this dataset was 52.87%. The rate of dropout was 47.13%.

The data was assumed to be completely independent, which means that the effect of each variable on the target variable (graduation) is not affected by any other variable. Also, missing data was assumed to be missing completely at random.

5 Data Preparation and Exploration

In this phase, we applied data cleaning and wrangling in order to transform original data set into another format with the purpose of making it more appropriate and valuable for modeling process.

Firstly, we joined the three source files: admission information, transcript records and graduation dates. Then, we performed a data profiling to get descriptive statistics which help us to understand the data.

The approach we used is grouping courses within similar academic field (systems engineering (SE), mathematics (MATH), physics (PH), management (MGMT), Language (LAN), Biology (BIO), etc.). For this, we aggregated the course grades, and course repetitions by student and by faculty. Additionally, we added the student age at the enrollment time and the standard deviation of academic term cumulative average to reflect the variance (irregularity) of academic performance. As a result, we obtained an aggregate dataset with 31 columns which contain the following relevant fields: gender, marital status, age at enrollment, academic terms (by academic field), academic cumulative average, standard deviation of academic term averages, course grades averages, course repetitions by faculty, and the dropout indicator (0,1).

Figure 2 presents the grade averages by academic fields, segmenting by dropout indicator. As expected, dropout students present lower grade averages across all subject. Management classes are the most challenging for both types of students, followed by mathematics and physics.

Categorical fields such as sex and marital status were transformed to numeric values because the most of models require numeric inputs. To do this transformation, we used dummy variables to represent each categorical value as a binary indicator column. Also, we scaled out the features to normalize the magnitudes and prevent that high magnitude fields skew the feature's weights into the machine learning models.

Although sampled from a diverse university, Fig. 3 shows a clear dominant demographic profile for SE students: single males between 17 and 19 years old, accounting for 71.5% of the dataset.

The correlation between significant variables can be visualized in Fig. 4. As expected, we find strong positive correlation between the averages of the different subjects, with mathematics and physics being the highest. We also have

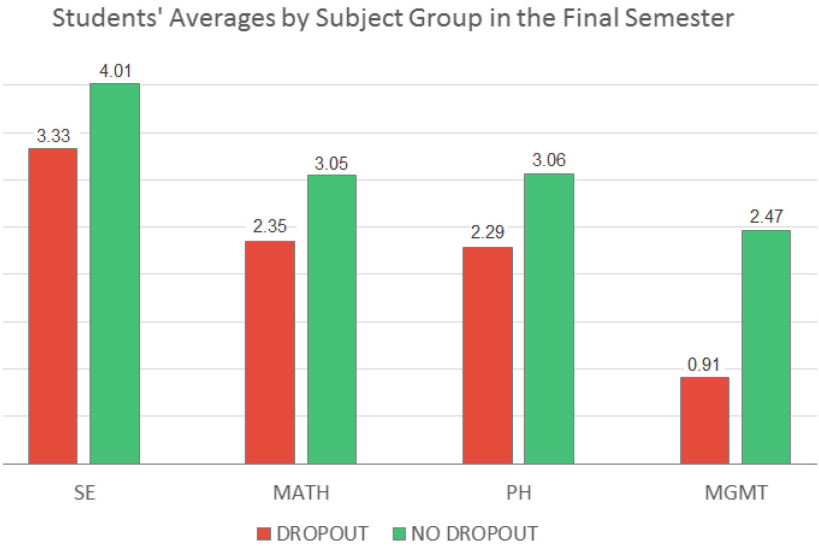

Fig. 2. Average grades by subject group of all students at the end of their program.

that GPA is most influenced by system engineering classes (since compose the majority of classes in the program) and has a negative correlation with the target variable. Surprisingly, there is a strong correlation between the amount of times a student fails a system engineering and a biology class.

As mentioned before, the frequency of dropouts in the dataset is almost equal (52.87% for *no* and 47.13% for *yes*) so no over/under sampling is required.

Finally, since it is of interest for school administrations to detect dropout as early as possible, we include the average grade progression for the first three semesters for both types of students in Fig. 5. Selected subjects where Mathematics, Physics and System Engineering since the represent the majority of courses in the program.

6 Modeling

According to previous works described in Sect. 2, models offering the best accuracy [2,8,14] are: Decision Tree, Logistic Regression and Naive Bayes. In addition, we used Random Forest model as a complemented technique because it is suitable for classification and regression, and it operate by constructing a multitude of decision trees at training time and outputting the class that is the mean prediction of the individual trees. All models where applied to data restricted to the first, second, third and last semester after enrollment.

We applied Cross Validation (CV) in every modeling tasks to avoid overfitting and we used all dataset in training and testing steps. In the approach called *k-fold* CV, the training set was split into k smaller sets, in our case, k=5 folds. And for each fold:

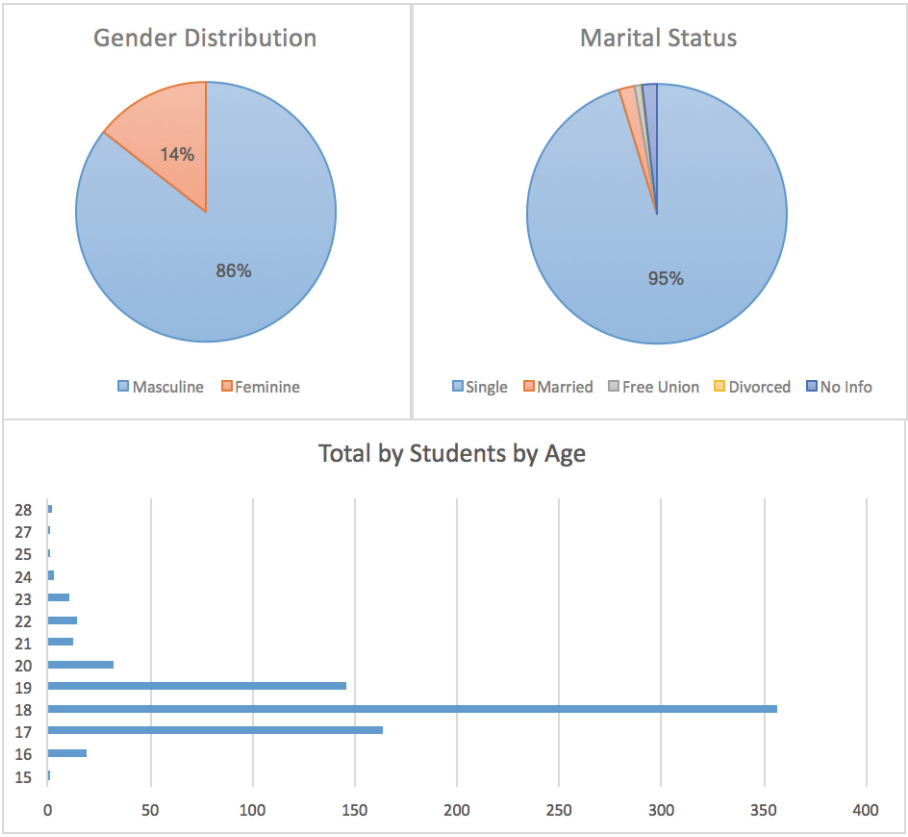

Fig. 3. Student's sex, marital status and age summary.

– A model is trained using *k–1* (4) of the folds as training data (4/5 of data).
– The resulting model is validated with the resting part of the data (1/5).

6.1 Models Setup

– **Decision Tree.** We trained a decision tree model with *gini* criteria and CV mentioned before. The model without pruning contains twelve levels and fifty-one leaf nodes. This tree model was pruned using max depth parameter = 4 levels and we got 7 leaf nodes tree, as shown in Fig. 6.
– **Logistic Regression.** We trained the logistic model with the following parameters: tolerance for stopping criteria = 0.0001, inverse of regularization strength = 1.0, and solver = *liblinear*.
– **Naive Bayes.** We used a Gaussian Naive Bayes algorithm.
– **Random Forest.** We trained this model with the following parameters: number of trees in the forest = 10, maximum depth of the tree = 4 and random state = 0, and a cross-validation generator = 6.

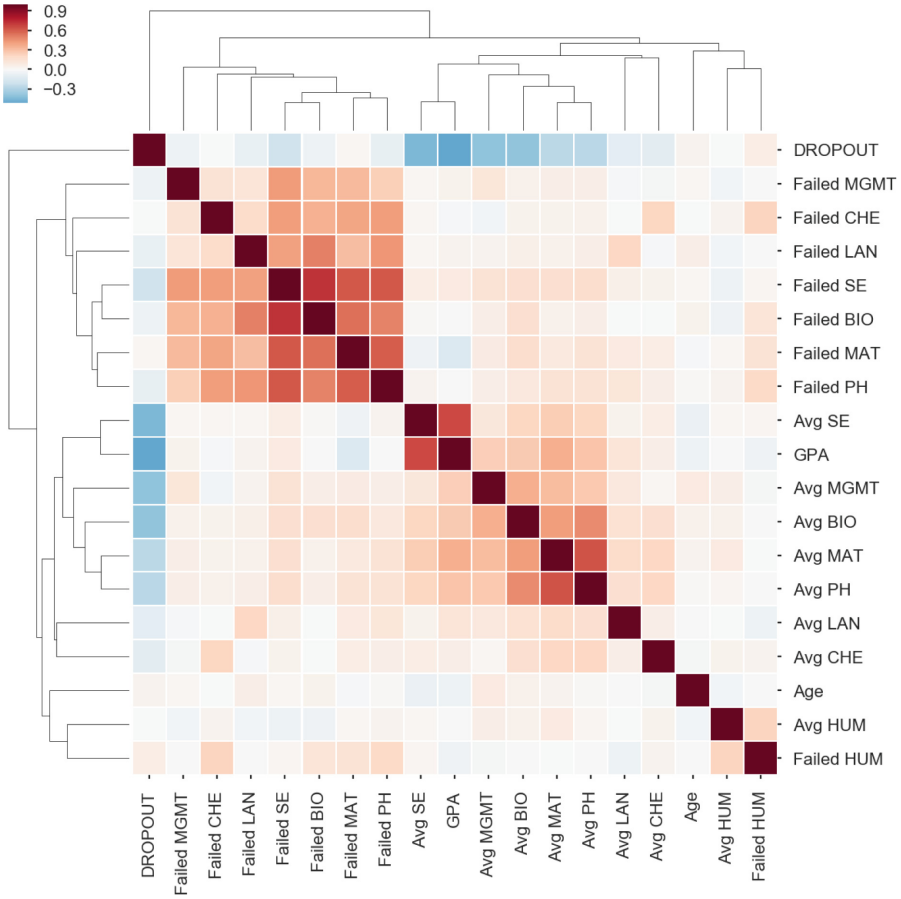

Fig. 4. Cluster map of feature's correlations

6.2 Watson Analytics

Watson Analytics is a smart service for analyzing and visualizing data to quickly discover patterns and meaning in data, without having any previous knowledge. Watson Analytics use guided data discovery, automated predictive analytics and cognitive capabilities to interact with data to get findings you understand. No previous configuration is required.

7 Model Evaluation and Results

In this section, models selected were evaluated in terms of Receiver Operating Characteristic (ROC) curve analysis. In a ROC curve the true positive rate, also called Sensitivity, is plotted in function of the false positive rate (i.e.: 100-Specificity) for different cut-off points of a parameter. Each one of the points on

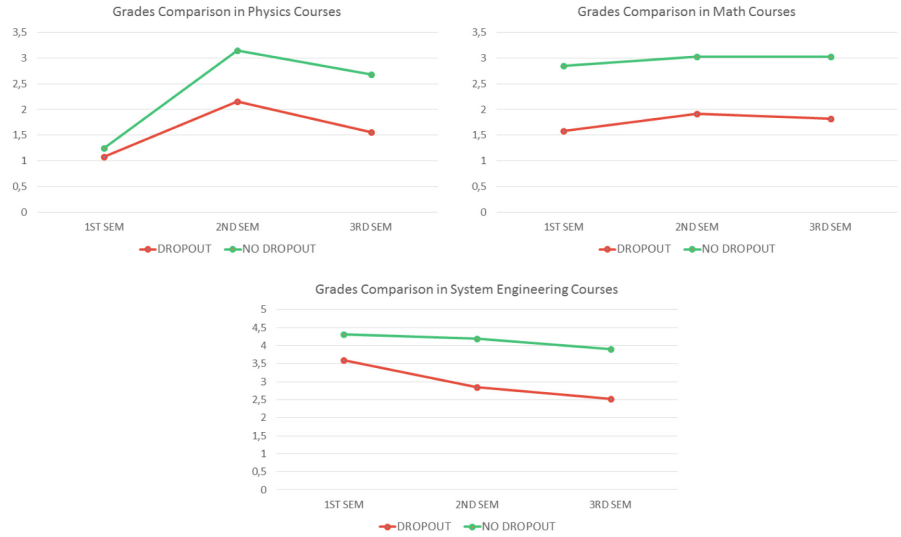

Fig. 5. Grading comparison: dropouts vs. non-dropouts. Tangent behavior among both types of students is very similar, showing differences in that non-dropouts have higher average across the three subjects.

the ROC curve represents a unique sensitivity/specificity pair corresponding to a decision threshold. ROC curve implies that top left corner point is the ideal point (i.e. a false positive rate of zero and a true positive rate of one), so, a larger area under the curve (AUC) will be better. The area under the ROC curve (AUC) is a measure of how well a parameter can distinguish between several groups.

Figure 7 presents the ROC–AUC of the different models evaluated by semester. All models show good results, having above 0.8 AUC value from second semester onwards and as expected, they all show that the longer the student is in enrolled the better the prediction gets. Random Forest shows the best results, giving an 0.91 AUC on the third and a 0.97 AUC on the last semester, making it the ideal model for this case study.

For the Decision Tree model, we found that:

- Case 1: a student with *Avg SE* lower or equal than 3.505, and *Failed SE* lower than or equal to 0.1389 (scaled 0 to 5), and *GPA* lower than or equal to 3.5164, has 100% (143/143) for dropping out.
- Case 2: a student with *Avg SE* lower than or equal to 3.505 and *Failed SE* lower than or equal to 0.1389 (scaled 0 to 5) and *GPA* greater than 3.5164, has 88.8% (48/54) for dropping out.

For the Logistic Regression model, the significant coefficients are displayed in Table 1. As expected, we have that *Failed MAT* and *Failed MGMT* increase the probability of dropping out, opposed to *Avg SE* that reduces the risk of not completing the degree. Surprisingly, *Failed SE* has a negative influence on the

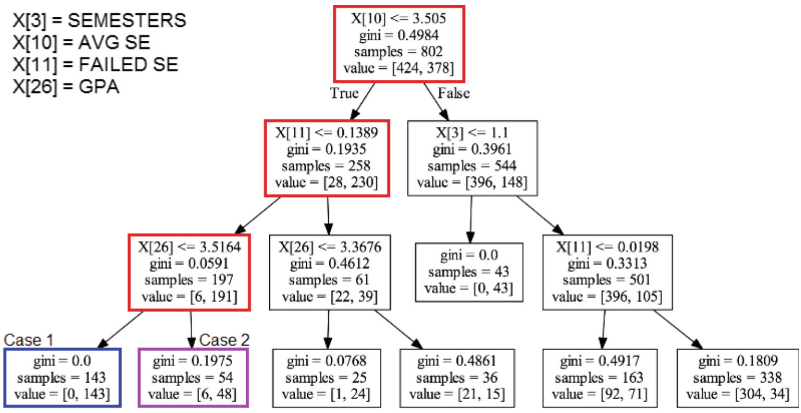

Fig. 6. Decision tree pruned to max depth = 3

dropping out factor. A possible explanation can be giving by taking into account that *Semesters* (the amount of semesters the student has been enrolled) also has a negative coefficient: Apparently, the longer the student is enrolled the lower the risk of dropping out. Hence, since SE subject are concentrated at the end of the program, failing these classes can be a sign of non dropouts students at the end of their careers. Most important, the standard deviation of a students grades has the most positive impact in a student's dropout factor.

Table 1. Significant coefficients for logistic regression (All Semesters)

Feature	Weight
Std GPA	2.84
Failed MAT	1.30
Failed MGMT	0.71
Avg SE	−1.28
Semesters	−1.62
Failed SE	−1.99

Finally, for Watson Analytics, the service produced several charts explaining the findings. We focused in the Decision Tree shown in Fig. 8, where Watson used *GPA* as the primary variable, detecting that 99.9% students below 3.57 drop out from university. Also, having *GPA* between 3.27 and 3.55 and variable *Failed SE* been equal to and lower than 1, students will have an 81% of probability of dropout. Watson uses the variables in a different way to those found by the different models of supervised learning, however, it is evident that the findings found by Watson serve in the same way to carry out the study.

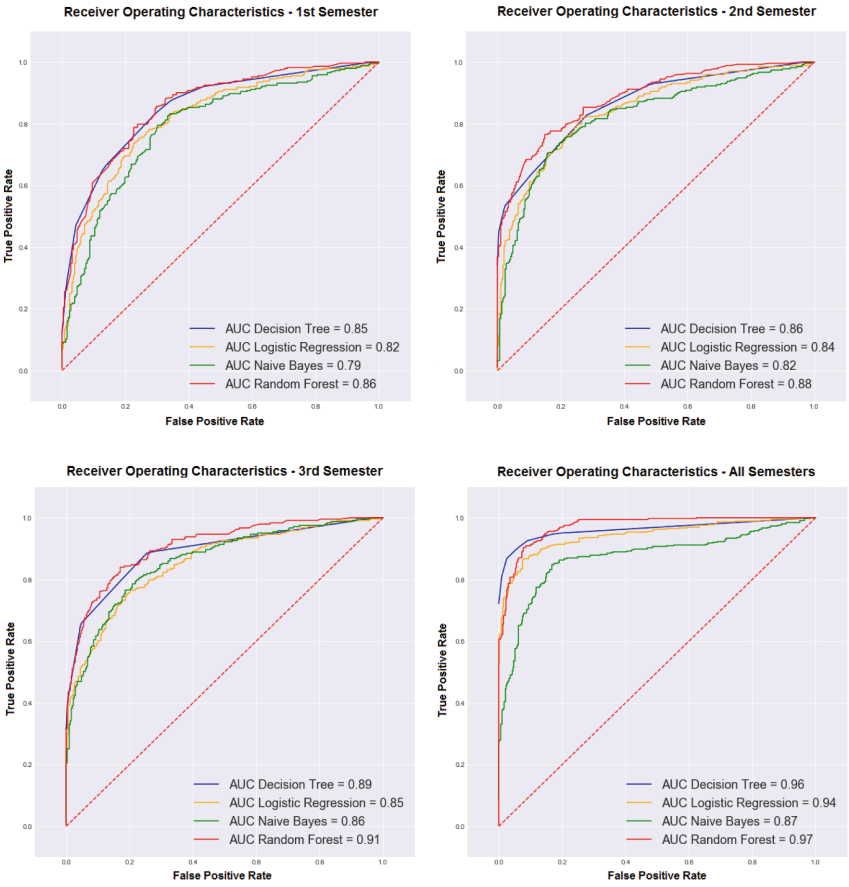

Fig. 7. Models evaluation using ROC - Area under curve

8 Deployment

This phase in CRISP–DM methodology implies to operationalize the model in the real–environment to detect risks and take decisions to prevent drop out rates. Depending on the requirements, this phase can be as simple as generating a report or as complex as implementing a repeatable data mining process. Ideally, it will be the user, not the data analyst, who will carry out the deployment steps.

However, this phase is out of scope of this work, but The findings will be shared and discussed with SE faculty in order to validate and refine the model, and implement it in a productive environment. This implementation may be deployed as a predictive web service to focus on potential dropping outs (based on prediction rules), generate early alerts and treat them properly. Afterward, the feedback of predictions and treatments should be new inputs to upgrade the model.

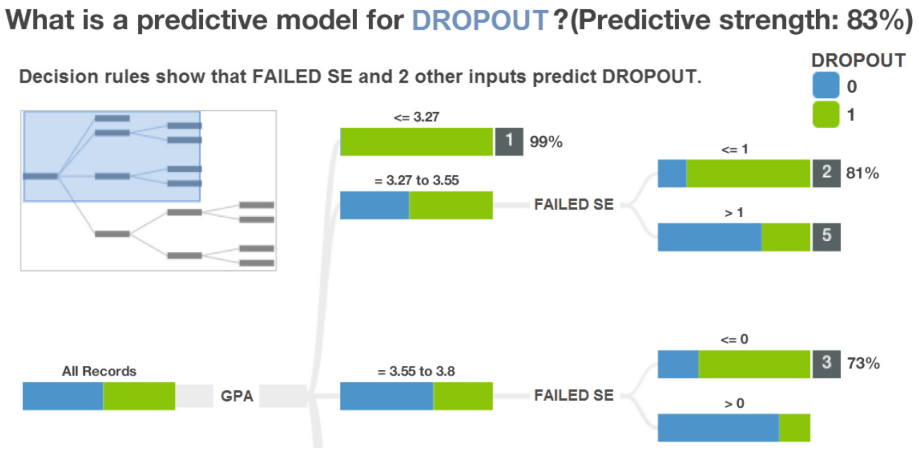

Fig. 8. Decision tree generated by Watson

9 Conclusions

Educational research has taken advantage of data mining. The current pace of applying data mining methods in this domain has increased for a variety of purposes, e.g. assessing student needs, predicting dropout rates, analyzing and improving student academic performance. Student drop out prediction is an important and challenging task.

In this paper, we showed preliminary results for predicting student attrition from a large dataset of student demographics and transcript records at different points in their degrees. In our findings, we discovered that systems engineering courses performance are correlated to physics and mathematics courses performances. The irregularity (standard deviation of term's averages) is positively correlated to dropout.

Our experimental results showed that the best AUC was achieved by random forest, from as early as the third semester of enrollment we get 0.91 AUC to last semester, where the model gave a 0.97 AUC. Four features were necessary (*Semesters, Avg SE, Failed SE, GPA*) to achieve this accuracy. It implies that courses related to SE have the greatest impact in dropout prediction.

One attractive future work is to collect a larger dataset from the whole university student database and apply the model using such data to see how it generalizes to other specific programs. In addition, other classification methods can be applied to find the most suitable method and give a better classification accuracy.

The findings must be shared and discussed with SE faculty in order to validate and refine the model and implement it in a productive environment.

References

1. Al-Radaideh, Q.A., Al-Shawakfa, E.M., Al-Najjar, M.I.: Mining student data using decision trees. In: International Arab Conference on Information Technology (ACIT 2006), Yarmouk University, Jordan (2006)
2. Aulck, L., Velagapudi, N., Blumenstock, J., West, J.: Predicting Student Dropout in Higher Education. arXiv preprint arXiv:1606.06364, June 2016
3. Baradwaj, B.K., Pal, S.: Mining educational data to analyze students' performance. arXiv preprint arXiv:1201.3417 (2012)
4. Bhardwaj, B.K., Pal, S.: Data mining: a prediction for performance improvement using classification. arXiv preprint arXiv:1201.3418 (2012)
5. Brunner, J.J., et al.: Higher Education in Regional and City Development Antioquia, Colombia (2016)
6. Brunsden, V., Davies, M., Shevlin, M., Bracken, M.: Why do he students dropout? A test of Tinto's model. J. Furth. High. Educ. **24**(3), 301–310 (2000). https://doi.org/10.1080/030987700750022244
7. Chapman, P., et al.: CRISP-DM 1.0. CRISP-DM Consortium **76**, 3 (2000)
8. Dekker, G.W., Pechenizkiy, M., Vleeshouwers, J.M.: Predicting students drop out: a case study. In: International Working Group on Educational Data Mining (2009). http://www.educationaldatamining.org/EDM2009/uploads/proceedings/dekker.pdf
9. Devasia, T., Vinushree T P, Hegde, V.: Prediction of students performance using educational data mining. In: 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE), pp. 91–95. IEEE, March 2016. https://doi.org/10.1109/SAPIENCE.2016.7684167, http://ieeexplore.ieee.org/document/7684167/
10. Durso, S.D.O., Cunha, J.V.A.D.: Determinant factors for undergraduate student's dropout in an accounting studies department of a Brazilian public university. Educação em Revista **34** (2018)
11. de Educacion, M.: Spadies - sistema de prevencion y analisis a la desercion en las instituciones de educacion superior. www.mineducacion.gov.co/1621/article-156292.html. Accessed 18 July 2017
12. Jing, L.: Data mining and its applications in higher education. New Dir. Inst. Res. **2002**(113), 17–36 (2002). https://doi.org/10.1002/ir.35, https://onlinelibrary.wiley.com/doi/abs/10.1002/ir.35
13. Kim, D., Kim, S.: Sustainable education: analyzing the determinants of university student dropout by nonlinear panel data models. Sustainability **10**(4), 954 (2018)
14. Kovacic, Z.: Early prediction of student success: mining students' enrolment data. In: Proceedings of Informing Science & IT Education Conference (InSITE) (2010)
15. Márquez-Vera, C., Cano, A., Romero, C., Noaman, A.Y.M., Mousa Fardoun, H., Ventura, S.: Early dropout prediction using data mining: a case study with high school students. Expert Syst. **33**(1), 107–124 (2016). https://doi.org/10.1111/exsy.12135, http://doi.wiley.com/10.1111/exsy.12135
16. Mishra, T., Kumar, D., Gupta, S.: Mining students' data for prediction performance. In: 2014 Fourth International Conference on Advanced Computing & Communication Technologies (ACCT), pp. 255–262. IEEE (2014)
17. Romero, C., Ventura, S.: Educational data mining: a survey from 1995 to 2005. Expert Syst. Appl. **33**(1), 135 – 146 (2007). https://doi.org/10.1016/j.eswa.2006.04.005, http://www.sciencedirect.com/science/article/pii/S0957417406001266

18. Seidman, A.: Retention revisited: R= E, Id+ E & In, Iv. Coll. Univ. **71**(4), 18–20 (1996)
19. Herzog, S.: Estimating student retention and degree completion time: decision-trees and neural networks vis-á-vis regression. New Dir. Inst. Res. **2006**(131), 17–33 (2006). https://doi.org/10.1002/ir.185, https://onlinelibrary.wiley.com/doi/abs/10.1002/ir.185
20. Tekin, A.: Early prediction of students' grade point averages at graduation: a data mining approach. Eurasian J. Educ. Res. **54**, 207–226 (2014). https://eric.ed.gov/?id=EJ1057301
21. Tinto, V.: Dropout from higher education: a theoretical synthesis of recent research. Rev. Educ. Res. **45**(1), 89–125 (1975)
22. Wirth, R.: CRISP-DM: towards a standard process model for data mining. In: Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining, pp. 29–39 (2000)
23. Yukselturk, E., Ozekes, S., Türel, Y.K.: Predicting dropout student: an application of data mining methods in an online education program. Eur. J. Open Distance E-Learn. **17**(1), 118–133 (2014)