Applying Data Mining Techniques to Predict Student Dropout: A Case Study

1st Boris Perez
Systems Engineering Department
Universidad de los Andes
Bogota, Colombia
br.perez41@uniandes.edu.co
Systems Department
Univ. Francisco de Paula Stder.
Cucuta, Colombia
borisperezg@ufps.edu.co

2nd Camilo Castellanos
Systems Engineering Department
Universidad de los Andes
Bogota, Colombia
cc.castellanos87@uniandes.edu.co

3rd Dario Correal
Systems Engineering Department
Universidad de los Andes
Bogota, Colombia
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 System Engineering (SE) undergraduate students after 7 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 and Naïve Bayes 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.**

***Index Terms*—Student drop out, student desertion prediction, educational data mining, prediction models**

I. Introduction

Understanding the behavior of successful students is essential at universities. One of the major concerns for them is the necessity to predict characteristics of dropouts [2] [13]. Consequences include financial loss, lower graduation rates and an inferior university reputation [11]. The quality of education is measured by the percentage of students graduated, and by the strategies of the university to retain its students [12]. If an institution loses a student for whatever means, the institution has a lower retention rate.

Early identification of students who are at risk of failure is critical for the success of any retention strategy [12]. It is necessary to detect these students as early as possible and thus provide some care to achieve student retention [11]. According to Seidman [14], the key to reducing dropout levels consists of early identification of students at risk, in addition to maintaining intensive and continuous intervention.

It is important to understand the data of the students that allow us to know why they drop out. There is a lot of research implying some common variables related to student graduation. Some of them are related to differences in location, student demographics, and funding [13]. There is no single reason why students drop out and in fact, it is a multifactorial problem called "the one thousand factors problem" [11]. Despite long-standing theory, student drop out continues to be a large concern to the educational community [2].

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) [6]. This initiative was designed by the Center for Economic Studies (SEDE) at the University of the Andes to follow the problem of dropout in higher education, to calculate the risk of desertion of each student, and to classify them by groups, facilitating the evaluation of strategies for each of the situations that influence dropout such as student status, academic program and institution; and promoting the consultation, consolidation, interpretation and use of attrition information (tables and graphs, each by various criteria).

Data mining techniques have been extensively applied to predict students' academic performance based on their socioeconomic status and previous academic performance [12]. These techniques help us to extend our understanding of the learning process by identifying key variables and evaluating them. The use of data mining in education is known as Educational Data Mining. 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 performance [9]. Specifically, it has been found that the factors like mother's education and family income are highly correlated with student performance [8].

In this paper, we want to understand the key determinants of dropout, to accurately identify students with high probability of dropping out, and to find the important characteristics associated with graduation level of students in a computer science program. To do this, we model student dropout using data gathered from academic databases from 2004 to 2010.

978-1-5386-6740-8/18/$31.00 © 2018 IEEE

We used two approaches to identify the key determinants of dropout. The first one following the CRISP–DM methodology and applying Decision Tree, Logistic Regression and Naive Bayes models. The second one, using Watson Analytics to automatically establish these determinants. We do not take into account data related to the enrollment process like demographic information. Our approach considered an extremely heterogeneous population at a private university.

This paper is organized as follows: Section II reviews the related work. Section III introduces the data mining methodology. Section IV describes the data set used. Section V presents the data preparation. Section VI describes the modeling process. Section VII reports results. Section VIII offers a preliminary proposal to use these results in a real context. Finally, Section IX outlines the conclusions.

II. Background

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 [8]. 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 [15].

Educational Data Mining (EDM) is used to study available data in educational context and extract value from the hidden information. 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 performance prediction [17] [15].

Tinto's model [16] 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.

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 [10], 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.

Finally, Gerben et al. [7] conducted a case study in which they used machine learning techniques to predict student success using features extracted from student pre-university academic records.

III. Methodology

The analytic task tackled in this work is a binary classification task where dropout (0, 1) is the target variable. To do this, we follow the Cross Industry Standard Process for Data Mining methodology (CRISP–DM) [5]. CRISP–DM involves six phases: *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. *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 experimentation, we use an open source tool in Python as our data science development environment: Jupyter notebook, Python data analysis library (Pandas) to deal with data structures, Scikit-learn (machine learning library), Seaborn (statistical data visualization tool) and Graphviz (graph visualization software to generate the decision tree charts).

IV. Data set

The data set used in this work comes from 802 students enrolled in the Computer Science Program at a private university in Bogotá, Colombia.

The data is organized in four tables, and together they include 43 columns:

- 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
- Financial aids, including all the financial aids in the required terms.

Our focus was on students entering the university from first term of 2004 through the second term of 2010. Although the institutional databases have the 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 about 47.5% based on the definition of non-completion (NC) presented below.

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 the effect of any other variable. Also, missing data was assumed to be missing completely at random.

A. *Defining Non-Completion*

Students who dropped out (non-completions or NCs) are defined as those students who did not complete at least one undergraduate degree within 6 years from first enrollment. In the dataset, it is a single binary feature.

Enrollment in this case was defined as when a student received at least one transcript grade (regardless of whether it is numeric or passing) for a term.

V. Data Preparation

In this stage, we apply 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 join the four source files: admission information, transcript records, financial aids and graduation dates. Then, we performed a data profiling to get descriptive statistics which help us to understand the data.

The approach we use is grouping courses within similar academic field (v. gr. system engineering, mathematics, physics, language, management, biology, etc). For this, we aggregate the course grades, and course repetitions by student and by faculty. Additionally, we add 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 obtain an aggregate dataset with 31 columns which contain the following fields: gender, marital status, financing type count, 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 a drop out indicator (0,1).

Categorical fields such as gender, marital status and financing type are transformed to numeric values because the most of models require numeric inputs. To do this transformation, we use a mapping of pairs of values where each category is assigned to a discrete number. Also, we scale out the features to normalize the magnitudes and prevent that high magnitude fields skew the feature's weights into the machine learning models.

Figure 1 shows an excerpt of highest *Pearson* correlations map among features extracted in previous steps. The color scale represents the *Pearson* correlation from -1.0 (dark blue) to 1.0 (dark red). This chart allows us to identify clusters

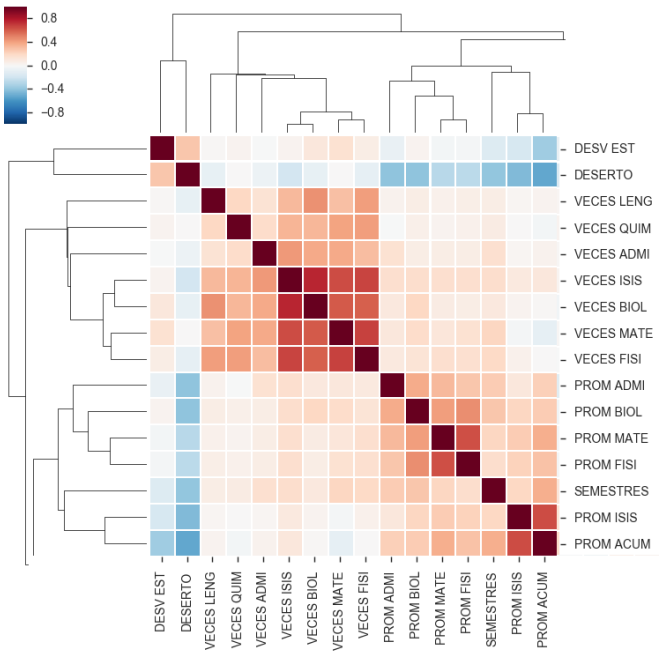

Fig. 1. Cluster map of Feature's Correlations

of correlations. For example, the cluster of course repetitions (upper left zone) shows that the pair *VECES ISIS* (number of courses retaken by a student in System Engineering program) – *VECES BIOL* (number of courses retaken by a student in Biology program); and pair *VECES MATE* (number of courses retaken by a student in Mathematics program) – *VECES FISI* (number of courses retaken by a student in Physics program) have the strong correlations. Another cluster related to academic averages (center zone) depicts that *PROM MATE – PROM FISI* (mathematics and physics average grades respectively) are highly correlated, and *PROM ISIS* (grade point average of the systems engineering program) presents the strongest correlation with *PROM ACUM* (cumulative grade point average). Analyzing the correlations of the target variable *DESERTO* (if the student drop out of the program - 0 or 1), we found a sort of features with significant negative correlation (blue): *PROM ADMIN* (grade point average of the Management program), *PROM BIOL* (grade point average of the Biology program), *SEMESTRES* (number of academic terms), *PROM ISIS*, *PROM ACUM*. On the other hand, the *DEV EST* (standard deviation of cumulative grade) shows a positive correlation with *DESERTO* (if the student drop out of the program - 0 or 1). This exploratory analysis offered us an initial insights about the potential predictors of drop out.

The frequency of dropouts in the dataset is 52.87% for *yes* (1) and 47.13% for *no* (0) and this shows that imbalancing between binary classes is small, so it is not required to over/down sampling the dataset.

Other step applied in this stage is Principal Component Analysis (PCA) to evaluate dimensionality reduction for this dataset. We use the scaled dataset because PCA is sensible to high magnitude variability of attributes.

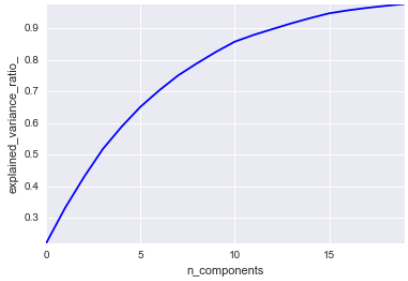

Fig. 2. Cumulative Explained Variance Ratio for 20 Principal Components

Figure 2 shows the cumulative explained variance proportion for 20 Principal Components (PCs). Fourteen PCs are necessary to explain more than 90% (91.54%) of variance, fifteen PCs explain 93.21% and twenty PCs cumulate the 97.66% of explained variance. In the next sections, we are going to use the first fifteen PCs and the original features to assess the model's behavior with both feature sets.

VI. Modeling

According to previous works described in Section II, models offering the best accuracy [2, 7, 10] are: Decision Tree, Logistic Regression and Naive Bayes. For each model, we use original dataset and PCA (fifteen PCs) datasets in order to compare the results. We choose PCA datasets with fifteen PCs because this enables us to reduce in 50% the dimensionality from 30 features to 15 features, and to keep a high percentage of data's variance (93.21%)

We applied Cross Validation (CV) in every following modeling tasks to avoid the 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:

- 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).

A. Decision Tree

Decision tree models are useful in binary and multi-class classification of an phenomenon like dropout (0 or 1) based on different features. These structures help to effectively understood the relationships among the variables regarding the target variable. An input set was grouped resulting from splitting the whole dataset in branches based on specific conditions.

In this case, 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 Figure 3.

The root node is the *PROM ISIS* (systems engineering courses average), and the other splitters are defined by *VECES ISIS*, *SEMESTRES* and *PROM ACUM*.

An important advantage of decision trees is the interpretability. We can specify the rules to predict the dropping out of a student with the boolean condition of each node. For example, analyzing the left branch only:

- A student with *PROM ISIS* lower or equal than 3.505, and *VECES ISIS* lower or equal than 0.1389 (scaled 0 to 5), and *PROM ACUM* lower or equal 3.5164, has 100% (143/143) for dropping out.
- *PROM ISIS* lower or equal than 3.505 and *VECES ISIS* lower or equal than 0.1389 (scaled 0 to 5) and *PROM ACUM* greater than 3.5164, has 88.8% (48/54) for dropping out.

B. Logistic Regression

The logistic regression is a regression classifier used to estimate the probability of a binary target based on independent features. It computes the occurrence probability of an event (drop out or not) based on predictors and weights or coefficients for these predictors.

We trained the logistic model with the following parameters: tolerance for stopping criteria=0.0001, inverse of regularization strength = 1.0, and solver=*liblinear*.

The coefficients of the model specify the features which contribute in positive or negative way to calculate the drop out likelihood: 0.832 VECES MATE, -0.509 SEMESTRES, -1.101 PROM ISIS, -2.440 VECES ISIS, 0.422 VECES HUMA, 0.636 VECES ADMI, 0.489 VECES QUIM and 2.225 DESV EST.

C. Naive Bayes

Naive Bayes is a probabilistic classifier based on Bayes' theorem with strong (naive) independence assumptions between the features. Naive Bayes classifiers can be faster compared to more sophisticated methods. The decoupling of the class conditional variable distributions allows that each distribution can be independently calculated as a one dimensional distribution. For this case study, we used a Gaussian Naive Bayes algorithm.

D. 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.

After included the data set in Watson, the service produce several charts explaining the discoveries. We focused in the decision tree as shown in Figure 4. In this tree, Watson used PROM ACUM as the primary variable, detecting that 99.9% students below 3.57 drop out from university. Also, if PROM ACUM is between 3.27 and 3.55 and variable VECES ISIS is 1 or lower, they had 81% of probability of dropout.

VII. Model Evaluation and Results

In this section, we compare the models evaluated in terms of Receiver Operating Characteristic (ROC) metric. ROC curve typically is described by false positive and true positive rate.

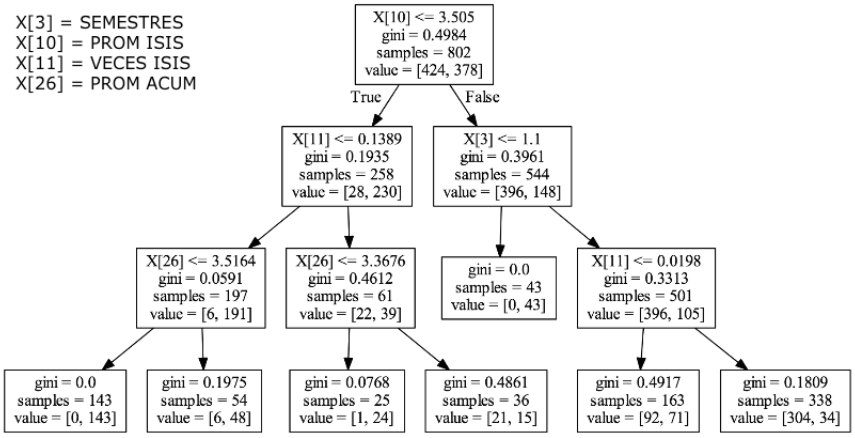

Fig. 3. Decision tree pruned to max depth=3

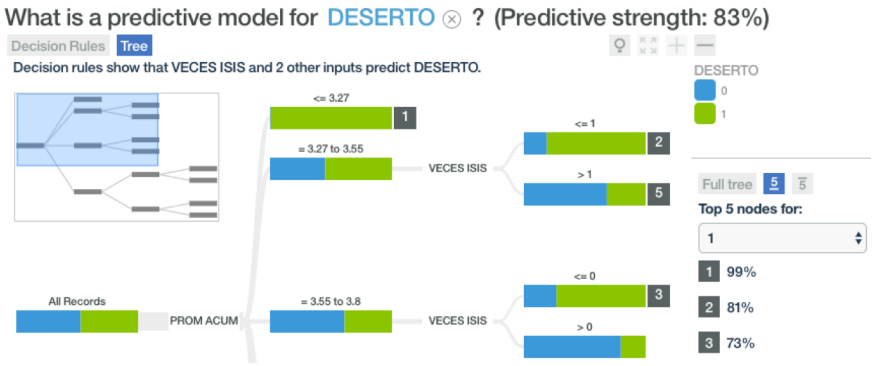

Fig. 4. Decision tree generated by Watson

The ROC curve implies that top left corner is the *ideal* point (i.e. a false positive rate of zero and a true positive rate of one), so it means that a larger area under the curve (AUC) is usually better.

Figure 5 presents the ROC AUC of tree models evaluated with and without PCA transformed dataset. In general, the AUC of models with original data (no-PCA) offered the best performance. It could be explained by the loss of variance implied by taking the first fifteen PCs (93.21%). Decision tree with original dataset achieved the best ROC–AUC (0.94), followed by Logistic Regression (0.92) and Naive Bayes (0.87).

VIII. Deployment

Deployment 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.

IX. 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, heterogeneous dataset of student demographics and transcript records. In the findings we discovered that system engineering courses performance are correlated to physics and mathematics courses performances. The irregularity (standard deviation of term's averages) is positively correlated to drop out.

Our experimental results showed that the best AUC was achieved by decision tree model (0.94), so this accuracy could be confident enough to help in early dropping out early detection. Four features were necessary (SEMESTRES, PROM SIS, VECES ISIS, PROM ACUM) 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 data set from the whole university student database and apply the model using such data. In addition, other classification methods can

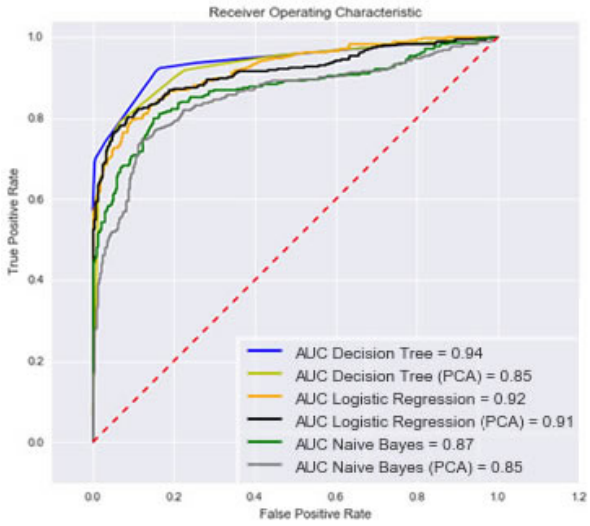

Fig. 5. Models evaluation using ROC - Area under Curve

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] Qasem A Al-Radaideh, Emad M Al-Shawakfa, and Mustafa I Al-Najjar. Mining student data using decision trees. In *International Arab Conference on Information Technology (ACIT'2006), Yarmouk University, Jordan*, 2006.

[2] Lovenoor Aulck, Nishant Velagapudi, Joshua Blumenstock, and Jevin West. Predicting Student Dropout in Higher Education. jun 2016.

[3] Brijesh Kumar Baradwaj and Saurabh Pal. Mining educational data to analyze students' performance. *arXiv preprint arXiv:1201.3417*, 2012.

[4] Brijesh Kumar Bhardwaj and Saurabh Pal. Data mining: A prediction for performance improvement using classification. *arXiv preprint arXiv:1201.3418*, 2012.

[5] Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rüdiger Wirth. Crisp-dm 1.0. *CRISP-DM Consortium*, 76, 2000.

[6] Ministerio de Educacion. Spadies - sistema de prevencion y analisis a la desercion en las instituciones de educacion superior. www.mineducacion.gov.co/1621/article-156292.html. Accessed: 2017-07-18.

[7] Gerben W Dekker, Mykola Pechenizkiy, and Jan M Vleeshouwers. Predicting Students Drop Out: A Case Study.

[8] Tismy Devasia, Vinushree T P, and Vinayak Hegde. Prediction of students performance using Educational Data Mining. In *2016 International Conference on Data Mining and Advanced Computing (SAPIENCE)*, pages 91–95. IEEE, mar 2016.

[9] Dorina Kabakchieva. Predicting Student Performance by Using Data Mining Methods for Classification. *Cybernetics and Information Technologies*, 13(1):61–72, jan 2013.

[10] Zlatko Kovacic. Early prediction of student success: Mining students' enrolment data. 2010.

[11] Carlos Márquez-Vera, Alberto Cano, Cristobal Romero, Amin Yousef Mohammad Noaman, Habib Mousa Fardoun, and Sebastian Ventura. Early dropout prediction using data mining: a case study with high school students. *Expert Systems*, 33(1):107–124, feb 2016.

[12] Tripti Mishra, Dharminder Kumar, and Sangeeta Gupta. Mining Students' Data for Prediction Performance. In *2014 Fourth International Conference on Advanced Computing & Communication Technologies*, pages 255–262. IEEE, feb 2014.

[13] Dheeraj Raju and Randall Schumacker. Exploring Student Characteristics of Retention that Lead to Graduation in Higher Education Using Data Mining Models. *http://dx.doi.org/10.2190/CS.16.4.e*, feb 2015.

[14] Alan Seidman. Retention revisited: R= e, id+ e & in, iv. *College and University*, 71(4):18–20, 1996.

[15] Ahmet Tekin. Early Prediction of Students' Grade Point Averages at Graduation: A Data Mining Approach. *Eurasian Journal of Educational Research*, 54:207–226, 2014.

[16] Vincent Tinto. Dropout from higher education: A theoretical synthesis of recent research. *Review of educational research*, 45(1):89–125, 1975.

[17] Erman Yukselturk, Serhat Ozekes, and Yalın Kılıç Türel. Predicting dropout student: an application of data mining methods in an online education program. *European Journal of Open, Distance and E-learning*, 17(1):118–133, 2014.