Data Mining Model to Predict Academic Performance at the Universidad Nacional de Colombia

Camilo E. López G.
Universidad Nacional de Colombia, Bogotá, Colombia, celopezg@unal.edu.co

Elizabeth León Guzmán
Universidad Nacional de Colombia, Bogotá, Colombia, eleonguz@unal.edu.co

Fabio A. González
Universidad Nacional de Colombia, Bogotá, Colombia, fagonzalezo@unal.edu.co

ABSTRACT

This paper presents the results of applying an educational data mining approach to model academic attrition (loss of academic status) at the Universidad Nacional de Colombia. Two data mining models were defined to analyze academic data. The models use two classification techniques, naïve Bayes and J-48, a decision tree classifier, in order to acquire a better understanding of the attrition during the first enrollments and to assess the quality of the data for the classification task, which can be understood as the prediction of the loss of academic status due to low academic performance. Different models were built to predict the loss of academic status in different scenarios including: attrition in any of the first four enrollments; at a specific enrollment using as input the admission process data and also using the historical academic records. Experimental results show that the prediction of the loss of academic status is improved when academic data is added.

Keywords: Educational Data Mining, Academic Analytics, Dropout, University Management.

1. INTRODUCTION

Education is a very important issue regarding the development of a country; specially in Colombia, where schooling is a factor strongly associated with social mobility, as stated by Gaviria in (Gaviria, 2002), therefore it is of great interest to identify the students who are at risk of dropping out as soon as possible, as well as to understand which factors have a larger influence on this. A Data Mining model is a suitable tool to encompass these tasks.

The application of Data Mining and other Analytics into the educational context have increased in the last decade. Ferguson presents in (Ferguson, 2012) three drivers for this to occur: first, the volumes of data that are collected in educational institutions have greatly augmented, whether from Course or Learning Management Systems or Student Information Systems; the second driver is the use of e-learning, although it have helped collecting data, it also have brought some learning issues such as possible lack of motivation and difficulties for the educators to receive direct feedback regarding the mood, level of interest, or even the understanding of the students; finally, the political concerns, countries are getting more understanding about the importance of higher education for their development and governments have an interest to improve it, to offer better learning opportunities that lead to better academic results.

Educational Data Mining (EDM) is "an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in" as defined by the International Educational Data Mining Society in www.educationaldatamining.org. Baker proposes in (Baker and Yazef, 2009) a classification for EDM as follows: Prediction, Clustering, Relationship mining, Distillation of data for human judgment, and Discovery with models.

Romero and Ventura, on the other hand, suggest in (Romero and Ventura, 2010) a different taxonomy based on the following educational tasks: Analysis & Visualization, Providing feedback, Recommendation, Predicting Performance, Student Modeling, Detecting Behavior, Grouping students, Social Network Analysis, developing Concept Map, Planning & Scheduling, and Constructing Courseware; however, being an application of Data Mining, its tasks are just the same: Classification, Clustering and Association Rules Analysis. They include exploratory tasks which precede Data Mining in the Knowledge Discovery Process.

In this paper, two data mining models are proposed to predict the loss of academic performance at a certain time by using, not only socio-economic data, but also the academic records. Several scenarios are tested regarding the data used.

The rest of the document is organized as follows: Section 2 presents the classification model for predicting academic success; section 3 describes the experimental setup and evaluation; results are presented in section 4; finally, the conclusions and future work are presented.

2. RELATED WORK

2.1 EDUCATIONAL DATA MINING

Dropout prediction and the analysis of its influencing factors is a well-studied subject since the late 1960s and early 1970s (Reason, 2009). Two of the most cited works were published in 1975 by Astin and Tinto. The former presents characteristics that increase the chances of completing the studies; these are individual student's characteristics at the time when he enters college and during the course, as well as institutional characteristics. The latter introduced a model of student retention at universities in which the event of dropping out is explained by the level of integration, both, social and academic, of an individual with the institution.

Another way to study academic success is to study the academic performance in a given course; it uses similar approaches for a different outcome and, instead of studying the failure at completing the course, it studies the failure at passing the course. Both of these use information about the student's past and present to predict his academic success in a class, a year, or a full program of studies.

Application of Data Mining techniques have been used to study this problem from a decade ago. Kotsiantis et al. applied in (Kotsiantis et al., 2003) different classification methods for predicting dropout from a class based on demographic and performance data from students with naive Bayes being the best option. Superby, Vandamme and Meskens (Superby et al., 2006) studied the phenomenon of academic failure of first-year students. They present the variables that are more correlated to academic success based on the model used by Parmentier (Parmentier, 2004), which explains that the academic result of a student is influenced by three set of factors: personal history, involvement in his own studies and the student's perceptions. Also, this work includes an application of Data Mining techniques to classify the first-year students into three categories: low, medium and high-risk students.

In (Dekker et al., 2009), three different datasets are used to predict dropout: Pre-university information, academic performance, and a combination of both. In general, the results were better for the third dataset, followed closely by the second. The authors implemented cost sensitive learning in order to avoid False Negatives. Kotsiantis goes further in (Kotsiantis, 2009) by implementing a local cost-sensitive technique to manage the imbalanced datasets; the results were better than those presented in his previous work (Kotsiantis et al., 2003). Bayer et al. used both (Bayer et al., 2012), student and social data from a Data Warehouse in the University to predict student dropout. Data Mining models had better results with the student and social data and the lowest results came from using social data only.

2.2 CLASSIFICATION METHODS

In a classification task, the groups are already known, so the objective is to assign a record to a predefined label or class. It can be seen as: Given a set of known attributes, estimate an unknown value; when this value is categorical, it is known as classification, when is numerical it is known as regression.

An important feature of a classification model is that it is built using part of the data, also known as the training set, which is used to learn the model. In this subset all the attributes are known, including the class. After the model is built, it is used to assign a label to new records where the class attribute is unknown.

In order to build the models, two widely used techniques are used, Decision Trees and a Bayesian Classifier; these were selected based on the results of previous work and the need for a predictive model that is descriptive at the same time, in order to acquire a better understanding of the event of loss of academic status.

A decision tree is a representation made out of nodes and arcs where an internal node presents a decision based on attribute values, and the arcs represent the choice made in the node. It ends on a leaf node, which represents the label or the class to be assigned. To classify a record with a decision tree, it starts by the root node and goes down one level at a time depending on the results of the conditions tested on every node; when it ends on a leaf node, the record is classified according to the label of that leaf node. In this work, the C4.5 algorithm is used which is based on Hunt's algorithm (Quinlan, 1993). It has an important feature which is its ability to manage both, discrete and continuous attributes.

On the other hand, a Bayesian Classifier (Tan et al., 2005) considers a probabilistic relationship between the class and the attributes, instead of a deterministic relationship where a given set of attributes not always have an identical label outcome. The classification task, to classify a record depending on its attributes values, can be expressed as the probability of a record of being from the class Y, given that the record has a set of attributes X. That is P(Y=y|X=x).

3. PROPOSED DATA MINING MODEL

This section introduces the student classification models to predict the loss of academic status due to low academic performance, which uses the student admission data: socioeconomic, demographic and the initial academic information (gathered from the admission process); and the academic records of previous academic periods. The general model can be seen in Figure 1.

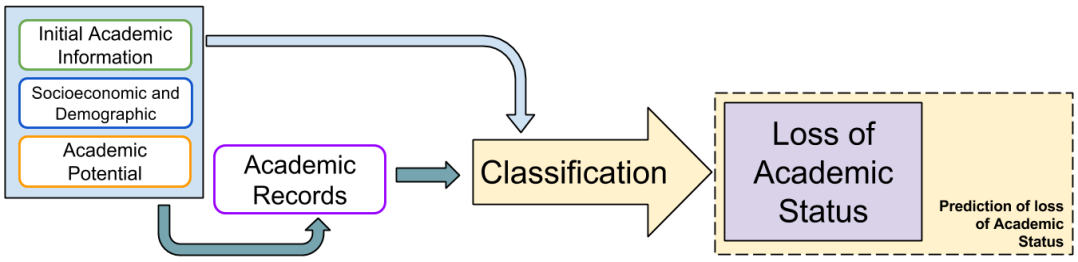

Figure 1: Data Mining Model

Different settings of the general model were trained and tested to predict the loss of academic status: first, the prediction is done regardless of the enrollment at which occurs; second, the prediction at a given enrollment is performed based on the initial information (data gathered during the admission process); and then, using the information known before the academic period starts, which includes the grades of the previous academic period when available. The different settings are explained below.

3.1 PREDICTING LOSS OF ACADEMIC STATUS

The most general case, in which the interest is to predict the occurrence of the loss of academic status at any time in the first four academic periods based on the initial information. Figure 2 shows this configuration.

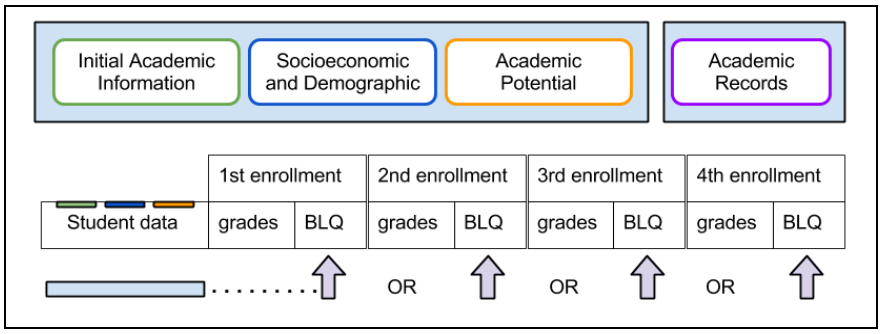

Figure 2: Predicting loss of academic status at any academic period based on initial information.

3.2 PREDICTING LOSS OF ACADEMIC STATUS AT A GIVEN SEMESTER

First, initial data is used to train a model to predict the loss of academic status at a particular academic period. The model is then complemented by adding academic information to the admissions data. The event of loss of academic status in a given period uses the academic information, grades and previous loss of academic status, all available before the current period. For instance, to make a prediction in the third semester, data from the first two are used with the admission data. A visual representation of these configurations is presented in Figure 3.

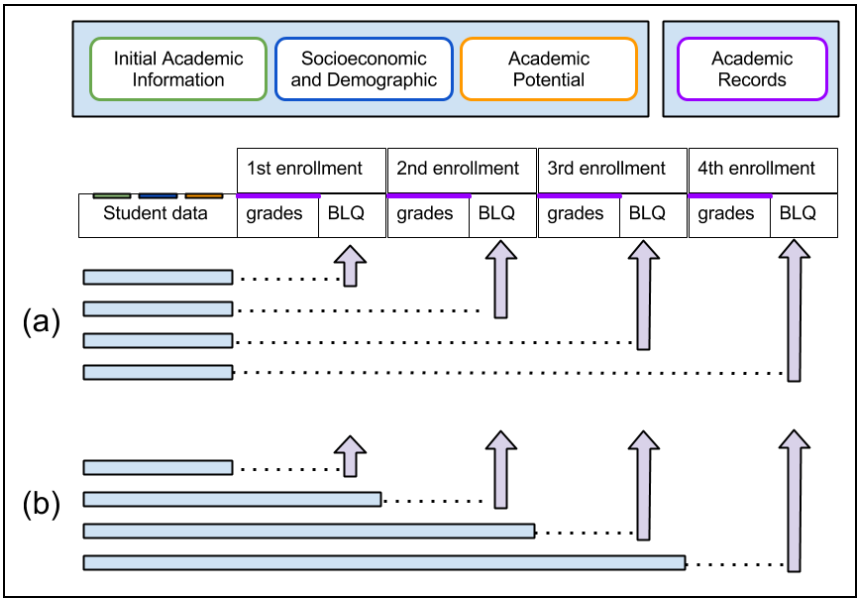

Figure 3: Predicting loss of academic status including academic data

The implementation of the models was done using Rapid Miner (Mierswa et al., 2006) with the operators NaiveBayes and Weka's W-J48 (Hall, 2009) for the Bayesian classifier and the decision tree respectively.

4. EXPERIMENTS

4.1 DATA SETS

Data was collected between the second academic period of 2007 (2007-II) and the second academic period of 2012 (2012-II) from two sources, the Direction of Admissions (DNA) and the Academic Information System. DNA collects the information from the biannual admission process and includes the admission test scores results, the options for enrollment and some socio-demographic attributes. On the other hand, the Academic Information

System includes data of the academic life of the student; three datasets were used regarding grades and credits, loss of academic status, and student enrollment records per academic period.

The admissions data set fields considered in this study can be grouped in three categories that will be described briefly:

- **Academic potential:** Admission test score in five modules (i.e. Sciences, Math, Image, Text and Social studies) and classification levels for Basic Math and Literacy.
- **Demographic and socio-economic:** Age at Admission, Gender, city of origin, 'estrato' (i.e. socio-economic classification), ethnicity.
- **Previous academic information:** high school type (e.g. public, private), type of access (e.g. regular, special admission program), option in which the student chose the program (from 1, first option, to 3), and the previous program, if exists.

All of the Academic Information System's datasets include fields to identify the academic period, the student and the program in which is enrolled.

The enrollment report is composed of records of the students enrolled at a given academic period. It includes different fields regarding student data, number of enrollments and general academic performance, such as GPA, and weighted GPA.

The grades report has data of the courses taken in each academic period by a student and their final results. Some of the fields regarding the courses are: course section, number of credits, numeric grade (0 to 5), alphabetic grade (approved and not approved) and the typology of the course, i.e. professional, foundation, optional electives, and leveling courses.

The loss of academic status report registers when a student's academic history is blocked. Some of its fields are the code of the blocking, the description, date and academic period; active, if the blocking is still active or not; and, in some cases, the information of the unlocking of the academic history: code, description, date and academic period.

The loss of academic status is considered academic when is related to a low academic performance, non-academic if the academic performance requirements were still fulfilled but the student didn't enroll in that academic period. The academic category is the only one considered in this research.

4.2 EXPERIMENTAL SETUP

For the experiments setup, 10-fold cross validation was used to train the model; in this, the data set is divided into ten equally distributed groups. The model was learned from nine of them, corresponding to the training set, and then is evaluated on the tenth group, corresponding to a validation set. The process is repeated ten times so that every group is used for learning and testing. This data set uses the data from the first ten academic periods.

The model is then applied to a previously unseen records corresponding to the 2012-II academic period, the test set; in order to test the model in a more realistic way, considering all possible known data to train the model for the current semester to apply it to a new academic period.

Additionally, the imbalance between the two classes was considered. To overcome this issue, a cost-sensitive technique was included in the model; the metaCost algorithm (Domingos, 1999) provides weights that represent the cost of classifying a record correctly or incorrectly, depending on the type of error that is more accepted. In this model, an error of classifying a student as No Risk when he is at risk is more critical than classifying a non-risk student as being at risk. Because of that, the following weights are considered in the model (Table 1).

The weights in (a) are the same for both types of errors; configuration in (b) and (c) consider the different acceptance regarding classification errors, (b) has a cost of misclassifying a student who is at risk as three times the error of misclassifying a non-at risk student as he were; finally, (c) presents a cost of misclassifying a student at risk but also considers a reward for classifying the BLQ class correctly.

Table 1: Weights in the cost-sensitive model.

		TRUE	
		No	BLQ
Predicted	No	0	1
	BLQ	1	0

(a)

	TRUE	
	No	BLQ
No	0	3
BLQ	1	0

(b)

	TRUE	
	No	BLQ
No	0	3
BLQ	1	-2

(c)

The performance of a classification model depends on the number of records of the validation set (academic period of 2012-II) correctly classified. These counts are commonly represented by a confusion matrix, a table that presents the number of records correctly and incorrectly classified.

In this case, the values correspond to:

- **Number of True Positives (TP):** The records correctly classified in the positive class (BLQ.Acad).
- **Number of True Negatives (TN):** The records correctly classified in the negative class.
- **Number of False Positives (FP):** The records incorrectly classified in the positive class. i.e. BLQ.Acad was incorrectly predicted.
- **Number of False Negatives (FN):** The records incorrectly classified in the negative class.

These values have a relation with the cost sensitive model mentioned above. In this work there is a cost to False Negatives and a reward to True Positives. The values are also used to construct the following measures:

- **Precision (P):** the fraction of instances classified as positive (TP + FP) that are correctly classified (TP).
- **True Positive Rate or Sensitivity (TPR):** The fraction of the instances of the positive class that are correctly classified. TPR = TP/(TP + FN)
- **True Negative Rate or Specificity (TNR):** The fraction of the instances of the negative class that are correctly classified. TNR = TN/(TN + FP)
- **Balanced Accuracy:** The average of the TPR and TNR. Bal. Acc. = (TPR) + TNR) / 2.

Balanced accuracy is the arithmetic mean of the accuracy of both classes. It is used instead of the regular accuracy to prevent the bias that is caused by the unbalanced dataset. Consider a dataset where the positive class is only 10% of the instances, a classifier that labels every instance as the negative class will have an accuracy of 90% but a balanced accuracy of only 50%.

5. RESULTS AND DISCUSSION

First experiments were intended to predict the loss of academic status in any of the first four academic periods using the initial data, but results were not satisfactory; the balanced accuracy ranged from 51 to 52% in the decision tree and between 54-57% in Naïve Bayes.

The next set of experiments was intended to predict, not only the event of loss of academic status, but also the semester in which occurs. First, only the initial data was used to predict an academic history blocking in the second, third and fourth enrollment. The algorithms were tested with the 2012-II academic period. The results have a similar behavior when using the decision tree with an increase in performance in the prediction at the first enrollment and a posterior decrease. Naïve Bayes, on the other hand, showed an irregular behavior in the predictions at enrollment 3 and a shift in the performance of the different cost matrices used. Figure 4 presents these results; the different lines represent the cost-sensitive configuration described in Table 1.

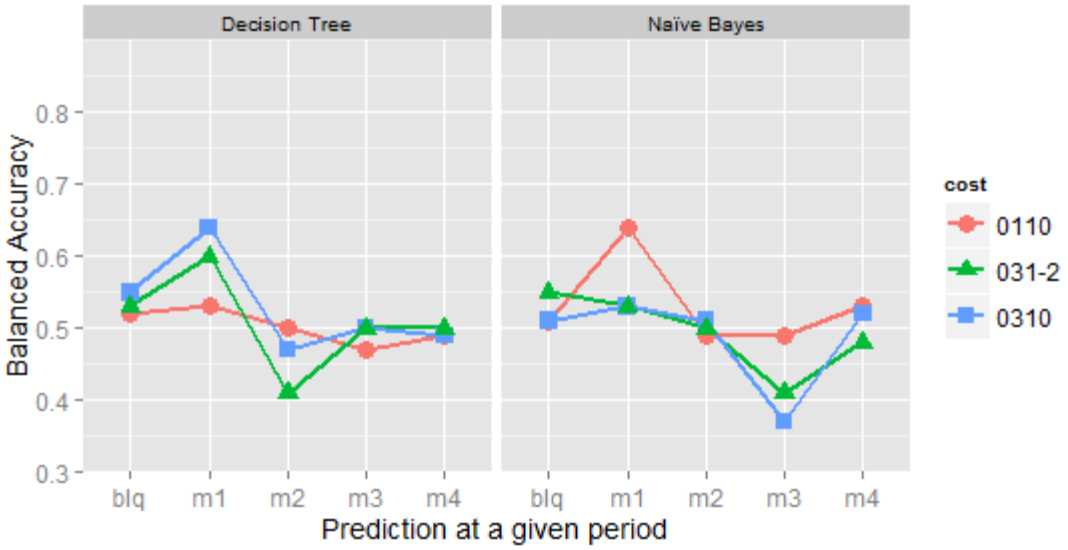

Figure 4: Predicting loss of academic status at a given semester using admissions data. Test results.

The next step was to include the academic records to the data. As it was described before, the academic data used are the grades and percentage of enrolled and approved credits in the previous academic period. Naïve Bayes had the best results, surpassing the 75% in balanced accuracy, up to 85% at the fourth enrollment on the test set. The decision tree didn't show much of an improvement, except for the second enrollment. This can be seen in Figure 5.

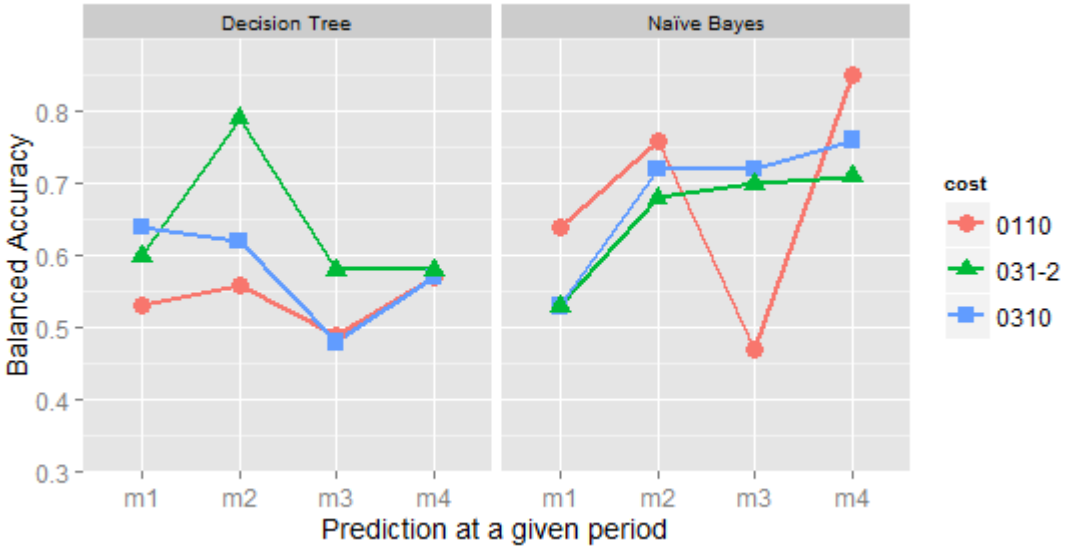

Figure 5: Predicting loss of academic status at a given semester using academic data. Test results.

Naïve Bayes results were better on the test set; however, there are differences between training and test data. The decision trees results were more consistent regarding that subject making it more reliable when testing on new data. The classifications results presented in this research are similar to those reported in the literature in problems

that used similar datasets, but those were mostly reported on validation sets, data that were used to learn the model.

Additionally, a systematic analysis was conducted to identify relevant factors related to the loss of academic status due to low academic performance according to the learned classification models. The interpretation depends on the selected model. The Bayesian classifier is interpreted based on the probabilities of the factors, or variables, and those with a higher probability are highlighted. When the attribute is continuous, a visual comparison of the density function is also taken into consideration. On a Decision Tree on the other hand, two approaches were followed: first, the features that are on the root of a tree are considered as more relevant; and second, the branches with more examples are also considered.

Results are described below.

- Admission test results: the academic potential shows an expected behavior considering the population under study, programs of the Faculty of Engineering. The components of Math and Science, along with the total score and the classification level for basic Math are the most relevant, poor performances are more related to loss of academic status. On the other hand, for the prediction at fourth enrollment, high scores in the social sciences component are more related to this loss.
- Age at enrollment: a first thought could lead us to think that younger students are at more risk, and they are, in absolute terms; however, the age rank of 23-28 presents a higher risk.
- Socioeconomic Status (Estrato): there are to variables used to measure this, PBM and estrato, according to the results the estrato was more telling than the PBM.
- Option for enrollment: The models show that this feature is relevant when the loss of academic status is predicted at the first enrollment but not so much when the prediction is at a later enrollment. A further evaluation shows that there is a relationship between the option for enrollment and the loss of academic status at first enrollment and that this relationship disappears at a later enrollment.
- Grades: the grade average and the percentage of the approved credits are relevant features and there is a difference according to the typology of the courses, the performance at professional subjects is most telling than the performance at foundational subjects and that the absence of elective courses is more related to the loss of academic status. It is also important to notice that the grades become more relevant as time progresses, when trying to predict the loss of academic status at a later enrollment; under this scenario, these features gain even more importance than socioeconomic and demographic data.

6. CONCLUSIONS AND FUTURE WORK

Two learning algorithms, naïve Bayes and a decision tree, were used to create classification models to predict the loss of academic status due to low academic performance. Different scenarios were evaluated changing the information available to the model. These scenarios include the prediction of the academic history block at any time of the first two years, at a specific enrollment using only admissions data and then including the academic information, i.e. grades and credits enrolled. The models were tested with previously unseen records corresponding to the 2012-II academic period.

The results are consistent with those reported in the literature on analogous scenarios. The accuracy of the classifiers improved when academic data was added; however, adding more academic data doesn't necessarily keep improving the classifier. It is important to notice that early dropout investigations suggest that retention is influenced by different factors involving the integration of the student to the University making the admissions data insufficient for making predictions.

Bayes classifier performance improved when academic data from the first enrollment were added; however the performance decreased after the addition of the academic data of the second enrollment. This may be caused by the assumption of independence required by the algorithm. Naïve Bayes results were better on the test set;

however, there are differences between training and test data. The decision tree results were more consistent regarding that subject making it more reliable when testing on new data.

The data used is gathered for each academic period, but the causes of low academic performance occur on a day to day basis. This leads to think that new, and possibly, non-traditional ways, for collecting information are needed. This work focused on the loss of academic status due to low performance; however, the academic performance can also be studied at a different level, perhaps at the course level. The classification model could also include the non-academic loss of student status, or a new model could be built to reflect this situation.

REFERENCES

J. Bayer, H. Bydzovská, J. Géryk, T. Obsıvac, and L. Popelınskỳ. (2012). "Predicting drop-out from social behaviour of students," in *Proceedings of the 5th International Conference on Educational Data Mining-EDM* 2012, Chania, Greece, pp. 103–109.

R. S. J. d Baker and K. Yacef. (2009). "The State of Educational Data Mining in 2009: A Review and Future Visions," *Journal of Educational Data Mining*, vol. 1, no. 1, pp. 3–17.

G. W. Dekker, M. Pechenizkiy, and J. M. Vleeshouwers. (2009). "Predicting Students Drop Out: A Case Study," in *Proceedings of the 2nd International Conference on Educational Data Mining*, Cordoba, Spain, vol. 9, pp. 41–50.

P. Domingos (1999). "MetaCost: A General Method for Making Classifiers Cost-Sensitive," in *In Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining*, pp. 155–164.

R. Ferguson. (2012). "The State of Learning Analytics in 2012: A Review and Future Challenges," Knowledge Media Institute, The Open University, UK, Technical Report KMI-12-01.

A. Gaviria. (2002). *Los que suben y los que bajan: Educación y movilidad social en Colombia*. Bogotá: Alfaomega.

M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. (2009). "The WEKA data mining software: an update," SIGKDD Explor. Newsl., vol. 11, no. 1, pp. 10–18.

S. Kotsiantis, C. Pierrakeas, and P. Pintelas. (2003). "Preventing student dropout in distance learning systems using machine learning techniques," in *Proc. Int. Conf. Knowl.-Based Intell. Inf. Eng. Syst.*, Oxford, U.K., pp. 3–5.

S. Kotsiantis. (2009). "Educational data mining: a case study for predicting dropout-prone students," *International Journal of Knowledge Engineering and Soft Data Paradigms*, vol. 1, no. 2, pp. 101–111.

I. Mierswa, M. Wurst, R. Klinkenberg, M. Scholz, and T. Euler. (2006). "YALE: rapid prototyping for complex data mining tasks," in Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, New York, NY, USA, pp. 935–940.

P. Parmentier. (1994.) "La réussite des études universitaires: facteurs structurels es processuels de la performance académique en première année en médicine.," PhD, Catholic University of Louvain.

J. R. Quinlan. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann.

R. D. Reason. (2009). "Student Variables that Predict Retention: Recent Research and New Developments," Journal of Student Affairs Research and Practice, vol. 46, no. 3.

C. Romero and S. Ventura. (2010). "Educational Data Mining: A Review of the State of the Art," Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, vol. 40, no. 6, pp. 601 –618, Nov.

J. F. Superby, J. P. Vandamme, and N. Meskens. (2006). "Determination of factors influencing the achievement of the first-year university students using data mining methods," in Workshop on Educational Data Mining, Boston, USA, pp. 37–44.

P.-N. Tan, M. Steinbach, and V. Kumar. (2005). *Introduction to data mining*. Boston: Pearson Addison Wesley, 2005.

Authorization and Disclaimer

Authors authorize LACCEI to publish the paper in the conference proceedings. Neither LACCEI nor the editors are responsible either for the content or for the implications of what is expressed in the paper.