A Novel Predictive Modeling System to Analyze Students at Risk of Academic Failure

Lotfi Najdi
GMES Laboratory, ENSA
Ibn Zohr University
Agadir, Morocco

Brahim Er-Raha
GMES Laboratory, ENSA
Ibn Zohr University
Agadir, Morocco

ABSTRACT

Supporting academic success is a central focus of higher education institutions. To address this challenge, predictive techniques could be applied in order to build models that predict academic performance such as student retention and graduation. This paper presents a predictive system for modeling and scoring students achievements. Based on student historical data, predictive models are developed to classify students who are at risk of dropping out and not graduating, by examining CART and random forest as an ensemble method. Generated models are then applied on freshman student's data to predict their academic behavior. The examination of the CART and Random forest algorithms on student data resulted tree based models with accuracy of 88%. The result of this work was an interactive system implemented using R language and shiny framework, including data preparation, model building and scoring engine. The predictive system, which is present in this paper, might help decision makers gaining a deeper insight in students' academic achievements and optimize their human and financial resources toward effective student support services.

Keywords

Educational Data Mining, Predictive modeling, Random Forest, Ensemble learning

1. INTRODUCTION

Over the last decades, there has been a dramatic rise in the amount of digital information stored in educational databases, due to the adoption of information technology. The main challenge facing higher education today is to transform these data into valuable and useful knowledge and patterns, which could help to better understand the paths of their students and to improve the quality of resources allocation. One promising solution to address these challenges is through knowledge discovery in databases or data mining in education called Educational Data Mining (EDM).

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 they learn in [1]. Educational data mining methods belong to a diversity of literatures such as data mining, machine learning, information visualization, and computational modeling. EDM methods fall into the following general categories: prediction, clustering, relationship mining, discovery with models, and distillation of data for human judgment [2].

The work presented in this paper aims to delivering predictive capabilities of students' risk assessment to higher education institutions. This objective is achieved by providing approach and tools for:

- Building customized predictive models, to assess a risk score by institution or department using Learning Algorithms.
- Enable scoring risk of academic failure for freshman students through Intuitive visual interfaces.

The rest of the paper is organized as follows: In Section 2, related works in predicting student's dropout and failure are summarized. Section 3 is devoted to the methods, data and tools of this study. Section 4 presents the resulting system in detail. Finally, this paper is concluded by a summary and an outlook for futures works.

2. RELATED WORK

In order to increase students' retention and graduation rates, universities should firstly identify factors that influence student performance and after that develop predictive models that accurately classify who are likely academically successful or at-risk. Over the years, there have been a considerable amount of research efforts addressing various educational issues. Among them, however, there are only a few published research papers in predicting student retention and progression in education process. Reference [3] has examined several educational data mining techniques to predict the Electrical Engineering (EE) students drop out after the first semester of their studies or even before they enter the study program as well as identifying success-factors specific to the EE program. This study shows that decision trees give a useful result with accuracies between 75 and 80%. Decision tree has been applied to evaluate student's performance for the last semester examination. Based on student data including Attendance, Class test, Seminar, and Assignment marks. This study earlier in identifying the dropouts and students who need special attention and allow the teacher to provide appropriate advising/counseling [4]. Reference [5] have employed several data mining methods like Decision Tree, Naïve Bayes and K-Nearest Neighbors in order classifies the students' grade into five categories, depending on their university performance and based on the student pre-university data, was considered. The results achieved by applying selected data mining algorithms for classification on the university sample data reveal that the prediction rates are not remarkable (between 52-67 %). Reference [6] has explored the sociodemographic variables (age, gender, ethnicity, education, work status, and disability) and study environment (course program and course block), that may influence persistence or dropout of the distance education students. Among classification tree growing methods Classification and Regression Tree (CART) was the most successful in growing the tree with an overall percentage of correct classification of 60.5%. The performance of the most frequently used decision tree classifiers are studied and the experiments are conducted to find the best classifier for retention data to predict the student's drop-out possibility, many attributes have been tested, and some of them are found effective on the prediction. An ID3 learning algorithm was selected to learn effective predictive models from the student dropout data with an accuracy of 85.7%[7]. Decision tree and Bayes as a classification technique were studied and evaluated to identify those students who are at the risk of failing; with highest prediction accuracy is that of decision tree at 92.34%. The research findings indicated that Entrance Certificate Examination) result, Gender, Number of students in a class, number of courses given in a semester, and field of study are the major factors affecting the student performances [8]. In [9] decision Tree and Naïve Bayes has been applied for the purpose of examining and predicting students' dropouts through their university programs. These methods were tested using 10-fold cross validation. The accuracy of decision Tree and Naïve Bayes classifiers were 98.14% and 96.86% respectively. [10] In this paper Logistic Regression, Naïve, and Decision Tree and stacking Ensemble were examined in order to build Student at-risk model. Stacking Ensemble based model had a recall of 75%. Student At-Risk Model has been deployed using SQL Server Analysis Services to populate a report hosted on a SharePoint site that serves as the front end for the analysts to access the data.

In an era of educational data mining patterns affecting student's outcomes are being continuously discovered and Most of the researchers have used Decision Tree for model building because of its ability to provide easy to understand classification rules. However there is a lack of research on investigating ensemble learning and converting models to educational practice. Hence, in this paper, a novel system for analyzing students' risk of drop out and non-graduation, is presented. This system combines building predictive models and using these patterns to score freshman data in order to make up-to-date decisions.

3. METHODOLOGY

3.1 System Architecture

The purpose of the presented system is to generate personalized models of students' risk; it is composed of three main parts:

- Model development engine: which consist of training learning algorithm based on students Historical Data (Student Demographic Data, Student Aptitude Data), unique to each studied area (faculty or bachelor). Data are automatically preprocessed and splitted into training testing sets, additionally k-fold cross validation is used. Once the model is validated, it could be saved for future uses, as standalone model.
- Model scoring engine: this engine use generated models by applying those patterns to freshman unlabeled data, so that a prediction about the risk of each student to drop out or fail.
- User interface: this component is responsible of the interaction with the end-users.

Those components are controlled by the code written with Shiny R. R is a programming language and data analysis software, used by scientific and academic community for performing statistical analysis, data visualization, and predictive modeling. Data analysis is done by writing scripts and routines based on objects and operators provided by R. Shiny is an R package that offers a web framework for building web applications using R. it helps to turn analyses into interactive web applications, by allowing developers to build intuitive interface using friendly controls like sliders, drop-downs, and text fields, and to implement server code with R language.

It also let incorporating a rich number of outputs components like plots, tables, and summaries. This tools is based on a reactive programming [11].

3.2 Learning methods

In supervised learning, the goal is to predict the value of an outcome measure based on a number of input measures [12]. In order to predict if a student is at risk of dropout or failure, classification was used. Supervised classification is one of the most commonly, applied data mining techniques. The purpose of this technique is to define rules for assigning objects into target categories or classes, based on qualitative or quantitative variables characterizing those objects, in order to accurately predict the target class for each case in the data. To build the model, a set of points with correct class labels, called a training set, is strongly required. After performing the learning, we are able to predict the class for new unlabeled data. Many different types of classification models have been proposed such as decision trees, probabilistic classifiers, support vector machines, and so on [13]. In this paper random forests and CART as classification method was investigated to predict students at risk of failure.

3.2.1 Random forest

Random forest is an ensemble learning method for both classification and regression, it consist in developing an ensemble of decision trees from randomly sampled subspaces of the input features, and final classification is obtained by combining results from the trees via voting[14]. In Random Forest, each tree is grown as follows:

- A sample of N cases in the training set is taken at random with replacement. This sample will be the training set for growing the tree.
- A number $m<M$ is fixed such that at each node (M is the number of input variables), m variables are selected at random out of the M, and the best split on m is used to split the node.
- Each tree is grown to the largest extent possible and there is no pruning.
- Predict new data is made by aggregating the predictions of the ntree trees (majority votes for classification, average for regression).

3.2.2 Classification and regression trees

Classification and regression trees (CART) is learning technique that produces either classification or regression trees, depending on whether the dependent variable is categorical or numeric, respectively[15]. One of the advantages of using CART tree is its ability to provide easy to understand classification rules.

3.2.3 K-fold cross validation

Cross-validation is a model validation technique for assessing how the results of learning algorithms will generalize to an independent data set. The goal of cross validation is to define a dataset to test the model in the training phase. K-fold cross consists in separating the training dataset into a number of equally sized groups of instances or folds. The model is then trained on all folds exception one that was left out and the prepared model is tested on that left out fold. The process is repeated so that each fold gets an opportunity at being left out and acting as the test dataset. Finally, the performance measures are averaged across all folds to estimate the capability of the algorithm on the studied problem.

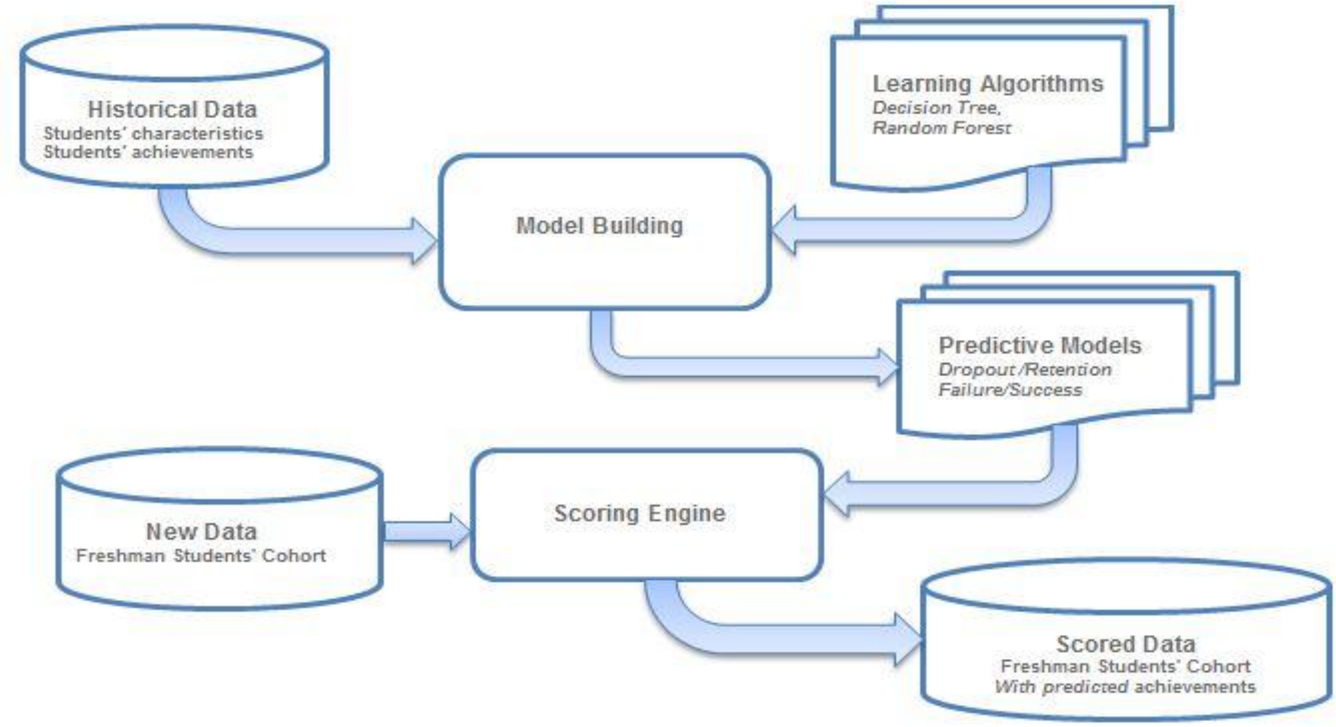


Figure 1 : System Architecture

3.3 Study area and data

The predictive modeling process is illustrated through a case study with the purpose of Predicting students at-risk of Drop out within one year, Drop out within two year, Non-Graduation. This study has been performed based on the student's characteristics of 6000 students who enrolled Economics and Management program Faculty of Law, Economic and Social Sciences. It mainly investigates sociodemographic variables (Age, Gender, distance from home), pre-enrolment (Secondary school grade, Secondary certificate type) and first term performance (First term GPA, Number of modules being validated at Semester1), to predict student success or failure based on Random Forest and CART.

4. RESULTS

4.1 Data classification using Random forests and CART algorithm

In the learning stage, different models were trained for predicting student achievements. The data was spited by default into training of 70% 30% for testing. Additionally 10-fold cross validation was used in order to test the accuracy of each model during the training stage. Random Forest algorithm is configured to explore all possible subsets of the independents variables. Thus By building predictive models based on it, the importance of features could be estimated. Figure 2 displays the plot of variables importance for the prediction of "Drop out within one year after first enrollment" model generated by random forest algorithm.

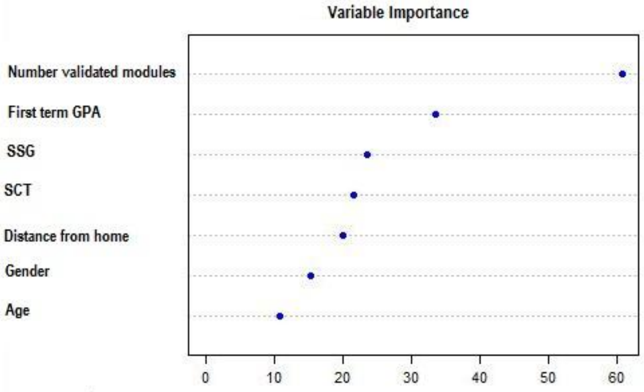


Figure 2 : Variable Importance Plot

Figure 2 shows that the "Number of validated modules at Semester1", "First term GPA", and "Secondary school grade" are the top 3 most important attributes in the dataset and the "Age" attribute is the least important separating persistent and non-persistent students.

Table 1 : Accuracy of classification algorithm

	CART	Random forests
Drop out within 1 year	87.4%	87.9%
Drop out within 2 year	88.2%	88%
Graduation	76.5%	76.8%

Table 2: Sensitivity and Specificity of classification algorithm

	Sensitivity		Specificity	
	CART	**RF**	**CART**	**RF**
Drop out within 1 year	92.12%	91.60%	78.55%	78.64%
Drop out within 2 year	93.09%	92.18%	79.37%	79.25%
Graduation	81.53%	82.43%	62.02%	66.44%

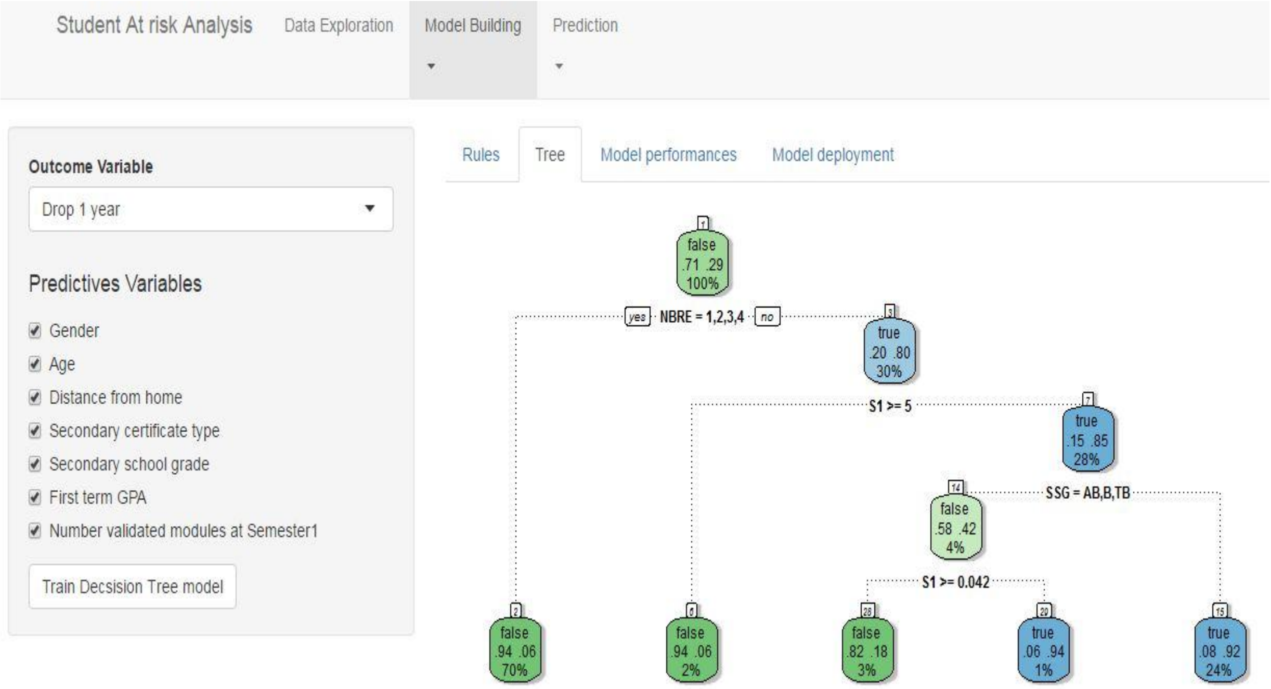
The model obtained by CART present an Accuracy of 87.4% in predicting the likelihood of a student drop out within 12 months after initial enrollment, which correspond to the overall percentage of correct classification. As shown in table 1 and table 2, CART and Random Forests, provide approximately the same performance measures in terms of accuracy, sensitivity and specificity for the studied outcomes. In the case of this dataset, the use of Random forests as an ensemble method doesn't make a remarkable gain in terms of accuracy.

4.2 Functional overview

The menu of the produced predictive tools includes three options: Data exploration, Model building and data scoring.

The first idem of the menu is "Data Exploration", which allow user to upload data file and explore it among different variables. The data received by server as a data frame is spited into training and test partitions before performing learning task.

The second option is "Model building". This option is used for generating predictive models to assess student drop out and failure.



Outcome Variable

Drop 1 year

Predictives Variables

☑ Gender

☑ Age

☑ Distance from home

☑ Secondary certificate type

☑ Secondary school grade

☑ First term GPA

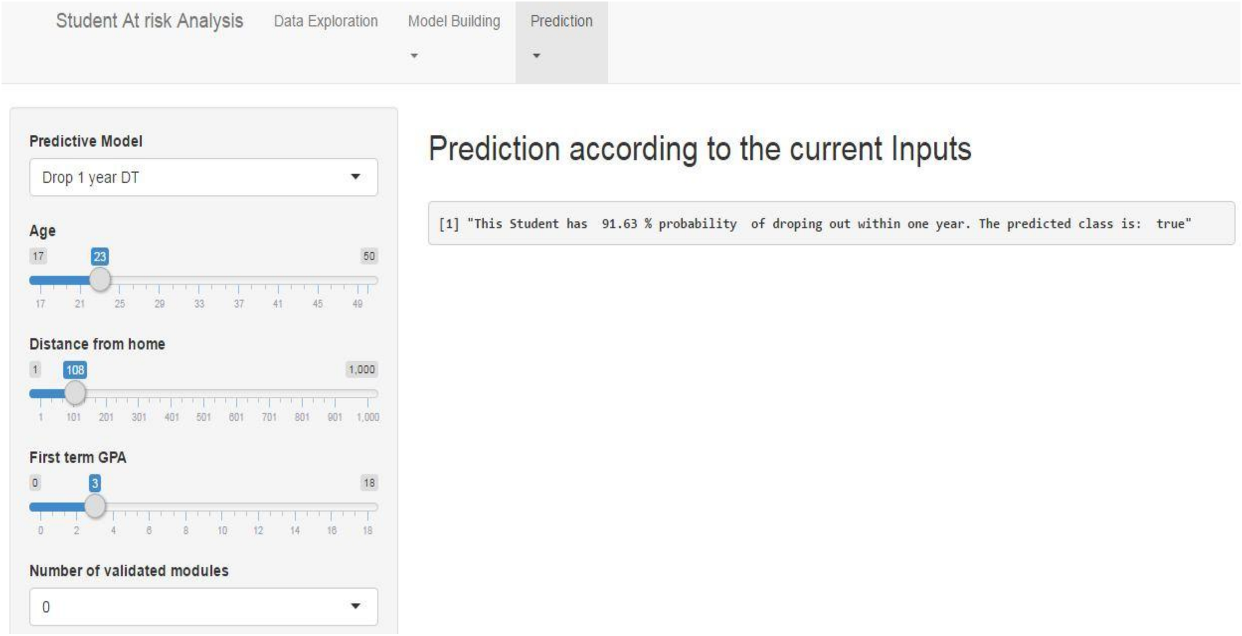
☑ Number validated modules at Semester1

Figure 3 : Model building with decision tree

From sidebar panel, user can select outcome variable to be studied, such as "Drop out within one year after first enrollment" and independents or predictive variables to be included in the learning process (see Figure 3).

After training the predictive model based on user choices, results from classification functions are summarized are rendered. Confusion Matrix and metrics for assessing the classification accuracy are displayed in the performance tab panel.

In addition to model performance there many other outputs that has been implemented. For example CART analysis displays a decision tree and a series of rules that can be used to assign individual records to child nodes based on the values of different predictors. Random forest shows a plot that indicates the relative importance of each predictor in estimating the generated model. Once the model is validated, user can save it on Disc, for future uses, as standalone model. This is done by saving the model R object to a file repository and restores that object again to performance analysis on new data without the need for further processing.



Predictive Model: Drop 1 year DT

Age: 23 (17–50)

Distance from home: 108 (1–1,000)

First term GPA: 3 (0–18)

Number of validated modules: 0

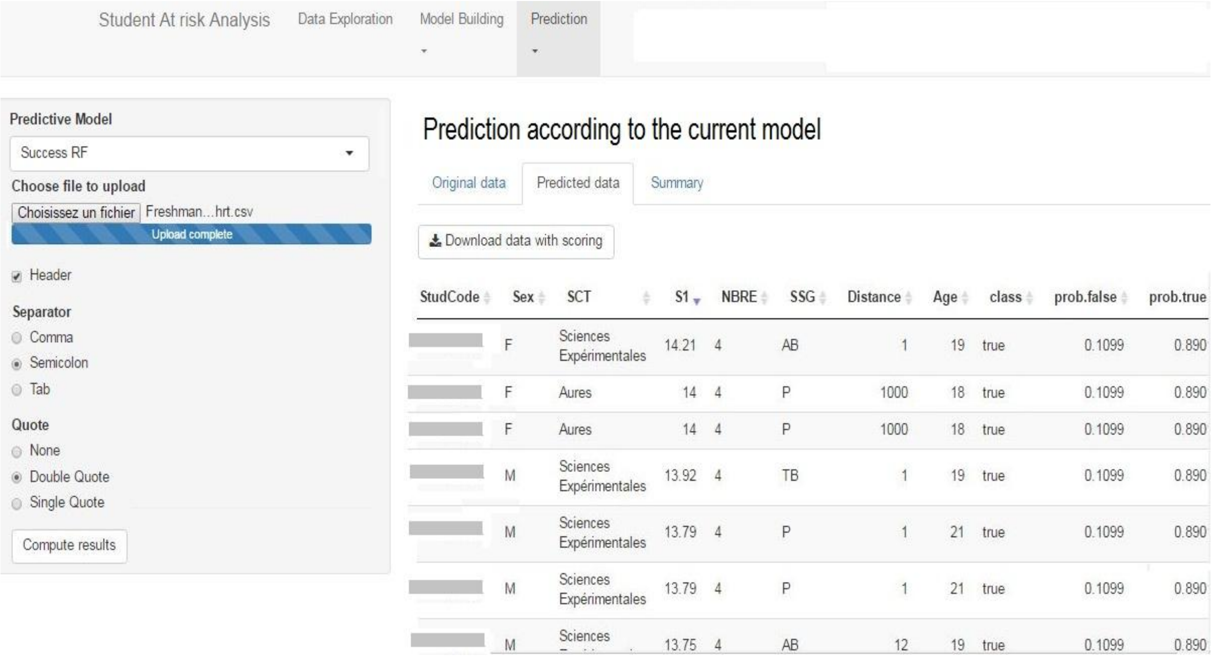
Prediction according to the current Inputs

```
[1] "This Student has  91.63 % probability  of droping out within one year. The predicted class is:  true"
```

Figure 4 : Predicting Student Drop out based on predictive models

The last option is “Prediction”. Through this menu users could take advantage of learning results in predicting student outcomes. Scoring new data can be done in two manners, by Using “individual prediction”, or via uploading Student data file and applying scoring function to all students of freshman cohort.

As shown in figure 4, the model to be run and the data that be scored are indicated in the side bar panel of “individual prediction” page. When any input of predictive variables changes, the classification function will automatically display the predicted class according to the selected model.



Predictive Model: Success RF

Choose file to upload: Freshman...hrt.csv

☑ Header

Separator
○ Comma
● Semicolon
○ Tab

Quote
○ None
● Double Quote
○ Single Quote

Prediction according to the current model

StudCode	Sex	SCT	S1	NBRE	SSG	Distance	Age	class	prob.false	prob.true
[redacted]	F	Sciences Expérimentales	14.21	4	AB	1	19	true	0.1099	0.890
[redacted]	F	Aures	14	4	P	1000	18	true	0.1099	0.890
[redacted]	F	Aures	14	4	P	1000	18	true	0.1099	0.890
[redacted]	M	Sciences Expérimentales	13.92	4	TB	1	19	true	0.1099	0.890
[redacted]	M	Sciences Expérimentales	13.79	4	P	1	21	true	0.1099	0.890
[redacted]	M	Sciences Expérimentales	13.79	4	P	1	21	true	0.1099	0.890
[redacted]	M	Sciences	13.75	4	AB	12	19	true	0.1099	0.890

Figure 5 : Scoring a group of new students at one time

The second way consist of scoring data for a group of students. This option provide more efficient way for scoring students data, it consist of uploading freshman Student data as csv file without dependent variables and then applying scoring function to all students of freshman cohort in just one time. As shown in figure 5, the preview of uploaded data can be seen in the “Original data” tab panel. After computing prediction a new data frame including predicted outcome variable is displayed in the “predicted data” panel. The results of this scoring task might be downloaded for subsequent use.

5. DISCUSSION AND CONCLUSION

Through this work, classification analysis as an EDM method has been discussed and implemented using R language and shiny framework. In order to predict student achievement and to identify the significant variables that affect educational success and failure, CART and Random forests as a bagging ensemble method was examined. Both of the two algorithms have accuracy around 88 % for predicting student drop out and 77 for student failure. The obtained results also reveal that the Number of validated modules at the first term, First term GPA, and Secondary school grade are the top three most important attributes separating persistent and non-persistent students. In addition to model building and validation, the produced web based system Include a scoring engine that could be accessed by the related decision makers, to predict freshman student risk of drop out and non-graduation. This system was based on an off-line learning. Nevertheless, it has a relevant potential to be used on-line for student prediction engine as part of the university decision support system.

Future work will focus on the following directions. First, the actual system will be enhanced by other Ensemble learning algorithms such as C5.0 and Stochastic Gradient Boosting and by performing tuning optimization for existent methods in order to boost the Accuracy of predictive models. Second, it will be expanded with other educational indicators such as student ‘grades and the time to degree completion.

6. REFERENCES

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

[2] R. S. Baker and K. Yacef, “The state of educational data mining in 2009: A review and future visions,” JEDM-Journal of Educational Data Mining, vol. 1, no. 1, pp. 3–17, 2009.

[3] G. W. Dekker, M. Pechenizkiy, and J. M. Vleeshouwers, “Predicting Students Drop Out: A Case Study.,” International Working Group on Educational Data Mining, 2009.

[4] S. K. Yadav, B. Bharadwaj, and S. Pal, “Mining Education data to predict student’s retention: a comparative study,” International Journal of Advanced Computer Science and Applications, vol. 2, no. 6, 2012.

[5] D. Kabakchieva, “Predicting Student Performance by Using Data Mining Methods for Classification,” Cybernetics and Information Technologies, vol. 13, no. 1, Jan. 2013.

[6] Z. J. Kovacic, “Predicting student success by mining enrolment data,” Research in Higher Education Journal, vol. 15, p. 1, 2012.

[7] S. Pal, “Mining Educational Data to Reduce Dropout Rates of Engineering Students,” International Journal of Information Engineering and Electronic Business, vol. 4, no. 2, pp. 1–7, Apr. 2012.

[8] M. A. Yehuala, “Application Of Data Mining Techniques For Student Success And Failure Prediction (The Case Of Debre_Markos University),” INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME, vol. 4, no. 4, 2015.

[9] B. R. Subedi and B. Johnson, “Predicting High School Graduation and Dropout Using a Hierarchical Generalized Linear Model Approach,” 2007.

[10] L. Agnihotri and A. Ott, “Building a student at-risk model: an end-to-end perspective,” in Proc. Int. Conference on Educational Data Mining Conference (EDM), 2014, pp. 209–212.

[11] C. Beeley, Web Application Development with R Using Shiny. Packt Publishing, 2013.

[12] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition, 2nd ed. 2009. Corr. 7th printing 2013 edition. New York, NY: Springer, 2011.

[13] M. J. Zaki and W. Meira, Data mining and analysis: fundamental concepts and algorithms. New York, NY: Cambridge University Press, 2014.

[14] A.-L. Boulesteix, S. Janitza, J. Kruppa, and I. R. König, “Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics,” Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 2, no. 6, pp. 493–507, 2012.

[15] J. Strickland, Predictive Analytics Using R. Lulu.com, 2015.