Identifying Risky Dropout Student Profiles using Machine Learning Models

Sharmin Sultana Prite
sharmin.prite5@gmail.com
Teresa Gonçalves
tcg@uevora.pt
Luís Rato
lmr@uevora.pt

Departamento de Informática, Universidade de Évora, Portugal

Abstract

Student dropout prediction is essential to measure the success of an education institute system. This paper focuses on identifying the dropout risk at the University of Évora based on student's academic performance. Educational data was collected from four different programs, from the academic years of 2006/2007 until 2018/2019. After gathering the raw data, some data pre-processing was done aiming to build a dataset capable of being used by Machine Learning algorithms. Decision trees, Naïve Bayes, Support Vector Machines and Random Forests were evaluated, with the best model reaching an accuracy of around 96% when distinguishing between risky dropout and non-dropout students.

keywords: Machine Learning, Data Mining, Educational Data, Random Forest, Support Vactor Machines

1 Introduction

Nowadays, we live in the information era where acquiring data is easy and storing is inexpensive. Information is also the primary ingredient to generate new knowledge. The data mining can be applied in various real-life application like market analysis, education, and scientific exploration [6]. The use of data mining technique to analyze an educational database is absolutely expected to be a great benefit to the higher educational institutions.

Student dropout in Higher Education Institutions (HEIs) is, nowadays, a crucial concern for educators and managers. It also became a focus for researchers. Knowing, beforehand, the students at risk of dropping out allow higher education players to take measures that can contribute to an improvement in the institution success rate. Reasons for a dropout can be related to economical, social and psychological issues [1].

Anupama Kumar *et al.* [7] used a decision tree to help tutors identify the weak students and improve their performance before dropouts. Similarly, William C. Blanchfield [3] described a method of identifying college dropouts tested at Utica College of Syracuse University; he used multiple discriminant analysis to identify dropouts, reaching an accuracy of around 73%. Researchers from the University of Wuppertal developed an Early Detection System (EDS) [2] using administrative student data from a state and private universities to predict student success as a basis for targeted intervention; the EDS used regression analysis, neural networks, decision trees, and AdaBoost to identify student characteristics which distinguish potential dropouts from graduates. Yujing Chen *et al.* [4] developed and evaluated a survival analysis framework for the early identification of students at the risk of dropping out. In summary, existing approaches including logistic regression, decision trees and boosting showed good performance for early prediction of at-risk students and were also able to predict when a student will dropout. Given existing approaches, authors of this article tried different machine learning algorithms namely Decision trees (DT), Naïve Bayes (NB), Support Vector Machines (SVM) and Random Forests (RF) over academic data.This work uses student academic data from 4 different programs at the University of Évora to build classification models able to identify students at risk of dropping out.

The rest of the paper is organized as follows: Section 2 introduces the data used in this work, while Section 3 presents the developed work: data preprocessing, dataset generation, experimental setup, and results and their discussion. Finally, Section 4 concludes the paper and discusses future work.

2 Study Data

For this study, the students' full academic record was gathered. It considers four undergraduate study programs: Management, Biology, Computer Science and Nursing, during 13 academic years (from 2006/2007 to 2018/2019).

The student academic record includes information about course enrollments and corresponding results during the student university life: from the first year when student register at the university until graduation or dropout. Students were anonymized, and updates on study programs were considered. The list of information gathered from the information system are: *school year, degree, department, course code, course unit, regime, course credits, course name, edition, speciality, semester, time, type, student id, student type, mark, result, final status*.

3 Developed work

As previously mentioned, this work aims at creating a classification model using Machine Learning techniques to identify students at risk of dropping out so it could be used by authorities of HEIs to take possible actions aiming to reduce the number of dropouts. Figure 1 presents the block diagram of the developed work.

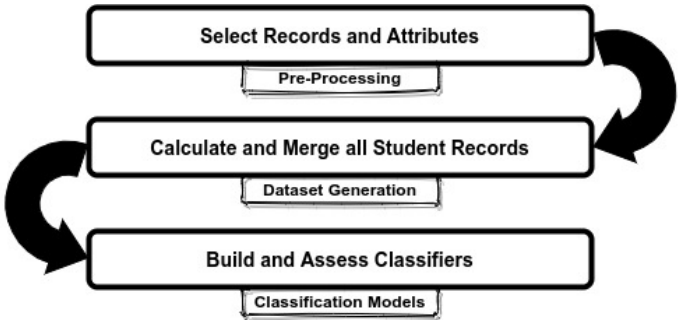

Figure 1: Developed work.

3.1 Pre-Processing

As already mentioned, student data was collected over a period of 13 years from four different undergraduate programs. In these programs, nursing is a four years program (totalling 240 credits), and the rest are three years of programs (totalling 180 credits).

The total number of enrollment records retrieved was 119407: 33731, 21328, 28689 and 35659 records for Management, Biology, Computer Science and Nursing, respectively. Information about the number of years taken to conclude the program is presented in Table 1.

Program	Min	Max	Avg	Stdev
Management	3	12	3.71	1.09
Biology	3	9	3.80	1.05
Computer Science	3	10	5.13	1.30
Nursing	3	13	4.25	0.64

Table 1: Information about the time taken to complete the programs.

Students enrolled at the university in the academic year of 2018/2019 were removed since at the time of data retrieval there was no academic record for them; this resulted in a total of 2934 students distributed as presented in Table 2.

From the data available at the University's information system, the following enrolment attributes were considered: *Academic_Year*, *Management*, *Biology*, *Computer Science*, *Nursing*, *Semester*, *Std_Id*, *Course*, *Credits*, *Mark*, *Final_Status*. Final_Status has two values: **S** means student pass the course, and **N** means student miss or fail the course. Course enrollment records without a value for Final_Status were removed because student enrolled the course but had not done any course activity.

Program	Number of students
Management	885
Biology	556
Computer Science	598
Nursing	895
Total	2934

Table 2: Number of students per study program

3.2 Dataset Construction

Using the retrieved data, and for each student, the annual student performance was calculated, and all the annual records were joint together to generate a single example; this example represents the academic path of a specific student.

The annual student performance is given by three attributes: the total number of enrolled and completed credits and average grade. This information was compiled for the student's five most recent academic years plus the performance calculated over the remaining student academic life. For students that successfully completed the program in less that 5 academic years, the values for attributes of oldest years were filled with zeros.

At the end, a dataset of 13 years composed by 21 attributes was built. Table 3 presents them.

Name	Number	Type
program_ects	1	int
program_name: man, bio, cs, nurse	4	bool (all)
year_0: enrol , avg_grade	2	int, float
year_1: enrol, complete, avg_grade	3	int, int, float
year_2: enrol, complete, avg_grade	3	int, int, float
year_3: enrol, complete, avg_grade	3	int, int, float
year_4: enrol, complete, avg_grade	3	int, int, float
year_rest: enrol, complete	2	int, int

Table 3: Dataset attributes.

A class label was then given to each example: success and unsuccess. The rule used was the following:

```
if registred = 2017 and completedCredit > 0
then SUCCESS
elseif registred < 2017 and completedCredit >= 210/150[a]
then SUCCESS
else UNSUCCESS
```

[a] 210 for nursing; 150 for other programs. This corresponds completing all except the credits of one semester.

The attributes and rules just described building the dataset were chosen considering a set of preliminary experiments that analysed other sets aiming to determine student success or unsuccess.

3.3 Classification Models

Four machine learning algorithms were used to build classifier models: Decision Tree (DT), Naïve Bayes (NB), Support Vector Machines (SVM), and Random Forest (RF). Weka 3.8.1 toolkit [5] was used for the experiments.

To tested the importance of the enrolled program and grade information, four different attribute subsets were used to build classification models:

- att_1: without *program_name*, without *avg_grade*
- att_2: with *program_name*, without *avg_grade*
- att_3: without *program_name*, with *avg_grade*
- att_4: with *program_name*, with *avg_grade*

The dataset was split into 70% of examples for training (2052 samples) and 30% for testing (882 samples). Then build the model using a training set and re-evaluated the model using the test set. To fine-tune the classifier algorithms, 10-folds cross-validation over the train set using the accuracy measure. Here, default parameter of all algorithms produce best results.

Table 4 shows the results obtained over the test set for each of the machine learning algorithms. As can be seen from the table, the overall performance by each algorithm over all the attributes is similar. The maximum difference of results is ranging from 0.67% to 1.71%, where RF has a minimum variation of 0.67%, and DT has a maximum of 1.71%. RF is outperforming all other algorithms by achieving 96.83% of accuracy.

Attributes	DT (%)	NB (%)	RF (%)	SVM (%)
Att_1	94.44	92.86	96.49	95.46
Att_2	94.90	92.74	96.15	96.15
Att_3	96.03	92.40	**96.83**	95.92
Att_4	**96.15**	**93.65**	96.60	**96.49**

Table 4: Accuracy results over test set.

Attributes	DT (%)	NB (%)	RF (%)	SVM (%)
Att_1	90.9	85.9	94.2	92.4
Att_2	91.7	88.4	93.7	93.6
Att_3	93.6	88.2	94.8	93.2
Att_4	93.8	89.9	94.4	94.2

Table 5: F-Measure Results over test set (Unsuccess class).

The F-measure results over unsuccess class of test set present in Table 5. The maximum difference of results is ranging from 1.1% to 4.0%, where RF has a minimum variation of 1.1%, and NB has a maximum of 4.0%. RF is out-performing all other algorithms by achieving 94.8% of F-measure.

From tables 4 and 5, it's not concluded that the best performance by RF is only achievable when all available attributes are not considered compared to the considering all attributes as the difference is the only 0.2% to 0.4%.

4 Conclusions and Future Work

This work presents an approach to identify dropout students by detecting risky profiles. It describes the available data, its preprocessing to generate a proper dataset and presents the results obtained using different machine learning algorithms. Using yearly enrollment information along with the study program and average grades an accuracy of around 96% for detecting risky dropout profiles was reached.

As future work, and to verify the results presented here we intend to enlarge the dataset to include more programs and, if possible, include student's personal, financial and social media information as attributes to improve the Machine Learning model.

Funding

This work was supported by the Erasmus Mundus LEADER (*Links in Europe and Asia for engineering, eDucation, Enterprise and Research Organization*) project.

References

[1] Jeff Allen, Steven B Robbins, Alex Casillas, and In-Sue Oh. Third-year college retention and transfer: Effects of academic performance, motivation, and social connectedness. *Research in Higher Education*, 49(7):647–664, 2008.

[2] Johannes Berens, Kerstin Schneider, Simon Görtz, Simon Oster, and Julian Burghoff. Early detection of students at risk–predicting student dropouts using administrative student data and machine learning methods. *CESifo Working Paper*, 2018.

[3] William C Blanchfield. College dropout identification: An economic analysis. *The Journal of Human Resources*, 7(4):540–544, 1972.

[4] Yujing Chen, Aditya Johri, and Huzefa Rangwala. Running out of stem: a comparative study across stem majors of college students at-risk of dropping out early. In *Proceedings of the 8th international conference on learning analytics and knowledge*, pages 270–279, 2018.

[5] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA data mining software: an update. *SIGKDD Explorations*, 11(1):10–18, 2009.

[6] Jiawei Han, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

[7] S Anupama Kumar and MN Vijayalakshmi. Implication of classification techniques in predicting student's recital. *Int. J. Data Mining Knowl. Manage. Process (IJDKP)*, 1(5):41–51, 2011.

*/