Identifying Risky Dropout Student Profiles using Machine Learning Models

Sharmin Sultana Prite, Teresa Gonçalves, Luís Rato

Departamento de Informática, Universidade de Évora, Portugal

sharmin.prite5@gmail.com, (tcg,lmr)@uevora.pt

Motivation and Objectives

⇒ Dropout prediction is essential to measure the success of an education institute system

⇒ In Portugal has the fourth highest rate of early school leaving in their academic year [2]

⇒ Reasons for a dropout can be related to economical, social and psychological issues [1]

⇒ Nowadays, Student dropout in HEIs is a crucial concern for educators and researchers

⇒ Requirement for fast and early predict dropout student

⇒ Automatic system that analysis student academic data and identify risky student profile

Study Data

⇒ Data from four different undergraduate programs:
Management, Biology, Computer Science and Nursing

⇒ Total 13 academic years Records (from 2006/2007 to 2018/2019)

⇒ Count yearly academic results

⇒ Information from university system

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

Table 1: List of information gathered from the information system

⇒ Total number of enrollment records was **119407**

Developed work

Figure 1 presents the block diagram of the developed work.

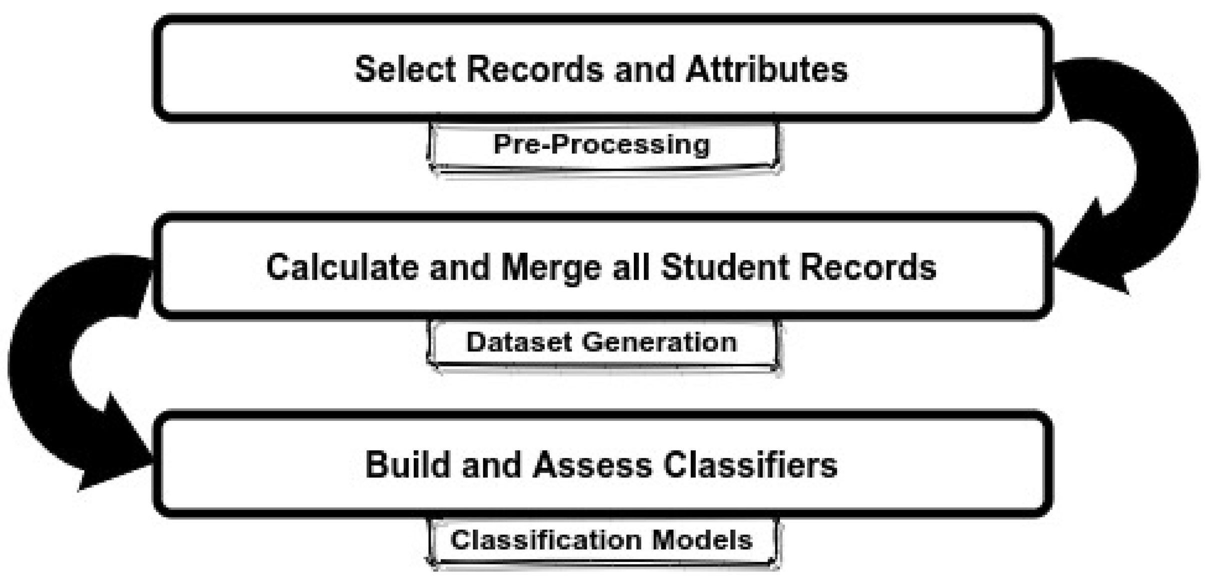

Figure 1: Developed work

Pre-Processing

⇒ Removed 2018/2019 enrolled student since they don't have academic record

⇒ Total 11 enrollment attributes considered

Academic Year	Management	Biology
Computer Science	Nursing	Semester
Student Id	Course	Credits
Mark	Final Status	

Table 2: Considered enrolment attributes list.

⇒ Removed enrollment records without a value for Final_Status

⇒ After pre-processing done, total students found **2934**

Dataset Construction

A dataset of 13 years composed by 21 attributes was built.

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.

Classification Models

Four machine learning algorithms used to build models:

1. Decision Tree (DT)
2. Naïve Bayes (NB)
3. Support Vector Machines (SVM)
4. Random Forest (RF)

Importance of enrolled program and grade, 4 different attribute subsets used to build models:

1. att_1: without *program_name*, without *avg_grade*
2. att_2: with *program_name*, without *avg_grade*
3. att_3: without *program_name*, with *avg_grade*
4. att_4: with *program_name*, with *avg_grade*

Experiment Setup

⇒ 70% of examples for training (2052 samples)

⇒ 30% of examples for testing (882 samples)

⇒ 70% training data used for build the model and 30% used for test the model

⇒ 10-folds cross-validation with default parameters

⇒ Weka 3.8.1 toolkit [3] used for experiments

Results

⇒ RF has a minimum variation of 0.67%

⇒ DT has a maximum of 1.71%

⇒ RF is out-performing 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.

⇒ Maximum difference of results is ranging from 1.1% to 4.0%

⇒ RF is out-performing all other algorithms by achieving 94.8% of F-measure.

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).

Conclusions and Future Work

⇒ Presents a machine learning approach to identify dropout students by detecting risky profiles

⇒ An accuracy of around 96% for detecting risky dropout profiles was reached.

⇒ Enlarge the dataset to include more programs

⇒ Include student's personal, financial and social media information

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] T Andrei, D Teodorescu, and B Oancea. Characteristics and causes of school dropout in the countries of the european union. *Procedia-Social and Behavioral Sciences*, 28:328–332, 2011.

[3] 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.