SECONDARY SCHOOL STUDENT PERFORMANCE PREDICTION USING DATA MINING

Bimurta Bismoy Sanchi
ID : 011151285
United International University
Dhaka, Bangladesh
bsanchi151285@bscse.uiu.ac.bd

Md. Easin Arafat
ID: 011152047
United International University
Dhaka, Bangladesh
marafat152047@bscse.uiu.ac.bd

Irfan Ahmed
ID: 011151174
United International University
Dhaka, Bangladesh
iahmad151174@bscse.uiu.ac.bd

***Abstract*—Predict student performance with past evaluations, some relevant features has done in this paper. Because of lack of success in the two important subjects like mathematics and Portuguese language and also early school leaving rate are the main causes. By the past evolution data (grades, demographic social or School rated features) and also some extra relevant features (Number of absences, parents jobs, parents education, Alcohol consumption) this paper predicts student achievements and finds out the reasons behind the failure. Four data mining(DM) models (Decision Tree, Random Forest, Neural Networks and Support Vector Machine) were used for the output. The two classes are Mathematics and Portuguese Language. The three input sections (with and without grades) were tested and the classes were modeled under binary /5 leveled classification and regression. The outputs will be helpful for finding out the failure reasons and so improving educational quality and also works as a good efficient student prediction tools.**

***Index Terms*—component, formatting, style, styling, insert**

I. INTRODUCTION

Introduction: Education is a must for all kinds of progresses. But Portuguese educational level is not so impressing. Student failure and dropping out rate make the difference between other European country and Portugal in 2006 early school leaving rate was 40% in Portugal but only 15% in average of other countries. the failure rate is extremely serious in mathematics and Portuguese language this paper finds out the reasons behind failure so that the reasons came out and could be solved. The research actually works on two questions. IS it possible to predict students performance? What are the factors that affect student achievements? Also there are so many other questions as domain that could be found out using DM models.There are several works on similar topics. Those works had different input variables, different accuracy rates of different methods. In 2003 Minaei Bidgoli, was modeled online student grades using binary 3- level and 9-level models. Decision tress and Neural network were given output 94% accuracy rate in binary, 72% in 3-level and 62% in 9-level. In 2004, Kotsiantis, predicted the performance of computer science students only using binary classifier. Naive Bayes was given 74% accuracy rate. In his paper, Mathematics and Portuguese modeled under 3 DM methods :- binary classifier, 5- level classifier and regression. Three input setups (with and without grades) is used. Four DM algorithms (Decision Tress, Random Forest, Naive Bayes, Support Vector Machine) is tested. This research worked for two goals. One, to predict student achievement. Two, identify the reasons that affect educational success. And finally the most relevant features and best methods are identify by an explanatory analysis.

Identify applicable funding agency here. If none, delete this.

II. DATASET

This data was collected during 2005-2006 school year from two Portuguese secondary public school where the paper was published in 2008, So it was recent dataset for that paper.

A. DATA COLLECTION AND RESOURCES

Final version contained 37 questions and it was answered by 788 students 111 answers were discarded due to lack of identification details Finally, the data was integrated into two datasets related to-
1. Mathematics
2. Portuguese Language
The two datasets were modeled under
- binary/five-level classification and
- regression tasks.
This data approach student achievement in secondary education of two Portuguese schools. The data attributes include student grades, demographic, social and school related features. And it was collected by using school reports and questionnaires. Two datasets are provided regarding the performance in two distinct subjects: Mathematics (mat) and Portuguese language (por). In [Cortez and Silva, 2008]

B. DATASET DESCRIPTION

Data Set Characteristics: Multivariate
Area: Social
Attribute Characteristics: Integer
Date Donated 2014-11-27
Associated Tasks: Classification, Regression

Missing Values? N/A

Among all the 33 attributes,

Binary(13 variables) e.g; male/female, Numeric(16 variables) e.g; from 1 very low to 5 very high, Nominal(4 variables) e.g; Job title, Name.

Attributes for both student-mat.csv (Math course) and student-por.csv (Portuguese language course) datasets:
1. school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
2. sex - student's sex (binary: 'F' - female or 'M' - male)
3. age - student's age (numeric: from 15 to 22)
4. address - student's home address type (binary: 'U' - urban or 'R' - rural)
5. famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
6. Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
7. Medu - mother's education (numeric: 0 - none, 1 - primary education
8. Fedu - father's education (numeric: 0 - none, 1 - primary education
9. Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police)
10. Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police)
11. reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
12. guardian - student's guardian (nominal: 'mother', 'father' or 'other')
13. traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
14. studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
15. failures - number of past class failures (numeric: n if 1<=n<3, else 4)
16. schoolsup - extra educational support (binary: yes or no)
17. famsup - family educational support (binary: yes or no)
18. paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
19. activities - extra-curricular activities (binary: yes or no)
20. nursery - attended nursery school (binary: yes or no)
21. higher - wants to take higher education (binary: yes or no)
22. internet - Internet access at home (binary: yes or no)
23. romantic - with a romantic relationship (binary: yes or no)
24. famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
25. freetime - free time after school (numeric: from 1 - very low to 5 - very high)
26. goout - going out with friends (numeric: from 1 - very low to 5 - very high)
27. Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
28. Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
29. health - current health status (numeric: from 1 - very bad to 5 - very good)
30. absences - number of school absences (numeric: from 0 to 93)
these grades are related with the course subject, Math or Portuguese:
31. G1 - first period grade (numeric: from 0 to 20)
32. G2 - second period grade (numeric: from 0 to 20)
33. G3 - final grade (numeric: from 0 to 20, output target)

In mathematics dataset, there are 33 features along with 395 instances. In Portugues language dataset, there are 649 instances also along with 33 features.

***Index Terms*—**

III. APPROACHES

In our Working stage we have applied 3 approaches Binary Classification, 5 level Classification, Regression.

- **For Binary Classification**
  if(G3 is greater than 10)pass;
  else fail;
- **For 5 level**
  16-20 excellent (A)
  14-15 good (B)
  12-13 Satisfactory (C)
  10-11 sufficient(D)
  0-9 Fail (F)
- **The regression is based on G3 value**
  Numeric out put between 0-20

A. Predictive Performance

Three input configuration were tested for each DM model

- **A Know G1 and G2**
- **B Drop G2**
- **C Drop G1 and G2**

B. Computational environment

- **Jupyter Note book**
- **Python 3**

IV. ALGORITHMS

A. Encode

For each feature, encode to categorical values .Encode G1, G2, G3 as pass or fail binary values

B. Split Data

- **Training and Test**
- **Test Size 0.2**
- **395 rows**
- **Test = (395*20)/100 =79**

C. Train our Classifaier

Support Vector Machine Classifier model

```
Clf  =  SVC ()
Clf.fit(x_train,y_train)
y_train =   pop(G3)
x_train = full data set with features
```

D. 3 step calculation

- **Model Accuracy Knowing G1 G2 Scores**
- **Model Accuracy Knowing Only G1 Score**
- **Model Accuracy Without Knowing Scores**

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	...
0	GP	F	18	U	GT3	A	4	4	at_home	teacher	...
1	GP	F	17	U	GT3	T	1	1	at_home	other	...
2	GP	F	15	U	LE3	T	1	1	at_home	other	...
3	GP	F	15	U	GT3	T	4	2	health	services	...
4	GP	F	16	U	GT3	T	3	3	other	other	...

5 rows × 33 columns

Fig. 1. Math Data Overhead

	age	Medu	Fedu	traveltime	studytime	failures
count	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000
mean	16.696203	2.749367	2.521519	1.448101	2.035443	0.334177
std	1.276043	1.094735	1.088201	0.697505	0.839240	0.743651
min	15.000000	0.000000	0.000000	1.000000	1.000000	0.000000
25%	16.000000	2.000000	2.000000	1.000000	1.000000	0.000000
50%	17.000000	3.000000	2.000000	1.000000	2.000000	0.000000
75%	18.000000	4.000000	3.000000	2.000000	2.000000	0.000000
max	22.000000	4.000000	4.000000	4.000000	4.000000	3.000000

Fig. 2. Math Data Describe

V. Output

A. Observed Knowledge

Before observing knowledge from code we tried to gain knowledge from our reference paper[1].This table describe the mathematics data set accuracy for different model. Here binary classification has been applied. For set up A and B NV has given the best output. for set up C RF has given the best output.

Input Setup	NV	NN	SVM	DT	RF
A	**91.9**	88.3	86.3	90.7	91.2
B	**83.8**	81.3	80.5	83.1	83.0
C	67.1	66.3	70.6	65.3	**70.5**

Table 1 : Binary classification accuracy result using **mat.csv** data set

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	...
0	GP	F	18	U	GT3	A	4	4	at_home	teacher	...
1	GP	F	17	U	GT3	T	1	1	at_home	other	...
2	GP	F	15	U	LE3	T	1	1	at_home	other	...
3	GP	F	15	U	GT3	T	4	2	health	services	...
4	GP	F	16	U	GT3	T	3	3	other	other	...

5 rows × 33 columns

Fig. 3. Portuguese Data Overhead

	age	Medu	Fedu	traveltime	studytime	failures
count	649.000000	649.000000	649.000000	649.000000	649.000000	649.000000
mean	16.744222	2.514638	2.306626	1.568567	1.930663	0.221880
std	1.218138	1.134552	1.099931	0.748660	0.829510	0.593235
min	15.000000	0.000000	0.000000	1.000000	1.000000	0.000000
25%	16.000000	2.000000	1.000000	1.000000	1.000000	0.000000
50%	17.000000	2.000000	2.000000	1.000000	2.000000	0.000000
75%	18.000000	4.000000	3.000000	2.000000	2.000000	0.000000
max	22.000000	4.000000	4.000000	4.000000	4.000000	3.000000

Fig. 4. Portuguese Data Describe

1) Output for Calculation 1:

```
Model Accuracy Knowing G1 & G2 Scores
Mean Model Accuracy 0.9178075
Confusion Matrix
[[23,2]
[4,50]]
False pass rate 0.08
False fail rate 0.0740740
```

2) Output for Calculation 2:

```
Model Accuracy Knowing only G1 Score
Mean Model Accuracy 0.832341269
Confusion Matrix
[[23,3]
[8,46]]
False pass rate 0.12
False fail rate 0.1481481
```

3) Output for Calculation 3:

```
Model Accuracy without Knowing Scores
Mean Model Accuracy 0.65808531
Confusion Matrix
[[6,19]
[2,52]]
False pass rate 0.76
False fail rate 0.03703
```

VI. Our Extended Works

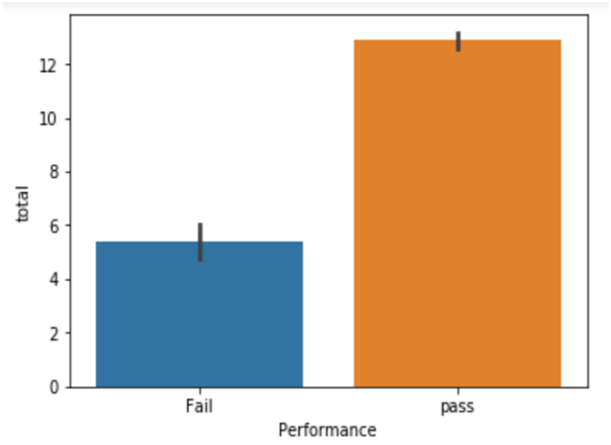

Fig. 5. Amount of pass and failure

We had tested every features and visualize the final out come and plot them in to a histogram. We applied our code on mathematics data set. For this Visualization we used binary classification.

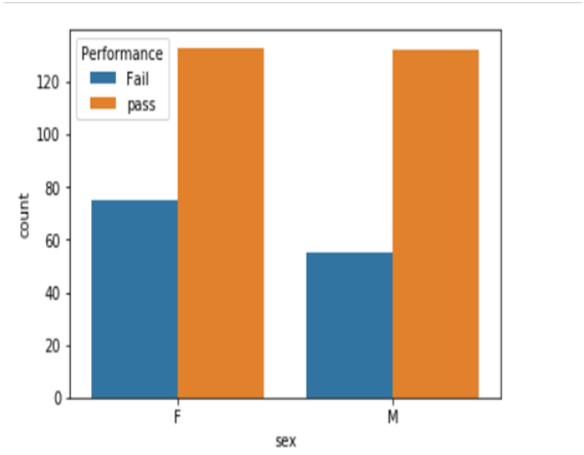

Fig. 6. Performance based on Gender

On **figure 6** we demonstrated the effects of result based on our feature no 2. sex - student's sex (binary: 'F'-female or 'M'-male)

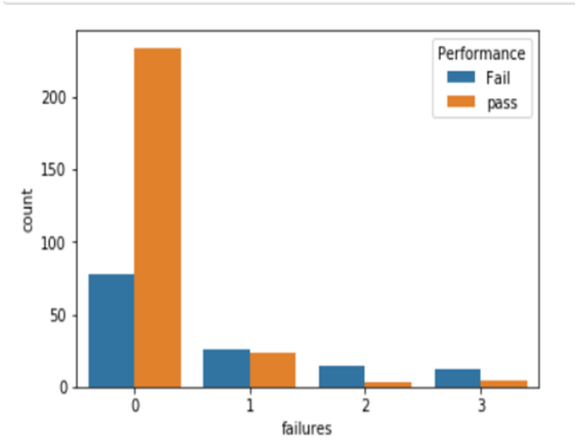

Fig. 7. Performance based on Fail in 0,1,2,3 subjects

On **figure 7** we ploted our data set result based on Failure.Here failures means number of past class failures.Here we can observe that the students who didn't fail in their previous classes their passing rate is high. On the other hand based on the previous failure history students need extra cares who failed on previous classes.

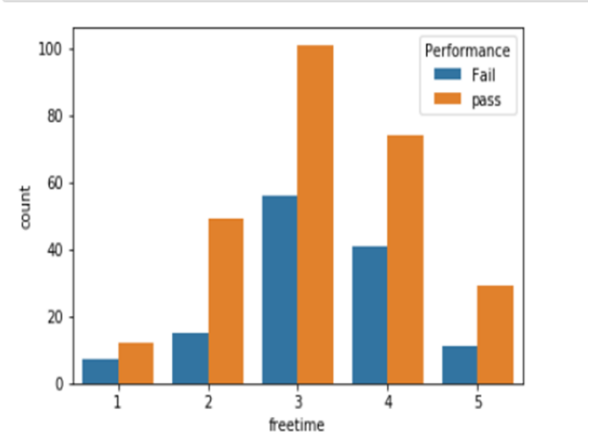

Fig. 8. Performance based on Free Time

On **figure 8** describe free time after school also affects the students performance. Free time was our 25 no feature it is in numeric form 1 means too low and 5 means too high.

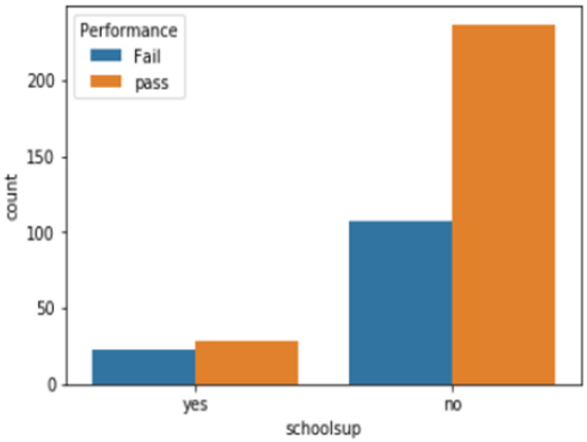

Fig. 9. Performance based on extra educational support

On **figure 9** shows that the students who didn't get any extra educational support their passing rate is higher.

VII. RESOURCES

Code links

- https://bit.ly/2GbzeQa
- https://github.com/sachanganesh/student-performance-prediction/blob/master/model.py

Dataset links

- https://archive.ics.uci.edu/ml/machine-learning-databases/00320/

REFERENCES

[1] Paulo Cortez and Alice Silva *USING DATA MINING TO PREDICT SECONDARY SCHOOL STUDENT PERFORMANCE*, Dep. Information Systems/Algoritmi RD Centre University of Minho.
4800-058 Guimaraes, PORTUGAL.

[2] Eurostat,2007.Early school-leavers. http://epp.eurostat.ec.europa.eu/

[3] Pritchard M. and Wilson S., 2003.Using Emotional andSocial Factors To Predict Student Success.Journal ofCollege Student Development.