EXPLORATORY ANALYSIS OF STUDENT PERFORMANCE IN MATHEMATICS

Imran Manzoor, Lorick Jain and Kumar Millind
Dep. Computer Science, PES University
Email: imran.manzoor31@gmail,com, lorick.jain@gmail.com

ABSTRACT

There are many factors that affect a students' performance in school. Research[1] has been done on the same dataset who's aim was to compare different techniques in predicting the final grade. It is clear that previous grades help in the prediction, but there are the other underlying factors. The predictive models used here are Decision Trees and SVM.

INTRODUCTION

The data chosen is Student Performance Dataset[2]. It contains grades as well as other information collected through questionnaires in two schools in Portugal. There were exams in three periods in the academic year: G1, G2, G3 for 1st, 2nd and 3rd periods respectively. Other columns include Mother Education, Father Education, Travel time, Study time etc.

The Research[1] mentioned before was to find a good model for this dataset taking all factors. Our aim is to find the key factors to get the best prediction results.

MATERIALS AND METHODS USED

The student dataset had no missing values. The columns G1, G2 and G3 had scores between 0 and 20. The scores were spilt into 5 classes following what was done in the previous research[1]. Scores between 16 to 20 got class 1(A), 14 to 15 got class 2(B), 12 to 13 got class 3(C), 10 to 11 got class 4(D) and 0 to 9 got class 5(F). This was achieved using python Pandas.

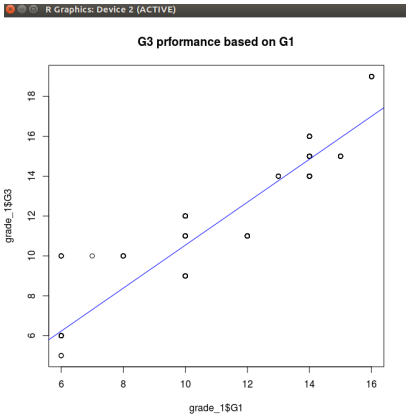
To find the key factors we tried drawing correlations from each factor to the final grade. This was done in MS Excel, R and Python.

The description of the dataset is given in table 1, thanks to [1].

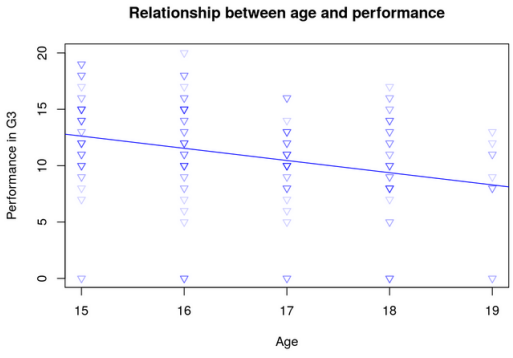
MODELS

The previous research[1] tried to model data using Decision Trees, Random Forsests, SVMs, Neural Networks and Naive Bayes. They used all the factors and compared the results by using G1 and G2, or only G1 or only G2 to predict G3. We tried to improve their accuracy in Decision Trees and SVMs using only the key factors.

RESULTS

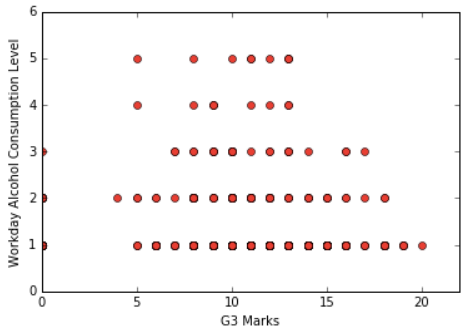


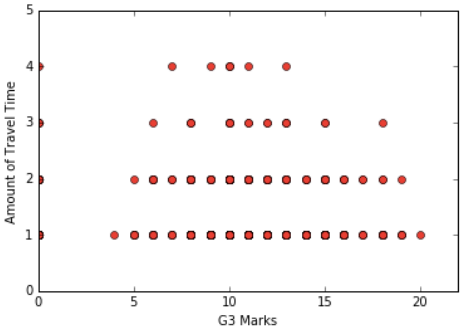
Plot Linear Regression



Attribute	Description (Domain)
sex	student's sex (binary: female or male)
age	student's age (numeric: from 15 to 22)
school	student's school (binary: *Gabriel Pereira* or *Mousinho da Silveira*)
address	student's home address type (binary: urban or rural)
Pstatus	parent's cohabitation status (binary: living together or apart)
Medu	mother's education (numeric: from 0 to 4[a])
Mjob	mother's job (nominal[b])
Fedu	father's education (numeric: from 0 to 4[a])
Fjob	father's job (nominal[b])
guardian	student's guardian (nominal: mother, father or other)
famsize	family size (binary: ≤ 3 or > 3)
famrel	quality of family relationships (numeric: from 1 – very bad to 5 – excellent)
reason	reason to choose this school (nominal: close to home, school reputation, course preference or other)
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).
studytime	weekly study time (numeric: 1 – $<$ 2 hours, 2 – 2 to 5 hours, 3 – 5 to 10 hours or 4 – $>$ 10 hours)
failures	number of past class failures (numeric: n if $1 \leq n < 3$, else 4)
schoolsup	extra educational school support (binary: yes or no)
famsup	family educational support (binary: yes or no)
activities	extra-curricular activities (binary: yes or no)
paidclass	extra paid classes (binary: yes or no)
internet	Internet access at home (binary: yes or no)
nursery	attended nursery school (binary: yes or no)
higher	wants to take higher education (binary: yes or no)
romantic	with a romantic relationship (binary: yes or no)
freetime	free time after school (numeric: from 1 – very low to 5 – very high)
goout	going out with friends (numeric: from 1 – very low to 5 – very high)
Walc	weekend alcohol consumption (numeric: from 1 – very low to 5 – very high)
Dalc	workday alcohol consumption (numeric: from 1 – very low to 5 – very high)
health	current health status (numeric: from 1 – very bad to 5 – very good)
absences	number of school absences (numeric: from 0 to 93)
G1	first period grade (numeric: from 0 to 20)
G2	second period grade (numeric: from 0 to 20)
G3	final grade (numeric: from 0 to 20)

a 0 – none, 1 – primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education.
b teacher, health care related, civil services (e.g. administrative or police), at home or other.





From our analysis we found that Travel Time, Mothers Education, Workday Alcohol Consumption level are the key factors and Internet too is included as a minor factor.

Now we use these factors along with G1 and G2 to predict G3 using Decision trees and SVMs

Model	Factors	Mean Prediction Score
Decision Tree	All Factors	0.663949191214
SVM	All Factors	0.607351965948
Decision Tree	Mother education, Daily Alcohol, Travel Time	0.620095938394
SVM	Mother education, Daily Alcohol, Travel Time	0.716147097747
Decision Tree	Daily Alcohol, Travel Time	0.660543144524
SVM	Daily Alcohol, Travel Time	0.759468073607
Decision Tree	Travel Time, internet	0.708723557784
SVM	Travel Time, internet	0.744006917848

Our SVM prediction score is better than previous research which had a score of 59.6% vs 75.9% in our classification model. Their DT score was better than ours. They had a score of 76.7% while our best is 70.8%.

CONCLUSION

- The reason to choose a school is not because they provide a special course but it is because student want to reduce the travel time.
- Most of the students study in Gabriel Pareira (males and females) and score better in this school compared to Mousinho de Silveria.
- The performance of children in a Math Course is dependent on different factors. First of all, we found a strong connection between the mother's job and the performance. It is best for the child when the mother works in the health sector and worst when the mother is at home. This result is counter-intuitive because one can assume that a mother who is at home has time for their children.
- In comparison to that the result that going out is negatively correlated with your performance in a Math Class is totally intuitive. Additionally, the results revealed that older children which failed once or several times have lower performance rates.
- While boys show lower performances when they grow older, girls remain relatively constant.
- Educated parents have children scoring good grades.
- Healthy females and males tend to score better grades.
- Internet access and grades scored show a positive correlation.

REFERENCES

[1] Paulo Cortez and Alice Silva, *Using Data mining to predict secondary school performance*

[2]https://archive.ics.uci.edu/ml/datasets/Student+Performance

Wikipedia: Decision trees

Wikipedia: SVM

sklearn machine learning libraries

Paulo Cortez, University of Minho, GuimarÃ£es, Portugal, http://www3.dsi.uminho.pt/pcortez