2/6/2019

Data modelling of student grades achieved in math and their relationship with various social, demographic and school related features.

Vinay Nagamangala Rame Gowda

Contents

Executive Summary

The aim of this report is to determine the relationship between various social, school related and demographic features with the final grades achieved by students in maths in a secondary Portuguese school. The data was collected from the data set obtained from UCI machine learning repository. For an easier understanding the visual representation of relationship was conducted. It was evident that more study time, extra paid classes and better family relations resulted in better pass percentage. Also student who were involved in extracurricular activities had more attendance percentage. Two data models (Decision tree classifier and K nearest neighbour) were created while taking nature of data into consideration. Although, these models had similar accuracy in predicting the results, Decision Tree Classifier is recommended because of the liberty of using certain percentage of the whole data as training set.

Introduction

The data obtained from UCI machine learning repository included various factors such as parent's education, residential area, access to internet, alcohol consumption, past failures etc. All these factors had some sort of impact on the student's final grades in maths. The relationship between these features and the results in general, various graphs were created. It was quite evident that some factors had more impact on the final results than the others. In order to predict a student's results accurately the decision tree classifier(to eliminate the non-deciding factors) and k nearest neighbour(grouping the data by common factors) models were used. The dataset was from classification category hence these models were highly successful.

Methodology

This project involves the use of Python in order to explore, prepare and model the data. After setting up the goal of predicting of correct result of students, following data preparation steps were followed in order to clean the data for modelling and for easier graphical representation –

- ❑ Data collection – It was downloaded from UCI machine learning repository
- ❑ Data exploration – The columns with most significance to the results were chosen.
- ❑ Data formatting – The columns were changed to categorical form for easier visual representation and then later to integers for machine learning.
- ❑ Data quality – The data was checked for data any anomalies, missing values or outliers.
- ❑ Data splitting – After graphical representation the data was split into training set and the testing set.

After this the modelling of data was done through two classifiers, Decision tree and K nearest neighbour. These classifiers take all the attributes into consideration and build a model to predict a result.

Results and Discussion

The data obtained from repository was in csv format. It was loaded onto Python and the columns that were not considered important were dropped . The data was checked for any missing any values and anomalies. The final dataset after processing had the following attributes.

sex	address	Parent status	Mother edu	Father edu	study time	past failures	school support	family support	extra paid classes	activities	wants higher	internet access	fam relation qual	workday alcohol	current health	abser
sex	address	Pstatus	Medu	Fedu	studytime	failures	schoolsup	famsup	paid	activities	higher	internet	famrel	Dalc	health	abser
Female	Urban	Separated	4	4	2	0	yes	no	no	no	yes	no	4	1	3	
Female	Urban	Together	1	1	2	0	no	yes	no	no	yes	yes	5	1	3	
Female	Urban	Together	1	1	2	3	yes	no	yes	no	yes	yes	4	2	3	
Female	Urban	Together	4	2	3	0	no	yes	yes	yes	yes	yes	3	1	5	

Figure 1 Headings of dataset

	sex	address	Parent status	Mother edu	Father edu	study time	\
count	395	395	395	395	395	395	
unique	2	2	2	5	5	4	
top	Female	Urban	Together	Higher	5th to 9th	2	
freq	208	307	354	131	115	198	

	past failures	school support	family support	extra paid classes	\
count	395	395	395	395	
unique	4	2	2	2	
top	0	no	yes	no	
freq	312	344	242	214	

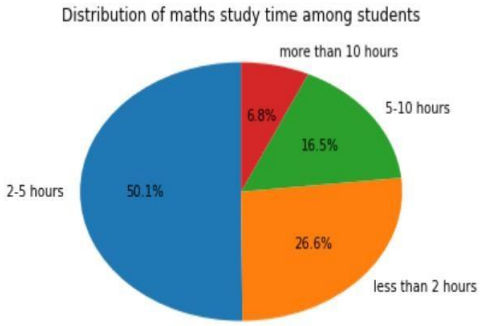
	activities	wants higher	internet access	fam relation qual	\
count	395	395	395	395	
unique	2	2	2	5	
top	yes	yes	yes	4	
freq	201	375	329	195	

	workday alcohol	current health	absences	Final Grades
count	395	395	395	395
unique	5	5	34	18
top	1	5	0	10
freq	276	146	115	56

Figure 2 Attribute descriptions of dataset

The above figure was obtained when the description of dataset was checked.

To understand the data more clearly major attributes were represented in form of graphs.



This pie chart shows the distribution of study time allocated by students towards maths. It is clear that most of the students(more than 50%) allocate 2-5 hours weekly in studying and the least percentage study more than 10 hours.

Figure 3 Time allocated to studies per week

If we compare the education status of mothers, most of the mothers of students have had higher education completed. Only a small percentage have had no education.

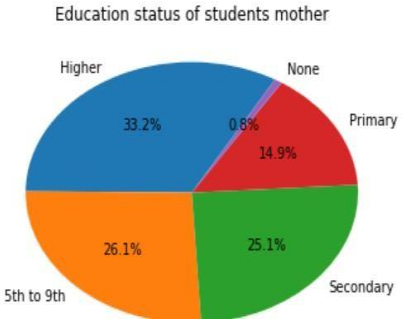
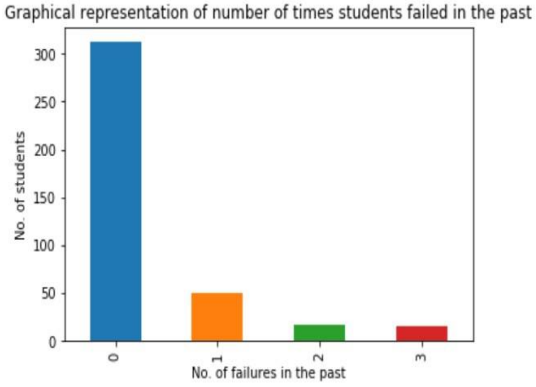


Figure 4 Highest education attained by students' mothers



This graph shows the number of students who have failed either once, twice, thrice or never. Most of the students (around 300) have never failed in the past.

Figure 5 number of students and their fail rate

Most of the students have their education supported by their family. The percentage of students who don't have their education supported by family is around 38.7%

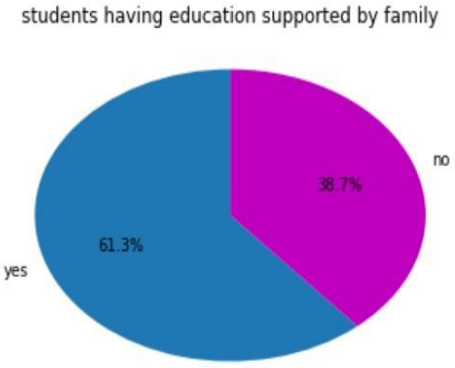


Figure 6 Percentage of students having education supported by their family

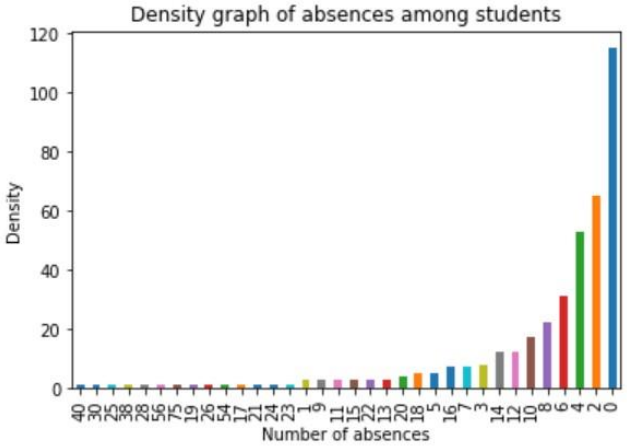


Figure 7 Absentee frequency of students

It is clear from this graph that most of the students have never been absent from school and approximately 60 students have been absent twice.

If we compare the current health status, than more than a third of students have excellent health condition. The students with bad and very bad health conditions are around 12 % each. The health condition can be related to absenteeism and can hence affect the grades.

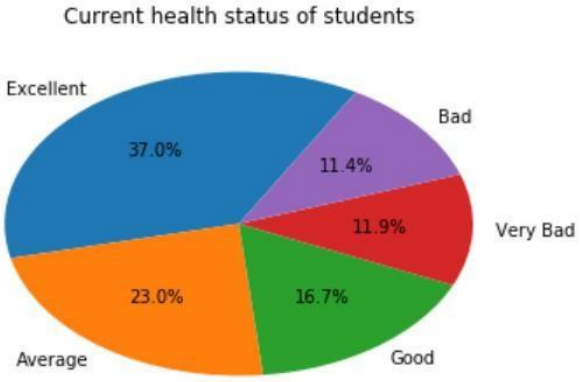


Figure 8 Percentage of current health status of students

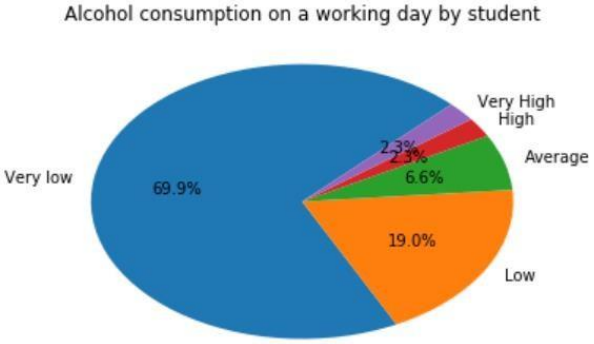


Figure 9 alcohol consumption on a weekday

High alcohol consumption can lead to impairment of memory power. Fortunately most of the students in the school have a very low intake of alcohol on the weekdays.

Almost 50% of students have good relationship with their family. Only 6.6% students have classified their relationship with family bad and very bad. Bad relationship can cause unnecessary stress to students and can affect their grades.

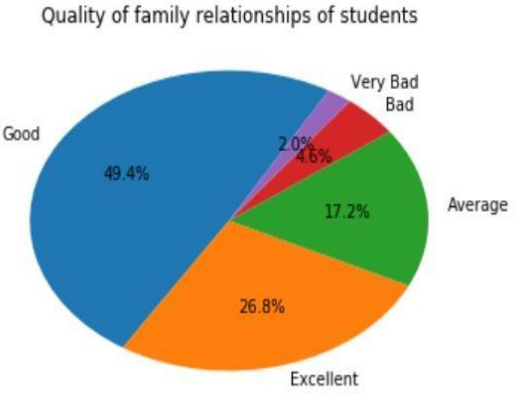
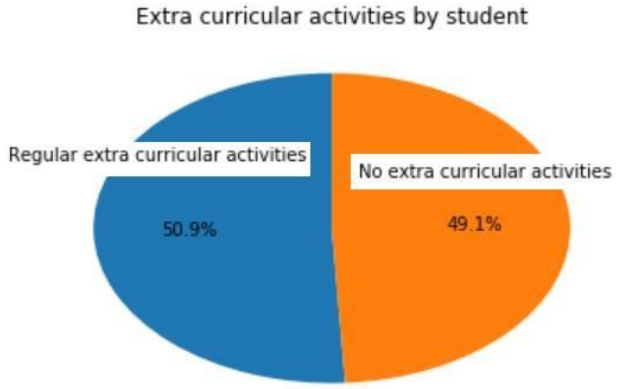


Figure 10 Family relationship quality of students



Percentage of students doing extracurricular activities is very similar to the ones who don't. Performing physical activities can stimulate the brain and hence affect the grades.

Figure 11 Extra curricular activities participation

This graph shows that number of students achieving 10 marks in the result is the highest. There are around 38 students in the data who have failed to score any marks. This can be further elaborated by the line graph which is plotted below.

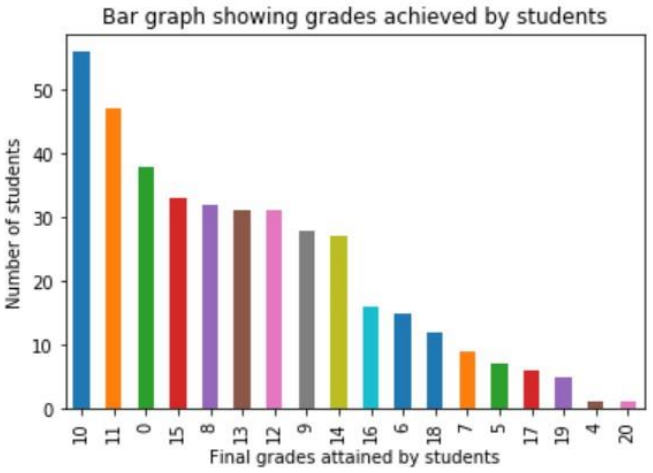


Figure 12 Distribution of grades among students.

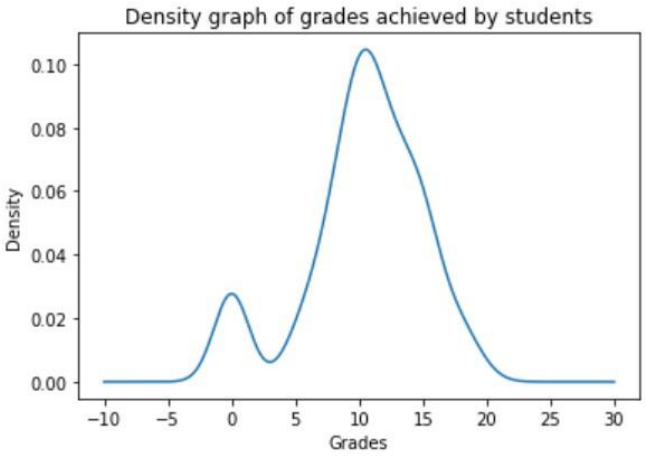
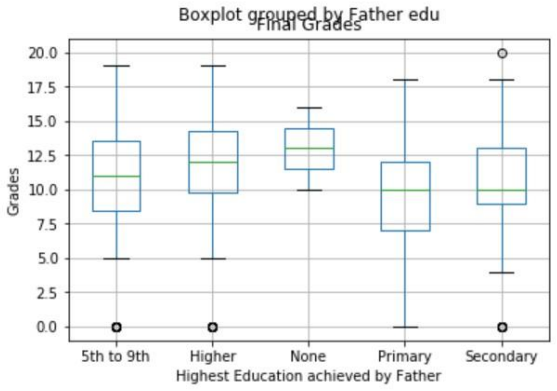


Figure 13 Line plot showing distribution of grades among students

If we take the relationship of these attributes into consideration the picture becomes clearer and we can further pin point the factor which most affect the grades among students.



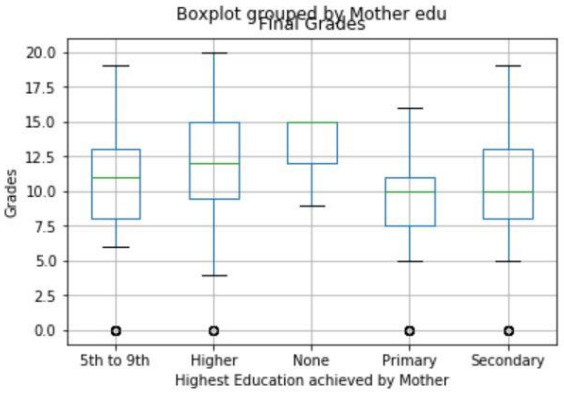


Figure 14comparison of grades with Father's education

Figure 15 comparison of grades with mother's education

Hypothesis – the children of parents' with high education should have better grades.

The above two box plots show that the students' whose parents have no education tend to have highest grades. But this statement can be discarded as the percentage of parents with no education was very small. If we see the second highest grades, they come from children whose parents have achieved at least high school education.

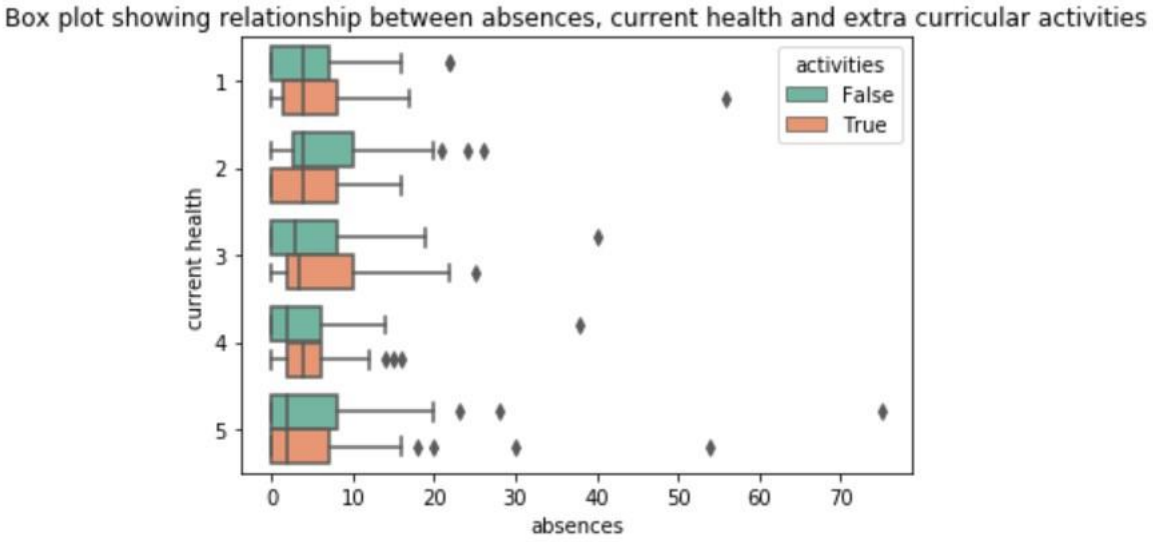


Figure 16 Relationship between absents, current health and extracurricular activities

Hypothesis – the students who perform extracurricular activities will have good health and will have less absents

This graph is a little unclear about this relationship. This is because there are some students who have abnormally high number of absents from the class. If we ignore that, the students who are physically active are less absents than the other students of same health status.

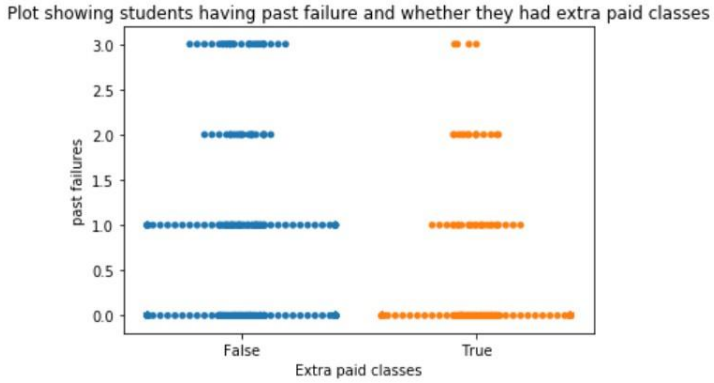


Figure 17 comparison of past failures with extra paid classes

Hypothesis – student with extra paid classes should have less failure rate.

Here the number of students who have no past failures is almost similar in both cases. But as go towards the student who had 1 or more than 1 failure tend to have no extra paid classes. Only a few students with extra paid classes have 3 failures which is much more than who haven't had any.

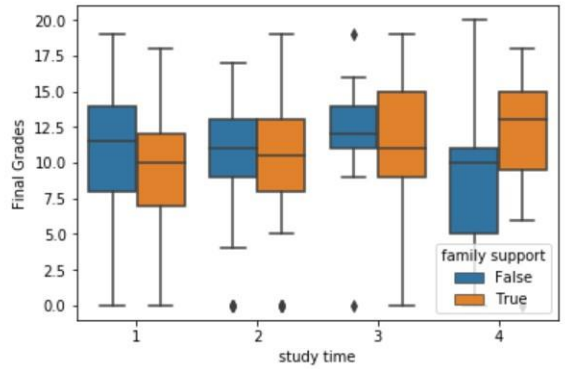


Figure 18 Relationship between final grades, study time and family supported education

Hypothesis – more study time and family supported education should result in better grades.

This graph clearly shows the highest average of grades of students in the last category (4, who study from 5-10 hours). There is an interesting trend among students in all the categories though. It looks like the students who don't get study supported by their family are getting higher grades than other students who study for similar time.

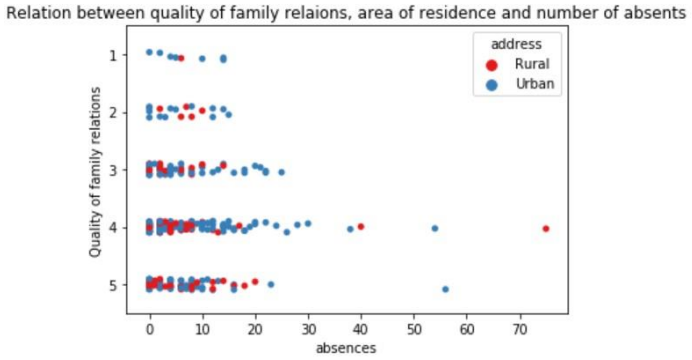


Figure 19 Relationship between Quality of family relation, address and number of absents.

Hypothesis – good quality of family relations should result in less absents

It looks like students who have bad family relationship looks reside in urban area more. They do have less 0 absents than any other group. But due to small percentage of them the results could not be conclusive.

Data modelling

The models that we used in this discussion were Decision Tree Classifier and K nearest neighbours. Here we tried to predict if the student will achieve equal to or more than 10 marks in their final grades or less than 10.

For the Decision Tree Classifier firstly we used 50% split, which means that 50% of data was used for testing and 50% was used for training. The predicted accuracy of this split was almost 99.5%. In regard to the confusion matrix, false positive value is 35 and false negative value is 39. The f1 score's weight average is 0.63 and the precision is 0.64, with pass result more accurately predicted. The weighted average of recall is 0.65.

For the 40% split, i.e. 40% data for testing and 60% for training, the predicted score was 99.57%. In the confusion matrix, false positive values are 29 and false negative values are 32. The f1 score's weight average is 0.62 and the precision is 0.62 as well, with pass result more accurately predicted. The weighted average of recall in this case is 0.62.

Similarly obtained from 20% split. Here the predicted accuracy was 99.27. In regard to the confusion matrix, false positives are 13 and false negatives are 16. The f1 score's weight average is 0.64 and the precision is 0.64, with pass result more accurately predicted. The weighted average of recall in this case is 0.63.

In K nearest neighbour the number of k was chosen by taking the square root of number of data sets in the testing split. It comes out to be number 9. We chose p's value to be 2 because of the two parameters that will come out as result. The accuracy of this model comes out to be 71.5%.

In the case of confusion matrix, the false positives are 18 and the false negatives are 3. The f1 score's weight average is 0.69 and the precision is 0.73, with pass result more accurately predicted, but only by a slight margin

Conclusion

From the above results and discussion we can say that most of the hypothesis that we predicted for our dataset turned out to be true. There were a few surprising results for example the residence of most families with bad relationship turned out to be urban areas.

The classification model that was more successful in our case was the kNN i.e. k nearest neighbour model, which had the accuracy of almost 73.5%. It even predicted the passes and the fails with equal precision. In case of decision tree classifier, the

most successful split was 50% split. The only problem with this model was the inability of predicting both pass and fails with equal accuracy.

Precision and recall are both very important in case of predicting the accuracy of data models. Precision gives us the proportion of results which are most relevant while recall gives us the proportion of total relevant results correctly classified by our model (Saxena, 2018). F1 score on the other hand is the mean of precision and recall. It was highest in the kNN model as well.

Confusion matric is a summary of prediction results on a classification problem (Brownlee, 2016). Error rate can be defined from confusion matrix and can be calculated from (1-correct predictions/total predictions)*100. The error rate was lower in the kNN model again.

kNN model has its drawbacks as well. It is often referred to as lazy learner (Genesis, 2018). As it doesn't learn anything from the dataset. Choosing of optimal number of neighbours is very necessary to get the accurate result as well.

It is safe to say that a combination of both models and even more if necessary, to accurately predicting the desire result.

Bibliography

Brownlee, J. (2016, Nov 18). *What is a Confusion Matrix*. Retrieved from Machine learning mastery: https://machinelearningmastery.com/confusion-matrix-machine-learning/

Genesis. (2018, Sept 25). *Pros and Cons of kNN*. Retrieved from From The Genesis: https://www.fromthegenesis.com/pros-and-cons-of-k-nearest-neighbors/

Saxena, S. (2018, 05 12). *Precision vs Recall*. Retrieved from Towards Data Science: https://towardsdatascience.com/precision-vs-recall-386cf9f89488