Analysing Factors for Students Performance Using Machine Learning

Samyak Jain, Subham Singh, Mayank Kestwal ,Manikant Roy
School of Computer Science & Engineering
Lovely Professional University, Phagwara Punjab, India

ABSTRACT

Student's performance is one of the major academic outcome in teaching and learning process. There are various factor which can affect the students' performance such as socio economic background, medium of instruction, habits etc. to name a few. This study presents the most prominent factor in students' academic success by applying machine learning approaches.
Keywords: Student Performance, Learning Analytics,

INTRODUCTION

In this modern era, education is highly influenced by technology and in same way teaching is also oriented around technology. MOOCs and other online teaching platform has changed the way of education as perceived earlier. And because of availability of this much technology it has led to storage, and maintenance of data about each student easy. Since data is available then it's obvious to use it for benefits of students and with help of this data we can draw insights, trends and patterns to understand performance and their behaviour towards each subject and this would result in better teaching and learning process and also would help to focus on specific factors responsible for such patterns and trends. So, some of the questions which can be asked from the data available. Which factors effects students' performance more towards their final grade?
Why some students are under performing and in which subjects their performance is good? Etc. a lot of questions can be asked and answered with the help of this data only to make education system more and more better. To find the answers of these questions we are going to visualize our data and use modern computing technologies such as Machine Learning. Better understanding of opaque insights would result in improvement of academic process decision and tackle the any academic related problem within the time.

The analysis of this case study is based on data available from UCI Machine Learning Repository dataset, "Student Performance Data Set" from https://archive.ics.uci.edu/ml/datasets/student+performance

2. Data Set Information

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).

2.1. Attribute Information

Data Set Characteristics:	Multivariate	**Number of Instances:**	649
Attribute Characteristics:	Integer	**Number of Attributes:**	33
Associated Tasks:	Classification, Regression	**Missing Values?**	N/A

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 (4th grade), 2 â€“ 5th to 9th grade, 3 â€“ secondary education or 4 â€“ higher education)
8 Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€“ 5th to 9th grade, 3 â€“ secondary education or 4 â€“ higher education)
9 Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
10 Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
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)
31 G2 - second period grade (numeric: from 0 to 20)
32 G3 - final grade (numeric: from 0 to 20, output target)

2. 2 Data Pre-processing:

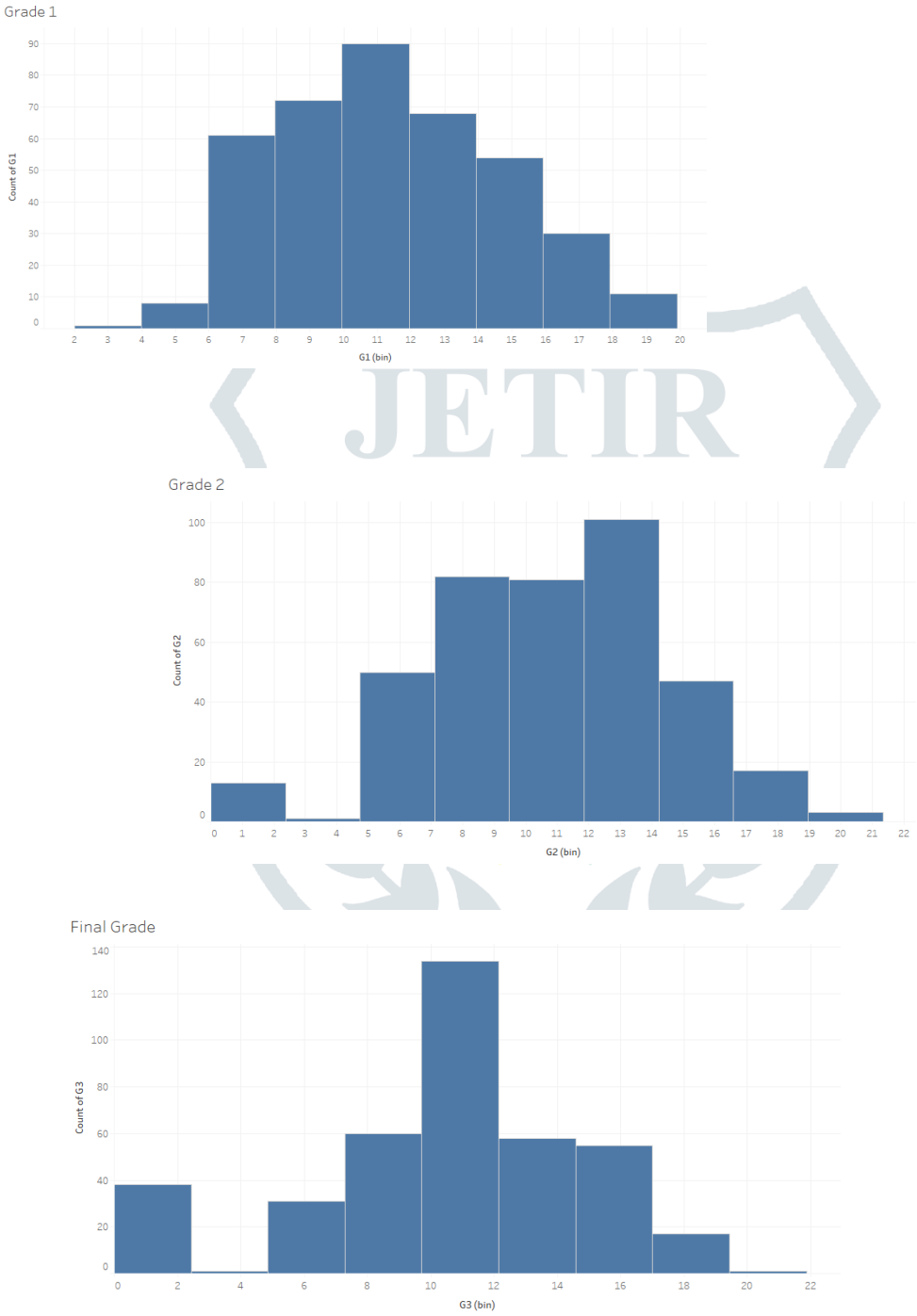
Original data may contain noise, there may be some attributes which are not going to contribute at all in Machine learning model, or they would not show any results while visualization to draw some insights, there may be missing values or range of values of data is too high also we can categorical features and mismatch in data types of same column. So, to handle all of these situations we need to pre process our data set prior to feeding it to machine learning model.
Some of the steps may involve are:

- Handling NULL or missing values.
- Normalization of real valued columns having high variance.
- One-Hot encoding or Label Encoding.
- Handling mismatch in Data types.

2.3. Exploratory Data Analysis

The purpose of EDA is to use summary statistics and visualizations to better understand data, and find clues about the tendencies of the data, its quality and to formulate assumptions and the hypothesis of our analysis.

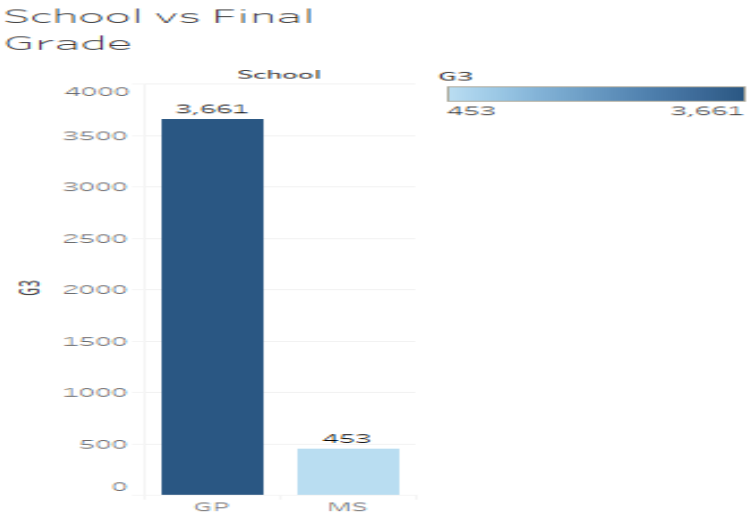
2.3.1. Grade Distribution



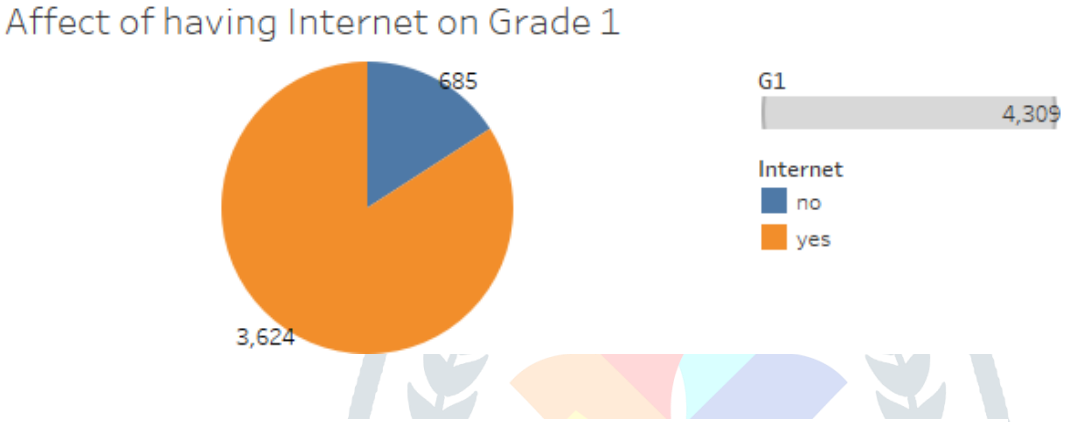
This shows us the distribution of all three grades are G1, G2 and G3. The distribution data is neither left skewed nor right skewed, it is normally distributed over the range.

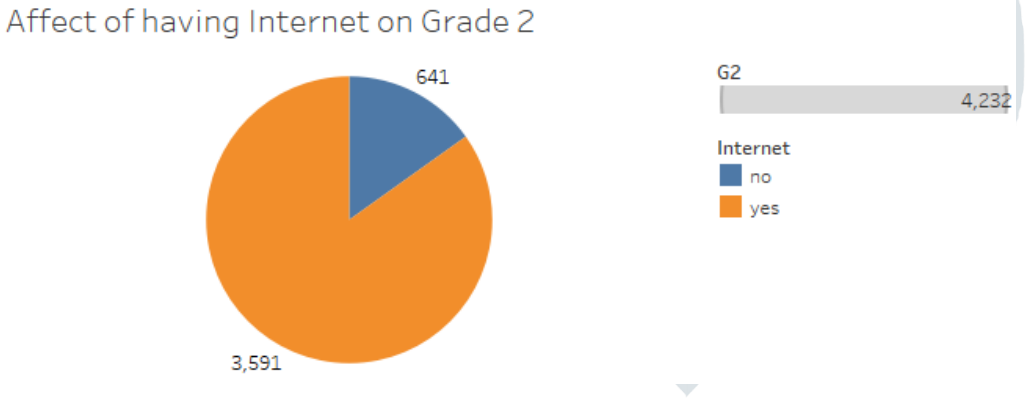
2.3.2. School vs Final Grade

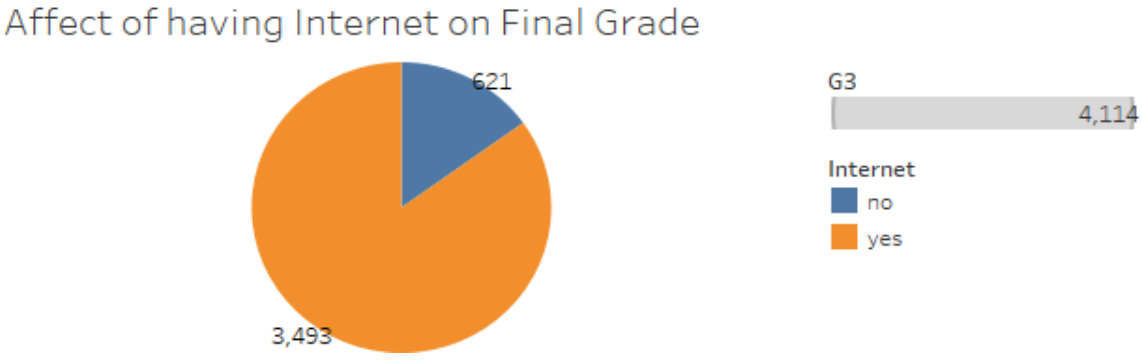
The school contains two categories one is ‘Gabriel Pereira’ and the other one is ‘Mousinho da Silveira’. The bar chart depicts that the Gabriel Pereira’s students performance in Mathematics is better than the students of Mousinho da Silveira.



2.3.3. Internet impact on Grades

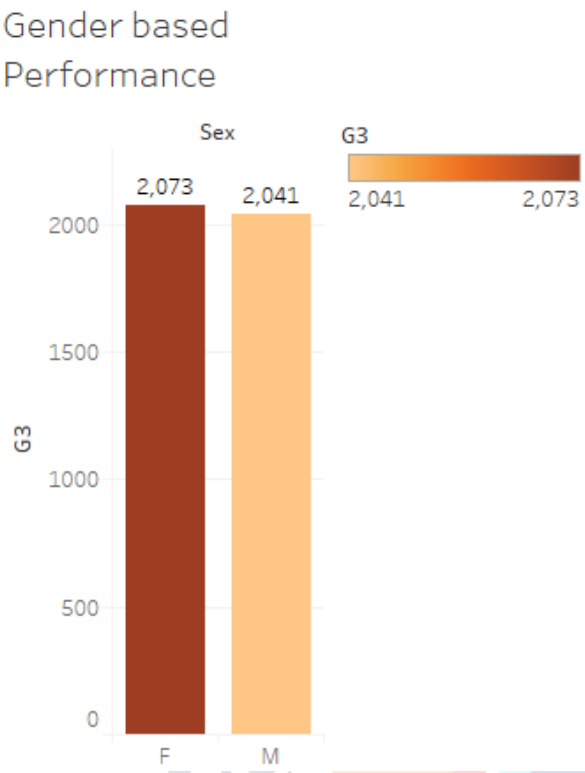






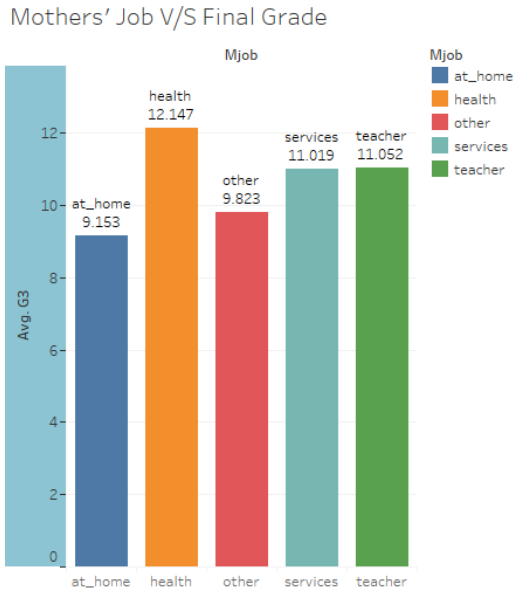
The pie chart clearly shows that the students who have internet scores more than those who don't have. The difference is visible, it is huge

2.3.4. Gender based performance



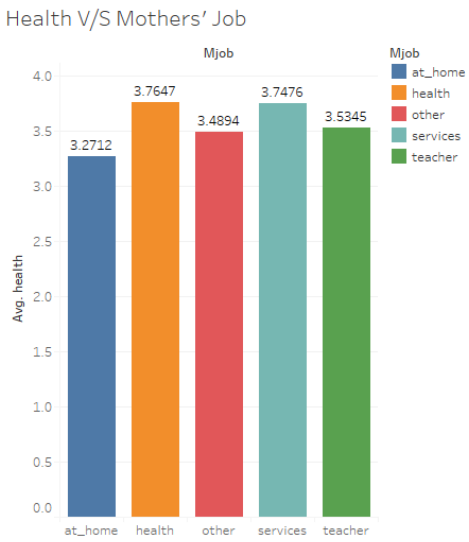
There is marginal difference between the scores of female and male students, both are performing well. Though Female students are doing a little better.

2.3.5. Mothers' Job V/S Final Grade



Mother is the best teacher in this world, so here we are comparing the occupation of mothers and their children performance. It depicts that the mothers who are working in health sector their child is performing better than the rest.

2.3.6. Health V/s Mothers' Job



Here also we can observe that total average health of those students are good whose mother is in Health sector as compared to health of other students. So we can conclude from above two graphs that healthy students are having good final average as compared to others student.

3. Grade Prediction Using Different Machine Learning Models.

3.1 Linear Regression:

```
from sklearn.linear_model import LinearRegression
regressor = LinearRegression()
regressor.fit(X_train.values,Y_train.values)
```

```
LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None, normalize=False)
```

3.1.1 Train and Test Score

```
regressor.score(X_train.values,Y_train.values)
```

```
0.8528204785553677
```

```
regressor.score(X_test.values, Y_test.values)
```

```
0.7953787140783359
```

3.2 Decision Tree

```
from sklearn.tree import DecisionTreeRegressor
regressor = DecisionTreeRegressor(random_state=0, max_depth=5, max_leaf_nodes=13)
regressor.fit(X_train.values,Y_train)
```

```
DecisionTreeRegressor(criterion='mse', max_depth=5, max_features=None,
                      max_leaf_nodes=13, min_impurity_decrease=0.0,
                      min_impurity_split=None, min_samples_leaf=1,
                      min_samples_split=2, min_weight_fraction_leaf=0.0,
                      presort=False, random_state=0, splitter='best')
```

3.2.1 Train and Test Score

```
regressor.score(X_train.values,Y_train)
```

```
0.9240404605421739
```

```
regressor.score(X_test.values,Y_test)
```

```
0.8490569061169286
```

3.3 Random Forest

```
from sklearn.ensemble import RandomForestRegressor
rf = RandomForestRegressor(n_estimators=30, min_samples_split=10)
rf.fit(X_train.values,Y_train)
```

```
RandomForestRegressor(bootstrap=True, criterion='mse', max_depth=None,
                      max_features='auto', max_leaf_nodes=None,
                      min_impurity_decrease=0.0, min_impurity_split=None,
                      min_samples_leaf=1, min_samples_split=10,
                      min_weight_fraction_leaf=0.0, n_estimators=30,
                      n_jobs=None, oob_score=False, random_state=None,
                      verbose=0, warm_start=False)
```

3.3.1 Train and Test Score

```
rf.score(X_train.values,Y_train)
```

```
0.943092532218054
```

```
rf.score(X_test.values,Y_test)
```

```
0.8695617842299628
```

4. Conclusion

From above we conclude that healthy students with their mother working in health sector are having more average marks than any other students. Also we can predict their final grade using Random Forest with an accuracy of 86.9% so that we can have prior information of students regarding their academics.

References

[1] P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.