ML PROJECT REPORT

Problem Statement:
Predicting grade of students for Maths and Portuguese subjects based on data collected from two portugues schools.

Dataset:-

Source:- https://archive.ics.uci.edu/ml/datasets/student+performance

Description of Dataset:-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).

Attribute Information:
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)
-
- 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)

Data Preprocessing:

- Both datasets are processed to check
    1. Null Values
    2. Duplicates
        &
    3. Invalid values

but fortunately there are no such irregularities in the dataset,implying that data is already clean and processed.

Since both datasets have same set of attributes and has similar kind of data ,the both datasets are merged vertically ,so as to make the dataset effective and increase the dataset,this process supposedly key aspect of data preprocessing.

Modification of Dataset:

A new column called ‘FinalGrade’ is asserted which reflects ‘Grade3’ and a broader level view of ‘Grade3’.

The column is inferred broadly as five categories under given conditions as follows:

- 'Excellent'[(data.G3 >= 18) & (data.G3 <= 20)]
- 'Good' [(data.G3 >= 15) & (data.G3 <= 17)]
- 'Satisfactory' [(data.G3 >= 11) & (data.G3 <= 14)]
- 'Poor' [(data.G3 >= 6) & (data.G3 <= 10)]
- 'Failure' [(data.G3 >= 0) & (data.G3 <= 5)]

Encoding:

There several attributes which are non-numeric and categorical ,so they must be encoded,because model cannot deal with non-numeric attributes. But as observed,the dataset has only few categories(max of 5) for every column,so best assumption would be one-hot-encoding which was our primary choice as categories are extracted as features which intensifies the effect of feature.

NOTE:We also tried Label encoder ,but results were obvious as expected,they weren’t as accurate as one hot encoding.So we continued with one-hot-encoding.

Exploratory Data Analysis(EDA):

To study and observe the behaviour of data,attributes ,relationship between attributes and target variable are graphically visualized ,so that pattern of data is keenly studied and to explore the dependency and weightage of attribute so as to extract reliable features to develop a reliant model with robust features.

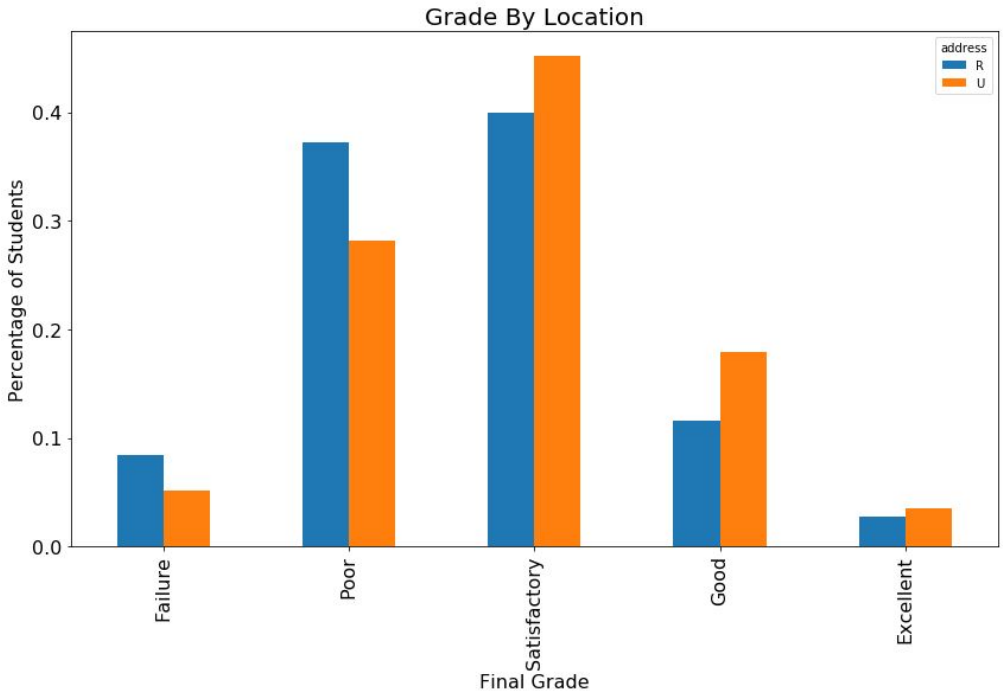
The primary plots that are visualized:

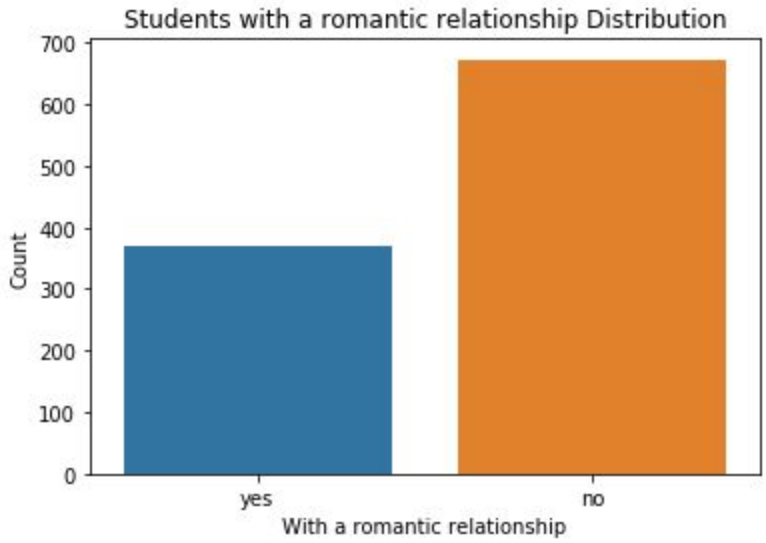
- Individual attributes plot against its frequency [to observe the data]

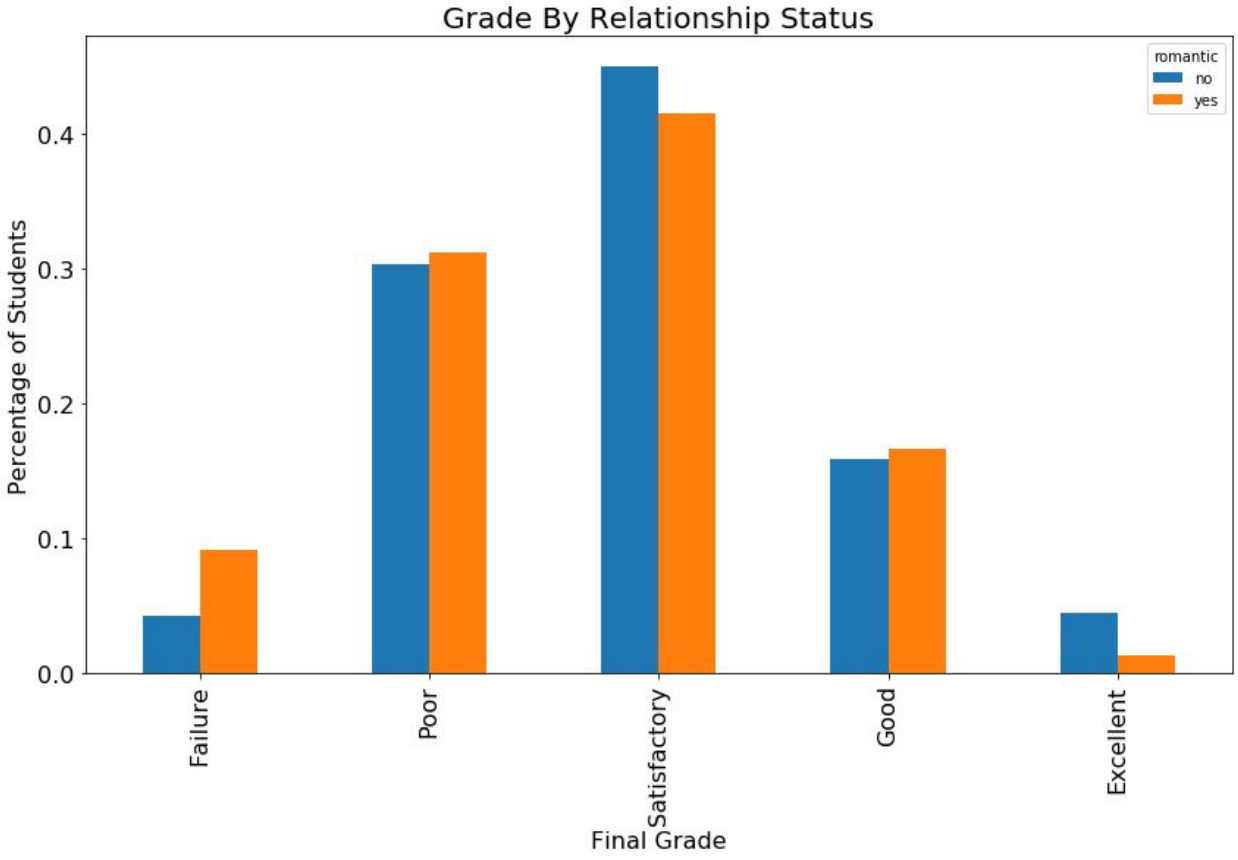
- Attributes against target variable(‘FinalGrade’)[to study the dependency and weightage]

Plot of Individual attributes and its frequency:

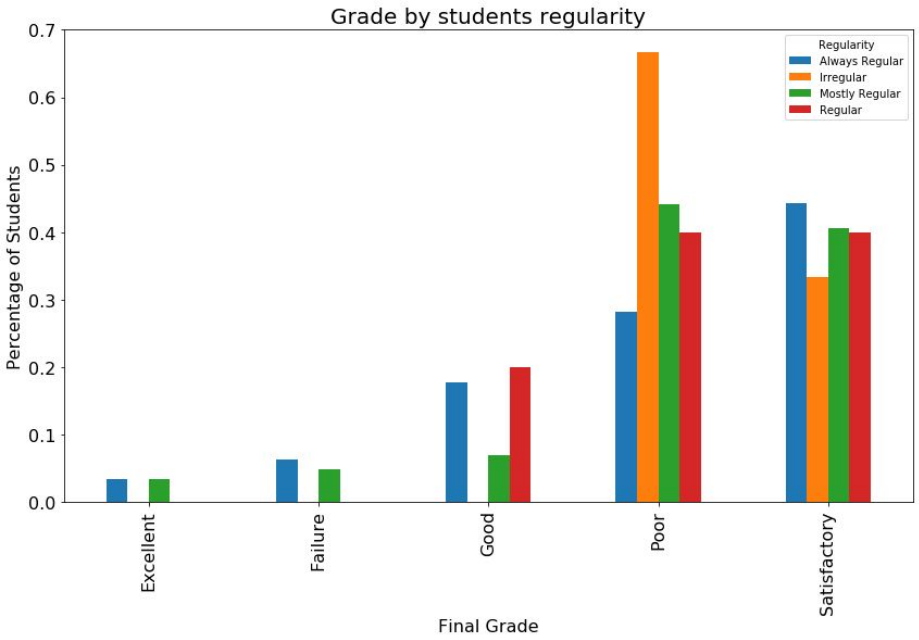
- We can infer from the below two graphs that most of students who live in urban area get more grades.



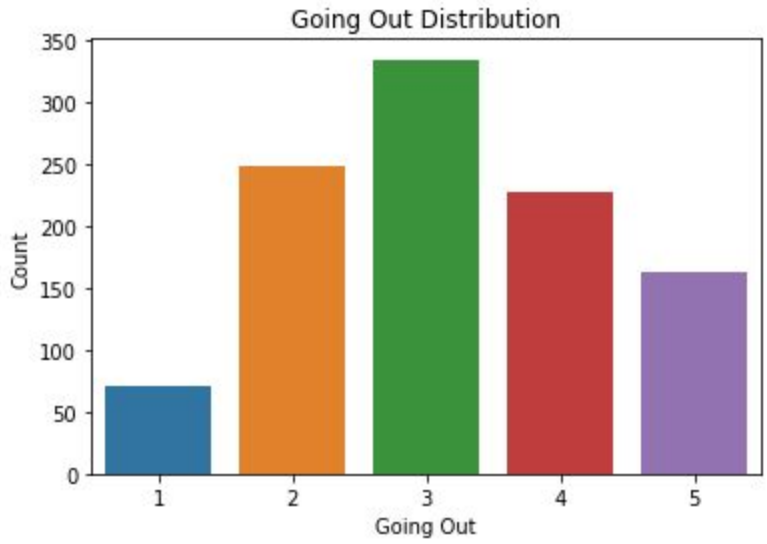


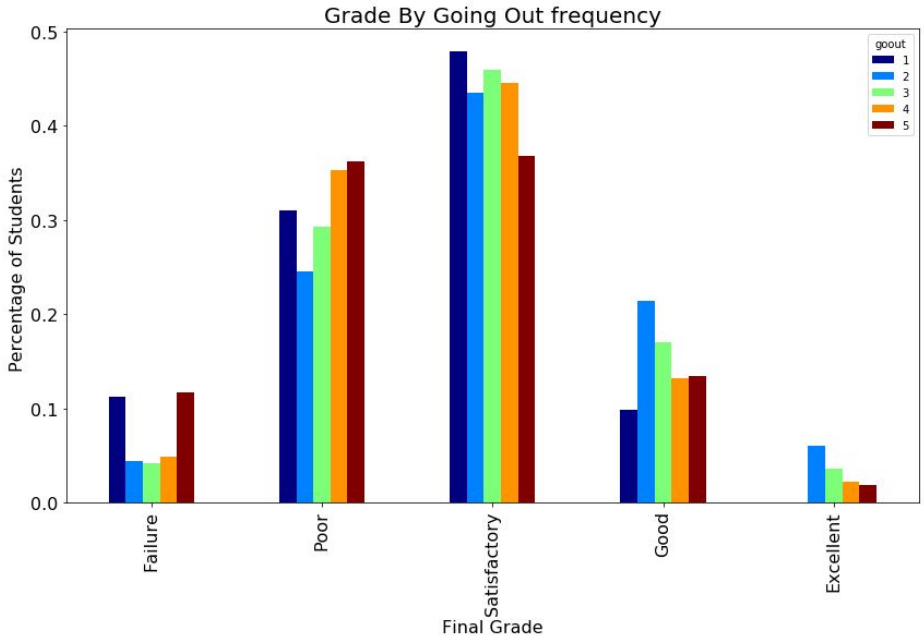


- From above two graphs we can infer that students who are not in relation get good grade.

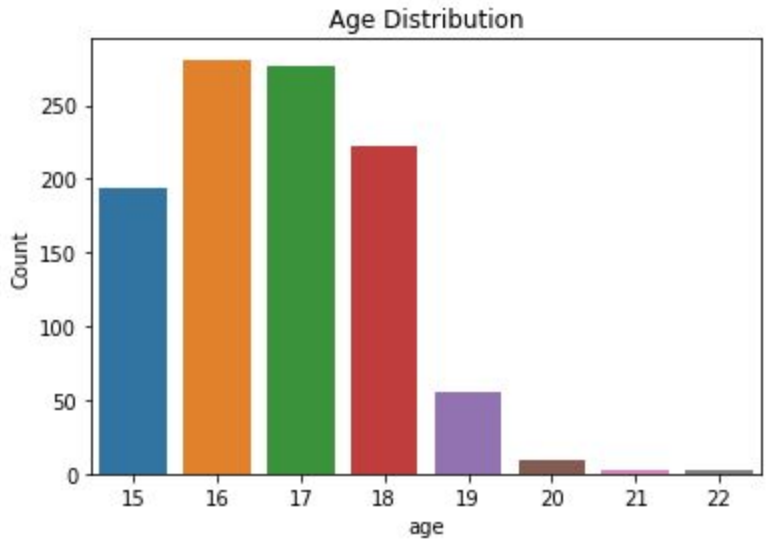


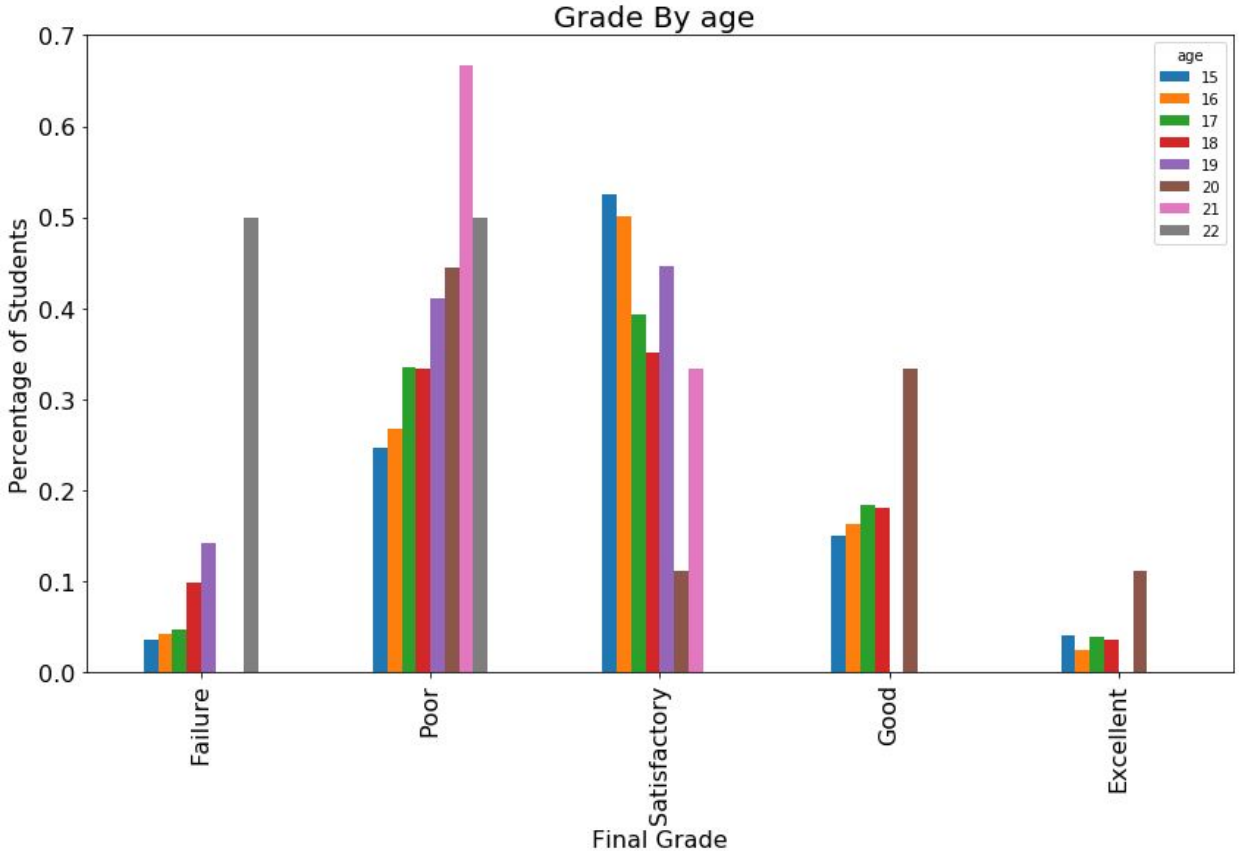
- Students who are mostly irregular endup with lesser grades than other students.



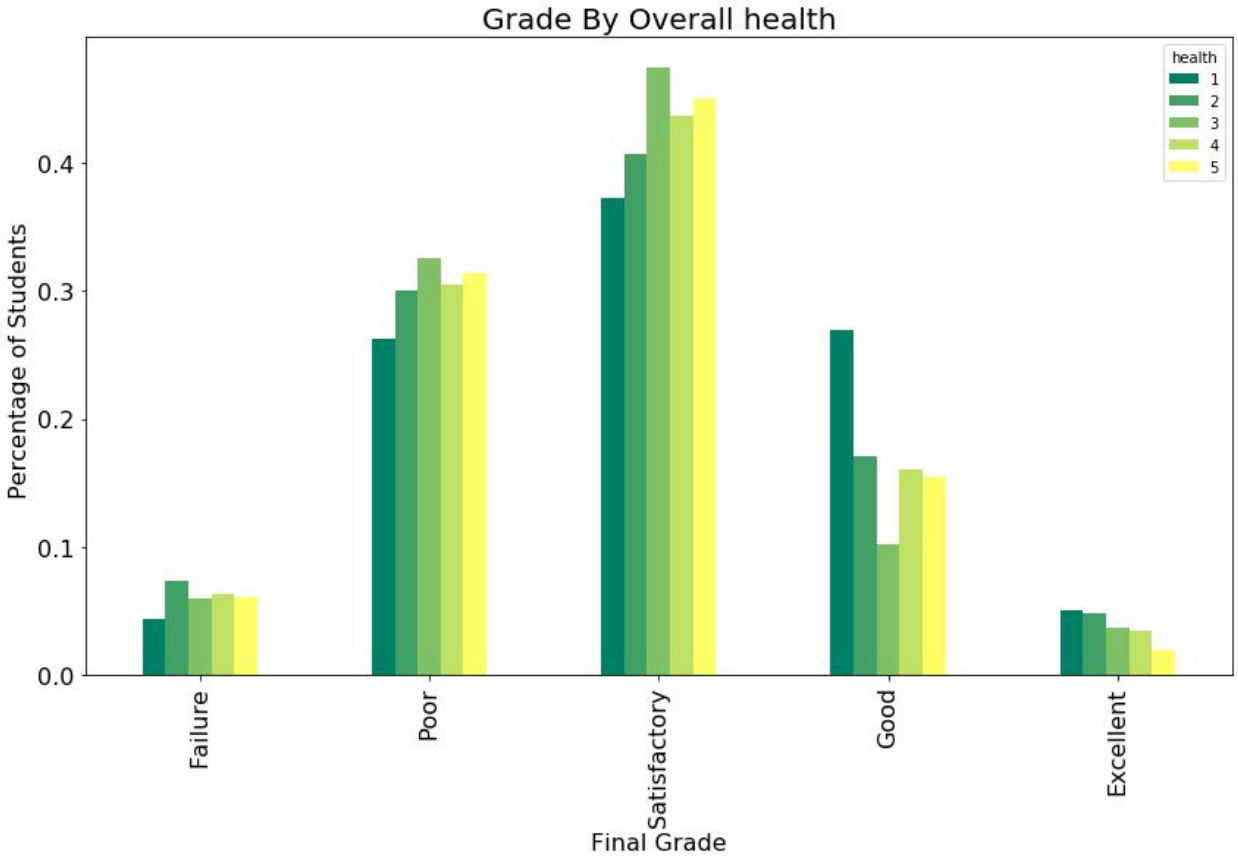


- Students need to decrease their frequency of going out to get good grades.

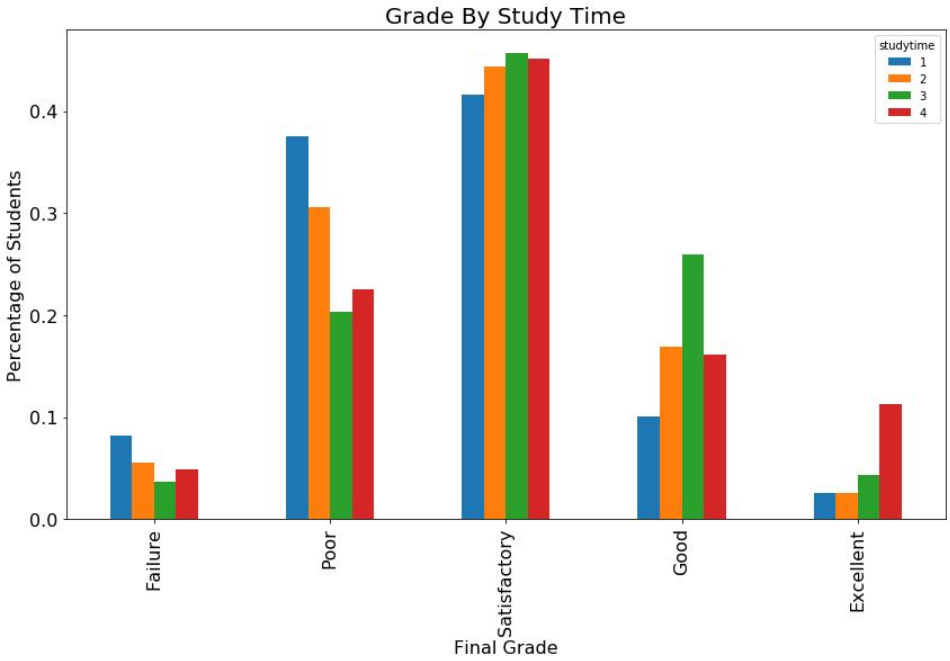
- 

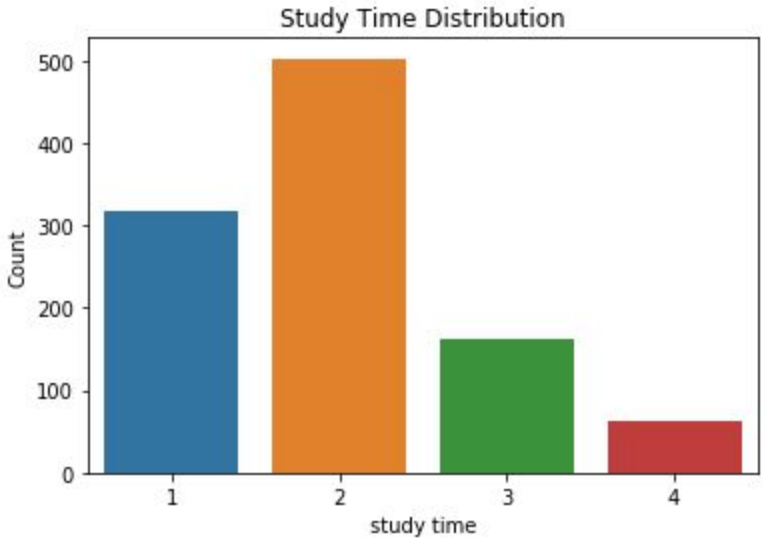


- Students who are between 15 to 20 are getting good grades.

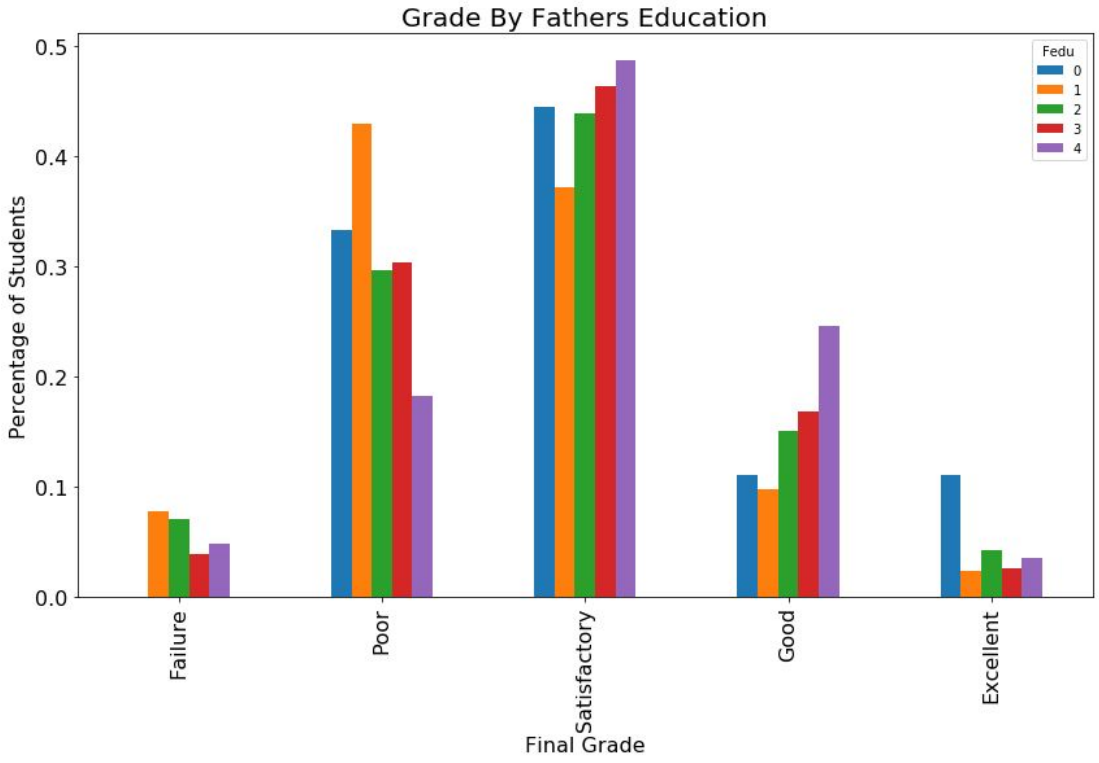


- Students with good health are usually better are studies.

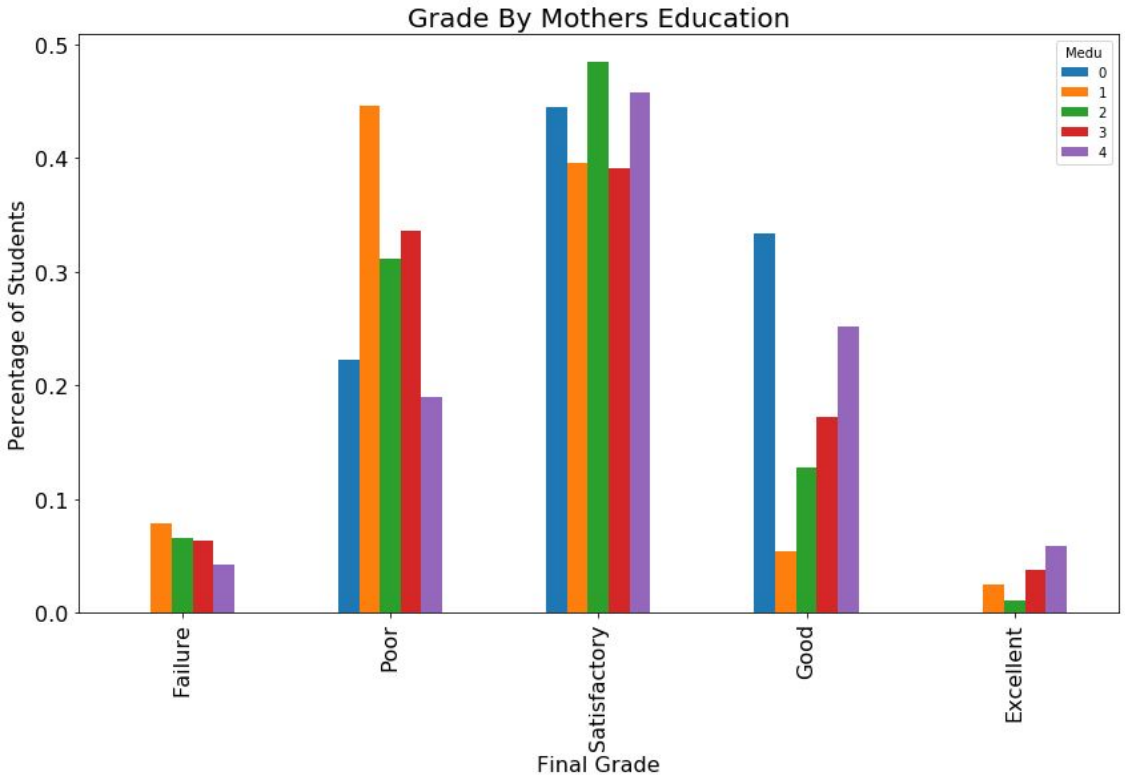




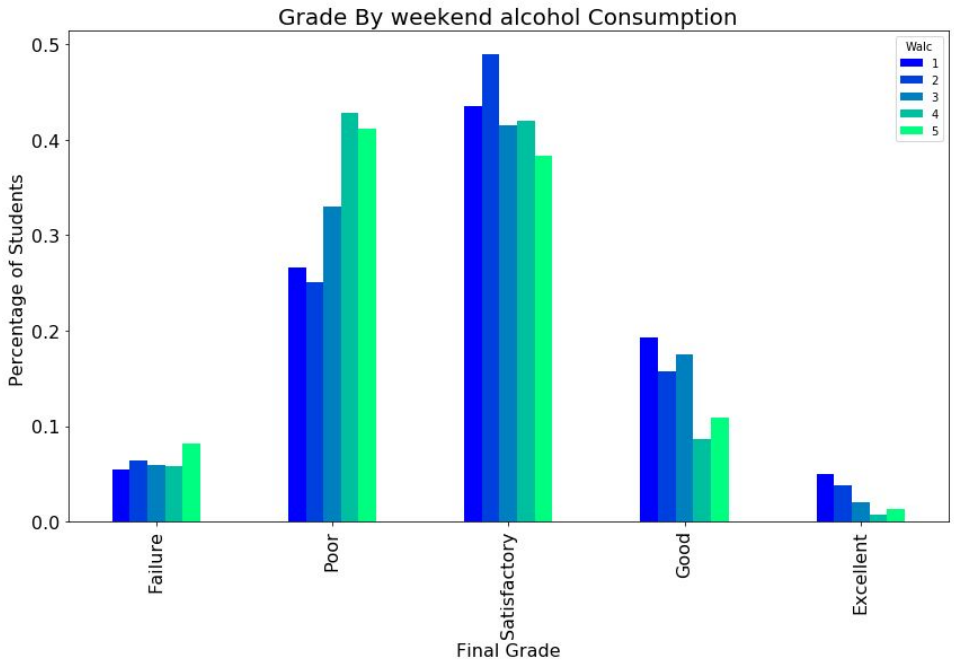
- Most of the students study 2 to 5 hours per week and who study more than 4 hours per week are good at studies.

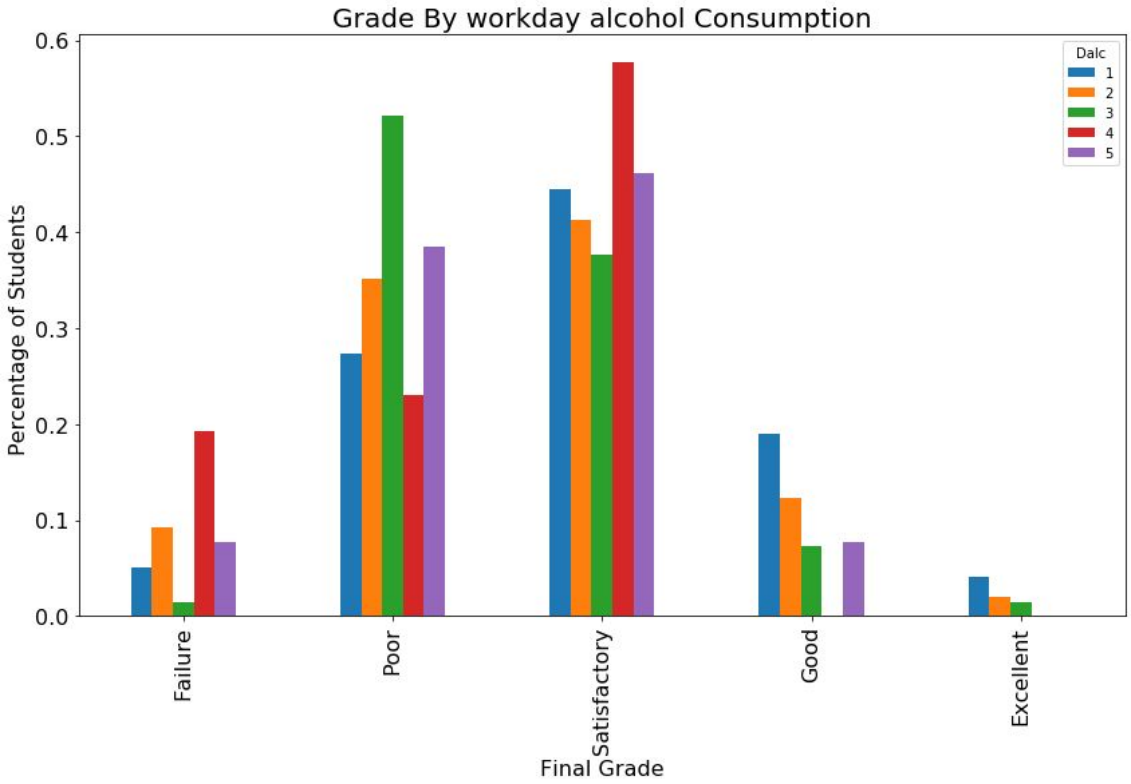


- Students who has well educated father are usually good at studies.

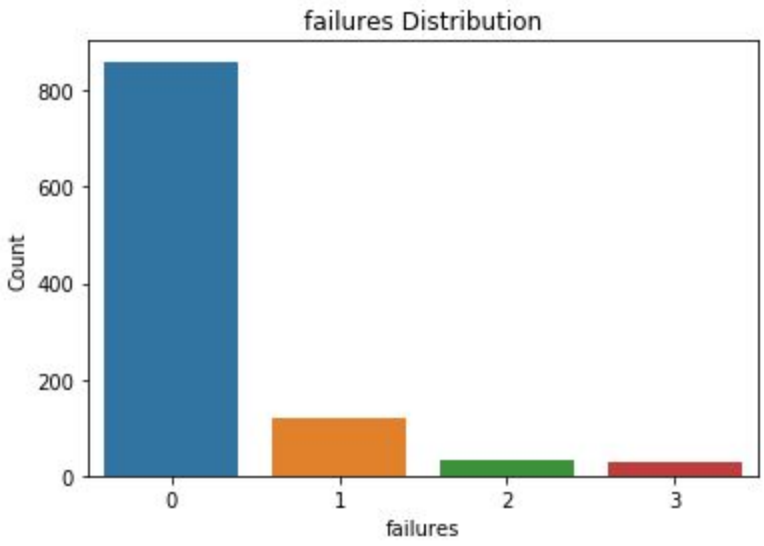


- Students who have well educated mother performs well at exams.

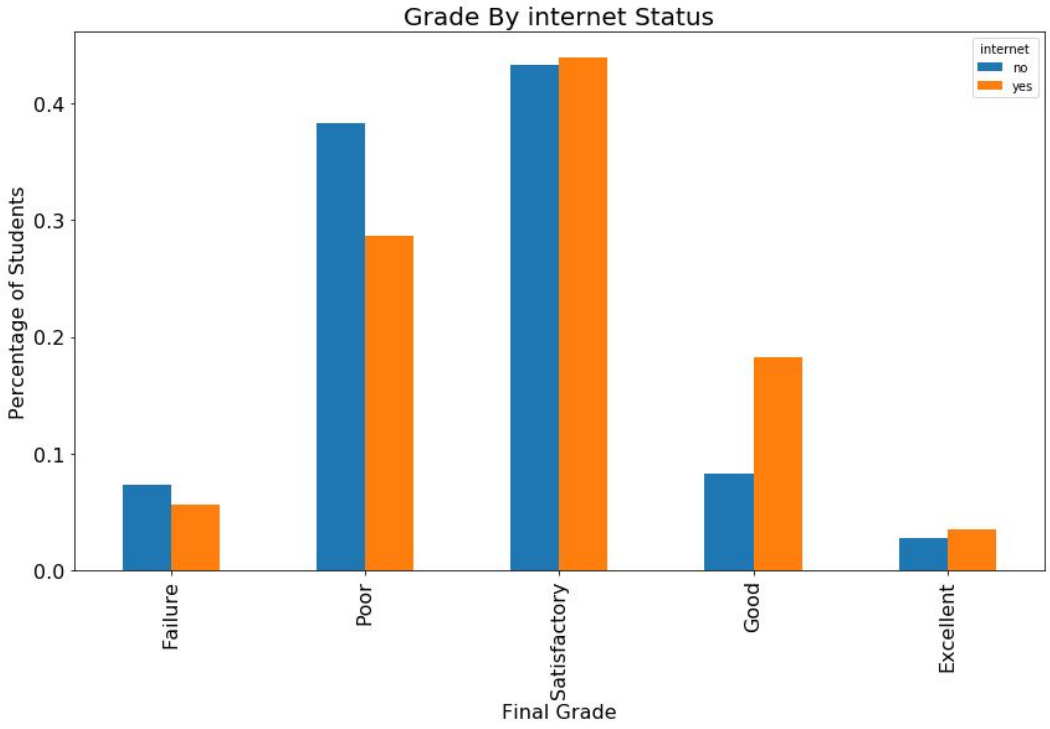


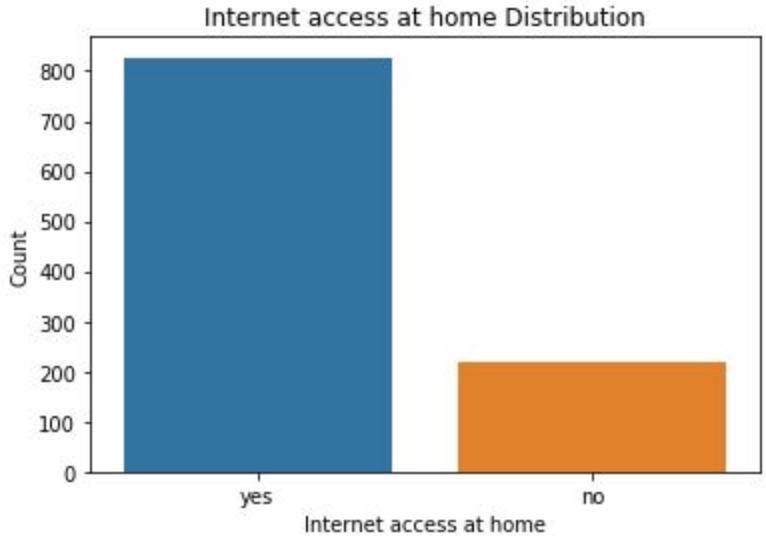


- Students consuming less alcohol during the weekend get good grades.

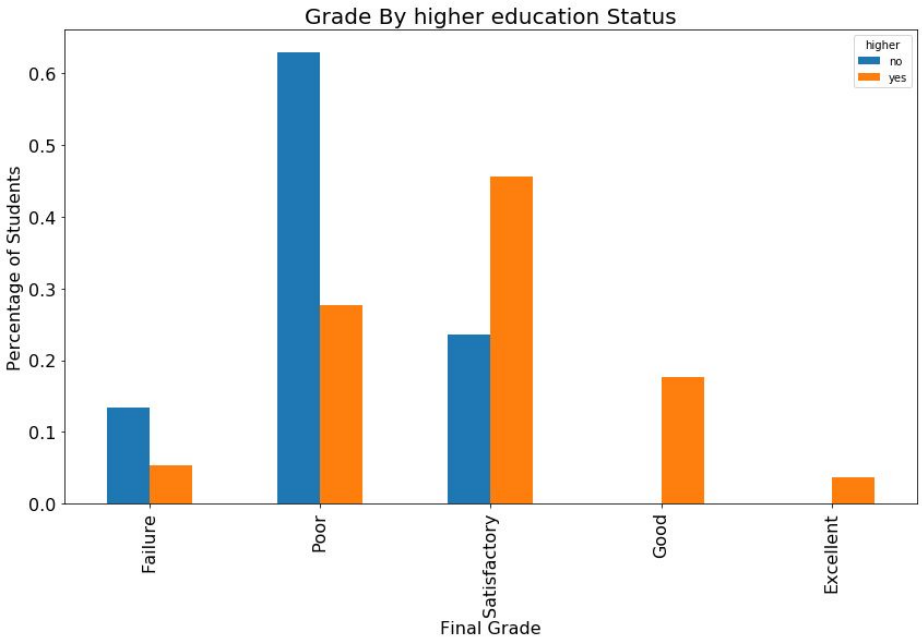


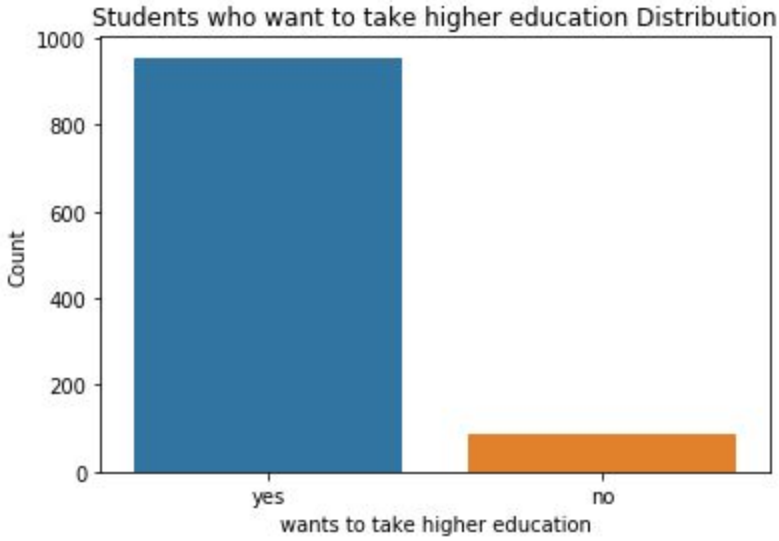
- Students who failed in previous exams are very few in number.



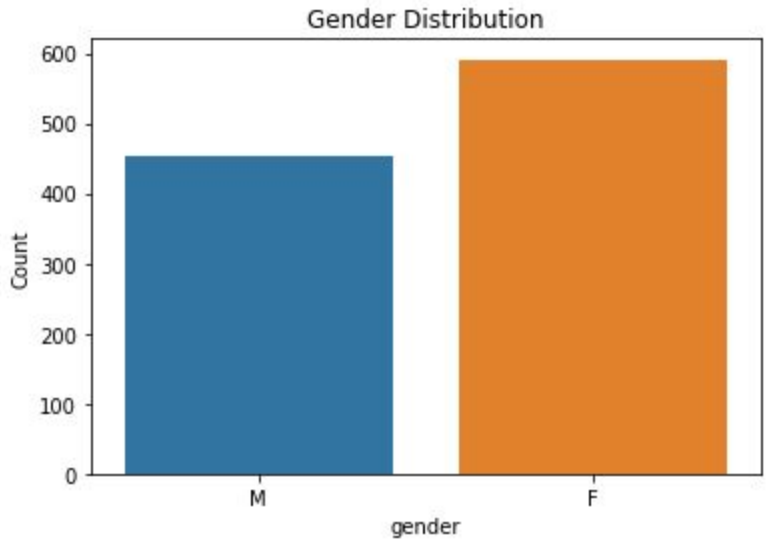


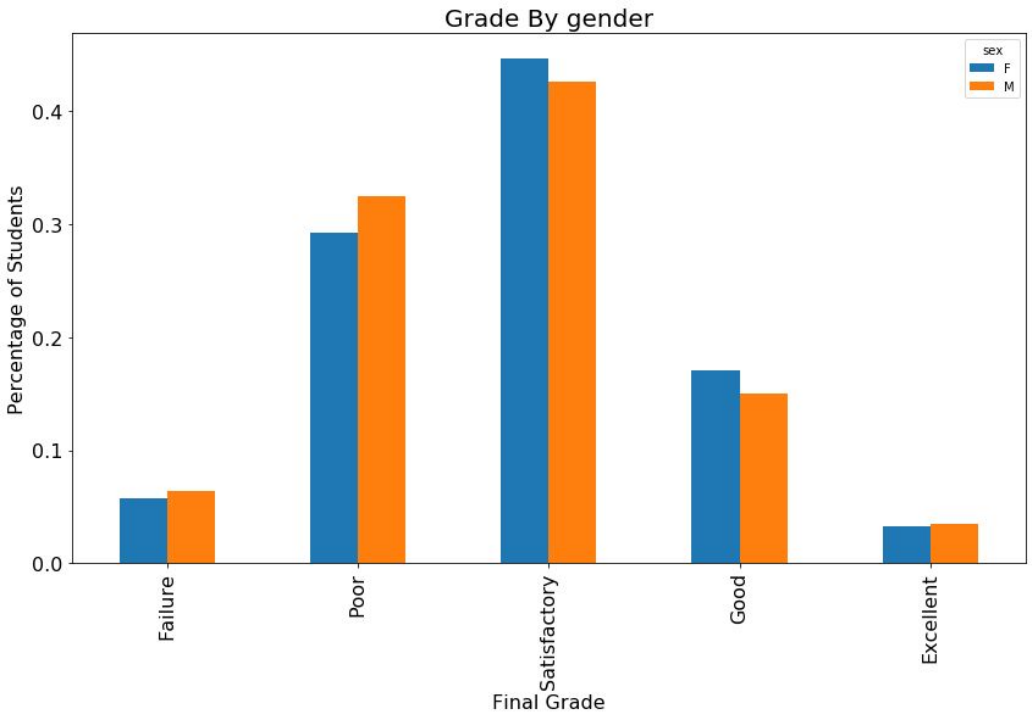
- Students who have internet are high in number and they are good at studies.



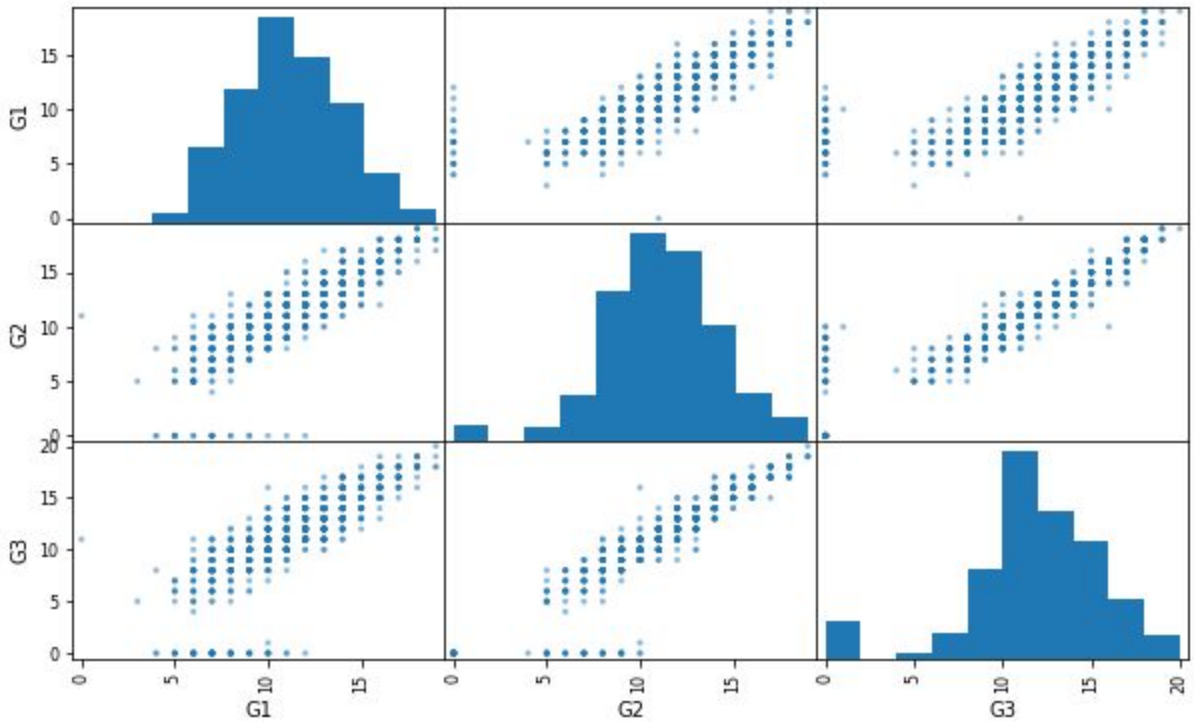


- Students who wants to take higher education are performing well at studies.





- Students who got ‘good’ or ‘satisfactory’ grades have girls in high number.



- **G3 is very much dependent on G1,G2.The pattern of G3 is highly influenced by G1 and G2.**

Feature Extraction:

Exploring and observing the data through visualizations resulted in some major conclusions,which made primary contribution to extract and structure important features.

So major conclusions from visualizations are-:

- ‘G1’ and ‘G2’ play a vital role in prediction of ‘FinalGrade’,so structuring these columns would hep.

- ‘absences’(number of absent days) has good weightage among features to decide the prediction.

- Information about Students whose father is working at home has least importance as per data analysis.

- Parental status information has negligible influence or effect on ‘FinalGrade’ prediction

Following feature extraction is done as part of implementing the results from above conclusions:-

- New features called ‘Grade1’ and ‘Grade2’ is included in feature set,to add weight to the significance of ‘G1’ ,’G2’ columns.

- Similarly a new feature called ‘ Requirement’ is extracted out of ‘absences’ column to subsequently add importance to the attribute.

  Similarly Following deductions are made on feature set:-

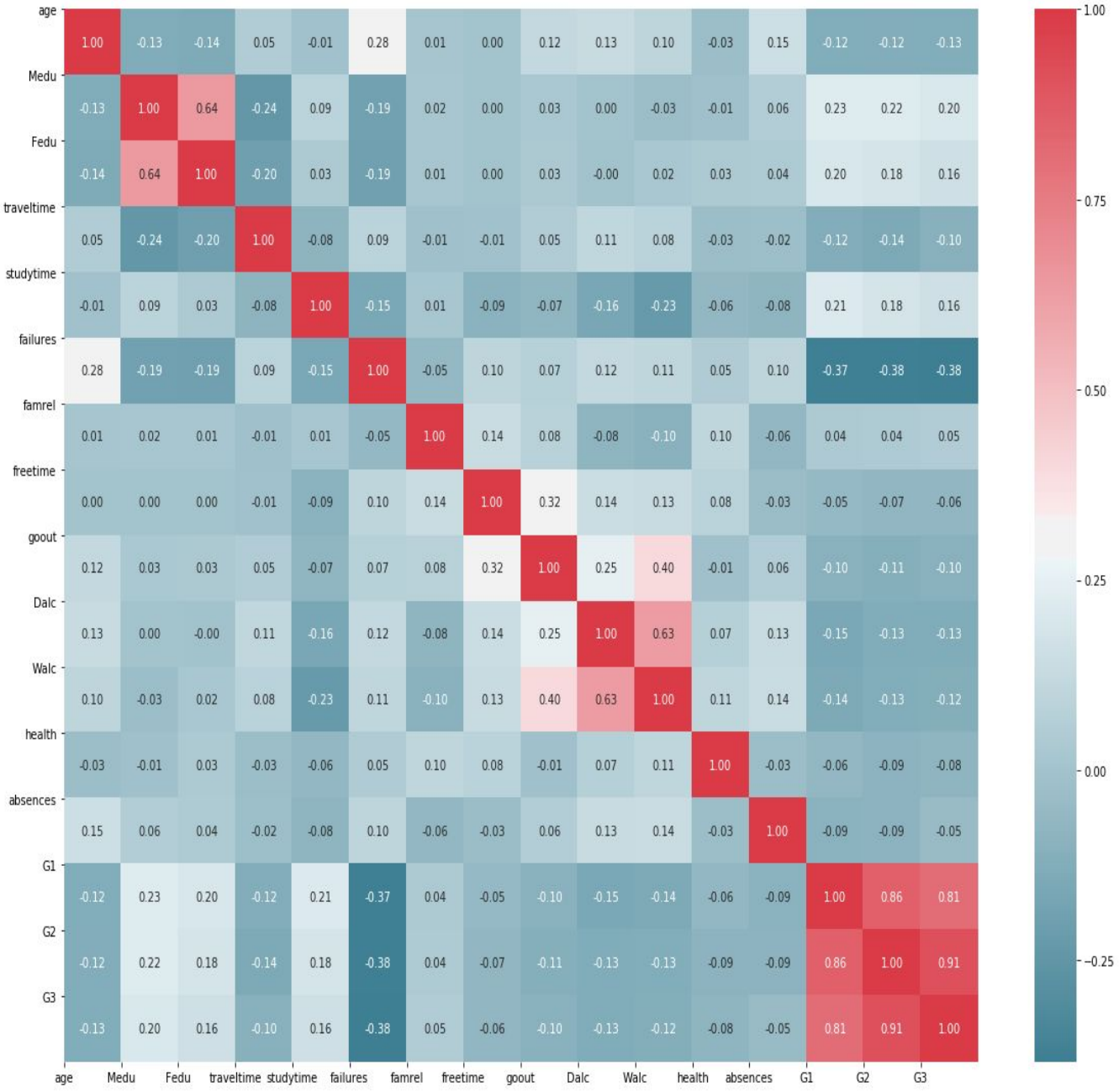
- ‘‘Fjob_teacher’ feature dropped from feature set.

- ‘Pstatus_A’ is removed from feature set.

- ‘Pstatus_T’ is dropped from feature set.

  because their presence is not making any significant difference, infact their absence making the model more effective and less sensitive.

Correlation Plot



Model Building:

So as per our analysis of data,our choices of model are-:

- Logistic Regression.
- Decision tree.
- Random forest.
- XGBoost.
- Support vector Machine(SVM)
- Ada boost

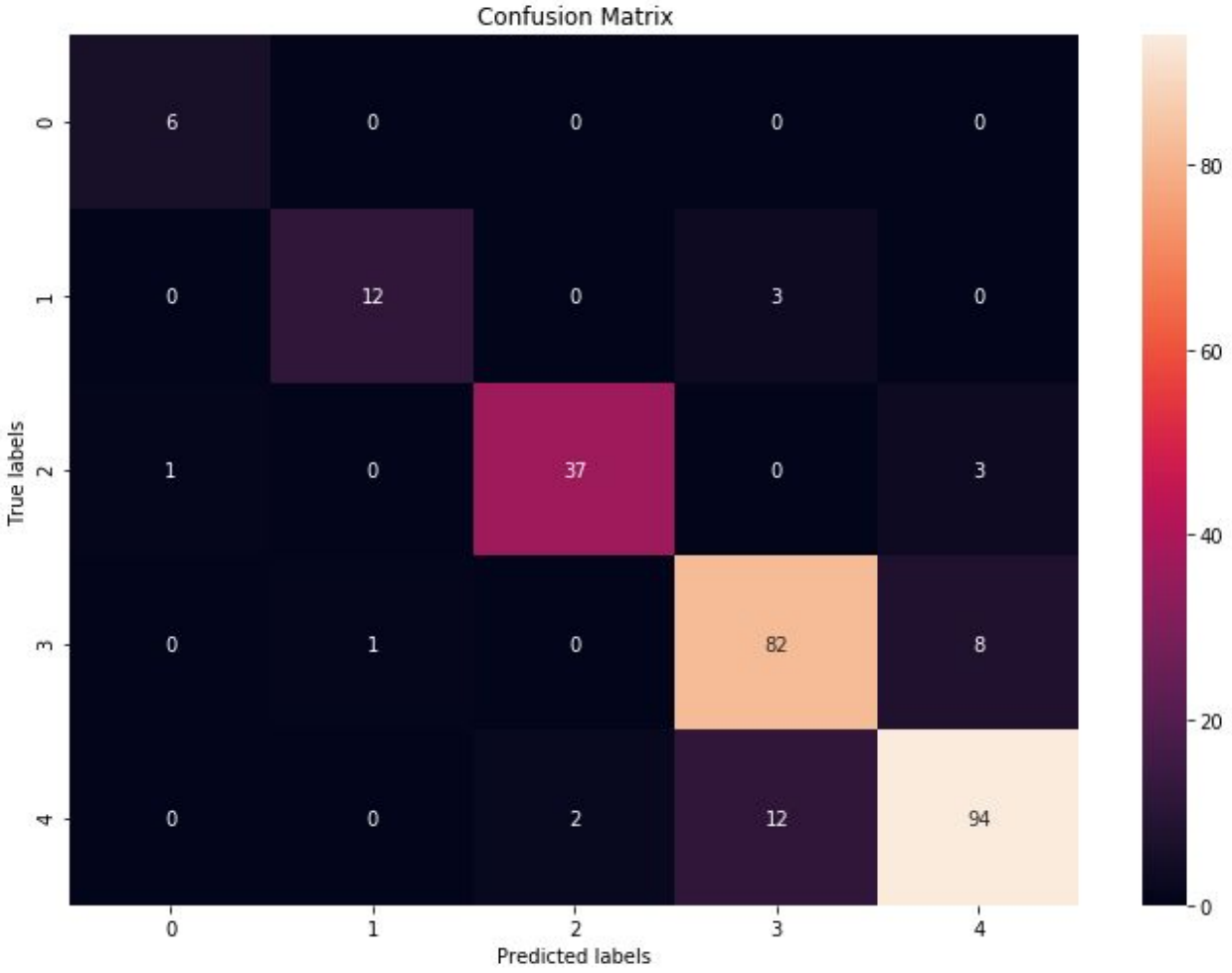
Analysis of models:- This table represents accuracies of different models when predicted on test data

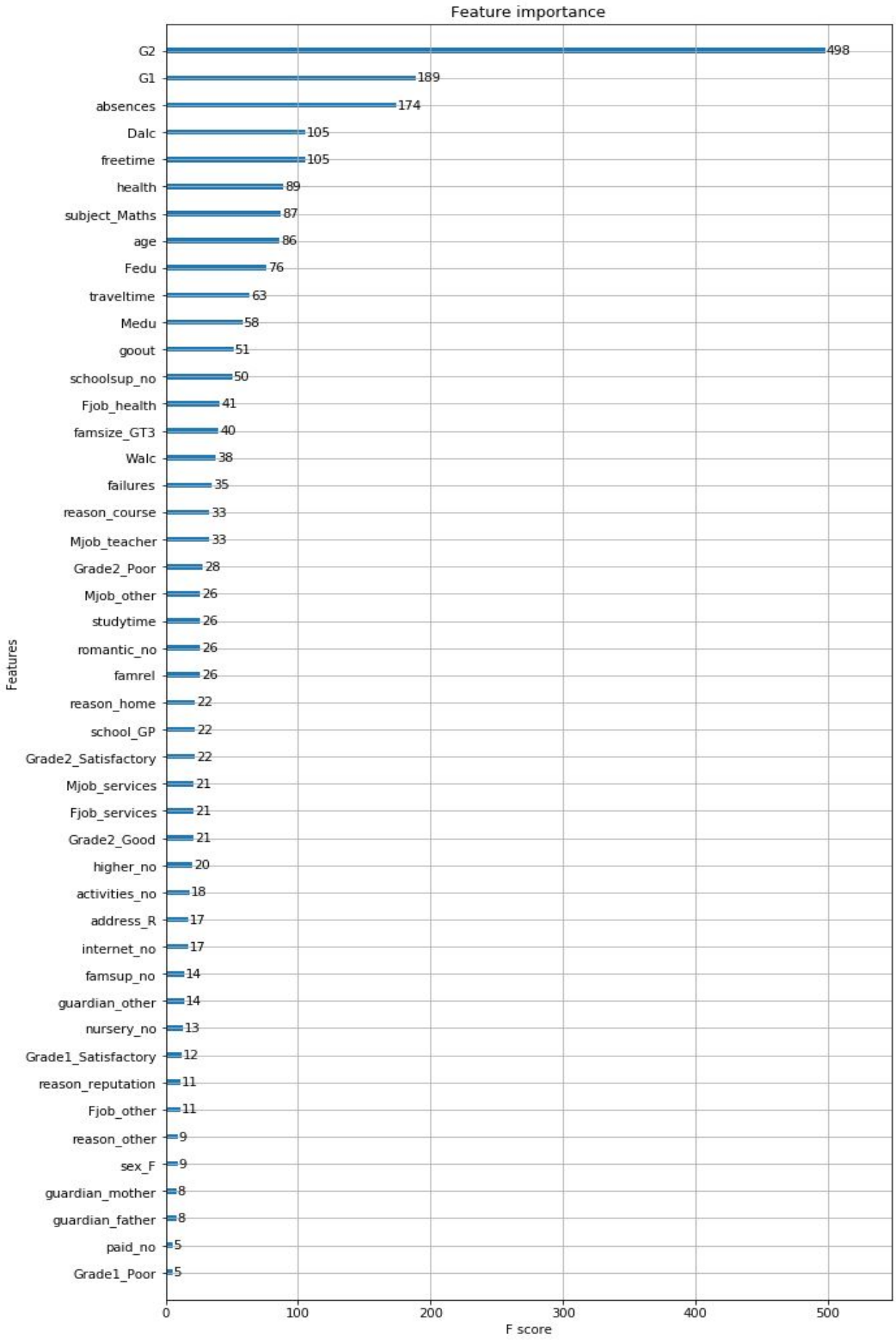
Model	FE1	FE2	FE3	FE4	FE5
LR	0.8103	0.8120	0.8161	0.8165	0.8182
RF	0.82	0.83	0.82	0.81	0.81
SVM	0.841	0.843	0.842	0.841	0.844
DT	0.801	0.801	0.803	0.802	0.83
ADA	0.791	0.7912	0.792	0.7911	0.7934
XG	0.86	0.87	0.87	0.87	0.88

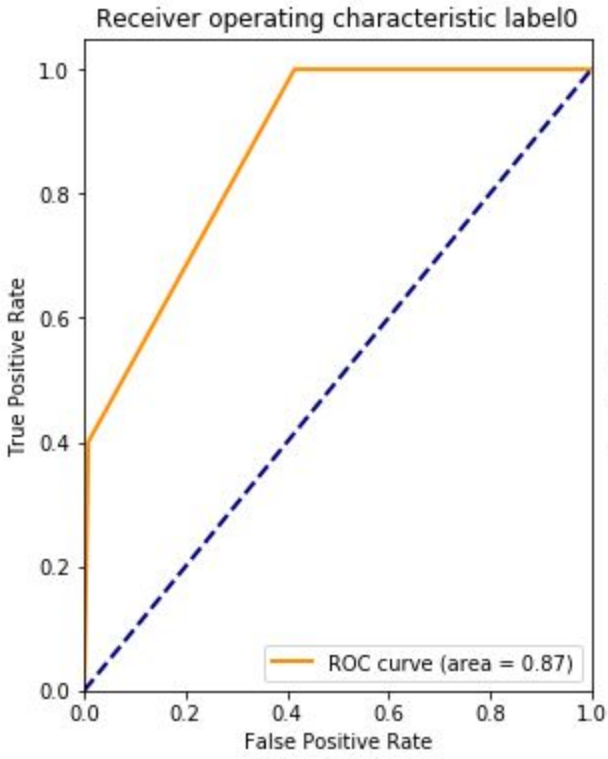
The following are different iterations over feature sets obtained by feature engineering:-
FE1: Contains all features
FE2: removing column name “Fjob_at_home”
FE3: also removing column name “Fjob_teacher”
FE4: also removing column name “Pstatus_A”
FE5: also removing column name “Pstatus_T”

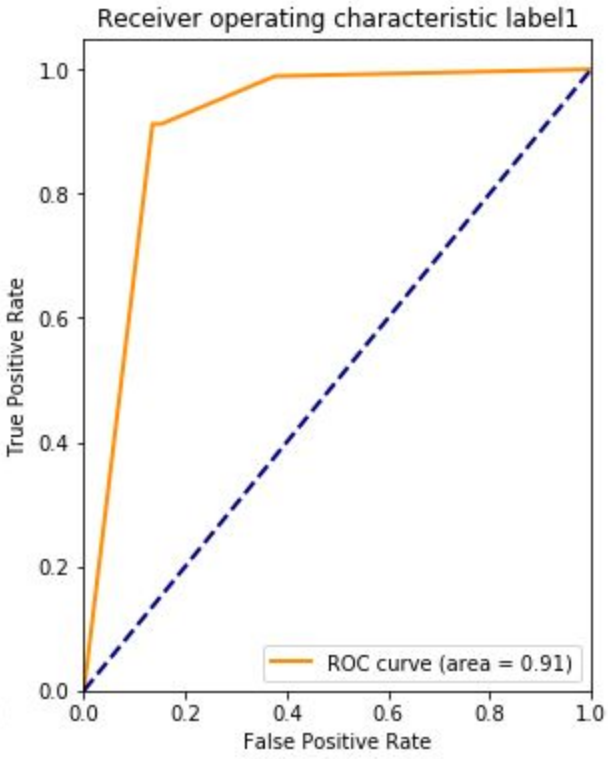
So as per observation over above models and it is clear that xg-boost dominates all other models and this model is more generalized based on our observations over train score and test score which has small difference between their predictions.

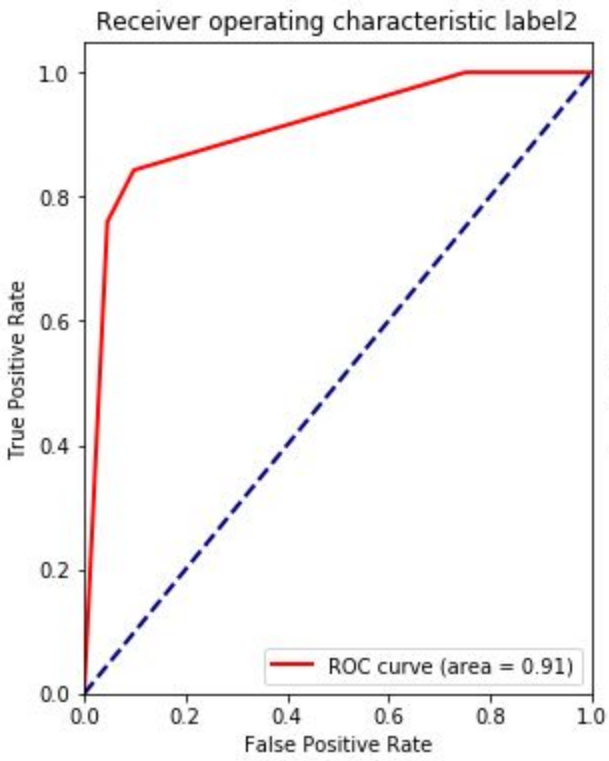
These are the confusion matrix and feature importance plots for our XGBoost model:

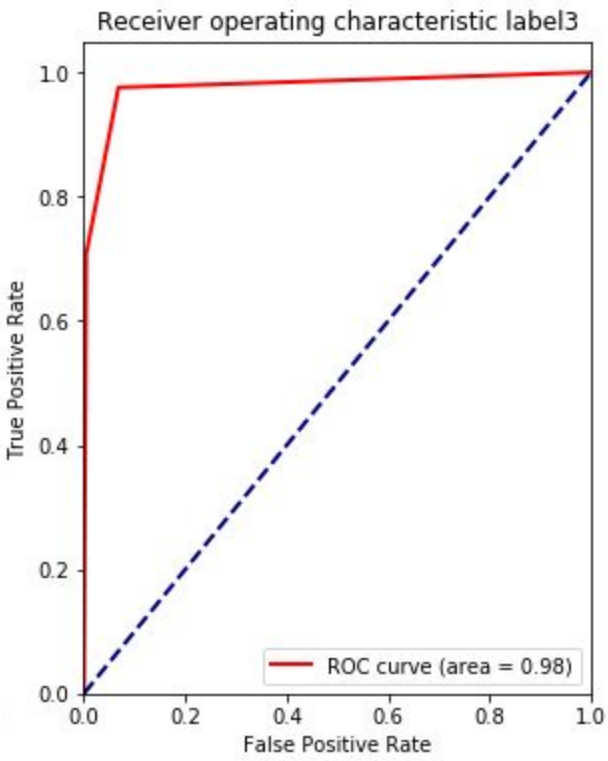


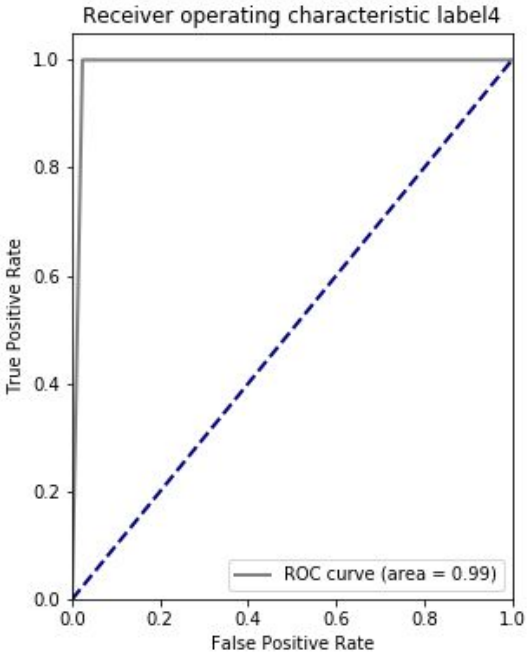

Feature importance
G2 498
G1 189
absences 174
Dalc 105
freetime 105
health 89
subject_Maths 87
age 86
Fedu 76
traveltime 63
Medu 58
goout 51
schoolsup_no 50
Fjob_health 41
famsize_GT3 40
Walc 38
failures 35
reason_course 33
Mjob_teacher 33
Grade2_Poor 28
Mjob_other 26
studytime 26
romantic_no 26
famrel 26
reason_home 22
school_GP 22
Grade2_Satisfactory 22
Mjob_services 21
Fjob_services 21
Grade2_Good 21
higher_no 20
activities_no 18
address_R 17
internet_no 17
famsup_no 14
guardian_other 14
nursery_no 13
Grade1_Satisfactory 12
reason_reputation 11
Fjob_other 11
reason_other 9
sex_F 9
guardian_mother 8
guardian_father 8
paid_no 5
Grade1_Poor 5
Features
0
100
200
300
400
500
F score

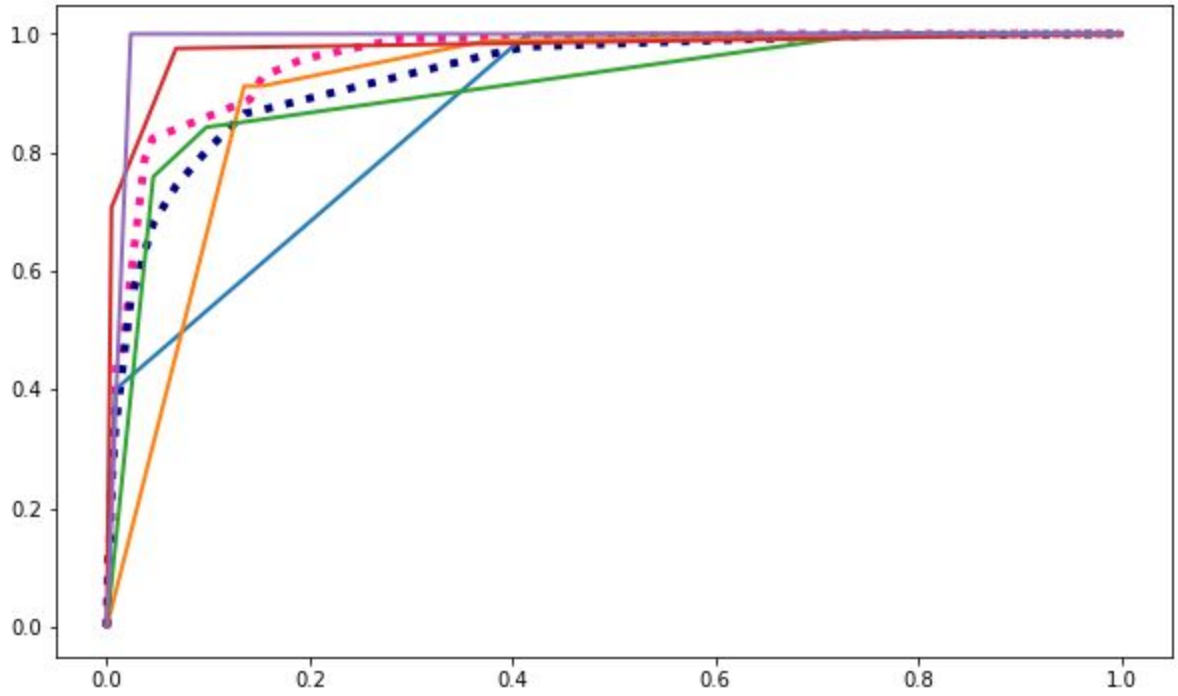

Receiver operating characteristic label0
True Positive Rate
False Positive Rate
ROC curve (area = 0.87)
0.0
0.2
0.4
0.6
0.8
1.0


Receiver operating characteristic label1
True Positive Rate
False Positive Rate
ROC curve (area = 0.91)
0.0
0.2
0.4
0.6
0.8
1.0


Receiver operating characteristic label2
True Positive Rate
False Positive Rate
ROC curve (area = 0.91)
0.0
0.2
0.4
0.6
0.8
1.0


Receiver operating characteristic label3
True Positive Rate
False Positive Rate
ROC curve (area = 0.98)
0.0
0.2
0.4
0.6
0.8
1.0





This is AUC for XGBoost model for all labels.

By:

Rohith Yogi N (IMT2016072)

Srujan Swaroop G (IMT2016033)

Siddarth Reddy D (IMT2016037)