A MINI PROJECT REPORT

ON

"Student Grade Analysis & Prediction"

Submitted by

Harshvardhan Deshmukh (Roll No. : 35)
Dhanashree Badgujar (Roll No. : 58)

Under the guidance of

Prof. Vishal Patil

For the Subject

Laboratory Practice II (410247) - Data Mining and Warehousing (410244 (D))

Submitted in partial fulfilment of the requirements for the award of the degree of

Bachelor in Computer Engineering

THE MET LEAGUE OF COLLEGES
MET
AS SHARP AS YOU CAN GET
Bhujbal Knowledge City

Institute of Engineering
Department of Computer Engineering

Academic Year 2021-22

CONTENTS

1 PROBLEM STATEMENT

The problem statement can be defined as follows:

Given a dataset containing attribute of 396 Portuguese students where using the features available from dataset and define classification algorithms to identify whether the student performs good in final grade exam, also to evaluate different machine learning models on the dataset.

2 ABSTRACT

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], the two data sets were modeled under binary/five-level classification and regression tasks. Important note: the target attribute G3 has a strong correlation with attributes G2 and G1.

This occurs because G3 is the final year grade (issued at the 3rd period), while G1 and G2 correspond to the 1st and 2nd period grades. It is more difficult to predict G3 without G2 and G1, but such prediction is much more useful (see paper source for more details).

Keywords - ***Data Mining, Regression tasks, Datasets, Classification, Binary Classification Performance***

---

3 INTRODUCTION

In higher educational institutes, many students have to struggle hard to complete different courses since there is no dedicated support offered to students who need special attention in the registered courses. Machine learning techniques can be utilized for students' grades prediction in different courses. Such techniques would help students to improve their performance based on predicted grades and would enable instructors to identify such individuals who might need assistance in the courses.
Here in this project we evaluate the grades of the students using various ML concepts, also we perform the analysis on the given datasets.

4 OBJECTIVE

- To Process and prepare the dataset for the better accuracy and also train the modules for the prediction of the output.
- Determining the Best Model.
- Identification of Relevant Attributes.

5 CLASSIFICATION OF ALGORITHMS

5.1 Regression Analysis

In statistical modelling, **regression analysis** is a set of statistical processes for estimating the relationships between a dependent variable (often called the 'outcome' or 'response' variable) and one or more independent variables (often called 'predictors', 'covariates', 'explanatory variables' or 'features'). The most common form of regression analysis is linear regression, in which one finds the line (or a more complex linear combination) that most closely fits the data according to a specific mathematical criterion. For example, the method of ordinary least squares computes the unique line (or hyperplane) that minimizes the sum of squared differences between the true data and that line (or hyperplane). For specific mathematical reasons (see linear regression), this allows the researcher to estimate the conditional expectation (or population average value) of the dependent variable when the independent variables take on a given set of values. Less common forms of regression use slightly different procedures to estimate alternative location parameters (e.g., quantile regression or Necessary Condition Analysis) or estimate the conditional expectation across a broader collection of non-linear models (e.g., nonparametric regression).

Regression analysis is primarily used for two conceptually distinct purposes. First, regression analysis is widely used for prediction and forecasting, where its use has substantial overlap with the field of machine learning. Second, in some situation regression analysis can be used to infer causal relationships between the independent and dependent variables. Importantly, regressions by themselves only reveal relationships between a dependent variable and a collection of independent variables in a fixed dataset. To use regressions for prediction or to infer causal relationships, respectively, a researcher must carefully justify why existing relationships have predictive power for a new context or why a relationship between two variables has a causal interpretation. The latter is especially important when researchers hope to estimate causal relationships using observational data.

5.2 CuffLinks

Cufflink is also a python library that connects plotly with pandas so that we can create charts directly on data frames. It basically acts as a plugin.

5.3 MatPlotLib

Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK. There is also a procedural "pylab" interface based on a state machine (like OpenGL), designed to closely resemble that of MATLAB, though its use is discouraged.[3] SciPy makes use of Matplotlib.

Matplotlib was originally written by John D. Hunter. Since then it has an active development community and is distributed under a BSD-style license. Michael Droettboom was nominated as matplotlib's lead developer shortly before John Hunter's death in August 2012 and was further joined by Thomas Caswell. Matplotlib is a NumFOCUS fiscally sponsored project.

6 CONFUSION MATRIX

A Confusion matrix is an N x N matrix used for evaluating the performance of a classification model, where N is the number of target classes. The matrix compares the actual target values with those predicted by the machine learning model. This gives us a holistic view of how well our classification model is performing and what kinds of errors it is making.

For a binary classification problem, we would have a 2 x 2 matrix as shown below with 4 values:

- The target variable has two values: Positive or Negative
- The columns represent the actual values of the target variable
- The rows represent the predicted values of the target variable

Understanding True Positive, True Negative, False Positive and False

Negative in a Confusion Matrix

True Positive (TP)

- The predicted value matches the actual value.
- The actual value was positive and the model predicted a positive value.

True Negative (TN)

- The predicted value matches the actual value.

- The actual value was negative and the model predicted a negative value.

False Positive (FP) – Type 1 error

- The predicted value was falsely predicted.
- The actual value was negative but the model predicted a positive value.
- Also known as the Type 1 error

False Negative (FN) – Type 2 error

- The predicted value was falsely predicted.
- The actual value was positive but the model predicted a negative value.
- Also known as the Type 2 error

Accuracy: he accuracy of a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier.

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision: Precision can be thought of as a measure of exactness (i.e., what percentage of tuples labelled as positive are actually such).

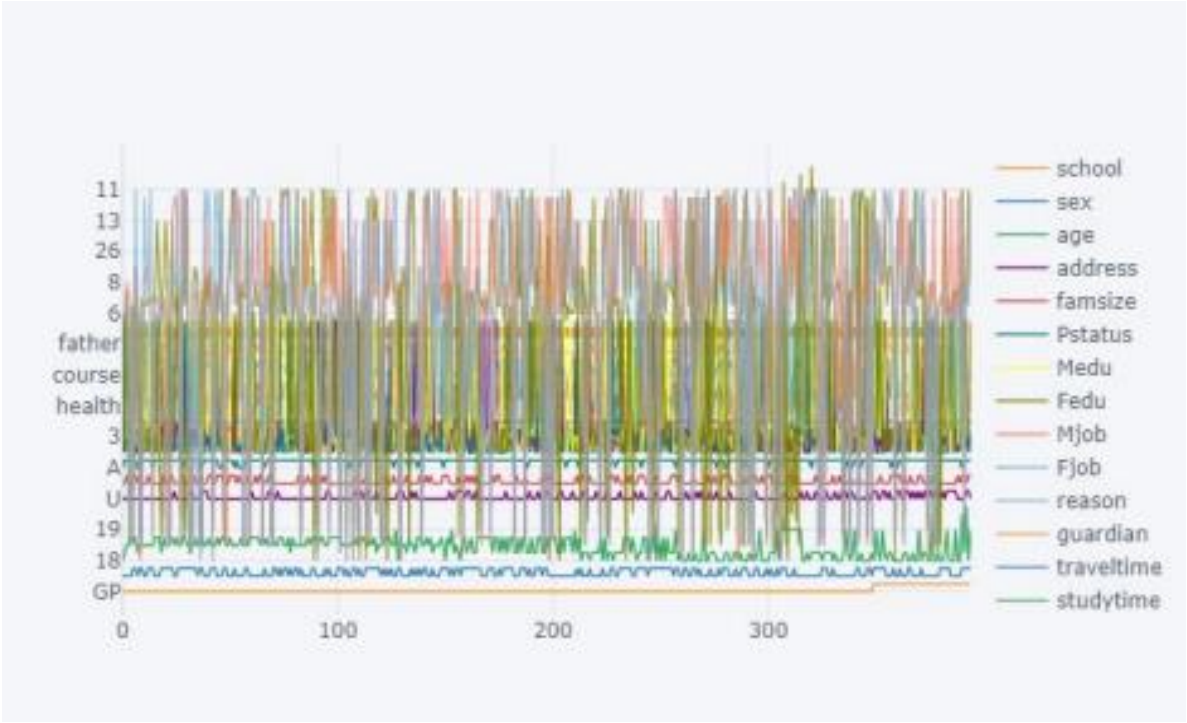
$$precision = \frac{TP}{TP + FP}$$

Recall: Recall is a measure of completeness (what percentage of positive tuples are labelled as such)

$$recall = \frac{TP}{TP + FN} = \frac{TP}{P}$$

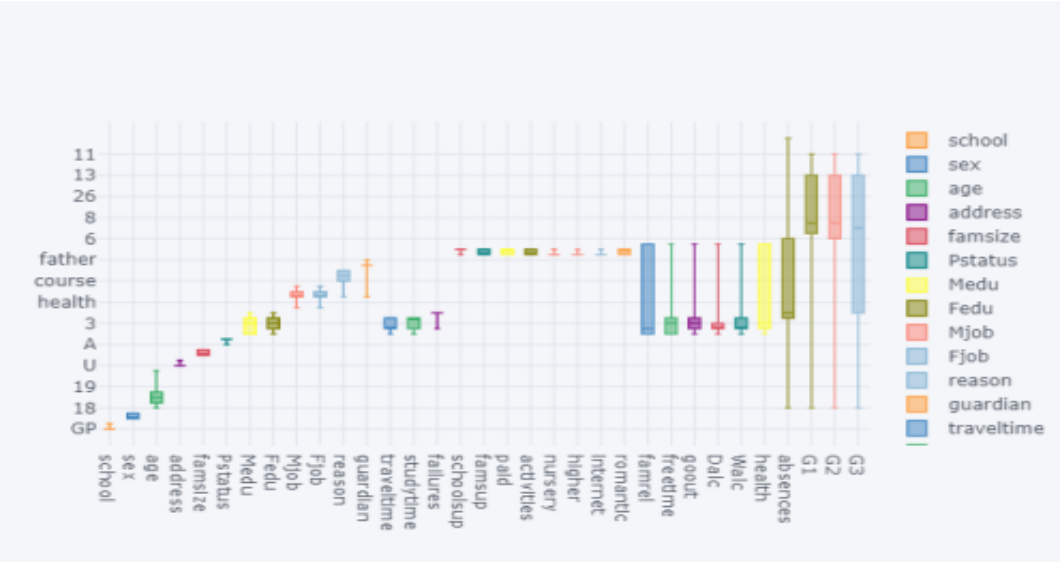
6 EXPERIMENTAL RESULTS

6.1 KDE Plot to view all attributes using cufflinks



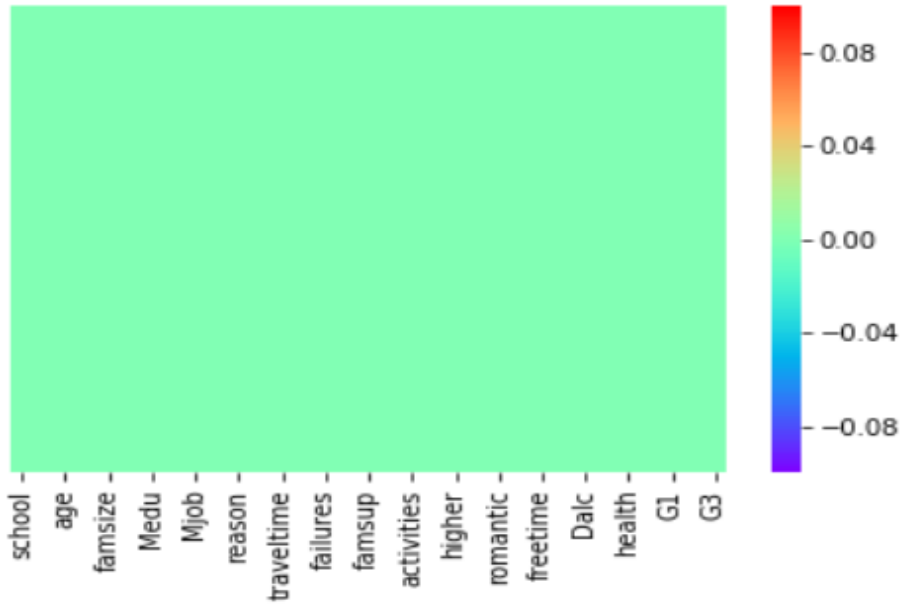
Observation: cufflink connects plotly with pandas to create graphs and charts of dataframes directly

4.2 - Box Plot to view all attributes using cufflinks

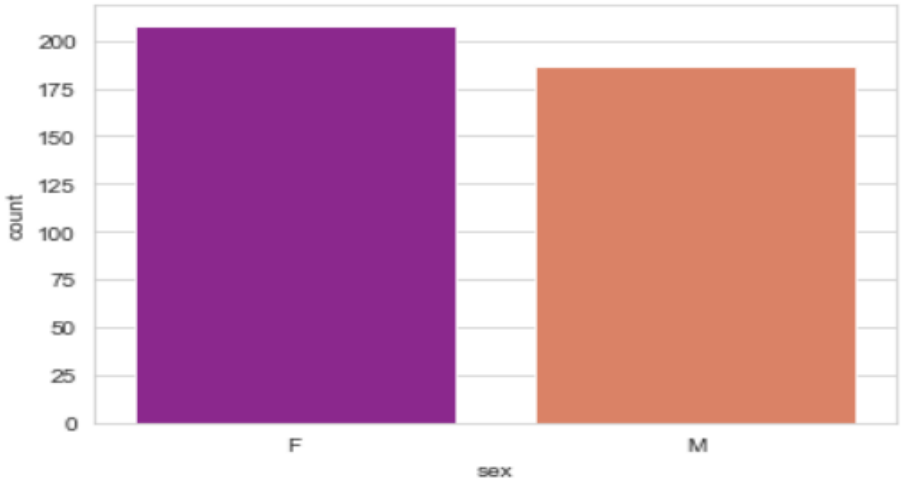


4.3 - Histogram Plot for G3 (Final Grade) using cufflinks

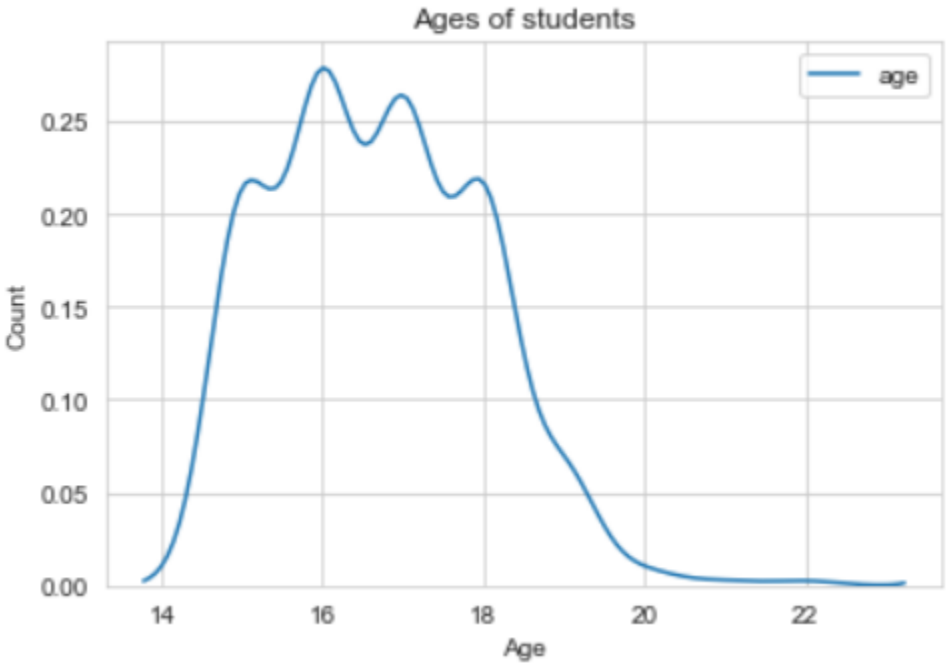
4.4 - Pictorial representation of any null data present in the dataset.



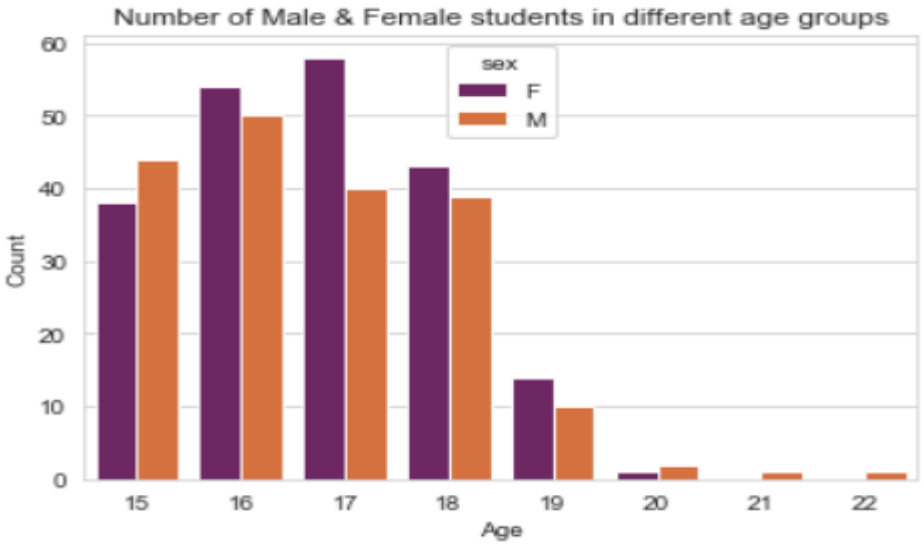
4.5 - Count Plot for Student Sex Attribute



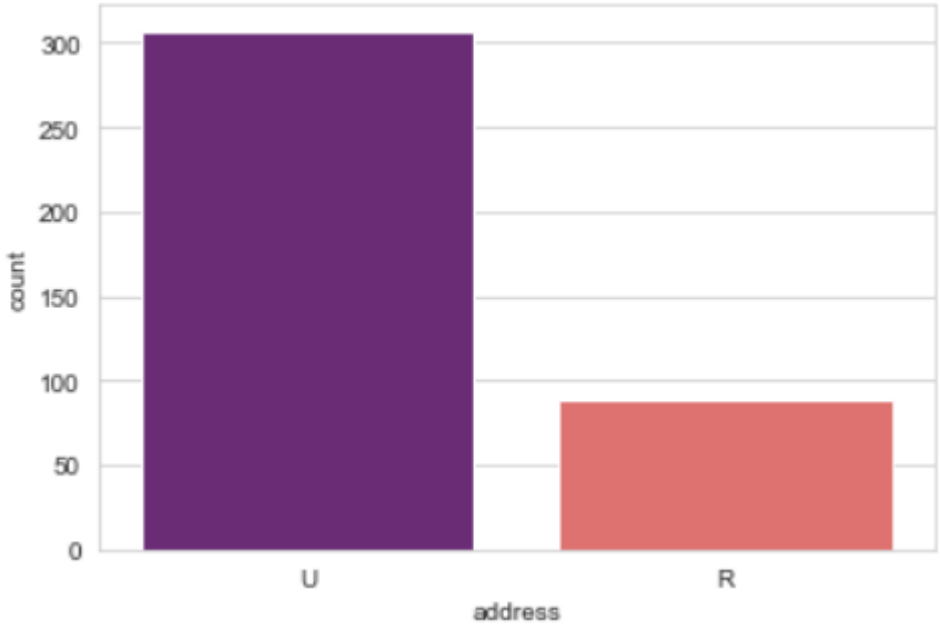
4.6 - Kernel Density Estimation for Age of Students



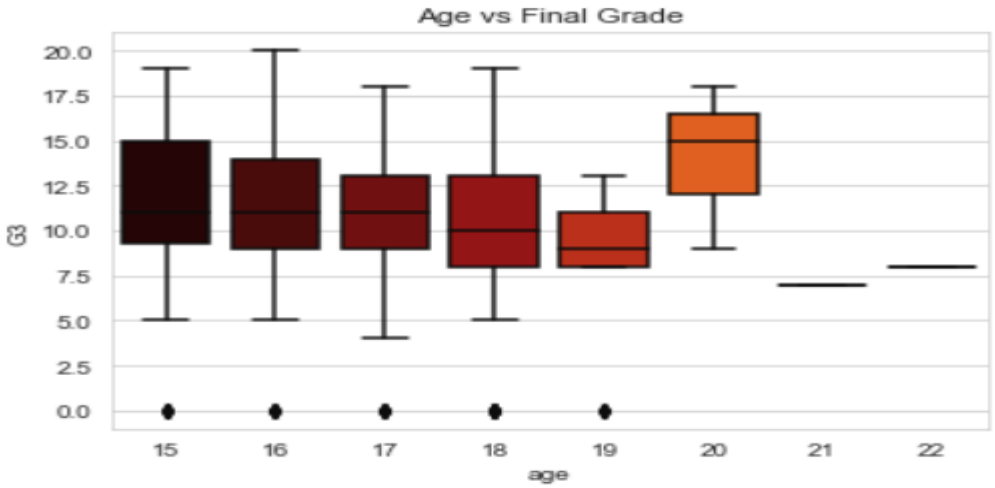
4.7 - Count Plot for Male & Female students in different age groups.



4.8 - Count Plot for students from Urban & Rural Region



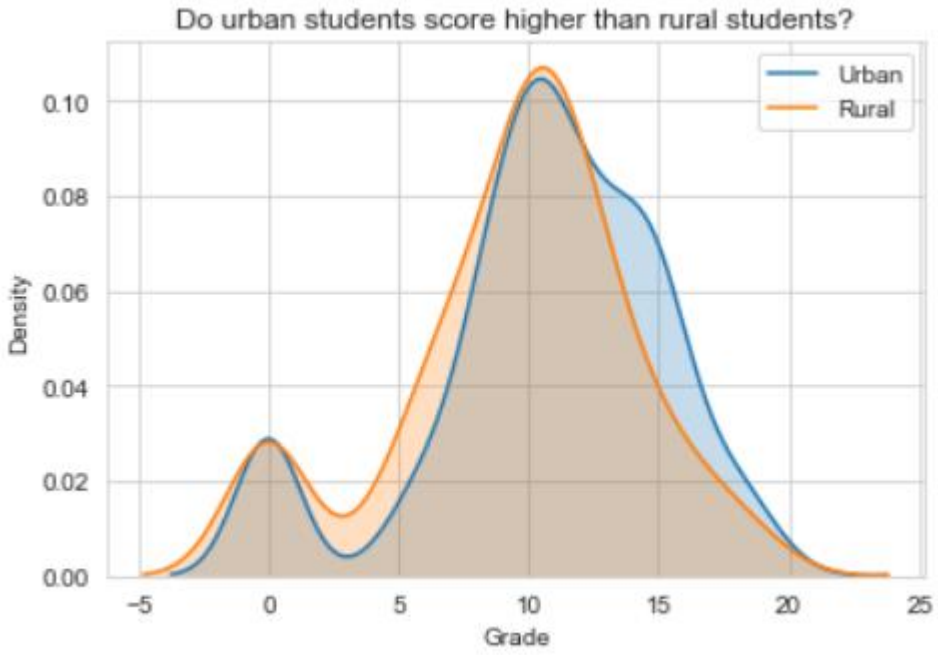
4.9 - Does age affect final grade?



Observation:

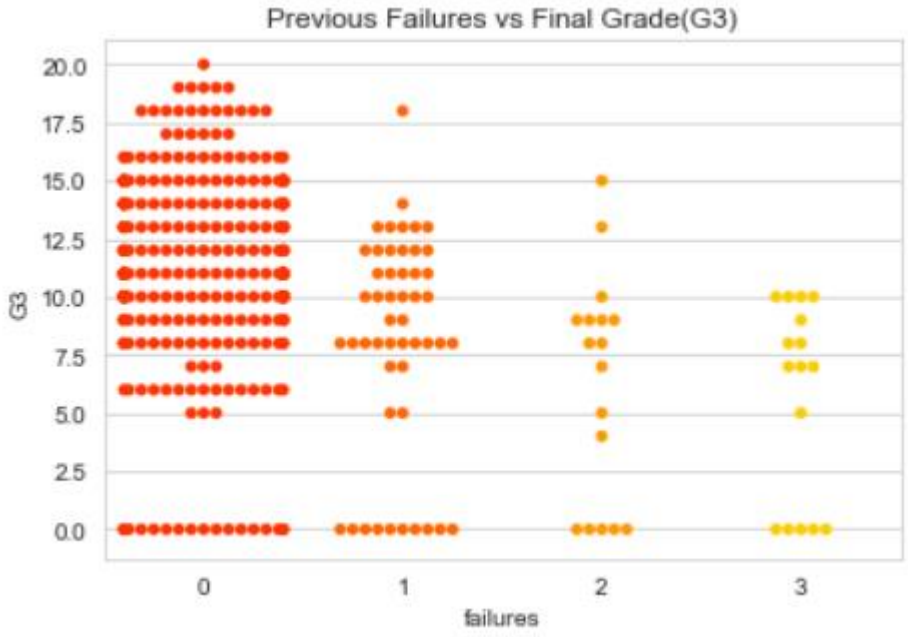
- Plotting the distribution rather than statistics would help us better understand the data.
- The above plot shows that the median grades of the three age groups(15,16,17) are similar. Note the skewness of age group 19. (may be due to sample size). Age group 20 seems to score highest grades among all.

4.10 - Do urban students perform better than rural students



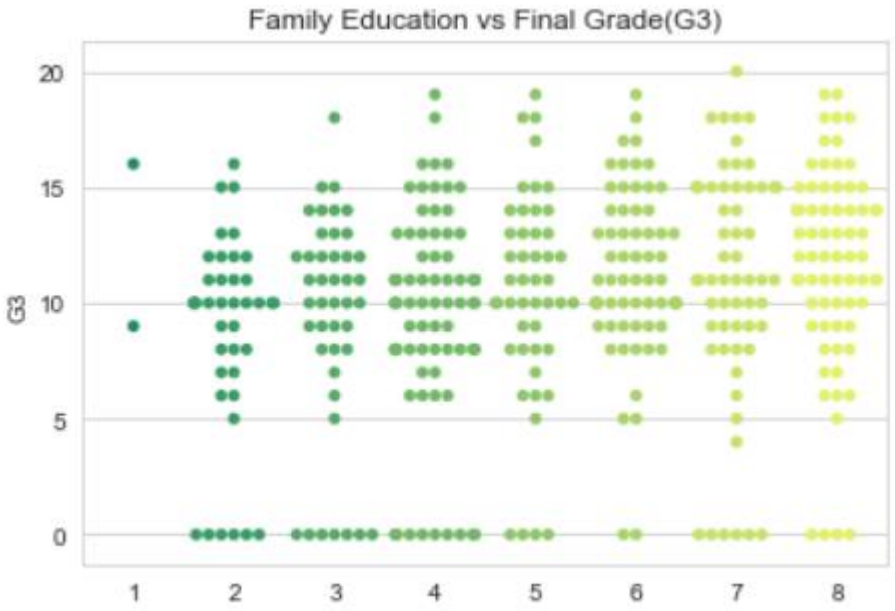
Observation: The above graph clearly shows there is not much difference between the grades based on location.

4.11 - Previous Failures vs Final Grade(G3)



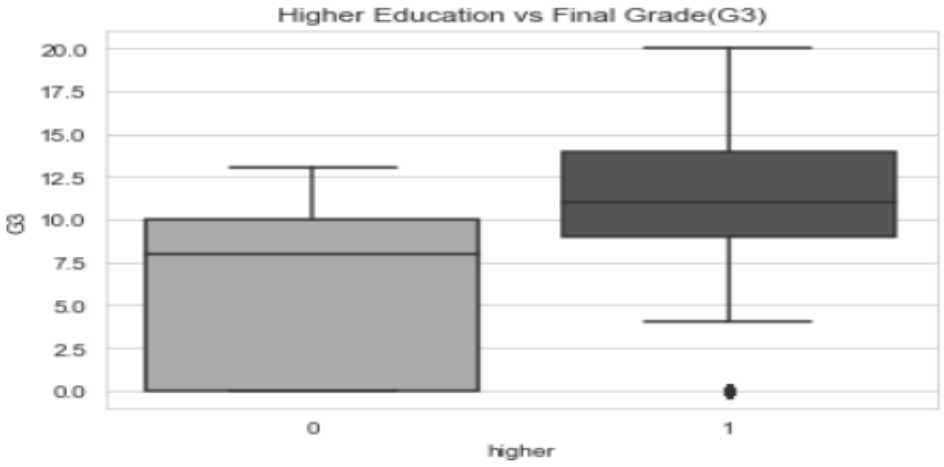
Observation: Student with less previous failures usually score higher

4.12 - Family Education vs Final Grade(G3)



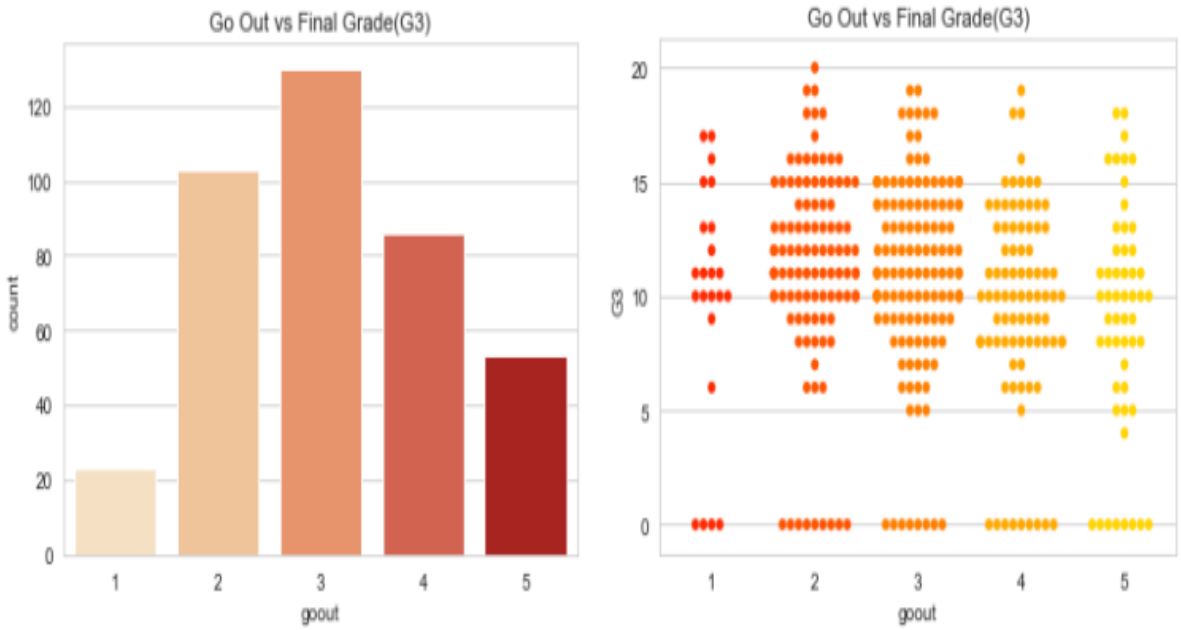
Observation: Educated families result in higher grades

4.13 - Higher Education vs Final Grade(G3)



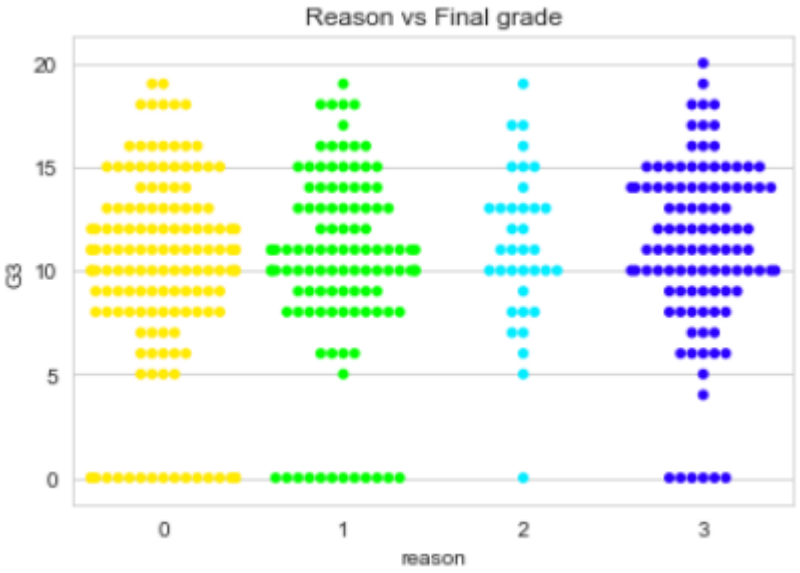
Observation: Students who wish to go for higher studies score more.

4.14 - Go Out vs Final Grade(G3)



Observation: The students have an average score when it comes to going out with friends & Students who go out a lot scoreless.

4.15 - Reason vs Students Count



Observation: The students have an equally distributed average score when it comes to reason attribute.

8 CONCLUSIONS

As we see both MAE & Model RMSE that the Linear Regression is performing the best in both cases