**Student Performance Prediction Model**

*Author:*

Ravikumar Panchal

# Contents

|  |  |
| --- | --- |
| **Declaration of Authorship** | **III** |
| **Abstract** | **VII** |
| **Acknowledgements** | **IX** |
| 1 **Introduction** | **4** |
| 1.1 Background and significance of the project | 4 |
| 1.2 Objectives | 4 |
| 1.3 Scope and limitations | 5 |
| 1.4 Structure of the Report | 5 |
| **2 Related Literature and Theoretical Focus** | **6** |
| 2.1 Literature Review | 6 |
| **3 Data Collection** | **10** |
| 3.1 Description of the dataset | 10 |
| 3.2 Source of the data | 11 |
| 3.3 Data pre-processing Steps | 13 |
| 3.3.1 Data cleaning and transformation | 13 |
| **4 Exploratory data analysis** | **18** |
| 4.1 Summary statistics of the dataset | 18 |
| 4.2 Data Visualization | 20 |
| 4.3 Identification of patterns and correlations | 24 |
| **5 Feature Engineering** | **26** |
| 5.1 Selection and engineering of relevant features | 26 |
| 5.2 Techniques used for feature selection and extraction | 32 |
| **6 Machine Learning Model Development** | **35** |
| 6.1 Description of the machine learning algorithms used | 35 |
| 6.2 Model selection criteria | 36 |
| 6.3 Model training and validation process | 38 |
| **7 Performance Evaluation** | **39** |
| 7.1 Evaluation metrics used | 39 |
| 7.2 Comparative analysis of different models | 40 |
| 7.3 Interpretation of results | 43 |
| **8 Conclusion** | **45** |
| 8.1 Summary of key findings | 45 |
| 8.2 Achievement of project objectives | 45 |
| 8.3 Recommendations for further implementation | 46 |
| **References** | 47 |
| **Appendices** | 48 |

# List of Figures

|  |  |
| --- | --- |
| 1.1 Mathematics Prediction result | 4 |
| 2.1 Performance Influence | 6 |
| 2.2 Student Under Probation | 7 |
| 2.3 Deep Learning Neural Network | 7 |
| 2.4 Novel Statistical Features | 8 |
| 2.5 Feature identification using ML | 8 |
| 3.1 Source dataset | 11 |
| 3.2 Python Library Import | 12 |
| 3.3 Dataset Records | 12 |
| 3.4 Missing Value Count | 13 |
| 3.5 Duplicate Value Count | 13 |
| 3.6 Null Data Count | 14 |
| 3.7 Unique Value Count | 14 |
| 3.8 Statistical Data Count | 15 |
| 3.9 Categorical data Count | 16 |
| 3.10 Numerical – categorical data count | 17 |
| 4.1 Summary statistical data count | 18 |
| 4.2 Histogram and KDE of Average value | 19 |
| 4.3 Histogram and KDE of total score | 20 |
| 4.4 Histogram and KDE of male and female performance | 21 |
| 4.5 Histogram and KDE of parental education male and female wise | 22 |
| 4.6 Histogram and KDE of ethenicity male and female wise | 23 |
| 4.7 Violin plot of math , reading and writing | 24 |
| 4.8 Identification of patterns and correlations | 25 |
| 5.1 Gender Distribution | 26 |
| 5.2 Marks Distribution | 27 |
| 5.3 Group wise distribition | 28 |
| 5.4 Race/Ethnecity impact | 29 |
| 5.5 Parental education impact | 30 |
| 5.6 Course preparation Test | 31 |
| 5.7 Course preparation Impact | 32 |
| 7.1 Plot Prediction Value | 44 |

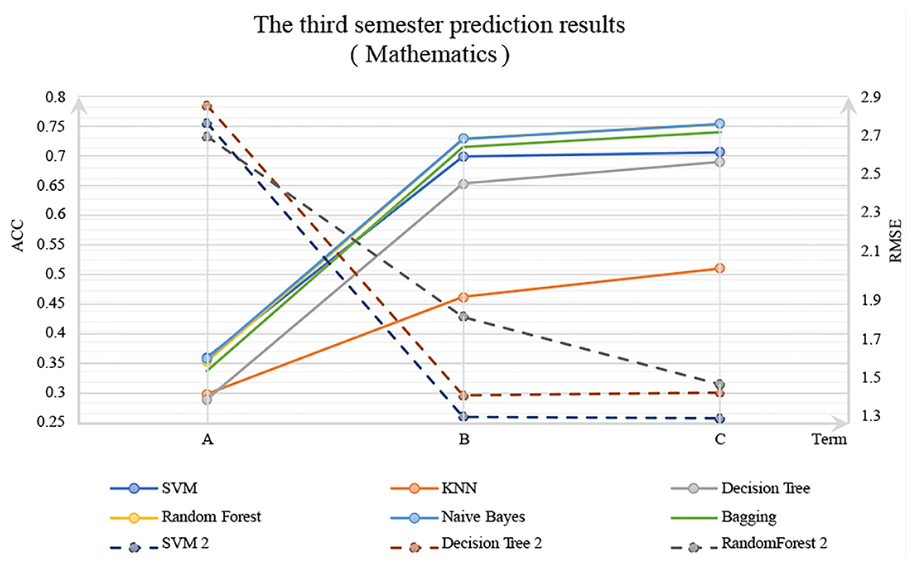
**Chapter 1**

# Introduction

### Background and significance of the project

In recent years, educational institutions have faced challenges in effectively identifying students who may be at risk of academic underperformance or dropout. This challenge is exacerbated by factors such as increasing student diversity, limited resources for personalized interventions, and the growing demand for accountability in education.

To address these challenges, many institutions are turning to data-driven approaches to gain insights into student behavior and performance. By leveraging the vast amounts of data available, including academic records, demographic information, and engagement metrics, educators can make informed decisions to support student success.



*Fig 1.1 mathematics Prediction result*

Despite the growing interest in data-driven education, there is a gap in research on developing comprehensive predictive models for student performance. While individual studies have explored various factors influencing student outcomes, there is a need for an integrated model that incorporates multiple predictors and utilizes advanced machine learning techniques for accurate predictions.

### Objectives

The primary objective of this project is to develop a student performance prediction model using machine learning techniques. This model aims to predict student outcomes, such as grades or graduation rates, based on a combination of academic, demographic, and behavioral factors.

Secondary objectives include evaluating the performance of different machine learning algorithms, such as Decision Trees and Support Vector Machines, and assessing the impact of demographic factors, such as socioeconomic status and ethnicity, on student performance.

### Scope and limitations

### This project focuses on developing and evaluating a predictive model using a specific dataset obtained from a single educational institution. The model will be trained and tested using historical student data, and the evaluation will be based on predefined performance metrics.

### Limitations of the study include the reliance on a single dataset, which may not fully represent the diversity of student populations, and the inability to account for external factors, such as changes in curriculum or teaching methods that may influence student performance over time.

### Structure of the report

The report consists of several chapters, including an introduction, literature review, methodology, results, discussion, and conclusion. Each chapter contributes to achieving the objectives of the project by providing context, reviewing relevant research, describing the methodology, presenting findings, and discussing implications.

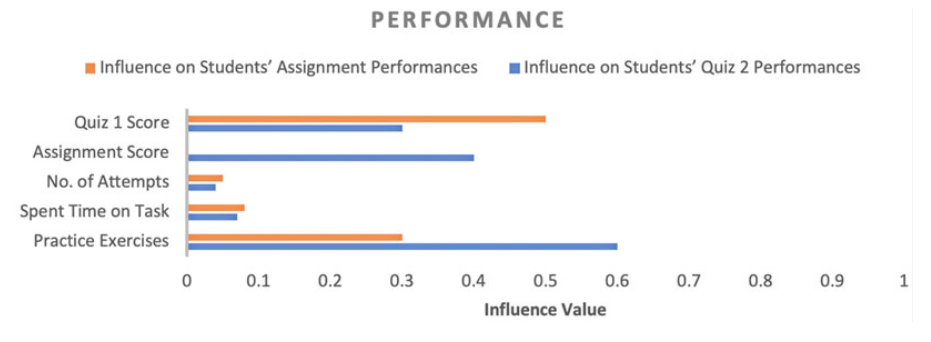
The chapters are interconnected, with each building upon the previous ones to provide a comprehensive analysis of the student performance prediction model. The literature review informs the development of the methodology, which in turn guides the analysis and interpretation of the results.

**Chapter 2**

# Related Literature and Theoretical Focus

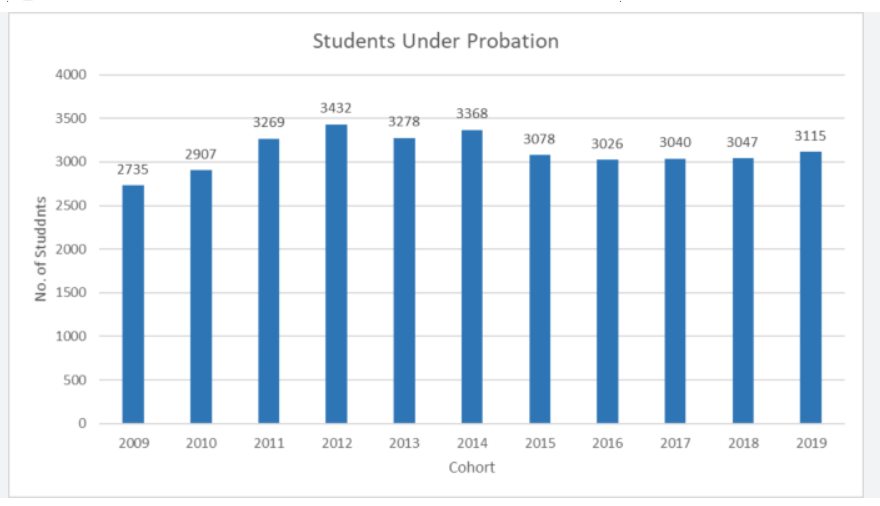
### Literature Review

The development and design of the student management system based on the network environment [2] discusses the method of the management information in higher education. On the basis of a comprehensive investigation and analysis on the student management in higher education, the authors establish the models of the college students’ management information by adopting the advanced information technology, and construct the student management information platform. Moreover, they analyse the characteristics of the information management in higher education, and elaborate the methods to solve the difficulties confronting in the student management of the higher education. Finally, the key method and technology to carry out the information management platform are presented.



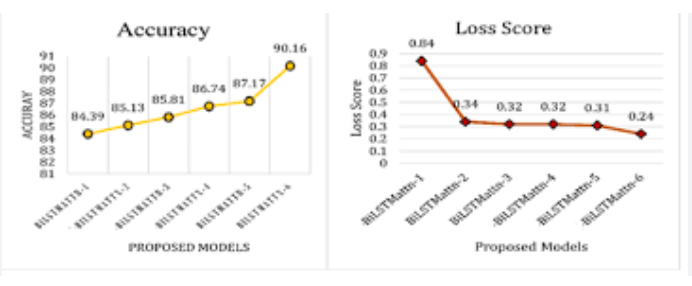
*Fig 2.1 Performance Influence*

In Android-based Attendance management system [3], a method of taking attendance by employing an application running on the Android platform is proposed. This application, once installed can be used to download the students list from a designated web server. Based on the downloaded list of students, the device will then act like a scanner to scan each of the student cards one by one to confirm and verify the students’ presence. The device’s camera will be used as a sensor that will read the barcode printed on the students’ cards. The updated attendance list is then uploaded to an online database and can also be saved as a file to be transferred to a PC later on. This system will help to eliminate the current problems, while also promoting a paperless environment at the same time. Since this application can be deployed on lecturers own existing Android devices, no additional hardware cost is required.



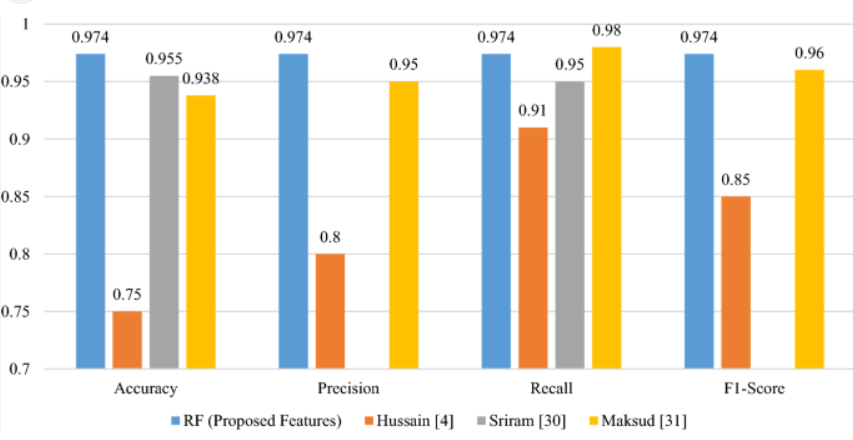
*Fig 2.2 Students under probation*

The paper, deskinment of student information management based on b/s architecture [4], uses the B/S structure to design the student information management system, and explains the system design principle, system plan and structure, the function module of information system according to current university student information management needs. It provides an interactive students management platform for the information of a large number of students and the management of students.



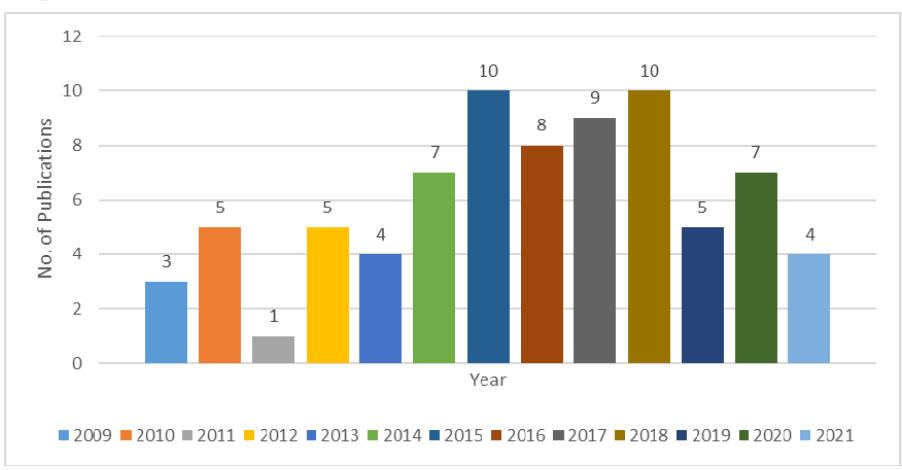
*Fig 2.3 Deep Learning Neural Network*

Research and implementations of web services in android network communication framework using Volley [5] includes combination of Web Services and mobile devices will promote the development of mobile applications. Volley framework (Google 2013 proposed) has the advantages of convenient use and network request faster, but it does not support Web Services. Extension of Volley, to support the Web Services, which can facilitate the Web Services application development, but also can improve the access performance of Web Services. On the basis of analysis and research of the Volley, Ksoap2 and Java Web Services, through the implementation of the HTTP Stack interface and the expansion of JSON Object Request to realize support for Web Services .The scheme uses JSON format to transfer data, support SSL/TLS protocol requests, custom parameter, sets or gets the request header. This scheme is good compatibility, easy to use, suitable for application on Android platform.



*Fig 2.4 Novel Statistical Features*

A Study of Student Information Management Software [6] provides the particulars to carry out the performance, management and decision-making functions of enterprises or organizations. Enormous grow of students is caused to expand the functionality in the respective educational institutions. As student added to the educational system it is difficult to manage and track student details. To overcome difficulties, the authors have come up with a new approach student information management system with additional features. This new approach will provide fast processing, efficient student tracking, and produces desired result. This approach will allow students to save their personal details. It is more secure, reliable and easy to use.



*Fig 2.5 Feature identification using ML*

The college student management system design using Computer Aided System [7] utilizes computer aided system. The model plays main role in an institution or in the college management. Initially, the system has developed with four layers based on the hierarchy such as Web display layer where application is deployed and displayed for end users. Business logic layer responsible for handling the functionality of the product. Data access layer is responsible for viewing the data. Database layer responsible for storing the student data. In Database layer ER diagram has been designed to provide data normalization. The process provides complete information about student, faculties and educational institution. Third thing in this project is to allowing user based on their categories.

Toward A Student Information System [8] provides end user a seamless navigation to the application and ease of access. The model provides information management storing of student academic reports. This model consists of various functionalities like information about the courses available in the college starting from first grade to graduation. It also enables students to enrol to particular course online, online fees payment, examination results, and also get notified when important events occur. All data stored and retrieved through the application is secure. So, to achieve this the authors have developed a powerful web based secured interface application which supports all type of request which are coming from the students also which gathers and corrects all student information. To achieve this, authors have used similarity (Euclidean distance) algorithm. The results showed that the new information gathered by the SDS has the ability to fill the requirement and done the error correction in the traditional model. The system provides seamless access through the web-based application to access and manage different department or all over the organization. This system is used to mainly monitor the attendance for the university. Students are provided access to login to application and view the progress report and attendance report. Initially faculties/students get registered with the system once they finish registration process, they can access the system as well as they are able to do the changes in the data. As per the requirement, users are granted with certain level permission to manage and track the student information. Either student or faculty can upload and copy the statistics from the database. Since it is a web-based application which is accessible from any part of the world it has certain features like accessibility, ease of use, etc. It is developed to suit the current environment which is rapidly growing in the student domain.

**Chapter 3**

# Data Collection

### Description of the Dataset

There are numerous approaches to data collection depending on the nature of the research being conducted. In this project, the methods adopted include the following: Interview, World Wide Web, references to published and unpublished collection. The data collected for this research can be broadly classified into two types, namely: the primary and secondary data.

##### Primary Data

Primary data can be defined as data collected directly from respondent relevant to the subject under investigation. The primary data obtained in this case is by interview method. The primary source data collection are the sources from where first-hand information can be obtained. The tools for gathering the primary source of data collection include; interview, observation and questionnaire etc.

##### Secondary Data

These are the sources of data collection in which an already made data is being obtained

i.e. the information that is already in printed form. Sources of secondary data include, textbooks, magazines, journals etc. In case of this project, most of the data is published documents and references.

##### The Interview Approach

We employed a combination of both oral interview, questionnaires and observation method by consulting staff, students, lecturers and downloading of information from websites to investigate the system. The oral interview and distribution of questionnaires were limited to the Gujarat Technological University.

### Source of the data

The dataset contains information about students, including their gender, parental education, study time, and academic performance. Exploring and preprocessing the data is the initial step in building an effective predictive model. The dataset contains variables such as Gender, race/Ethnicity, parental level of education, lunch, test preparation course, and the scores which includes math, reading and writing scores.

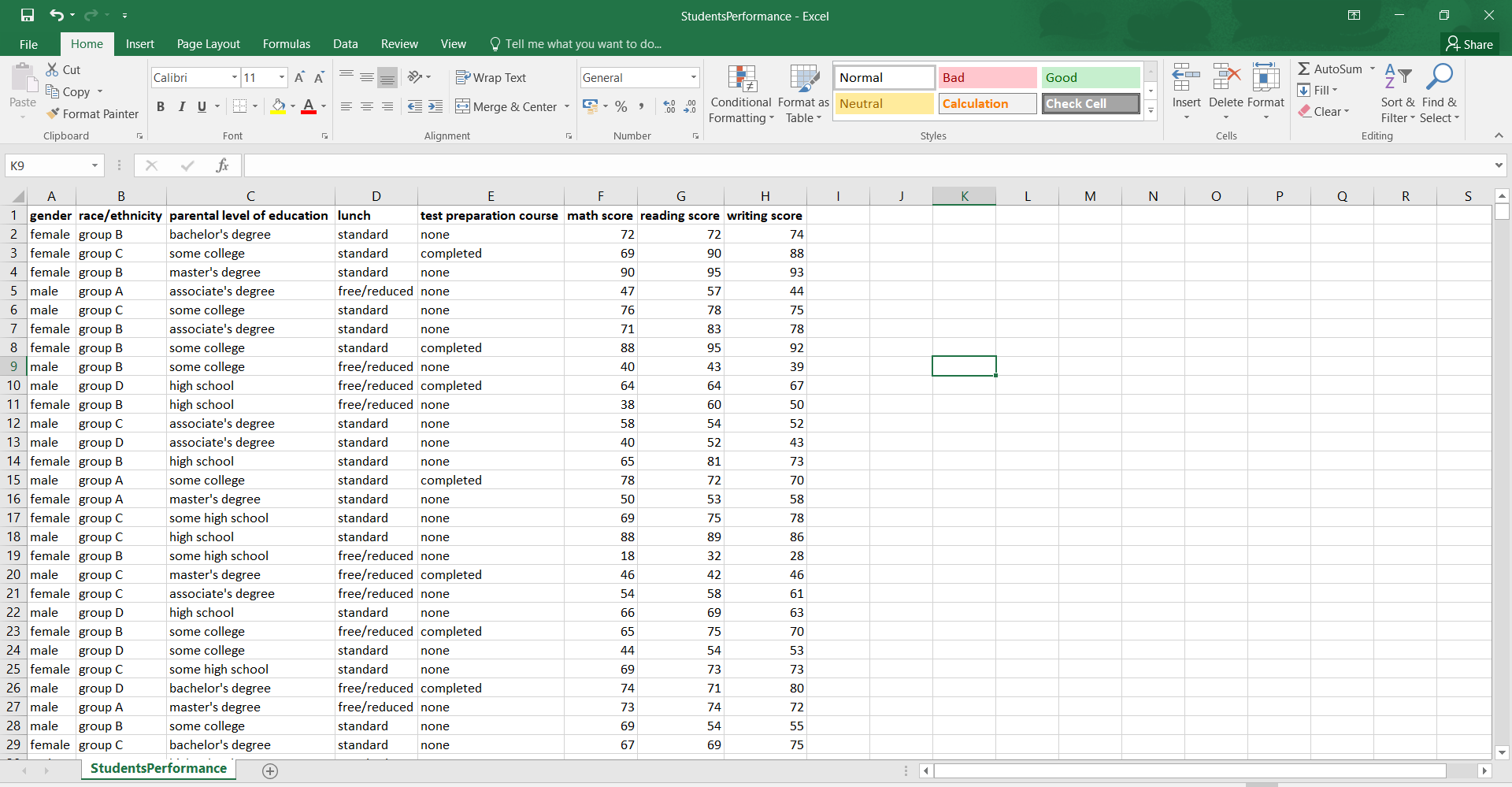
**Dataset Source -**<https://www.kaggle.com/datasets/spscientist/students-performance-in-exams?datasetId=74977>

The data consists of 8 column and 1000 rows.

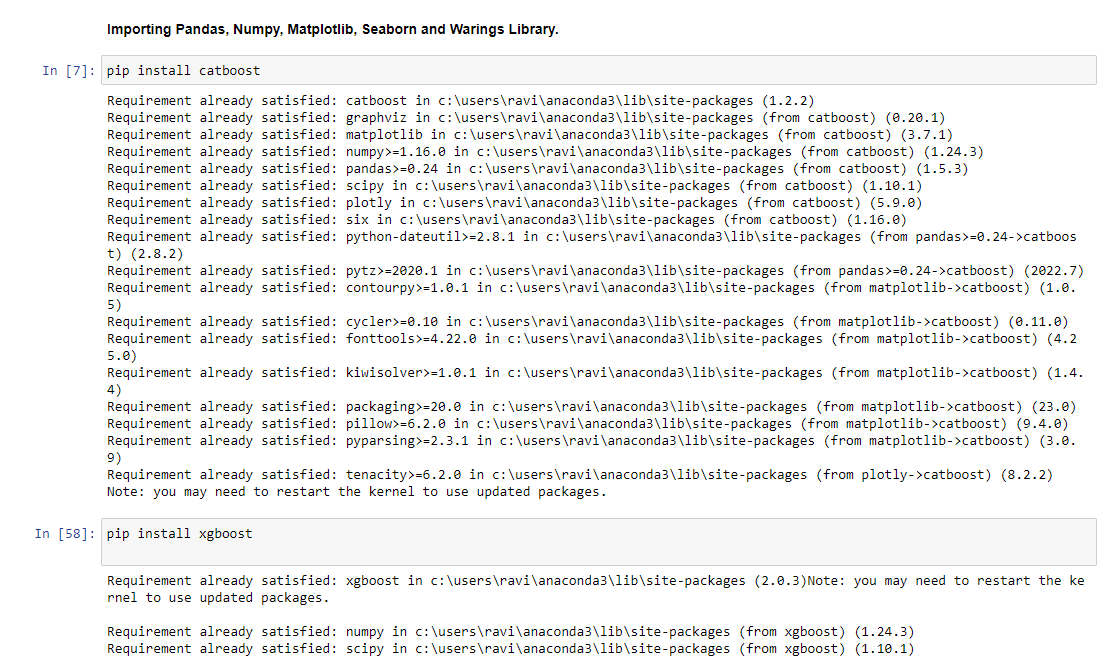
### Dataset information

* gender : sex of students -> (Male/female)
* race/ethnicity : ethnicity of students -> (Group A, B,C, D,E)
* parental level of education : parents' final education ->(bachelor's degree, some college, master's degree, associate's degree, high school)
* lunch : having lunch before test (standard or free/reduced)
* test preparation course : complete or not complete before test
* math score
* reading score
* writing score

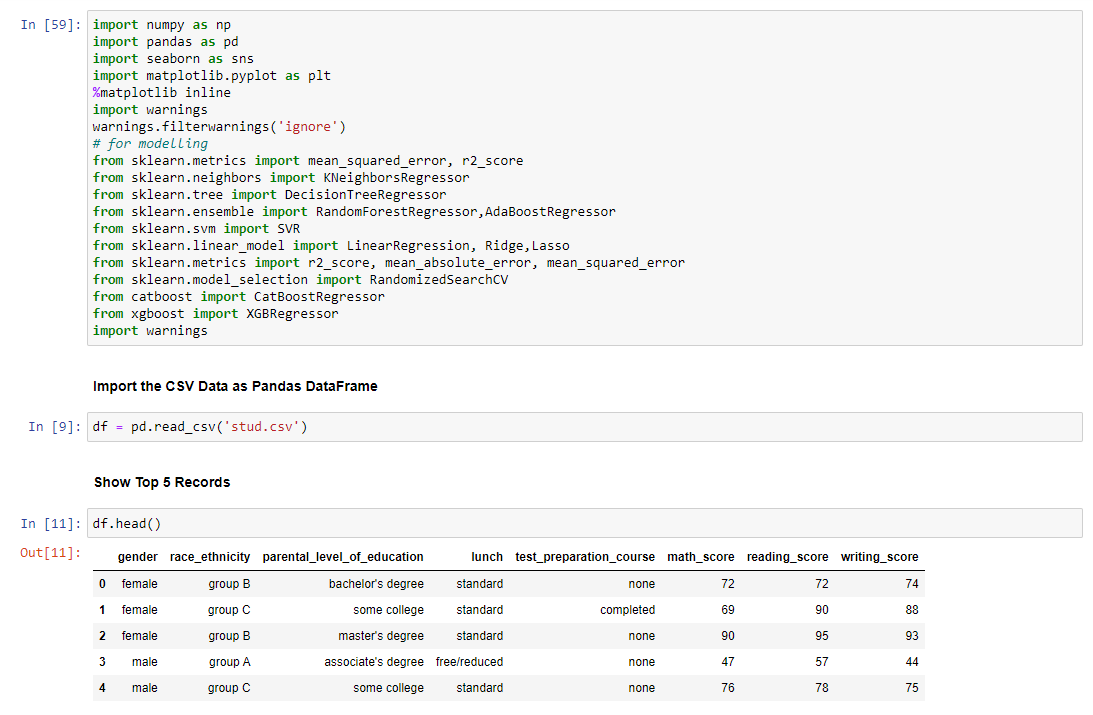
Data is collected over multiple academic years, allowing for longitudinal analysis of student performance and behavior.



*Fig 3.1 Source dataset*



*Fig 3.2 Python library Import*



*Fig 3.3 Dataset record*

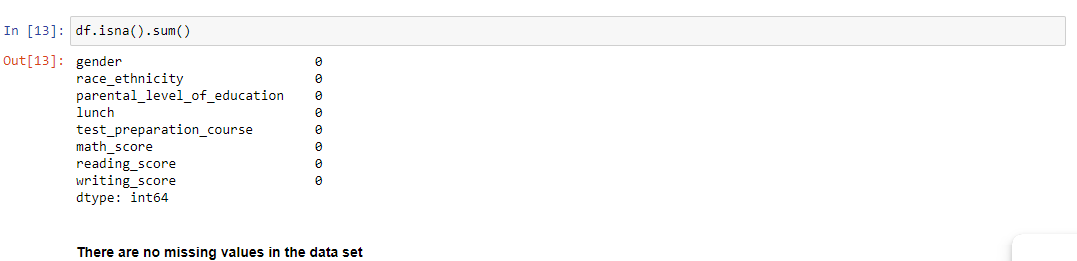
### Data Pre-Processing Steps

#### Data Cleaning and transformation

The dataset undergoes cleaning to address missing values, outliers, and inconsistencies.

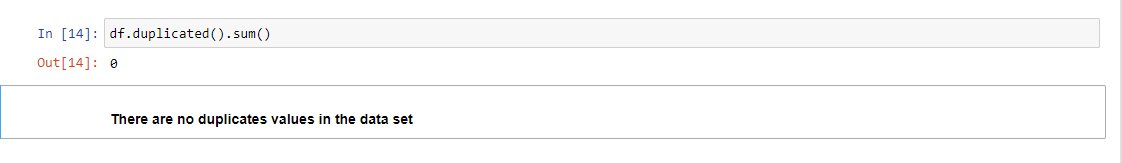
For instance

* Check Missing values
* Check Duplicates
* Check data type
* Check the number of unique values of each column
* Check statistics of data set
* Check various categories present in the different categorical column



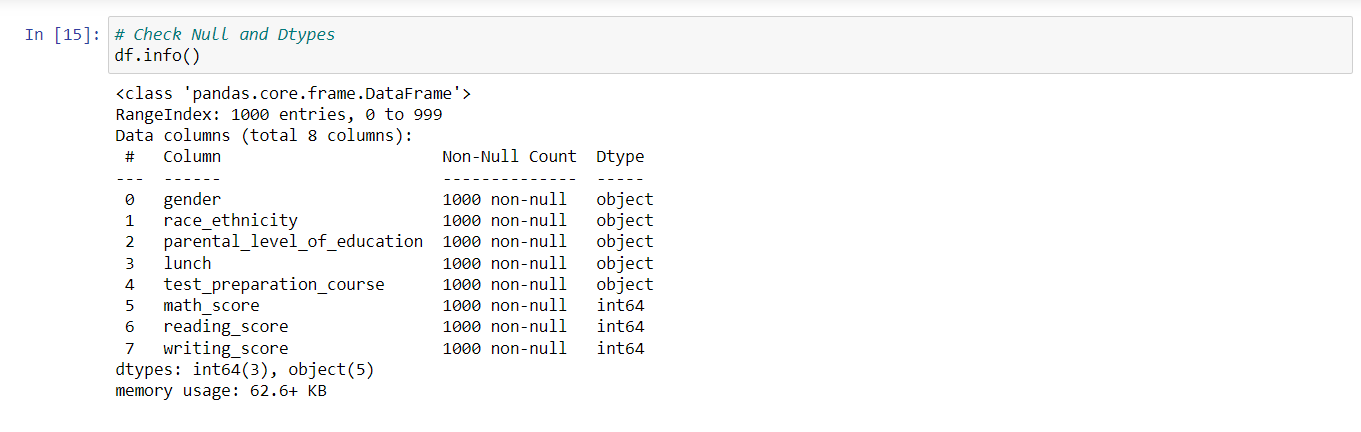
*Fig 3.4 Missing values Count*

In the dataset the feature values consists of Gender, race Ethnicity, Parental level of education, lunch, test preparation course. It also include the student scores in total three categories among math-score, writing- score and reading -scores. Here we can see that the missing value score is **0**. There for the dataset having complete values without any missing feature.



*Fig 3.5 Duplicate values Count*

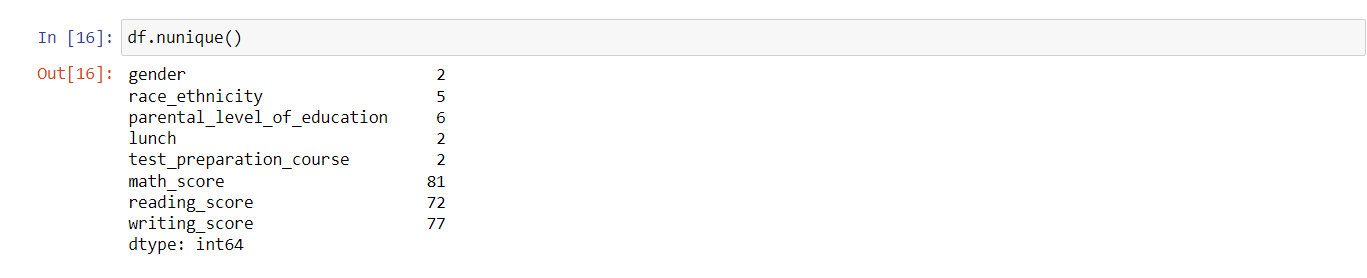
In the dataset the feature values consists of Gender, race Ethnicity, Parental level of education, lunch, test preparation course. It also include the student scores in total three categories among math-score, writing- score and reading -scores. Here we can see that the duplicate value score is **0**. There for the dataset having complete values without any repeating feature.



*Fig 3.6 Null data Count*

In the dataset the feature values consists of Gender, race Ethnicity, Parental level of education, lunch, test preparation course. It also include the student scores in total three categories among math-score, writing- score and reading -scores. Here we can see that the Null value score is **0**.

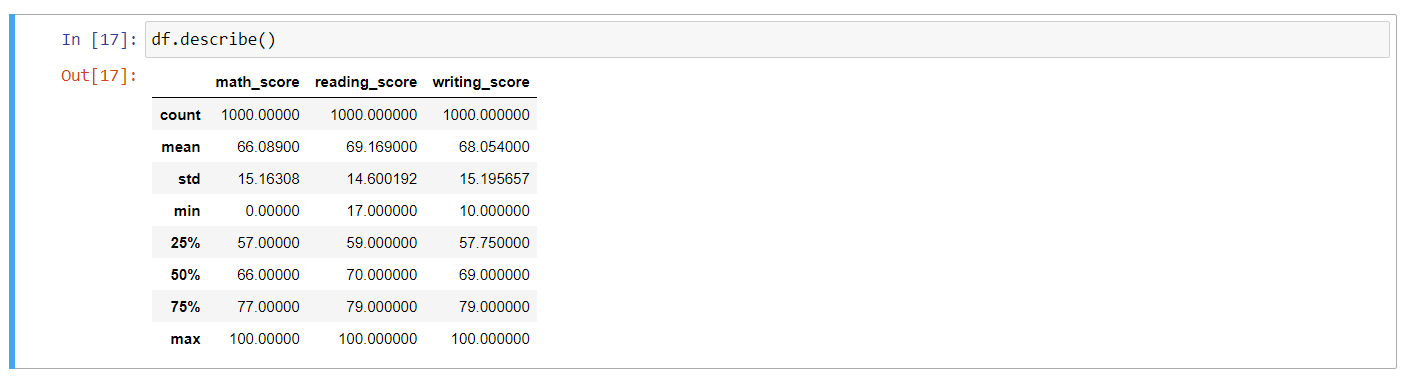
Here the dataset consists of two types of data. The feature Gender, race Ethnicity, Parental level of education, lunch and test preparation course is of “object type” while math-score, writing- score and reading –scores is of “ Integer type”.



*Fig 3.7 Unique value Count*

In the dataset the feature values consists of various unique values. I.E Gender has 2 and race Ethnicity has 5 unique values. Same way Parental level of education has 6, lunch and test preparation course has 2 unique values. It also include the student scores in total three categories among math-score, writing- score and reading -scores. Here we can see, that it consists of highest unique values as

Respectively.



*Fig 3.8 Statistical Data Count*

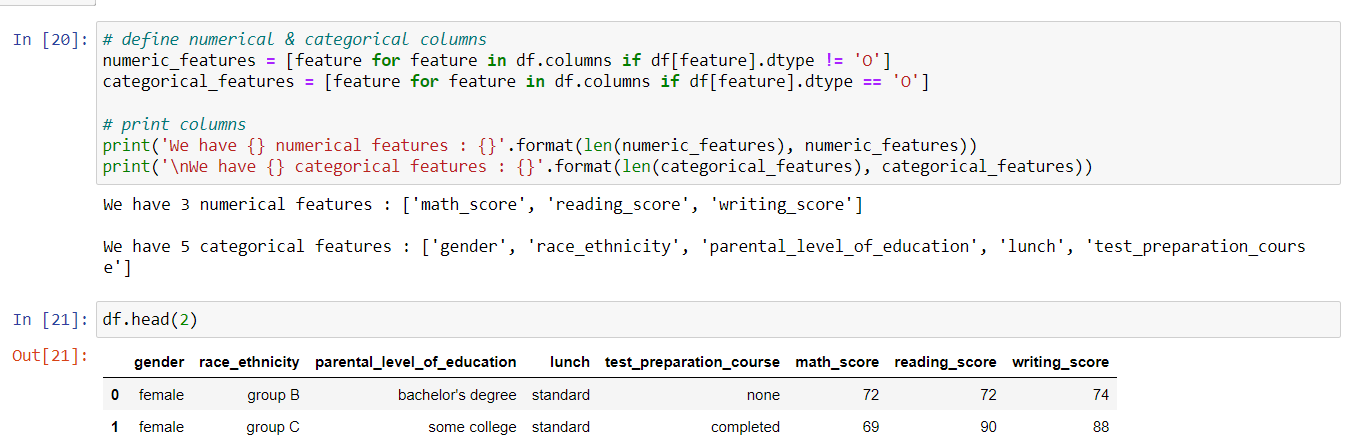
**Summary:**

* From above description of **numerical data**, all means are very close to each other - between **66 and 68.05.**
* All **standard deviations** are also close - **between 14.6 and 15.19.**
* While there is a **minimum score 0 for math**, for **writing minimum is much higher = 10** and for **reading met higher = 17**

**Transformation:** Certain variables may require transformation to facilitate analysis. For example, categorical variables such as race/ethnicity may be encoded using one-hot encoding for compatibility with machine learning algorithms.



*Fig 3.9 categorical Data Count*



*Fig 3.10 Numerical- categorical Data Count*

**Aggregation:** Data aggregation may be performed to summarize information at different levels of granularity. For instance, daily attendance records may be aggregated to calculate average absenteeism rates for each student over the academic year.

****

*Fig 3.10 Numerical- categorical Data Count*

The above data summaries the followings.

* There are total **7 students** who gets **full marks in maths.**
* There are total **14 students** who gets **full marks in writing.**
* There are total **17 students** who gets **full marks in reading.**
* If we compare the performance of the students then students are mainly **weak in maths** compare to writing and reading.

**Chapter 4**

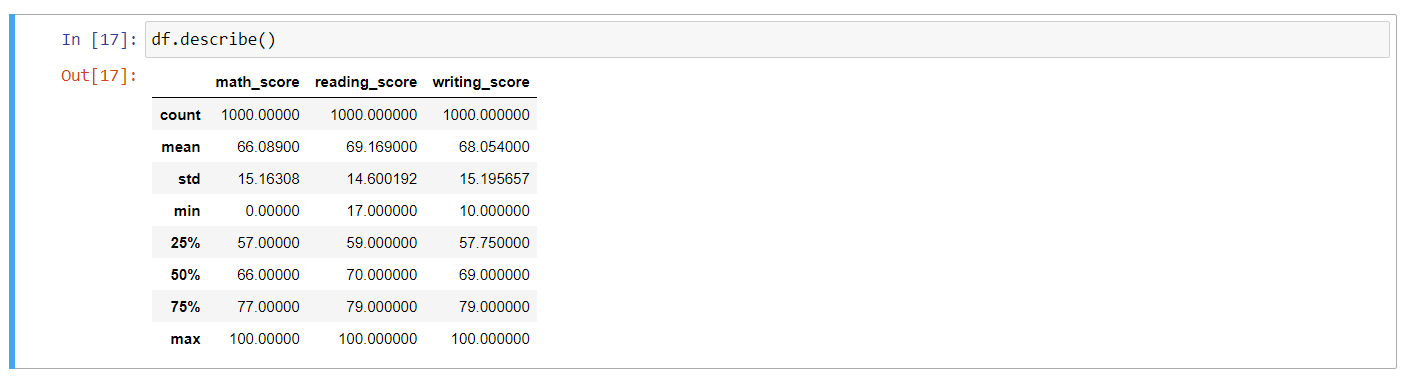
# Exploratory data Analysis

### Summary statistics of the dataset

#### The main objective of doing statistical analysis is to identify the relation in existing data features. Basically, the statistics consists of mean, max, count and Standard deviation count that gives the best result.

#### In the dataset the feature values includes math-score, reading-score and writing-score. Regardless of which strategies the analysts follow, there are preferred practices to ensure that the statistical analysis must be useful. All standard deviations are also close - between 14.6 and 15.19.The levels of tests and types of test data, combined with testing libraries are important aspects of the actual test process. From above description of numerical data, all mean are very close to each other - between 66 and 68.05.Among the various testing practices or strategies that are followed by analysts, the two important ones are unit testing and system testing. While there is a minimum score 0 for math, for writing minimum is much higher = 10 and for reading met higher = 17.

#### Here the writing-score has maximum Standard deviation (SD) of 15.195657. While the reading-score has 14.600192 and the math-score contains the least SD among all, 15.16308.



*Fig 4.1 Summary Statistical Data Count*

### Data Visualization

There are quite a few outcomes and takeaways that are there after going through all the heavy development work. Primary results of having such a solution would be:

Visualize average score distribution to make some conclusion

- Histogram

- Kernel Distribution Function (KDE)

### Histogram and KDE

### The below snippet creates a visual comparison of the distribution of 'average' values from the Data Frame with one subplot showing a single histogram and KDE curve and the other subplot showing separate histograms and KDE curves for each gender.

### 

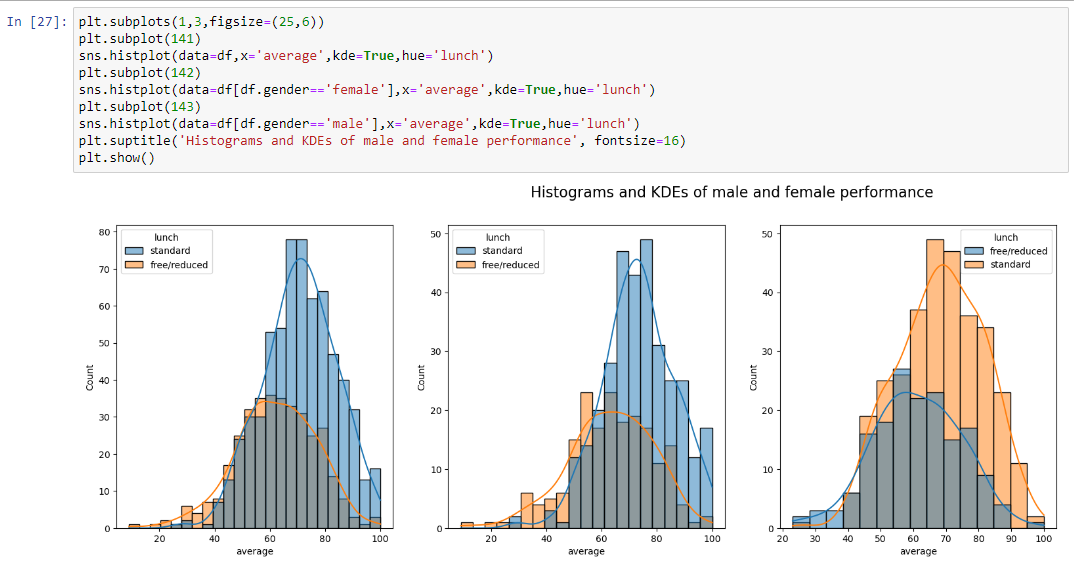
*Fig 4.2 Histogram and KDE of Average values*

The below snippet creates a visual comparison of the distribution of ' total score' values from the Data Frame with one subplot showing a single histogram and KDE curve and the other subplot showing separate histograms and KDE curves for each gender.

### *Fig 4.3 Histogram and KDE of total score*

The below snippet creates a visual comparison of the distribution of ' male-female performance' values from the Data Frame with one subplot showing a single histogram and KDE curve and the other subplot showing separate histograms and KDE curves for each gender. The result shows that Female students tend to perform well then male students.

### Here we summaries that the standard lunch intake helps to perform well in exams. As a result there is a strong relation between Lunch feature and performance. We can see that it varies between males and females.



*Fig 4.4 Histogram and KDE of male and female performance*

The below snippet creates a visual comparison of the distribution of ' male-female wise parental education ' values from the Data Frame with one subplot showing a single histogram and KDE curve and the other subplot showing separate histograms and KDE curves for each gender. The result shows that Female students tend to perform well then male students.

##### Insights

* In general parent's education don't help student perform well in exam.
* 2nd plot shows that parent's whose education is of associate's degree or master's degree their male child tend to perform well in exam
* 3rd plot we can see there is no effect of parent's education on female students.

### 

*Fig 4.5 Histogram and KDE of parental education male and female wise*

The below snippet creates a visual comparison of the distribution of ' male-female wise parental education ' values from the Data Frame with one subplot showing a single histogram and KDE curve and the other subplot showing separate histograms and KDE curves for each gender. The result shows that Female students tend to perform well then male students.

##### Insights

* Students of group A and group B tends to perform poorly in exam.
* Students of group A and group B tends to perform poorly in exam irrespective of whether they are male or female

### 

*Fig 4.6 Histogram and KDE of race ethnicity male and female wise*

### Maximum score of the students in all three subjects

### This code snippet creates a figure with four subplots arranged horizontally, each displaying a violin plot for different scores (math, reading, and writing) from a DataFrame df. In summary, this code visually compares the distribution of math, reading, and writing scores using violin plots, with each type of score represented in a separate subplot within the same figure.

### 

### *Fig 4.7 violin plot of math, reading and writing score*

**Insights**

* From the above three plots its clearly visible that most of the students score in between 60-80 in Maths whereas in reading and writing most of them score from 50-80.

### Identification of patterns and Correlations

### This code snippet creates a figure with five subplots arranged horizontally, each displaying a pie chart representing the distribution of different categorical variables from a Data Frame. This code visually represents the distribution of various categorical variables (gender, race/ethnicity, lunch type, test preparation course completion, parental level of education) using pie charts, with each variable represented in a separate subplot within the same figure.

##### Insights

* Number of Male and Female students is almost equal
* Number students are greatest in Group C
* Number of students who have standard lunch are greater
* Number of students who have not enrolled in any test preparation course is greater
* Number of students whose parental education is "Some College" is greater followed closely by "Associate's Degree"

### C:\Users\Ravi\Desktop\Identification of patterns and correlations.png

### *Fig 4.8 Identification of patterns and Correlations*

**Chapter 5**

# Feature Engineering

### Selection and engineering of relevant features

### By carefully engineering features, we can improve the performance of machine learning model and make it more robust to variations in the data. Feature engineering is often an iterative process, where we experiment with different transformations and combinations of features to find the most effective representation for your problem. Here by considering the following list of attributes which helps in optimum ML model design.

### 5.1.1 Gender Column

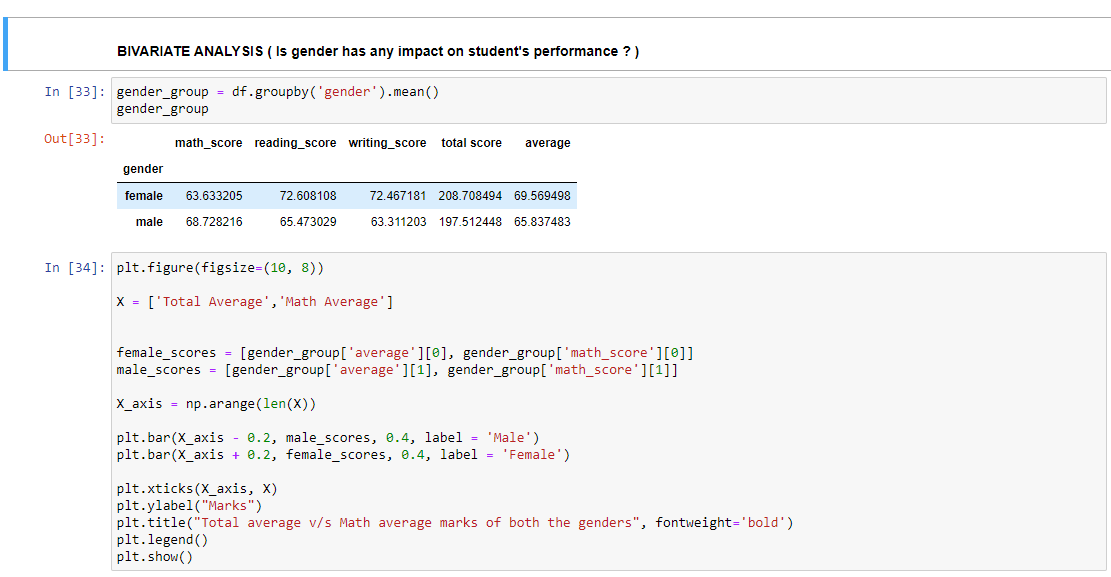
The main selection of gender feature is to identify two major things. The first most thing is how the distribution of Gender exist? The second key aspect is to check the relation between the student performance Vs Gender identification.

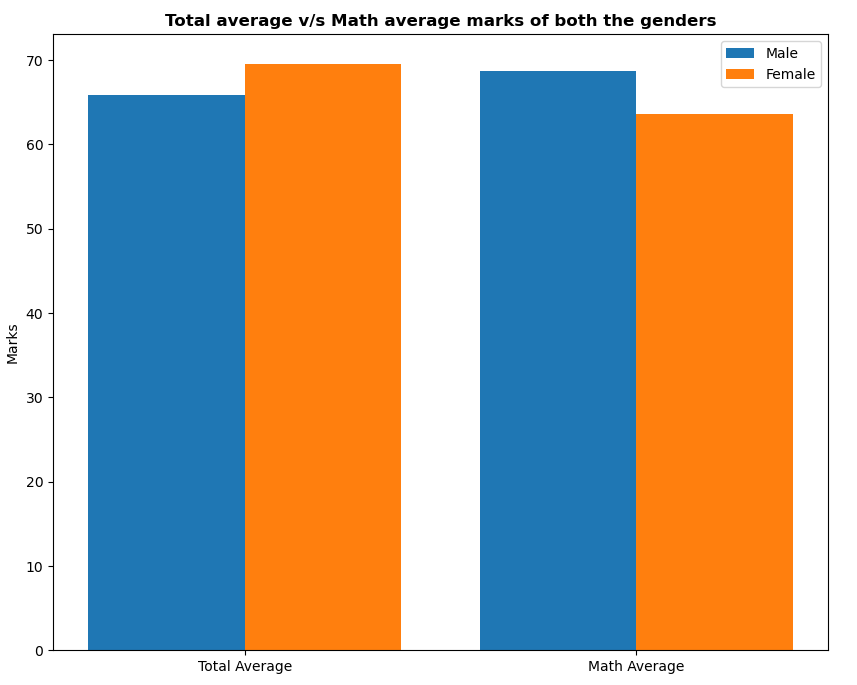


*Fig 5.1 Gender Distribution*

#### Insights

* Gender has balanced data with female students are 518 (48%) and male students are 482 (52%).



**

*Fig 5.2 Marks Distribution*

#### Insights

* On an average females have a better overall score than men.
* Whereas males have scored higher in Maths.

### 5.1.2 Race / Ethnicity Column

### 

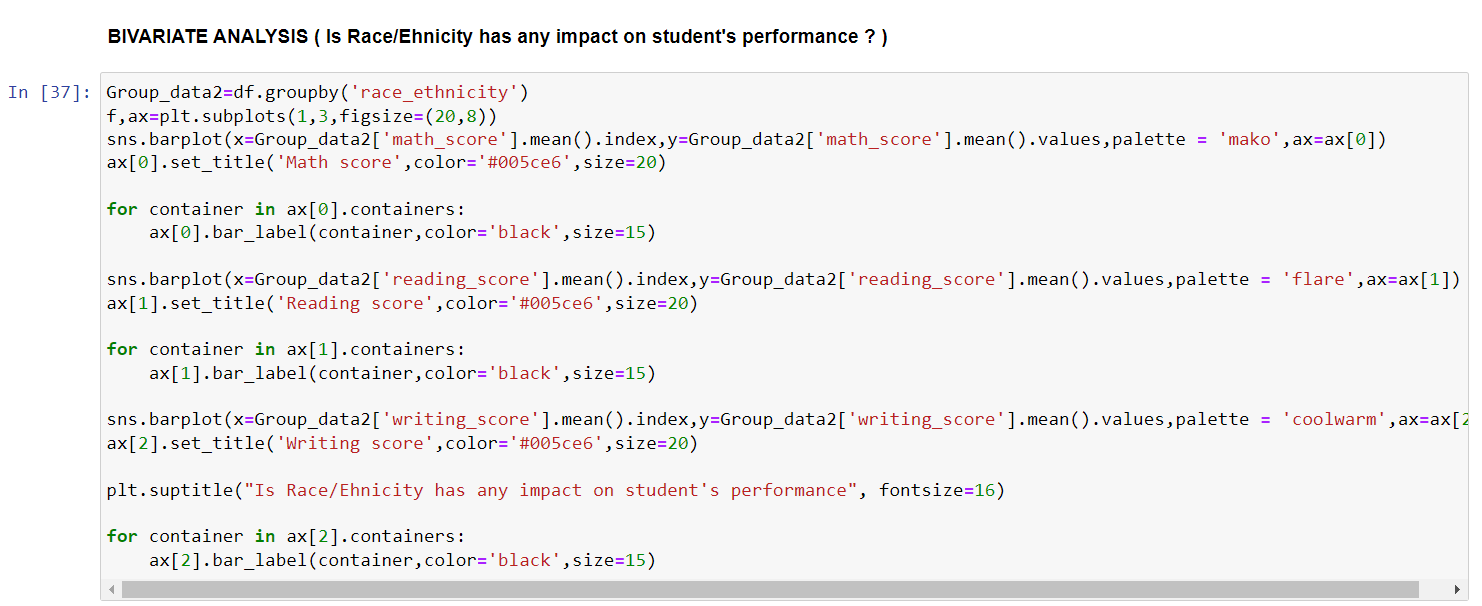
### The main selection of gender feature is to identify two major things. The first most thing is how the distribution in group exist? The second key aspect is to check the relation between the race/ethnicity Vs student performance.

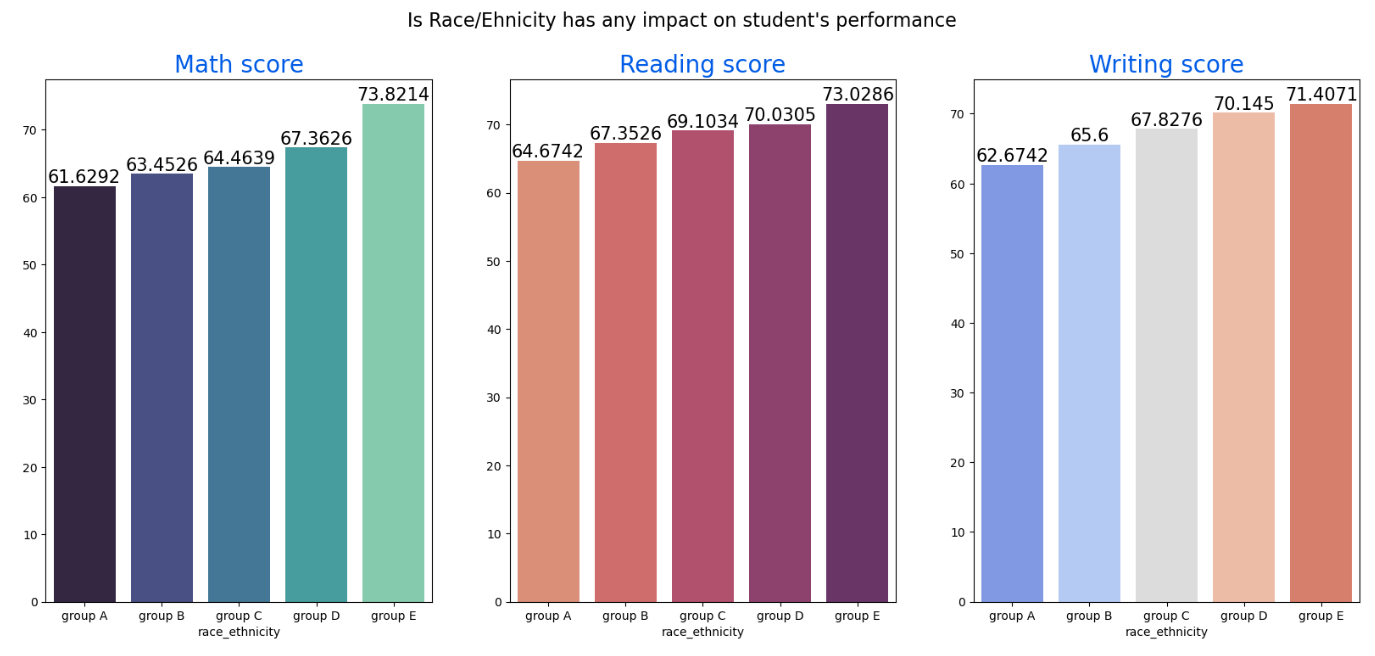
### 

*Fig 5.3 Group wise Distribution*

#### Insights

* Most of the student belonging from group C /group D.
* Lowest number of students belong to group A.



****

*Fig 5.4 Race/Ethnicity Impact*

#### Insights

* Group E students have scored the highest marks.
* Group A students have scored the lowest marks.
* Students from a lower Socioeconomic status have a lower avg in all course subjects

### 5.1.3 Parental level of Education Column

### The main selection of parental level education feature is to identify two major things. The first most thing is how the parental educational background in group exist? The second key aspect is to check the relation between the parental educational background Vs student performance.

### 

### 

*Fig 5.5 Parental Education Impact*

#### Insights

* The score of student whose parents possess master and bachelor level education are higher than others.

### 5.1.4 Lunch Column

### The main selection of Lunch intake feature is to identify two major things. The first most thing is which type of lunch is common among the students? The second key aspect is to check the relation between the lunch type Vs student performance and the result.

### 

*Fig 5.6 Test Preparation Course*

#### Insights

* Students being served Standard lunch was more than free lunch
* Students who get Standard Lunch tend to perform better than students who got free/reduced lunch.

### Techniques used for feature selection and extraction

### 5.2.1 Test preparation course column

### The main selection of test preparation feature is to identify two major things. The first most thing is which type of lunch is common among the students? The second key aspect is to check the relation between the test preparation Vs student performance and the result.

### 

*Fig 5.7 Test preparation Impact*

#### Insights

* Students who have completed the Test Preparation Course have scores higher in all three categories than those who haven't taken the course

### 5.2.2 Outlier Detection

### 

*Fig 5.8 Outlier Detection*

### 5.2.3 Pair-plot Analysis

### 

### 

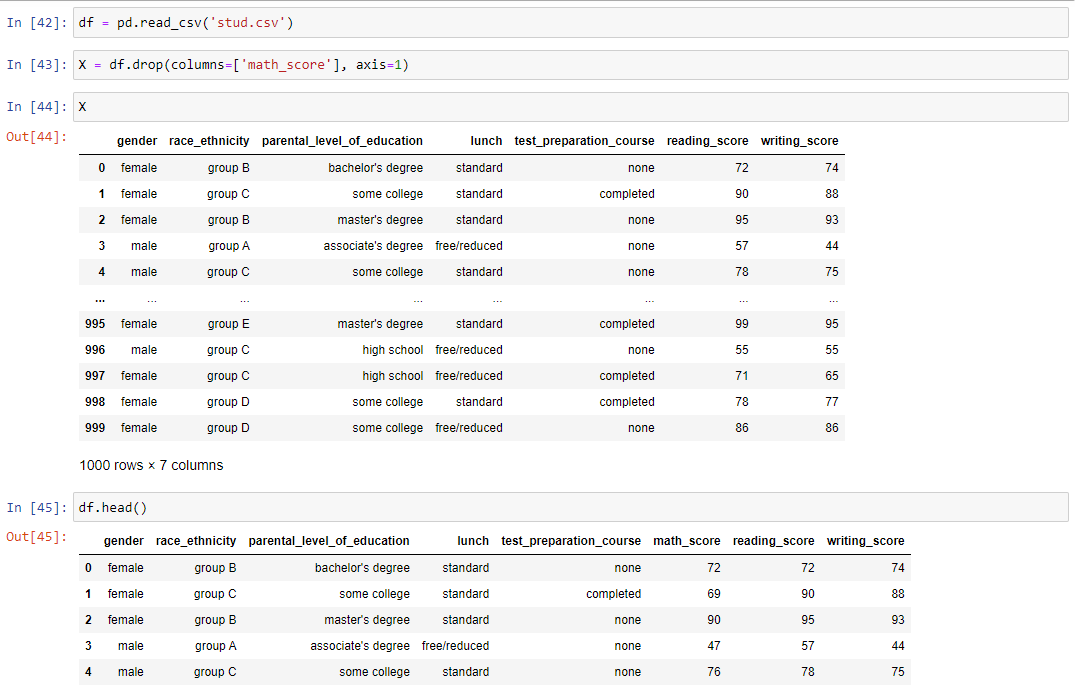
#### Insights

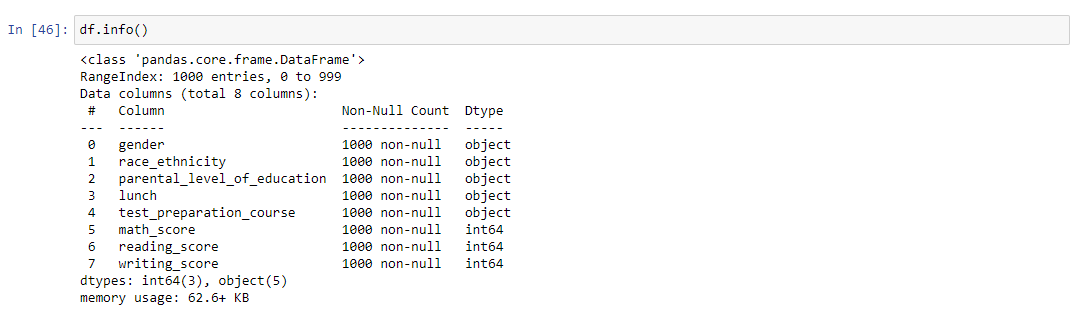
* From the above plot it is clear that all the scores increase linearly with each other.
* Student's Performance is related with lunch, race, parental level education
* Females lead in pass percentage and also are top-scorers
* Student's Performance is not much related with test preparation course
* Finishing preparation course is beneficial.

**Chapter 6**

# Machine Learning model Development

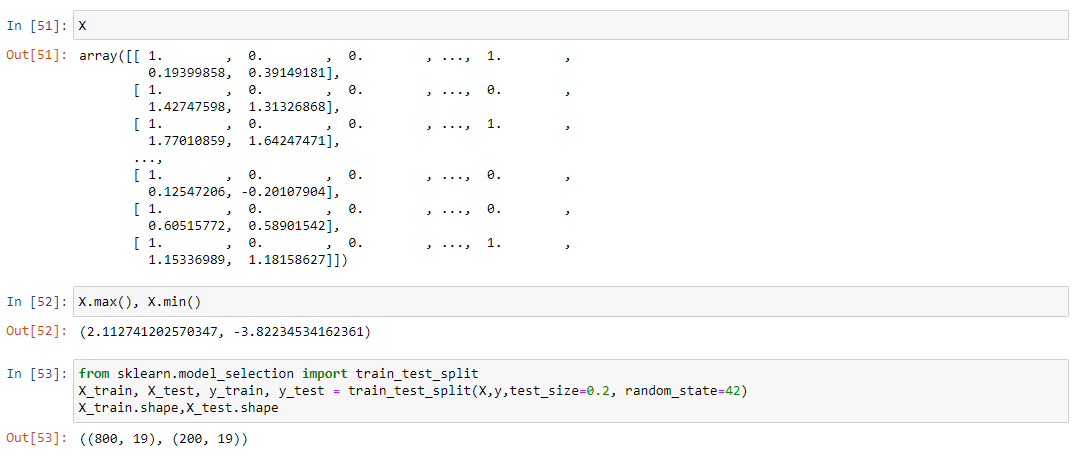
### 6.1 Description of the machine learning algorithm Used





### 6.2 Model selection criteria





### 6.3 Model Training and validation process

Before building models, we preprocess the data, splitting it into training and testing sets.

****

**Chapter 7**

# Performance Evaluation

### Evaluation Metric Used

The **Linear Regression** model is selected as the final model based on its performance metrics on both the training and testing sets:

* Root mean squared error: 5.3940
* Mean absolute error: 4.2148
* R2 score: 0.8804

### 

### Comparative analysis of different models

**Linear Regression**

Model performance for Train set:

* Root mean squared error: 5.3231
* Mean absolute error: 4.2667
* R2 score: 0.8743

Model performance for Test set:

* Root Mean squared error: 5.3940
* Mean absolute error: 4.2148
* R2 score: 0.8804 ===================================

**Lasso**

Model performance for Train set:

* Root mean squared error: 6.5938
* Mean absolute error: 5.2063
* R2 score: 0.8071

Model performance for Test set:

* Root Mean squared error: 6.5197
* Mean absolute error: 5.1579
* R2 score: 0.8253 ===================================

**Ridge**

Model performance for Train set:

* Root mean squared error: 5.3233
* Mean absolute error: 4.2650
* R2 score: 0.8743

Model performance for Test set:

* Root Mean squared error: 5.3904
* Mean absolute error: 4.2111
* R2 score: 0.8806 ===================================

**K-Neighbors Regressor**

Model performance for Train set:

* Root mean squared error: 5.7091
* Mean absolute error: 4.5175
* R2 score: 0.8554

Model performance for Test set:

* Root Mean squared error: 7.2583
* Mean absolute error: 5.6370
* R2 score: 0.7835 ===================================

**Decision Tree**

Model performance for Train set:

* Root mean squared error: 0.2795
* Mean absolute error: 0.0187
* R2 score: 0.9997

Model performance for Test set:

* Root Mean squared error: 8.2592
* Mean absolute error: 6.4750
* R2 score: 0.7197 ===================================

**Random Forest Regressor**

Model performance for Train set:

* Root mean squared error: 2.2728
* Mean absolute error: 1.8171
* R2 score: 0.9771

Model performance for Test set:

* Root Mean squared error: 6.0423
* Mean absolute error: 4.6931
* R2 score: 0.8500 ===================================

**XGBRegressor**

Model performance for Train set:

* Root mean squared error: 1.0073
* Mean absolute error: 0.6875
* R2 score: 0.9955

Model performance for Test set:

* Root Mean squared error: 6.4733
* Mean absolute error: 5.0577
* R2 score: 0.8278 ===================================

**CatBoosting Regressor**

Model performance for Train set:

* Root mean squared error: 3.0427
* Mean absolute error: 2.4054
* R2 score: 0.9589

Model performance for Test set:

* Root Mean squared error: 6.0086
* Mean absolute error: 4.6125
* R2 score: 0.8516 ===================================

**AdaBoost Regressor**

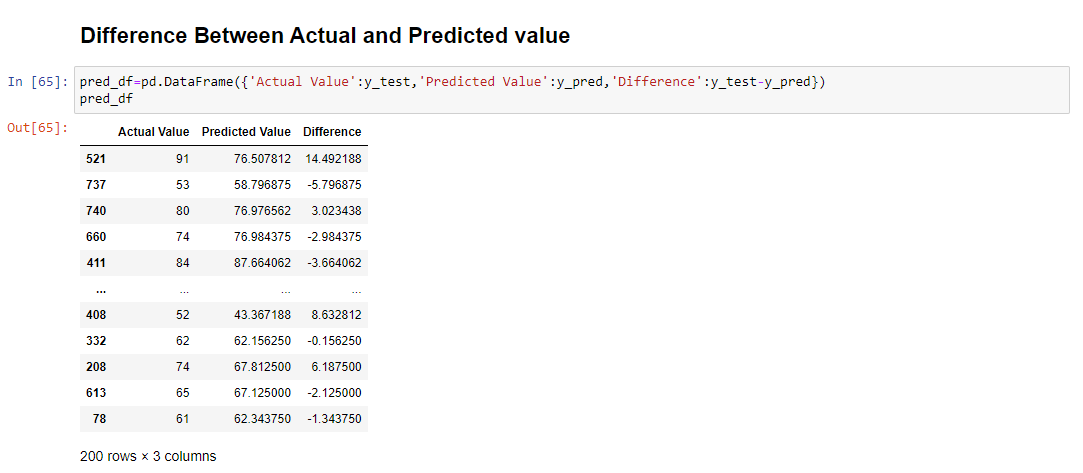
Model performance for Train set:

* Root mean squared error: 5.7847
* Mean absolute error: 4.7346
* R2 score: 0.8516

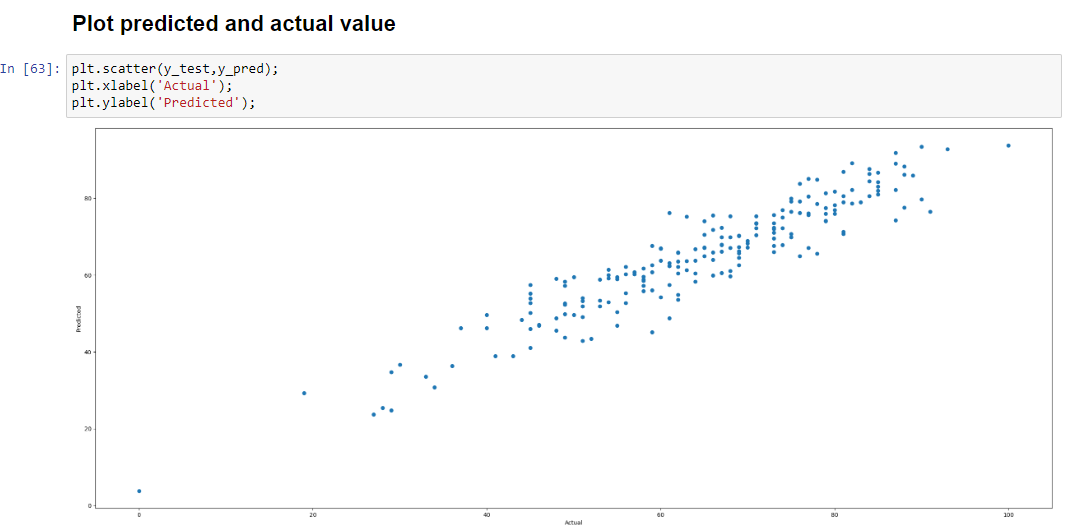
Model performance for Test set:

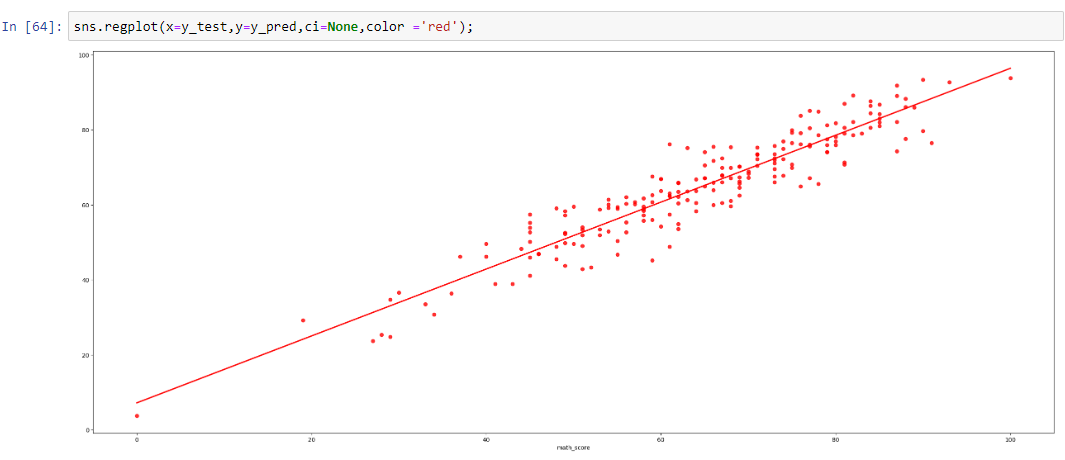
* Root Mean squared error: 5.9128
* Mean absolute error: 4.5969
* R2 score: 0.8563 ===================================

### Interpretation of the results

****

****

****

****

*Fig 7.1 Plot Prediction value*

**Chapter 8**

# Conclusion

### 8.1 Summary

### This end-to-end project aims to predict student performance using machine learning techniques. By analyzing a dataset containing various student attributes and academic details, we strive to gain insights into factors influencing academic success and build a predictive model.

### 8.2 Comparison with Existing Solutions

There are similar applications available: many of them having very less functionalities, some of them having a lot of bugs, lot of them not being managed anymore, quite a few of them being limited to newer versions of android, a number of them with cluttered and complex UI, and a very infinitesimal number delivering what is promised.

Our research suggests that there is no existing solution available specifically for this university that performs all the operations which our product promises. Although there are handful of application for other universities which are supposed to be quite similar to this product, but every one of them have their own drawbacks and limitations. We are trying to take the best features from various existing solutions and adding few unique features of our own to create a product that makes the life of the user easier.

While we accept that few of the features and functionalities that are provided in the product have already been implemented in various products independently, it is equally important to note that our product is a mixture of those features with quite a few added unique features and functionalities.

The product is designed by keeping the students in the centre. But while the app may seem to be limited for students, it is just as equally useful for the faculties, principal, HODs, administrative staff, parents and guardians.

### 8.3 Future Enhancements

It is absolutely absurd to concede that a product is perfect, complete or finished in the computer world. The evolution constantly takes place no matter what we do. *“There’s always a way to do it better - find it!”* – the simple quote from Thomas Alva Edison seems very appropriate for the invariably advancing world of programming.

We always have to keep in mind the features that can be added in order to raise the level of the application, and the methods that can be improved upon to make the application, as a whole, a lot faster than it currently is.

### The project extends to deployment using Flask application and Docker images, with the model deployed on an EC2 instance.

# References

1. EDUCAUSE Research: <http://net.educause.edu/ir/library/pdf/ERB1210.pdf>
2. Z. Yue and Y. Jin, "The Development and Design of the Student Management System Based on the Network Environment," 2010 International Conference on Multimedia Communications, Hong Kong, 2010, pp. 5-8. doi: 10.1109/MEDIACOM.2010.23
3. S. A. M. Noor, N. Zaini, M. F. A. Latip and N. Hamzah, "Android-based attendance management system," 2015 IEEE Conference on Systems, Process and Control (ICSPC),

Bandar Sunway, 2015, pp. 118-122. doi: 10.1109/SPC.2015.7473570

1. J. Mei-shan, Q. Chang-li and L. Jing, "The designment of student information management system based on B/S architecture," 2012 2nd International Conference on Consumer Electronics, Communications and Networks (CECNet), Yichang, 2012, pp. 2153-2155. doi: 10.1109/CECNet.2012.6201529
2. Y. Shulin and H. Jieping, "Research and implementation of Web Services in Android network communication framework Volley," 2014 11th International Conference on Service Systems and Service Management (ICSSSM), Beijing, 2014, pp. 1-3. doi: 10.1109/ICSSSM.2014.6943373
3. F. Yue, "A study of student information management software," 2016 IEEE International Conference of Online Analysis and Computing Science (ICOACS),

Chongqing, 2016, pp. 393-396. doi: 10.1109/ICOACS.2016.7563123

1. L. Meng, "College Student Management System Design Using Computer Aided System," 2015 International Conference on Intelligent Transportation, Big Data and Smart City, Halong Bay, 2015, pp. 212-215. doi: 10.1109/ICITBS.2015.59
2. A. Alshareef, A. Alkilany, M. Alweshah and A. A. Bakar, "Toward a student information system for Sebha University, Libya," Fifth International Conference on the Innovative Computing Technology (INTECH 2015), Pontevedra, 2015, pp. 34-39. doi: 10.1109/INTECH.2015.7173362

# Appendix