Project Report

Khusroo Kalim Ansari: **300347313**

**What have influences on your exam scores?**

There are many factors that can lead to one achieving a good exam score like number hours one studies, the amount of preparation, prior experience or knowledge about the subject and many more such factors that can lead to a better exam score.

In our analysis we have also tried to predict such observation using a dataset called Students Performance from Kaggle https://[www.kaggle.com/datasets/sonukumari47/students-performance-in-exams.](http://www.kaggle.com/datasets/sonukumari47/students-performance-in-exams)

**Introduction**

* We have chosen the Student Performance dataset taken from Kaggle.
* The dataset contains marks obtained by students in college from the United States.
* The goal of our analysis is to predict variables that provides us with better prediction for gender identification i.e., which gender is likely to perform better in the given test score

The focus of our prediction is based on whether the gender is male or female and identifying features which may be more relatable to the gender classification.

Note: This analysis is not to build a perception or assumption about any particular gender it is to just show how machine learning can be used to analyze data, and datasets can be of a certain sample size which do not include all the population and scenarios and judging a one’s capabilities of doing well in any exam based on their gender is not a very accurate depiction of one’s characteristics and capabilities.

**Data Preparation**

The name of the dataset used in this project is “Students Performance in Exams” and it was obtained from Kaggle. This is the link to the data.

https://[www.kaggle.com/datasets/sonukumari47/students-performance-in-exams.](http://www.kaggle.com/datasets/sonukumari47/students-performance-in-exams)

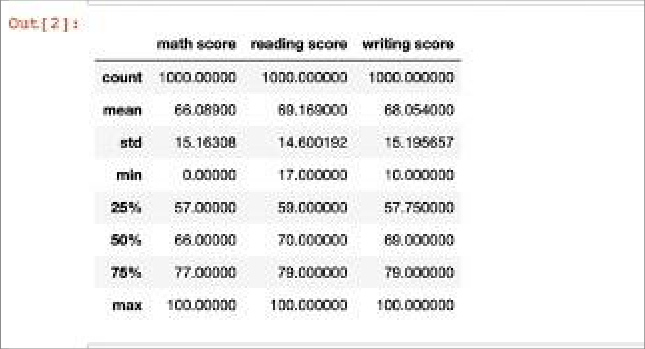
The dataset has 8 columns named as

1. Gender – Describes the gender of student
2. Ethnicity- Ethnicity of the student
3. Parental education – Level of parental education
4. Lunch – Type of lunch student took before the exam
5. Test preparation- Whether the student prepared for exam and took the preparation test
6. Math score – Score in math
7. Reading score - Score in reading
8. Writing score- Score in writing

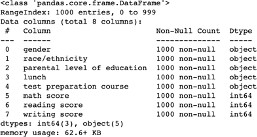
## Data Processing

We used various functions to understand our dataset in detail and some of them are as follows.

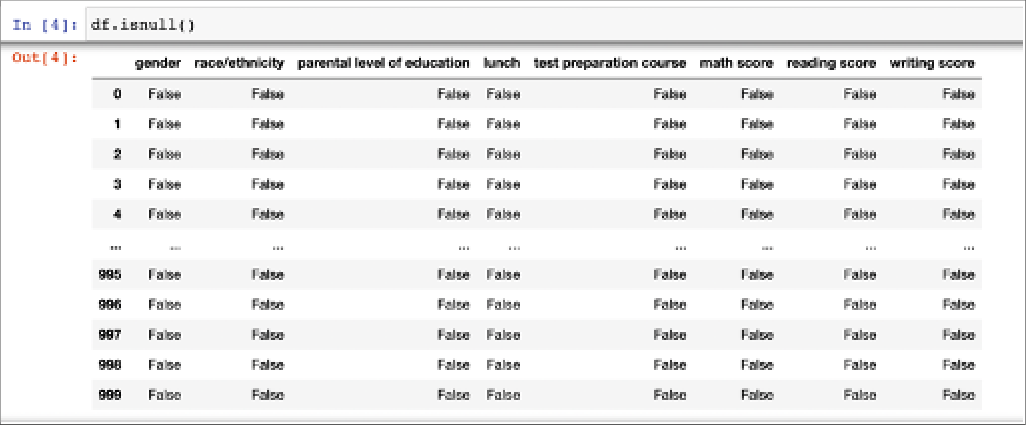
df.describe(): to learn about the statistics of the dataset



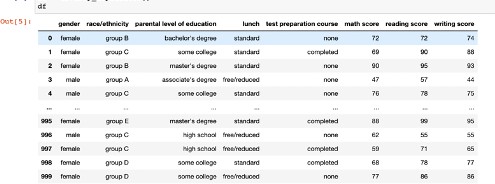
* df.info (): to check the data types of datasets



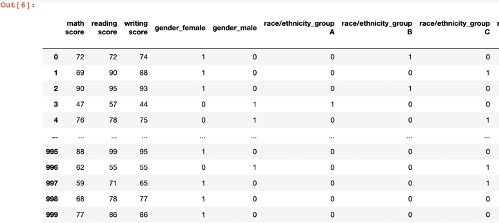
* df.isnull(): to check null values



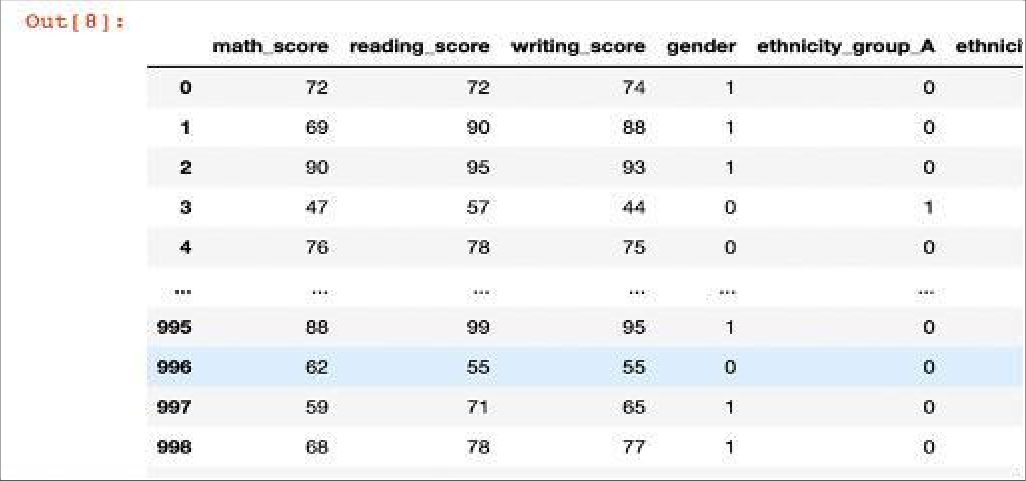
* df.drop\_duplicates() to drop the duplicate values from the dataset



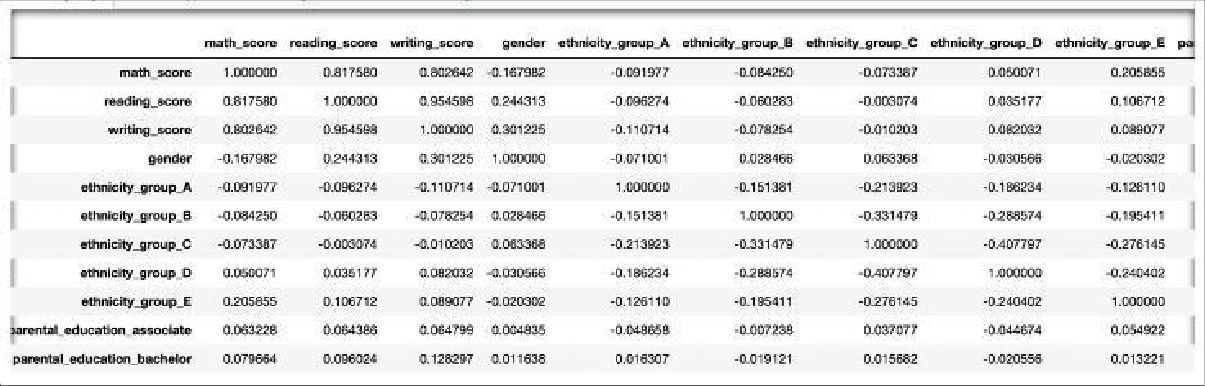
* pd.get\_dummies(df) as our dataset contains some categorical values, we created dummies



* df.rename(): to rename some of the columns in the dataset for easy interpretation

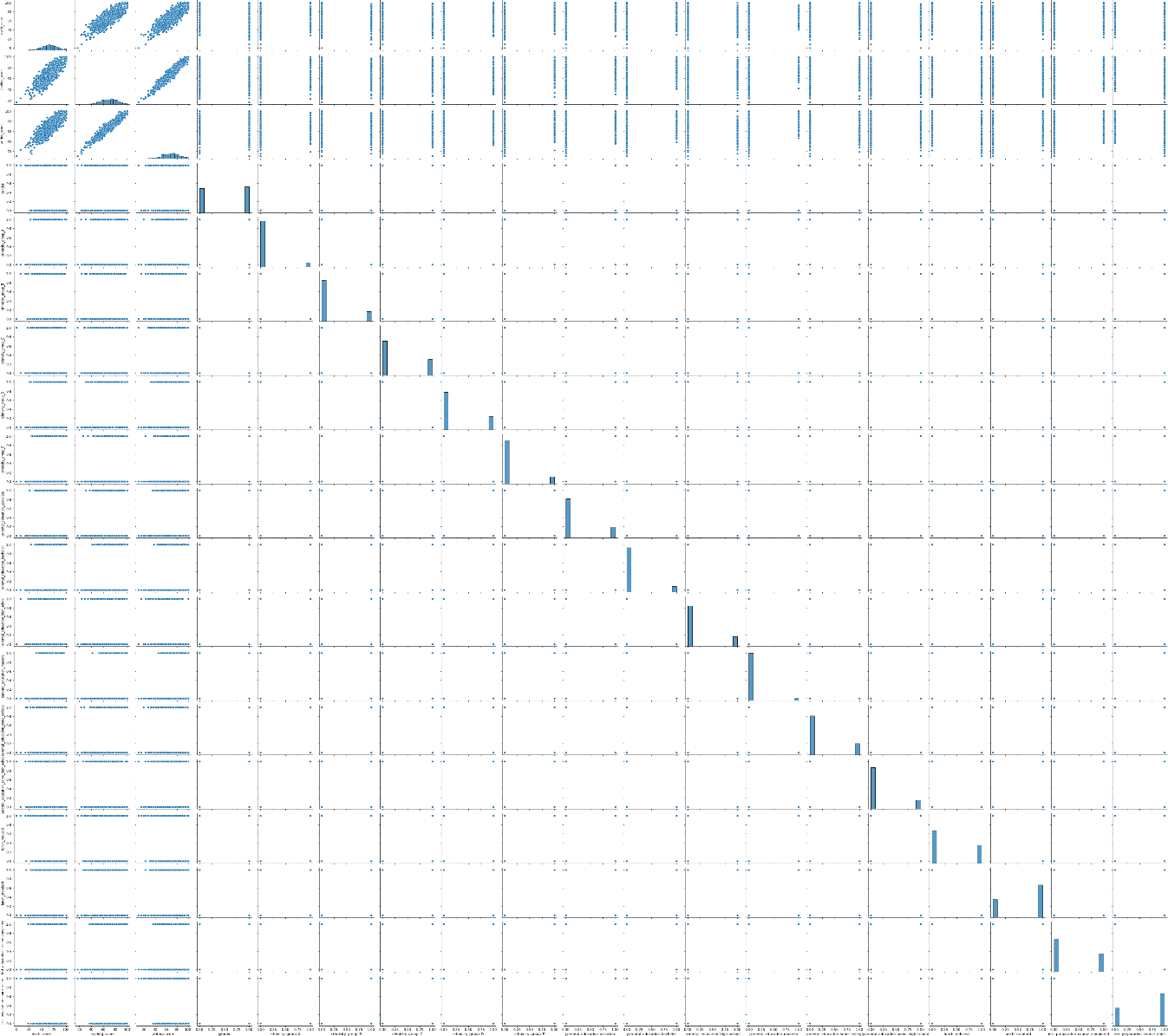


* df.corr() : to calculate correlation among all the features to be able to find the best features.



**Data Visualization**

* Plotting Gender with Reading : Using seaborn we have plotted all the correlation below.

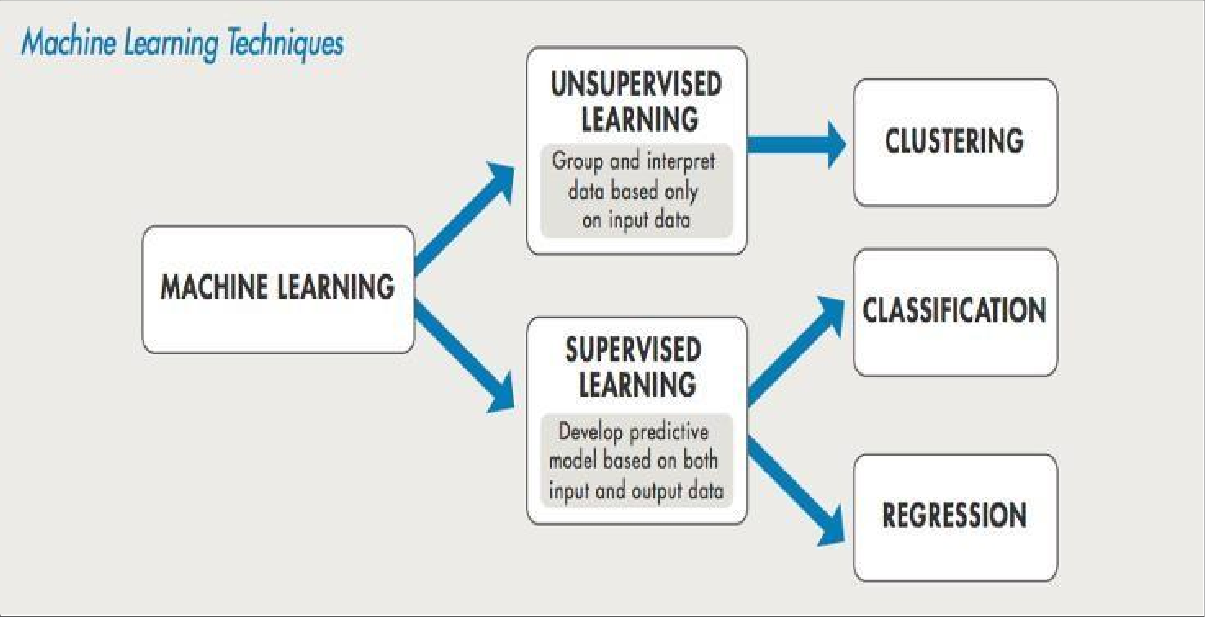


* Heat Map: We created a heat map to understand our features

Chart

Description automatically generated

# Model Planning and Implementation



**Supervised Learning Method**

### Supervised Learning can be divided into three main steps

1. Feature Selection
2. Feature Scaling
3. Modelling

### Feature Selection

Before implementing any mode, to select the best and most important features of the dataset, we used feature selection techniques like Linear SVC, Logistic Regression, Random Forest, RFE Logistic and Decision tree. Every feature selection method selected a different number of features, but we selected the one with the best performance.

### Feature Scaling

After that, we performed feature scaling on our selected features using Robust, Standard and Minmax Scaler.

### Modelling

Finally, the output from the best feature scaling technique was then fed to build our machine learning model. We used a pipeline method to feed all the classifiers namely KNN, Decision Tree, Random Forest, AdaBoost, XgBoost to our model.

Let us now see the detailed implementation of the mentioned models in the further report below.

**Feature Selection**

There are number of ways to select features based on different attributes. Firstly, we are trying to use the corelation method to find the best selected feature by using the code.

### df.corr()

We get a bunch of corelation, which we can see in the screenshot below.



Through this co-relation statistics of dataset, we can notice that our target variable **gender** has low positive relationship with **reading\_score** and **writing\_score** which reads to **0.2444313** and **0.301225** respectively.

Whereas the remaining variable has no co-relation with target variable and may not provide better predictions. Thus, we decide to not proceed with this method as only two features have low positive corelation with our target variable

Let us now use scatter plot to see if we can pick features based on the plots

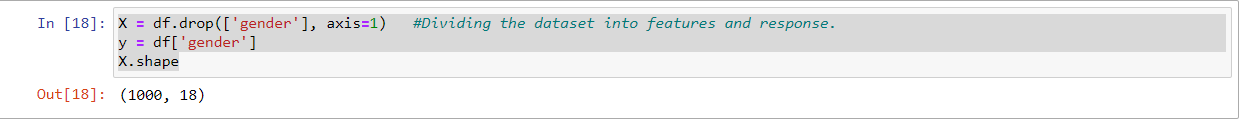
## gender VS reading\_score gender VS writing\_score



**gender VS ethnicity\_group\_A**



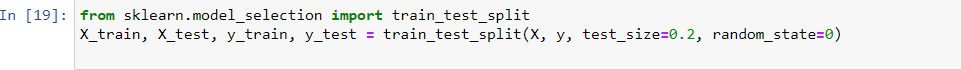
* Based on corelation, we can consider writing\_score and reading\_score as our features which influence our target variable. In simple words, we can say that based on all features in the dataset gender gets more influenced by its reading and writing scores. But this not the right technique which can provide better prediction. Hence, we will use some feature selection methods to identify features.
* To implement the feature selection method, we divide the dataset into features and response by using the below code.



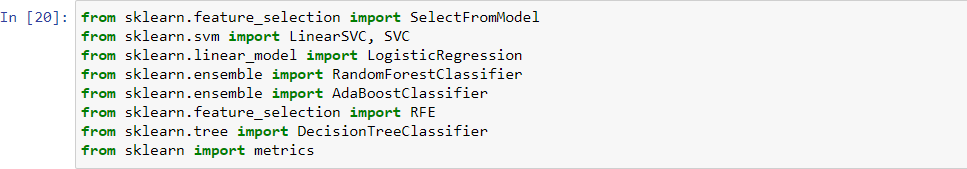
[

### We have a total of 18 features

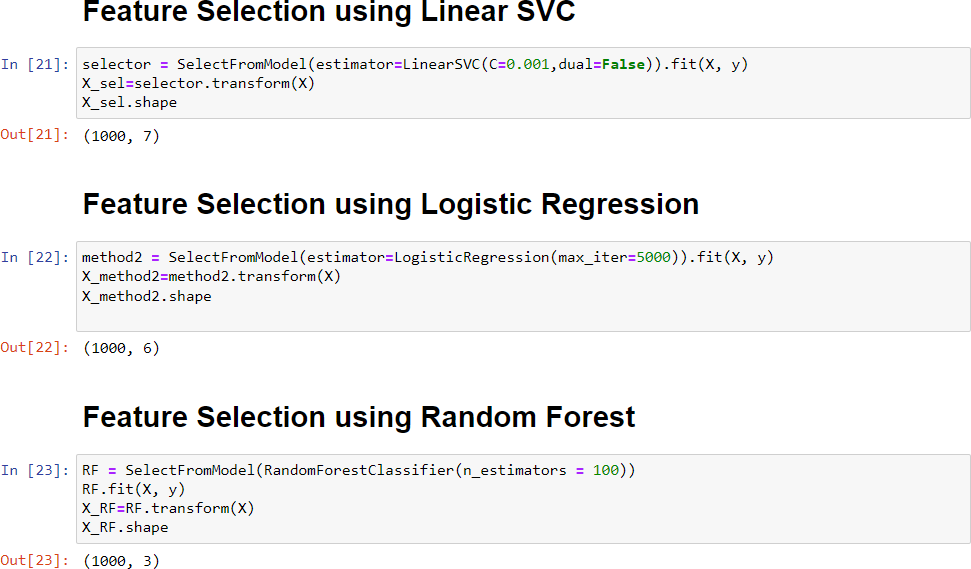
* After dividing the dataset into features and response we divide the dataset in training and testing data samples

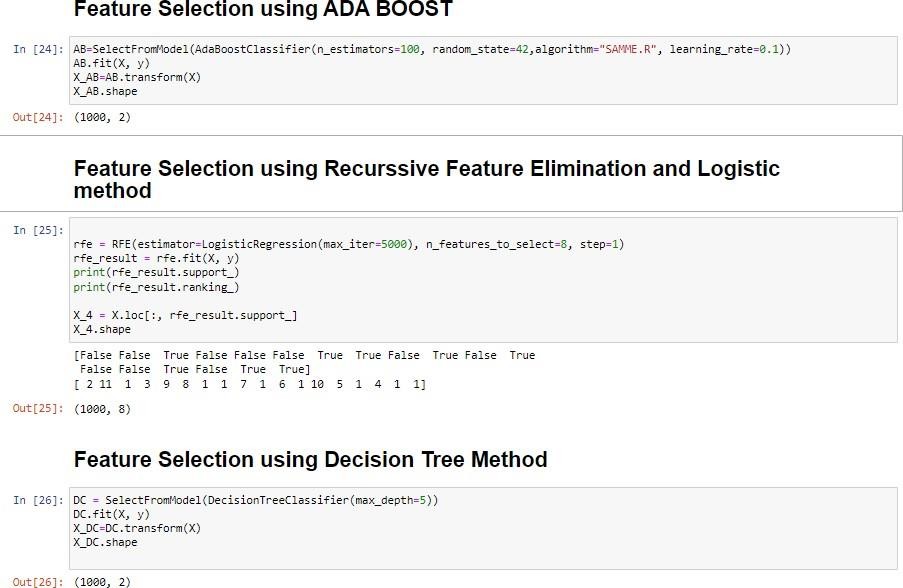


* We now import all the important libraries to use feature selection methods



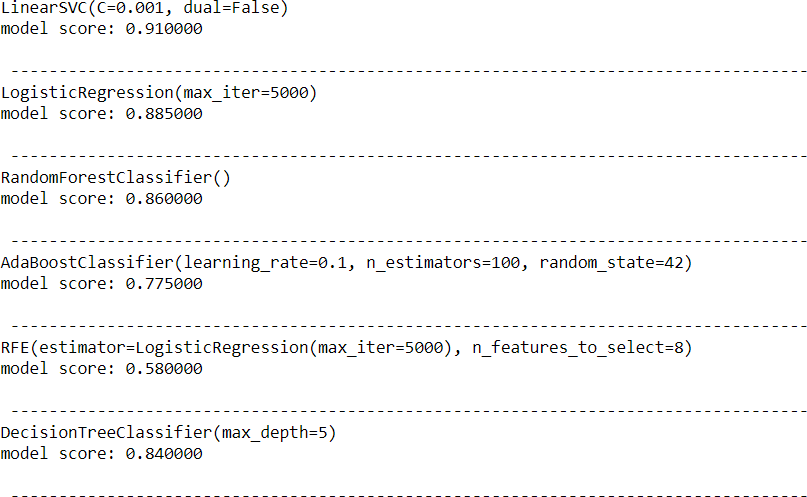
* We are using 6 different feature selection methods, to know which method gives us the best feature.
  1. Linear SVC
  2. Logistic Regression
  3. Random Forest
  4. ADA Boost
  5. Recursive Feature Elimination and Logistic method
  6. Decision Tree



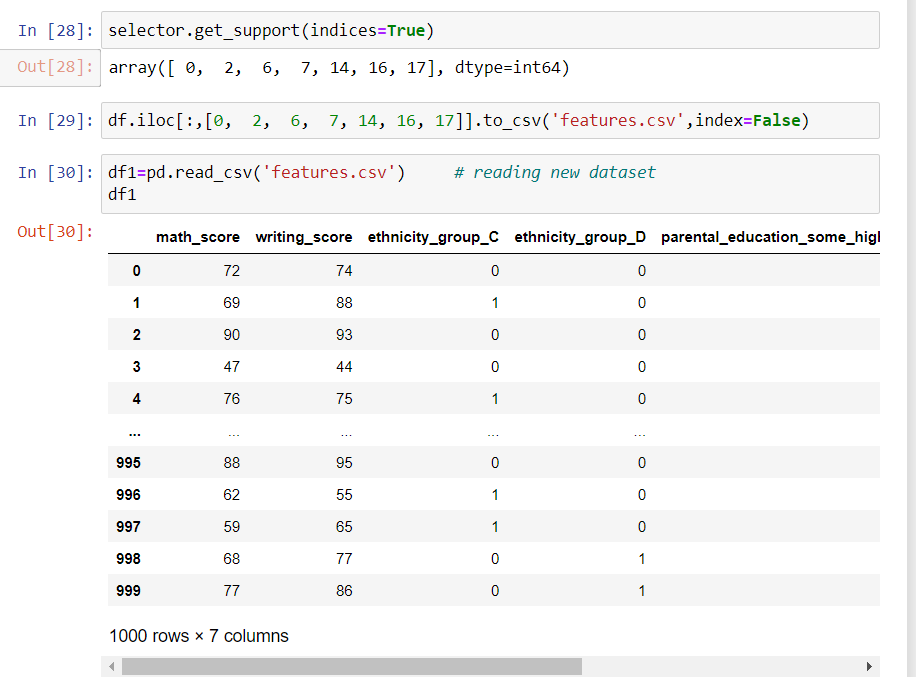


* Now, let us print the score of all the above feature selection method using pipeline to know which feature selection method gives the best result.





* In the above result we can see that the Linear SVC gives us the best results, Hence we use Linear SVC for feature selection
* Let us now saved the features as a new dataset file ‘features.csv’ and then also read it to see the new dataset.



## Feature Scaling

Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data pre-processing step. We expect to see an improved model performance with feature scaling under KNN and SVR and a constant model performance under decision trees with or without feature scaling.

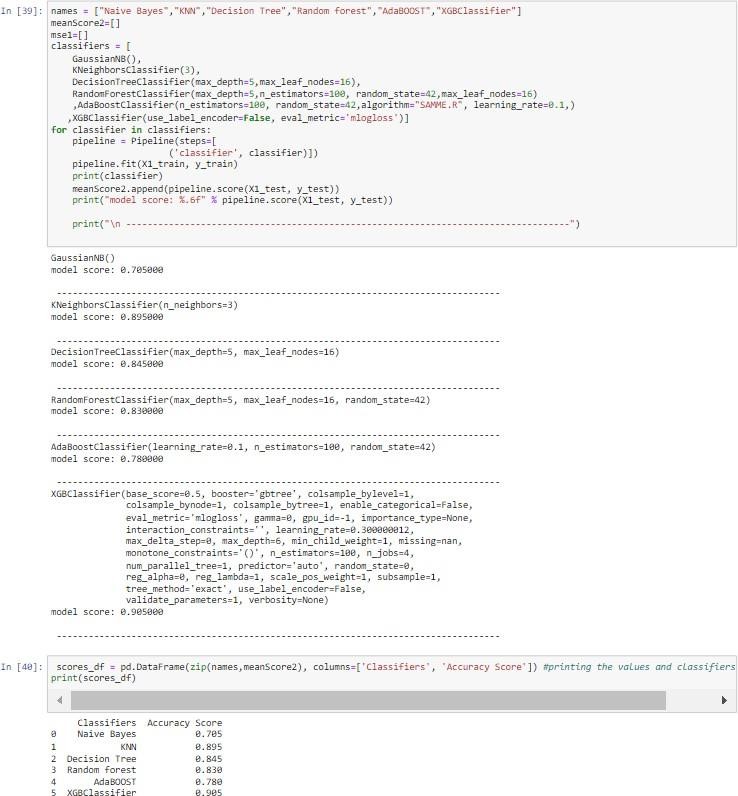
* + To start with our feature scaling let us our new data ‘df1’ into ‘X1’ and df[‘gender’] in y
  + Then split the new dataset into train and test set, and apply 3 scaling methods using Linear SVC to find the lowest RMSE
    1. Standard Scaler
    2. MinMax Scaler
    3. Robust Scaler



* + - We can notice that our selected features perform better without data scaling.
    - The original features subset error rate is low with the RMSE score of

**0.339116**, but when we use different scalers the error rate increases.

* + - Let us now apply classification on the features using different classification models to find the best accuracy score
    - We have applied 6 classification models
      1. Naïve Bayes
      2. K- Nearest Neighbors
      3. Decision Tree
      4. Random Forest
      5. AdaBoost
      6. XGBoost

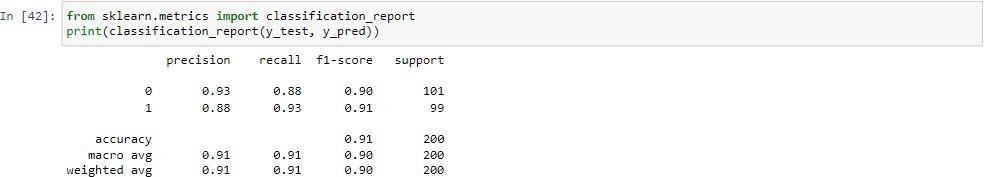


* We can see that from all the applied classifiers on our selected features, subset’s mean accuracy score is the highest in XGBClassifier = 0.905. Hence the best classifier to predict the accuracy is XGBoost Classifier.
* We can now further calculate the classification report, confusion matrix and ROC curve based on our results.



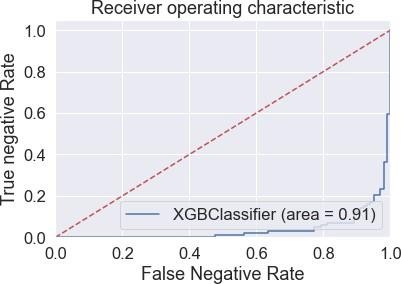
The result prompts us that we 181 (89+92) correct predictions and 19 (12+7) incorrect predictions.

### Classification Report



* From the report we can see that above precision is recorded higher, which means that a classifier cannot label all positive samples if the negative samples exist.
* In other words, certain factors can predict more positive impact on gender of students, while certain factors have no impact on the gender.

### ROC Curve



* We can see that area under curve is close to 1, which means the classifier applied shows better classification of data.
* We can conclude that it will give more accurate results of our prediction.

**Out Of Sample Prediction**

It is common to evaluate the performance of a machine learning algorithm on a dataset using a resampling technique such as k-fold cross-validation.

The k-fold cross-validation procedure involves splitting a training dataset into k groups, then using each of the k groups of examples on a test set while the remaining examples are used as a training set. This is called out of sample prediction.

This procedure can be summarized as follows:

1. Shuffle the dataset randomly.

2. Split the dataset into k groups.

3. For each unique group:

a. Take the group as a holdout or test data set.

b. Take the remaining groups as a training data set.

c. Fit a model on the training set and evaluate it on the test set.

d. Retain the evaluation score and discard the model.

4. Summarize the skill of the model using the sample of model evaluation score

Graphical user interface, text, application

Description automatically generated

# Results Interpretation, Implications and Conclusion

* After carefully performing the steps of feature selection, scaling, transforming, and modelling using supervised machine learning algorithms, we have deduced that our dataset produces 90.5% accuracy without feature scaling and following supervised learning methods so far as compared to other computations we did on our dataset.
* The reason our analysis doesn’t include Clustering (an unsupervised learning method) is because since the dataset we have, is been well organized and well labeled and performing clustering on our dataset will result in incorrect analysis and prediction.
* Also, we can see that gender classification whether male or female highly depended on most important features those are math\_score (20) and writing\_score(21.25)while other features put minimal impact on gender category as per our scaled data subset and feature selection. This selection has given us accuracy of 0.905
* In our analysis we can aslo conclude that if the gender is male there's a high chance they perform better in Maths and if the gender is female they have a higher chance to perform better in Writing.