# Predictive modeling

PRESENTED BY : JYOTI DEVI

# TASK DESCRIPTION

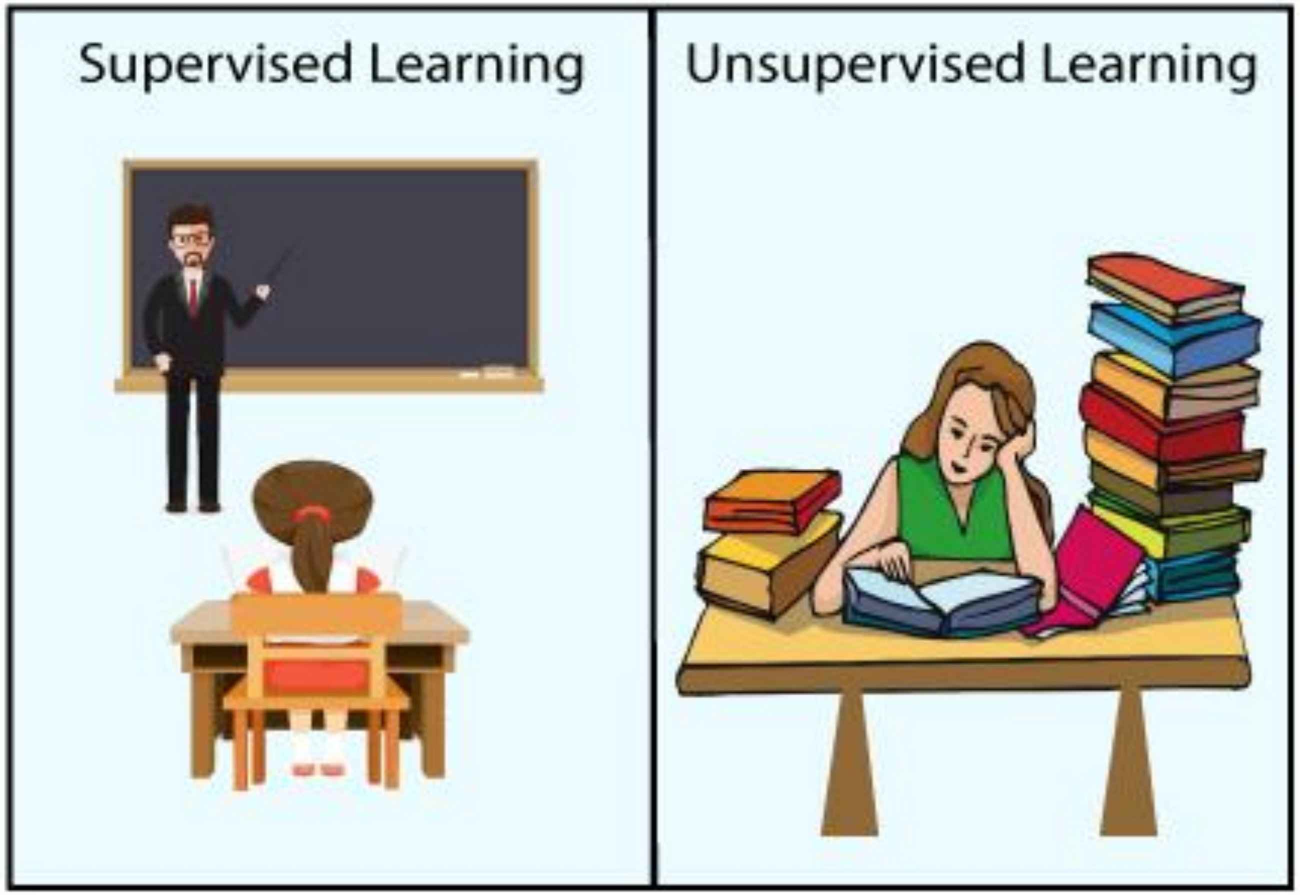
In this task, I will build a predictive model using a given dataset to predict a target variable using supervised learning algorithms such as decision trees, logistic regression, and random forests to build the model.

We will Build a predictive model to predict the math score of a student based on other variables such as gender, race/ethnicity, parental level of education, lunch, and test preparation course.

# What is Supervised learning ?

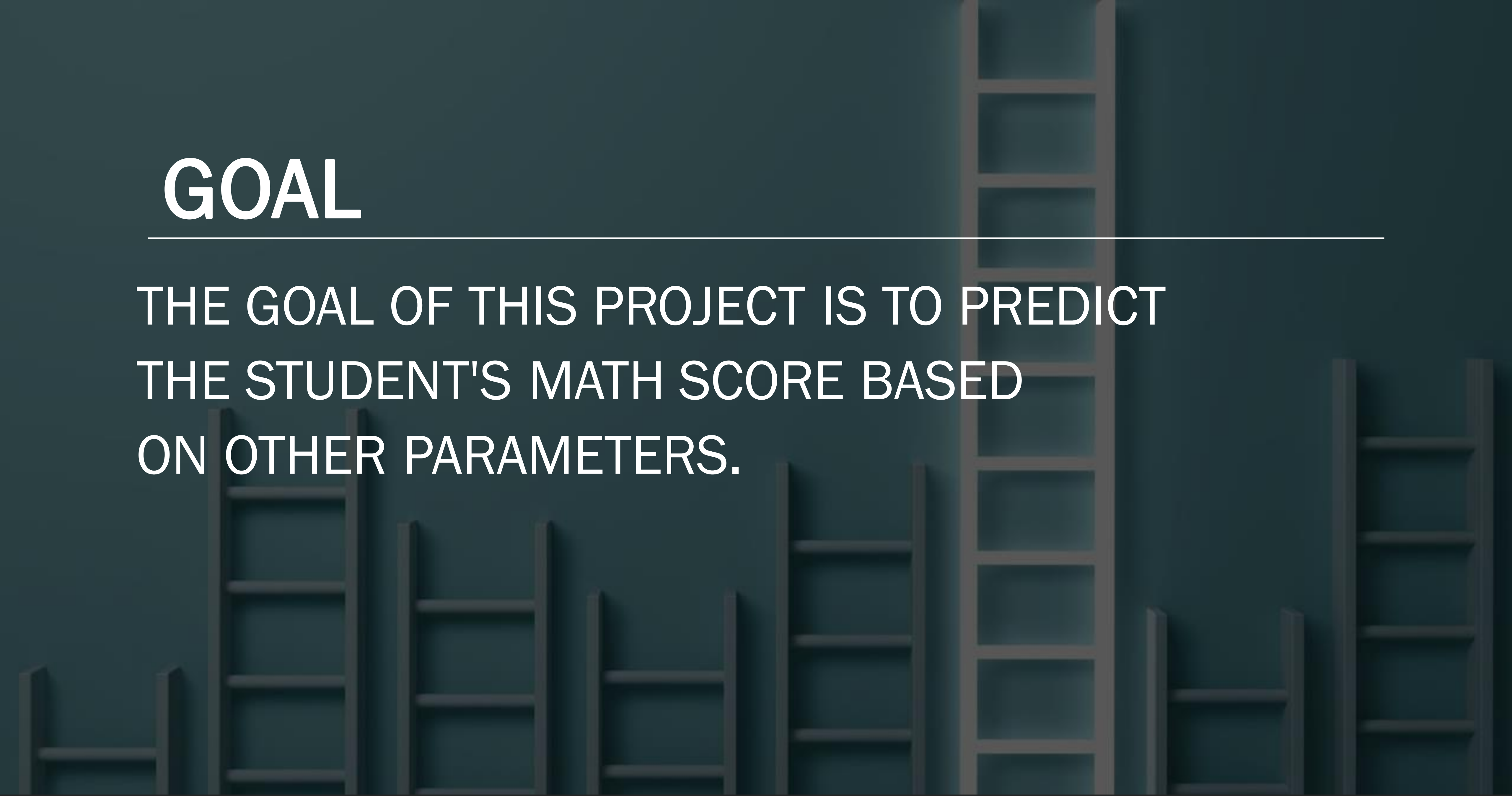
In this technique a computer algorithm is trained on input data that has been labeled for a particular output. The model is trained until it can detect the underlying patterns and relationships between the input data and the output labels

# Example:



# DATASET

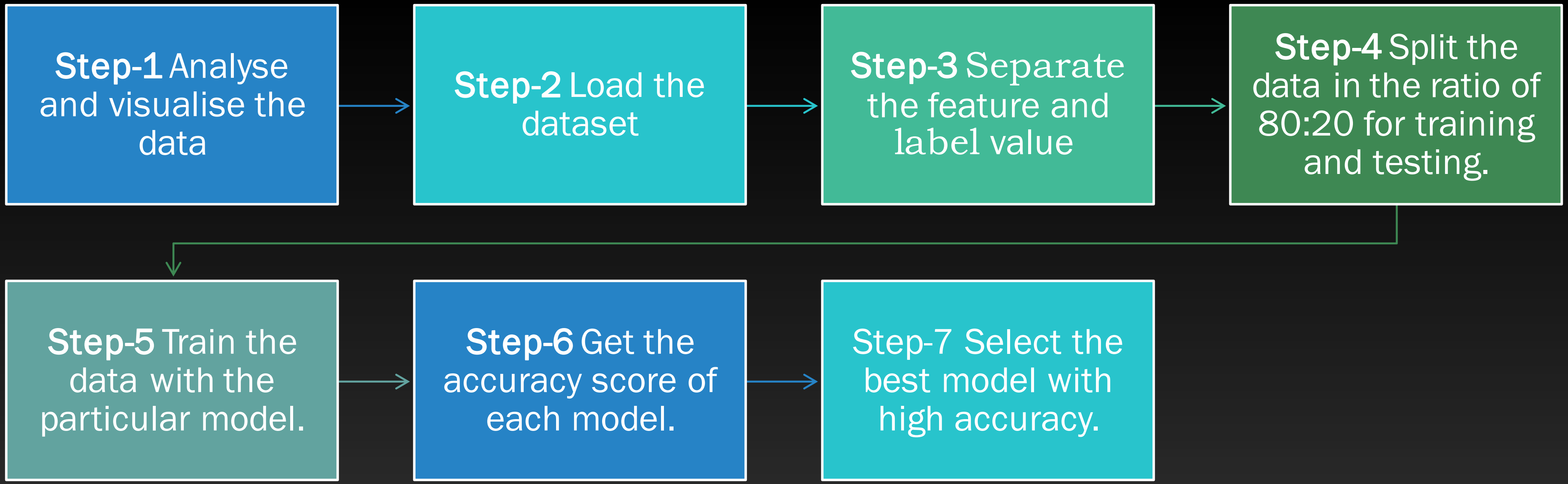
- The dataset is provided by Hackveda and this dataset contain the information of student.
- We can download the dataset from here :
- https://www.kaggle.com/spscientist/students-performance-in-exams



# GOAL

THE GOAL OF THIS PROJECT IS TO PREDICT THE STUDENT'S MATH SCORE BASED ON OTHER PARAMETERS.

# Steps to achieve this goal :



**Step-1** Analyse and visualise the data

**Step-2** Load the dataset

**Step-3** Separate the feature and label value

**Step-4** Split the data in the ratio of 80:20 for training and testing.

**Step-5** Train the data with the particular model.

**Step-6** Get the accuracy score of each model.

Step-7 Select the best model with high accuracy.

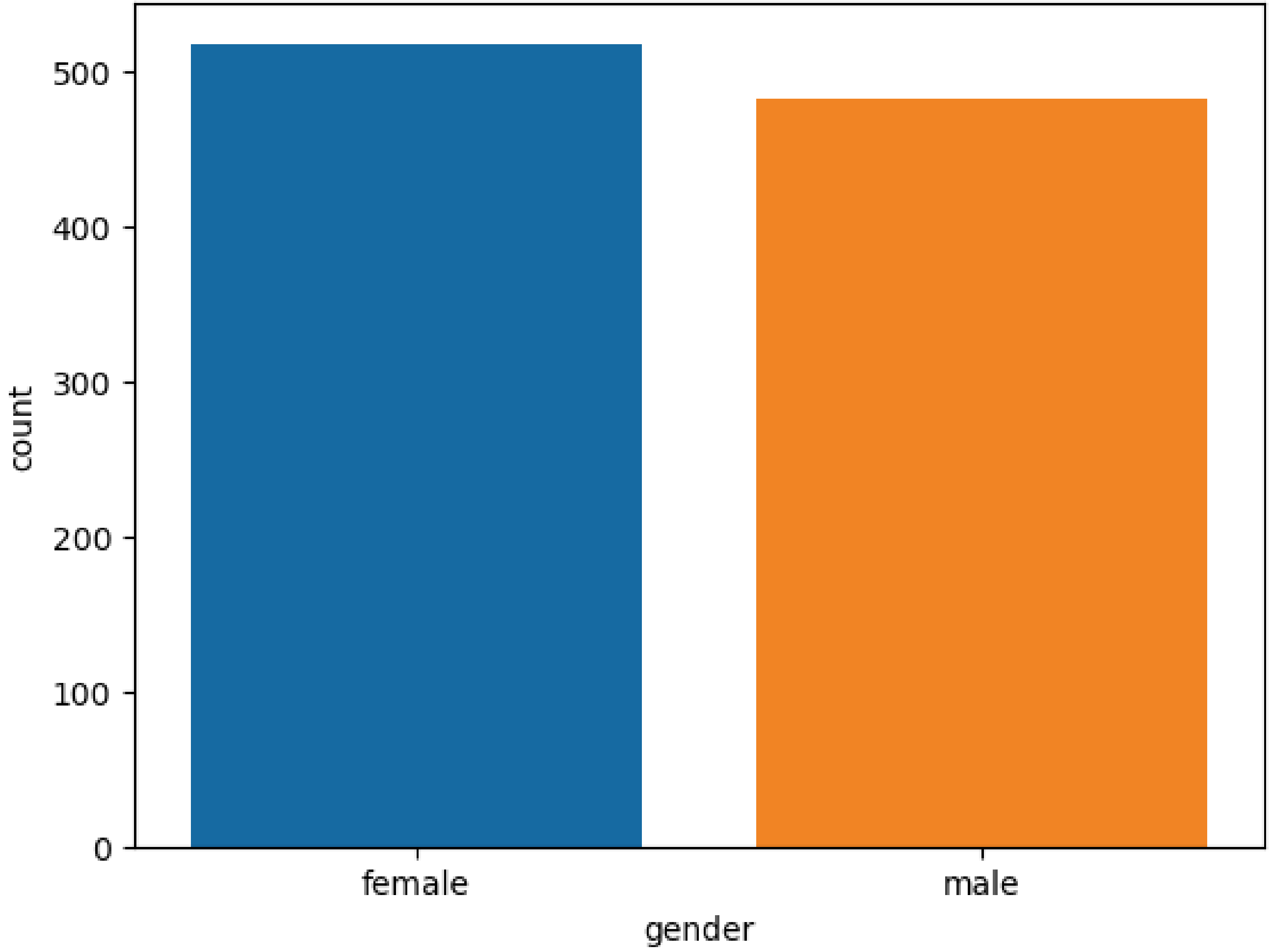
# Feature values

After **analysis** the dataset we got some feature values on which the students's math score depends.

1. gender
2. Race/ethnicity
3. parental level of education
4. Lunch
5. test preparation course
6. reading score
7. writing score

# Data Visualisation

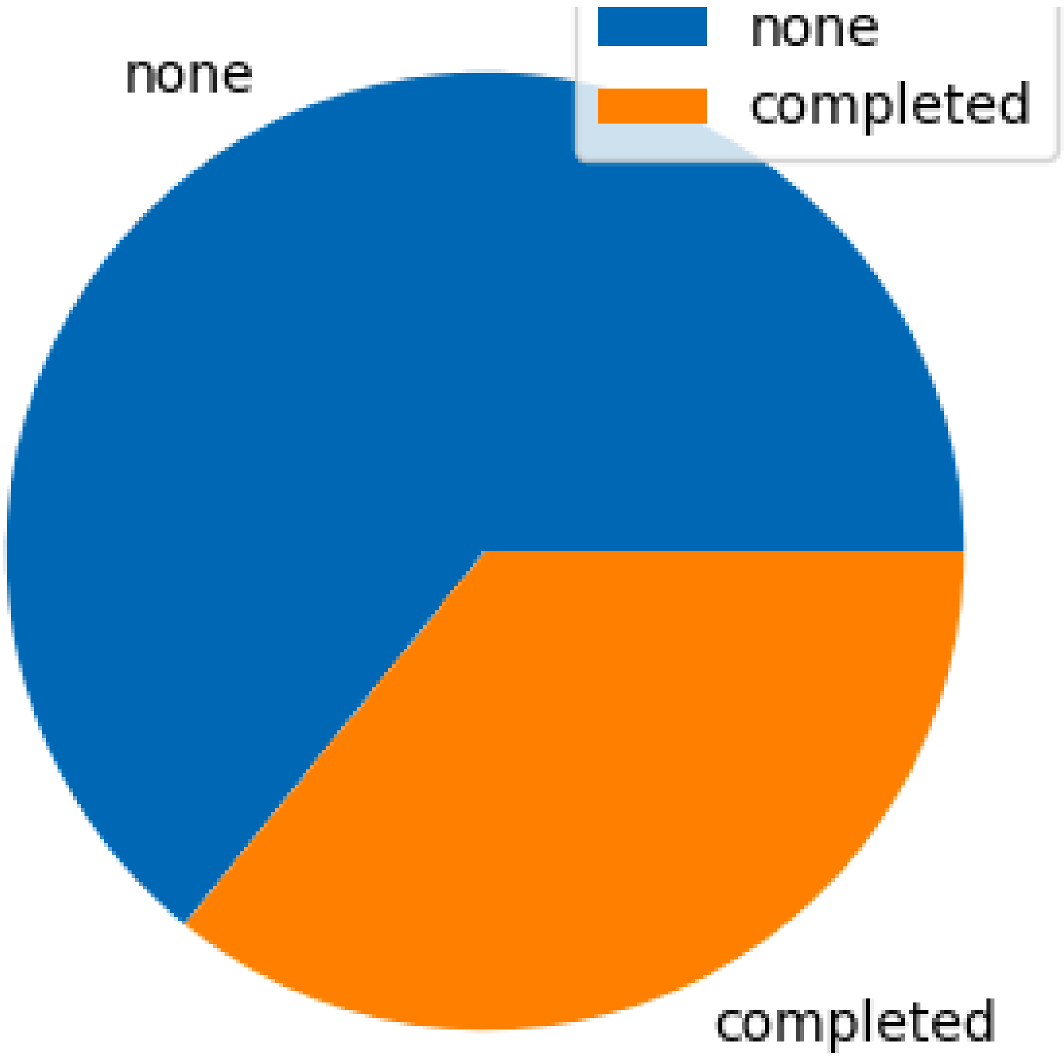
Checking  the gender column  is balance or not?



# Data Visualisation

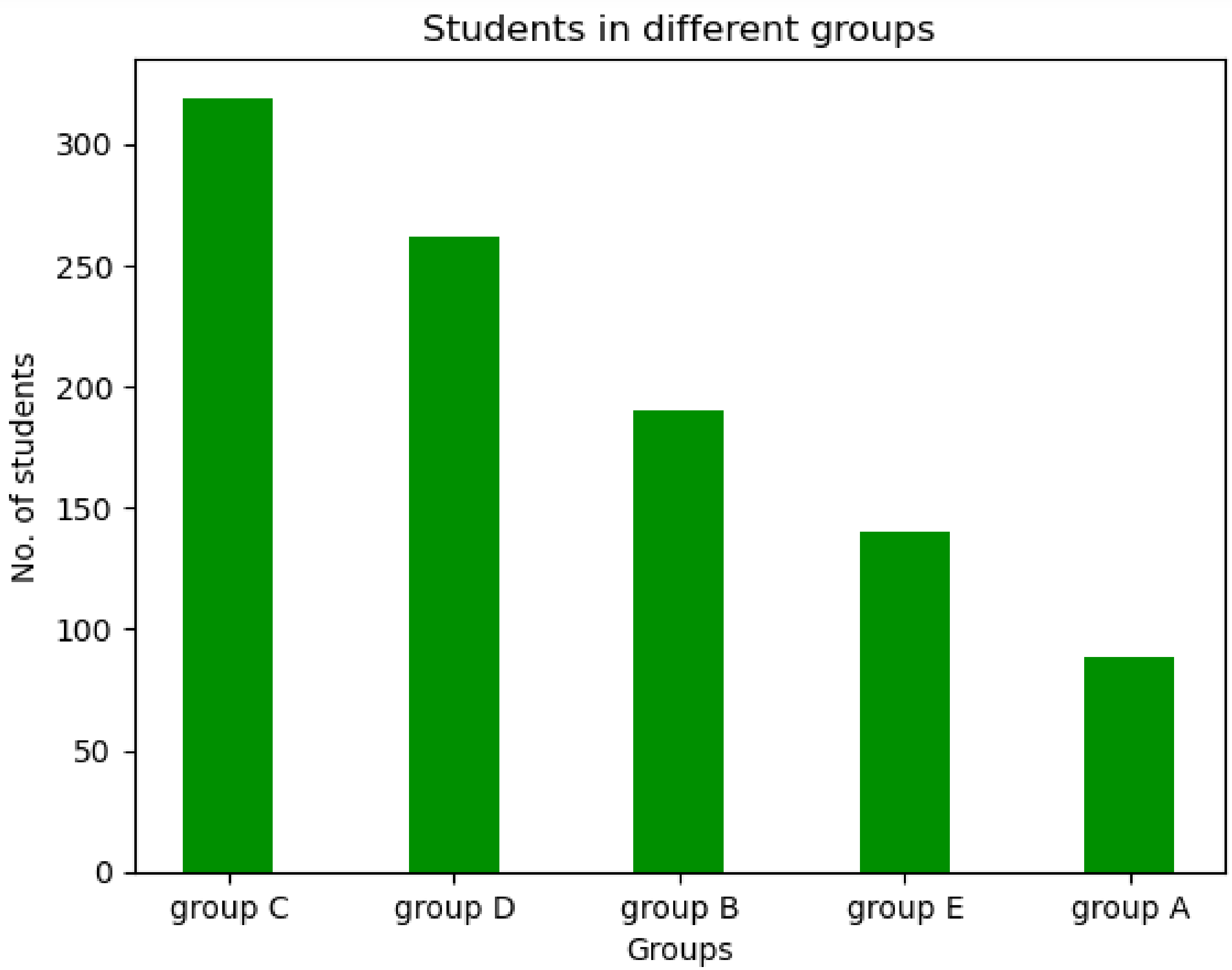
Checking How many of the students took course.

- Significantly more students didn't take course



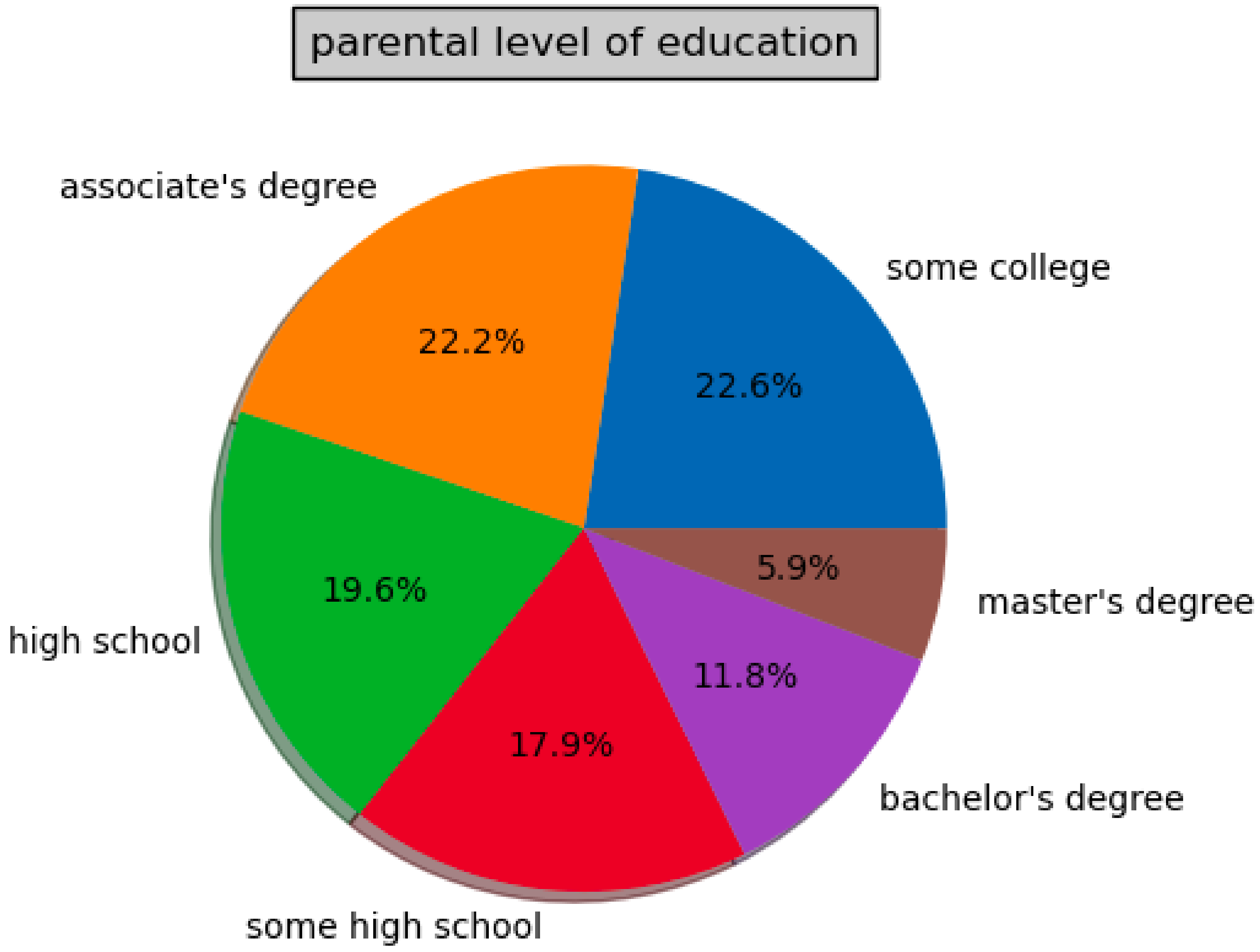
# Data Visualisation

No. of students in each group



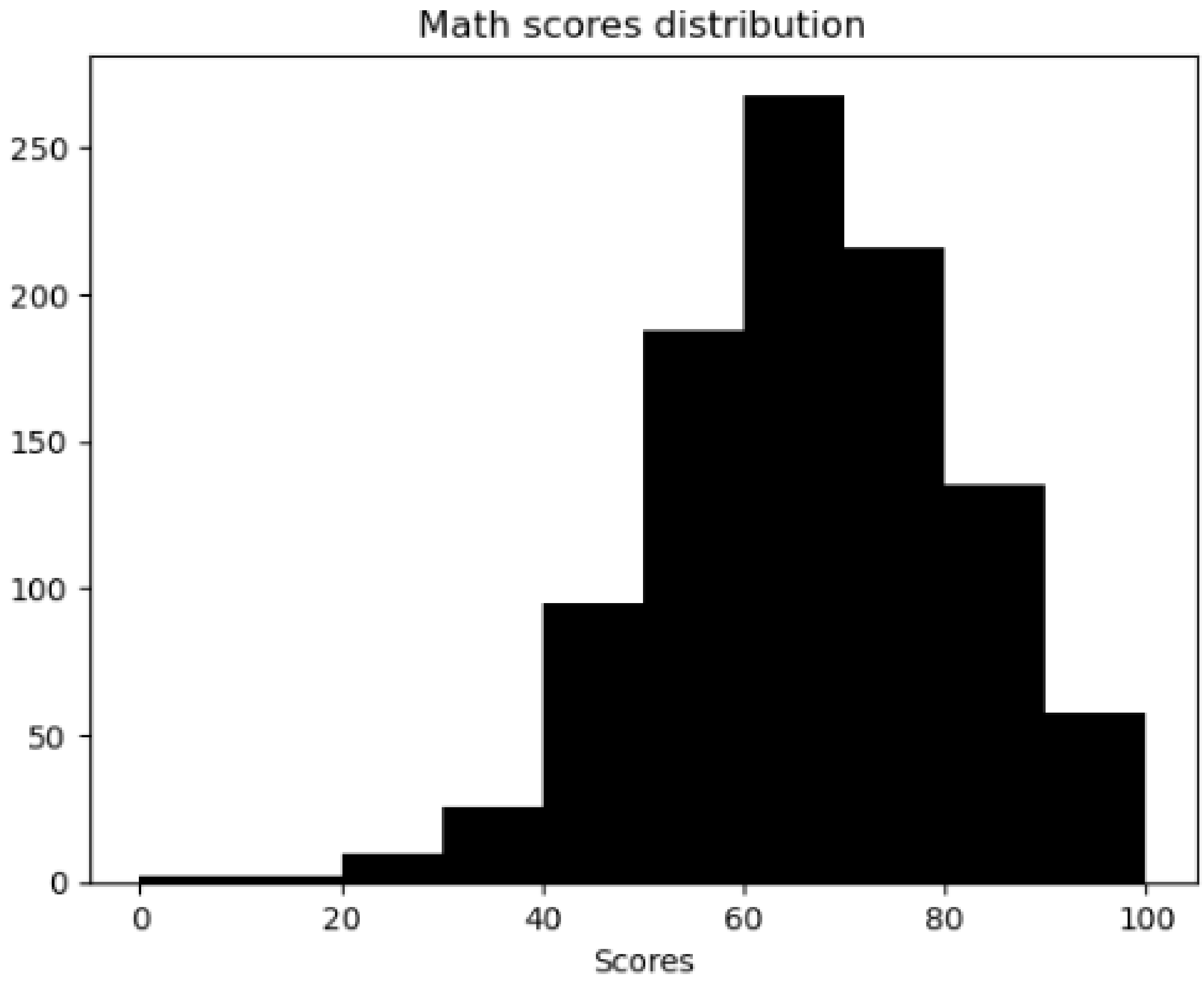
# Data Visualisation

Parental education

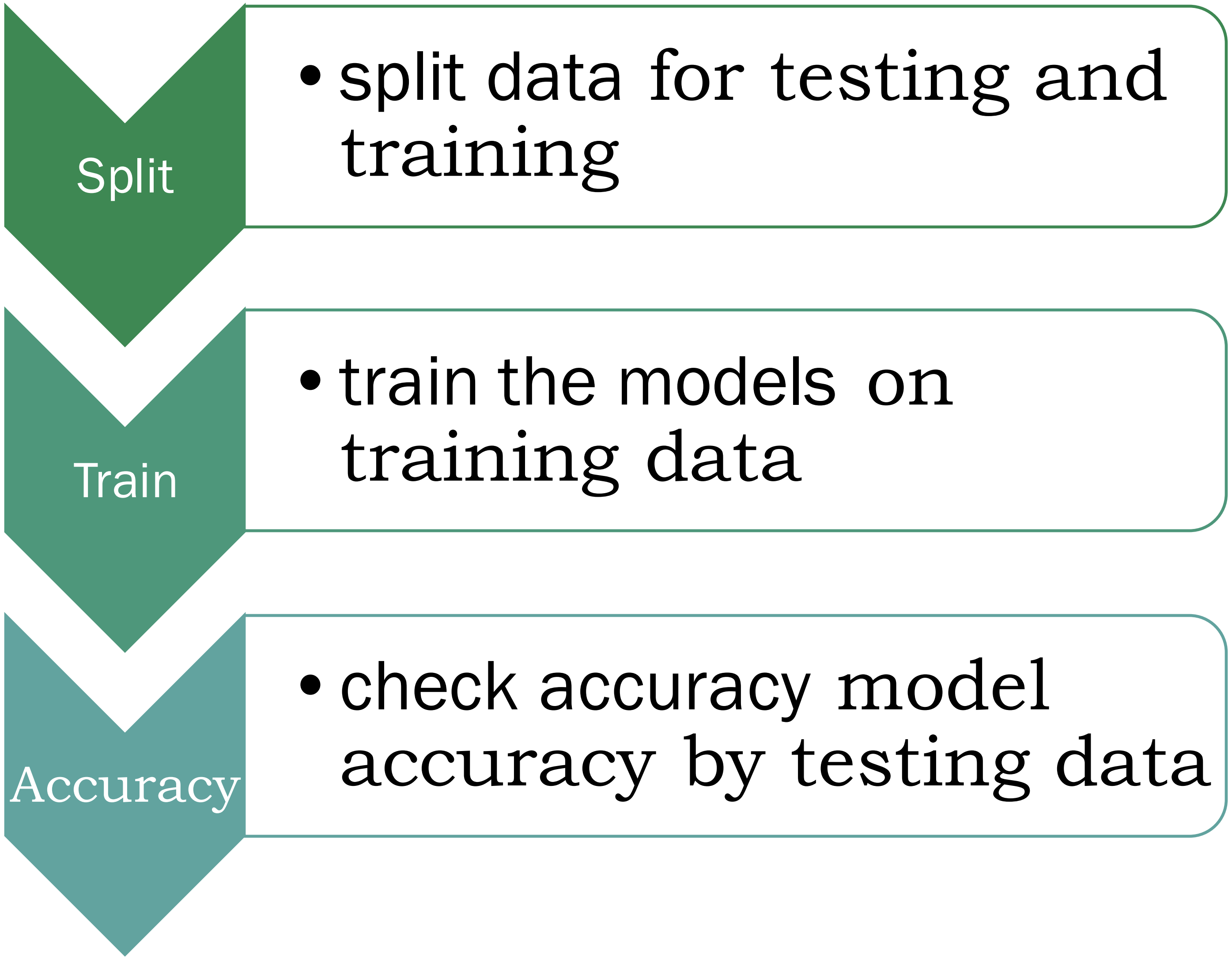


# Data Visualisation

Scores distribution visualization by histogram



# Steps to get best result



Split

- split data for testing and training

Train

- train the models on training data

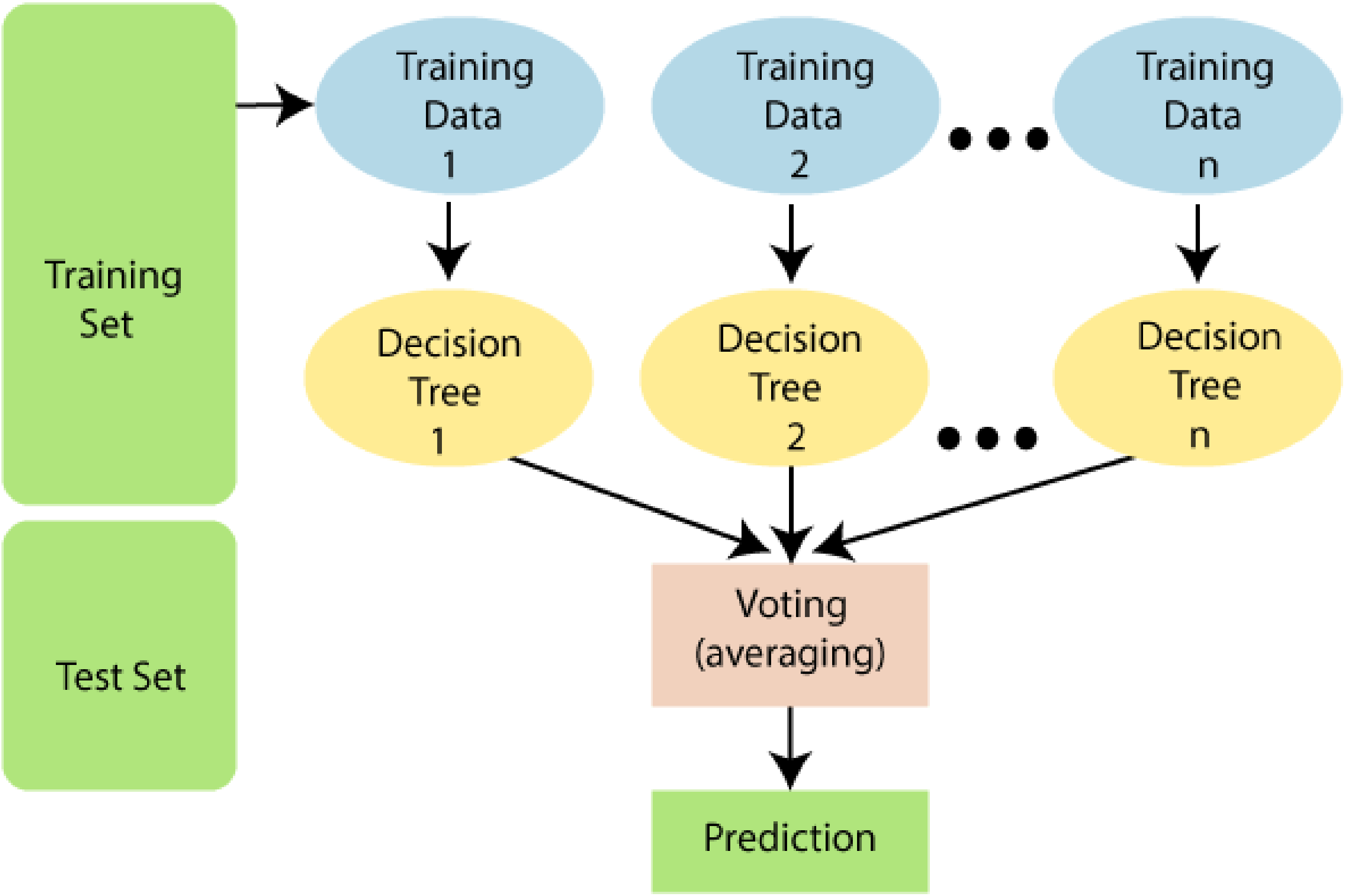
Accuracy

- check accuracy model accuracy by testing data

# Machine learning algorithms used

1. random forest model
2. Logistic Regression
3. Linear regression

# Random forest model



Training Set

Test Set

Training Data 1

Training Data 2

Training Data n

Decision Tree 1

Decision Tree 2

Decision Tree n

Voting (averaging)

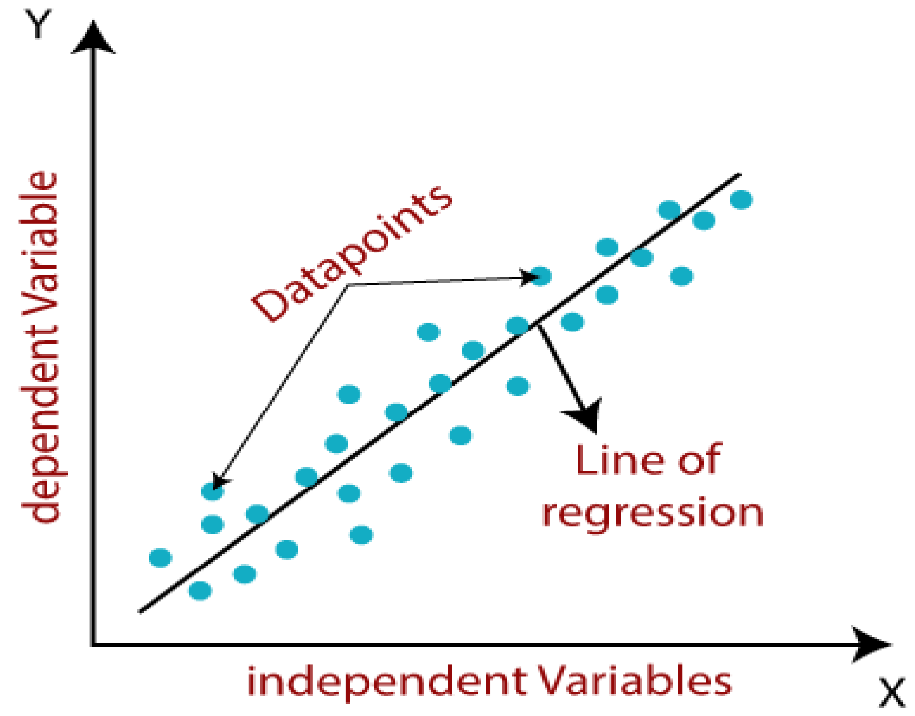
Prediction

# Logistic regression model

Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a  discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, **it gives the probabilistic values which lie between 0 and 1**.

# Linear regression model

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable



# Accuracy of each model

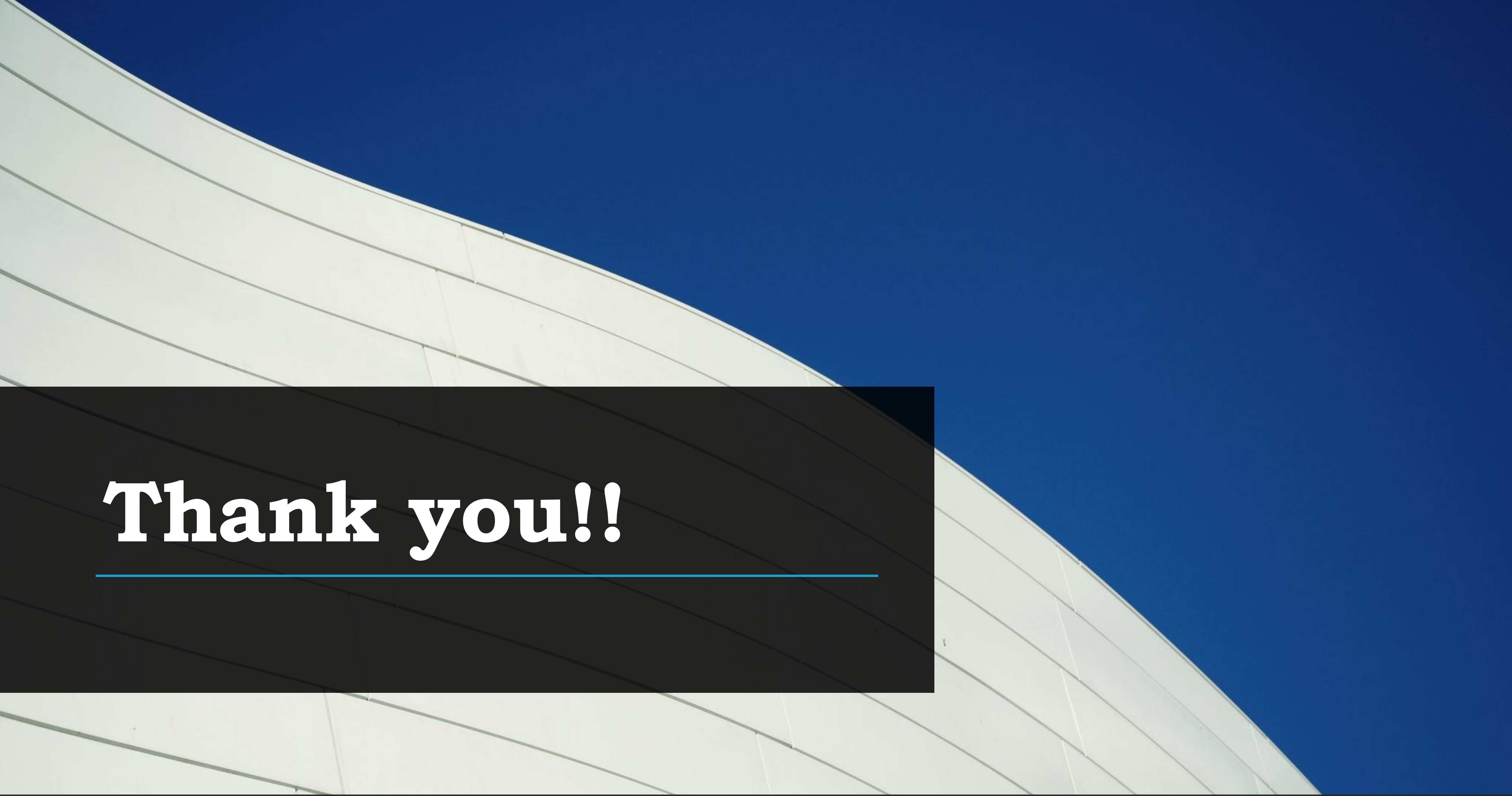
Linear regression model : 88.38026201112224%

Random forest model  : 84.89196832714583 %

Logistic Regression model : 67.54519631919993 %

# Conclusion

The accuracy of linear regression model is more as compare to random forest and logistic regression model so, Linear regression model is more suitable for the given dataset.



# Thank you!!