Diabetes Prediction System

Group 74

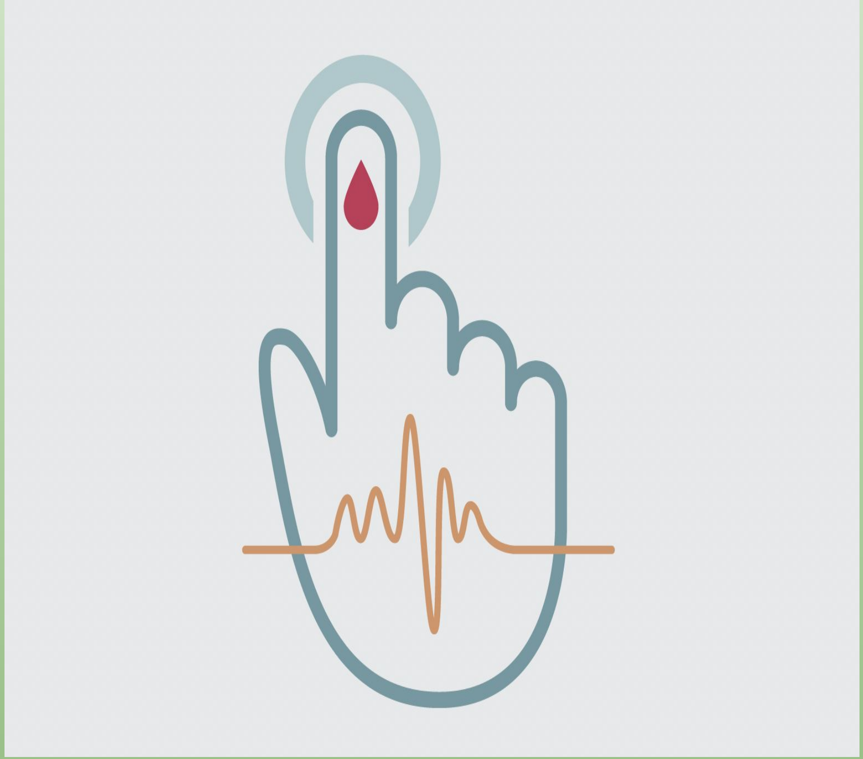
Shraddha Padalkar
Surabhi Shende

Presentation Date: 25th April 2022

IE 7275- Data Mining in Engineering

Table of Contents

- Problem Setting
- Objective
- Data Source & Data Description
- Data Preprocessing
- Exploratory Data Analysis
- Model Building
- Final Result
- Challenges
- Future Scope



Project overview visuals

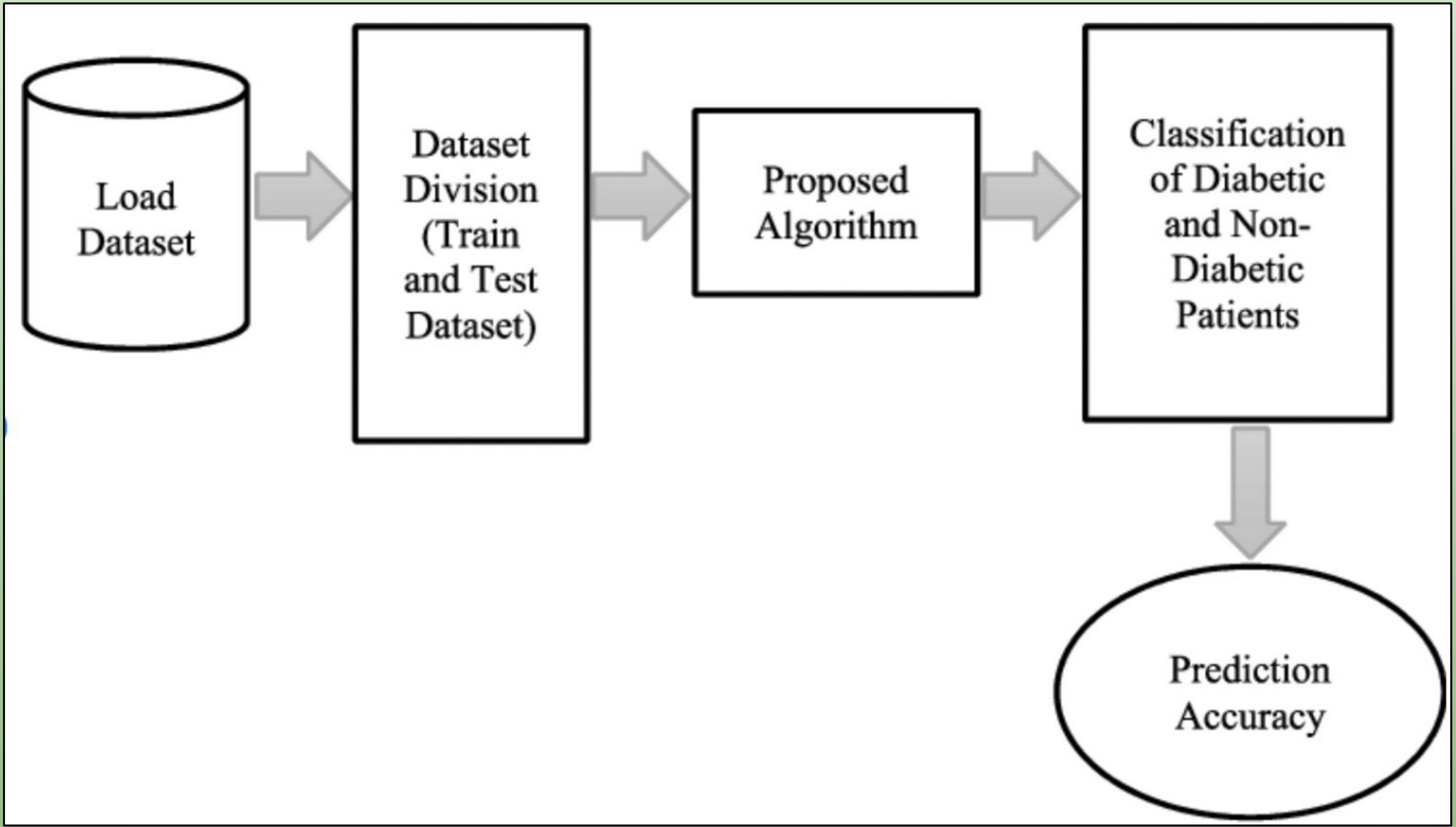
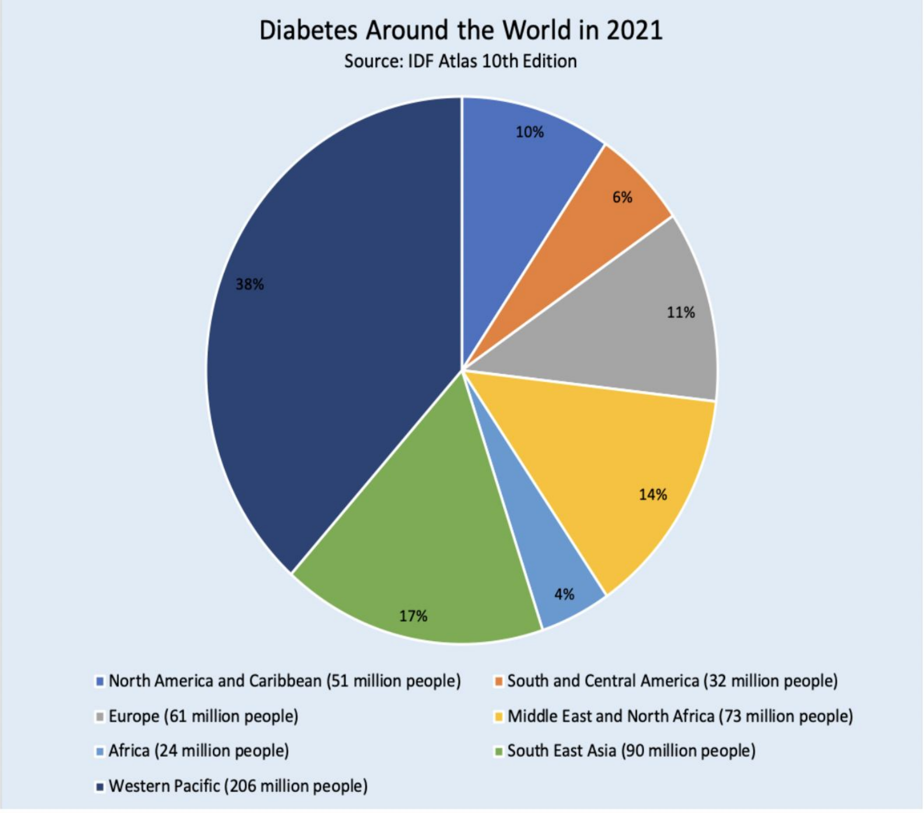

Load Dataset
Dataset Division (Train and Test Dataset)
Proposed Algorithm
Classification of Diabetic and Non-Diabetic Patients
Prediction Accuracy

Fig 1: Flow Chart for Diabetes Prediction

Problem Setting

- Diabetes is a chronic disease that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces.
- It has the potential to induce a variety of serious illnesses, including stroke, kidney failure, & heart attacks.



Diabetes Around the World in 2021

Source: IDF Atlas 10th Edition

Fig 2: Diabetes around the world in 2021

Objective

- The main goal of this project is to predict Diabetes using data mining technologies such as Logistic Regression, Random Forest, Decision Trees etc

- Our aim is to develop a model that will recognize whether a person is diabetic or not based on parameters such as Glucose level, blood pressure, BMI etc.

- The study's major goal is to create a system that can predict diabetic individuals with better accuracy.

Data Source & Data Description

- The dataset comprises of 768 instances with 8 predictor attributes and one target attribute.
- The outcome will be binary where 0 represents a patient without diabetes and 1 represents otherwise.
- Dataset has been taken from Kaggle which is an open-source, secure online repository for data

Data Source: https://www.kaggle.com/mathchi/diabetes-data-set

Data Preprocessing

- Handling of missing or null values should be taken care of because it creates problems while classification using Machine learning models.
- The data cleaning was performed and found that there are 0 null values in the dataset.
- As there were no missing and no null values the dataset comprises of 768 instances with 8 attributes and one target attribute.

Exploratory Data Analysis

- After processing the data pre-processing steps, there were 768 instances.
- The numeric variables are scaled and we proceeded with the Model building phase.
- We found out which attributes were correlated and we created a count plot to count number of people who have diabetes right from young to old people.

Exploratory Data Analysis - Numeric Variables

We visualized these attributes using heatmap which significantly contributed to predicting of diabetes.

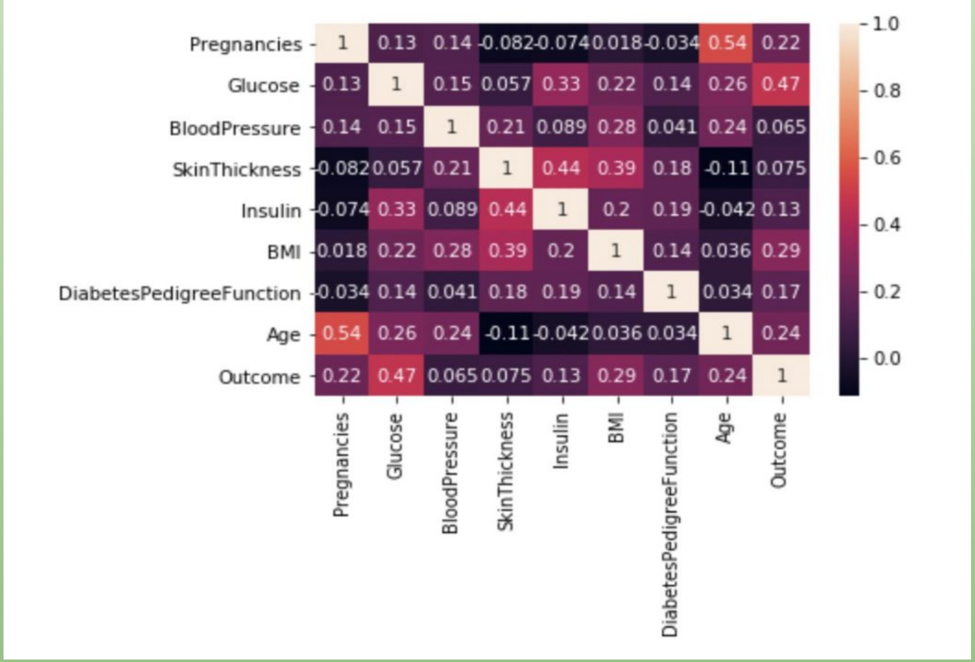


Fig 3: HeatMap containing attributes contributing to diabetes

Exploratory Data Analysis - Numeric Variables

- We counted people on the basis of their Body Mass Index(BMI).
- We also counted number of women who are pregnant and are diabetic.

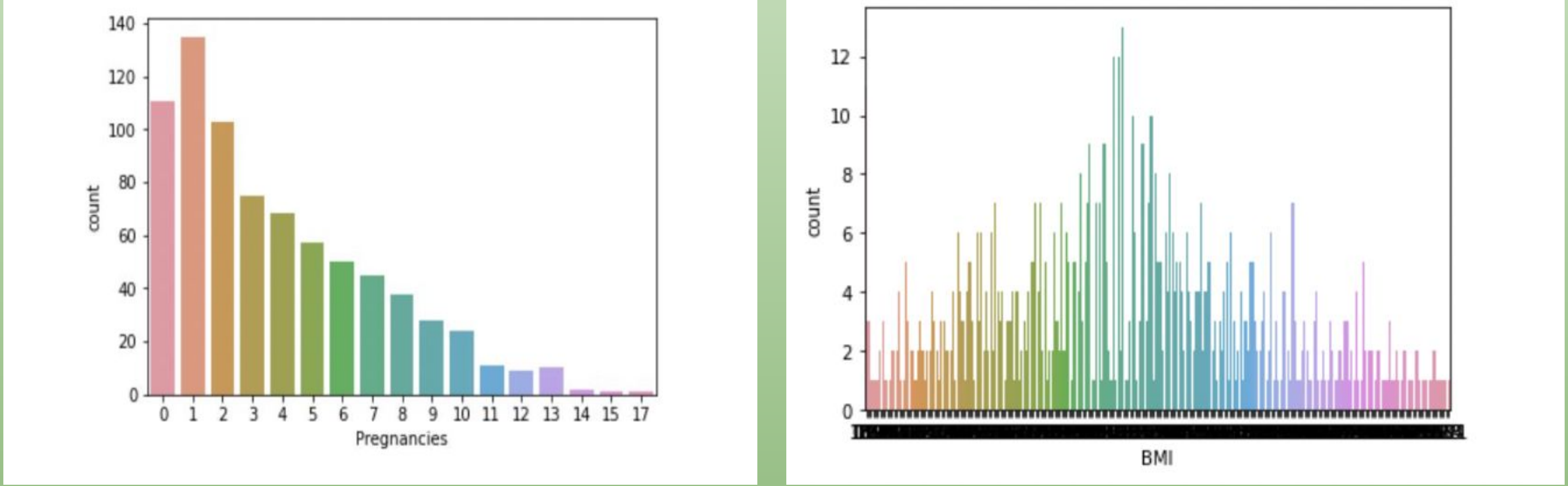
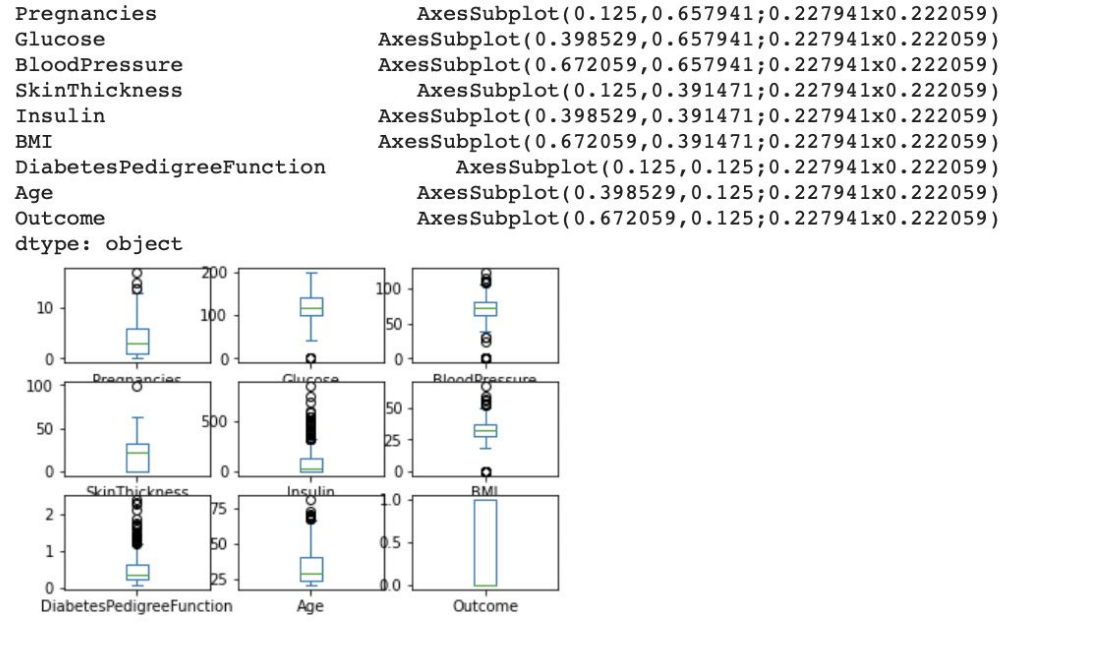


Fig 4: Count Plot

Exploratory Data Analysis - Numeric Variables

We also created a Box plot to visualise all the variables.



```
Pregnancies                     AxesSubplot(0.125,0.657941;0.227941x0.222059)
Glucose                      AxesSubplot(0.398529,0.657941;0.227941x0.222059)
BloodPressure                AxesSubplot(0.672059,0.657941;0.227941x0.222059)
SkinThickness                   AxesSubplot(0.125,0.391471;0.227941x0.222059)
Insulin                      AxesSubplot(0.398529,0.391471;0.227941x0.222059)
BMI                          AxesSubplot(0.672059,0.391471;0.227941x0.222059)
DiabetesPedigreeFunction           AxesSubplot(0.125,0.125;0.227941x0.222059)
Age                             AxesSubplot(0.398529,0.125;0.227941x0.222059)
Outcome                         AxesSubplot(0.672059,0.125;0.227941x0.222059)
dtype: object
```

Fig 5: Box Plot

Exploratory Data Analysis - Numeric Variables

We also created a Dist plot to visualise density distribution wrt all the variables.

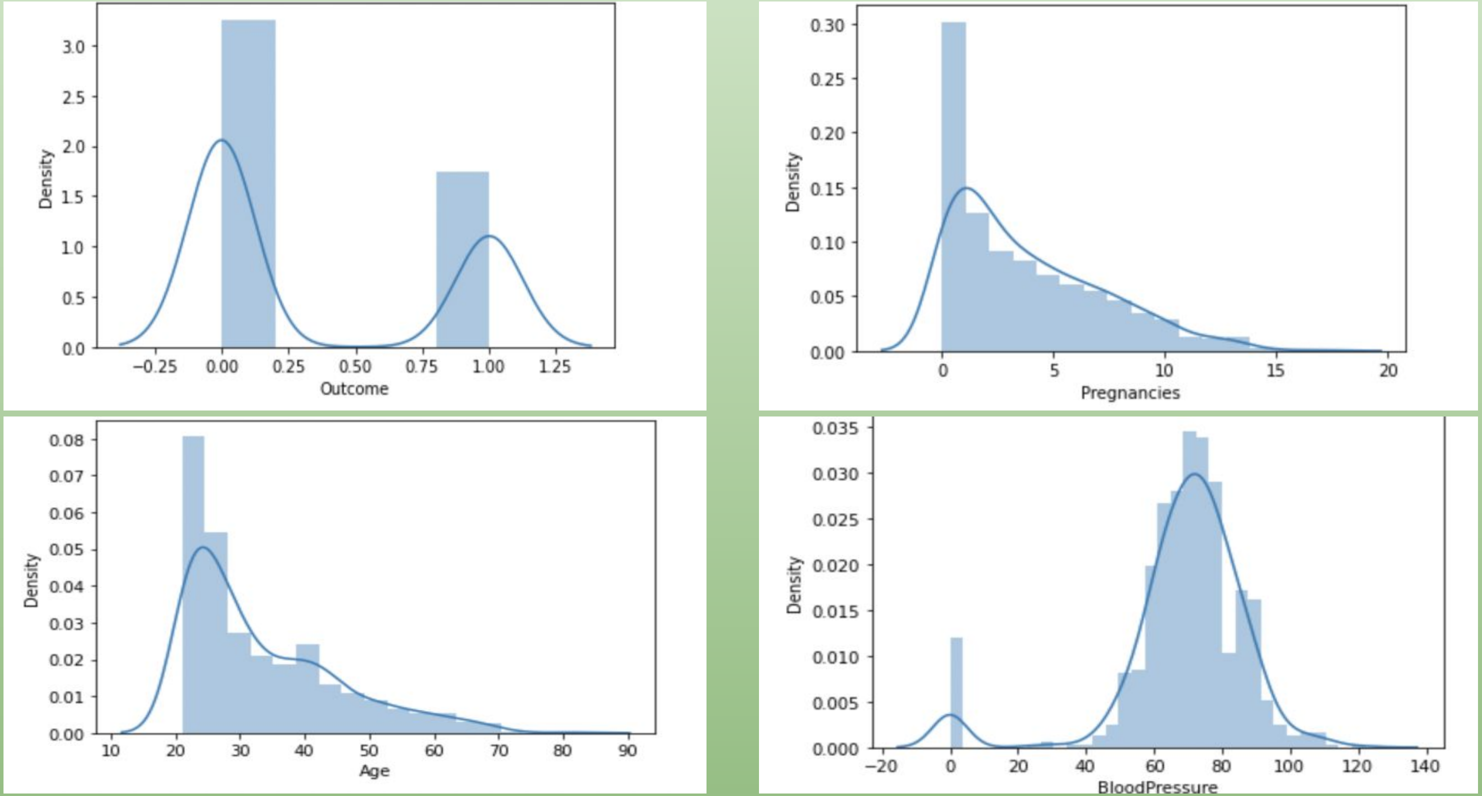


Fig 6: Dist Plot

Exploratory Data Analysis - Numeric Variables

We also created a Dist plot to visualise density distribution wrt all the variables

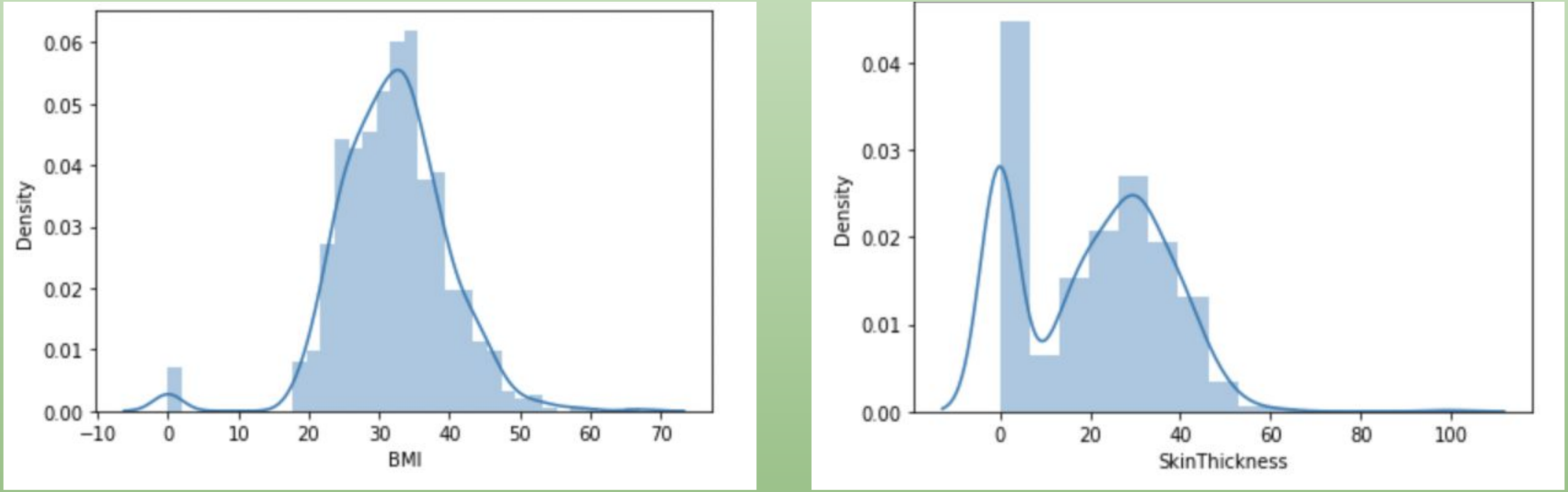


Fig 6: Dist Plot

Exploratory Data Analysis - Numeric Variables

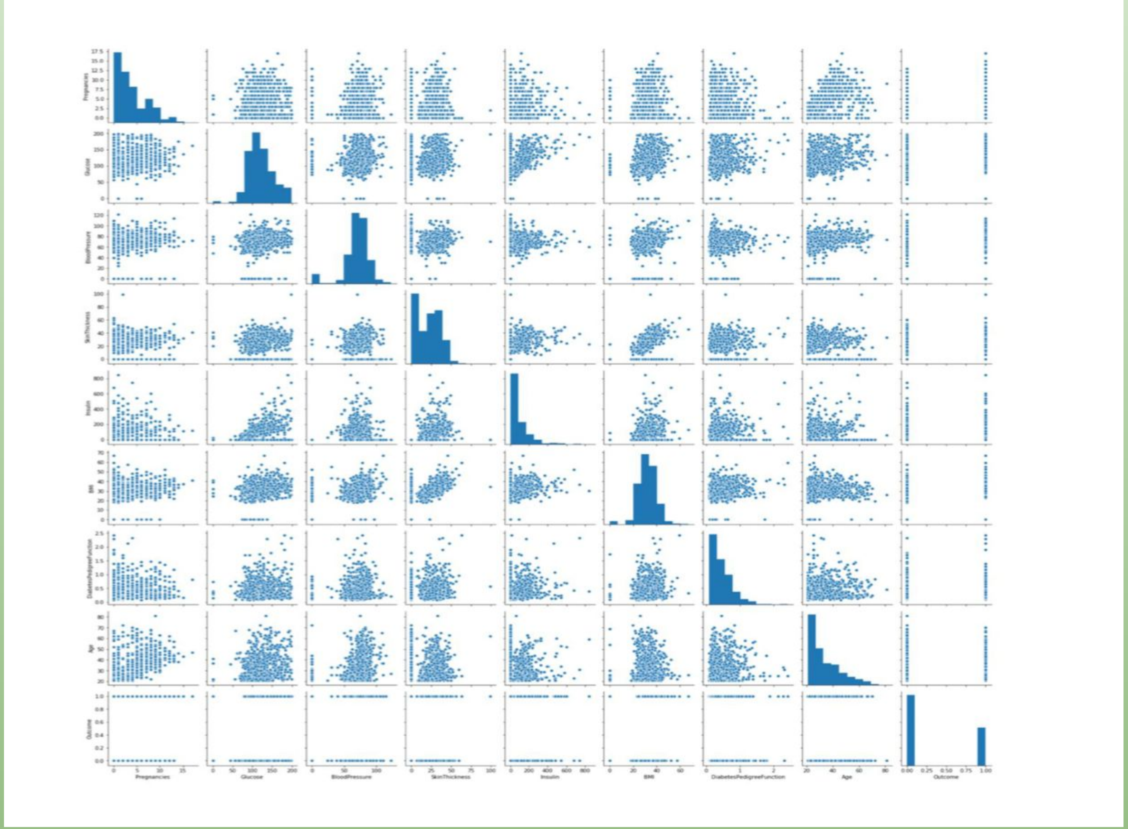


Fig 7: Pair Plot

We used a pair plot to visualize positively contributing attributes and negatively contributing attributes.

Exploratory Data Analysis - Numeric Variables

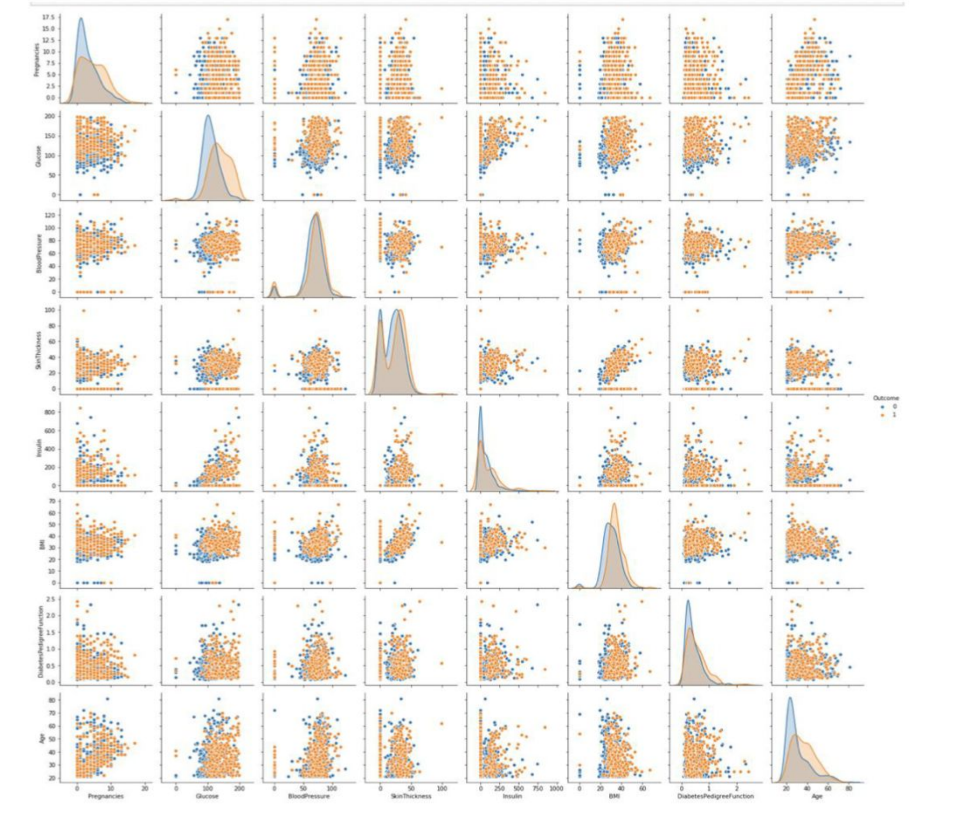


Fig 7: Pair Plot

Exploratory Data Analysis - Numeric Variables

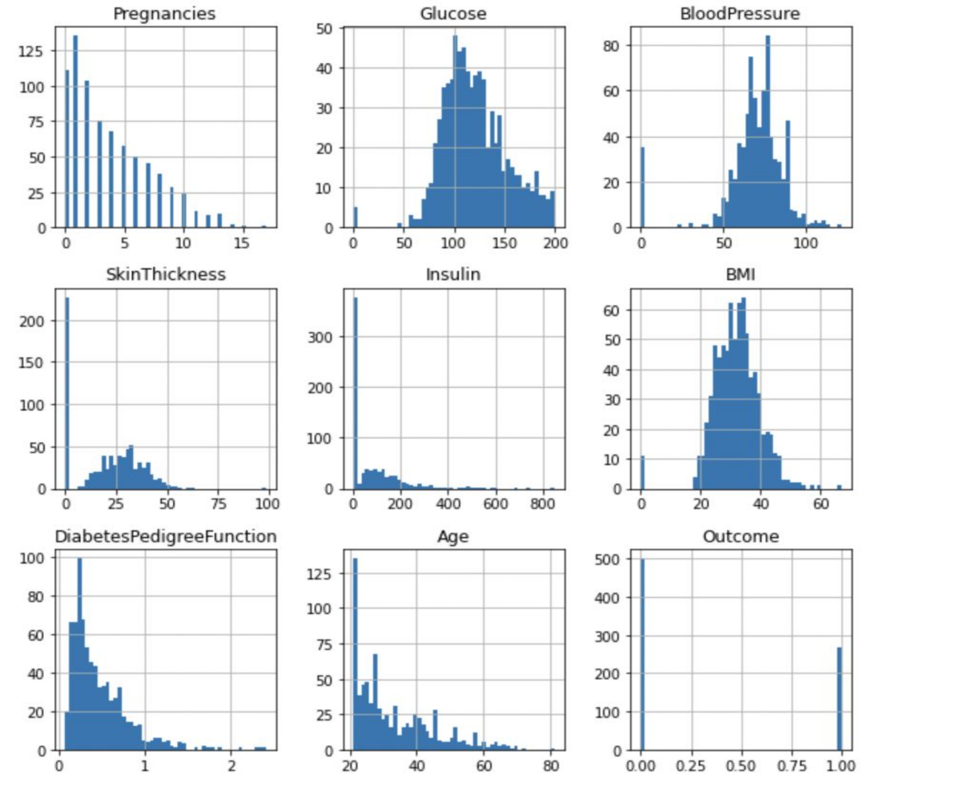


Fig 8: Histogram

We created a histogram which showed all the variables and how they are spread out.

Exploratory Data Analysis - Numeric Variables

We also created a box plot for outlier visualization with respect to attributes such as pregnancies, Glucose Level, Blood Pressure, Skin Thickness, Insulin, BMI, Age etc.

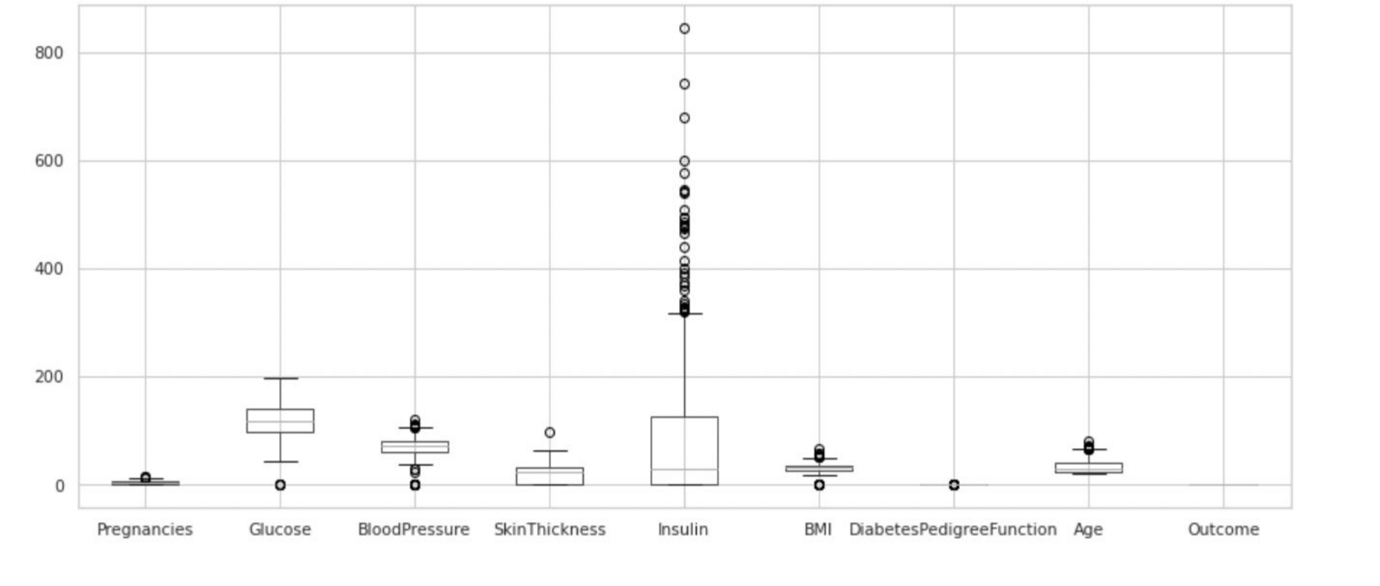


Fig 9: Box Plot

Exploratory Data Analysis - Categorical Variables

- We have only one categorical variable, the target variable ‘Outcome’, which is binary in nature.
- Therefore, to visualize this variable graphically we made use of the count plot() function from the ‘seaborne’ library.
- From count plot(), we can see that about 500 instances denote people who don’t have diabetes (indicated as ‘0’) & about 300 instances contain diabetes (indicated as ‘1’).

Exploratory Data Analysis - Categorical Variables

The countplot shows the occurrences of the diabetic & non-diabetic patients that are represented in the outcome column of the diabetes dataset

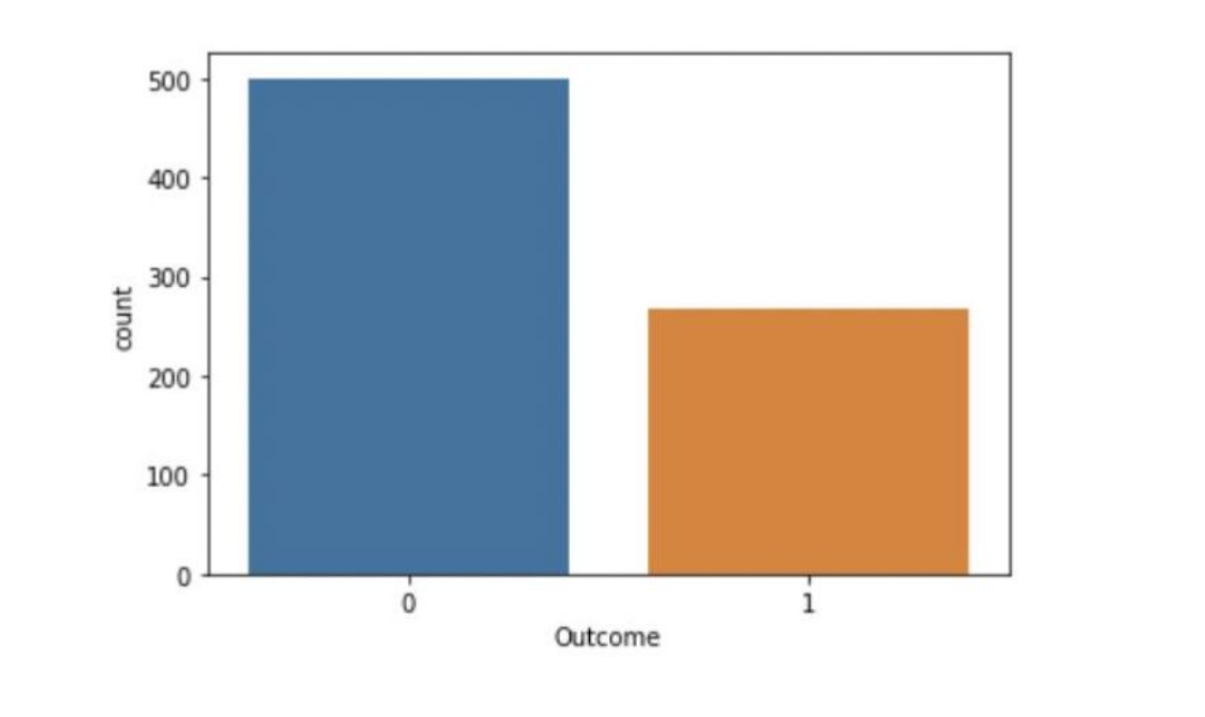


Fig 10: Count Plot

Model Building

We split the dataset into **80% training set** and **20 % testing set.**

The following models were used:

- Logistic Regression
- K-NN Classifier
- Random Forest
- Decision Tree

Model Building - Logistic Regression

Logistic Regression is a statistical model that uses the logistic function to model a binary dependent variable. It estimates the parameters of the logistic model and the outcome variable is categorical in nature.

Advantages:

Easier to implement, interpret, and very efficient to train

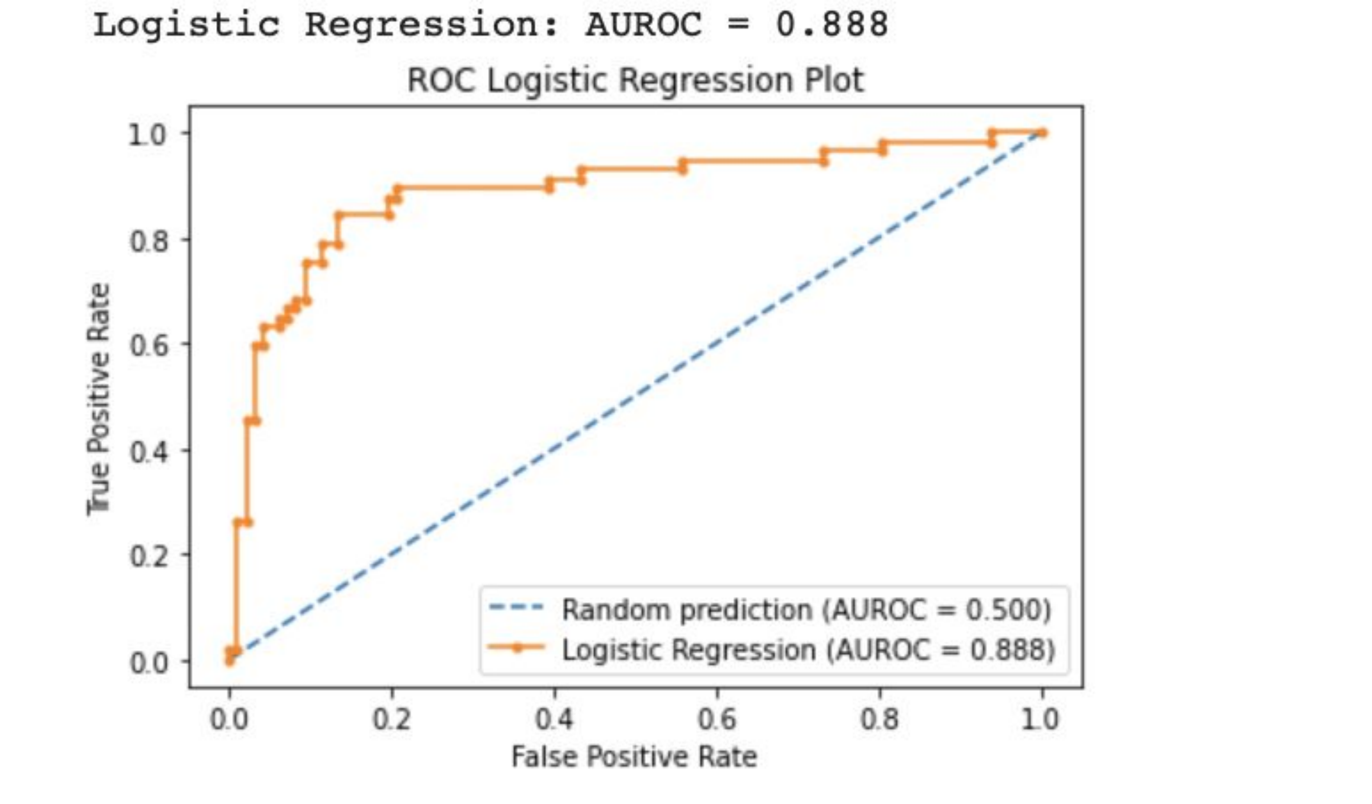
Performs efficiently when the dataset is linearly separable

Disadvantages

It constructs linear boundaries

Implementation: The base logistic regression model was executed, resulting in an accuracy of **83.7** %.

Performance Evaluation - Logistic Regression



```
Logistic Regression: AUROC = 0.888
```

Fig 11: ROC Curve of Logistic Regression

Model Building - K-NN Classifier

k-NN stands for k- Nearest Neighbours. It uses data and classifies new data points based on similarity measures.

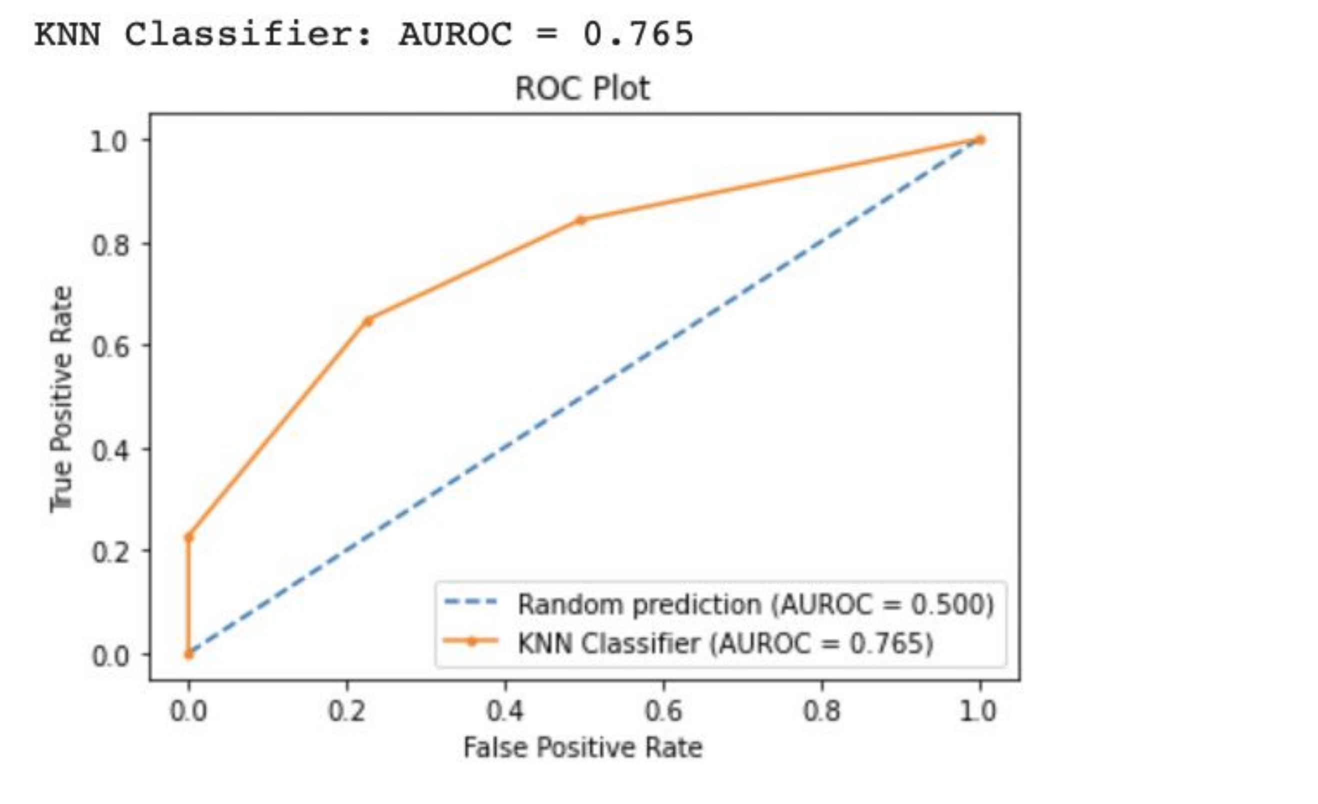
Advantages:

It is an instance-based learning method and does not learn anything during the training period

Disadvantages:
It does not work well with large datasets and high dimensions
Needs feature scaling(standardization) before applying kNN to any dataset

Implementation: The base K-NN model was executed, resulting in an accuracy of 72.7%.

Performance Evaluation - K-NN Classifier



```
KNN Classifier: AUROC = 0.765
```

Fig 12: ROC Curve of K-NN Classifier

Model Building - Random Forest

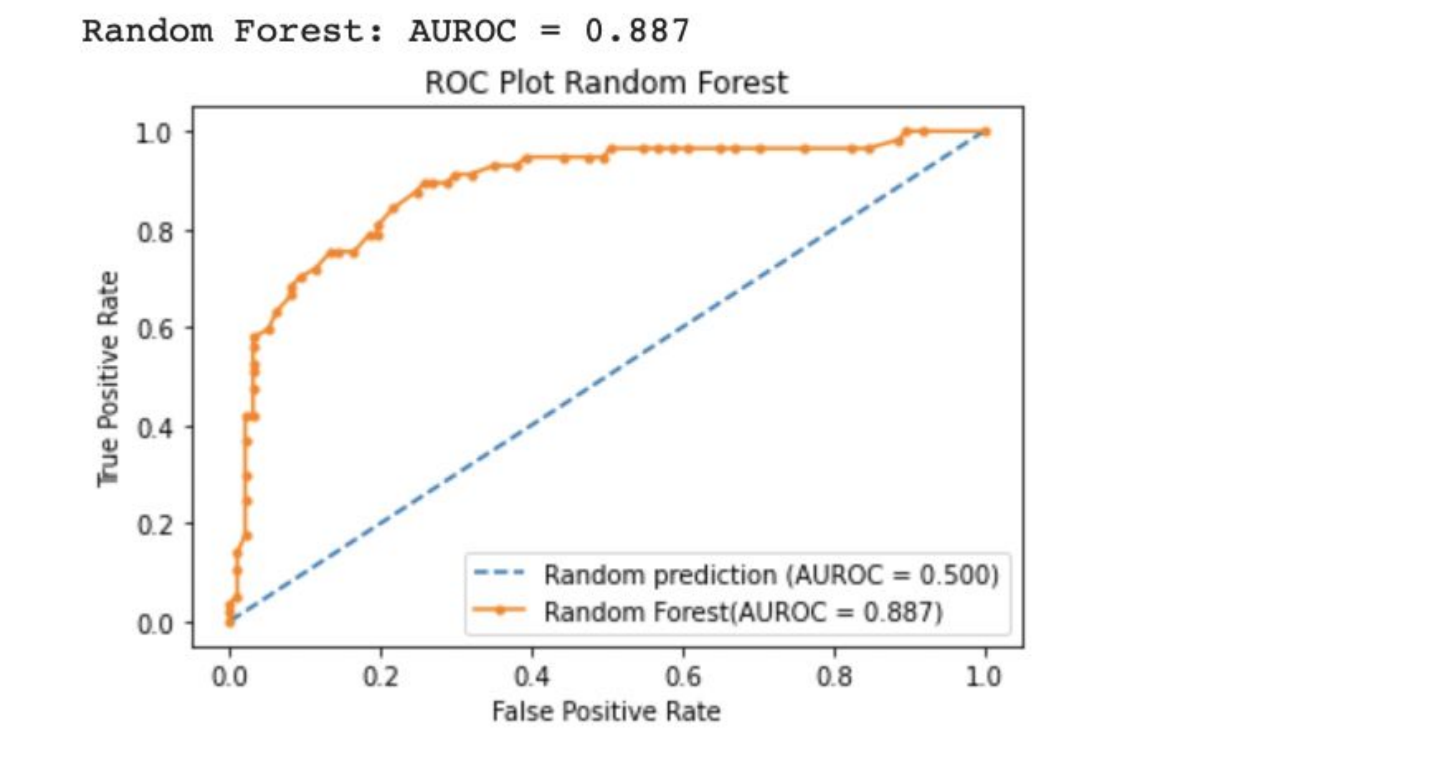
It is a classification algorithm consisting of many decision trees. It builds decision trees on different samples and takes their majority vote for classification and average in the case of regression.

Advantages:
It can be used to solve both classification and continuous variables
No feature Scaling is required

Disadvantages:
It can create a lot of trees and combine their outputs thus increasing its complexity

Implementation: The base Random Forest model was executed, resulting in an accuracy of 82.4%.

Performance Evaluation - Random Forest



```
Random Forest: AUROC = 0.887
```

Fig 13: ROC Curve of Random Forest

Model Building - Decision Tree

Decision Trees are a type of Supervised Machine Learning where the data is continuously split according to a certain parameter. The goal is to create a training model that can use to predict the class or value of the target variable.

Advantages:

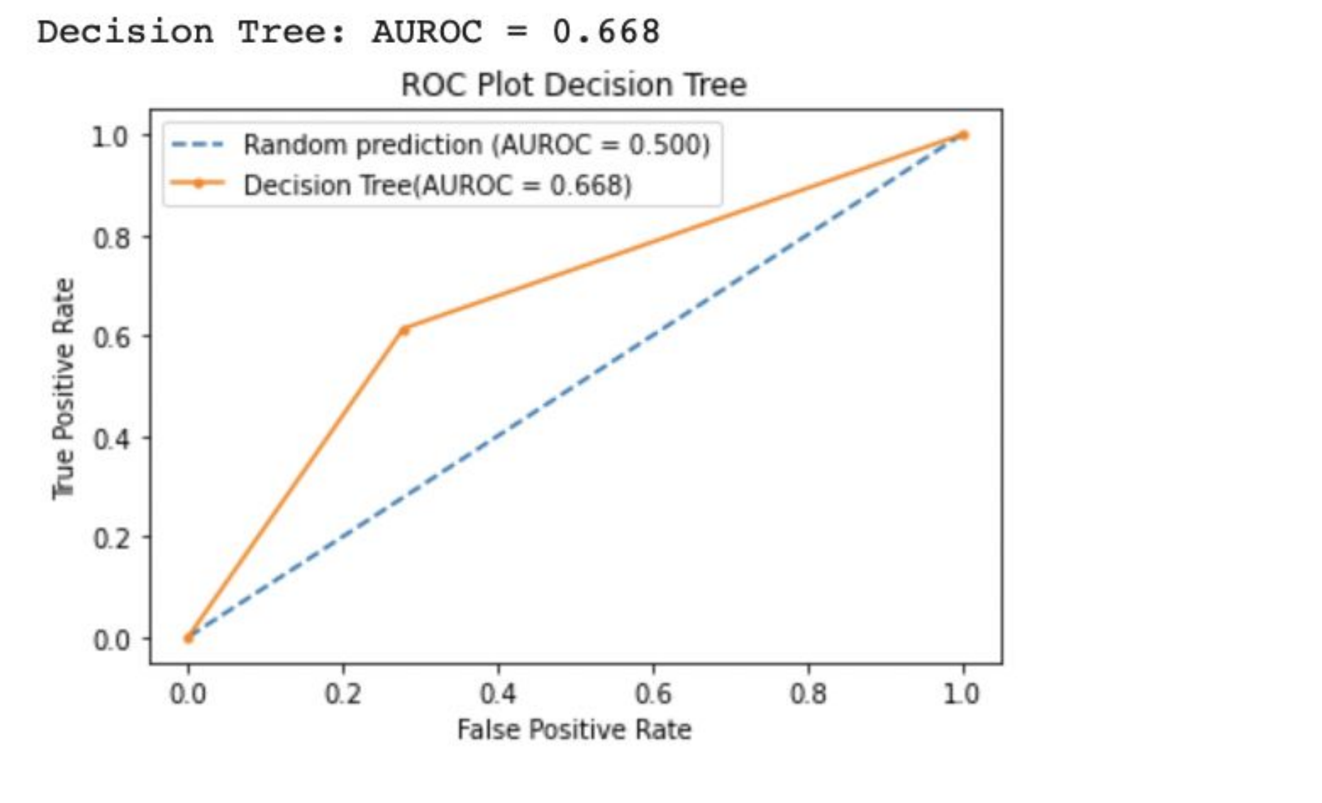
It requires less effort for data preparation during pre-processing compared to other algorithms

Disadvantages:

It often involves higher time to train the model

Implementation: The base Decision Tree model was executed, resulting in an accuracy of 68.6%.

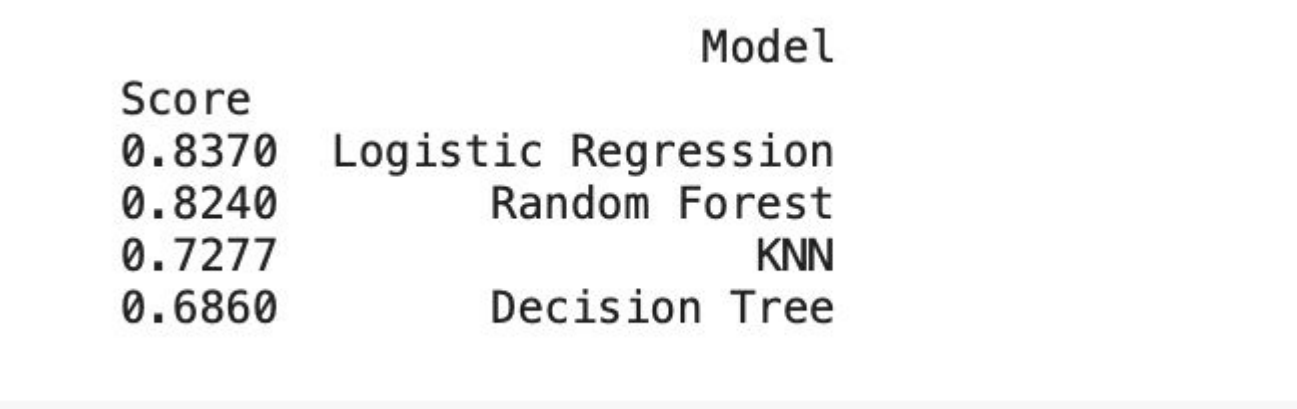
Performance Evaluation - Decision Tree



```
Decision Tree: AUROC = 0.668
```

Fig 14: ROC Curve of Decision Tree

Final Result



```
                          Model
Score
0.8370  Logistic Regression
0.8240        Random Forest
0.7277                  KNN
0.6860        Decision Tree
```

Fig 15: Final Result showing scores of the models

As shown in the figure, Logistic Regression has greatest accuracy of 83.70% making it the best model to consider while predicting diabetes in a person.

Random Forest ranks second as it has less accuracy than logistic regression i.e. 82.40%

Challenges Faced during the project

- Getting a dataset of the desired size was difficult. Either the datasets were too huge or too small.
- Initially we were not able to obtain desired accuracy of the model, but after performing all data mining steps carefully we were able to obtain desired accuracy.

Future Scope

- Proposed system uses "Logistic Regression algorithm" to find the diabetes disease, in data science we have many algorithms for classification such as Naive Bayes, SVM, ID3 etc... in future we can add more algorithms to find outputs and algorithms can be compared to find the efficient algorithm.

- Further we can build a Diabetes prediction website where we can add treatment module & doctors can upload treatment details for patients & patient can view those treatment details.

Thank You!

REGULATION
T2D
INSULIN
MELLITUS
INJECT
BLOOD SUGAR
ENERGY
EXERCISE
COMPLICATION
WEIGHT
BETA CELLS
DIET
DIABETES
SUGAR
HORMONE
BODY
HEALTH
PANCREAS
PREVENTION
BLOOD PRESSURE
BLINDNESS
AUTOIMMUNE DISEASE
FOOD
BLOOD GLUCOSE