Predicting the occurence of Diabetes in Pregnant Women

Authors:
R. Harini (18BCE 1010)
Aanya Jain (18BCE1067)
Shraddha Nair (18BCE1070)

Guide: Prof. Pattabiraman V.

INTRODUCTION

Diabetes is a chronic disease associated with abnormally high levels of the sugar glucose in the blood. Although it has no cure, it can still be predicted in advance using the emerging technologies of the day. Our project tries to predict the occurrence of diabetes in pregnant women by using various prediction and classification models like Logistic Regression, Decision Tree, Naïve Bayes, K Nearest Neighbours and Random Forest

OBJECTIVE

The key objectives of this project are:

-- Learning and Understanding the working of Logistic Regression, Decision Tree, Naive Bayes, KNN and Random Forest

-- Applying these models on the dataset used, to predict the occurence of diabetes in women

-- Comparing the accuracy of these models in predicting diabetes in pregnant women

-- Visualizing the relationship and correlation between the input features, the effect of each feature on the outcome, and the accuracy of all the models in predicting diabetes occurence in women

-- Finding the model that gives the best accuracy for the used dataset.

DATASET

The data was collected and made available by "National Institute of Diabetes and Digestive and Kidney Diseases" as part of the Pima Indians Diabetes Database. Its features are:

Pregnancies: Number of times pregnant
Glucose: Plasma glucose concentration
BloodPressure: Diastolic blood pressure (mm Hg)
SkinThickness: Triceps skin fold thickness (mm)
Insulin: 2-Hour serum insulin (mu U/ ml)
BMI: Body mass index (kg/ m2)
DiabetesPedigreeFunction: Function which scores likelihood of diabetes based on family history Age: Age (years)
Outcome: Class variable (0 if non-diabetic, 1 if diabetic)

WORKFLOW

Data preprocessing

Data Visualization

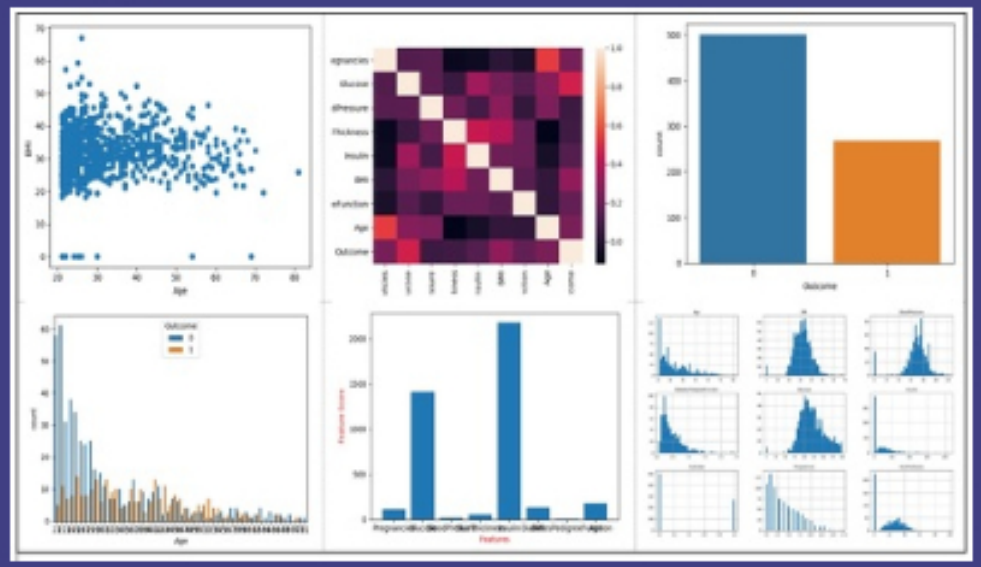
Prediction using models

Comparison of Accuracy

Result Visualization

Conclusion

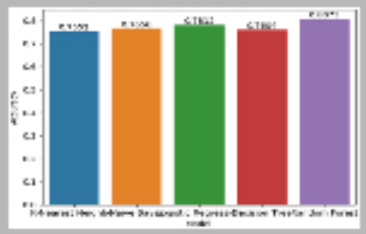
DATA VISUALIZATION



RESULT AND CONCLUSION

On applying all the models on the dataset, the calculated accuracies were as follows:

KNN : 0.7552 Naive Bayes : 0.7656
Logistic Regression : 0.7812
Decision Tree : 0.7604
Random Forest : 0.8073



Thus, based on the result a conclusion can be made that Random Forest Model makes the most accurate predictions for the chosen dataset. On the other hand, the KNN model does not give a very accurate prediction when compared to the other models. Thus, diabetes in women can now be predicted with an accuracy of about 0.81, at an earlier stage.This will allow pregnant women to start taking medications at an earliear stage and thus prevent the disease from causing too much harm.