Case Study: Enhancing Predictive Accuracy of Diabetes Diagnosis Among Female Patients

Background

Diabetes is a chronic condition that affects millions of people worldwide, leading to severe health complications if not properly managed. Early diagnosis is crucial for effective treatment and management, particularly in high-risk populations. The Pima Indian community has been identified as having a higher prevalence of diabetes, making it a significant area of study.
This case study focuses on using machine learning techniques to improve the predictive accuracy of diabetes diagnosis among female patients from the Pima Indian community. The dataset used for this analysis was obtained from the National Institute of Diabetes and Digestive and Kidney Diseases and is publicly available on Kaggle.

Objective

The primary goal of this project is to investigate and implement various machine learning algorithms to enhance the accuracy of predicting diabetes among female patients. The project aims to identify the most effective model for this purpose and to understand the key predictors of diabetes within this population.

Problem statement

Investigate and implement machine learning algorithms to improve the predictive accuracy of diabetes diagnosis among female patients.

Methodology

Dataset Overview

- Source: The dataset was sourced from https://kaggle.com and originally comes from the National Institute of Diabetes and Digestive and Kidney Diseases. Link: https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset/data
- Structure: The dataset consists of 768 rows and 9 columns.

Features

- Pregnancies: Number of pregnancies the subject has had.
- Glucose: Glucose concentration in the blood.
- Blood Pressure: Blood pressure measurement.
- Skin Thickness: Triceps skinfold thickness.
- Insulin: Serum insulin levels.
- BMI: Body Mass Index (BMI).
- DiabetesPedigreeFunction: A function that quantifies the genetic predisposition to diabetes.
- Age: The age of the subject.
- Outcome: The target variable indicating whether the subject is diabetic (1) or not (0).

Exploratory Data Analysis

- Null Values: The dataset contains no missing values.
- Duplicates: No duplicate rows were found in the dataset.
- Descriptive Statistics: Summary statistics were computed to understand the distribution of each variable.
- Correlation Analysis: Correlation among predictor variables and with the outcome variable was analyzed.

Preprocessing

- Outliers Removal: Outliers were identified and removed to reduce their impact on model performance. Techniques such as IQR and Z-score were employed to detect outliers.
- Feature and Target Splitting: The dataset was divided into features (independent variables) and the target variable (Outcome).
- Data Scaling: Features were scaled to ensure uniformity in the model training process, which helps in improving model performance, especially for algorithms like Support Vector Machines and Logistic Regression.
- Train-Test Split: The dataset was split into training and testing sets to evaluate the model's generalization performance.

Model Building

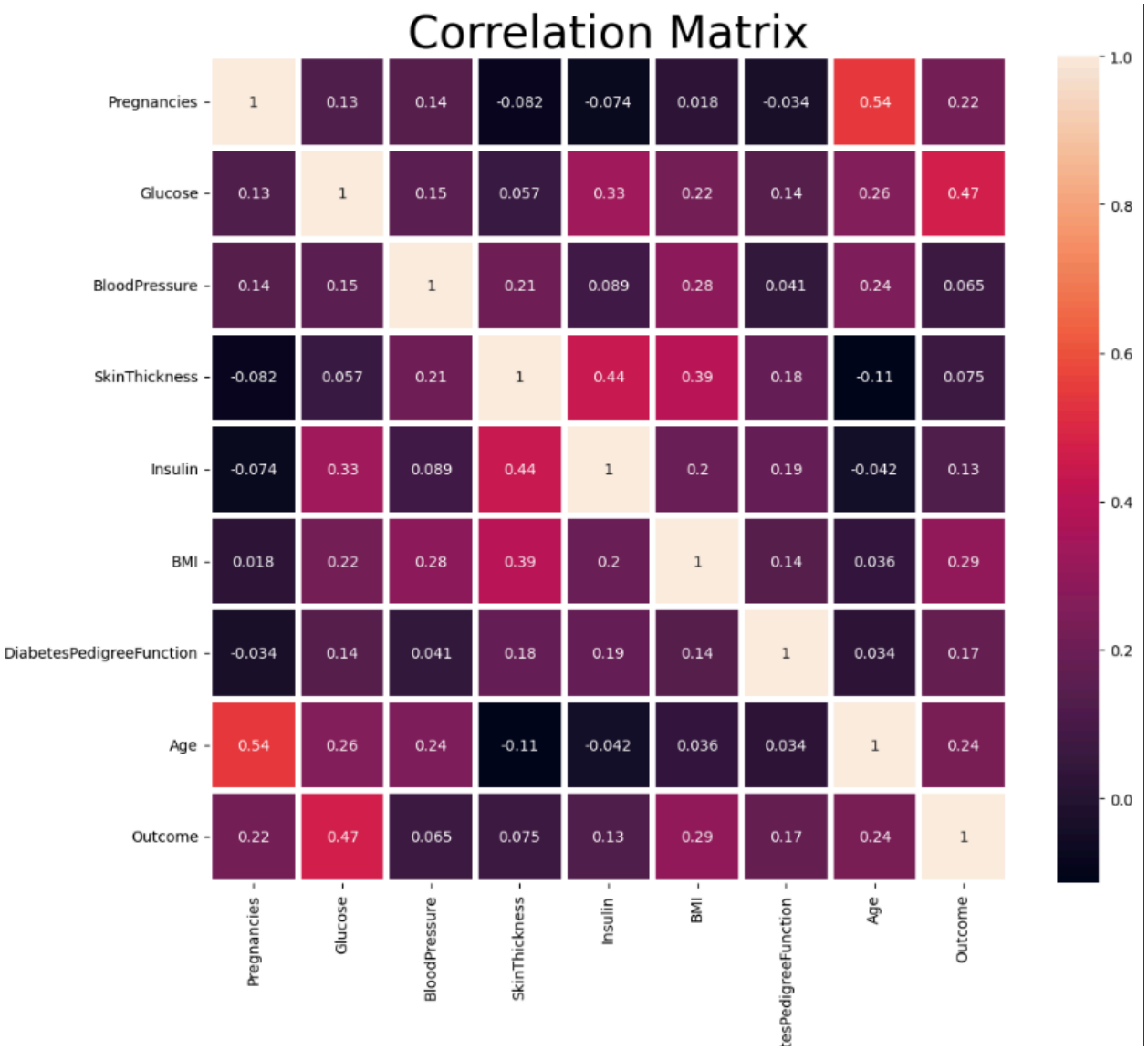
- Logistic Regression: A model was built using logistic regression to predict the probability of diabetes. Logistic regression was chosen for its simplicity and interpretability.
- Random Forest Classifier: A random forest classifier was employed to leverage the power of ensemble learning, improving prediction accuracy and robustness by combining multiple decision trees.
- Support Vector Machines (SVM): SVM was used to find the optimal hyperplane that maximizes the margin between classes, effective for classification tasks with a clear margin of separation.

Model Evaluation

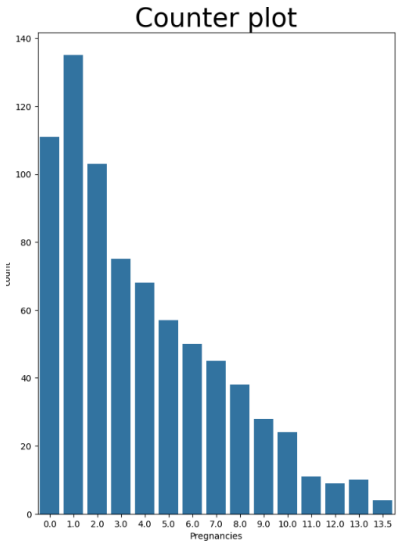
- Accuracy: The overall correctness of the model's predictions was measured as the ratio of correctly predicted instances to the total instances.
- Recall: The ability of the model to correctly identify positive cases (patients with diabetes) was evaluated to understand how well the model performs in detecting the target condition.
- F1 Score: The harmonic mean of precision and recall was calculated to provide a balance between the two, especially useful in cases where the class distribution is imbalanced.

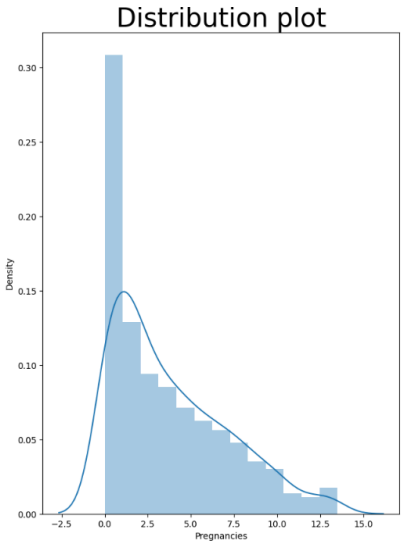
Model	**Accuracy**	**Recall**	**F1 Score**	**Average**
Logistic Regression	81.82%	74.36%	67.44%	74.54%
Random Forest Classifier	79.87%	69.05%	65.17%	71.36%
Support Vector Classifier	80.52%	71.79%	65.12%	72.48%

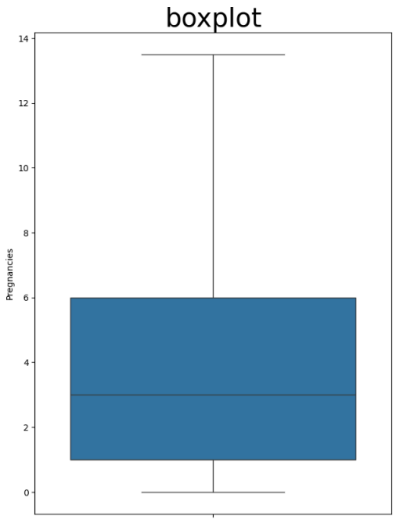
Visualizations



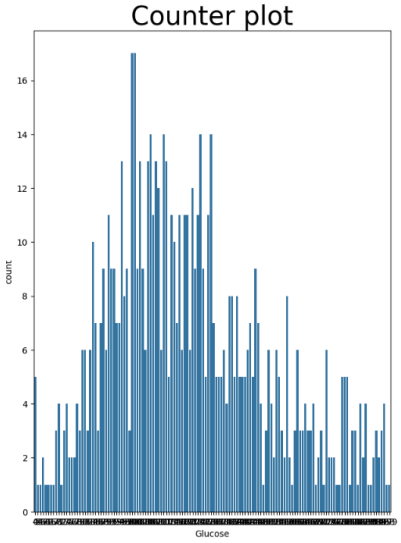
Pregnancies Visualizations:

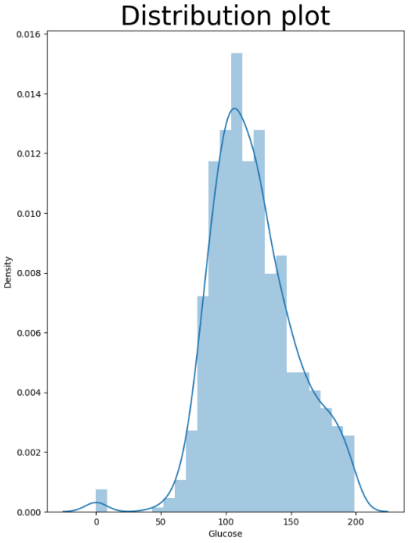


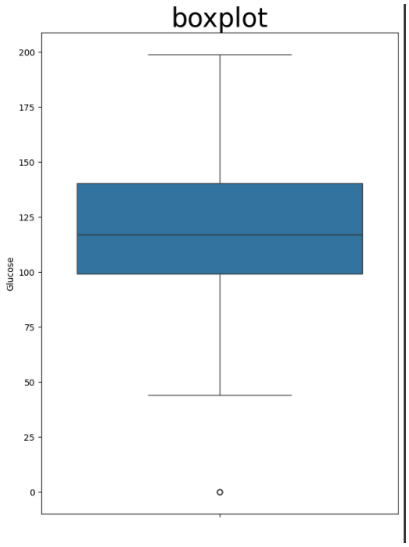




Glucose Visualizations:

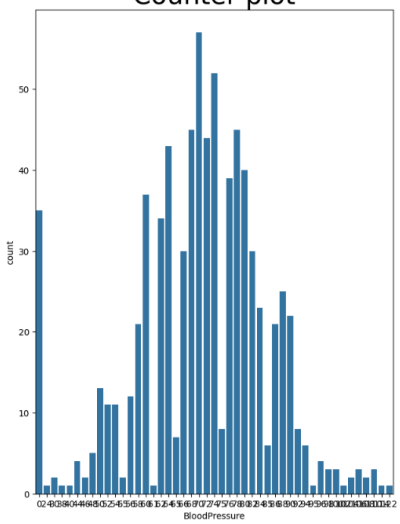




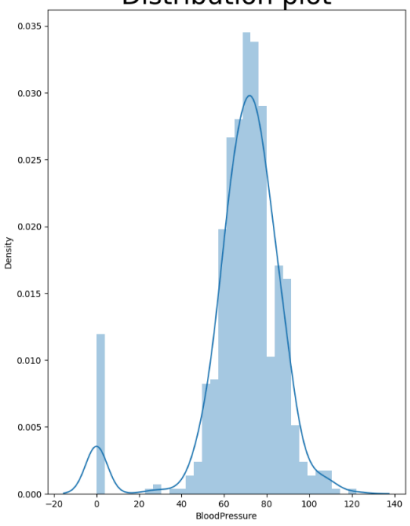


Blood Pressure Visualizations:

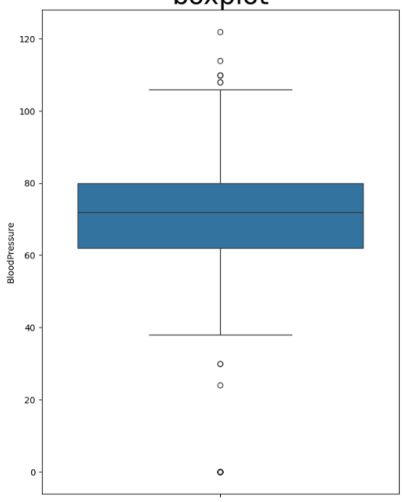
Counter plot



Distribution plot

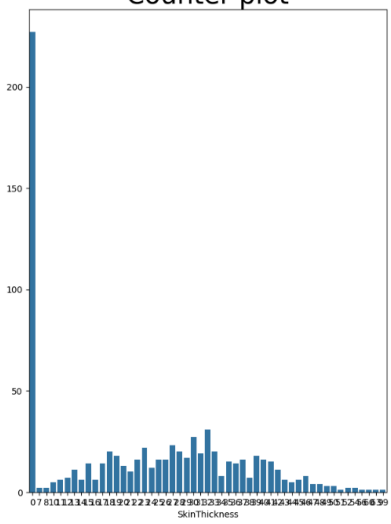


boxplot

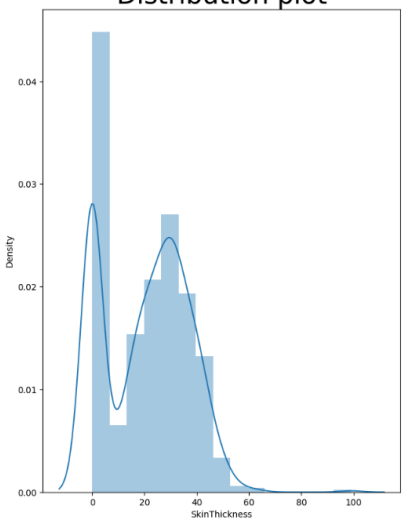


Skin Thickness Visualizations

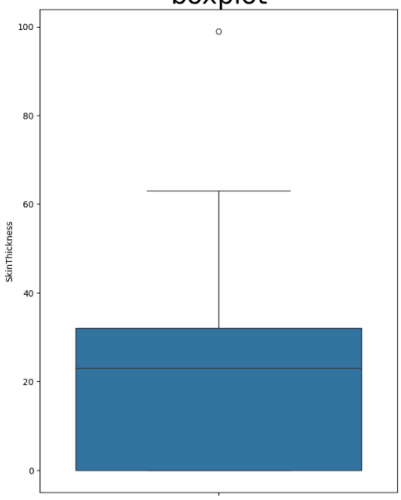
Counter plot

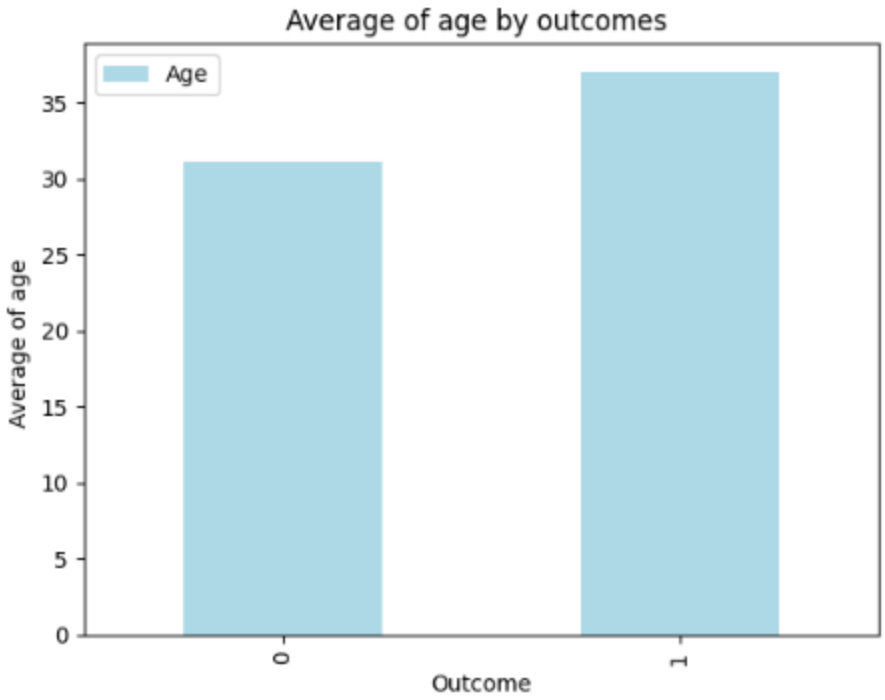


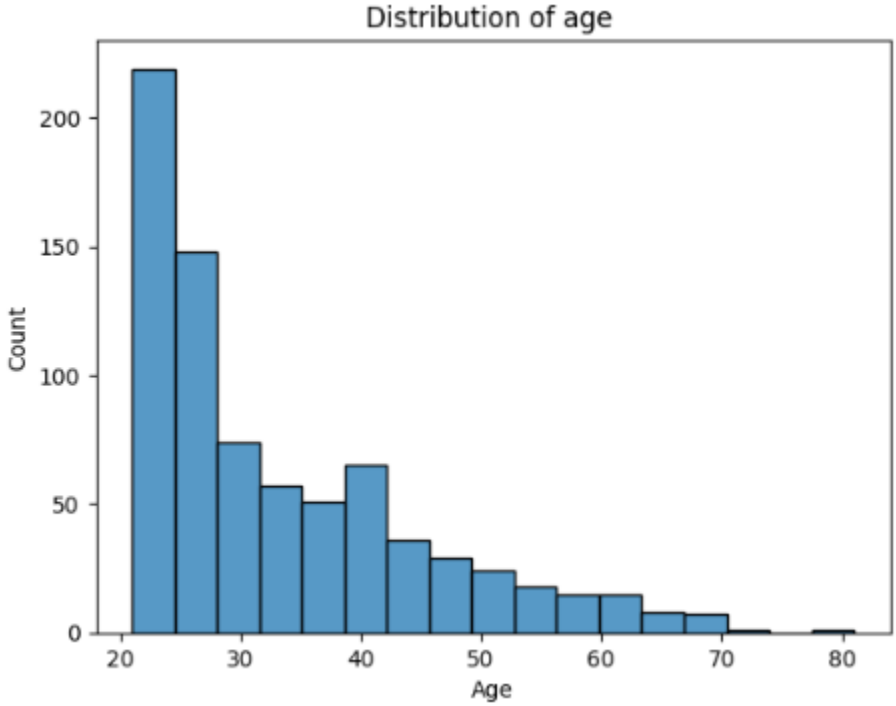
Distribution plot

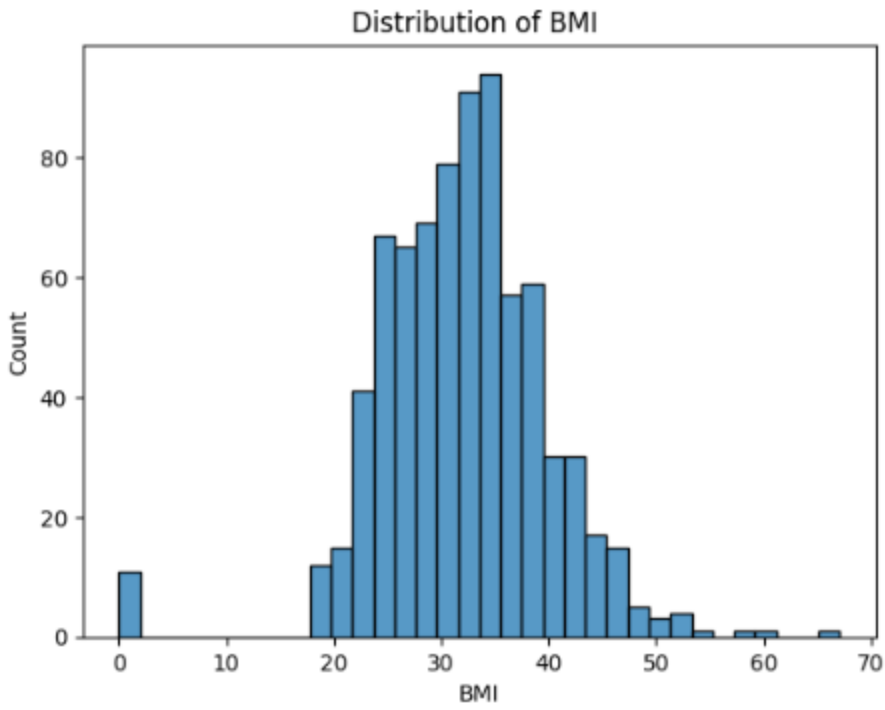


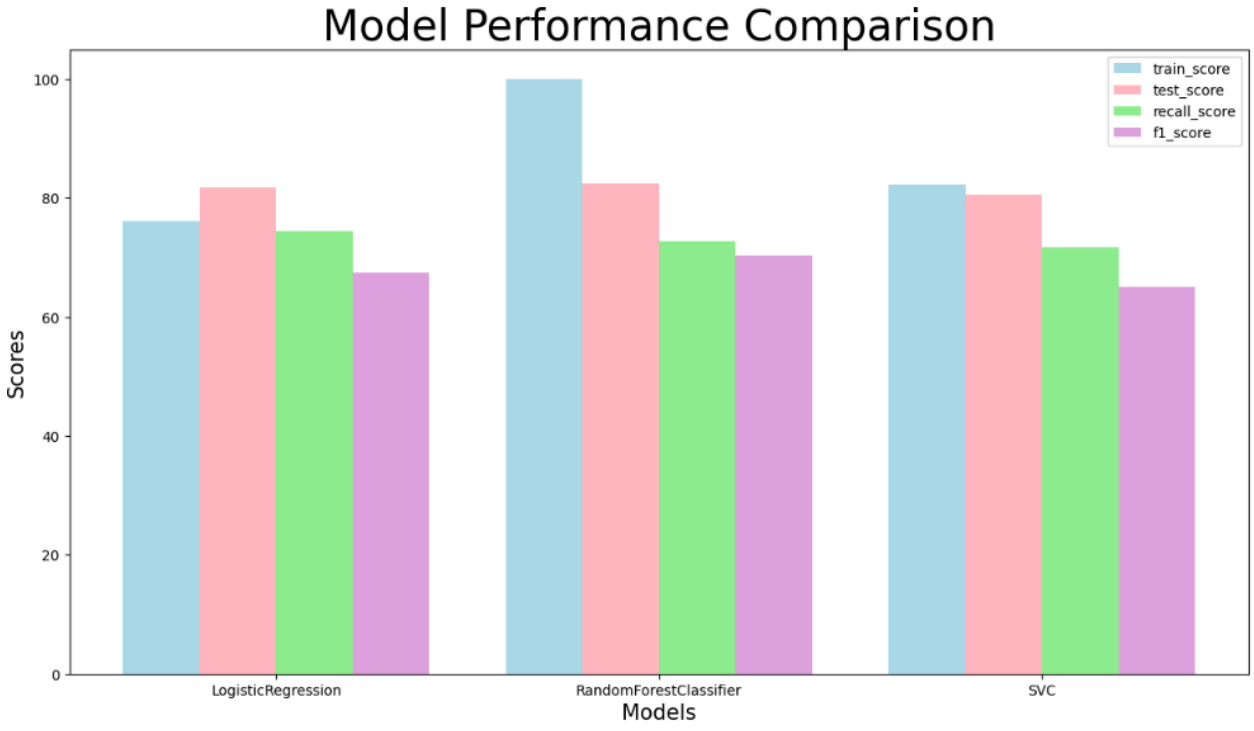
boxplot

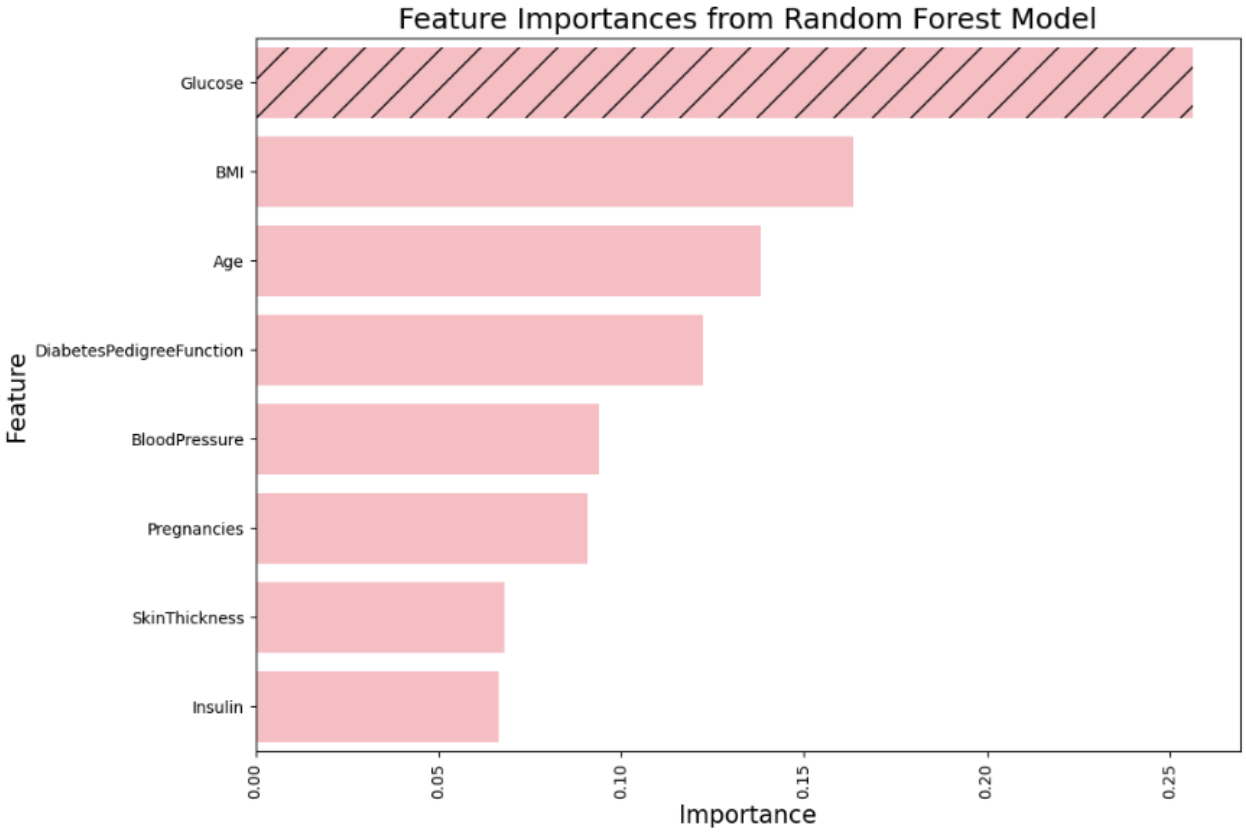


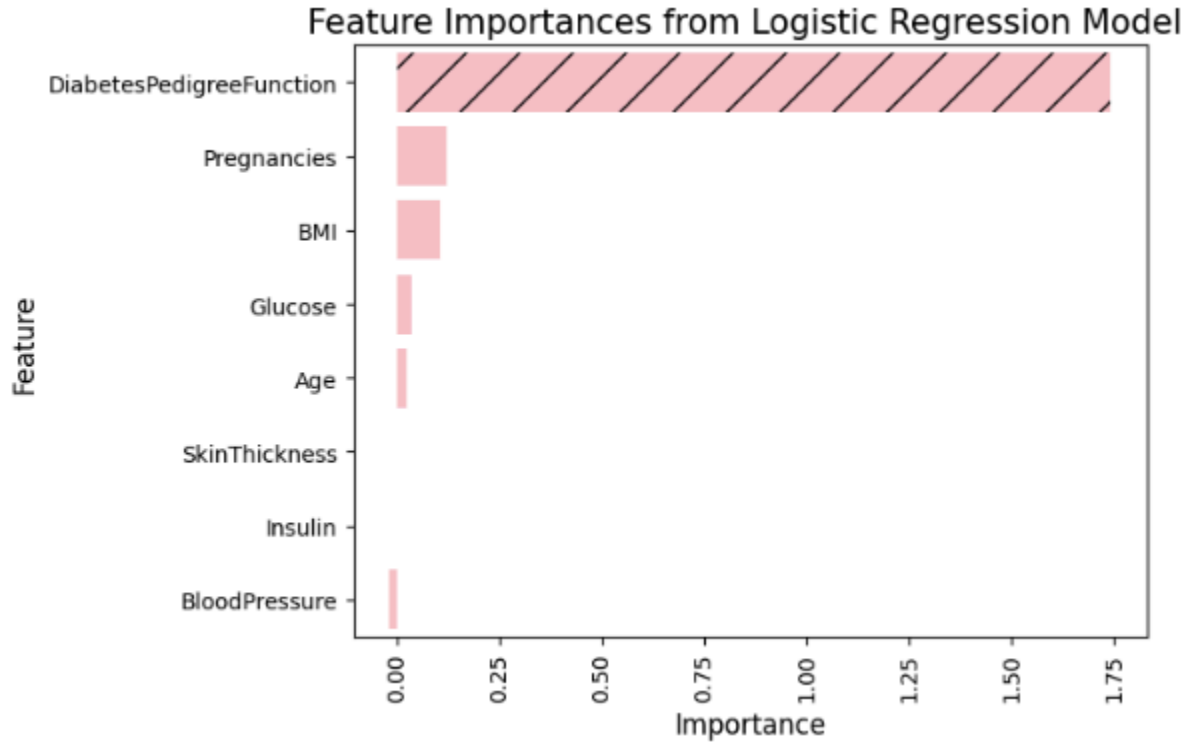

Average of age by outcomes
Age
35
30
25
20
15
10
5
0
Average of age
0
1
Outcome


Distribution of age
200
150
100
50
0
Count
20
30
40
50
60
70
80
Age


Distribution of BMI
Count
BMI
0
20
40
60
80
0
10
20
30
40
50
60
70


Model Performance Comparison
Scores
Models
train_score
test_score
recall_score
f1_score
0
20
40
60
80
100
LogisticRegression
RandomForestClassifier
SVC


Feature Importances from Random Forest Model
Glucose
BMI
Age
DiabetesPedigreeFunction
BloodPressure
Pregnancies
SkinThickness
Insulin
Feature
0.00
0.05
0.10
0.15
0.20
0.25
Importance


Feature Importances from Logistic Regression Model
DiabetesPedigreeFunction
Pregnancies
BMI
Glucose
Age
SkinThickness
Insulin
BloodPressure
Feature
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
Importance

Key Findings

- Among all the features, glucose had the highest correlation with the outcome target.
- Women with a high number of pregnancies have a significantly low skin thickness, low insulin, low BMI and low diabetes pedigree function.
- Older women have a significantly higher number of pregnancies than younger women - as would ideally be the case.
- All the models performed fairly well. However, the model that performed best was Logistic Regression with an average score of 74.54% while the one that performed worst was Random Forest Classifier with an average score of 71.36%.
- The average age of the women who were diabetic was 37 while the average age of those who weren't diabetic was 31. Hence, we can deduce that older women are more susceptible to diabetes than younger ones.
- The Random Forest Classifier scores pretty highly on the training set (99% +) but generalizes relatively low on the test set (around 82%), This means the model overfits the training set.