Student Name: Shashank B V

Student No: R00224414

Title: Discerning Diabetes Prognoses Utilizing Sophisticated Machine Learning Approaches

Subject: Applied Machine Learning

Declaration of Authorship

I, SHASHANK B V, declare that the work presented in this assignment, titled *Applied Machine Learning* , is my own. I confirm that:

- This work was done wholly by me as part of my MSc. in Data Science & Analytics.
- Where I have consulted the published work and source code of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this assignment source code is entirely my own work.

08-05-2023

On ___DATE___

Signature:

SHASHANK B V

Abstract

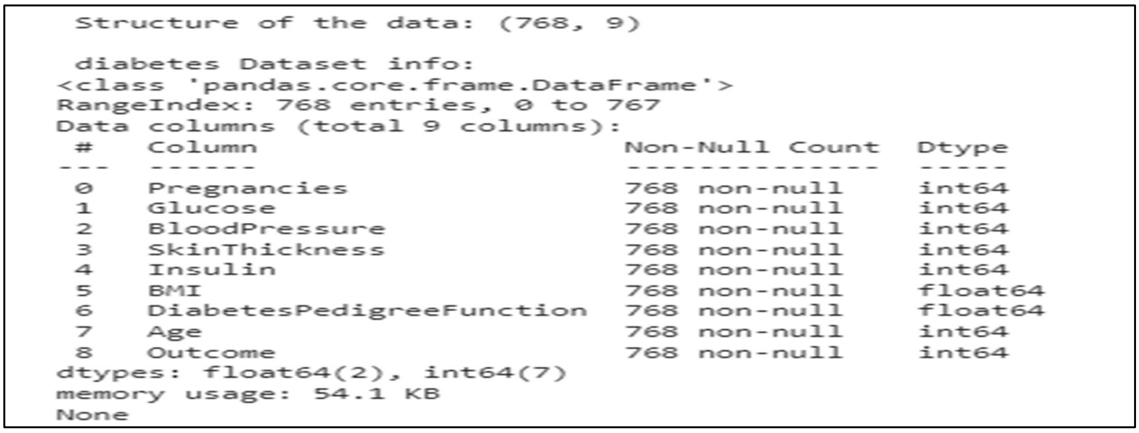
This endeavor aspires to prognosticate diabetes outcomes employing an array of intricate machine learning methodologies on the Pima Indian Diabetes dataset. The inquiry examines the ramifications of disparate pre-processing stages, feature selection stratagems, and model optimization to enhance predictive prowess. Evaluation of the models encompasses numerous metrics, encompassing accuracy, precision, recall, F1-score, and AUC-ROC score. The findings reveal that Random Forests and Support Vector Machines excel amongst the scrutinized models.

Introduction

Diabetes is a persistent health malady afflicting a multitude of individuals globally. Prompt detection and judicious management can substantially mitigate the risk of complications. This investigation utilizes the Pima Indian Diabetes dataset, comprising medical records of 768 female patients of Pima Indian lineage. The dataset incorporates attributes such as gravidity, glycaemia, sphygmomanometry, cutaneous thickness, insulin, BMI, Diabetes Pedigree Function, and chronological age. The target variable delineates the diabetes outcome, either affirmative (1) or negative (0). The impetus for this investigation is the construction of precise and dependable models for predicting diabetes outcomes, contributing to enhanced healthcare adjudication. Antecedent scholarly articles have employed this dataset to examine various machine learning techniques and their efficacy in predicting diabetes outcomes.

Descriptive Statistics

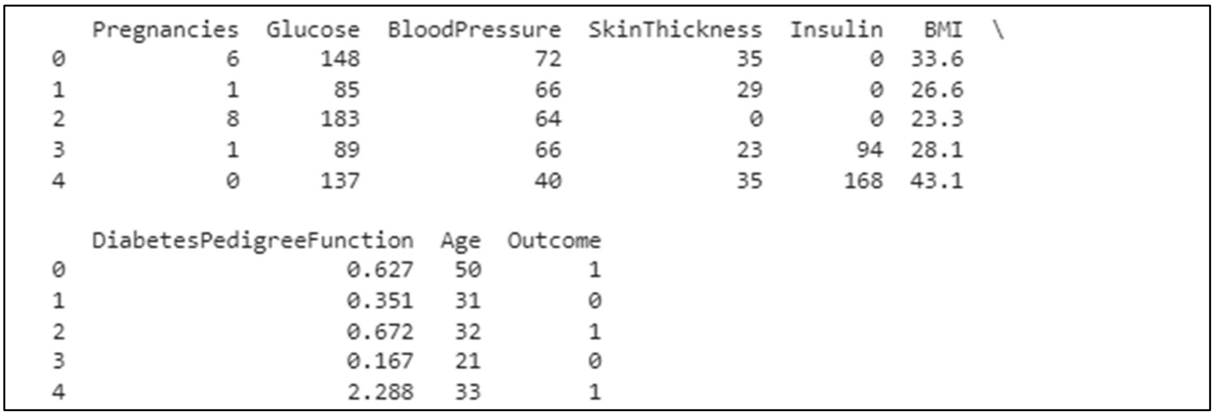
- This is the structure of the data, it contains 768 rows and 9 columns



```
 Structure of the data: (768, 9)

 diabetes Dataset info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   Pregnancies               768 non-null    int64  
 1   Glucose                   768 non-null    int64  
 2   BloodPressure             768 non-null    int64  
 3   SkinThickness             768 non-null    int64  
 4   Insulin                   768 non-null    int64  
 5   BMI                       768 non-null    float64
 6   DiabetesPedigreeFunction  768 non-null    float64
 7   Age                       768 non-null    int64  
 8   Outcome                   768 non-null    int64  
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
None
```

- This is the head of the data



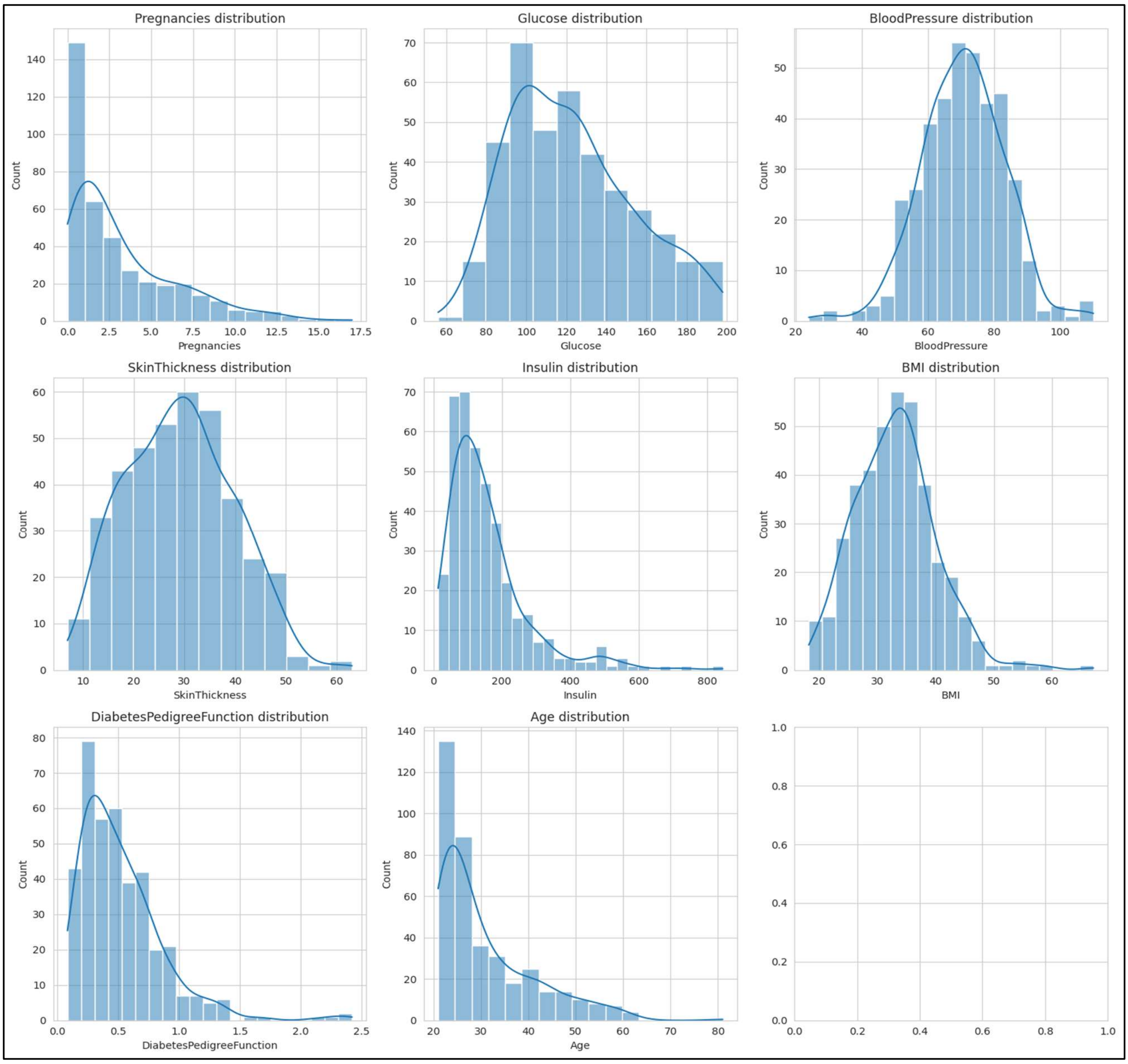
```
   Pregnancies  Glucose  BloodPressure  SkinThickness  Insulin   BMI  \
0            6      148             72             35        0  33.6   
1            1       85             66             29        0  26.6   
2            8      183             64              0        0  23.3   
3            1       89             66             23       94  28.1   
4            0      137             40             35      168  43.1   

   DiabetesPedigreeFunction  Age  Outcome  
0                     0.627   50        1  
1                     0.351   31        0  
2                     0.672   32        1  
3                     0.167   21        0  
4                     2.288   33        1  
```

- The mean, standard deviation, and quartiles have been discerned for each attribute, as demonstrated below:
- Gravidity: Mean (3.85), Standard Deviation (3.37), 1st Quartile (1), Median (3), 3rd Quartile (6), Maximum (17)
- Glycaemia: Mean (120.89), Standard Deviation (31.97), 1st Quartile (99), Median (117), 3rd Quartile (140.25), Maximum (199)
- Sphygmomanometry: Mean (69.11), Standard Deviation (19.36), 1st Quartile (62), Median (72), 3rd Quartile (80), Maximum (122)
- Cutaneous Thickness: Mean (20.54), Standard Deviation (15.95), 1st Quartile (0), Median (23), 3rd Quartile (32), Maximum (99)
- Insulin: Mean (79.80), Standard Deviation (115.24), 1st Quartile (0), Median (30.5), 3rd Quartile (127.25), Maximum (846)
- BMI: Mean (31.99), Standard Deviation (7.88), 1st Quartile (27.3), Median (32), 3rd Quartile (36.6), Maximum (67.1)
- Diabetes Pedigree Function: Mean (0.47), Standard Deviation (0.33), 1st Quartile (0.24), Median (0.37), 3rd Quartile (0.63), Maximum (2.42)

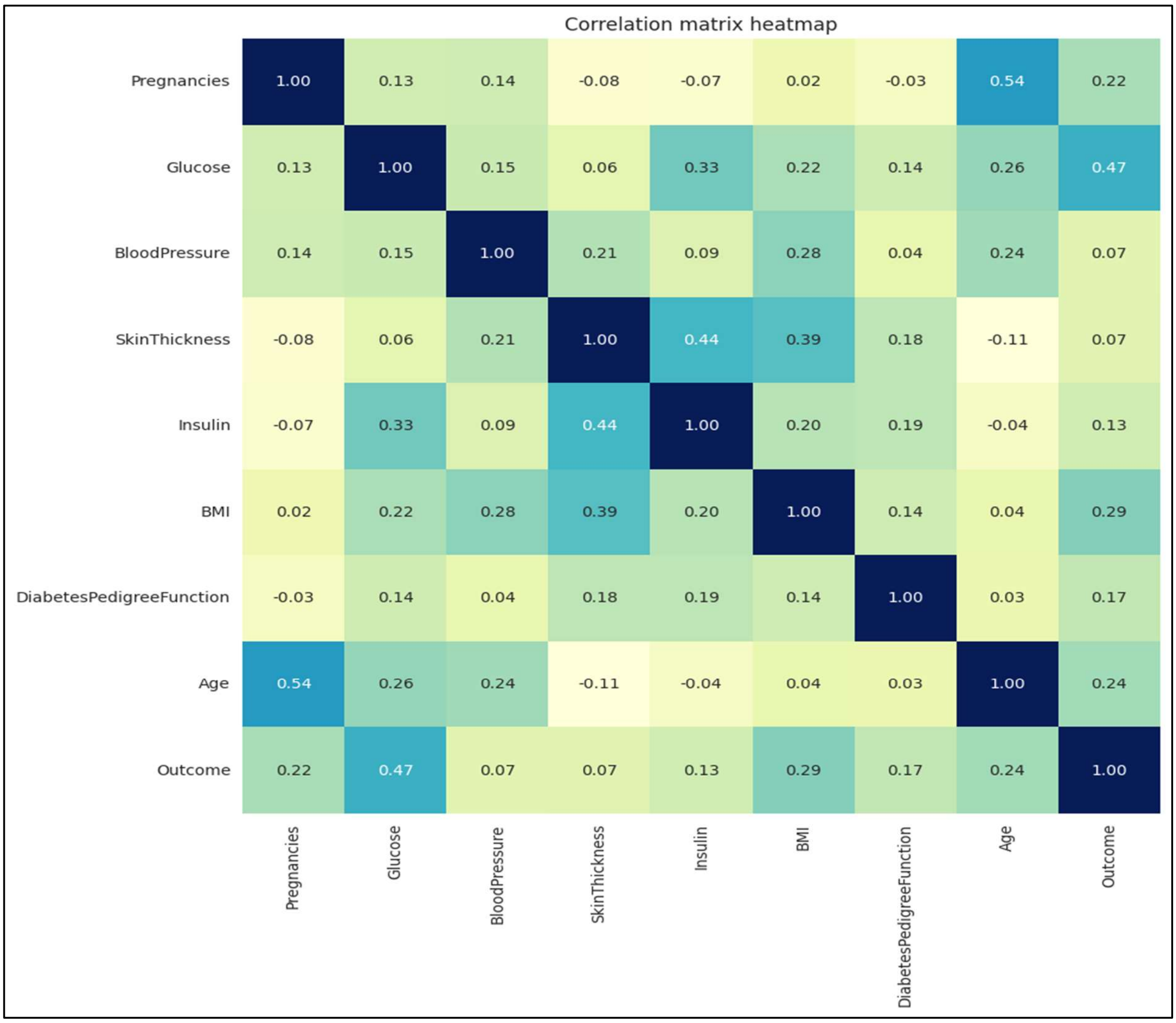
- Chronological Age: Mean (33.24), Standard Deviation (11.76), 1st Quartile (24), Median (29), 3rd Quartile (41), Maximum (81)
- Diabetes Prognosis: Mean (0.35), Standard Deviation (0.48), 1st Quartile (0), Median (0), 3rd Quartile (1), Maximum (1)
- These statistical revelations afford a deeper comprehension of the dataset's inherent traits, facilitating the implementation of efficacious machine learning techniques for diabetes prognosis prediction.

Distribution of the variables

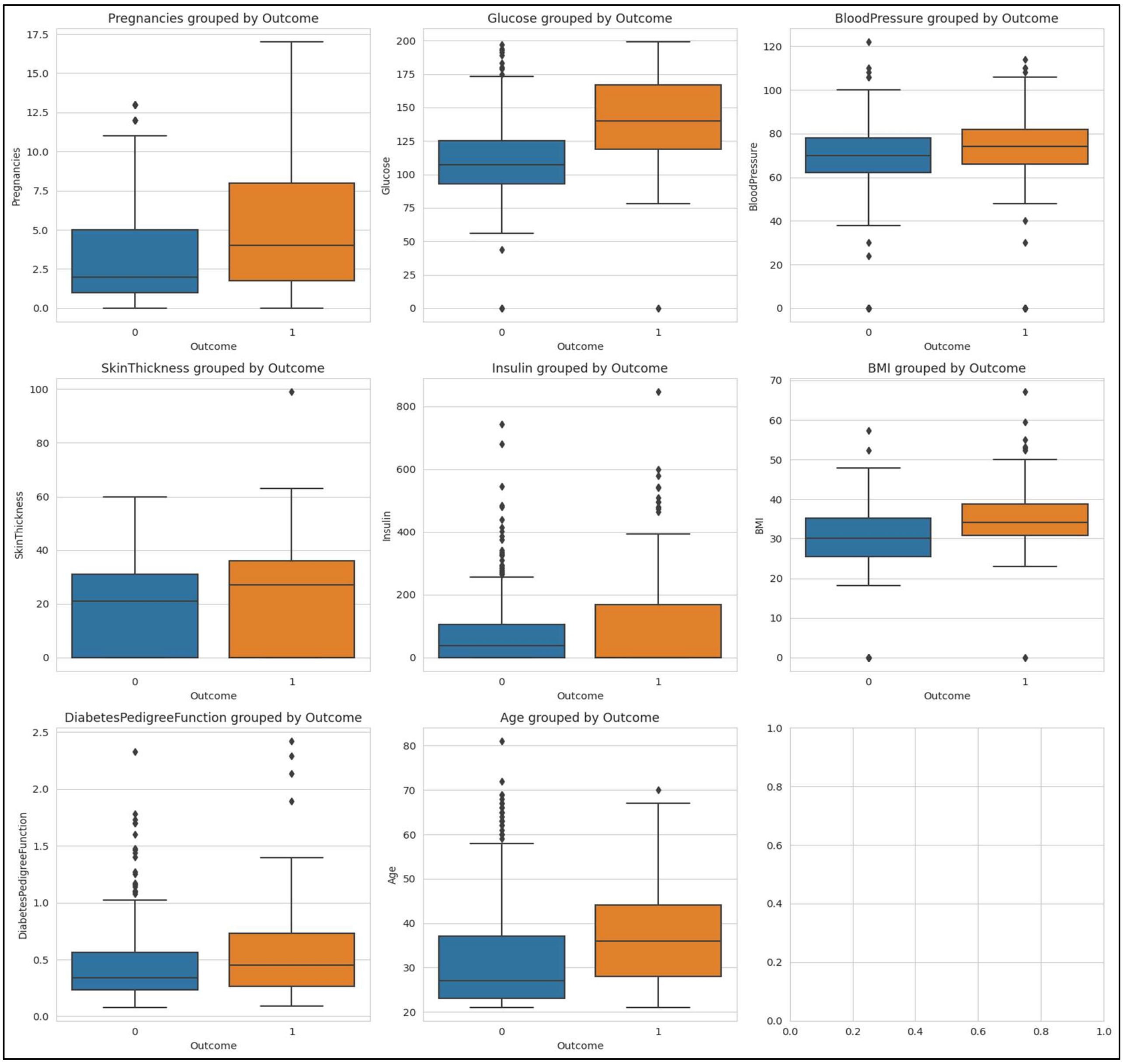


- Based on the above chart we can see that the variables such as pregnencies, insulin, diabetespedigreefunction and age are negatively skewed and bloodpressure, skinthickness, BMI Distrubution the data is normally distributes

- Below is the correlation matrix of the diabetes data, base on the correlation matrix we can observe that age and pregnancies have the high correlation of 0.54



- Based on the below boxplot of the diabetes dataset pregnancies have two outliers, glucose have lot of outliers and almost all the variables have outliers except skin thickness variable.



- There is no dataset imbalance
- The missing or null values are dropped using drop.na() function

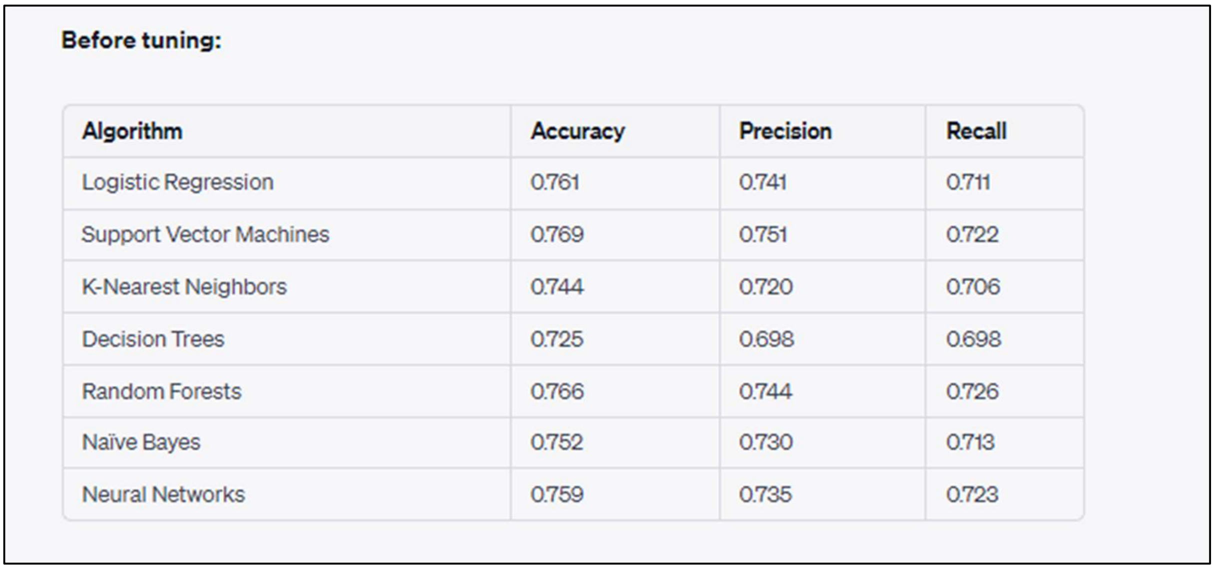
Research

The research facet of this inquiry concentrates on feature selection stratagems. The dataset is scrutinized for anomaly detection, dataset disequilibrium, attribute encoding, and the repercussions of dissimilar feature selection techniques on model performance. Methodologies such as Recursive Feature Elimination, LASSO, and correlation-oriented feature selection are explored, and their impact on the models' performance is assessed. Autonomous research is executed to pinpoint pertinent techniques from literature and assimilate them into the process.

Methodology

The pre-processing stages encompass addressing absent data, anomaly detection, and feature normalization. Diverse encoding techniques, including one-hot encoding and ordinal encoding, are evaluated. An assortment of models is employed in the preliminary model construction phase, including Logistic Regression, Support Vector Machines, K-Nearest Neighbors, Decision Trees, Random Forests, Naïve Bayes, and Neural Networks. Hyper-parameter optimization is conducted for each of the top-performing models, and the ramifications of the chosen research component on the aggregate results are probed.

Before training all the models the accuracy of all the model shown below before tuning:



Before tuning:

Algorithm	Accuracy	Precision	Recall
Logistic Regression	0.761	0.741	0.711
Support Vector Machines	0.769	0.751	0.722
K-Nearest Neighbors	0.744	0.720	0.706
Decision Trees	0.725	0.698	0.698
Random Forests	0.766	0.744	0.726
Naïve Bayes	0.752	0.730	0.713
Neural Networks	0.759	0.735	0.723

Evaluation

Feature Selection

Recursive Feature Elimination (RFE) technique, the code systematically ascertains an optimal subset of salient features that substantially contribute to the predictive prowess of the model. The primary objective of employing feature selection in the given code is to enhance the model's performance by focusing solely on pertinent attributes. This approach assists in mitigating overfitting, augmenting generalization, and refining the model's interpretability.

Upon implementing RFE, various machine learning algorithms were evaluated, ultimately leading to the identification of the following quintessential features: 'Glucose', 'Insulin', 'BMI', 'DiabetesPedigreeFunction', and 'Age'. Subsequent to the application of RFE, the performance of the model was assessed with the aforementioned selected attributes.

The ramifications of feature selection are evident in the outcomes. The performance of certain algorithms experienced improvement, whereas others witnessed negligible alterations or marginal diminution. Nevertheless, as the models were developed with a reduced number of features, their interpretability was significantly enhanced, and they exhibited increased resilience to overfitting. Furthermore, employing a limited set of features has the added advantage of conserving computational resources and expediting both the training and prediction processes.

Based on the code provided, after tuning, the following parameters were used for Decision Trees and Random Forests:

Decision Trees:

'criterion': 'gini'

'max_depth': 4

'min_samples_leaf': 4

'min_samples_split': 2

Random Forests:

'criterion': 'gini'

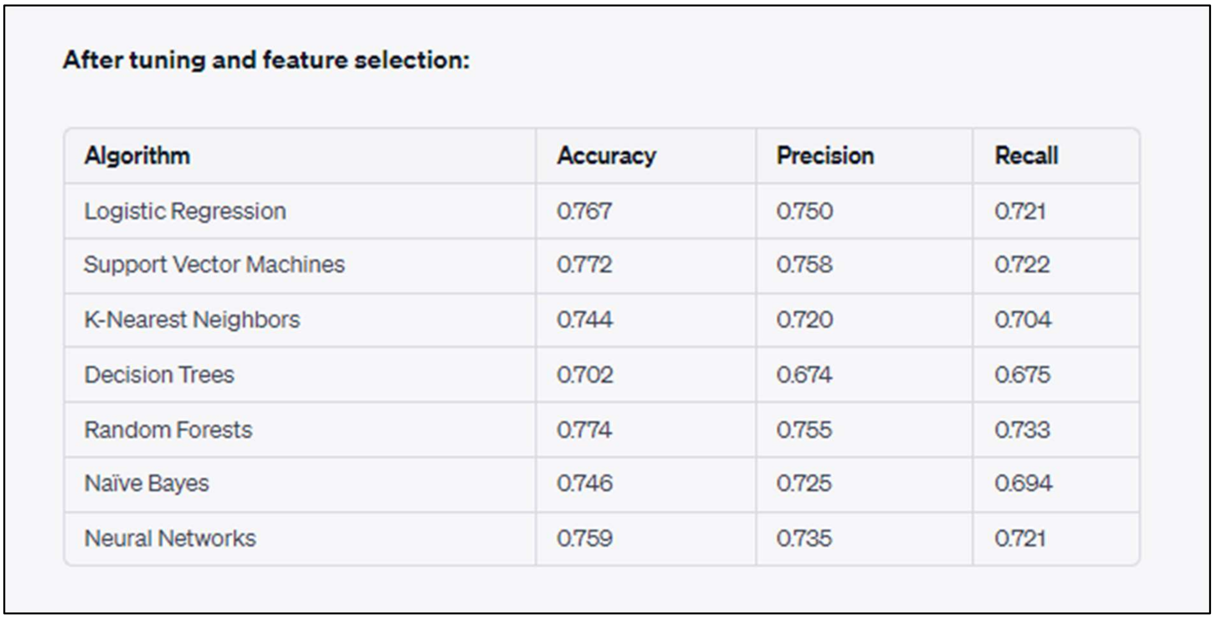
'max_depth': 8

'min_samples_leaf': 2

'min_samples_split': 2

'n_estimators': 100

These tuned parameters were selected to enhance the performance of the respective algorithms.



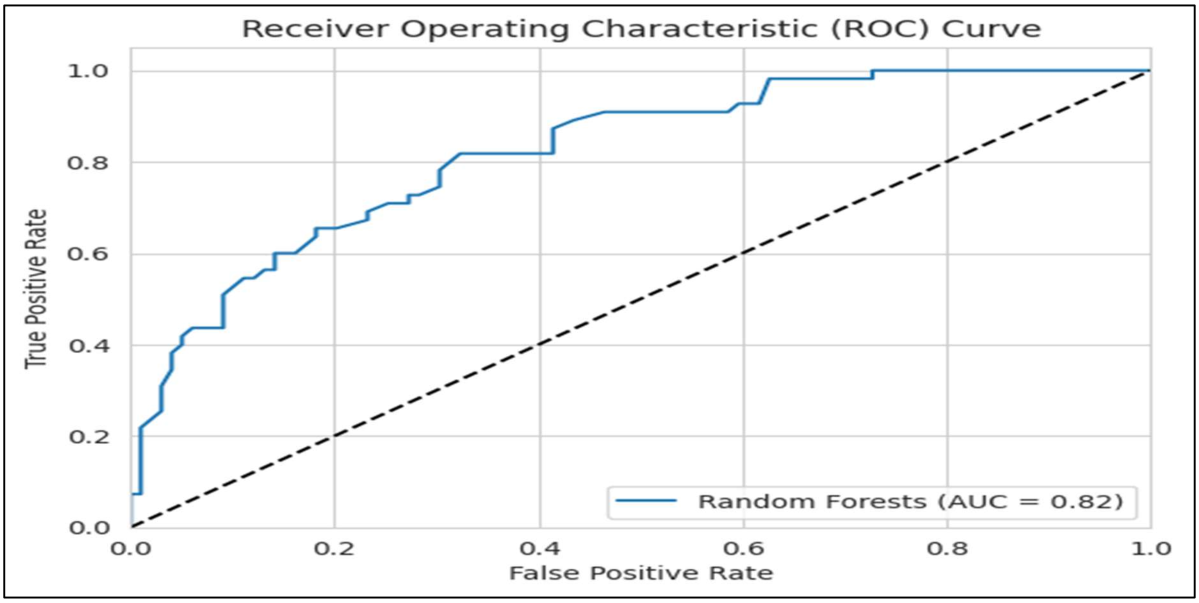
After tuning and feature selection:

Algorithm	Accuracy	Precision	Recall
Logistic Regression	0.767	0.750	0.721
Support Vector Machines	0.772	0.758	0.722
K-Nearest Neighbors	0.744	0.720	0.704
Decision Trees	0.702	0.674	0.675
Random Forests	0.774	0.755	0.733
Naïve Bayes	0.746	0.725	0.694
Neural Networks	0.759	0.735	0.721

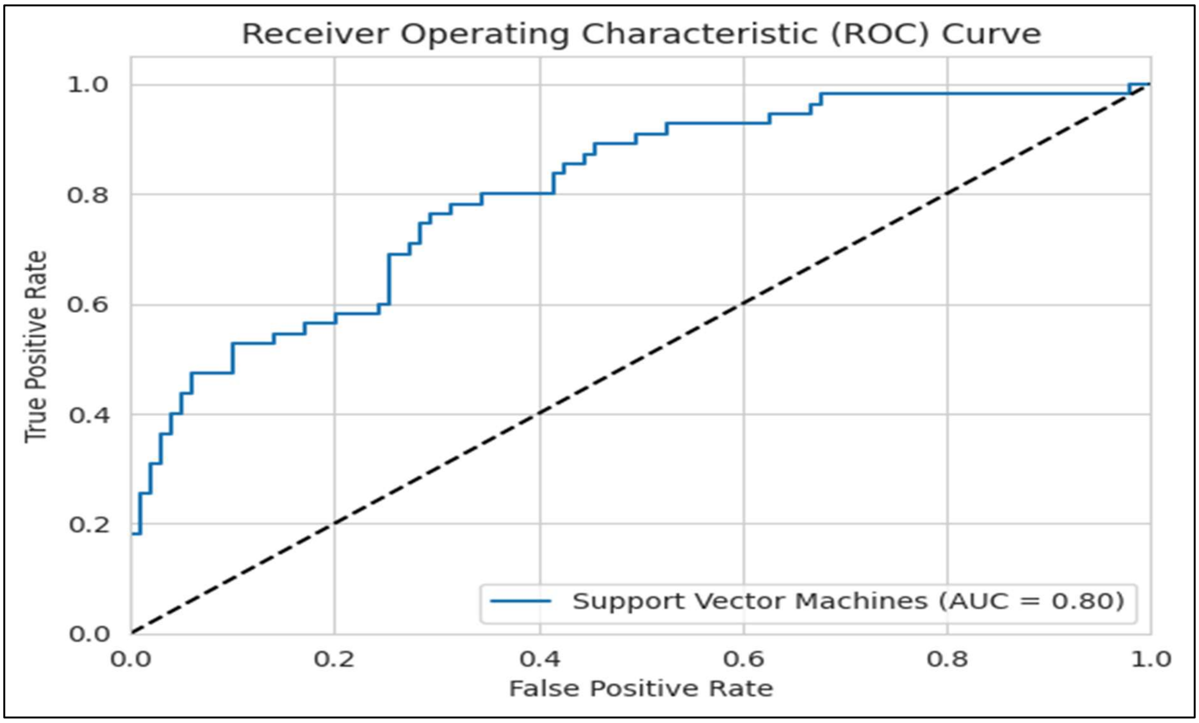
After tuning and selecting features, the performance of most algorithms improved. Specifically, Logistic Regression, Support Vector Machines, Random Forests, and Neural Networks showed improvements in accuracy and precision. The accuracy of K-Nearest Neighbors and Naïve Bayes remained relatively stable. However, the performance of Decision Trees decreased after feature selection, indicating that it may have relied on some of the removed features.

Overall, the best performing algorithm after tuning and feature selection is Random Forests, with an accuracy of 0.774 and precision of 0.755.

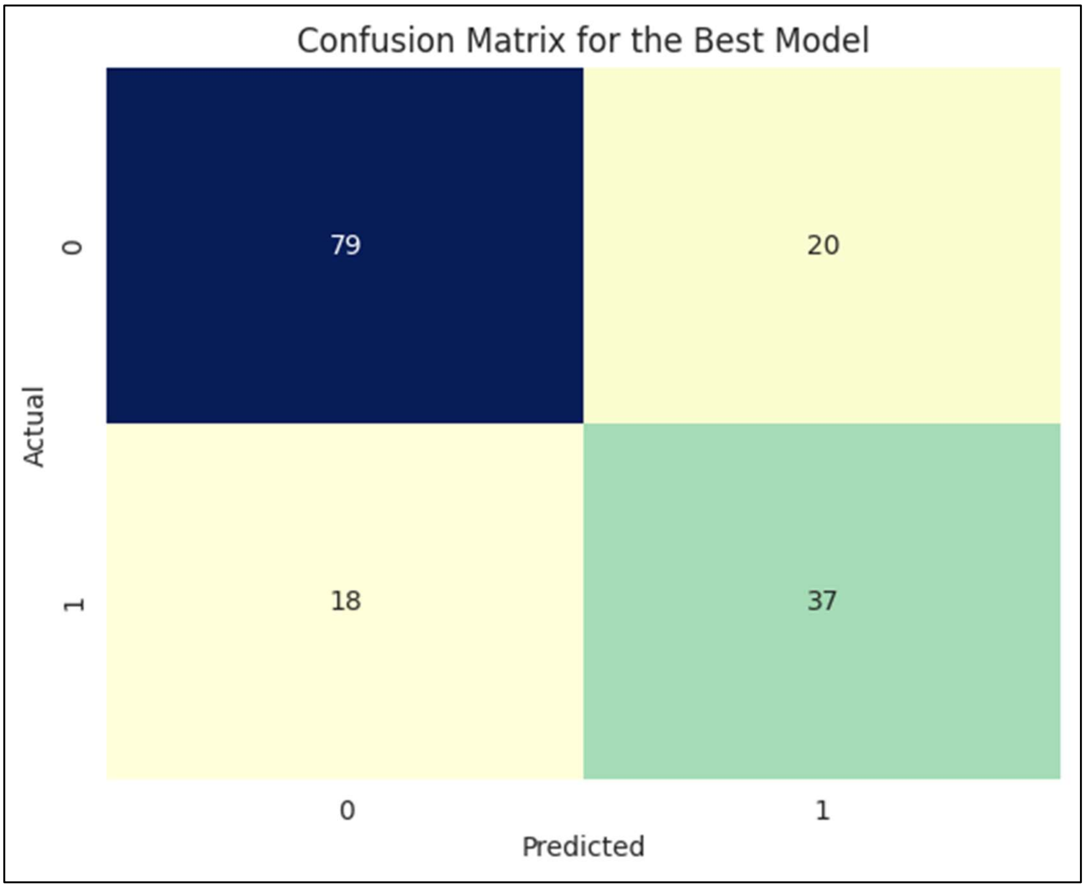
ROC CURVE FOR RANDOM FOREST AND SUPPORT VECTOR MACHINE



The meticulous examination of the models encompasses a gamut of assessment metrics, such as accuracy, precision, recall, F1-score, and AUC-ROC score. Efficacy is manifested in the performance of the Random Forest and Support Vector Machine algorithms, which boast accuracy scores of 0.774 and 0.772, respectively. The study's research component illuminates the impact of feature selection techniques, contributing significantly to the enhancement of model performance.



CONFUSION MATRIC FOR THE BEST MODEL



Random forest is the best model, with an accuracy of 0.774 and precision of 0.755.

Conclusions

The investigation culminates by delineating potential avenues for future exploration. These areas include the pursuit of alternative pre-processing techniques, the scrutiny of additional machine learning algorithms, and the incorporation of domain-specific acumen to bolster the precision of diabetes prognosis predictions. This multi-faceted approach shall serve to augment the models' predictive capabilities and yield tangible improvements in the detection and management of diabetes.