“DIABETES PREDICTION”
AN ANALYSIS OF CLASSIFICATION MODELS

Presented by
Bahador Mirzazadeh,
Mohammad Matin Parvanian
Sadaf Jamali

University of Padova | May 2023

Department of Mathematics

INTRODUCTION

This paper presents a systematic analysis of classification models for diabetes prediction, such as logistic regression, LDA, QDA, Naive Bayes, and KNN. We assess their performance on a given dataset, employing performance metrics and feature selection techniques. Our study includes the description of model coefficients and identifies the most informative features for diabetes prediction. Early detection and prediction of diabetes can greatly improve disease management and enhance the quality of life for affected individuals.

BACKGROUND

The prevalence of diabetes worldwide has led to increased research and study on predicting the disease. Early detection and prediction are crucial for managing diabetes and reducing complications. Statistical learning techniques and large datasets have facilitated the development of accurate predictive models. These models utilize clinical and demographic features, such as age, BMI, blood pressure, glucose levels, and genetic markers, to identify individuals at risk or already affected by diabetes.

DATA COLLECTION

The project utilized the Diabetes dataset, which was originally sourced from the National Institute of Diabetes and Digestive and Kidney Diseases. The dataset aims to predict whether a patient has diabetes or not, based on various diagnostic measurements. It consists of information from 768 female patients of Pima Indian heritage, aged 21 years and above. Each data point in the dataset represents a patient and includes multiple features or attributes. Additionally, there is a binary outcome variable indicating the presence (1) or absence (0) of diabetes for each patient.

FEATURES

- Pregnancies: Number of times pregnant.
- Glucose: Plasma glucose concentration.
- BloodPressure: Diastolic blood pressure (mm Hg).
- SkinThickness: Triceps skinfold thickness (mm).
- Insulin: 2-Hour serum insulin (mu U/ml).
- BMI: Body mass index (weight in kg/(height in m)^2).
- DiabetesPedigreeFunction: Diabetes pedigree function, which represents the genetic influence of diabetes based on family history.
- Age: Age of instances in years.

PREPROCESSING

There are some variables with zero values like glucose and blood pressure. These variables cannot take zero values since it is not defined for their range of them. So in this dataset, missing values are considered as zero. The number of zeros for each variable is shown below:

[1] "Column Glucose has 5 zeros."
[1] "Column BloodPressure has 35 zeros."
[1] "Column SkinThickness has 227 zeros."
[1] "Column Insulin has 374 zeros."
[1] "Column BMI has 11 zeros."

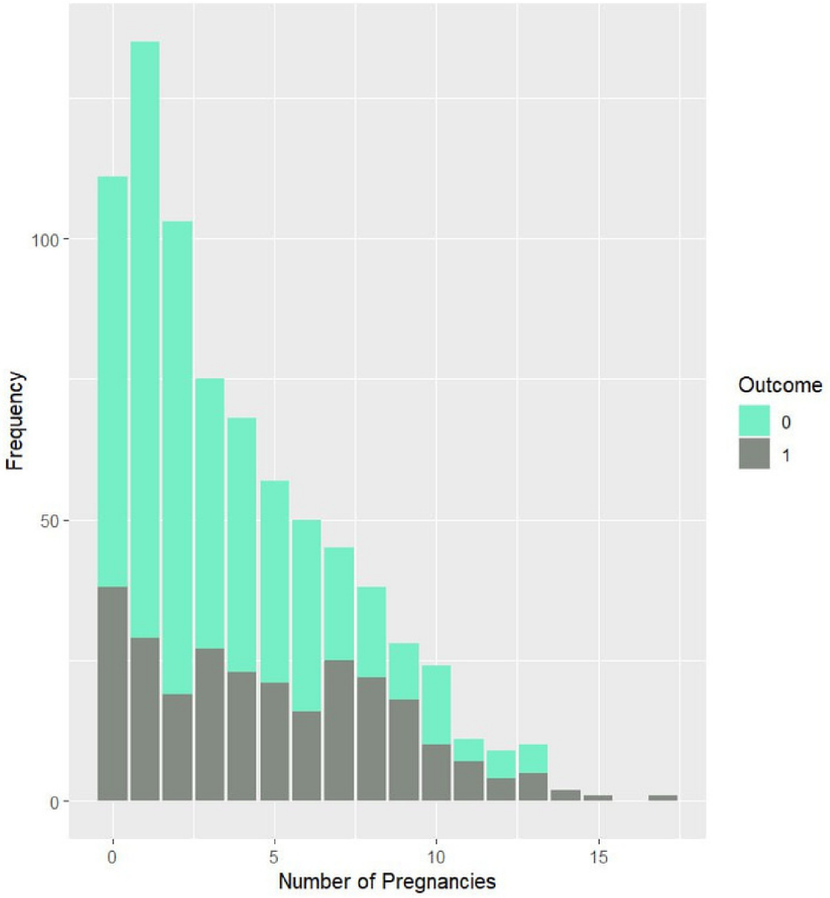
PREPROCESSING

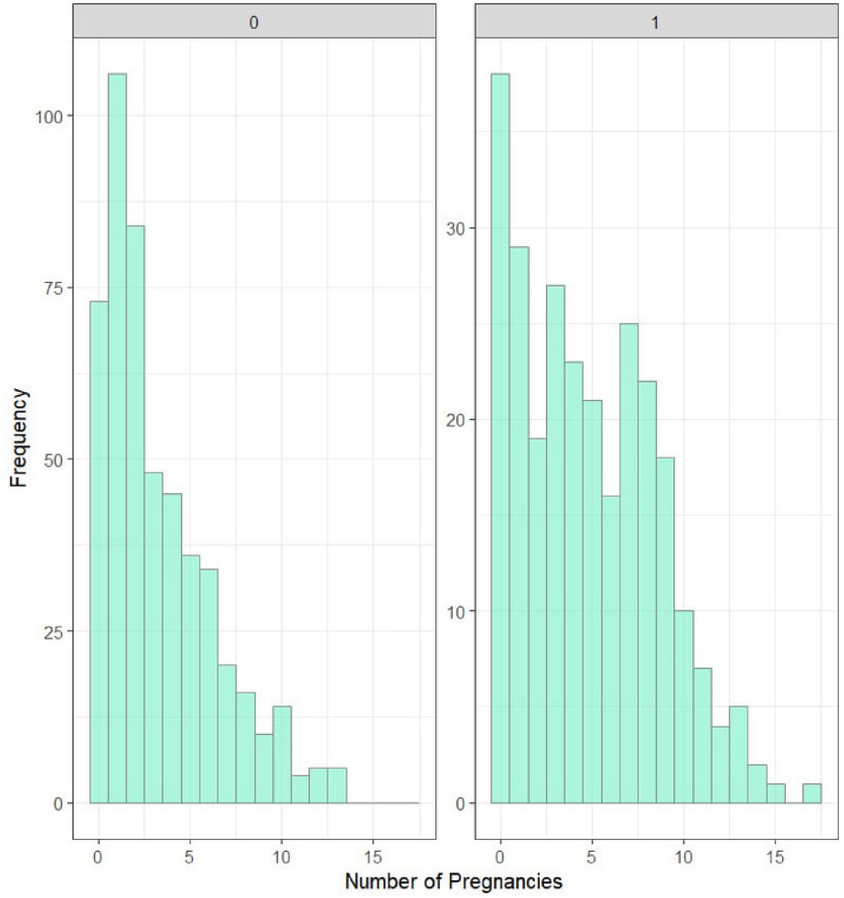
The project considers three methods for handling missing values in the dataset: removal, mean replacement, and median replacement. Due to a large number of missing values, removal is not efficient. Mean replacement is sensitive to outliers, as a single extreme value can skew the mean significantly. In contrast, using the median for replacement is more robust against outliers. Given the presence of outliers in the dataset, median replacement is considered a better choice for handling missing values.

EDA

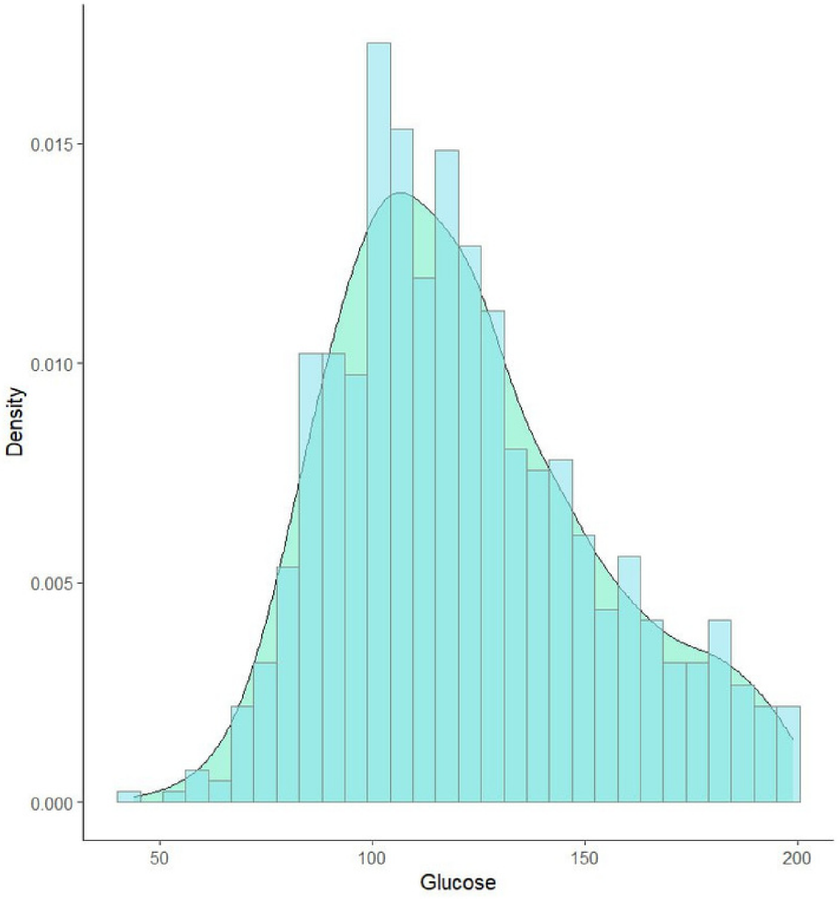
Exploratory Data Analysis (EDA) involves analyzing and visualizing data to gain insights and understand its characteristics. It helps understand the dataset's underlying patterns, relationships, and distributions. Through EDA, we can identify missing values and outliers and examine the data's overall structure. In the following slides, we demonstrate some of them.

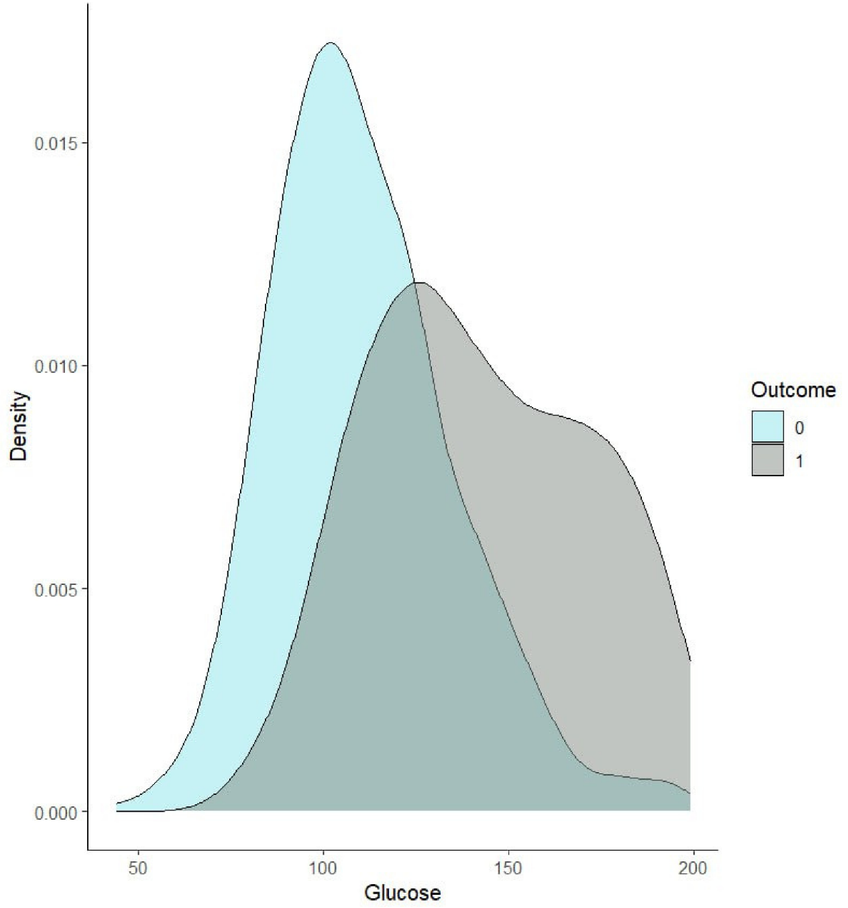
EDA


Frequency
100
50
0
0
5
10
15
Number of Pregnancies
Outcome
0
1

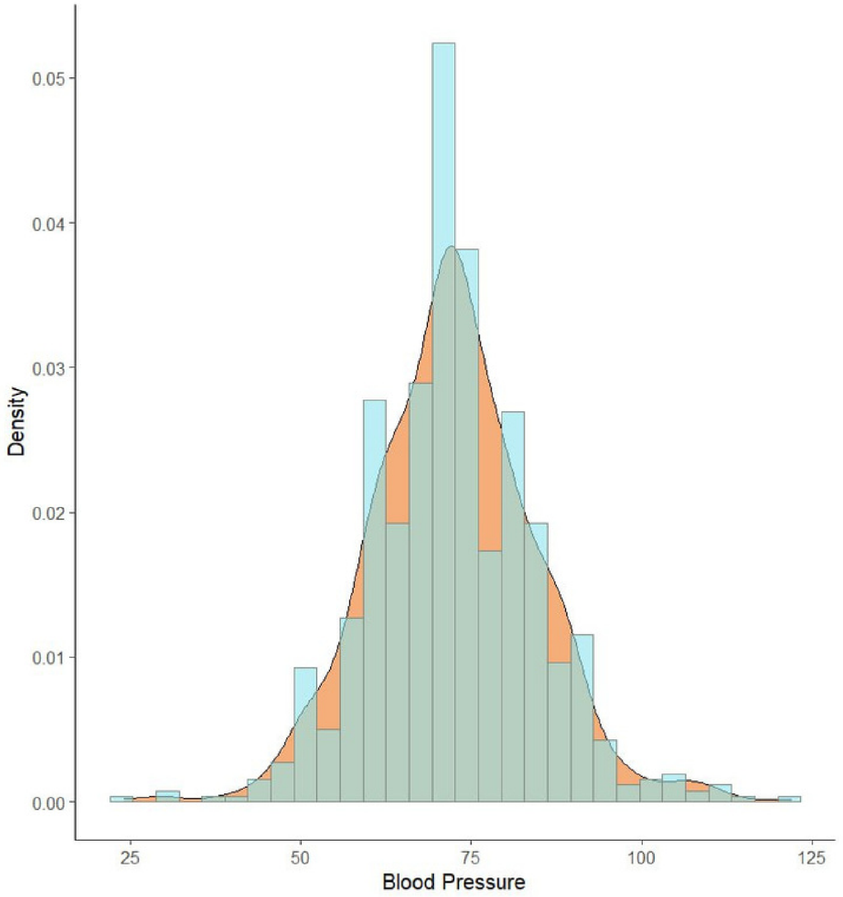

0
1
Frequency
100
75
50
25
0
30
20
10
0
0
5
10
15
Number of Pregnancies

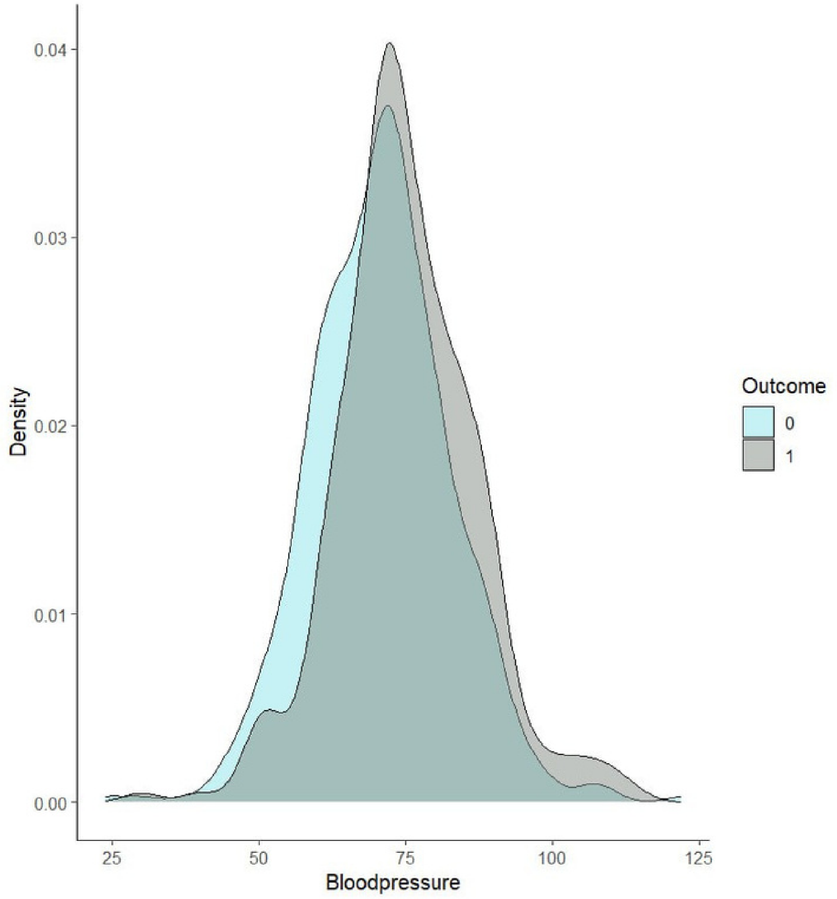
EDA


0.015
0.010
0.005
0.000
Density
50
100
150
200
Glucose

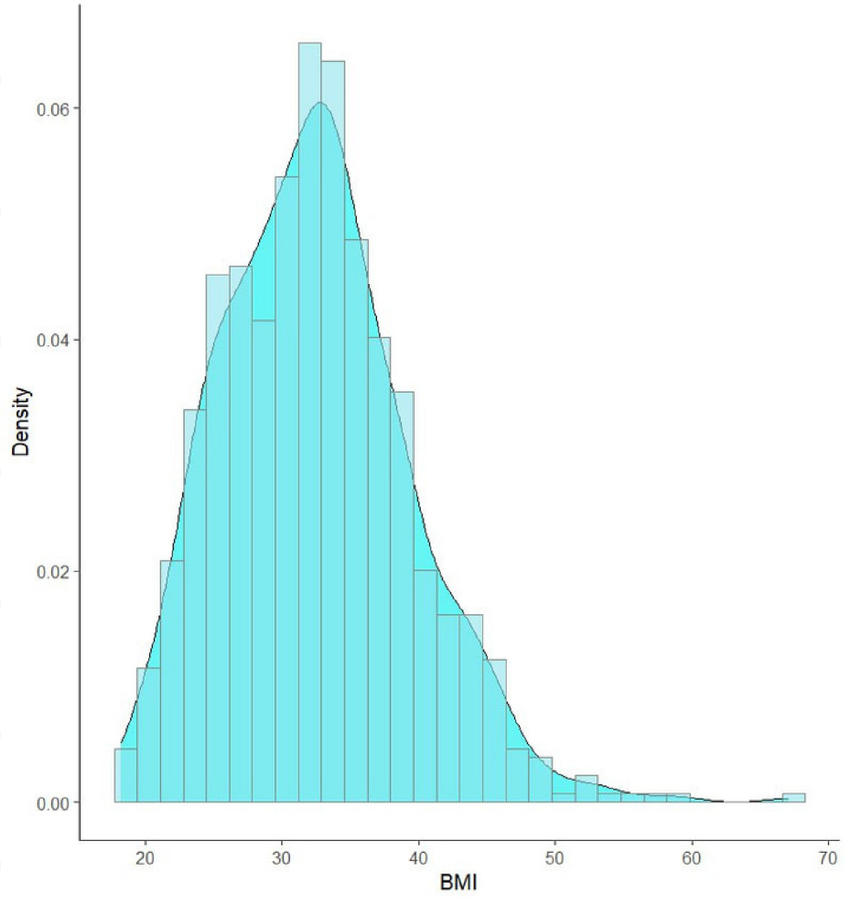

0.015
0.010
0.005
0.000
Density
50
100
150
200
Glucose
Outcome
0
1

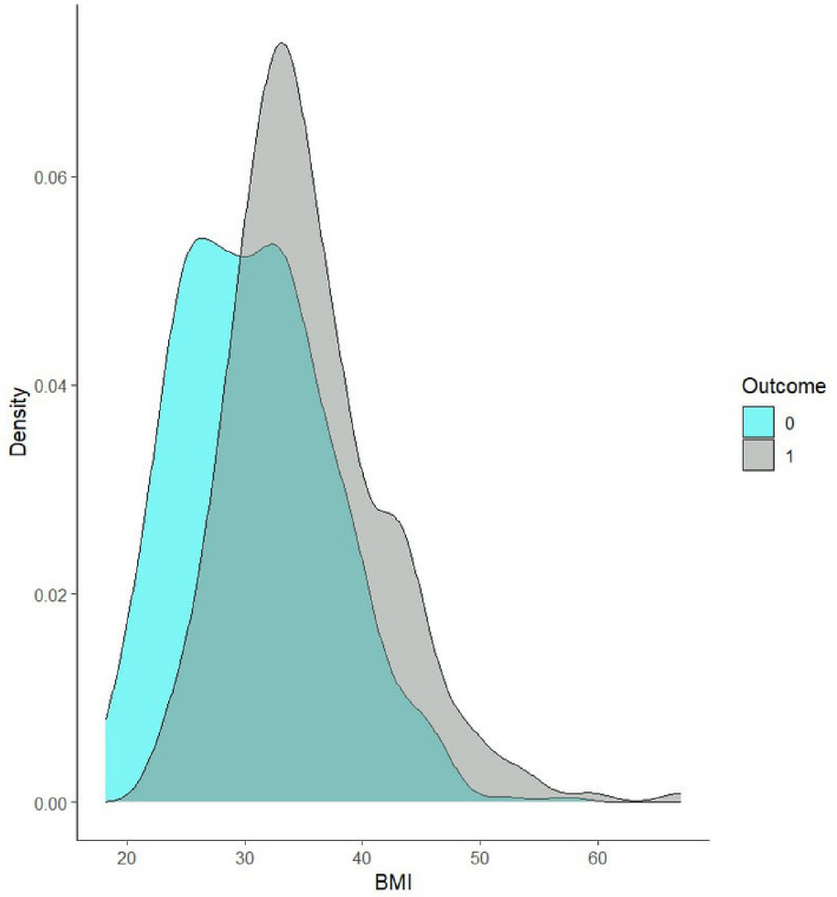
EDA


0.05
0.04
0.03
0.02
0.01
0.00
Density
25
50
75
100
125
Blood Pressure

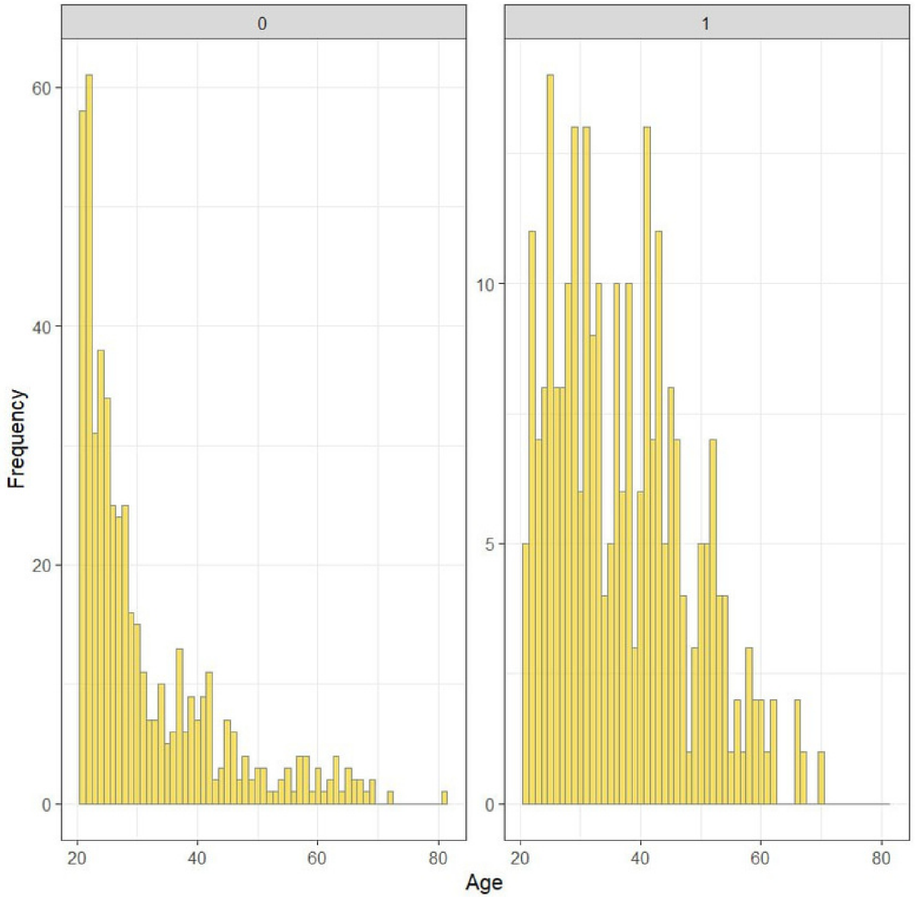

0.04
0.03
0.02
0.01
0.00
Density
25
50
75
100
125
Bloodpressure
Outcome
0
1

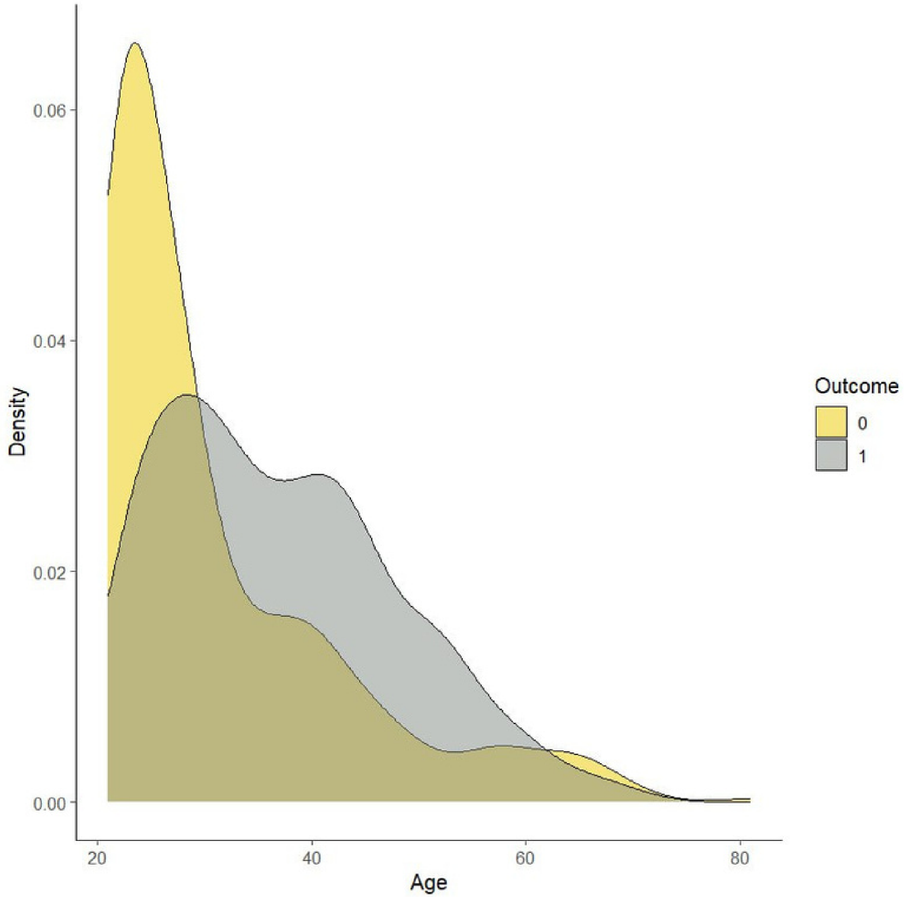
EDA


0.06
0.04
0.02
0.00
Density
20
30
40
50
60
70
BMI

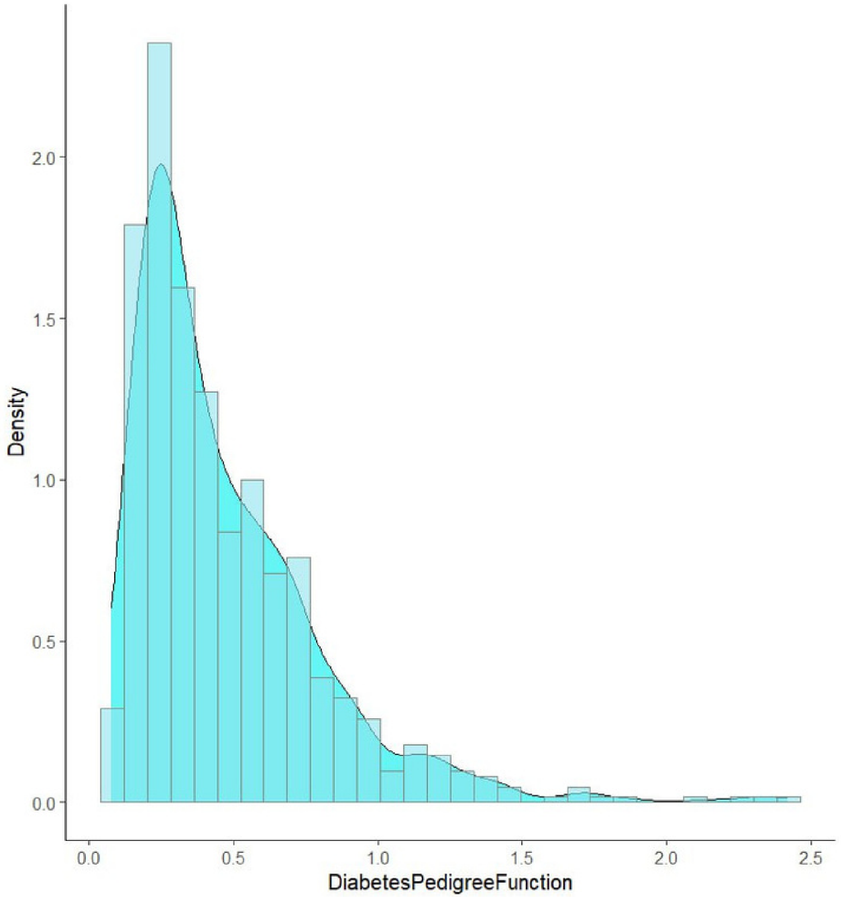

0.06
0.04
0.02
0.00
Density
20
30
40
50
60
BMI
Outcome
0
1

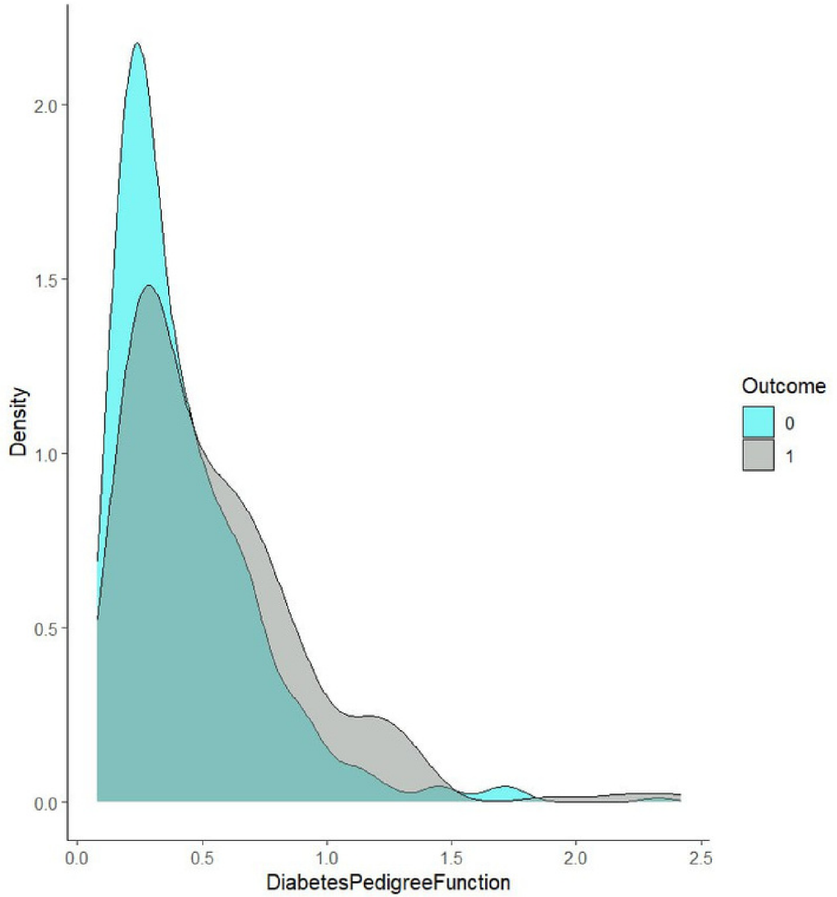
EDA


0
1
Frequency
Age
60
40
20
0
10
5
20
40
60
80

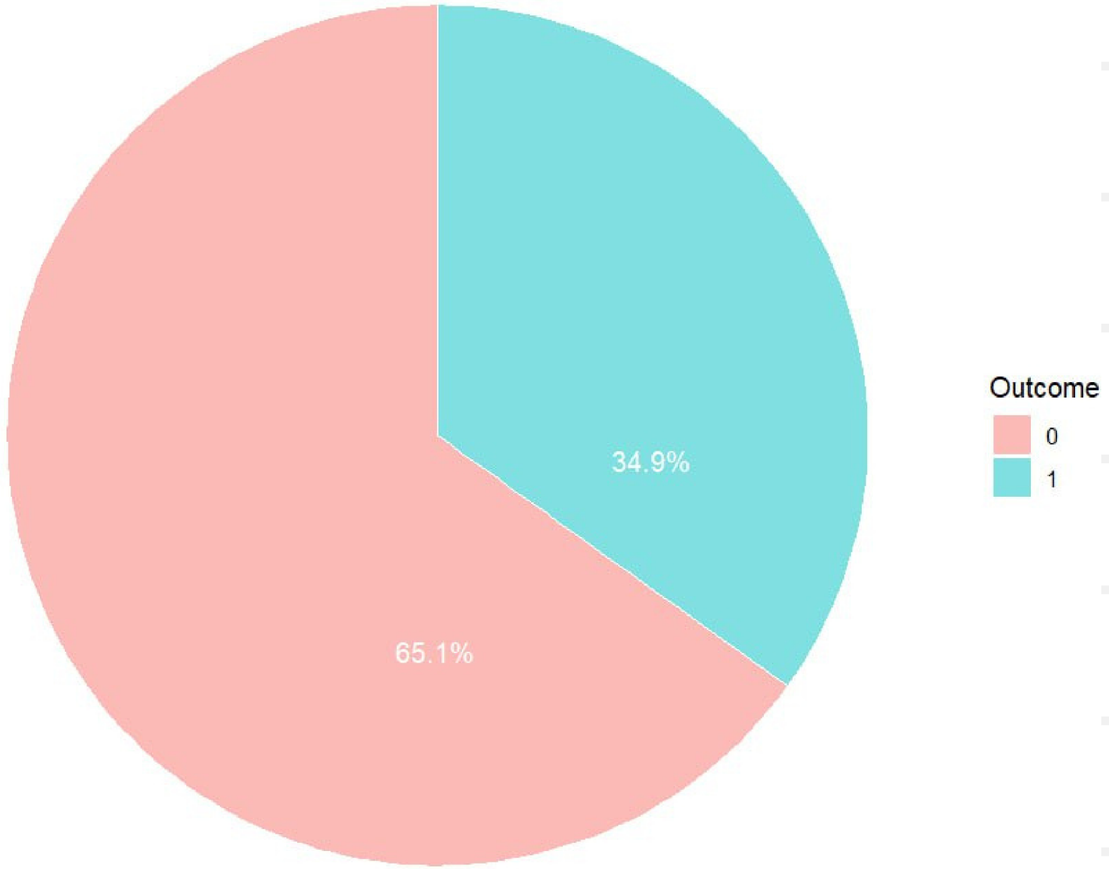

Density
Age
Outcome
0
1
0.06
0.04
0.02
0.00
20
40
60
80

EDA

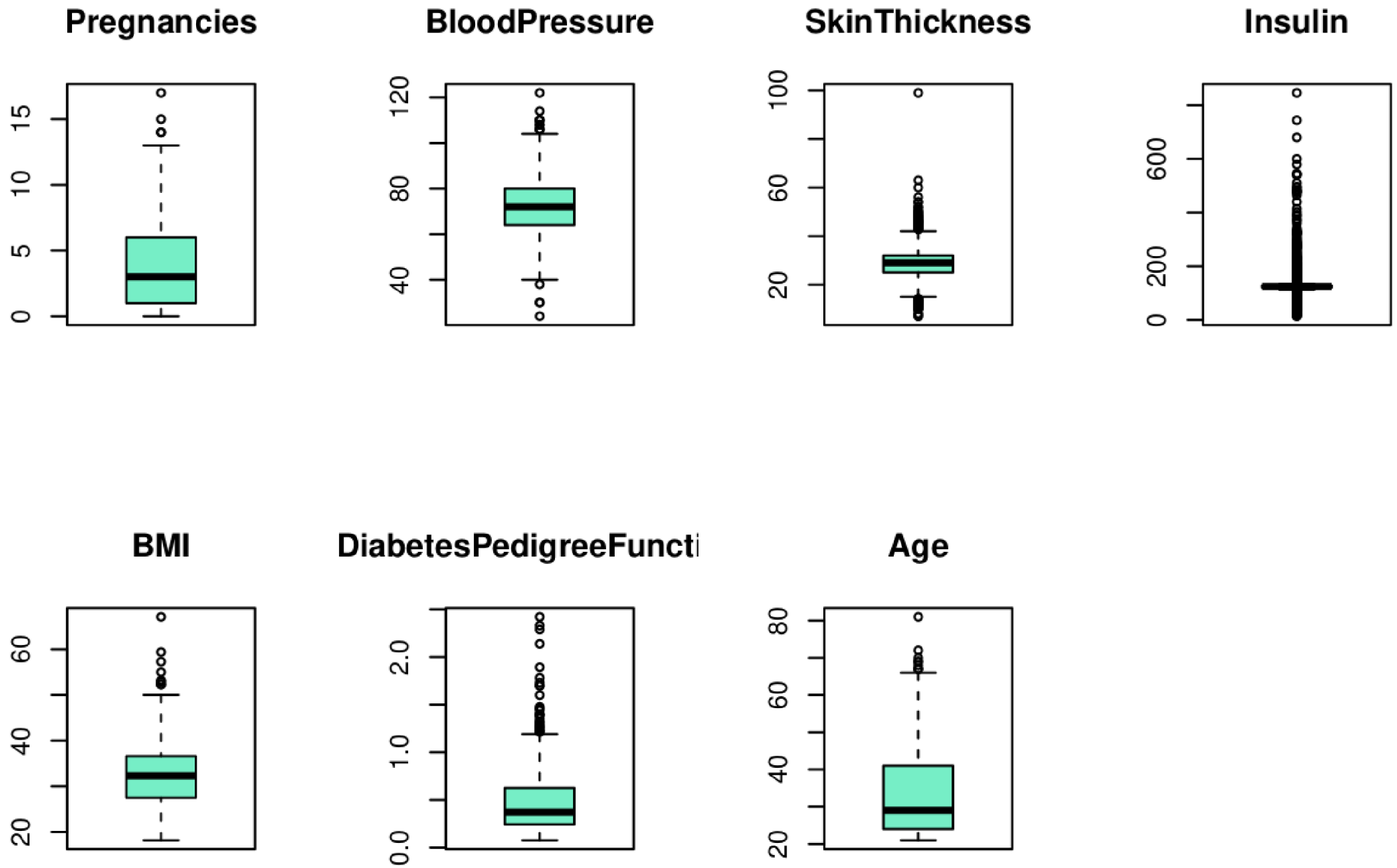

Density
2.0
1.5
1.0
0.5
0.0
0.0
0.5
1.0
1.5
2.0
2.5
DiabetesPedigreeFunction


Density
2.0
1.5
1.0
0.5
0.0
0.0
0.5
1.0
1.5
2.0
2.5
DiabetesPedigreeFunction
Outcome
0
1

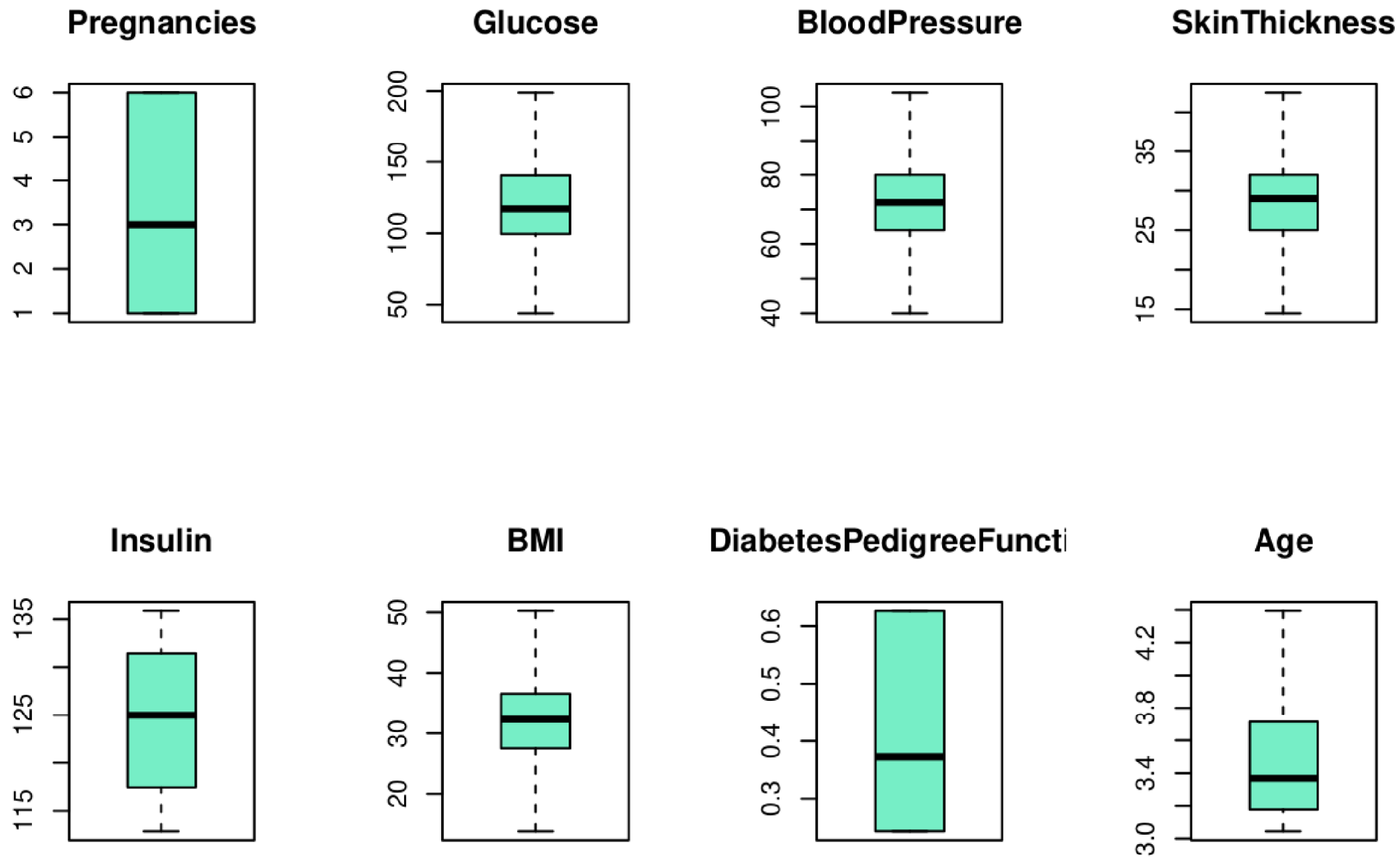
EDA



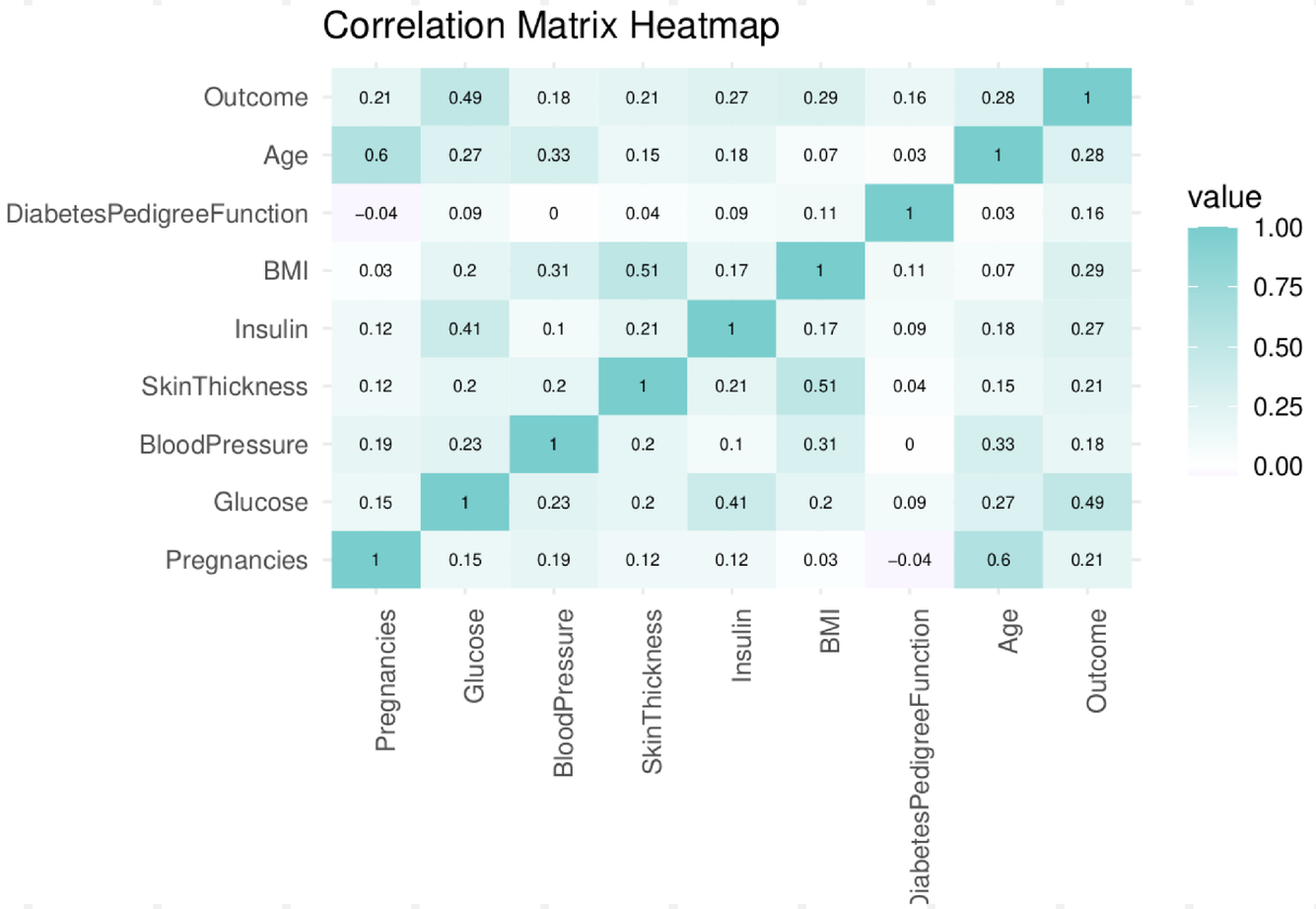
EDA


Pregnancies
0
5
10
15
BloodPressure
40
80
120
SkinThickness
20
60
100
Insulin
0
200
600
BMI
20
40
60
DiabetesPedigreeFuncti
0.0
1.0
2.0
Age
20
40
60
80

EDA


Pregnancies
1 2 3 4 5 6
Glucose
50 100 150 200
BloodPressure
40 60 80 100
SkinThickness
15 25 35
Insulin
115 125 135
BMI
20 30 40 50
DiabetesPedigreeFuncti
0.3 0.4 0.5 0.6
Age
3.0 3.4 3.8 4.2

EDA


Correlation Matrix Heatmap
Outcome 0.21 0.49 0.18 0.21 0.27 0.29 0.16 0.28 1
Age 0.6 0.27 0.33 0.15 0.18 0.07 0.03 1 0.28
DiabetesPedigreeFunction −0.04 0.09 0 0.04 0.09 0.11 1 0.03 0.16
BMI 0.03 0.2 0.31 0.51 0.17 1 0.11 0.07 0.29
Insulin 0.12 0.41 0.1 0.21 1 0.17 0.09 0.18 0.27
SkinThickness 0.12 0.2 0.2 1 0.21 0.51 0.04 0.15 0.21
BloodPressure 0.19 0.23 1 0.2 0.1 0.31 0 0.33 0.18
Glucose 0.15 1 0.23 0.2 0.41 0.2 0.09 0.27 0.49
Pregnancies 1 0.15 0.19 0.12 0.12 0.03 −0.04 0.6 0.21
Pregnancies
Glucose
BloodPressure
SkinThickness
Insulin
BMI
DiabetesPedigreeFunction
Age
Outcome
value
1.00
0.75
0.50
0.25
0.00

MODELS

In this section, we will fit our models to the Diabetes dataset. As mentioned earlier, we will consider five different algorithms: Logistic Regression, K-Nearest Neighbors (KNN), Naive Bayes, Linear Discriminant Analysis (LDA), and Quadratic Discriminant Analysis (QDA). For each model, we will explore different scenarios by applying feature selection methods and cross-validation techniques. Our goal is to identify the best model that is most suitable for our Diabetes dataset.

METRICS

The recall formula indicates that a higher recall value implies a lower rate of false negatives, which is particularly important when identifying positive instances. Therefore, this metric holds significant importance for us. Similarly, a higher precision value signifies a lower rate of false positives. Another vital metric is the F1 score, which allows us to consider both precision and recall simultaneously. While we strive for a higher recall value, we also aim for higher precision. Hence, the F1 score is of great significance to us. Lastly, we have the accuracy, which represents the proportion of correctly classified instances (both positive and negative) out of the total number of instances.

LOGISTIC REGRESSION

Logistic regression is a statistical algorithm used for binary classification tasks. It models the relationship between independent variables and a binary outcome using a logistic function. It estimates the probabilities of the outcome class and predicts the most likely class based on a predefined threshold. Logistic regression is widely used in various fields, such as healthcare, finance, and marketing, to predict and analyze binary outcomes based on input features. We combined logistic regression with cross-validation and also backward feature selection.

LOGISTIC REGRESSION COEFFICIENTS

```
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -11.788484   2.148732  -5.486 4.11e-08 ***
## Pregnancies                0.190786   0.062758   3.040  0.00237 **
## Glucose                    0.036873   0.004310   8.554  < 2e-16 ***
## BloodPressure             -0.003562   0.005635  -0.632  0.52732
## SkinThickness              0.005284   0.012675   0.417  0.67675
## Insulin                    0.010045   0.014509   0.692  0.48871
## BMI                        0.039715   0.010199   3.894 9.87e-05 ***
## DiabetesPedigreeFunction   2.015679   0.658233   3.062  0.00220 **
## Age                        0.716324   0.415684   1.723  0.08485 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 788.34  on 614  degrees of freedom
## Residual deviance: 564.40  on 606  degrees of freedom
## AIC: 582.4
```

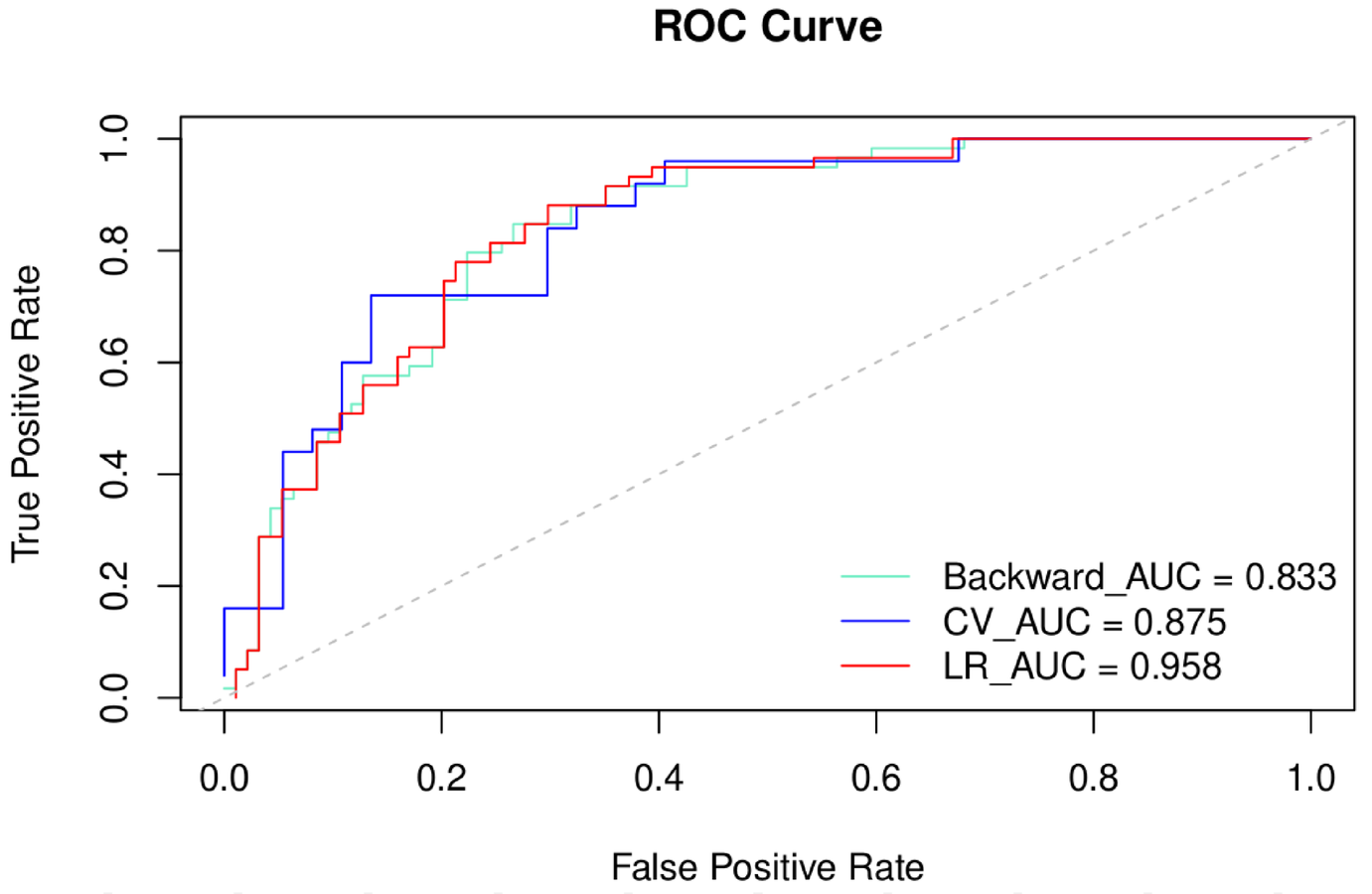
LOGISTIC REGRESSION WITH CV COEFFICIENTS

```
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -11.151862   2.166953  -5.146 2.66e-07 ***
## Pregnancies                0.148833   0.063131   2.358  0.01840 *
## Glucose                    0.033943   0.004425   7.670 1.72e-14 ***
## BloodPressure              0.004011   0.005866   0.684  0.49409
## SkinThickness              0.003505   0.012758   0.275  0.78354
## Insulin                    0.013032   0.014595   0.893  0.37191
## BMI                        0.042593   0.010354   4.114 3.90e-05 ***
## DiabetesPedigreeFunction   2.113587   0.685459   3.083  0.00205 **
## Age                        0.420931   0.425804   0.989  0.32288
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 716.98  on 551  degrees of freedom
## Residual deviance: 529.20  on 543  degrees of freedom
## AIC: 547.2
```

LOGISTIC REGRESSION WITH BFS COEFFICIENTS

```
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -10.562913   1.410756  -7.487 7.02e-14 ***
## Pregnancies               0.192942   0.062734   3.076   0.0021 **
## Glucose                   0.037630   0.004021   9.358  < 2e-16 ***
## BMI                       0.040994   0.008650   4.739 2.15e-06 ***
## DiabetesPedigreeFunction  2.039643   0.656826   3.105   0.0019 **
## Age                       0.650746   0.401606   1.620   0.1052
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 788.34  on 614  degrees of freedom
## Residual deviance: 565.50  on 609  degrees of freedom
## AIC: 577.5
```

LOGISTIC REGRESSION MODELS
ROC CURVES



RESULTS COMPARISON

Metric	Logistic_Regression	Cross_Validation	Backward_Feature_Selection
Recall	0.75	0.68	0.73
Precision	0.50	0.58	0.50
F1-Score	0.60	0.62	0.60
Accuracy	0.74	0.76	0.73
AIC	582.40	547.20	577.50

KNN

KNN, or k-nearest neighbors, is a simple machine learning algorithm used for classification and regression tasks. It determines the class or value of a data point by considering the k nearest neighbors in the training dataset. The algorithm assigns the majority class or calculates the average value of the k nearest neighbors to make predictions. KNN is a non-parametric algorithm that does not make any assumptions about the underlying data distribution and can handle both numerical and categorical features. Also, we considered KNN model with cross-validation.

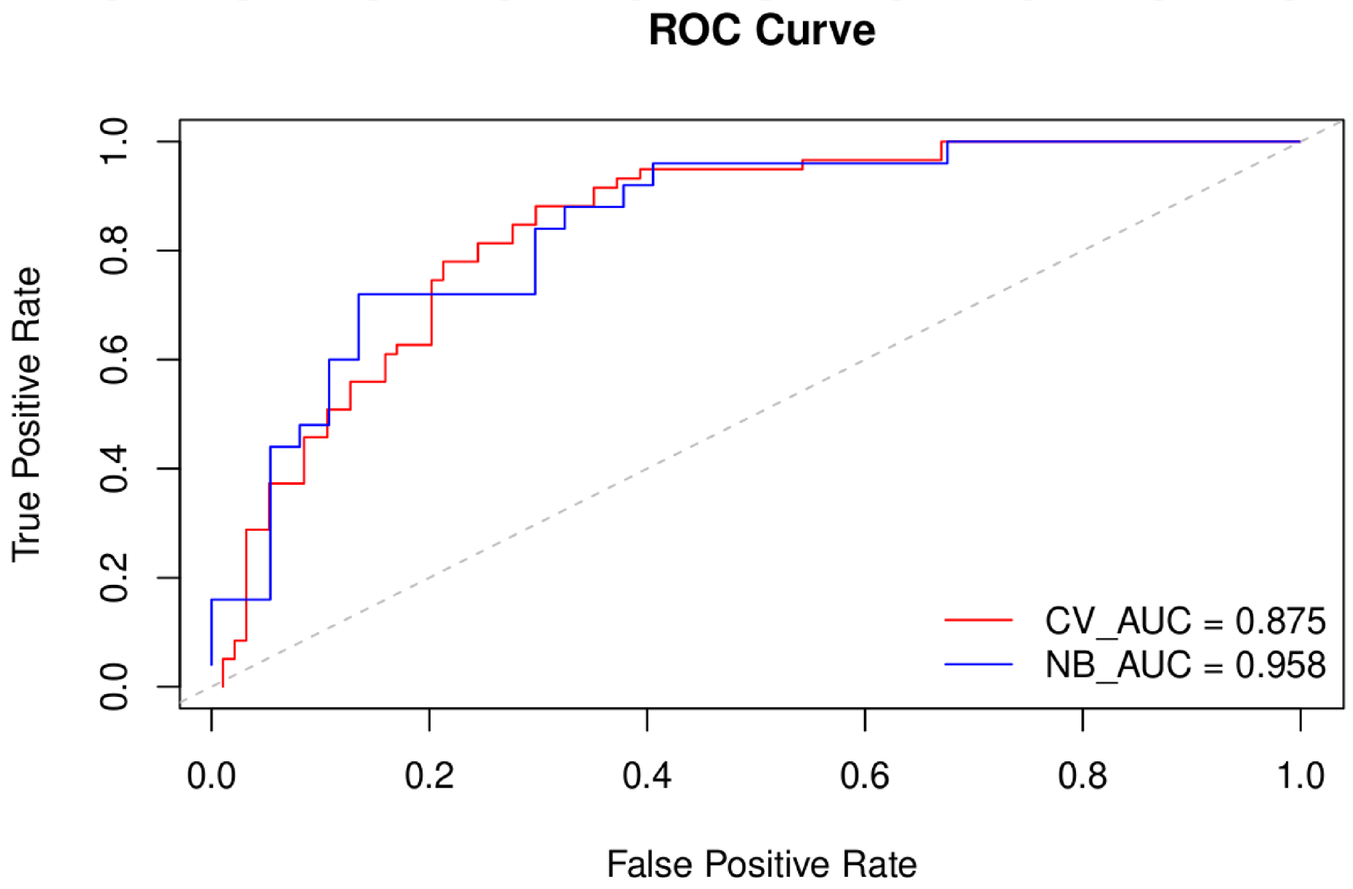
RESULTS COMPARISON

Metric	KNN	KNN_Cross_Validation
Recall	0.63	0.60
Precission	0.49	0.57
F1-Score	0.55	0.58
Accuracy	0.69	0.71

NAIVE BAYES

Naive Bayes is a probabilistic machine learning algorithm used for classification tasks. It assumes that features are conditionally independent given the class label, which is a naive assumption but simplifies the computation. It calculates the probability of each class for a given set of features using Bayes' theorem and selects the class with the highest probability as the predicted class. Naive Bayes is computationally efficient, works well with high-dimensional data, and is commonly used for text classification and spam filtering tasks. Also, we considered the Naive Bayes model with cross-validation.

NAIVE BAYES MODELS ROC CURVES



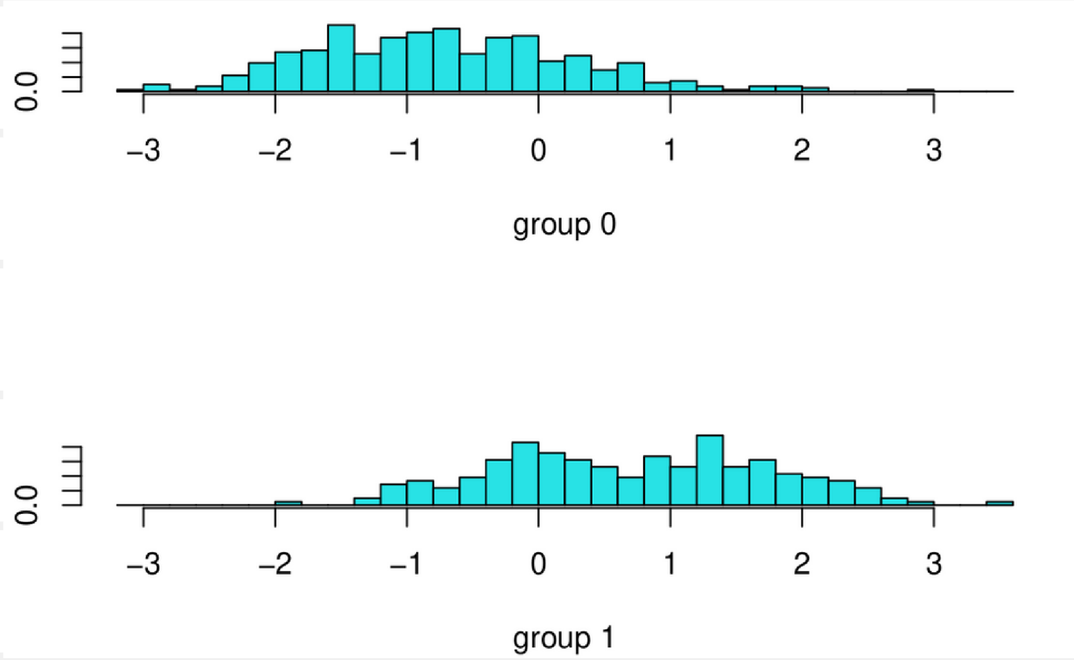
RESULTS COMPARISON

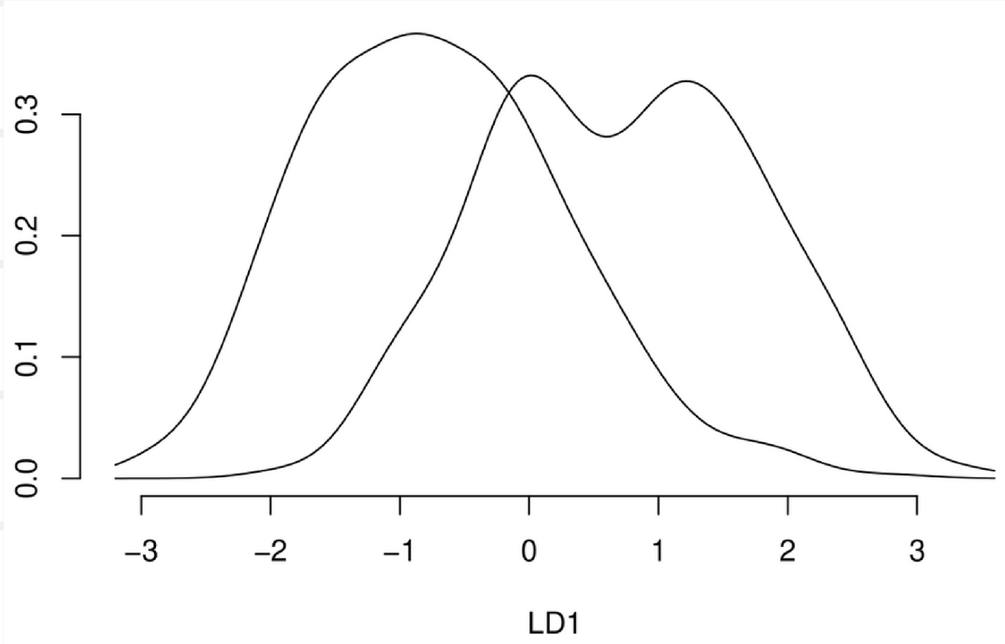
Metric	Naive_Bayes	NB_Cross_Validation
Recall	0.64	0.64
Precission	0.59	0.65
F1-Score	0.61	0.64
Accuracy	0.71	0.75

LDA

Linear Discriminant Analysis (LDA) is a dimensionality reduction and classification technique used for pattern recognition and classification tasks. It aims to find a linear combination of features that maximizes the separation between different classes in the data. LDA assumes that the data follows a Gaussian distribution and calculates class-specific means and covariance matrices. It then projects the data onto a lower-dimensional space, where the classes are well-separated. LDA is particularly effective when the classes have distinct distributions and can be used for both binary and multi-class classification problems. Also, we considered the LDA model with cross-validation.

LDA PLOTS





RESULTS COMPARISON

Metric	LDA	LDA_Cross_Validation
Recall	0.75	0.68
Precission	0.52	0.56
F1-Score	0.62	0.61
Accuracy	0.75	0.76

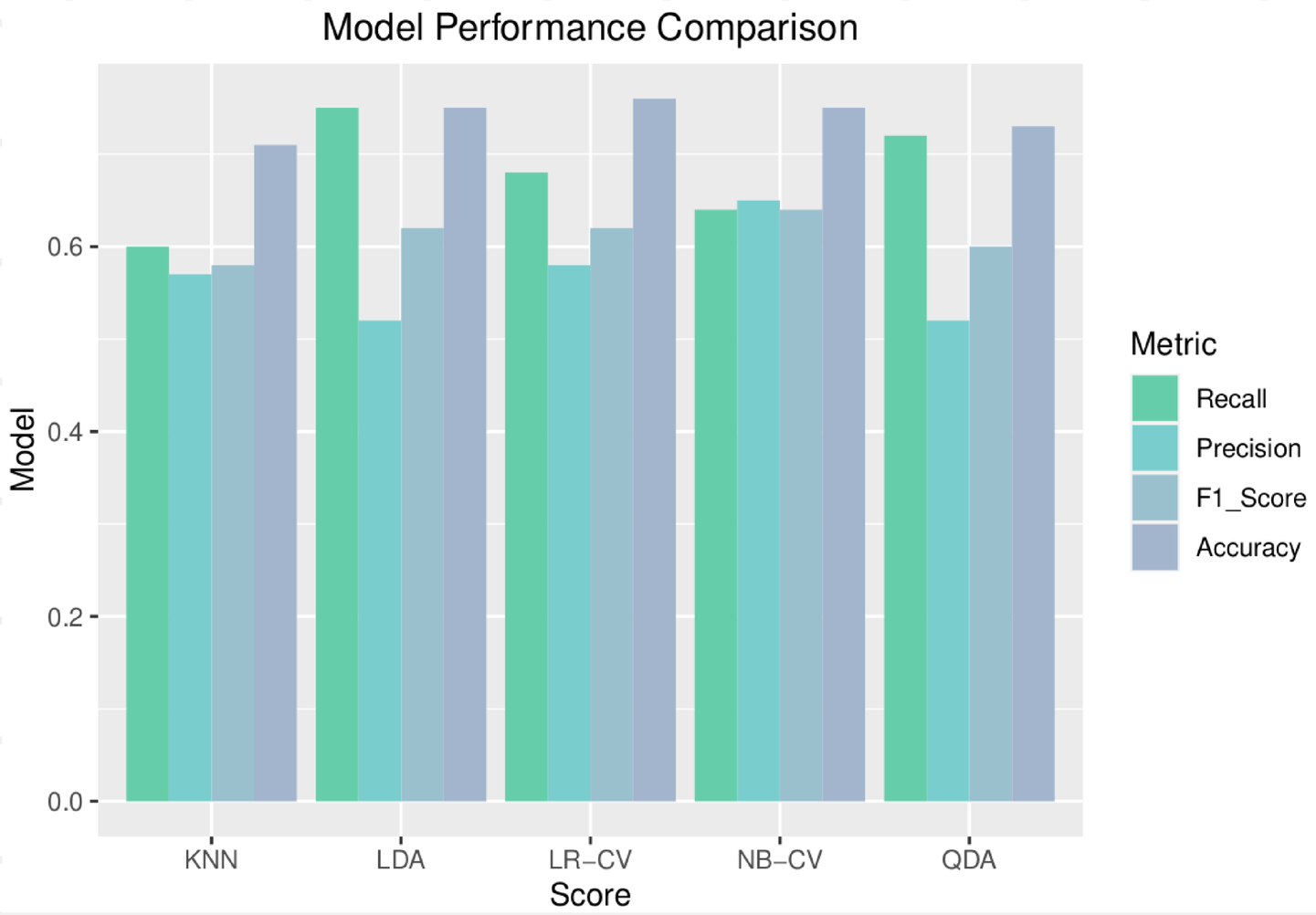
QDA

QDA, or Quadratic Discriminant Analysis, is a classification algorithm that assumes each class has its own covariance matrix. It calculates the probability of each class for a given set of features using the quadratic discriminant function. QDA can handle non-linear decision boundaries and is suitable when the covariance matrices of different classes are different. It is a flexible algorithm but requires a larger number of training samples compared to other classifiers. Also, we considered the LDA model with cross-validation.

RESULTS COMPARISON

Metric	QDA	QDA_Cross_Validation
Recall	0.72	0.65
Precission	0.52	0.54
F1-Score	0.60	0.59
Accuracy	0.73	0.75

COMPARING MODELS



CONCLUSION

Now that we have fitted different models, it’s time to choose the best model among all the available options. We have chosen the LDA model as our preferred choice. This model demonstrates a good recall and accuracy value, along with satisfactory precision and F1-score. However, it’s worth noting that the choice of the best model depends on various factors, including the specific requirements and priorities of the problem at hand.

RECOMMENDATIONS

- Consider exploring other classification models: Apart from the models mentioned earlier (logistic regression, KNN, and Naive Bayes), it is recommended to explore other classification architectures such as random forest, SVM, decision tree, and Ada boosting. These models may have different strengths and weaknesses and could potentially provide different results for the diabetes prediction problem.

- To enhance the performance of our model, it is recommended to improve the data collection process. The presence of a considerable number of missing values poses a challenge, which can be addressed by employing a more meticulous data collection strategy to minimize the occurrence of missing values. Additionally, expanding the dataset by collecting additional data would contribute to achieving more accurate and reliable results. By enhancing the quality and quantity of the dataset, we can potentially enhance the overall effectiveness of our model.

THANK YOU

Presented by

Bahador Mirzazadeh

Mohammad Matin Parvanian

Sadaf Jamali