Supervised Learning -- Classification Use Case with Diabetes Dataset

Business Understanding

- **Problem Statement**

  The problem we are addressing is the prediction of diabetes in patients based on various medical attributes. This is a classification problem where the goal is to accurately categorize patients as diabetic or non-diabetic.

- **Importance of the Problem**

  Predicting diabetes is critical due to its increasing prevalence and the severe health complications associated with it. Early detection can lead to better management and prevention strategies, improving patient outcomes and reducing healthcare costs.

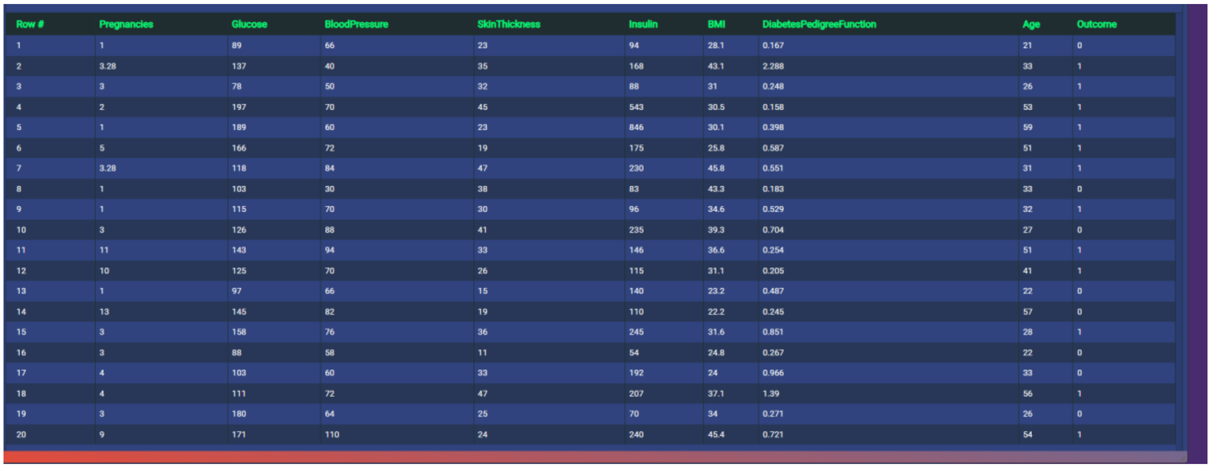
- **Data Source and Description**

  The dataset used for this analysis is sourced from the National Institute of Diabetes and Digestive and Kidney Diseases. The data includes various medical predictors such as age, body mass index (BMI), blood pressure, and others, which are used to predict whether a patient has diabetes (Outcome variable).

Data Collection

The dataset was downloaded from Kaggle, a popular data science and machine learning platform. The specific dataset used is the "Diabetes" dataset.

The dataset is named diabetes.csv and contains the following columns:



Row #	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
1	1	89	66	23	94	28.1	0.167	21	0
2	3.28	137	40	35	168	43.1	2.288	33	1
3	3	78	50	32	88	31	0.248	26	1
4	2	197	70	45	543	30.5	0.158	53	1
5	1	189	60	23	846	30.1	0.398	59	1
6	5	166	72	19	175	25.8	0.587	51	1
7	3.28	118	84	47	230	45.8	0.551	31	1
8	1	103	30	38	83	43.3	0.183	33	0
9	1	115	70	30	96	34.6	0.529	32	1
10	3	126	88	41	235	39.3	0.704	27	0
11	11	143	94	33	146	36.6	0.254	51	1
12	10	125	70	26	115	31.1	0.205	41	1
13	1	97	66	15	140	23.2	0.487	22	0
14	13	145	82	19	110	22.2	0.245	57	0
15	3	158	76	36	245	31.6	0.851	28	1
16	3	88	58	11	54	24.8	0.267	22	0
17	4	103	60	33	192	24	0.966	33	0
18	4	111	72	47	207	37.1	1.39	56	1
19	3	180	64	25	70	34	0.271	26	0
20	9	171	110	24	240	45.4	0.721	54	1

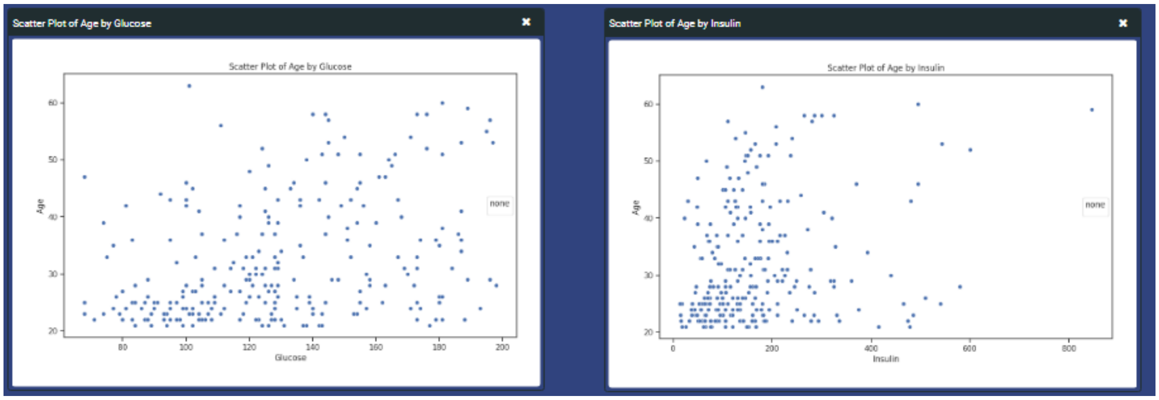
- Pregnancies: Number of times pregnant
- Glucose: Plasma glucose concentration
- BloodPressure: Diastolic blood pressure (mm Hg)
- SkinThickness: Triceps skinfold thickness (mm)
- Insulin: 2-Hour serum insulin (mu U/ml)
- BMI: Body mass index (weight in kg/(height in m)^2)
- DiabetesPedigreeFunction: Diabetes pedigree function
- Age: Age (years)
- Outcome: Class variable (0 or 1)

Data Understanding

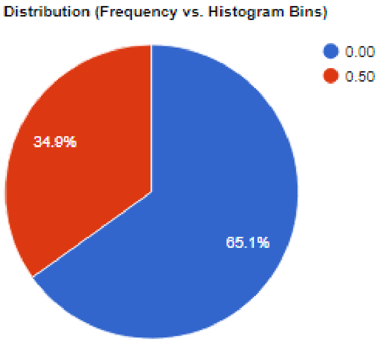
Exploratory Data Analysis (EDA)

Initial data exploration reveals the following:

- The dataset contains 768 rows and 9 columns.
- The target variable Outcome is binary, with 1 indicating diabetes and 0 indicating no diabetes.
- There are several missing or zero values in columns like Glucose, BloodPressure, SkinThickness, Insulin, and BMI.


Scatter Plot of Age by Glucose
✖
Scatter Plot of Age by Glucose
Age
Glucose
20
30
40
50
60
80
100
120
140
160
180
200
none
Scatter Plot of Age by Insulin
✖
Scatter Plot of Age by Insulin
Age
Insulin
20
30
40
50
60
0
200
400
600
800
none

Data Split



Obervations (Count of zero value)

Glucose: 5 instances
BloodPressure: 35 instances
SkinThickness: 227 instances
Insulin: 374 instances
BMI: 11 instances

Data Preparation

Handling Missing Values

Replaced zero values in Glucose, BloodPressure, SkinThickness, Insulin, and BMI with the mean values of their respective columns.

Data Splitting

The dataset was split into training and testing sets:

- 80% of the data was used for training.
- 20% of the data was reserved for testing.

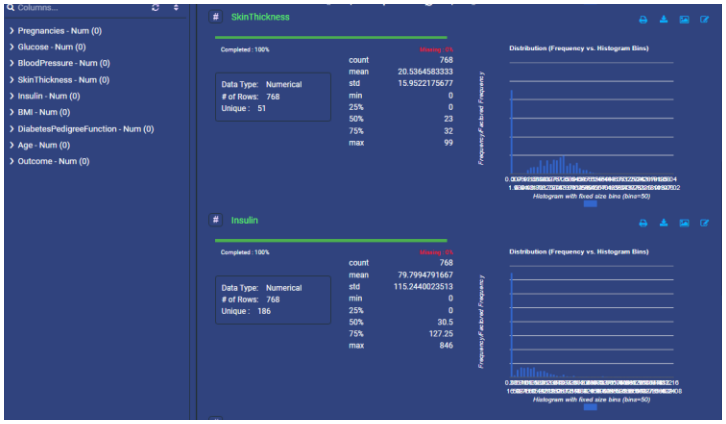
Methodology

Model Selection

For this classification task, we used the following algorithms:

- **Extra Trees Classifier**
  ExtraTrees Classifier is an ensemble ML approach that trains numerous decision trees and aggregates the results from the group of decision trees to output a prediction
- **Quadratic Discriminant Analysis**
  Quadratic Discriminant Analysis (QDA) is a powerful classification technique used in machine learning to distinguish between different groups or classes based on their features. It is particularly useful for handling heteroscedastic data, where the variability within each group is different.
- **Random Forest Classifier**
  Random forest is a machine learning algorithm used for classification and regression tasks. It excels at prediction accuracy by leveraging the power of aggregating decision trees.

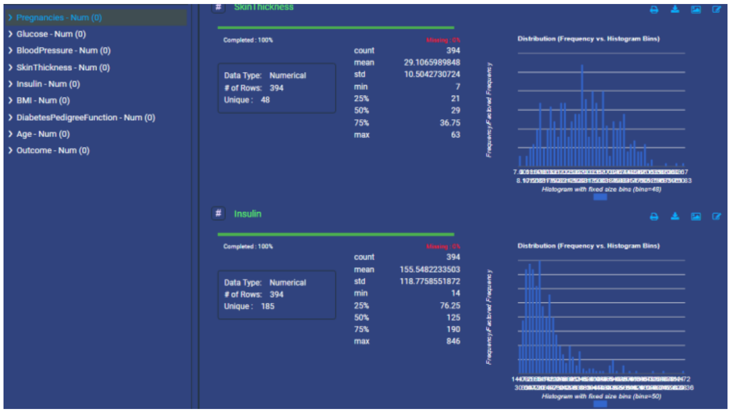
Original Data Distribution of Skin Thickness and Insulin


Columns...
Pregnancies - Num (0)
Glucose - Num (0)
BloodPressure - Num (0)
SkinThickness - Num (0)
Insulin - Num (0)
BMI - Num (0)
DiabetesPedigreeFunction - Num (0)
Age - Num (0)
Outcome - Num (0)
Distribution (Frequency vs. Histogram Bins)
Frequency/Factored Frequency
Histogram with fixed size bins (bins=50)
Distribution (Frequency vs. Histogram Bins)
Frequency/Factored Frequency
Histogram with fixed size bins (bins=50)

```
# SkinThickness
Completed : 100%                         Missing : 0%
                          count                  768
Data Type:   Numerical    mean       20.5364583333
# of Rows:   768          std        15.9522175677
Unique :     51           min                    0
                          25%                    0
                          50%                   23
                          75%                   32
                          max                   99

# Insulin
Completed : 100%                         Missing : 0%
                          count                  768
Data Type:   Numerical    mean       79.7994791667
# of Rows:   768          std       115.2440023513
Unique :     186          min                    0
                          25%                    0
                          50%                 30.5
                          75%               127.25
                          max                  846
```

Wrangled Data Distribution of Skin Thickness and Insulin


Pregnancies - Num (0)
Glucose - Num (0)
BloodPressure - Num (0)
SkinThickness - Num (0)
Insulin - Num (0)
BMI - Num (0)
DiabetesPedigreeFunction - Num (0)
Age - Num (0)
Outcome - Num (0)
Distribution (Frequency vs. Histogram Bins)
Frequency/Factored Frequency
Histogram with fixed size bins (bins=48)
Distribution (Frequency vs. Histogram Bins)
Frequency/Factored Frequency
Histogram with fixed size bins (bins=50)

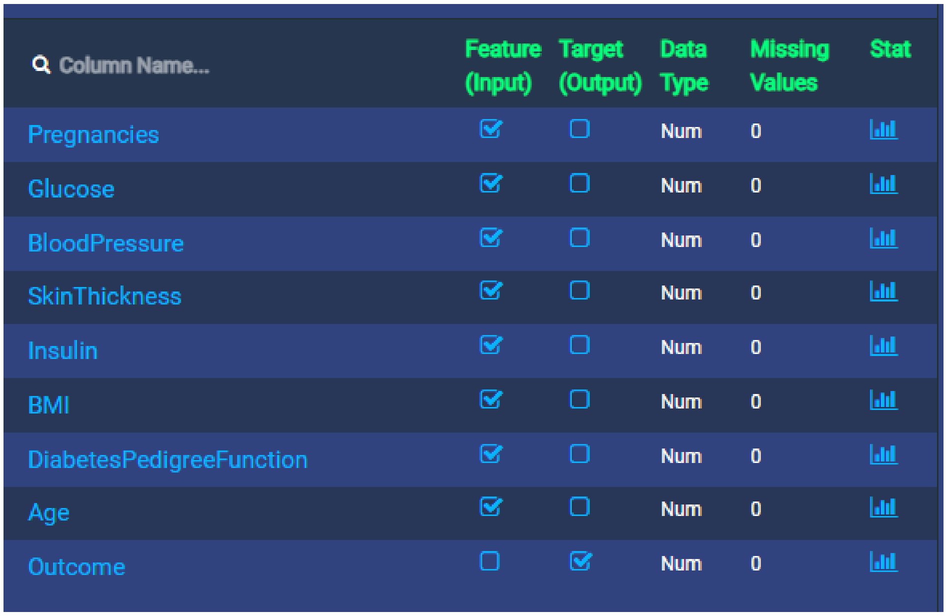
```
# SkinThickness
Completed : 100%                         Missing : 0%
                          count                  394
Data Type:   Numerical    mean       29.1065989848
# of Rows:   394          std        10.5042730724
Unique :     48           min                    7
                          25%                   21
                          50%                   29
                          75%                36.75
                          max                   63

# Insulin
Completed : 100%                         Missing : 0%
                          count                  394
Data Type:   Numerical    mean      155.5482233503
# of Rows:   394          std       118.7758551872
Unique :     185          min                   14
                          25%                76.25
                          50%                  125
                          75%                  190
                          max                  846
```

Model Training

Each model was trained using the training dataset.

Feature and Target:



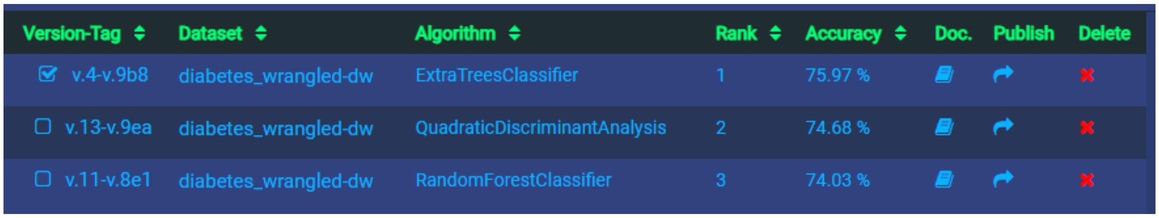
Column Name...	Feature (Input)	Target (Output)	Data Type	Missing Values	Stat
Pregnancies	☑	□	Num	0	📊
Glucose	☑	□	Num	0	📊
BloodPressure	☑	□	Num	0	📊
SkinThickness	☑	□	Num	0	📊
Insulin	☑	□	Num	0	📊
BMI	☑	□	Num	0	📊
DiabetesPedigreeFunction	☑	□	Num	0	📊
Age	☑	□	Num	0	📊
Outcome	□	☑	Num	0	📊

Model Evaluation

The models were evaluated on the test dataset using the following metrics:

- Accuracy
- Precision
- Recall
- F1 Score
- Jaccard Score

Results



Version-Tag ⬍	Dataset ⬍	Algorithm ⬍	Rank ⬍	Accuracy ⬍	Doc.	Publish	Delete
☑ v.4-v.9b8	diabetes_wrangled-dw	ExtraTreesClassifier	1	75.97 %	📘	➦	✖
□ v.13-v.9ea	diabetes_wrangled-dw	QuadraticDiscriminantAnalysis	2	74.68 %	📘	➦	✖
□ v.11-v.8e1	diabetes_wrangled-dw	RandomForestClassifier	3	74.03 %	📘	➦	✖

Performance Metrics

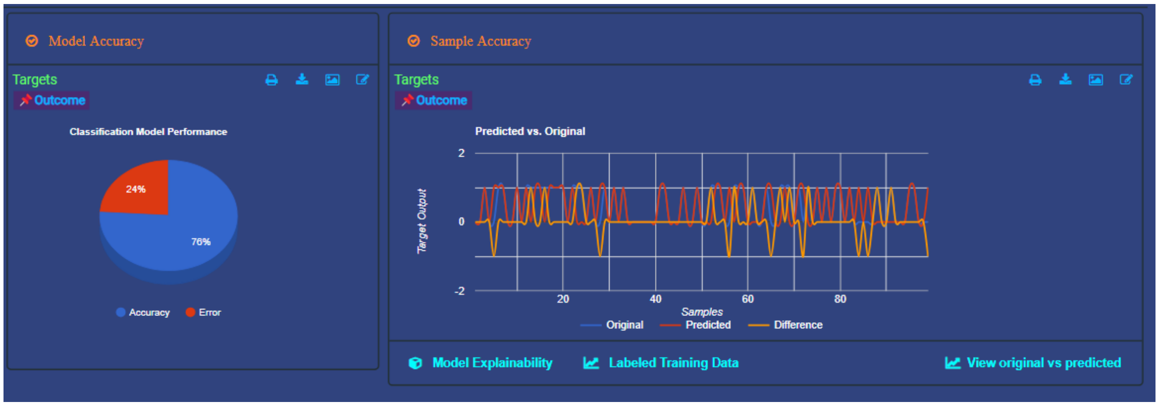
The performance metrics for each model are summarized as follows:

Model	Accuracy	Precision	Recall	F1 Score	Jaccard Score
Extra Trees Classifier	75.97%	0.76	0.76	75.97	0.6
Quadratic Discriminant Analysis	74.68%	0.75	0.75	74.68	0.59
Random Forest Classifier	74.03%	0.74	0.74	74.03	0.58

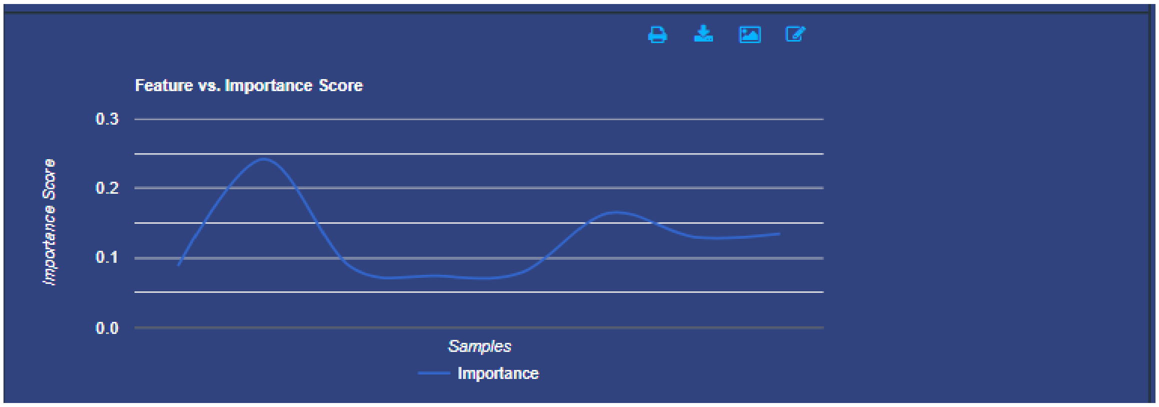
Best Model

The Extra Trees Classifier performed the best with the highest accuracy.

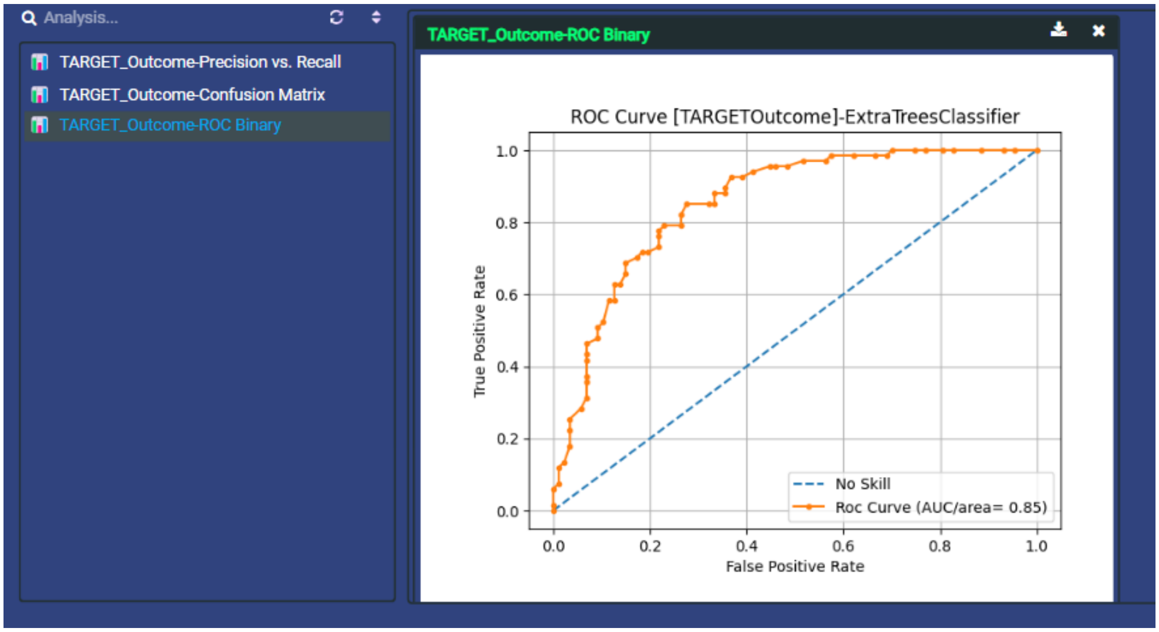
Model Accuracy



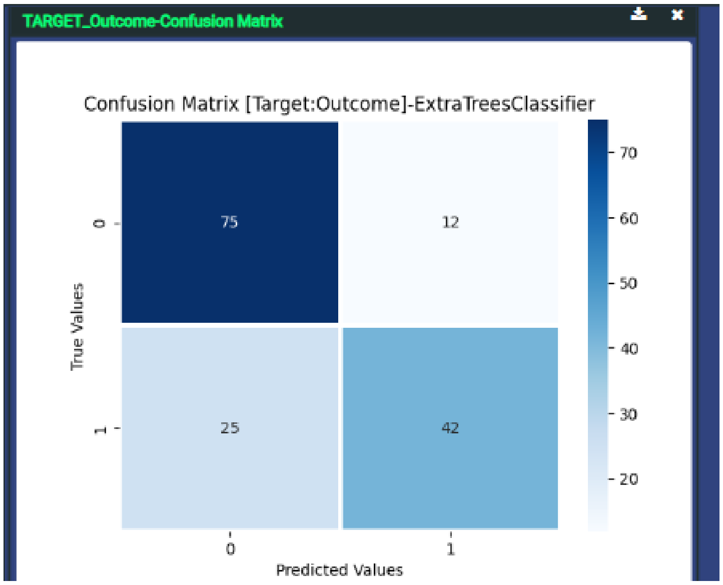
Feature Importance



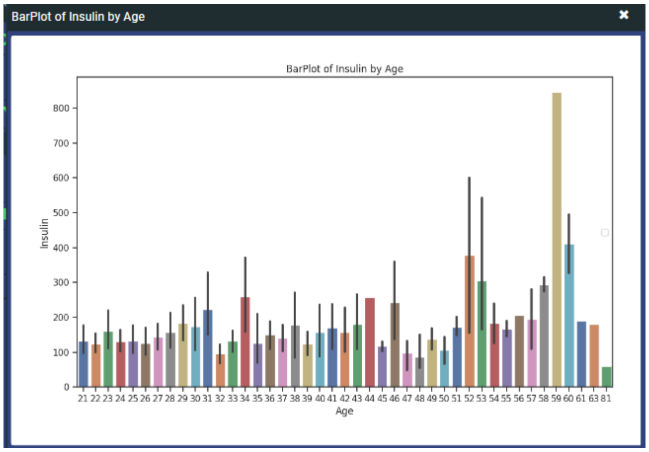
ROC Binary

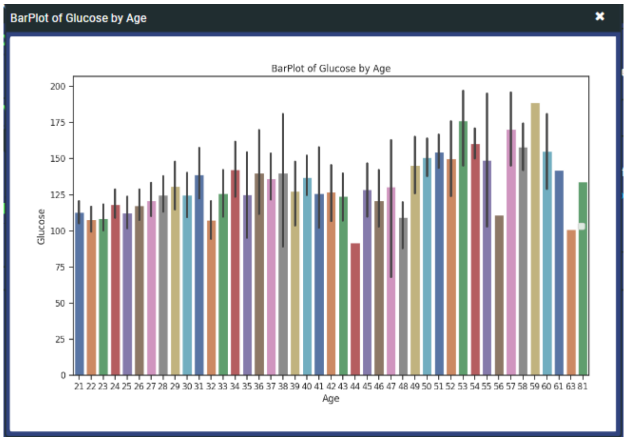

Analysis...
TARGET_Outcome-Precision vs. Recall
TARGET_Outcome-Confusion Matrix
TARGET_Outcome-ROC Binary
TARGET_Outcome-ROC Binary
ROC Curve [TARGETOutcome]-ExtraTreesClassifier
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
No Skill
Roc Curve (AUC/area= 0.85)

Confusion Matrix

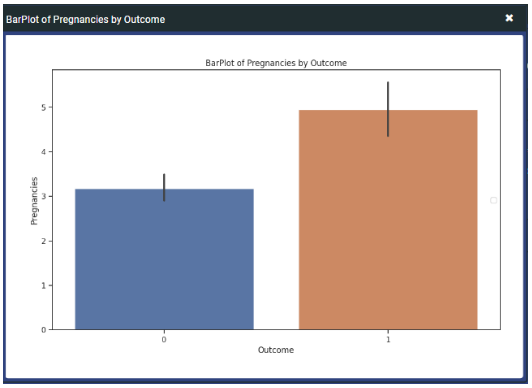

TARGET_Outcome-Confusion Matrix
Confusion Matrix [Target:Outcome]-ExtraTreesClassifier
True Values
Predicted Values
0
1
75
12
25
42
70
60
50
40
30
20

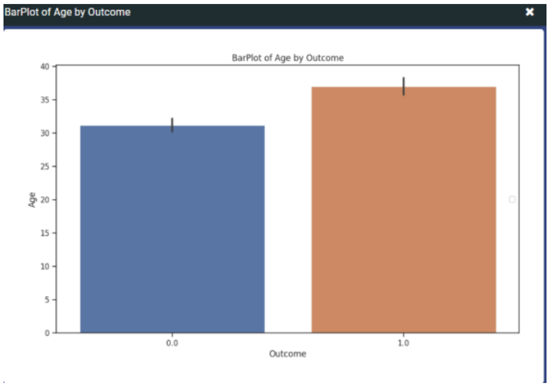
BarPlot of Glucose and Insulin by Age

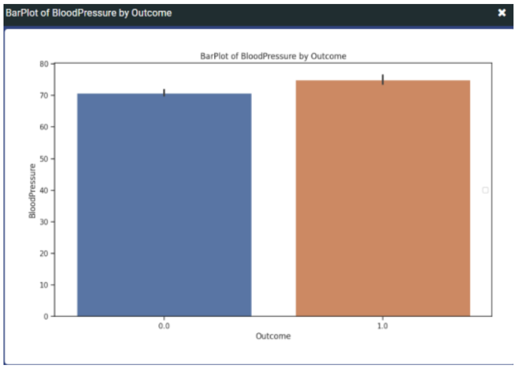


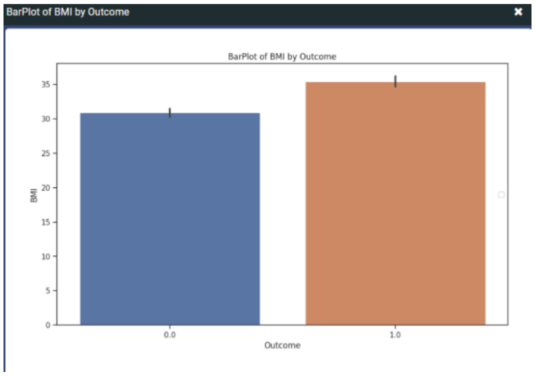


Target (Outcome) by Features (Pregnancies, BMI, Age, Blood Pressure)









Conclusions

Improvements

What improvements would you like to make in future?

Include more relevant features to improve the model's predictive power.

Apply advanced techniques like ensemble learning to boost model performance.

Real-life Application

How do you think the solution could be used in real life?

The solution can be integrated into healthcare systems to assist doctors in early diabetes diagnosis, leading to timely intervention and better patient management.

Value to Client

What value do you think the solution will have to the client?

The model provides a reliable tool for predicting diabetes, which can enhance patient care and reduce long-term healthcare costs by enabling early detection and prevention strategies.

Key Learnings

What did you learn through this project?

The importance of data preprocessing and feature engineering in building effective models.

The value of model evaluation metrics in selecting the best model for deployment.

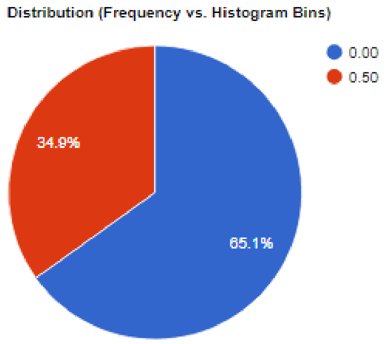
This project provided valuable insights into the application of machine learning in healthcare, demonstrating the potential impact of data-driven solutions in real-world scenarios.

Iteration 2

Analysis before Iteration 2

Before iteration 2 our data accuracy was 74% with ROC was 85 and our data was not balanced.

Not Balanced Data



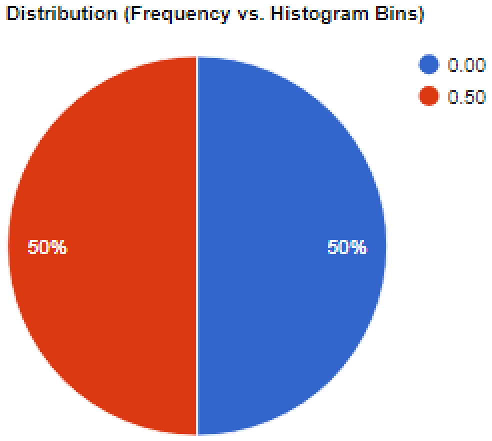
Analysis after Iteration 2

In Iteration 2 we did data balancing by under sampling, now our data is balanced.

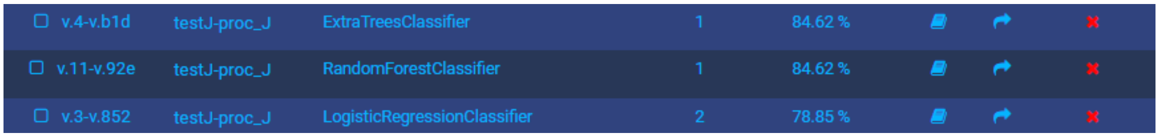
Accuracy improved to 88% from 74%.

ROC improved to 88 from 85.

Balanced Data

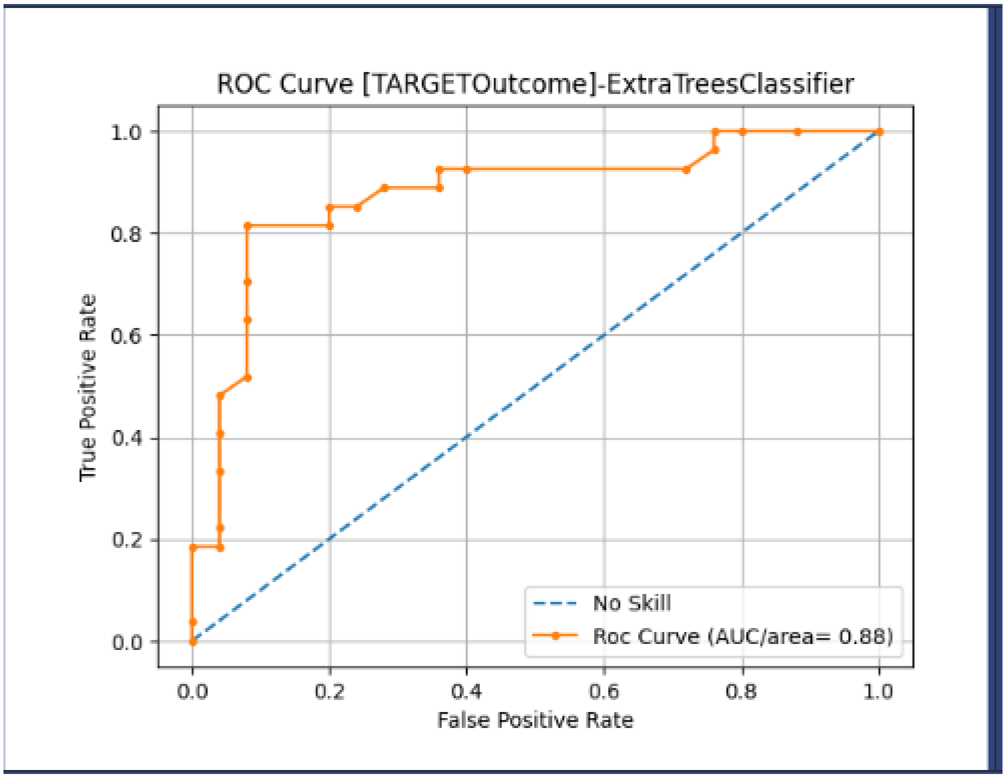


Top 3 Classifier



□ v.4-v.b1d	testJ-proc_J	ExtraTreesClassifier	1	84.62 %	📘	↪	✖
□ v.11-v.92e	testJ-proc_J	RandomForestClassifier	1	84.62 %	📘	↪	✖
□ v.3-v.852	testJ-proc_J	LogisticRegressionClassifier	2	78.85 %	📘	↪	✖

ROC


ROC Curve [TARGETOutcome]-ExtraTreesClassifier
True Positive Rate
False Positive Rate
No Skill
Roc Curve (AUC/area= 0.88)