**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.

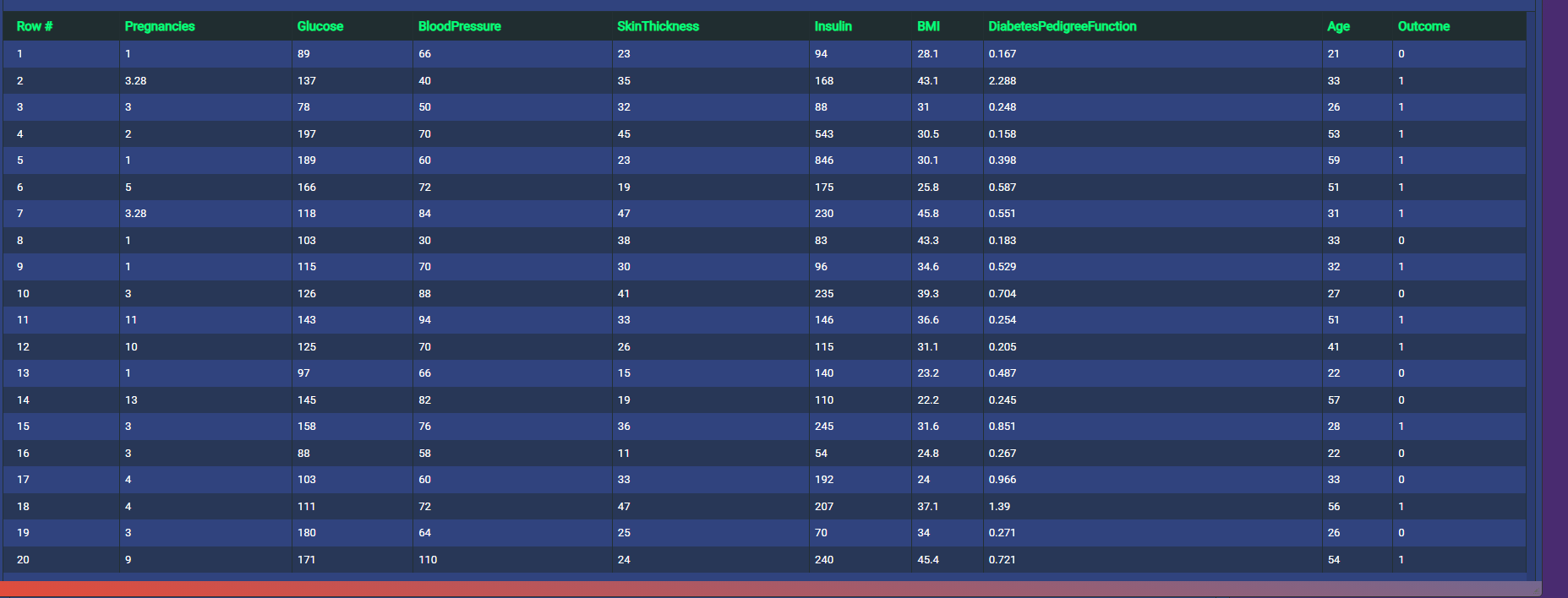
* **Data Source and Description**

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](https://www.kaggle.com/datasets/mathchi/diabetes-data-set), 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:**

****

* 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)**

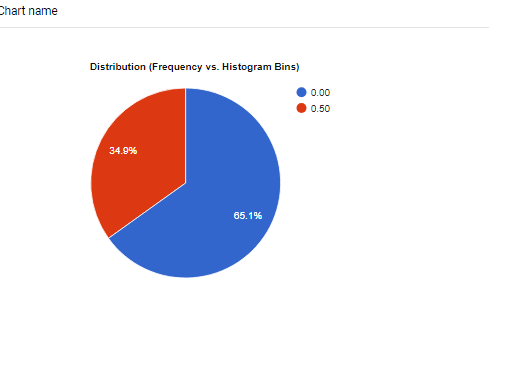
Initial data exploration reveals the following:

* 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.



**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

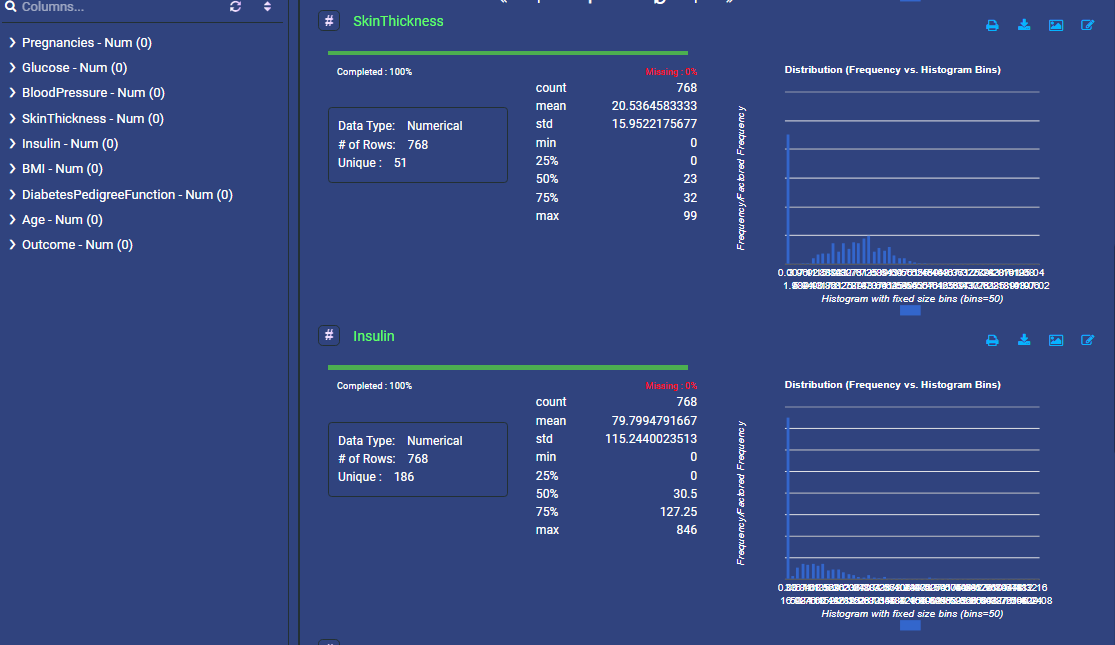
* **Quadratic Discriminant Analysis**

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.

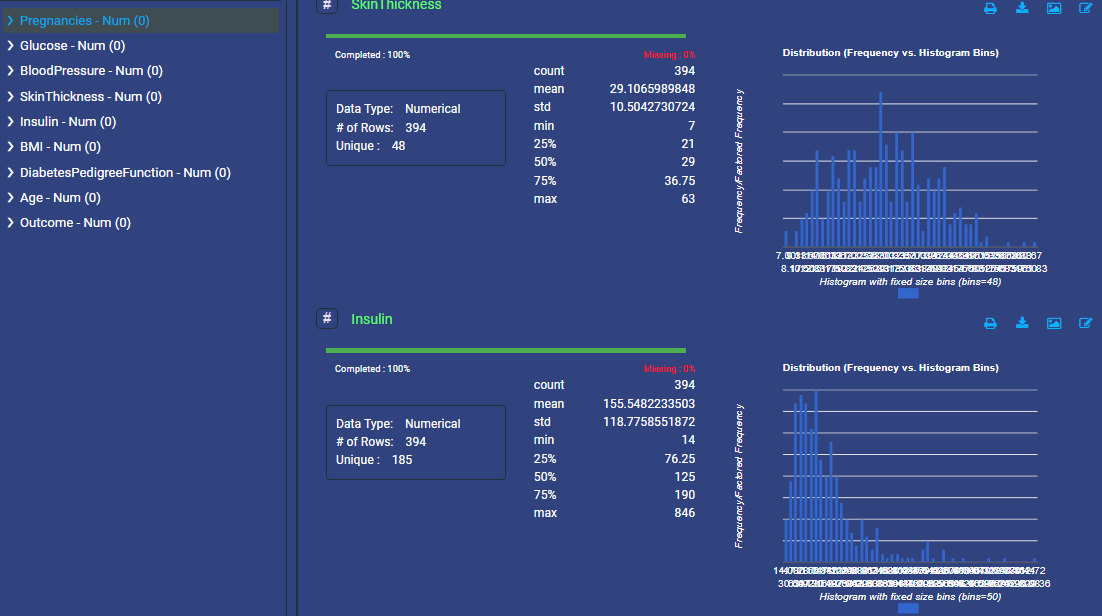
* **Random Forest Classifier**

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**

****

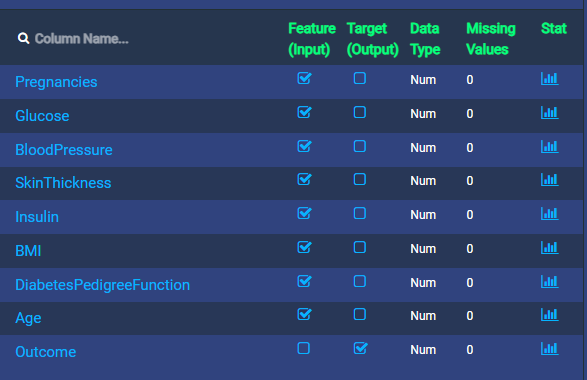
**Wrangled Data Distribution of Skin Thickness and Insulin**

****

**Model Training**

Each model was trained using the training dataset.

**Feature and Target:**

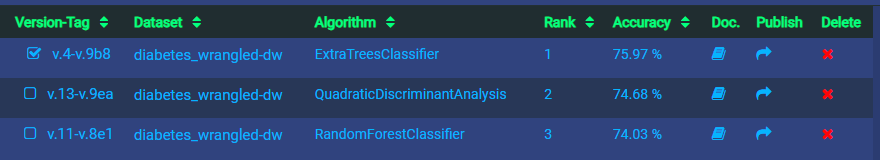


**Model Evaluation**

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

* Accuracy
* Precision
* Recall
* F1 Score
* Jaccard Score

**Results**

****

**Performance Metrics**

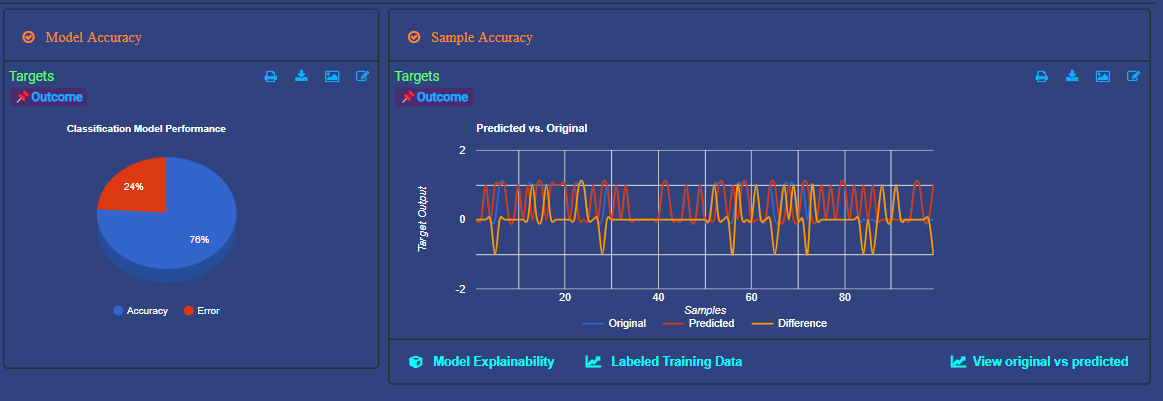
The performance metrics for each model are summarized as follows:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 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 |

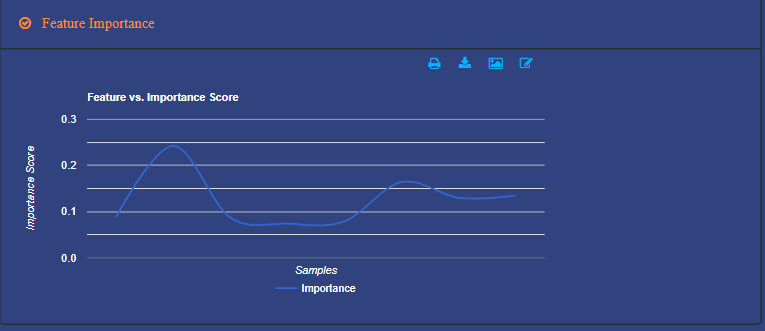
**Best Model**

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

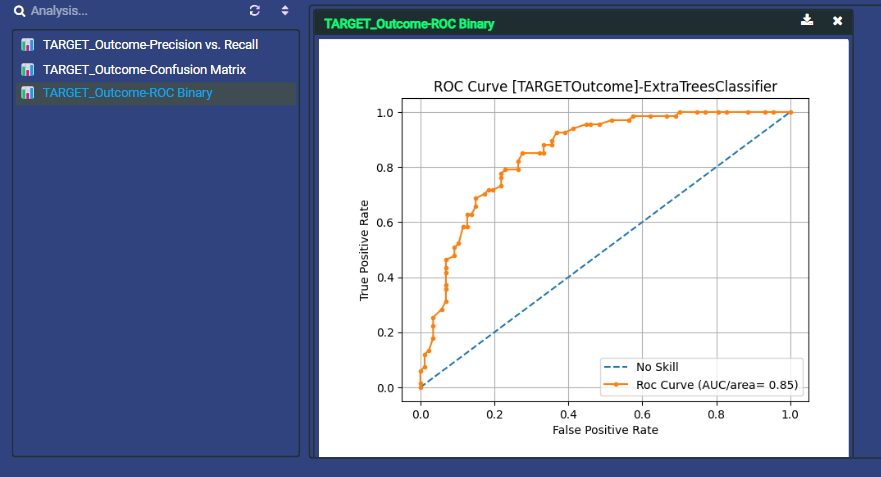
**Model Accuracy**

****

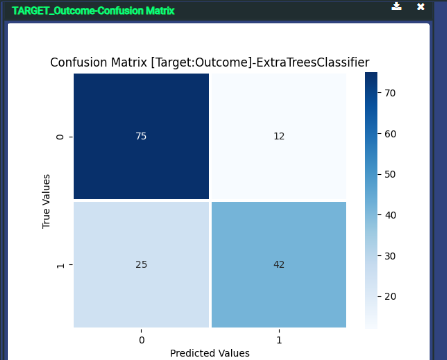
**Feature Importance**

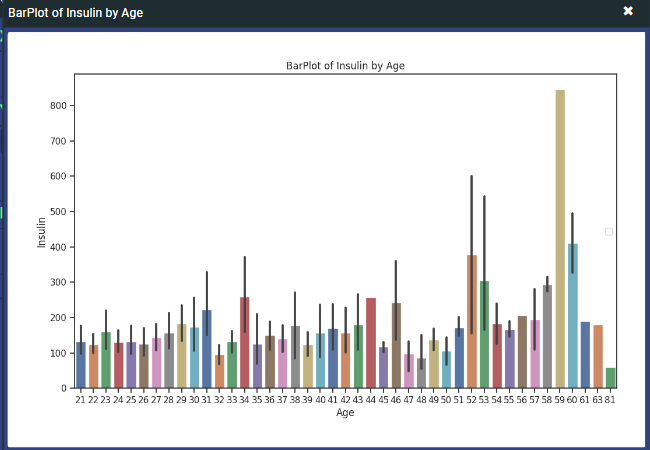
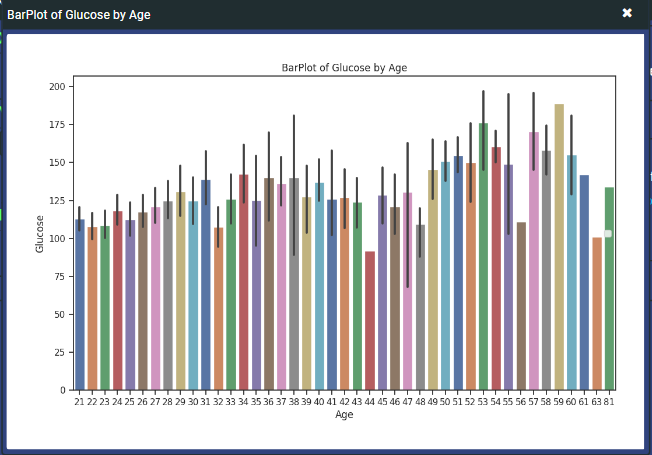
****

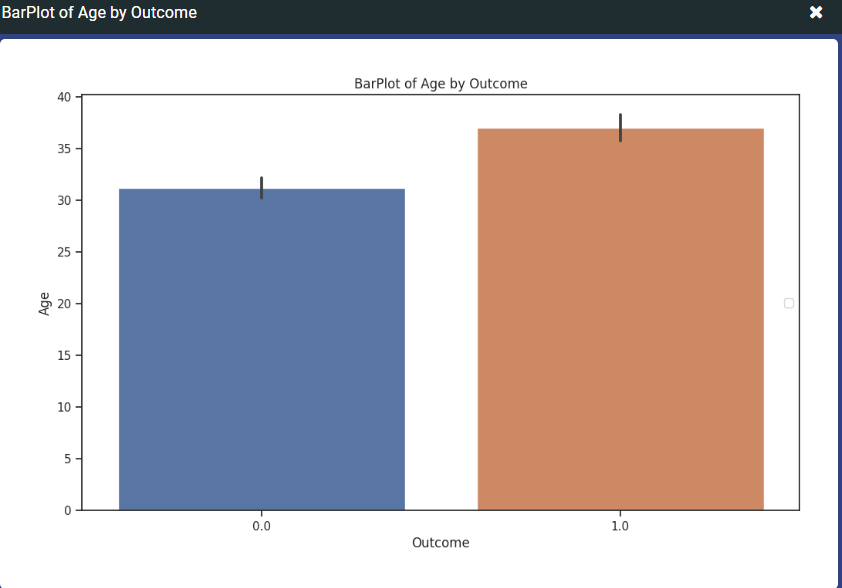
**ROC Binary**

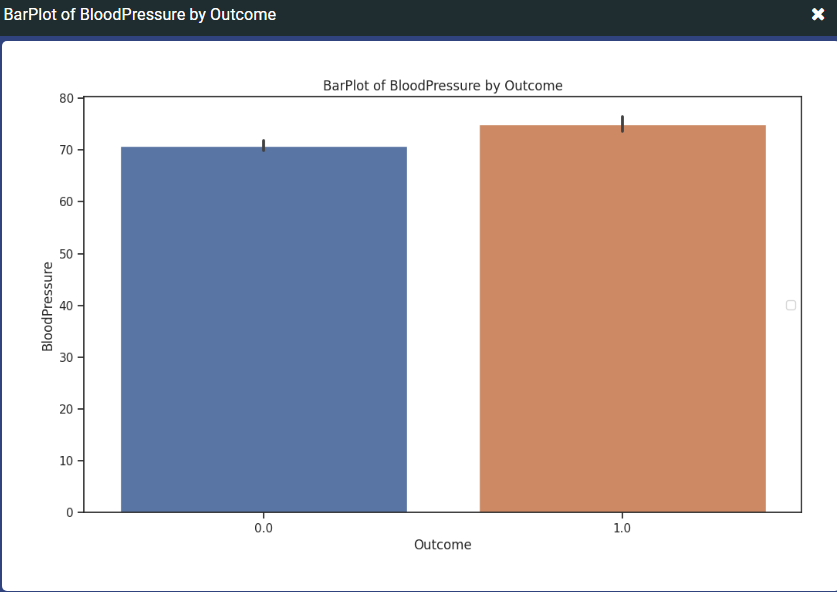
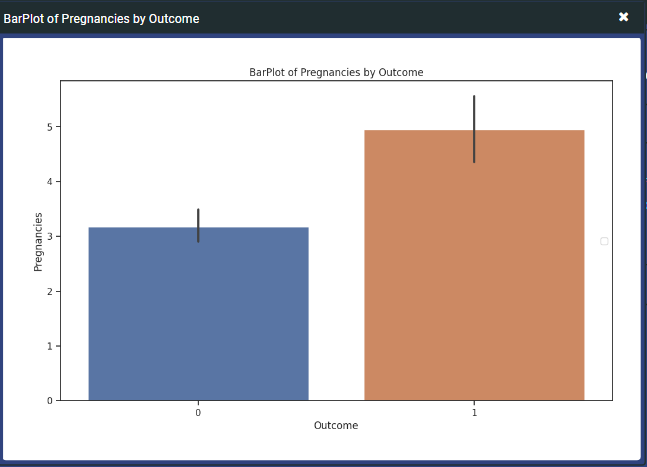
****

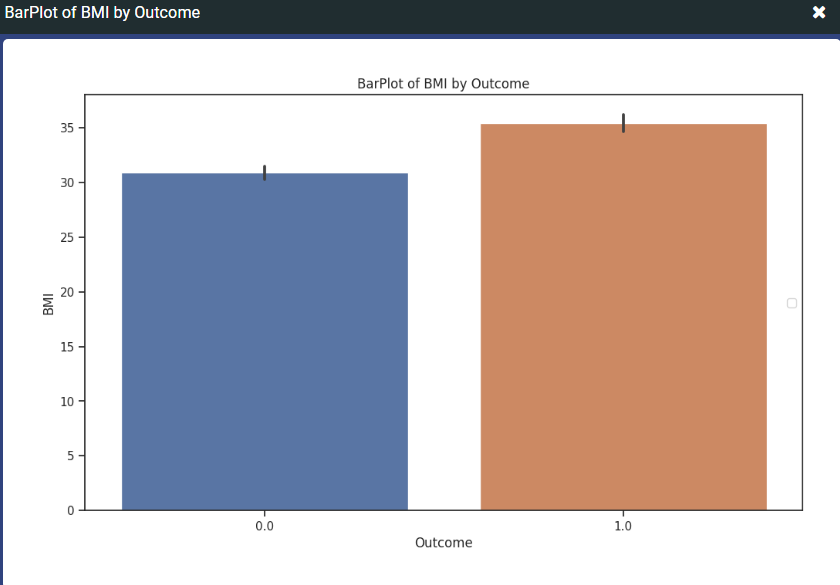
**Confusion Matrix**

****

**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.

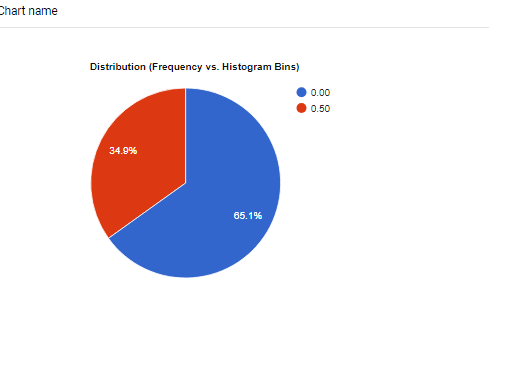
This project provided valuable insights into the application of machine learning in healthcare, demonstrating the potential impact of data-driven solutions in real-world scenarios.

**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**



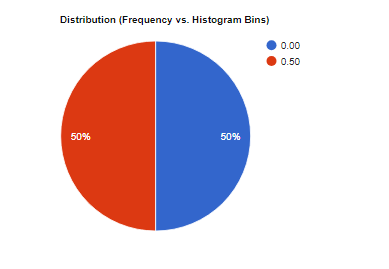
**Analysis after Iteration 2**

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

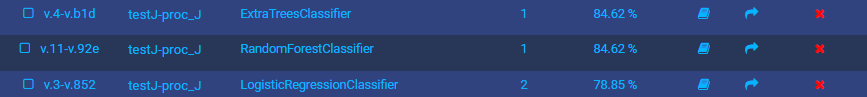
Accuracy improved to 88% from 74%.

ROC improved to 88 from 85.

**Balanced Data**



**Top 3 Classifier**



**ROC**

