# Detecting Diabetes

**Prepared By: Erum Akmal**

1. **Problem Statement for Project:**
   In this project, I have predicted diabetes using Naive Bayes Classification.

2. **Data Sources and Description:**
   For data sources I used the data set for diabetes list shared by the instructor for the assignment.

   For this project, I worked on a diabetes project using Naïve Bayes classification. My data set included diabetes as dependent variable and glucose and blood pressure as independent variables.

3. **Data Mining Operations:**
   -Data wrangling:
   I started the project by loading and describing the data set. Preprocessing was not required in this project since there were no null values and all values were integers.

   Data was partitioned into 80% training and 20% testing. Model was trained using GaussianNB and target predictions were made.

   After the confusion matrix, classification report was generated.

   -Libraries Used: I have used multiple libraries in this project including pandas, NumPy, matplotlib, seaborn, scikit-learn.

   -Data Modeling: I have performed the analysis below in my project:

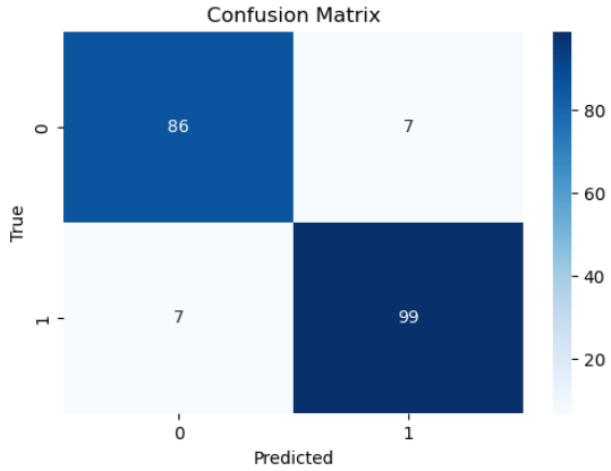
   - Splitting dataset into 80-20
   - Modelling with GaussianNB
   - Confusion Matrix
   - ROC Curve
   - Correlation Heatmap
   - Histogram of diff or target and prediction
   - Box plot for dataset
   - Scatter Plot for NB Model

4. **Model Outputs:**
   Below is my evaluation of the model performance:

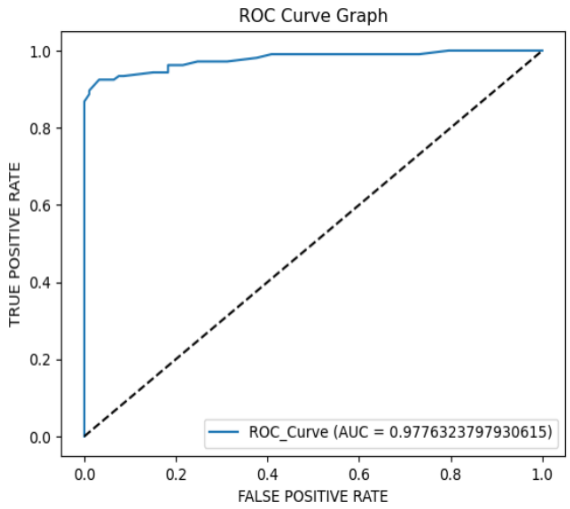
```
Confusion Matrix:
 [[86  7]
 [ 7 99]]
Accuracy:  0.9296482412060302
```

Overall, this model is performing well since accuracy is 0.92

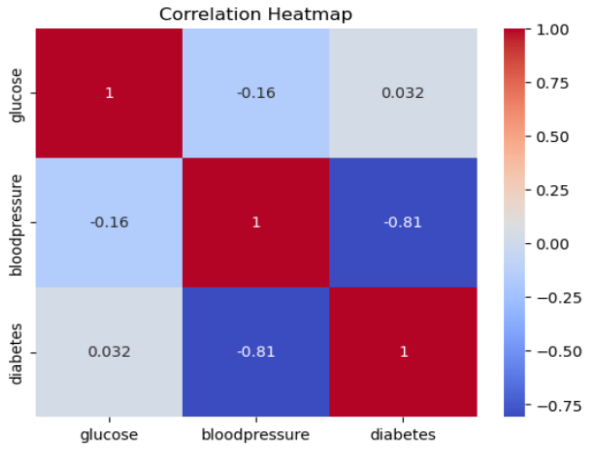
The graph for Confusion matrix was generated first.




After that the ROC curve was generated.

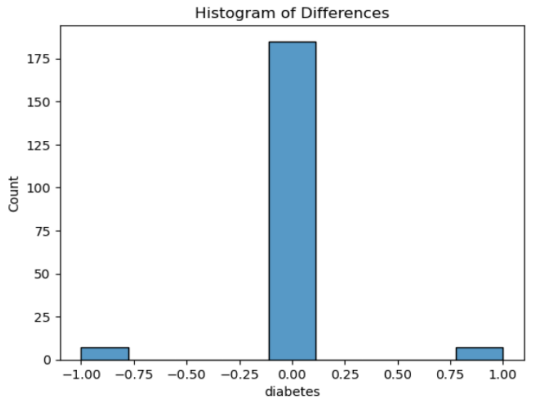


After that, the correlation heatmap was plotted.

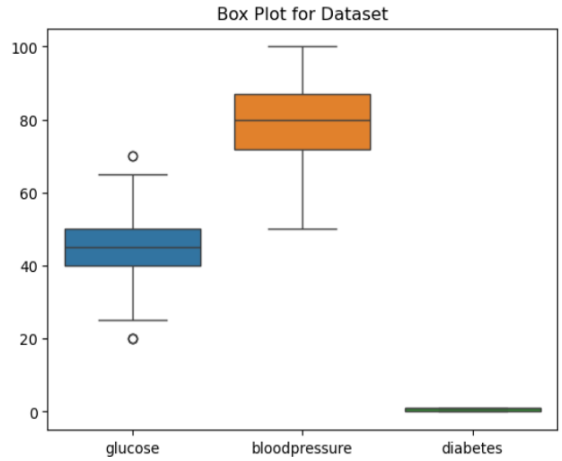



After the analysis, I can observe a positive correlation between diabetes and glucose. The correlation between diabetes and blood pressure is negative but strong. A 92% model accuracy is explained above.

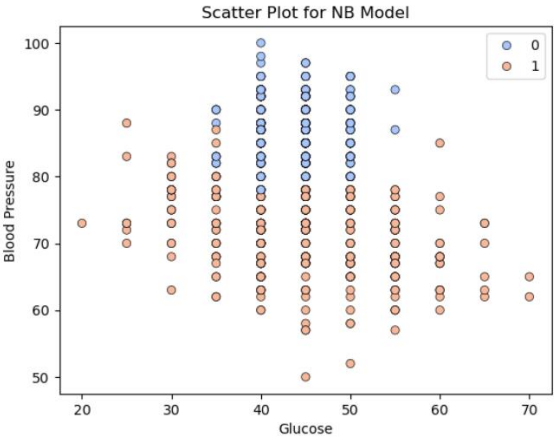
Graph for histogram of differences was generated after that.




Box plot for data set was plotted next.




At the end, the scatter model for NB model was generated.




## 5. Limitations and Challenges:

Some of the limitations of this study are that I have only used 2 variables in this study for predicting diabetes. There could be multiple other factors that can potentially affect diabetes. This model worked well but there might be multiple other reasons that can affect diabetes which were not covered in the data set.

## 6. Conclusion:

I was able to solve the problem partially. By adding more data and variables in this study, a better model could have been predicted for diabetes. By adding a greater number of features and rows, we can test with multiple factors that can play a role in improving diabetes.

Model performance is good over here which can be observed in ROC Curve blue line in graph shown above and accuracy is 92% which confirms that model has performed well.