Assignment Report: Predicting Diabetes Risk using Logistic Regression

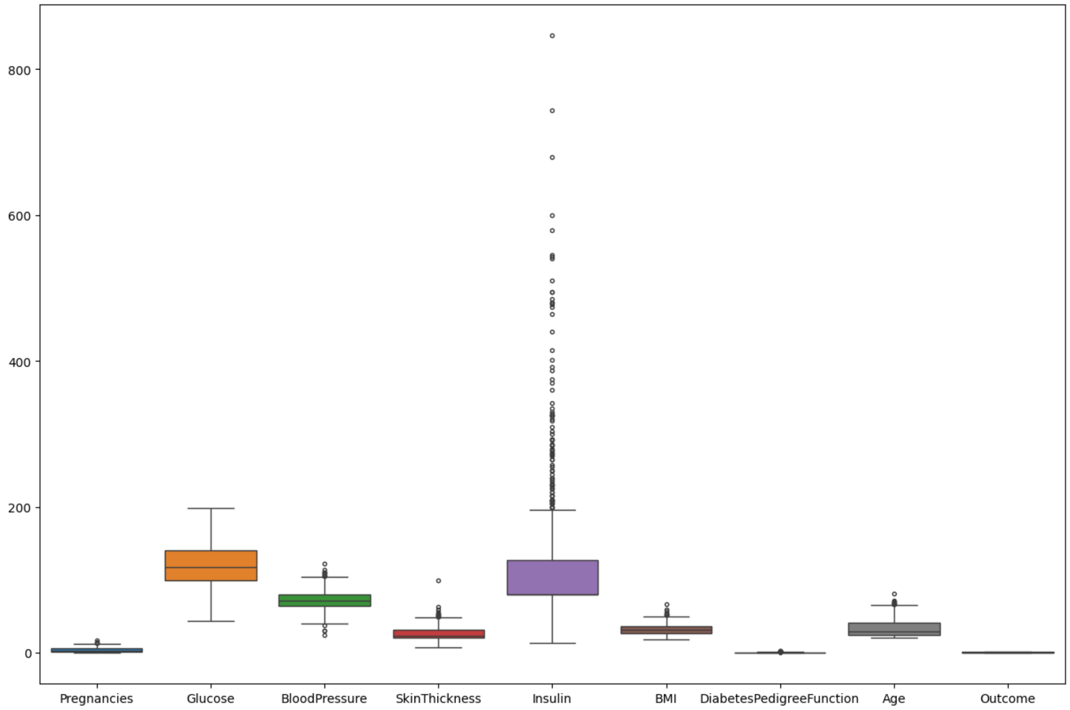
Submitted by: Shree Prada(C0887257)

Introduction:

In this assignment, I aimed to develop a predictive model to assess the risk of diabetes using logistic regression. The dataset used contains various health-related features such as pregnancies, glucose levels, blood pressure, skin thickness, insulin levels, BMI, diabetes pedigree function, and age. The objective was to explore the dataset, preprocess it, build a predictive model, optimize its performance, and interpret the results.

Data Exploration and Preprocessing:

The dataset was initially explored to understand its structure and contents. I observed that some features had missing values represented as 0. These missing values were replaced with the mean values of their respective columns to ensure data integrity. Furthermore, I examined the distribution of the target variable, 'Outcome', and found a significant class imbalance. To address this issue, I employed the Synthetic Minority Over-sampling Technique (SMOTE) to balance the class distribution.



Model Development and Optimization:

We chose logistic regression as our predictive model due to its simplicity and interpretability. Before fitting the model, I split the dataset into training and testing sets. I then standardized the features to ensure that they were on the same scale.

Model Evaluation and Performance Metrics:

After training the model, I evaluated its performance using various metrics including accuracy, precision, recall, F1-score, confusion matrix, and ROC AUC. These metrics provided insights into how well the model performed in predicting diabetes risk. Additionally, I plotted the Receiver Operating Characteristic (ROC) curve and calculated the Area Under the Curve (AUC) to assess the model's discriminatory power.

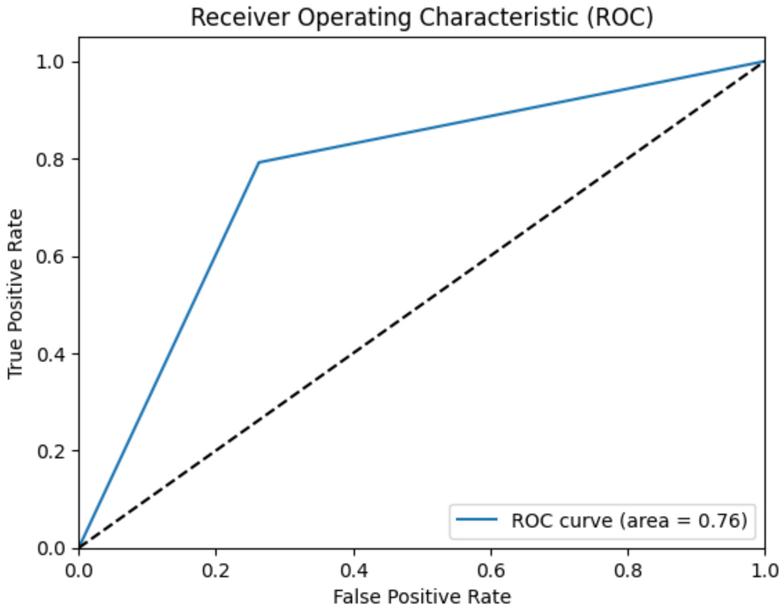
- Accuracy: 76.5%

- Precision: 75.5%

- Recall: 79.2%

- F1-score: 77.3%

- ROC AUC Score: 76.5%



Interpretation of Model Coefficients:

To gain further insights into the factors influencing diabetes risk, I examined the coefficients of the logistic regression model. I interpreted these coefficients to understand the relationship between each feature and the likelihood of diabetes. This analysis provided valuable insights into which features were most predictive of diabetes risk.

Insights from Model Coefficients:

- Pregnancies: An increase in the number of pregnancies is associated with an increase in the likelihood of diabetes.
- Glucose: Higher glucose levels are associated with an increased likelihood of diabetes.
- Blood Pressure: Higher blood pressure is associated with a decreased likelihood of diabetes.
- Skin Thickness: This feature does not have a significant impact on the likelihood of diabetes.
- Insulin: Higher insulin levels are associated with a decreased likelihood of diabetes.
- BMI: Higher BMI is associated with an increased likelihood of diabetes.
- Diabetes Pedigree Function: An increase in this function is associated with an increased likelihood of diabetes.
- Age: Older age is associated with an increased likelihood of diabetes.

Conclusion:

In conclusion, I successfully developed a logistic regression model to predict the risk of diabetes based on various health-related features. The model demonstrated moderate performance in terms of accuracy and other evaluation metrics. Glucose, BMI, and Diabetes Pedigree Function were identified as significant predictors of diabetes risk. This information can be valuable for healthcare professionals in identifying individuals at risk of developing diabetes and implementing preventive measures.

References:

- Python Documentation

- Scikit-learn Documentation

- Pandas Documentation

- Seaborn Documentation

- Matplotlib Documentation

- Stack Overflow (for specific coding queries)