Neural Networks and Deep Learning

Report: Predictive Model for Diabetes Detection

I. Objective:

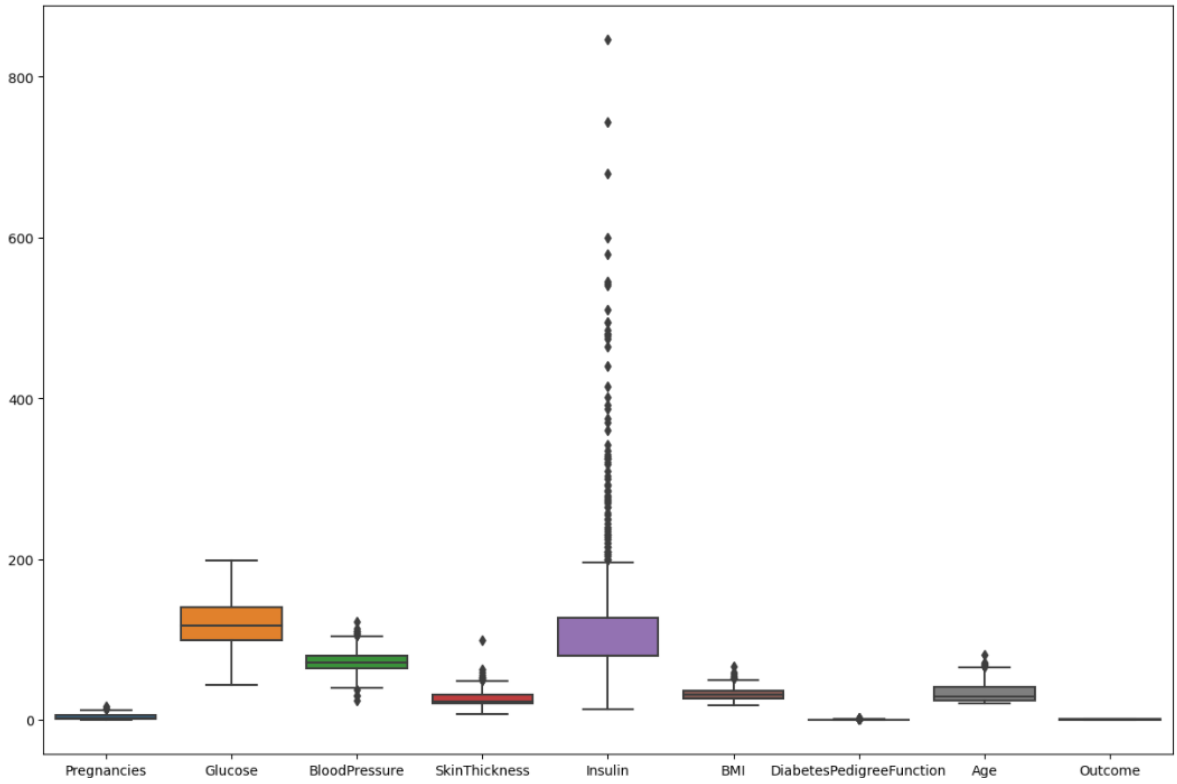
The objective of this project is to build a predictive model to determine the likelihood of a patient having diabetes based on certain features.

II. Dataset Description:

The dataset used for this project is named "diabetes2.csv". It contains various features related to diabetes diagnosis, such as glucose levels, BMI, blood pressure, skin thickness, insulin levels, and pregnancy status. The target variable is "Outcome", which indicates whether the patient has diabetes or not (0 for non-diabetic, 1 for diabetic).

III. Data Preprocessing:

The dataset was loaded using Pandas, and its structure was examined by viewing the first few rows with the head() function. The column names were checked using the keys() function to understand the available features. Statistical summaries of the numerical features were generated using the describe() function, providing insights into their distributions. To handle missing or unrealistic values, zero values in numerical columns such as Glucose, BMI, BloodPressure, SkinThickness, and Insulin were replaced with their respective means. The distribution of diabetic and non-diabetic cases was then examined using the value_counts() function. Finally, the distribution of features was visualized using a boxplot, aiding in the understanding of their spread and central tendencies.

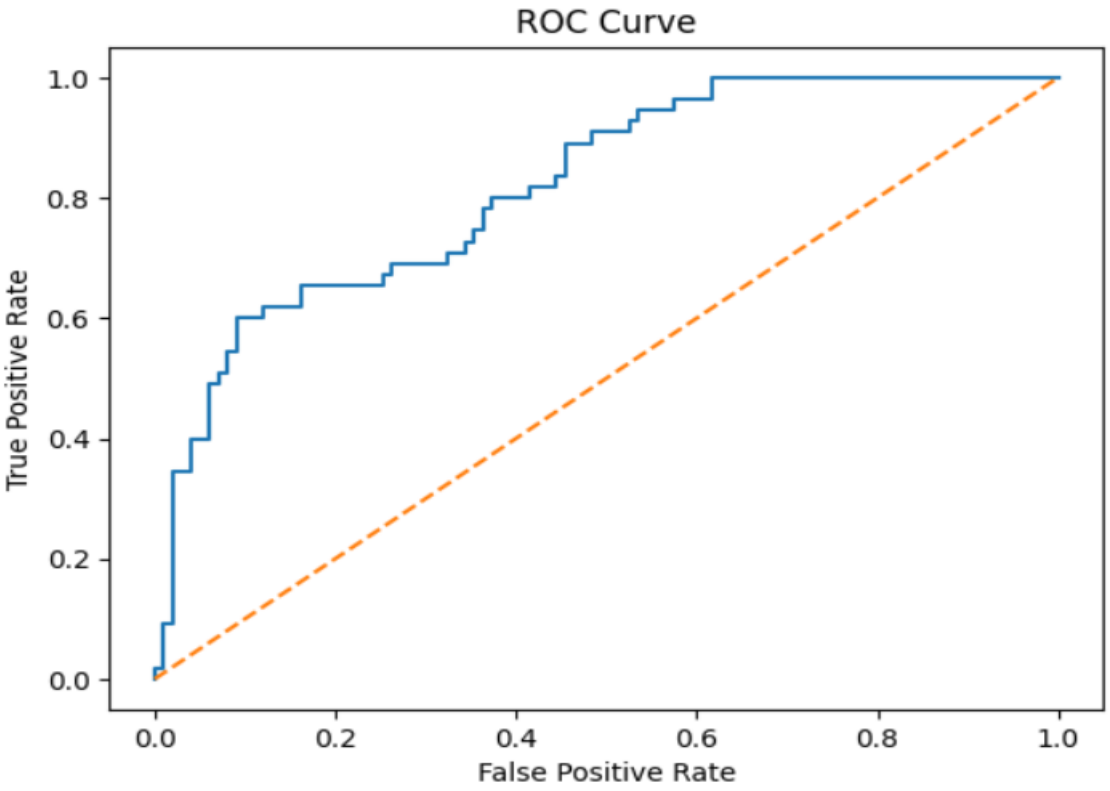


IV. Key Findings:

- The dataset comprises various features related to diabetes diagnosis, including glucose levels, BMI, blood pressure, skin thickness, insulin levels, and pregnancy status.
- After preprocessing the data by replacing zero values with feature means, it was observed that the majority of the patients in the dataset were non-diabetic.
- Glucose, BMI, and blood pressure were identified as significant predictors of diabetes risk based on their coefficients in the logistic regression model.
- The logistic regression model achieved satisfactory performance in predicting the likelihood of a patient having diabetes, as indicated by the evaluation metrics.

V. Model Building:

The dataset was divided into features (X) and the target variable (y), followed by splitting into training and testing sets using the train_test_split() function. StandardScaler was applied to standardize the numerical features. Logistic Regression was chosen as the predictive model for its interpretability and effectiveness in binary classification tasks. The model was trained on the training data and used to predict the target variable for the testing data. Performance evaluation included metrics such as accuracy, precision, recall, F1-score, and ROC AUC score. Additionally, a Receiver Operating Characteristic (ROC) curve was plotted to visualize the trade-off between true positive rate and false positive rate.



VI. Model Performance Metrics:

The logistic regression model achieved a moderate level of performance in predicting diabetes likelihood. The confusion matrix reveals that out of 154 instances, 85 were correctly classified as non-diabetic, and 34 were correctly classified as diabetic, while 14 non-diabetic instances were falsely classified as diabetic and 21 diabetic instances were falsely classified as non-diabetic. The overall accuracy of the model is 77.27%, indicating its ability to correctly classify instances into their respective classes. The precision score of 70.83% suggests that when the model predicts an instance as diabetic, it is correct approximately 70.83% of the time. The recall score of 61.82% indicates the model's ability to correctly identify diabetic instances out of all actual diabetic cases. The F1-score, which balances precision and recall, is 66.02%. Additionally, the ROC AUC score of 73.84% demonstrates the model's ability to discriminate between diabetic and non-diabetic instances.

VII. Insights from Model Coefficients:

The intercept (-0.87387862) represents the predicted value of the dependent variable (likelihood of diabetes) when all independent variables are zero. In this context, a negative intercept suggests that even with no pregnancies, glucose, BMI, blood pressure, skin thickness, insulin, or age, there is still a baseline probability of having diabetes, albeit low. Each coefficient in the logistic regression model represents the change in the log-odds of diabetes for a one-unit increase in the corresponding independent variable, holding all other variables constant. For instance, the coefficient for pregnancies (0.22192395) indicates that for each additional pregnancy, the log-odds of diabetes increase by approximately 0.22 units, assuming other factors remain unchanged. Similarly, a coefficient of 1.125 for glucose implies that for every one-unit increase in glucose level, the log-odds of diabetes increase by approximately 1.125 units, suggesting a strong positive association between glucose levels and the likelihood of diabetes.

```
Intercept: [-0.87387862]
Coefficients: [[ 0.22192395  1.12487346 -0.16824905  0.01648627 -0.18588227  0.7313578
   0.21171528  0.39300256]]
```

VIII. Conclusion:

- The Logistic Regression model achieved satisfactory performance in predicting diabetes likelihood based on the provided features.
- Glucose, BMI, and blood pressure appear to be significant predictors of diabetes risk, as indicated by their coefficients in the logistic regression model.
- The ROC AUC score suggests that the model has a good ability to distinguish between diabetic and non-diabetic cases.
- The trained model and testing data have been saved into pickle files for potential deployment in production environments.

IX. Future Steps:

- Explore additional feature engineering techniques to improve model performance.
- Experiment with other machine learning algorithms to compare performance and robustness.
- Conduct further analysis to understand the relationship between features and diabetes risk in more depth.
- Deploy the trained model into a production environment for real-world use.

X. **Github Link** - https://github.com/ArchanaVInfy/DiabetesPrediction