Summary (Healthcare Diabetes Data Analysis)

Objective: The objective of this project is to predict the likelihood of a patient having diabetes using diagnostic data. The dataset, sourced from the **National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK)**, includes several medical predictor variables such as the number of pregnancies, BMI, insulin levels, age, and more, with the target variable being the diabetes outcome.

Key Steps involved in the Solutions:

1. **Data Preprocessing:**

    - Handled missing values and standardized the dataset to improve model accuracy.
    - Split the data into training and testing sets to facilitate model training and validation.

2. **Data Visualization and Exploration:**

    - Created histograms to understand the distribution of various features like glucose, insulin levels, BMI , and age.
    - Used scatter plot to analyze the relationships between pairs of variables, highlighting correlations.
    - Plotted the correlation heatmap to identify relationships among all the features.

3. **Model Building and Evaluation:**

- Applied several classification algorithms, including:
  i) **Logistic Regression**: Achieved an accuracy of 78% on the test set with an AUC score of 0.84.

  ii) **K-Nearest Neighbors (KNN):** Achieved an accuracy of 77% on the test set with improved performance after tuning.

  iii) **Support Vector Classification (SVC):** Showed good accuracy of 76% with consistent results across different kernel types.

  iv) **Random Forest Classifiers:** Performed well with an accuracy of 82% and showed robustness in prediction.

- Ensemble learning techniques like **AdaBoost** and **Random Forest** further enhanced the prediction capabilities.

- Leveraged tools like **LazyPredict** to automate and compare the performance of multiple models efficiently.

4. **Model Comparison:**

- Evaluated models based on metrics such as precision, recall, F1-score, and ROC-AUC.

- Compared model performance, with **Random Forest, SVC, and Logistic Regression** showing the highest accuracy scores.

- Identified models like **Random Forest** and **XGBoost** as top performers with the highest accuracy of over 81%.

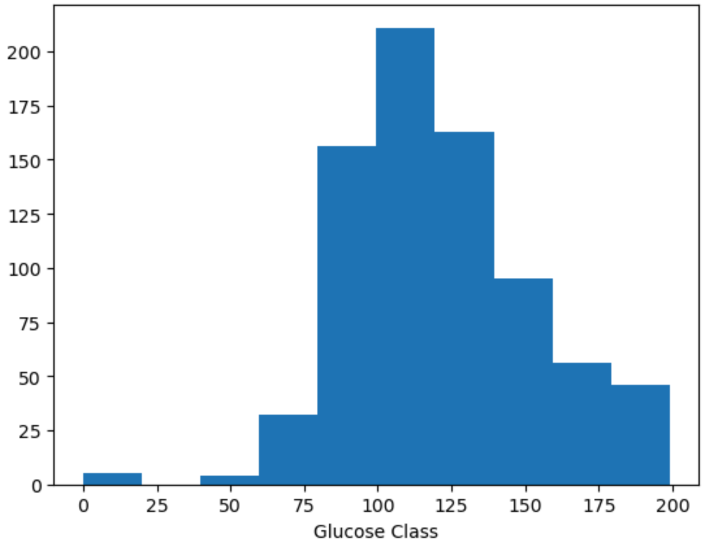
5. **Key Findings:**

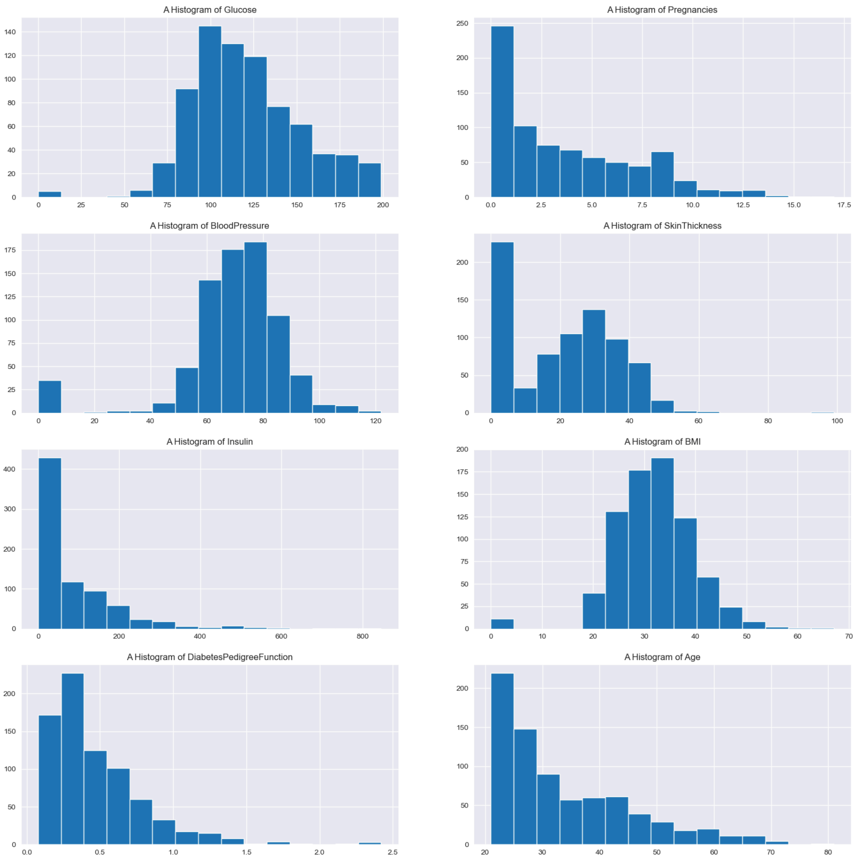
- The **Random Forest Classifier** and other ensemble methods demonstrated the most reliable performance.

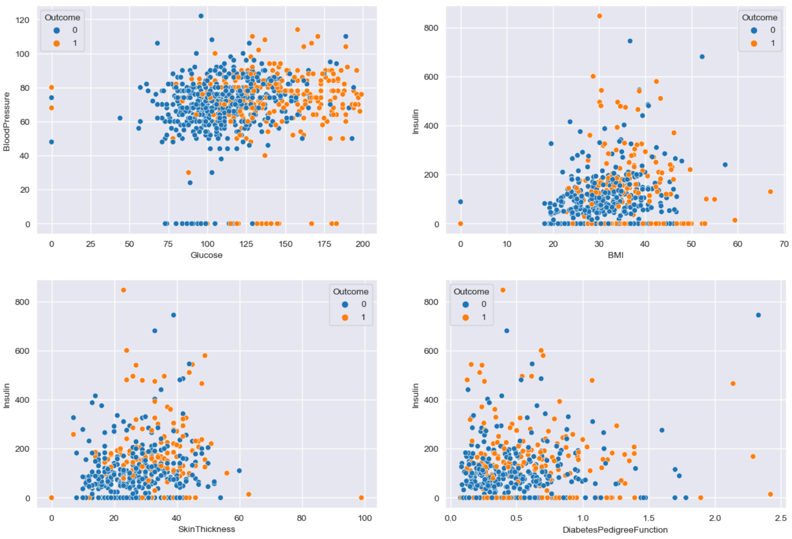
- Standardization of data led to improved accuracy in models such as **KNN** and **Logistic Regression.**

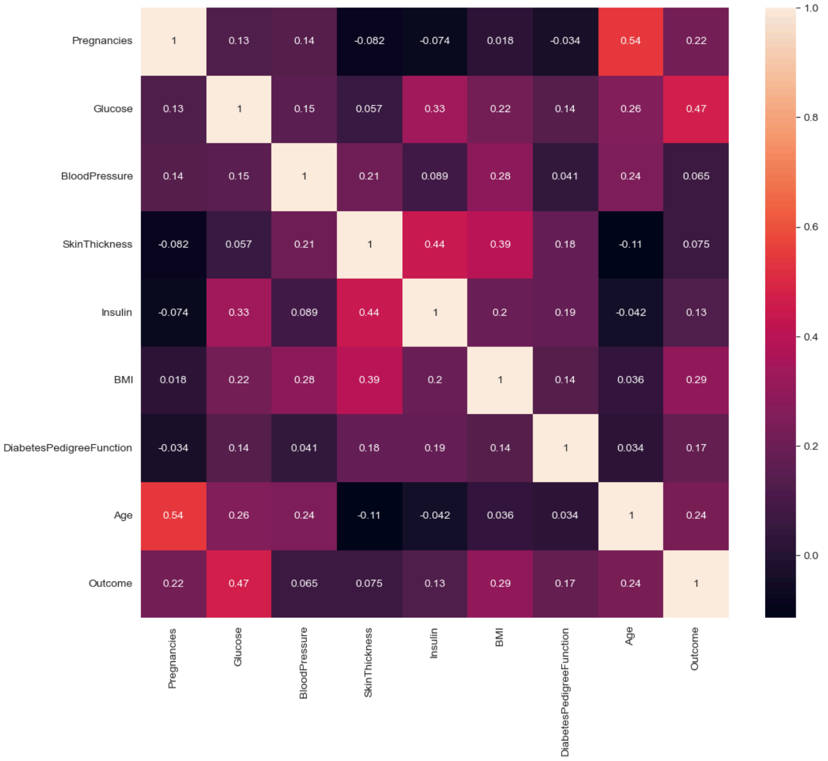
Conclusion: The project successfully developed a predictive model for diabetes diagnosis using various machine learning techniques. Ensemble methods like **Random Forest** achieved the highest accuracy. Data standardization and model comparisons highlighted key insights, making the solution robust and effective for identifying diabetic patients on medical diagnostic data.

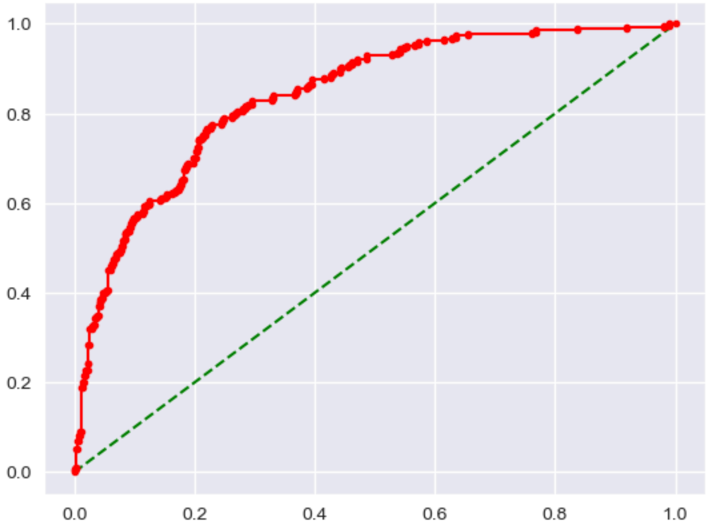
Visualization of Data(Output)

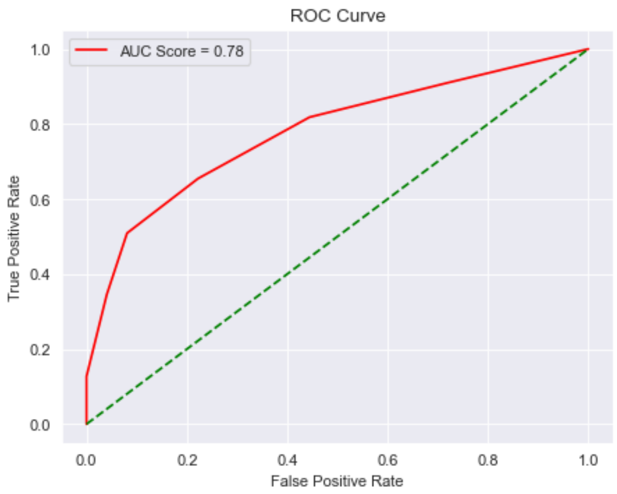




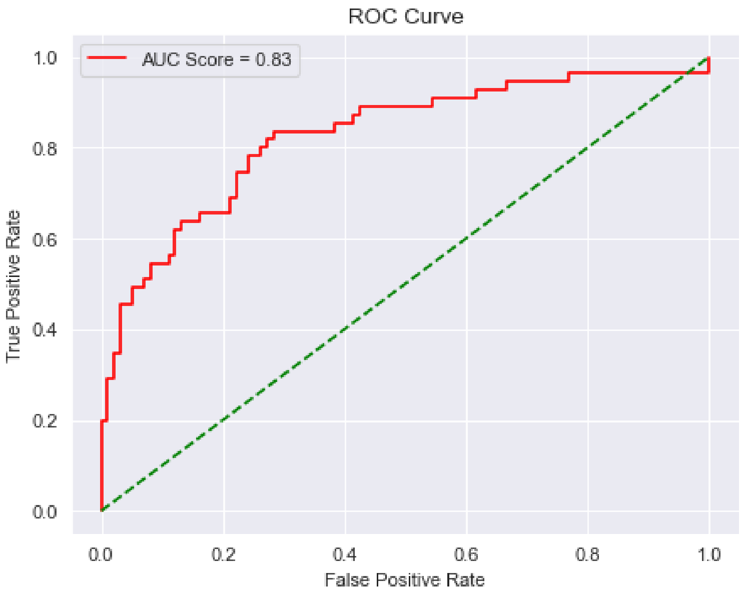


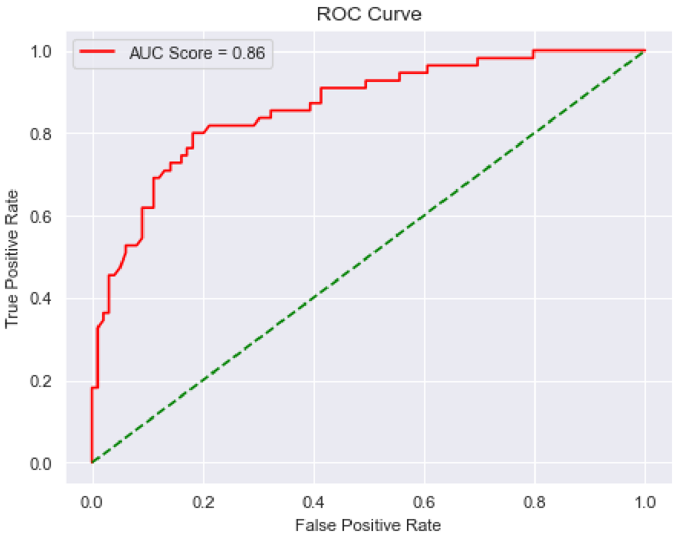






Logistic Regression **KNN Classification**





Support Vector Classifier

Ensemble Learning