Project Report

Objective

The goal of this project was to learn the fundamentals of a machine learning workflow by building a model to predict the onset of diabetes using a patient diagnostic dataset.

Methodology

1. Data Preprocessing

The Pima Indians Diabetes Dataset was loaded and cleaned to ensure model accuracy. Key preprocessing steps included:

- **Missing Value Imputation:** Zero values in columns like Glucose, BloodPressure, and BMI (which are biologically impossible) were treated as missing data and replaced with the median of their respective columns.
- **Feature Scaling:** All numerical features were normalized using MinMaxScaler to bring them into a consistent range of 0 to 1.

2. Exploratory Data Analysis (EDA)

A correlation heatmap was generated to visualize the relationships between all features and the target variable (Outcome). This analysis revealed that Glucose, BMI, and Age were the features most strongly correlated with a diabetes diagnosis.

3. Model Training & Evaluation

Two different machine learning models were trained on the preprocessed data and evaluated to determine the best performer.

- **Models:** Logistic Regression and Random Forest Classifier.
- **Evaluation Metrics:** Accuracy and Confusion Matrix were used to compare performance on a held-out test set.

Results & Analysis

Model Performance

The models' performance on the test data was as follows:

- **Logistic Regression:** Achieved an accuracy of **0.78** (78%).
- **Random Forest:** Achieved an accuracy of **0.74** (74%).

Confusion Matrix Insights

A deeper look at the confusion matrices provided critical insights into each model's predictive behavior.

- **False Negatives:** The Logistic Regression model had fewer False Negatives (22) compared to the Random Forest model (19), indicating it was more successful at correctly identifying patients who actually had the disease.
- **False Positives:** The Logistic Regression model also had fewer False Positives (12) compared to the Random Forest model (21), meaning it made fewer incorrect diagnoses of diabetes in healthy individuals.

Conclusion

Based on its higher overall accuracy and superior performance in minimizing critical errors such as false negatives, the **Logistic Regression** model is the better choice for this task. It provides a more reliable prediction and is therefore the preferred model for this patient dataset.