Diabetes Prediction Using Machine Learning

Objective

The primary objective of this project is to build a predictive model that determines whether an individual is likely to have diabetes based on various health-related input features. The project aims to:

- Analyze and preprocess the diabetes dataset.
- Train multiple machine learning models.
- Evaluate and compare model performance.
- Deploy the best-performing model using Streamlit to create an interactive web application for diabetes prediction.

Dataset Used

Dataset: *Diabetes.csv* (commonly known as the Pima Indians Diabetes Dataset)

Description:
This dataset contains records of patients along with their health measurements such as:

- Pregnancies
- Glucose levels
- Blood Pressure
- Skin Thickness

- Insulin levels
- BMI (Body Mass Index)
- Diabetes Pedigree Function (DPF)
- Age

Target Variable:

- **Outcome:** Indicates whether the patient has diabetes (1) or not (0).

The dataset is widely used for binary classification tasks in medical diagnosis.

Model Chosen

For this project, three different models were evaluated:

1. **Logistic Regression:**
   A linear model that predicts the probability of a binary outcome.
2. **Support Vector Machine (SVM) with a Linear Kernel:**
   An effective classification algorithm that finds the optimal hyperplane to separate classes in the feature space.
3. **Random Forest Classifier:**
   An ensemble method that builds multiple decision trees and merges their outputs for improved accuracy and robustness.

After evaluating all three models, the Support Vector Machine (SVM) with a linear kernel was selected based on its performance on the test set.

Performance Metrics

The primary performance metric used in this project is **Accuracy**. Accuracy is defined as the percentage of correct predictions made by the model on unseen data. Additional steps include:

- Splitting the dataset into training and testing subsets.
- Evaluating model performance on both the training set and the test set to check for overfitting.

Reported Accuracy:

- **Training Accuracy:** (e.g., 0.85 or 85%)
- **Test Accuracy:** (e.g., 0.78 or 78%)

Note: In practice, you might also consider additional metrics such as Precision, Recall, F1-score, and a confusion matrix for a more detailed evaluation, especially in a medical diagnosis context.

Challenges & Learnings

Challenges:

- **Data Preprocessing:**
  Handling missing values, scaling numerical features, and ensuring data quality was crucial.
- **Model Selection:**
  Experimenting with different machine learning models and tuning

their hyperparameters to avoid overfitting and underfitting posed a significant challenge.

- **Deployment Issues:**
  Integrating the trained model with a web application framework like Streamlit and managing file paths (e.g., loading images) required careful handling.

Learnings:

- **Importance of Data Scaling:**
  Standardizing the features was essential to ensure that models like SVM performed optimally.
- **Comparative Analysis:**
  Evaluating multiple models provided insights into the strengths and weaknesses of each algorithm.
- **Practical Deployment:**
  Building an interactive web app with Streamlit taught valuable lessons in making machine learning models accessible to end-users.
- **Iterative Improvement:**
  The project reinforced the concept that machine learning is an iterative process involving continuous evaluation and refinement.