Project: Diabetes classification Using Machine Learning

Project Overview:

Objective:

The goal of this project is to classify whether a patient has diabetes based on specific health metrics. Students are expected to analyze the dataset, choose appropriate preprocessing steps, apply classification models, and evaluate their performance.

Dataset:

- **Diabetes dataset**: The dataset consists of various medical predictor variables and a target variable (diabetes outcome: 0 or 1). The features include factors like glucose level, blood pressure, BMI, age, etc.
- You can provide the Pima Indians Diabetes Dataset as a resource or upload a similar dataset from Kaggle (be careful about features names).

Project Requirements:

1. **Data Preprocessing:**
   - Analyze the dataset to understand the distribution of classes.
   - Handle missing values, if any.
   - Encode categorical features appropriately.
2. **Feature Selection:**
   - Perform feature selection or dimensionality reduction to improve model performance.
3. **Model Selection:**
   - Experiment with at least three different machine learning algorithms (e.g., Decision Trees, Neural Networks, LR, …).
   - Justify your choice of algorithms based on theoretical knowledge and the characteristics of the dataset.
4. **Model Evaluation:**
   - Evaluate the performance of your models using metrics such as accuracy, precision, recall, F1-score, and confusion matrix.
   - Compare the performance of different models and provide a comprehensive analysis.
5. **Hyperparameter Tuning:**
   - Use techniques like Grid Search or Random Search to optimize the hyperparameters of your chosen models.

6. **Report:**
    - Document your approach, decisions, and results in a detailed report.
    - Include visualizations of your findings, such as feature importance, performance metrics, and decision boundaries.

Deliverables:

1. A Python notebook (or script) with your code and results.
2. A report (PDF) explaining your approach, decisions, results, and conclusions.

Evaluation Criteria:

1. **Data Understanding and Preprocessing (20%)**
2. **Algorithm Choice and Justification (30%)**
3. **Model Performance and Evaluation (20%)**
4. **Hyperparameter Tuning and Optimization (20%)**
5. **Quality of the Report and Visualizations (10%)**