Project 1: Predicting Diabetes Using Decision Trees and Gradient Boosting

Introduction

This project aimed to predict the onset of diabetes in individuals using machine learning techniques, specifically Decision Tree and Gradient Boosting classifiers. The dataset used for this task included variables such as pregnancies, glucose levels, blood pressure, skin thickness, insulin levels, BMI, diabetes pedigree function, age, and outcome (indicating the presence or absence of diabetes).

Data Exploration and Preprocessing

The initial step involved loading and exploring the dataset using Pandas. Key steps included checking for missing values, understanding data types, and summarizing the unique counts for each feature:

- The dataset was found to contain no missing values, and all features had appropriate data types (integers or floats).
- The feature distribution was analyzed to ensure data readiness for model training.

Model Building

1. Decision Tree Classifier

- **Splitting Data**: The data was split into training and testing sets using a 80-20 split, ensuring a robust evaluation.
- **Training**: A Decision Tree Classifier was instantiated with a random state for reproducibility and trained on the dataset.
- **Evaluation**: The model's performance was assessed using accuracy score, confusion matrix, and classification report. The decision tree achieved an accuracy of approximately 74.67%.
- **Visualization**: The tree structure was visualized using Matplotlib, which provided insights into decision-making at various nodes.

2. Gradient Boosting Classifier

- **Training**: Gradient Boosting, a more sophisticated ensemble method, was applied to enhance prediction accuracy.
- **Evaluation**: Similar to the Decision Tree, the Gradient Boosting model was evaluated using accuracy scores. It achieved an accuracy identical to the Decision Tree at 74.67%.

Results and Insights

Both models provided a clear understanding of feature importance and model accuracy. While the Decision Tree offered interpretability, Gradient Boosting suggested potential for better

performance with hyperparameter tuning. The accuracy levels indicate room for improvement, possibly through feature engineering or ensemble methods.

Conclusion

This project demonstrated the application of basic and advanced machine learning models to predict diabetes. While the initial results were promising, further work is needed to enhance model performance.