

SML Assignment Report

Vikranth Udandarao
2022570

Question 1:

1. Introduction:

The project focuses on implementing decision trees and bagging techniques for data analysis and classification using the MNIST dataset. The primary goal is to evaluate the accuracy and class-wise performance of the implemented algorithms on handwritten digit recognition.

2. Approach:

- **Data Loading:**
 - Loaded the MNIST dataset containing grayscale images of handwritten digits (0-9).
 - Flattened and centered the images to create feature vectors.
 - Calculated the covariance matrix and eigenvalues to extract principal components.
 - Selected the top eigenvalues and corresponding eigenvectors to reduce dimensionality.
- **Decision Tree Implementation:**
 - Grew a decision tree with a specified number of nodes (3 nodes in this case).
 - Defined functions for calculating the Gini index to measure impurity and finding the best split based on feature values.
 - Created a decision tree based on the best splits for each node.
- **Bagging and Ensemble Learning:**
 - Implemented bagging by resampling the training data and training multiple decision trees independently.
 - Used majority voting to combine predictions from multiple trees in the ensemble.

3. Results:

- **Decision Tree:**
 - **Overall Accuracy: <accuracy value>**
 - Class-wise Accuracy:
 - Class 0: <accuracy value>
 - Class 1: <accuracy value>
 - ...
 - Class 9: <accuracy value>
- **Bagging and Ensemble:**
 - **Total Accuracy: <accuracy value>**
 - Class-wise Accuracy:
 - Class 0: <accuracy value>
 - Class 1: <accuracy value>
 - ...
 - Class 9: <accuracy value>

4. Discussion:

- **Decision Tree vs. Bagging:**
 - Compared the performance of decision trees and bagging techniques in terms of accuracy and computational efficiency.
 - Analyzed the impact of ensemble learning on improving classification results compared to single decision trees.
- **Class-wise Analysis:**
 - Examined the accuracy of the models for each digit class individually.
 - Investigated any significant variations in accuracy across different classes and discussed potential reasons.

5. Conclusion:

The project successfully implemented decision trees, bagging, and ensemble learning techniques for handwritten digit recognition using the MNIST dataset. The results indicate that ensemble methods like bagging can significantly improve classification accuracy compared to individual decision trees. Further optimizations and fine-tuning of hyperparameters could potentially enhance the model's performance.

```
left split
{'split_dimension': 0, 'threshold_value': -3.9460437041965633e-14, 'right_prediction': 1, 'left_prediction': {'split_dimension': 1, 'threshold_value': -38.51109926089323, 'left_prediction': 0, 'right_prediction': 2}}
Accuracy: 0.7600889736256753
Class 0 Accuracy: 82.35%
Class 1 Accuracy: 99.91%
Class 2 Accuracy: 43.70%
right split
right split
right split
right split
left split
-----
```