**MACHINE LEARNING FROM DATA**

**Report: Lab Session 8 – Decision Trees and Random Forests**

**Names: Abdullah Qureshi, S M Rakib Hasan and Niina Hietamäki**

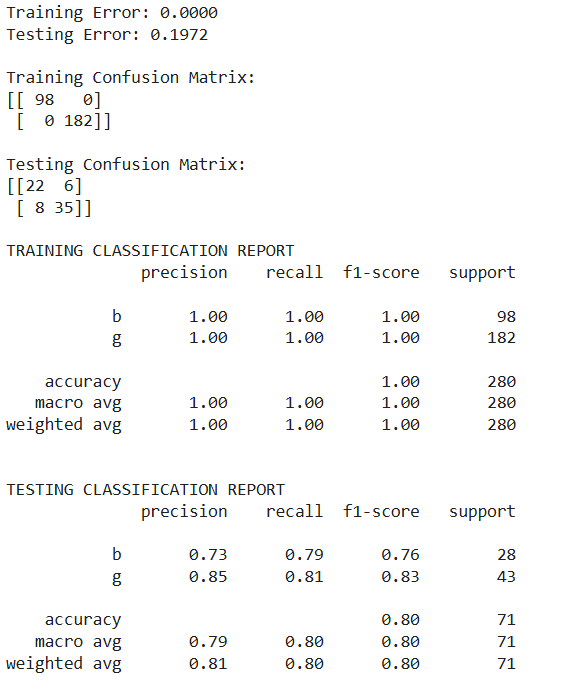
**Instructions**

* Download and uncompress the file Mlearn\_Lab8.zip
* Answer the questions in the document Mlearn\_Lab8\_report\_surname.pdf

**Questions**

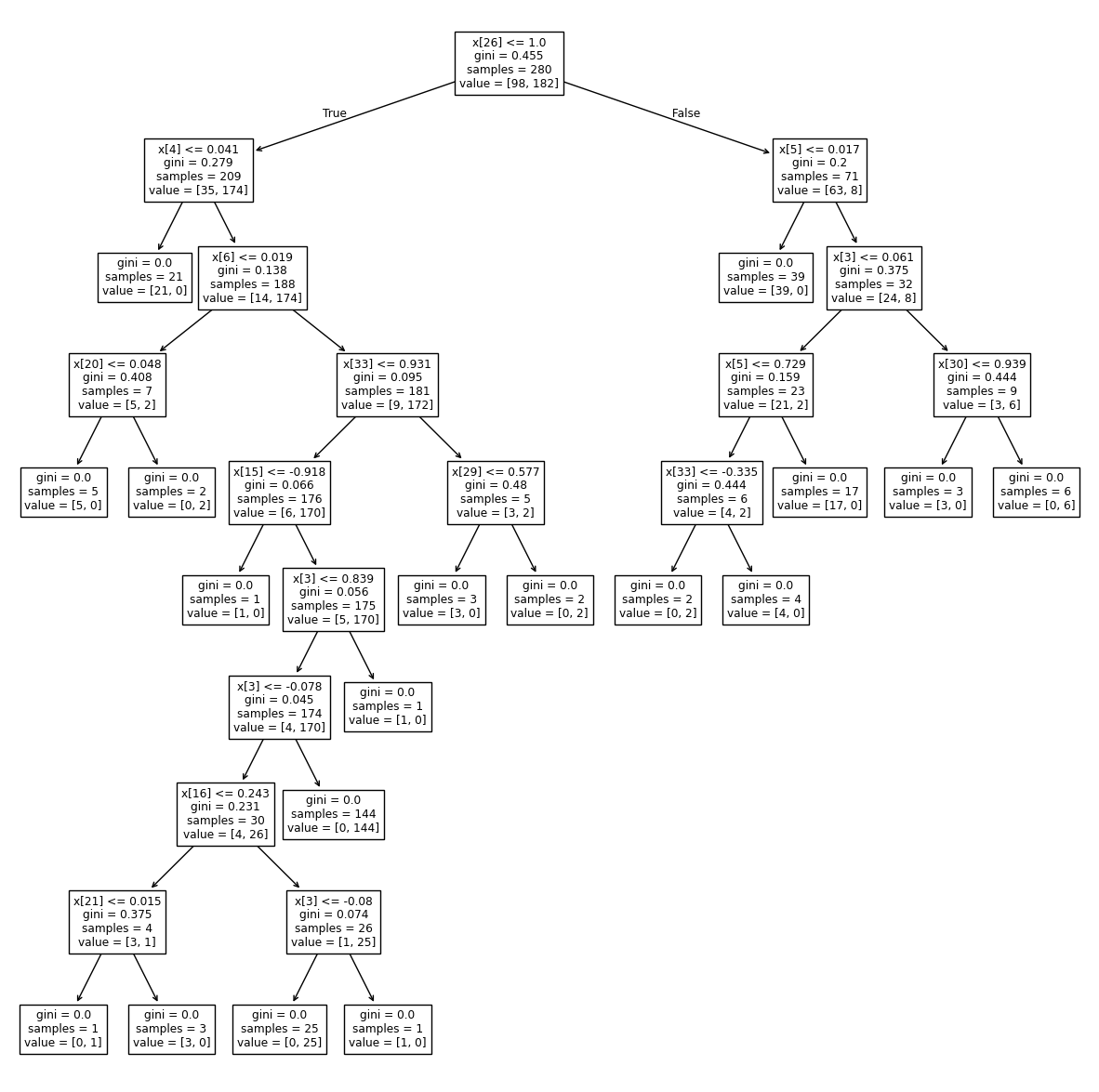
Default parameters:

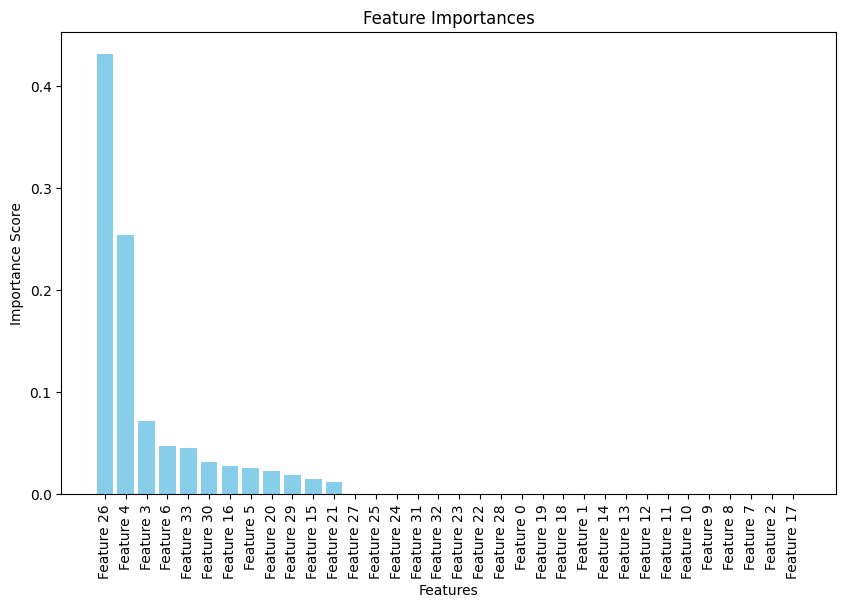
Q1: For the tree trained with the default parameters, copy the training, and test classification errors and the confusion matrices.



Feature importances:

Q2: Visualize the tree and analyze the questions at each node. Compute the feature-importances. Which are the most relevant features for the classification task?





Feature Importances:

Feature Importance

26 Feature 26 0.431125

4 Feature 4 0.254024

3 Feature 3 0.071051

6 Feature 6 0.046725

33 Feature 33 0.045375

30 Feature 30 0.031397

16 Feature 16 0.027553

5 Feature 5 0.024982

20 Feature 20 0.022427

29 Feature 29 0.018838

15 Feature 15 0.014730

21 Feature 21 0.011774

27 Feature 27 0.000000

25 Feature 25 0.000000

24 Feature 24 0.000000

31 Feature 31 0.000000

32 Feature 32 0.000000

23 Feature 23 0.000000

22 Feature 22 0.000000

28 Feature 28 0.000000

0 Feature 0 0.000000

19 Feature 19 0.000000

18 Feature 18 0.000000

1 Feature 1 0.000000

14 Feature 14 0.000000

13 Feature 13 0.000000

12 Feature 12 0.000000

11 Feature 11 0.000000

10 Feature 10 0.000000

9 Feature 9 0.000000

8 Feature 8 0.000000

7 Feature 7 0.000000

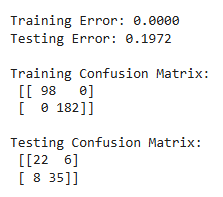
2 Feature 2 0.000000

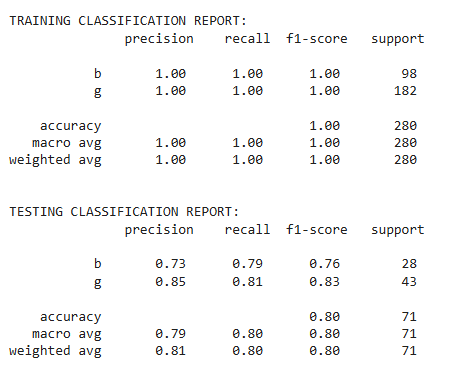
17 Feature 17 0.000000

We can see that Feature 26, Feature 4, and Feature 3 are the most important features in making decisions in the decision tree with scores of 0.4311, 0.2540, and 0.0711 respectively. They are being used in the first few levels of the tree

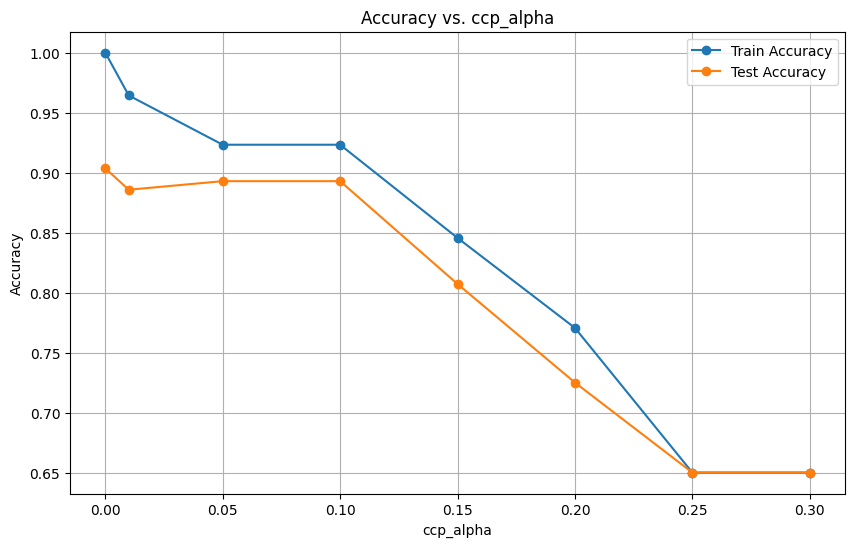
Regularization:

Q3: Copy and analyze the results (errors, confusion matrices and plots) and compare with the results obtained with the default parameters.

Regularization helped the model work better. On the training data, it got everything right (error = 0), which basically means it memorized the data. On the test data, the accuracy was 80 %, but the model still made some mistakes. It didn’t do as well with class b (precision 73 %, recall 79 %) compared to class g (precision 85 %, recall 81 %).

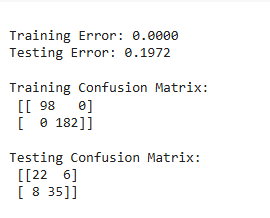
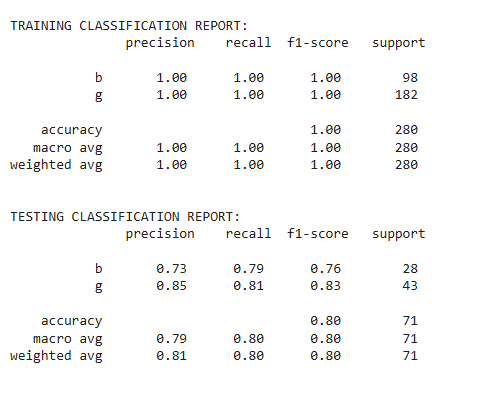
Cost-complexity pruning helped reduce overfitting. The model handled the test data better than when using the default settings. Pruning made the model simpler without losing much performance.

The plot shows that a very small ccp\_alpha leads to overfitting, while a very high value oversimplifies the model. The best test accuracy was achieved when ccp\_alpha was between 0.05 and 0.10, where the model maintained a good balance between simplicity and performance.

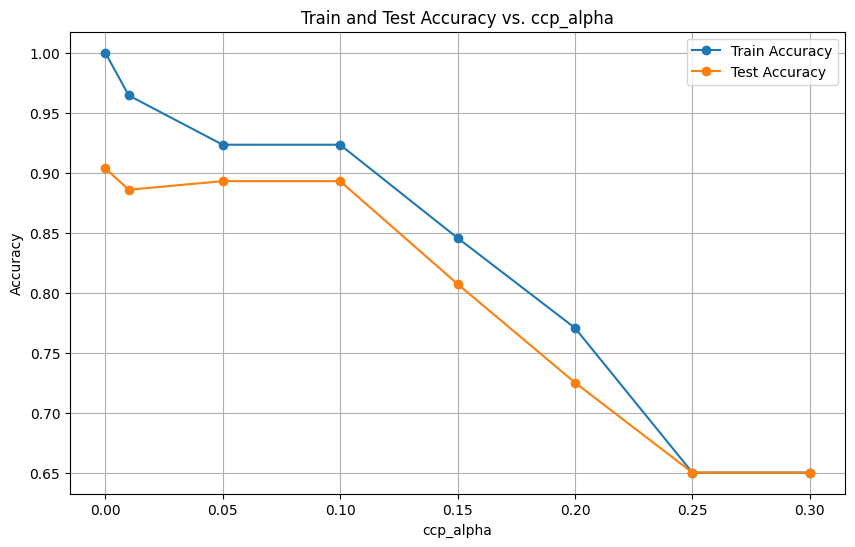


Tree pruning:

Q4: Copy and analyze the results (errors, confusion matrices and plots) and compare with the results obtained with the previous strategy (regularization).



Tree pruning didn’t change anything because the best ccp\_alpha was 0, so no pruning was applied. The test accuracy stayed at 80%, and training accuracy was 100%, showing no overfitting. Both pruning and regularization gave the same results, and the pruned and unpruned trees were identical.



Random Forests:

Q5: Copy and analyze the results (errors, confusion matrices and plots) and compare with the results obtained with decision trees.

Answer:

COMPARISON OF MODELS:

Decision Tree - Train Error: 0.0000, Test Error: 0.1972 (From Question 4) and 0.1690 (Newly done with all training samples)

Random Forest - Train Error: 0.0286, Test Error: 0.0986

**Training Error:**

Decision Tree: Train Error = 0.0000

The Decision Tree perfectly classifies the training data, which is a sign of overfitting. The model memorized the training set without generalizing well.

Random Forest: Train Error = 0.0286

The Random Forest model does not perfectly fit the training data due to its ensemble approach, which helps in regularizing the model and improving generalization.

**Test Error:**

Decision Tree: Test Error = 0.1972

The Decision Tree exhibits poor generalization on the test set. The high test error (19.72%) compared to its perfect training accuracy is a clear indication of overfitting.

Random Forest: Test Error = 0.0986

The Random Forest significantly improves test accuracy, with a test error of only 9.86%. This is almost half the test error of the Decision Tree, showcasing Random Forest's ability to handle variance and noise in the data.

So, the gap between train and test errors is much smaller for the Random Forest (around 7%) than for the Decision Tree (about 20%), highlighting that Random Forest generalizes better to unseen data.

Random Forests on MNIST:

Q6: Copy and analyze the results (errors, confusion matrices and plots) and compare with the results obtained in Lab6 using Neural Networks.

Answer:

Analyzing the results,

**1. Training Results:**

**Metrics**:

Precision, recall, and F1-score are all 1.00 for every class.

Overall accuracy is 1.00 (100%).

**Confusion Matrix:**

No errors were made during training. The diagonal of the matrix shows all predictions match the true labels.

So, the model performs perfectly on the training data, which could indicate potential overfitting. The model may have learned the training data too well, potentially at the cost of generalization to unseen data.

**2. Testing Results:**

**Metrics:**

Overall accuracy is 0.94 (94%), which is a good score but not as perfect as on the training data.

Precision, recall, and F1-scores are high across most classes, but some classes (e.g., Class 9) show slightly lower performance:

Class 9: Precision = 0.89, Recall = 0.91, F1-score = 0.90.

**Confusion Matrix:**

Most predictions are along the diagonal (correctly predicted labels), but there are minor misclassifications spread across several classes.

For example:

For Class 0, only 2 samples were misclassified.

For Class 9, more misclassifications are evident compared to other classes (e.g., samples being classified as Class 8).

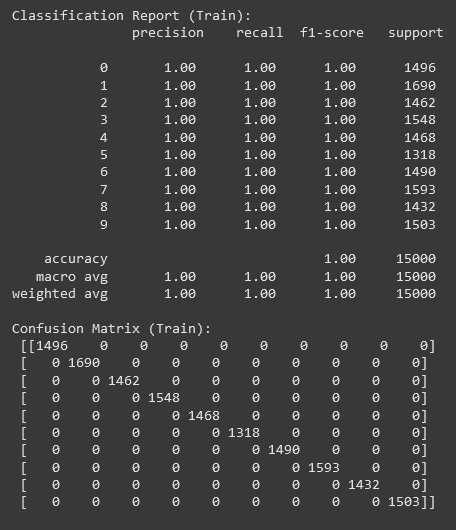
So, While the test results are strong, the drop in performance compared to the training set indicates overfitting. Some classes (like Class 9) are harder for the model to distinguish.

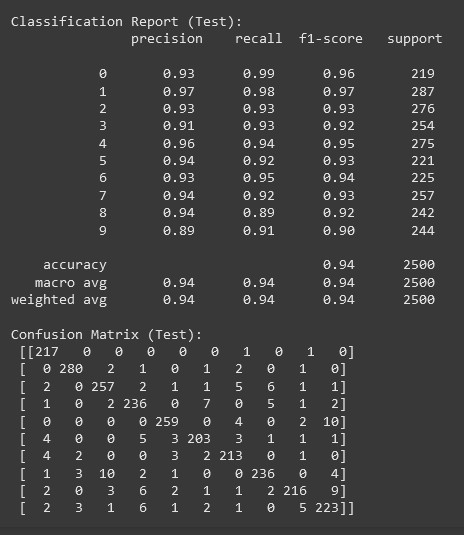
**3. Feature Importance Plot:**

The heatmap indicates the importance of different regions/features for the model's decision-making:

Bright regions (yellow) are the most important, showing where the model focuses to make predictions.

Darker regions (red to black) are less important.





Feature Importance-  
