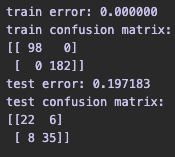
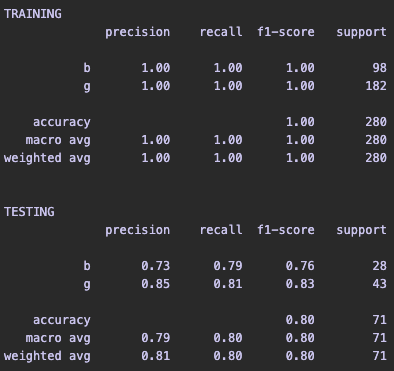
**MACHINE LEARNING FROM DATA**

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

**Questions**

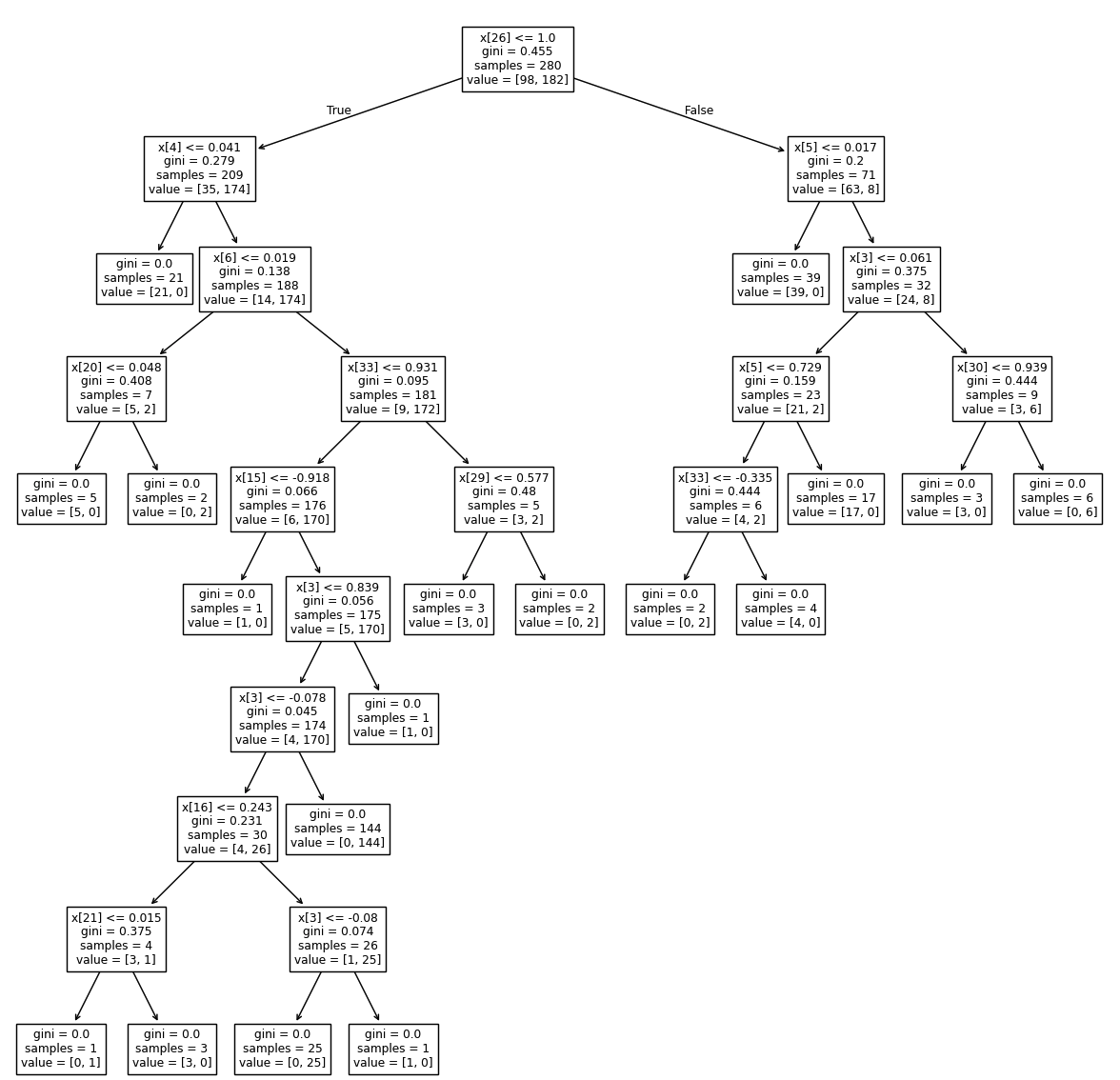
**Default parameters:**

**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 importance:  
array([0.02695519, 0. , 0.0770718 , 0.03016054, 0.15408633,

0.03119368, 0.08295619, 0.02198569, 0.01699943, 0.01532283,

0.00537967, 0.02733841, 0.01215537, 0.02396685, 0.01488081,

0.0182003 , 0.01238203, 0.03518953, 0.01047553, 0.02967175,

0.01397684, 0.0127529 , 0.00934536, 0.03358974, 0.00575936,

0.0071419 , 0.14342256, 0.02668298, 0.01931411, 0.0049968 ,

0.02066255, 0.01676339, 0.02481927, 0.01440031])

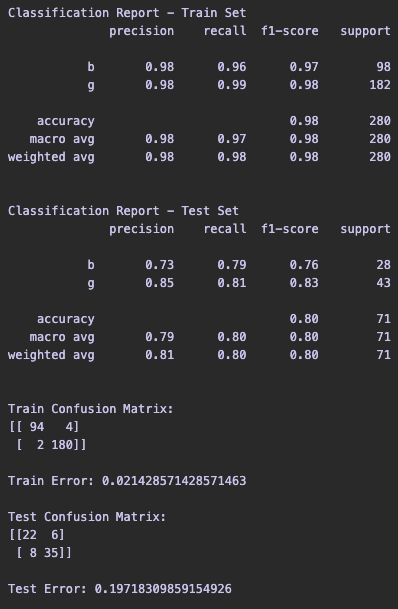
Most important feature is in index: 4

Most important feature importance is: 0.15408632749226353

The features with a higher value are the most important features, the ones with value = 0, are not contributing to the classification task.

**Regularization:**

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

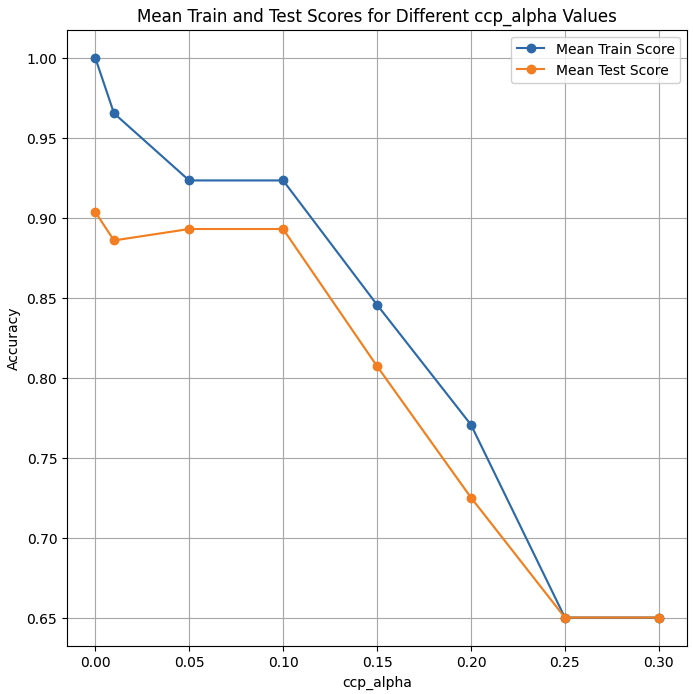
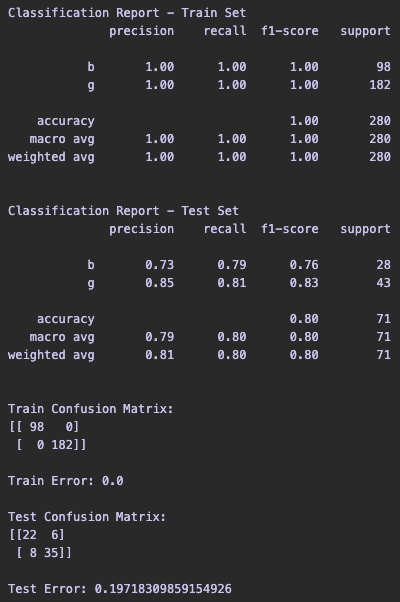


With the default parameters the model is clearly overfitting.

With grid search, the error for the train test is bigger, but the error for the test set is almost the same.

**Tree pruning:**

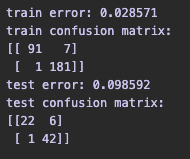
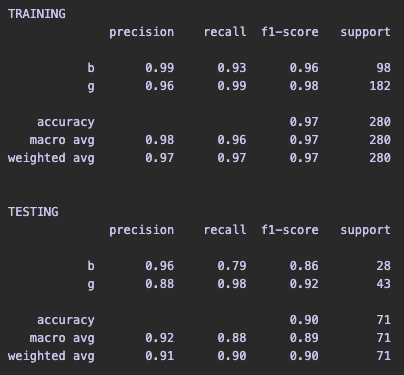
**Q4**: Copy and analyze the results (errors, confusion matrices and plots) and compare with the results obtained with the previous strategy (regularization).



The train error is 0 again, and the test error is almost the same as well. This is a clear case of overfitting.

**Random Forests:**

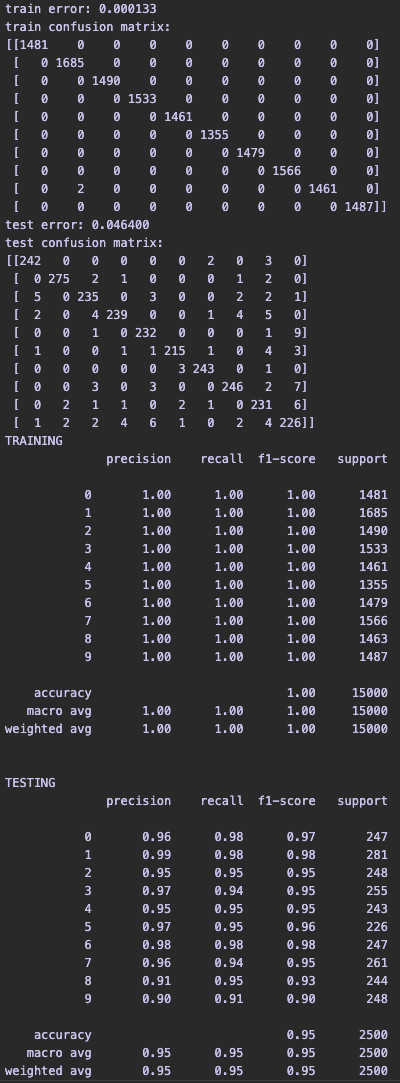
**Q5**: Copy and analyze the results (errors, confusion matrices and plots) and compare with the results obtained with decision trees.



The test error is lower than with the previous experiments. The train error is not 0, so we can assume that this model generalizes better and does not overfit the data.

**Random Forests on MNIST:**

Q6: Copy and analyze the results (errors, confusion matrices and plots) and compare with the results obtained in Lab7 using Neural Networks.



**LAST LAB:**  
train error: 0.12719999999999998

test error: 0.1644

Comparing with the model from last lab, the train and test error are much lower, so we can assume that the classifier works better for this problem.

