1. Prerequisites:

- **pip install sklearn matplotlib graphviz**

Before running the ipynb file, it is required to install three libraries via the above command.

- **import pandas as pd**
- **import os**

You need to import two libraries to execute cells in the notebook.

- **pip install ucimlrepo**

This command is important to download the dataset in Section 1 of the notebook.

2. Structure of folders:

- Charts: containing 4 images that visualize the distribution of labels for each (train/test) 40/60, 60/40, 80/20, and 90/10.
- ConfusionMatrix: containing 4 confusion matrixes of 40/60, 60/40, 80/20, and 90/10.
- ClassificationReport: containing 4 classification reports of 40/60, 60/40, 80/20, and 90/10 in csv format.
- Data: containing the original dataset (2 csv files: Features, Labels) and 4 folders (each folder contains 4 csv files: feature_train, feature_test, label_train, label_test) of 40/60, 60/40, 80/20, 90/10.
- Graph: containing images that visualize the resulting decision tree of 40/60, 60/40, 80/20, 90/10.
- MaxDepth: containing Accuracy_Max_Depth folder (table demonstrates the accuracy equivalent with max_depth) and images that visualize the resulting decision trees for each max_depth.

3. Preparing the data sets

- ***Comment***: For each set of different proportions, including (train/test) 40/60, 60/40, 80/20, and 90/10, the distributions of labels in all the data sets (the original set, training set, and test set) is nearly identical for each type in the labels. Therefore, the division of the train and test sets is appropriate. (figures in Charts folder)

4. Building the decision tree classifiers

- Use an instance of sklearn.tree.DecisionTreeClassifier to fix feature_train, label_train for each split dataset in Data folder. With information gain, "entropy” is assigned to criterion (parameter in DecisionTreeClassifier.

- For each set of different proportions, including (train/test) 40/60, 60/40, 80/20, and 90/10, the resulting decision trees using graphviz is saved in Graph folder.

5. Evaluating the decision tree classifiers

• 40/60

Comment:

- Label 1 has the best results from this model. Precision, recall and f1-score are so high. All instances that were predicted as 1 were actually label 1, and about 91.43% of the total actual label 1 instances were predicted correctly (3 labels are confused with label 2).
- Label 2 has a lower precision about 79.25% (3 confused labels 1, 8 confused labels 2). Its recall is high about 97.67%. However, there were some instances from other labels that were incorrectly predicted as label 2.
- Label 3 has a high precision about 95.45% but lower recall about 72.41%, indicating it's conservative in predicting label 3 instances.
- The confusion matrix shows that there are 8 instances where label 3 was incorrectly predicted as label 2. Also, there are 3 instances where label 1 was incorrectly predicted as label 2 and there are 1 instance where label 2 was incorrectly predicted as label 3.
- Overall, the accuracy of the model is about 88.79% , which means it made the correct prediction for 88.79% of the instances. However, the macro average F1 score is 88.46%, which is slightly lower than the accuracy, indicating that there might be an imbalance in the performance of the classifier across different labels. The weighted average F1 score is about 88.73%, which is slightly lower than the accuracy, suggesting that the classifier's performance is not uniformly good across all labels.

• 60/40

Comment:

- Label 1 has the best results from this model with a perfect precision score, high recall and F1 scores. All instances that were predicted as label 1 were actually label 1, and about 95.83% of the total actual label 1 instances were predicted correctly (1 label is confused with label 2).
- Label 2 has a lower precision about 85.29%, but its recall is perfect. All of the total actual label 2 instances were predicted correctly, but there were some instances from other labels that were incorrectly predicted as label 2 (1 confused label 1, 4 confused labels 3).
- Label 3 has a perfect precision but lower recall about 78.95%, indicating it's conservative in predicting label 3.
- The confusion matrix shows that there are 4 instances where label 3 was incorrectly predicted as label 2. Also there are 1 instance where label 1 was incorrectly predicted as label 2.

- Overall, the accuracy of the model is 93.06%, which means it made the correct prediction for 93.06% of the instances. However, the macro average F1 score is 92.72 which is slightly lower than the accuracy, indicating that there might be an imbalance in the performance of the classifier across different classes. The weighted average F1 score is 92.99, which is slightly lower than the accuracy, suggesting that the classifier's performance is not uniformly good across all classes.

- **80/20**

Comment:

- Label 1 has high precision score, perfect recall and F1 scores. About 85.71% instances were actually in label 1 (2 confused labels 2). While all of the total actual label 1 instances were predicted correctly.
- Label 2 has a high precision about 91.67% (only 1 confused label 3), but its recall is lower about 78.57% of the total actual Class 2 instances were predicted correctly.
- Class 3: Class 3 has a high precision and recall, indicating it's good at predicting Class 3. This means that about 90% of the total actual Class 3 instances were predicted correctly.
- The confusion matrix shows that there is 1 instance where label 2 was incorrectly predicted as label 3. Also, there are 2 instances where label 2 was incorrectly predicted as label 1.
- Overall, the accuracy of the model is about 88.89, which means it made the correct prediction for 88.89% of the instances. However, the macro average F1 score is 88.97, which is slightly higher than the accuracy, indicating that the performance of the classifier is balanced across different classes. The weighted average F1 score is 88.67, which is slightly lower than the accuracy, suggesting that the classifier's performance is not uniformly good across all classes.

- **90/10**

Comment:

- Class 1: has a high precision score about 85.71%, perfect recall and F1 scores about 92.31%. All of the total actual label 1 instances were predicted correctly.
- Class 2: Class 2 has a perfect precision, but its recall is lower about 85.71% of the total actual Class 2 instances were predicted correctly.
- Class 3: Class 3 has a lower precision about 80% and recall about 80%, indicating it's conservative in predicting label 3. About 80% of the total actual label 3 instances were predicted correctly.
- The confusion matrix shows that there is 1 instance where Class 3 was incorrectly predicted as Class 1. Also, there is 1 instance where Class 2 was incorrectly predicted as Class 3.
- Overall, the accuracy of the model is 88.89%, which means it made the correct prediction for 88.89% of the instances. However, the macro average F1 score is 88.21%, which is slightly lower than the accuracy, indicating that there might be an imbalance in the performance of the classifier across different classes. The weighted average F1 score is

88.89%, which is the same as the accuracy, suggesting that the classifier's performance is uniformly good across all classes.

- **Summary:**

Comment:

- Model 2 seems to perform the best overall, with the highest accuracy and macro average F1 score. However, each model has its strengths and weaknesses, and the best model to choose would depend on the specific requirements of your task.

6. The depth and accuracy of a decision tree:

Table:

Max_depth	None	2	3	4	5	6	7
Accuracy	88.89	91.67	94.44	88.89	88.89	88.89	94.44

Comment:

- The model achieves the highest accuracy of 94.44% at max_depth values of 3 and 7. This suggests that these depths allow the model to generalize well from the training data to unseen data. However, the complexity at 3 is lower than 7.
- The accuracy is 91.67% at max_depth of 2, which is slightly lower than the peak but still quite high.
- The accuracy drops to 88.89% for max_depth values of None, 4, 5, and 6. This could be due to overfitting or underfitting.

7. Grading:

No	Specifications	Scores (%)	Complete (%)
1	Preparing the data sets	30	30
2	Building the decision tree classifiers	20	20
3	Evaluating the decision tree classifiers		
	Classification report and confusion matrix	10	10
	Comments	10	10
4	The depth and accuracy of a decision tree		
	Trees, tables, and charts	20	20
	Comments	10	10