A brief description for each dataset (what is the task, what are the features and the target)

Dataset 1: ID: 1462 banknote-authentication

URL;
https://www.openml.org/search?type=data&sort=runs&status=active&qualities.NumberOfClasses=%3D_2&id=1462

Data were extracted from images that were taken from genuine and forged banknote-like specimens. For digitization, an industrial camera usually used for print inspection was used. The final images have 400x 400 pixels. Due to the object lens and distance to the investigated object gray-scale pictures with a resolution of about 660 dpi were gained. A Wavelet Transform tool was used to extract features from these images. (Source: https://www.openml.org/d/1462).

Task: This dataset is about distinguishing genuine and forged banknotes. The task is to model the probability that a banknote is fraudulent as its features function. The number of instances (rows) in the data set is 1372, and the number of features (columns) is 5.

The features are:

1. V1 (variance of Wavelet Transformed image (continuous)): variance finds out how each pixel varies from the neighboring pixels and classifies them into different regions;
2. V2 (skewness of Wavelet Transformed image (continuous)): skewness is the measure of the lack of symmetry;
3. V3 (kurtosis of Wavelet Transformed image (continuous)): kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution;
4. V4 (entropy of image (continuous)): image entropy is a quantity which is used to describe the amount of information which must be coded for, by a compression algorithm;

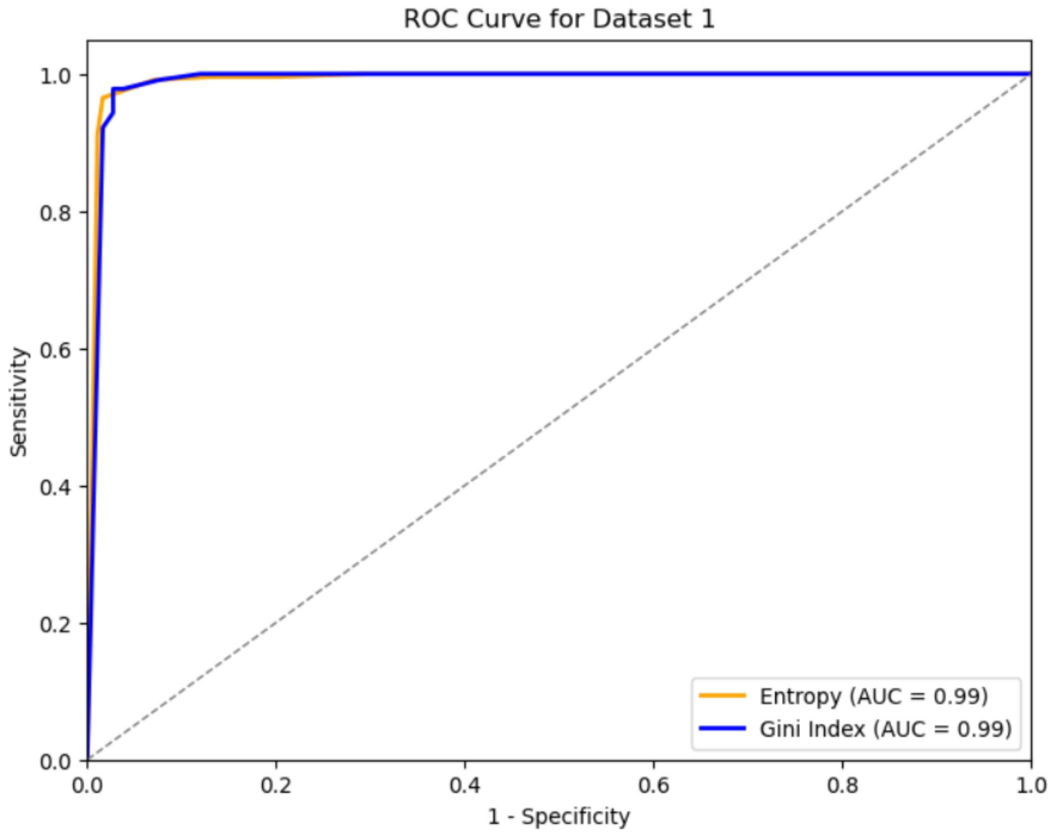
Class (**target**): 1 for genuine and 2 for forged

For each dataset, results in the form of a graph of ROC curves and a table of AUC values.

```
AUC Values (10 Folds) for Dataset 1
    Entropy  Gini Index
0  0.975428    0.965116
1  0.987275    0.977622
2  0.992541    0.993638
3  0.978499    0.987714
4  0.987714    0.974550
5  0.995393    0.986836
6  0.988372    0.987714
7  0.993607    0.973986
8  0.998016    1.000000
9  0.983466    0.985670

=================================================

Best Parameter For Entropy: {'min_samples_leaf': 20}
Accuracy For Entropy: 0.9880309771837489
Best Parameter For Gini: {'min_samples_leaf': 10}
Accuracy For Gini: 0.9832846080268196
```



```
AUC Values for Dataset 1
    Entropy  Gini Index
0  0.991624    0.989405
```

Discussion of the results and conclusions

The code employs pre-fit cross-validation within the grid search process to tune on min_samples_leaf parameter and select the best-performing model configuration before fitting the model to the complete training dataset. Then, the model's performance is evaluated on the hold-out test set using ROC curves and AUC values.

Result

The optimal hyper-parameter for min_sample_leaf determined via grid search is 20 for entropy and 10 for Gini, resulting in an accuracy of 98% for both. The values indicate the parameters that resulted in the best performance, as determined by the supplied scoring metric ('roc_auc') in cross-validation. The models were trained with the best hyperparameters and then tested on a separate hold-out test set. The assessment yielded satisfactory AUC values of 99% for both models. AUC values close to 1 indicate excellent discrimination between classes, implying that both models successfully distinguish between the positive and negative classes in the test data.

An effective classifier displays a ROC curve positioned around the top-left corner of the graph, demonstrating higher true positive rates and lower false positive rates at various thresholds. Both ROC curves exhibit excellent discrimination between true positive and false positive rates.

Conclusion

Both models demonstrate strong performance in distinguishing between classes and making precise predictions on new data, as evidenced by the accuracy and AUC values obtained.

The difference in the optimal min_samples_leaf hyperparameters between entropy and Gini index criteria shows the decision tree models' sensitivity to hyperparameter tuning and criterion selection.

The effectiveness of decision trees constructed utilizing both entropy and Gini index criteria indicates their ability to capture the underlying patterns and relationships in the dataset.

A brief description for each dataset (what is the task, what are the features and the target)

Dataset 2: ID: 40983 (Wilt Data Set)

URL; https://www.openml.org/search?type=data&sort=runs&status=active&qualities.NumberOfClasses=%3D_2&id=40983

This data set involved detecting diseased trees in Quickbird satellite images. The high-resolution QuickBird images of 165 different area were acquired in 27 August 2012 to detect the diseased trees. The QuickBird images contains four 2.4 m resolution MS bands; G, R, NIR and PAN band.

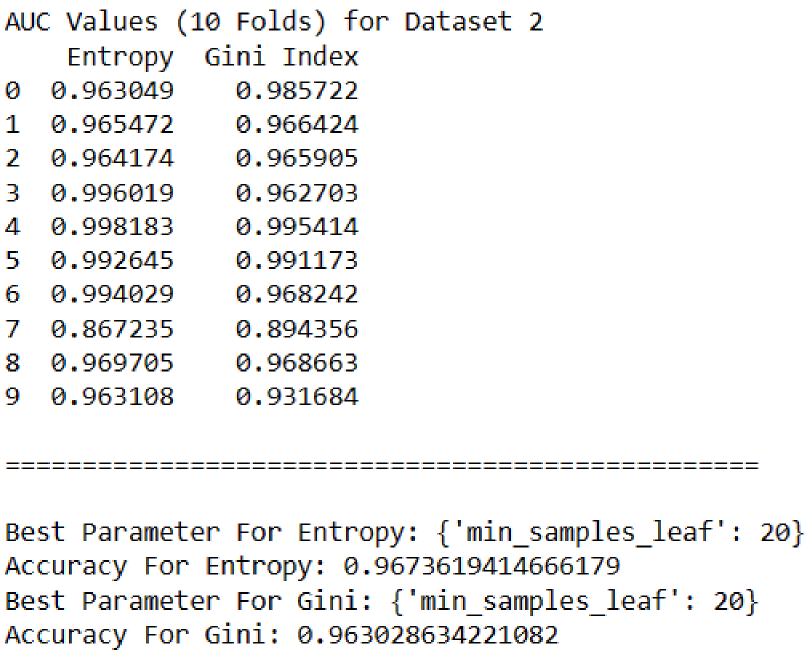
Task: The dataset is about the detection of Pine Wilt Disease (PWD) infected trees and non-affected. The number of instances (rows) in the data set is 4839, and the number of features (columns) is 6.

The features:

1. GLCM_Pan: GLCM mean texture (Pan band)
2. Mean_G: Mean green value
3. Mean_R: Mean red value
4. Mean_NIR: Mean NIR value
5. SD_Pan: Standard deviation (Pan band)

Class (**Target**): ‘infected’ correspond ‘1’, and ‘non-affected’ to 0

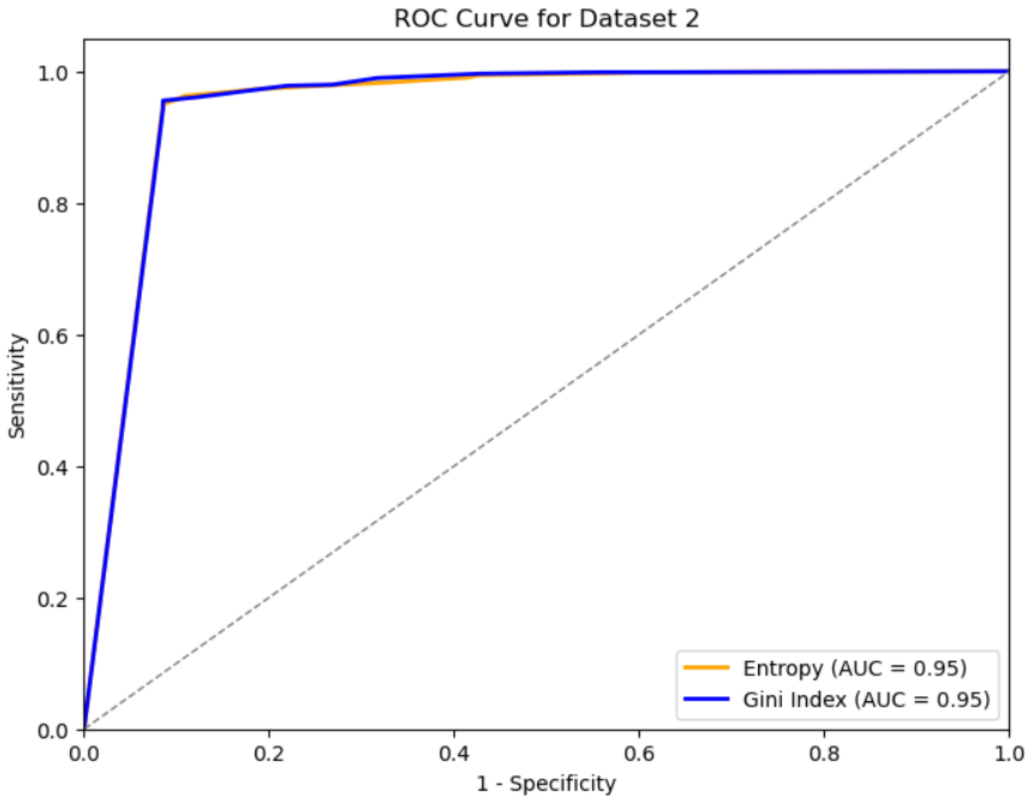
For each dataset, results in the form of a graph of ROC curves and a table of AUC values.

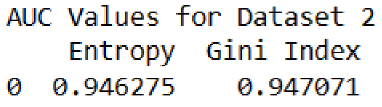


```
AUC Values (10 Folds) for Dataset 2
    Entropy  Gini Index
0  0.963049    0.985722
1  0.965472    0.966424
2  0.964174    0.965905
3  0.996019    0.962703
4  0.998183    0.995414
5  0.992645    0.991173
6  0.994029    0.968242
7  0.867235    0.894356
8  0.969705    0.968663
9  0.963108    0.931684

=================================================

Best Parameter For Entropy: {'min_samples_leaf': 20}
Accuracy For Entropy: 0.9673619414666179
Best Parameter For Gini: {'min_samples_leaf': 20}
Accuracy For Gini: 0.963028634221082
```





```
AUC Values for Dataset 2
    Entropy  Gini Index
0  0.946275    0.947071
```

Discussion of the results and conclusions

The code employs pre-fit cross-validation within the grid search process to tune on min_samples_leaf hyper-parameter and select the best-performing model configuration before fitting the model to the complete training dataset. Then, the model's performance is evaluated on the hold-out test set using ROC curves and AUC values.

Result & Conclusion

The best hyper-parameter for min_sample_leaf obtained using grid search is 20 for both entropy and Gini, resulting in an accuracy of 96% for both. These results are lower compaired to dataset 1. The AUC values of 94% is also lower. This shows that the classifiers constructed on Dataset 2 are less effective in distinguishing between the positive and negative groups compared to Dataset 1. Lower AUC values may imply that that the underlying patterns in the data are more complex or noisy.

The ROC curves for both entropy and Gini index criteria in Dataset 2 are stated as demonstrating some difference between true positive and false positive rates, although not as distinct as those in Dataset 1. This means that the classifiers' capacity to correctly categorize positive cases while limiting false positives is not as evident in Dataset 2 compared to Dataset 1. The underlying relationships between features and the target variable may be more obvious and evident in Dataset 1, leading to higher classifier performance. In contrast, Dataset 2 may bring some challenges, such as vagueness in feature significance, which could affect classifier performance negatively.