Name: Sushruth Danivasa Sridhar

Assignment – 1

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

   Dataset 1130 - OVA_Lung

   - Task: Binary classification - The key objective is to classify tissue samples using gene expression profiles. The aim is to leverage machine learning algorithms to accurately predict the type of tissue from gene expression data.

   - Features: There are 10936 numeric features in the dataset, each representing a unique gene expression level, identified by probe set IDs such as '1560622_at', '200699_at', etc.

   - Target: The target variable for classification is the tissue type, which is nominal and has two distinct values.

   Dataset 1134 - OVA_Kidney

   - Task: Binary classification - Dataset 1134 from GEMLeR is similarly used for the classification of tissues using gene expression data.

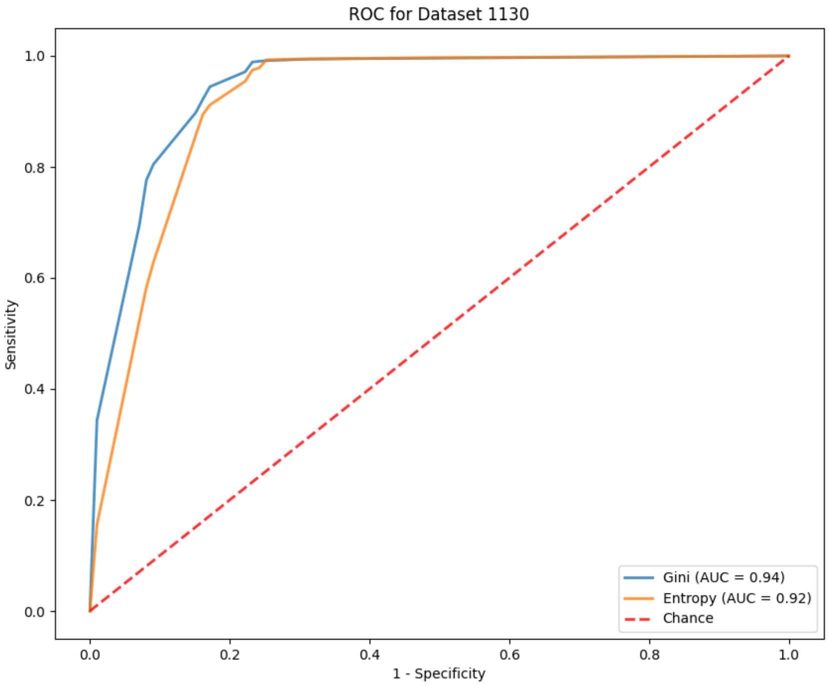
   - Features: There are 10936 numeric features in the dataset, each representing a unique gene expression level, identified by probe set IDs such as '1560622_at', '200699_at', etc.

   - Target: The target variable for classification is the tissue type, which is nominal and has two distinct values.

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

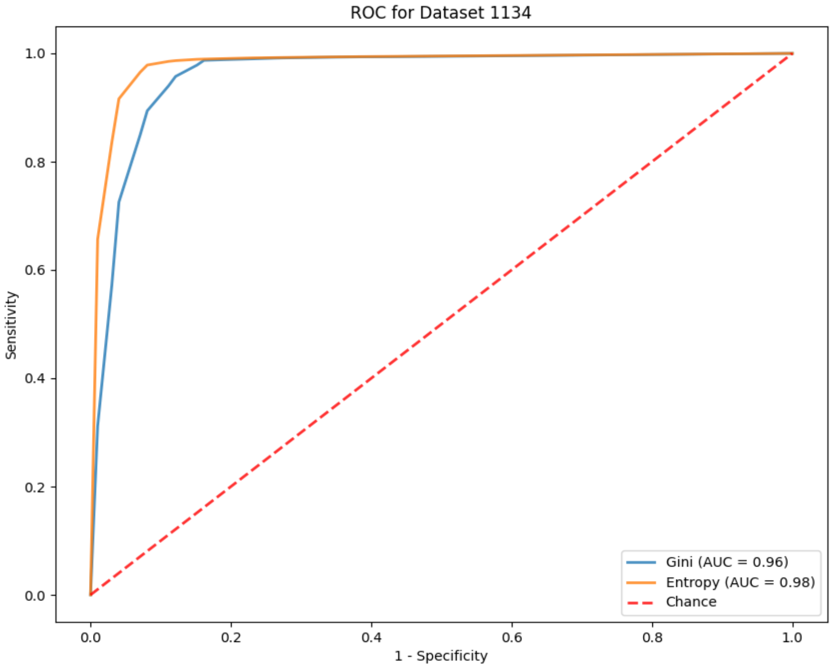
   Dataset 1130 - OVA_Lung:

Criterion	AUC Value
Gini	0.94
Entropy	0.92



Dataset 1134 - OVA_Kidney:

Criterion	AUC Value
Gini	0.96
Entropy	0.98



3. Discussion of the results and conclusions

From the ROC curves and AUC values, we can infer the following:

High AUC Values: Both dataset's OVA_Lung and OVA_Kidney show high AUC values for both criteria, indicating a strong performance of the decision tree classifiers in distinguishing between the two tissue types.

Criterion Comparison: The Entropy criterion appears to slightly outperform the Gini criterion, particularly in Dataset 1134, suggesting that Entropy may be more suited for this classification task.

Performance Indicators: The high AUC values close to 1.0 suggest that the classifiers have a high true positive rate and a low false positive rate, which is ideal for medical diagnosis and bioinformatics applications where false negatives and positives can have significant consequences.

In conclusion, the decision tree classifiers applied to the GEMLeR gene expression datasets have demonstrated excellent performance. The slight variation in AUC between the Gini and Entropy criteria could suggest that for specific datasets, one criterion may be more optimal than the other. This level of performance indicates the potential for applying these classifiers to real-world gene expression data for diagnostic and research purposes. However, it is important to validate these findings further with out-of-sample testing to ensure that the models generalize well to new data and to mitigate any potential overfitting.