# Assignment – 1 Report

## NEERAJA VUDHANTHI

The datasets used for this assignment are OVA_Lung and OVA_Endometrium. These datasets are a collection of gene expression datasets, related to predictive modeling or analysis in the context of lung and endometrial research. The aim is to predict the lung and the endometrium based on available data i.e. the features. The features are numeric and the target is nominal belonging to the binary classification category. The features for each of the datasets are the genes and gene characteristics which are 1000 in number. The most important gene of each dataset is as follows:

Genes associated with the endometrium:

1. HOXA10
2. ESR1 (Estrogen Receptor 1)
3. PGR (Progesterone Receptor)
4. IGF1 (Insulin-like Growth Factor 1)
5. LIF (Leukemia Inhibitory Factor)
6. VEGFA (Vascular Endothelial Growth Factor A)
7. MMPs (Matrix Metalloproteinases)
8. IL6 (Interleukin 6)
9. PTGS2 (Prostaglandin-Endoperoxide Synthase 2)
10. IGFBP1 (Insulin-like Growth Factor Binding Protein 1)

Genes associated with the lung:

1. SP-A (Surfactant Protein A)
2. SP-B (Surfactant Protein B)
3. SCGB1A1 (Secretoglobin Family 1A Member 1)
4. CFTR (Cystic Fibrosis Transmembrane Conductance Regulator)
5. ACE2 (Angiotensin-Converting Enzyme 2)
6. TTF1 (Thyroid Transcription Factor 1)
7. MUC5B (Mucin 5B, Oligomeric Mucus/Gel-Forming)
8. SFTPC (Surfactant Protein C)
9. HIF1A (Hypoxia Inducible Factor 1 Subunit Alpha)
10. NKX2-1 (NK2 Homeobox 1)

The target for the OVA_Lung is to find whether the tissue is a Lung or not, and the target for the OVA_Endometrium is to find whether the tissue is an Endometrium or not.

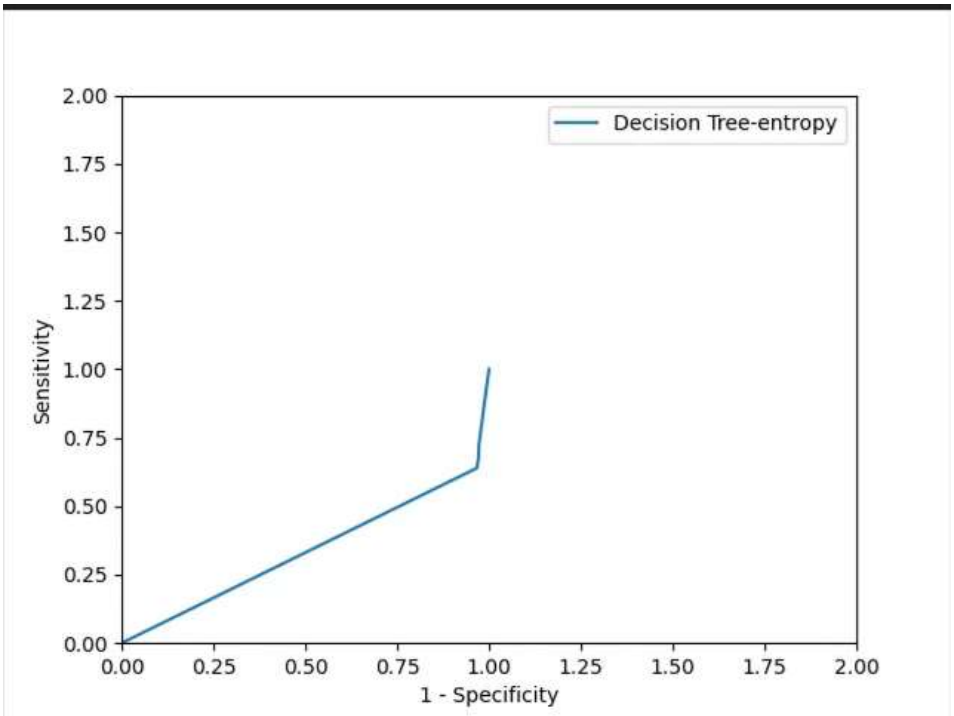


**Fig-1: ROC curve for the decision tree based on "entropy" on the OVA_Endometrium dataset**

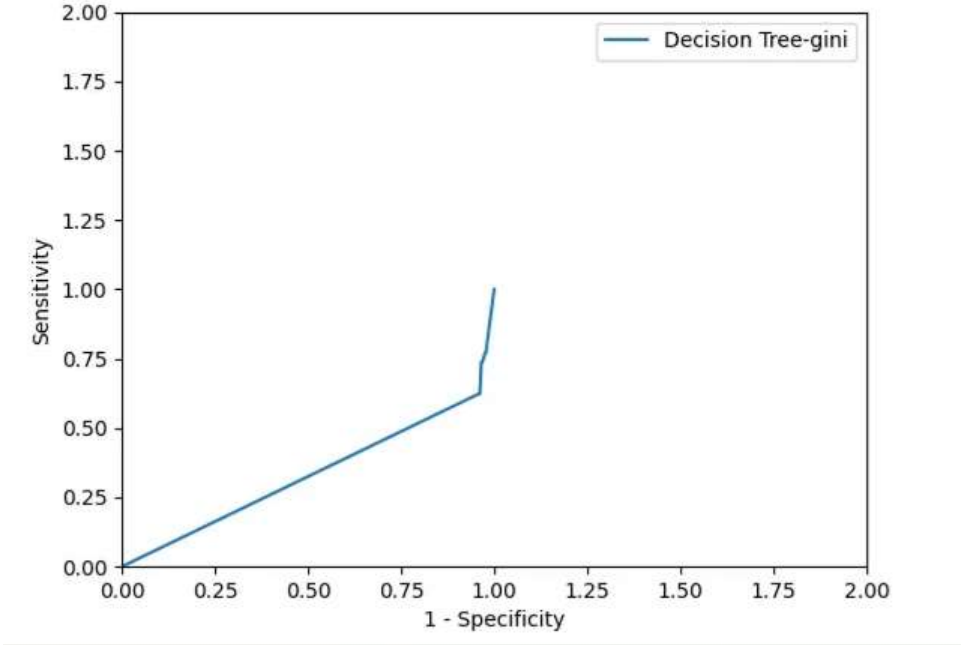


**Fig-2: ROC curve for the decision tree based on "gini-index" on the OVA_Endometrium dataset**

Based off of the ROC curves of the decision trees trained on OVA_Endometrium dataset, the entropy-based decision tree has a higher True Positive Rate and thus making it a model that correctly identifies positive instances among all actual positive instances in the dataset.

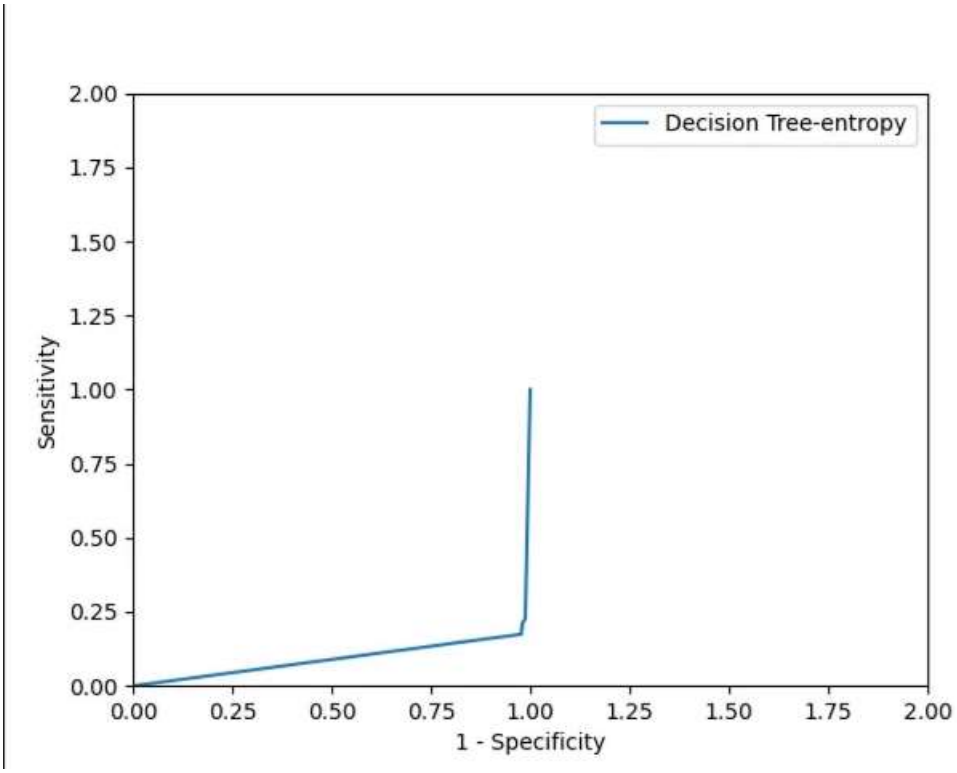


**Fig-3: ROC curve for the decision tree based on "entropy" on the OVA_Lung dataset**

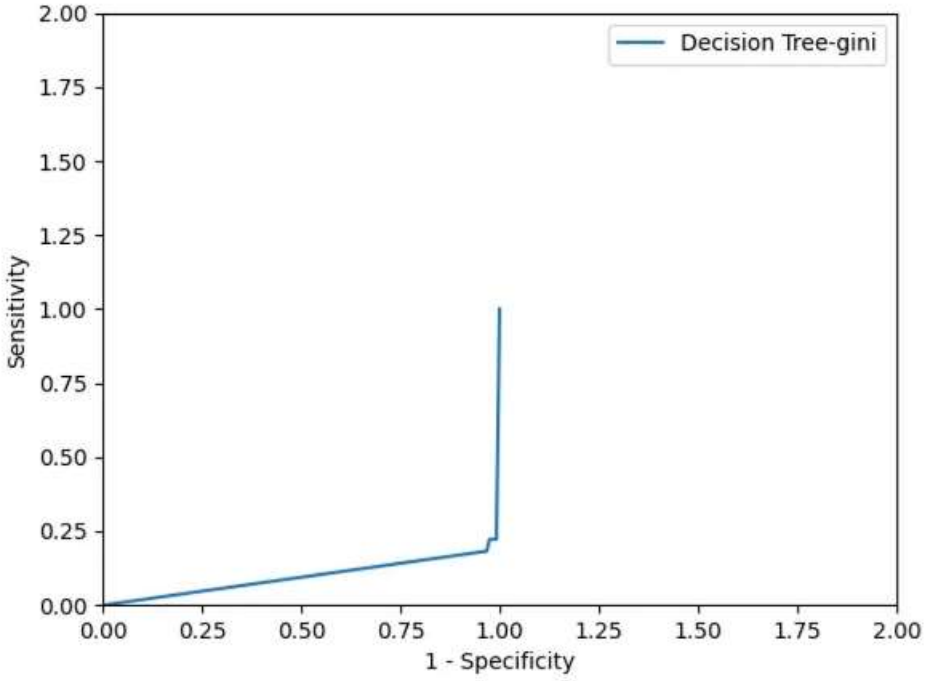


**Fig-4: ROC curve for the decision tree based on "gini-index" on the OVA_Lung dataset**

Based off of the ROC curves of the decision trees trained on OVA_Lung dataset, the entropy-based decision tree has a higher True Positive Rate and thus making it a model that correctly identifies positive instances among all actual positive instances in the dataset.

| Tree | Accuracy | F1-Score | Precision | Recall | AUC score |
|---|---|---|---|---|---|
| Entropy | 0.941747 | 0.100000 | 0.090909 | 0.111111 | 0.538888 |
| Gini-index | 0.932038 | 0.086956 | 0.071428 | 0.111111 | 0.533888 |

**Table-1: Metric values for both entropy-based and gini-index based decision trees trained on OVA_Endometrium dataset.**

**1. Accuracy:** This tells us how often the model correctly predicts whether something belongs to a certain category. In this case, the decision tree with the "Entropy" method was right about 94.17% of the time, while the one using "Gini-index" was right about 93.20% of the time. So, the "Entropy" model has a slightly better overall accuracy.

**2. F1-Score:** This is a way to combine both precision and recall into a single measure. It helps us understand how well the model balances between correctly identifying positive cases and not mislabeling negative cases. Here, both models have quite low F1-scores: 10.00% for "Entropy" and 8.70% for "Gini-index". This suggests that both models struggle with either finding true positive cases, avoiding false positives, or both.

**3. Precision:** Precision tells us how many of the instances predicted as positive are actually correct. For the "Entropy" model, it's about 9.09%, and for the "Gini-index" model, it's about 7.14%. This means that both models tend to label things as positive incorrectly quite often.

**4. Recall:** Recall measures how many of the actual positive instances the model managed to find. For both models, this is around 11.11%. This indicates that both models miss a significant portion of the actual positive instances.

**5. AUC Score:** The Area Under the ROC Curve (AUC) is a way to measure how well the model can distinguish between different classes. Both models have AUC scores slightly above 50%, indicating that they perform better than random guessing but not by much. This suggests that the models have limited ability to differentiate between the classes.

| Tree | Accuracy | F1-Score | Precision | Recall | AUC score |
|---|---|---|---|---|---|
| Entropy | 0.977346 | 0.867924 | 0.821428 | 0.920000 | 0.951197 |
| Gini-index | 0.980582 | 0.880000 | 0.880000 | 0.880000 | 0.934718 |

**Table-2: Metric values for both entropy-based and gini-index based decision trees trained on OVA_Lung dataset.**

**1. Accuracy:** This tells us how often the model gets its predictions right. For the "Entropy" model, it's about 97.73%, while for the "Gini-index" model, it's slightly higher at 98.06%. Both models seem to be quite accurate, with the "Gini-index" model having a slight edge.

**2. F1-Score:** This is a combined measure of precision and recall. It gives us an idea of how well the model balances between correctly identifying positive cases and not mislabeling negative cases. For the "Entropy" model, it's around 86.79%, and for the "Gini-index" model, it's slightly higher at 88.00%. These are both quite high, indicating that both models perform well in terms of finding positive cases and avoiding false positives.

**3. Precision:** Precision tells us how many of the instances predicted as positive are actually correct. For the "Entropy" model, it's about 82.14%, and for the "Gini-index" model, it's the same at 88.00%. This means that both models are quite good at correctly labelling positive cases.

**4. Recall:** Recall measures how many of the actual positive instances the model managed to find. For the "Entropy" model, it's about 92.00%, and for the "Gini-index" model, it's also 88.00%. This suggests that the "Entropy" model is slightly better at capturing all the positive instances, while the "Gini-index" model might miss a few.

**5. AUC Score:** The Area Under the ROC Curve (AUC) measures the model's ability to distinguish between different classes. For the "Entropy" model, it's about 95.12%, and for the "Gini-index" model, it's slightly lower at 93.47%. Both models have high AUC scores, indicating that they perform well in differentiating between classes.

**The best parameters for entropy-based and gini-index-based decision trees trained on OVA_Endometrium are given as:**

entropy best parameters: {‘min_samples_leaf’: 10}

gini best parameters: {‘min_samples_leaf’: 8}

- The best settings for making decisions about the "OVA_Endometrium" dataset show that the decision tree using entropy likes to have at least 10 samples in each group, while the one using Gini index prefers 8.

**The best parameters for entropy-based and gini-index-based decision trees trained on OVA_Lung are given as:**

entropy best parameters: {‘min_samples_leaf’: 6}

gini best parameters: {‘min_samples_leaf’: 6}

- The best settings for making decisions about the "OVA_Lung" dataset show that the decision tree using entropy likes to have at least 6 samples in each group, along with the one using Gini index preferring 6 as well.