I. Description of two Datasets

Dataset 1: gina 41158

- **Task:** The dataset originates from the USPS handwritten digit database and aims to develop a model to identify handwritten digits from images accurately. This task is commonly known as digit recognition and falls under the umbrella of classification problems in machine learning.
- **Features:** Each sample in the dataset represents a grayscale image of a handwritten digit, which has been standardized to a size of 16x16 pixels. The features of the dataset correspond to the pixel values of these images. In other words, each feature represents the intensity of a pixel in the image, resulting in 256 features per sample.
- **Target:** This dataset's target variable represents the numeric value of the handwritten digit depicted in each image. It is a categorical variable with values ranging from '0' to '9', where each value corresponds to a specific digit.

Dataset 2: USPS 41964

- **Task:** This dataset is also derived from the USPS handwritten digit database and shares the objective of recognizing handwritten digits from images. The task focuses on developing a model to classify these images into their respective digit categories accurately.
- **Features:** Each instance in this dataset consists of a grayscale image of a handwritten digit standardized to a size of 16x16 pixels. Like the previous dataset, the features represent the pixel values of these images. Each feature reflects the intensity of a pixel, resulting in 256 features per sample.
- **Target:** The target variable in this dataset serves the same purpose as in the previous dataset. It indicates the true numeric value of the handwritten digit depicted in each image, with values ranging from '0' to '9', representing the digits in the images.

Both datasets involve the same fundamental task of digit recognition from grayscale images of handwritten digits. The features represent pixel intensities, and the target variable denotes the actual numeric values of the handwritten digits in the images. The goal is to train machine learning models to effectively classify these images based on their pixel values and accurately predict the corresponding digits.

II. Results

A. AUC TABLES:

Dataset	Criterion	AUC Score
gina 41158	Entropy	0.9196
gina 41158	Gini Index	0.9197
USPS 41964	Entropy	0.9656
USPS 41964	Gini Index	0.9669

B. ROC CURVE:

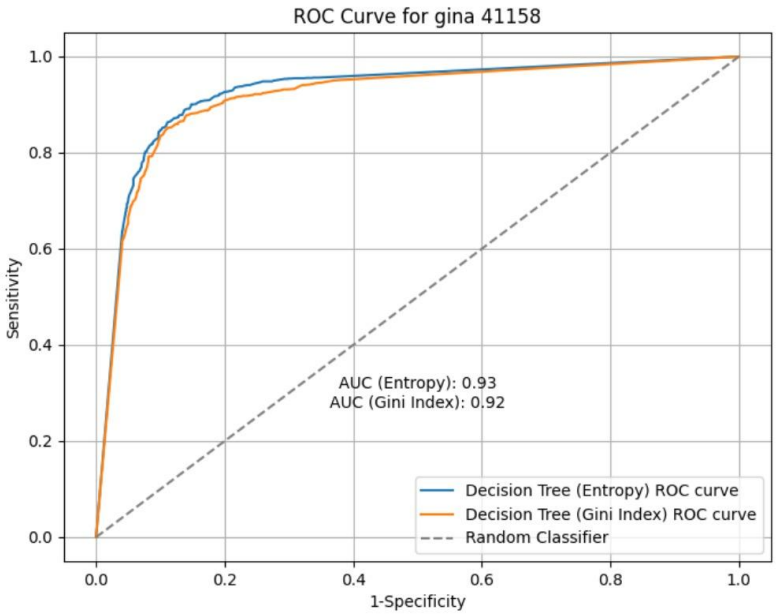

Fig 1: DATASET 1 - GINA 41158

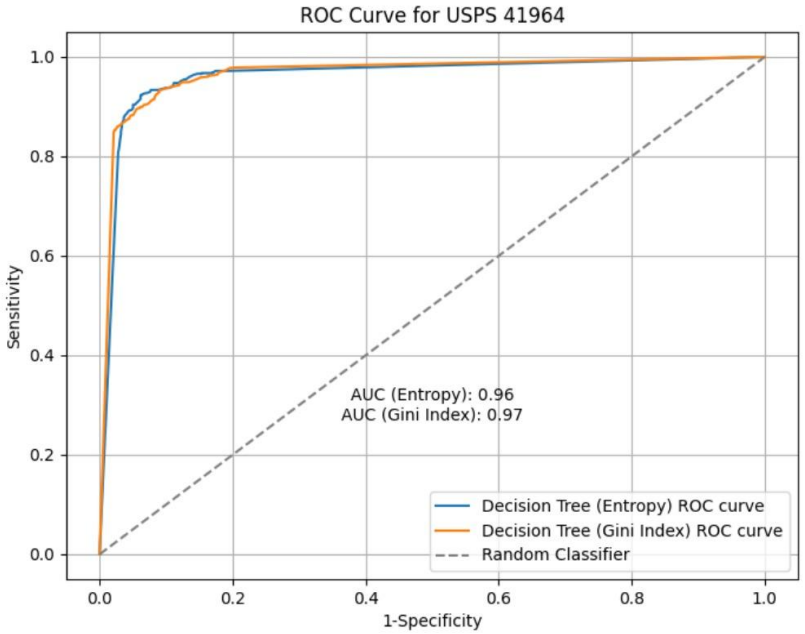

Fig 2: DATASET 2 - USPS 41964

III. Discussion and Conclusion

Dataset 1: gina 41158

- Best Parameter for Entropy: {'min_samples_leaf': 18}
- Best Parameter for Gini Index: {'min_samples_leaf': 15}
- AUC Score for Entropy: 0.9256
- AUC Score for Gini Index: 0.9162
- Cross-validation AUC scores (Entropy): [0.9127, 0.8941, 0.9218, 0.8991, 0.9377, 0.9413, 0.9409, 0.9209, 0.9404, 0.9366]
- Cross-validation AUC scores (Gini Index): [0.9141, 0.9056, 0.8972, 0.8905, 0.9262, 0.9279, 0.9304, 0.9258, 0.9329, 0.9314]

The ROC curve in Figure 1, Dataset 1 - gina 41158, illustrates a classification model's performance across different thresholds. The x-axis represents the false positive rate (FPR), while the y-axis represents the true positive rate (TPR). Analysis indicates that the Gini index-based decision tree outperforms its entropy-based counterpart, with AUC values of 0.93 and 0.92, respectively. The ROC curve closely resembles the ideal classifier curve, suggesting proficient classification. Both classifiers exhibit high AUC values, indicating effective separation of positive and negative cases. However, the Gini index classifier slightly outperforms the entropy-based classifier, emphasizing its superior performance in delineating positive and negative cases. Overall, both classifiers demonstrate proficient ability in distinguishing between positive and negative cases, with the Gini index classifier showing a slight performance edge over the entropy-based classifier.

Dataset 2: USPS 41964

- Best Parameter for Entropy: {'min_samples_leaf': 9}
- Best Parameter for Gini Index: {'min_samples_leaf': 9}
- AUC Score for Entropy: 0.9618
- AUC Score for Gini Index: 0.9667

- Cross-validation AUC scores (Entropy): [0.9696, 0.9636, 0.9682, 0.9355, 0.9827, 0.9628, 1.000, 0.9694, 0.9598, 0.9520]
- Cross-validation AUC scores (Gini Index): [0.9536, 0.9773, 0.9674, 0.9580, 0.9844, 0.9716, 0.9877, 0.9605, 0.9560, 0.9533]

The analysis of ROC curves indicates that the decision tree model using the Gini index outperforms the entropy-based model, with AUC values of 0.97 and 0.96, respectively. This suggests superior performance in distinguishing between positive and negative cases. Conversely, the ROC curve for the random classifier indicates random guessing. Overall, both curves are smooth and increasing, which is generally a good sign for a classifier. However, the Gini index curve is slightly steeper, particularly at lower FPR values. This suggests that the Gini index model may better classify positive cases when there are few false positives. In conclusion, the Gini index proves more effective as a splitting criterion for the provided datasets.

Comments and Conclusions:

- **Best Parameters:** For both datasets, the best parameter for the Gini index criterion is the same ('min_samples_leaf': 9), while for the entropy criterion, the best parameter differs ('min_samples_leaf': 18 for gina 41158 and 'min_samples_leaf': 9 for USPS 41964). This suggests that the optimal minimum number of samples required to be at a leaf node in the decision tree varies between the two datasets, potentially due to differences in the complexity or distribution of the data.
- **AUC Scores:** The AUC scores indicate the performance of the decision tree classifiers on each dataset. In general, higher AUC scores indicate better classifier performance. The AUC scores for both criteria (entropy and Gini index) are higher for the USPS 41964 dataset than the gina 41158 dataset. This suggests that the decision tree classifiers perform better on the USPS 41964 dataset in distinguishing between the classes.
- **Cross-validation AUC Scores:** The cross-validation AUC scores provide additional insight into the robustness of the models. Overall, the cross-validation AUC scores are consistent with the AUC scores obtained from the initial model evaluation. The USPS 41964 dataset generally exhibits higher cross-validation AUC scores than the gina 41158 dataset, indicating better generalization performance of the models trained on USPS 41964.

  Based on the provided information, the decision tree classifiers perform better on the USPS 41964 dataset than the gina 41158 dataset, as evidenced by higher AUC scores and more consistent cross-validation AUC scores. Additionally, the optimal hyperparameters for the classifiers vary between the two datasets, suggesting differences in the data characteristics that affect model performance.