MACHINE LEARNING
Assignment 2

Golla Meghanandh Manvith Prabhash
22CS30027

Part 1 :- Decision Tree

1. USING ENTROPY :

1.1. diabetes.csv

1.1.1. Before Pruning

Accuracy : 0.7467532467532467
Precision : 0.7240740740740741
Recall : 0.7222222222222222

1.1.2. After Pruning

Accuracy : 0.7662337662337663
Precision : 0.7541346973572037
Recall : 0.7737373737373738

1.2. diabetes_noise.csv

1.2.1. Before Pruning

Accuracy: 0.4810810810810811
Precision : 0.4593166175024582
Recall : 0.4600434572670208

1.2.2. After Pruning

Accuracy : 0.5783783783783784
Precision : 0.5660196501317997
Recall : 0.5665137614678899

Observations :

The decision tree model exhibited improved performance after pruning on the noiseless dataset, with increases in accuracy, precision, and recall. Conversely, the noisy dataset showed a significant drop in performance, but pruning still led to improvements across all metrics, although it remained inferior to the noiseless dataset's performance.

Performance Comparison:

- **Pre-pruning**: The accuracy of the noisy dataset plummeted from 0.7468 to 0.4811, indicating difficulties in class distinction due to noise.
- **Post-pruning**: Pruning recovered some performance metrics, with accuracy and precision increasing by approximately 10% and 11%, respectively, yet still falling short compared to the noiseless dataset.

For **Bonus Part**:

2. USING GINI IMPURITY

2.1. diabetes.csv

2.1.1. Before Pruning

Accuracy : 0.7727272727272727

Precision : 0.7551100628930818

Recall : 0.7383838383838384

2.1.2. After Pruning

Accuracy : 0.7727272727272727

Precision : 0.7528735632183908

Recall : 0.7585858585858586

2.2. diabetes_noise.csv

2.2.1. Before Pruning

Accuracy: 0.5783783783783784

Precision : 0.5486552920763447

Recall : 0.540620473201352

2.2.2. After Pruning

Accuracy : 0.6162162162162163

Precision : 0.6019447162426614

Recall : 0.6006156446161275

Comparison of Entropy and Gini Impurity Methods :

1. Overall Accuracy:

   Gini impurity generally yields **higher accuracy** than **entropy** across both datasets. Entropy shows a more significant improvement post-pruning in noiseless conditions but does not surpass Gini's performance in noisy conditions.

2. Precision:

   Gini impurity consistently achieves **higher precision**, especially in **noisy datasets**, indicating better performance in minimising false positives.

3. Recall:

   Entropy outperforms Gini in noiseless datasets regarding recall, making it preferable when identifying true positives is critical. However, Gini demonstrates higher recall in noisy datasets, showcasing its robustness against noise.

Conclusion :

- **Entropy** tends to provide better recall in noiseless environments, making it suitable when identifying true positives is crucial.

- **Gini Impurity**, however, offers overall better accuracy and precision, particularly in noisy datasets, making it more robust when handling data with noise.

In **summary**, while **entropy** may be better suited for **noiseless environments** due to its higher recall, **Gini impurity** is more effective in **noisy conditions**, providing superior accuracy and precision. Thus, for datasets with noise, Gini impurity is the recommended choice over entropy.