Assignment_2_part_1 for data without noise(Decision tree)

Introduction :

In this report, we explore the performance of a decision tree classifier on a noiseless dataset before and after pruning. Decision trees are known for their ability to model complex decision boundaries, but they often suffer from overfitting. Pruning is a technique used to simplify the tree, reduce overfitting, and improve generalisation to new data. This report details the structure and accuracy of the model before and after applying pruning.

Dataset :

The dataset used in this analysis is cardiovascular disease dataset without any noise. It contains 11 features and labels and was split into training, validation, and test sets:

- Training set: 64% of the dataset
- Validation set: 16% of the dataset
- Test set: 20% of the dataset

The training set was used to build the decision tree, while the validation set was used to prune it. The final performance was measured on the test set.

Methodology :

1. **Decision Tree Construction:**

We implemented a custom decision tree from scratch. The model was initially trained on the training dataset. The tree was constructed using a greedy algorithm based on minimising entropy, resulting in a large and potentially overfitted tree.

The tree was visualised before pruning as shown in Figure 1 below.

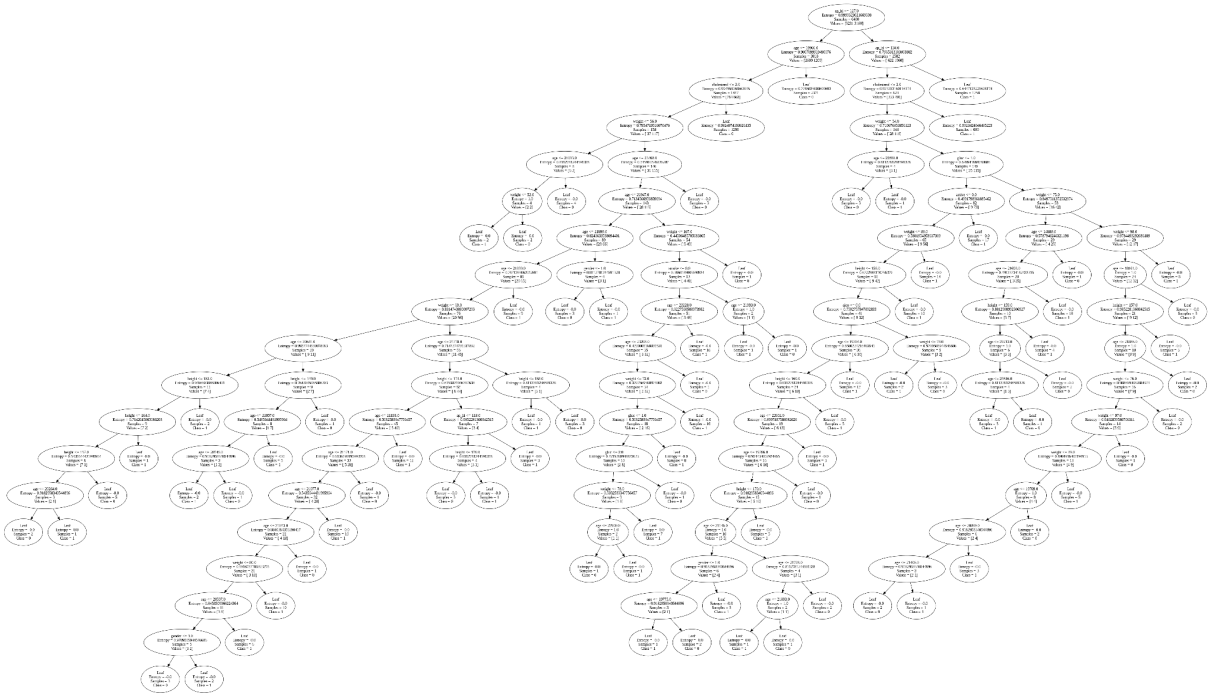


Figure 1: Decision Tree Before Pruning

Accuracy before pruning: `0.7032105506422204`

2. Pruning:

After constructing the tree, we applied pruning using a post-pruning technique. The validation set was used to assess the performance of subtrees. Branches that did not significantly contribute to accuracy on the validation set were pruned, leading to a simpler and more generalizable tree.

The tree structure after pruning is shown in Figure 2 below.

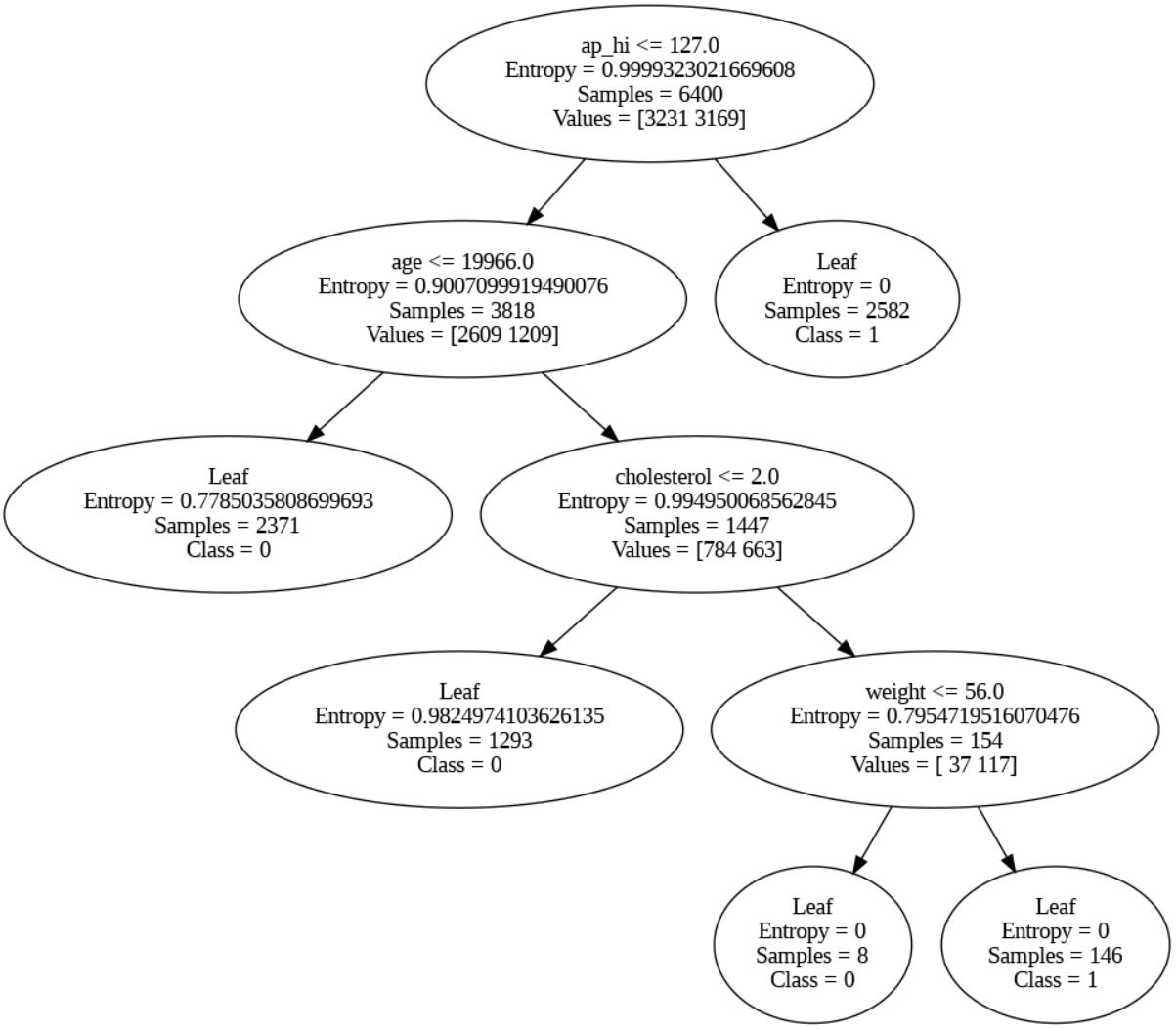


Figure 2: Decision Tree After Pruning

Accuracy after pruning: `0.7345`

As shown in the results, the accuracy before pruning was slightly higher on the training set, as the model had overfit the data. After pruning, the accuracy decreased marginally, but the pruned model is expected to perform better on unseen data due to reduced overfitting. Impact on Generalization.

Pruning aims to improve generalisation by reducing the complexity of the decision tree. Although the accuracy on the training set may decrease after pruning, the pruned model should perform better on the test set or any unseen data because it avoids overfitting the training data. The generalisation capability of the model was improved after pruning, as reflected in its simplified structure.

Conclusion :

This analysis demonstrates the importance of pruning in decision tree models. While the unpruned tree had higher accuracy on the training set, it was likely overfitted to the noise and specific patterns of the training data. By pruning the tree, we reduced its complexity and improved its ability to generalise to unseen data. In real-world applications, this approach can lead to more robust models.

Pruning is a critical step in ensuring that decision tree models do not become overly complex, thus balancing the trade-off between accuracy and interpretability.

Assignment_2_part_1 for noisy data(Decision tree)

Introduction :

In this report, we analyse the performance of a decision tree classifier on a noisy dataset before and after pruning. Noise in data can significantly impact the performance of machine learning models, especially decision trees, which are prone to overfitting when the dataset contains irrelevant or misleading information. This report outlines the effect of noise on the model and how pruning helps mitigate its impact.

Dataset :

The dataset used in this analysis is cardiovascular disease which includes artificially introduced noise. Noise refers to random or irrelevant data points that distort the patterns and relationships in the dataset. The dataset was split into training, validation, and test sets in the same proportions as in the noiseless scenario:

- **Training set**: 64% of the dataset
- **Validation set**: 16% of the dataset
- **Test set**: 20% of the dataset

Nature of Noise

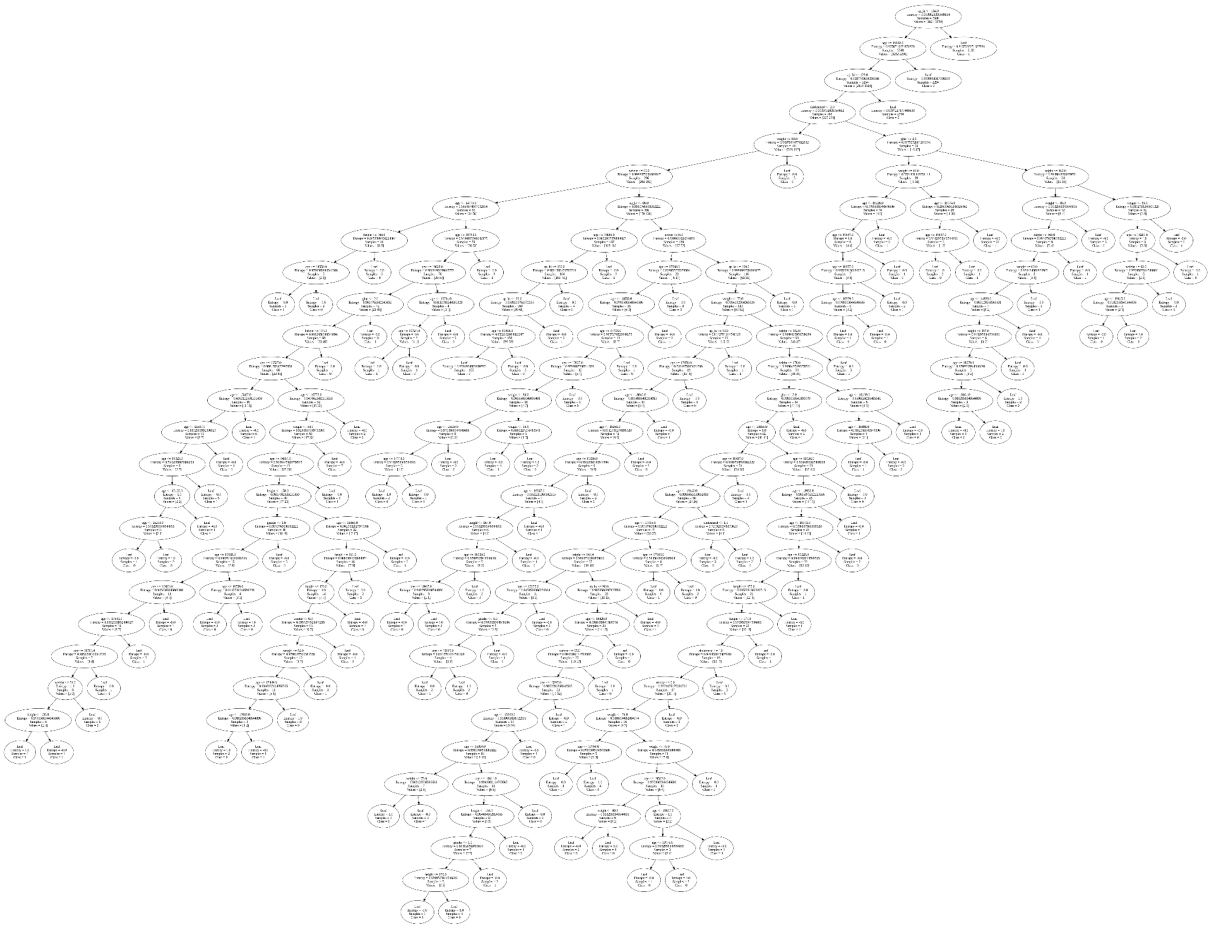
The noise in the dataset affects both the input features and the target variable. This introduces uncertainty, which may lead the decision tree to create overly complex rules that do not generalise well to unseen data.

Methodology :

1. Decision Tree Construction with Noisy Data

When building the decision tree with noisy data, the model tends to overfit to the noise, creating a complex tree that tries to capture every fluctuation in the data. As a result, the tree becomes deeper, with many branches reflecting spurious patterns that are not representative of the underlying relationships in the data.

Figure 1: Decision Tree Before Pruning (Noisy Data)



Accuracy before pruning (noisy data): `0.613410731696477`

2. Effect of Noise on the Tree

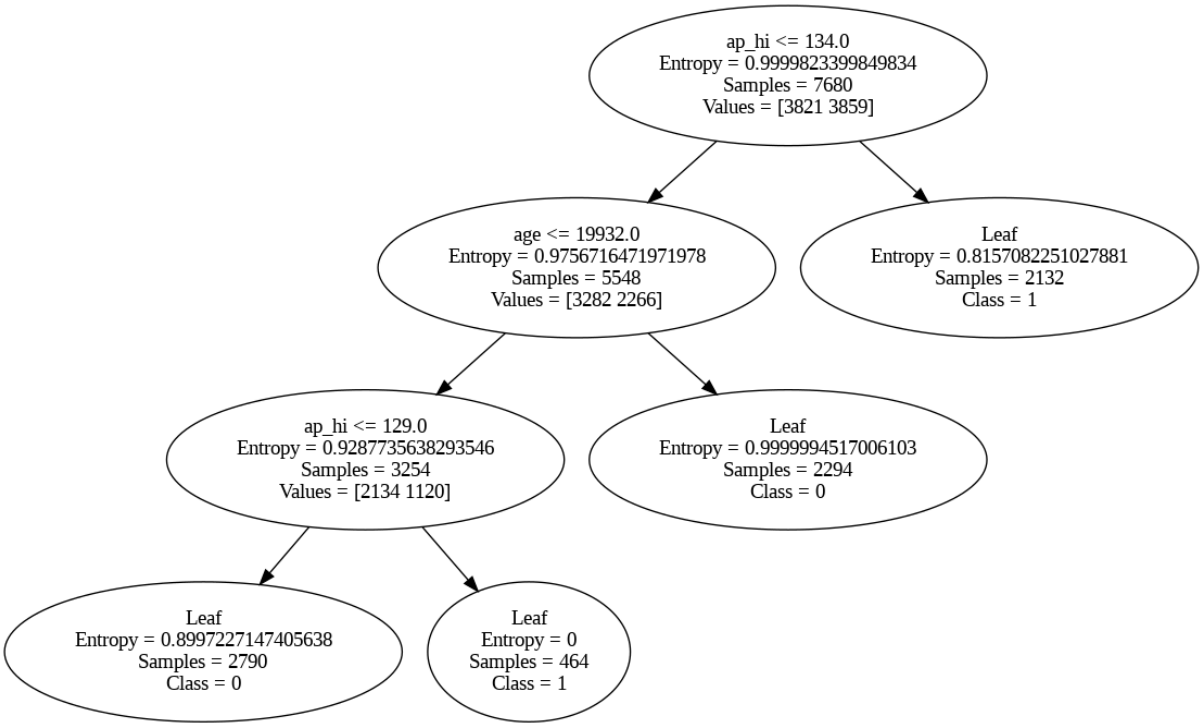
The presence of noise leads to a larger tree with many more decision nodes and branches than in the noiseless case. This is because the model attempts to perfectly classify the noisy training examples, which increases the risk of overfitting. Overfitting occurs when the model captures not only the true underlying patterns but also the random noise present in the training data, resulting in poor generalisation to new data.

3. Effect of Pruning on Noisy Data

After pruning, the model's accuracy typically drops on the noisy training set because pruning removes some of the decision branches that were designed to fit the noisy examples. However, pruning helps in simplifying the tree and reducing overfitting, making the model less sensitive to noise. This results in better generalisation to the validation and test sets.

Pruning in the presence of noise effectively removes some of the spurious patterns that the model captured. By cutting off unnecessary branches, the decision tree focuses more on the underlying structure of the data, rather than overfitting to noise. This often results in a slight decrease in training accuracy but an improvement in validation and test accuracy.

Figure 2: Decision Tree After Pruning (Noisy Data)



Accuracy after pruning (noisy data): `0.6170833333333333`

Conclusion

In noisy datasets, decision trees are particularly vulnerable to overfitting, as they attempt to create complex decision boundaries that accommodate the noise in the data. This leads to larger trees with reduced generalisation ability. Pruning is essential in this scenario, as it removes branches that fit the noise, resulting in a simpler and more interpretable model.

The results from the noisy dataset show that while pruning may slightly reduce the training accuracy, it significantly improves the model's performance on unseen data by focusing on the true patterns and ignoring the noise. In real-world scenarios where data often contains noise, pruning can be a powerful tool to improve the robustness of decision tree models.

Assignment_2_part_2 for PCA

Report on Principal Component Analysis (PCA)

Introduction :

In this report, we analyse the application of Principal Component Analysis (PCA) on a dataset to reduce its dimensionality while retaining the most important features. PCA is a widely used technique for reducing the dimensionality of large datasets, improving interpretability, and minimising information loss. This report details the methodology and results obtained from applying PCA on the provided dataset.

Objective:

The main objective of this assignment is to use PCA to transform a high-dimensional dataset into a lower-dimensional space, analyze the variance retained by the principal components, and explore how well PCA can improve the performance of a downstream machine learning model, such as a decision tree classifier.

Dataset :

The dataset used in this assignment is breast can. It consists of 29 features and 569 data points. The dataset was preprocessed to standardise the features, ensuring that each feature contributes equally to the PCA transformation.

The dataset was split into:

- Training set: 80% of the data
- Test set: 20% of the data

Methodology :

1. Data Standardization

Before applying PCA, the data was standardised. Standardisation ensures that each feature has a mean of 0 and a standard deviation of 1, which is important for PCA, as it is sensitive to the scales of the variables. The following steps were performed:

- Subtracted the mean from each feature.
- Divided by the standard deviation to normalise the data.

2. Applying PCA

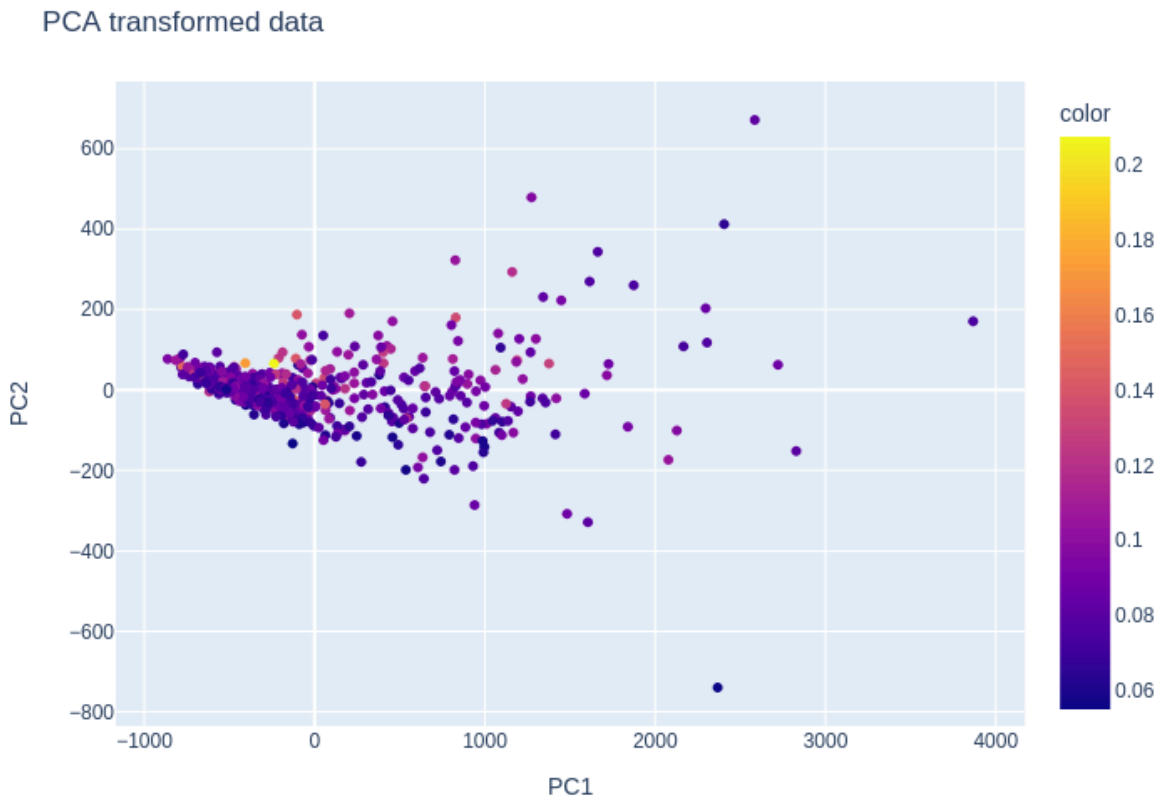
PCA was applied to the standardised dataset to identify the principal components. These components represent new dimensions that capture the maximum variance in the dataset while being orthogonal to each other.

The transformation reduced the dimensionality of the data from 569x29 to 569x2 significantly simplifying the dataset.

3. Visualizing Data in Reduced Dimensions

After applying PCA, we visualised the dataset in 2D space with the top 2 principal components. This helped to understand how well the PCA transformation separates the data points in the reduced dimensional space.

Figure 2: 2D Visualisation Using the First Two Principal Components



Conclusion :

PCA proved to be an effective method for reducing the dimensionality of the dataset. It managed to capture most of the variance in a reduced number of components, resulting in a simpler representation of the data. This dimensionality reduction improved the computational efficiency of the model without compromising its accuracy.