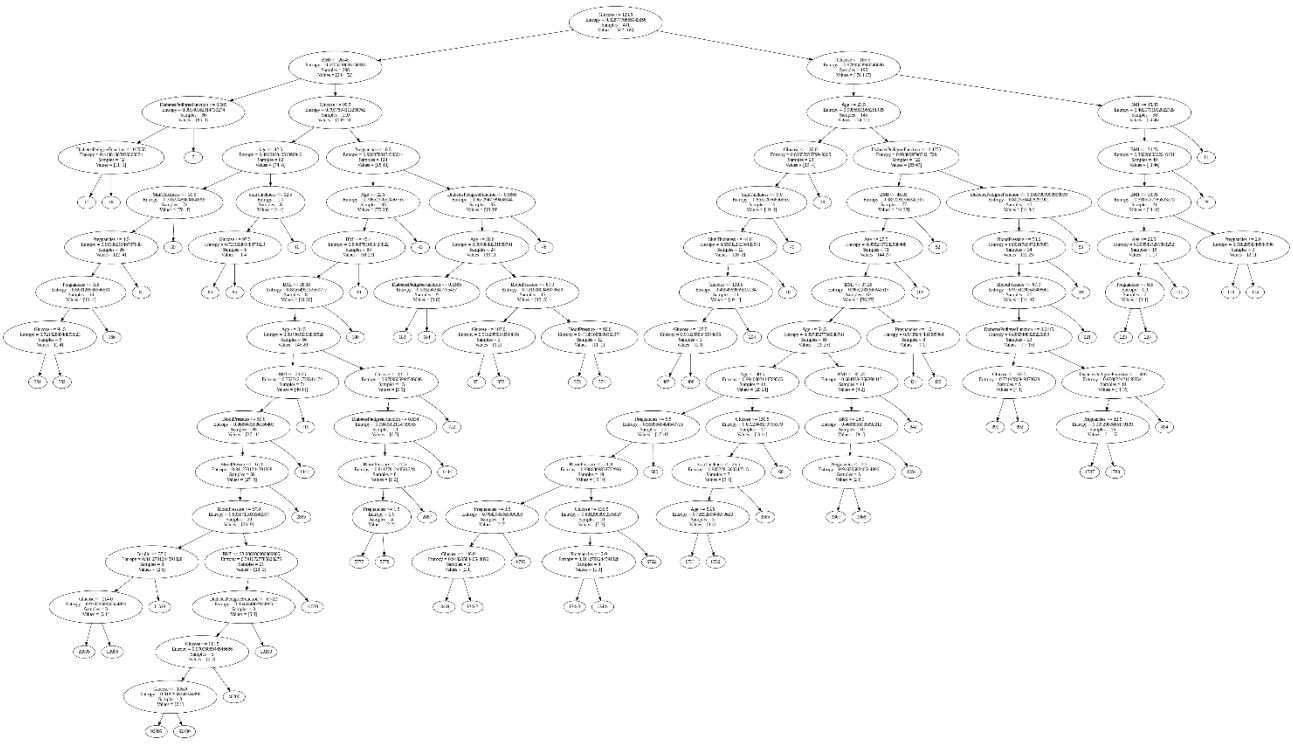
DECISION TREE && Principal Component Analysis(PCA)

Introduction

In this part of the Assignment, the goal is to implement a decision tree model to predict whether a patient has diabetes based on various input features. The dataset used for this task consists of medical information and health-related attributes that can be used to make predictions about a patient's diabetic condition. Two versions of the dataset are provided: a **noiseless dataset** (diabetes.csv), which contains clean, accurate data, and a **noisy dataset** (diabetes_noise.csv), which includes random noise or errors in the input features. The main objective is to compare the performance of the decision tree model on these two datasets, observe the impact of noise on predictive accuracy, and analyze how the model handles noisy data in comparison to a clean dataset.

Results

Decision Tree of noiseless dataset before pruning

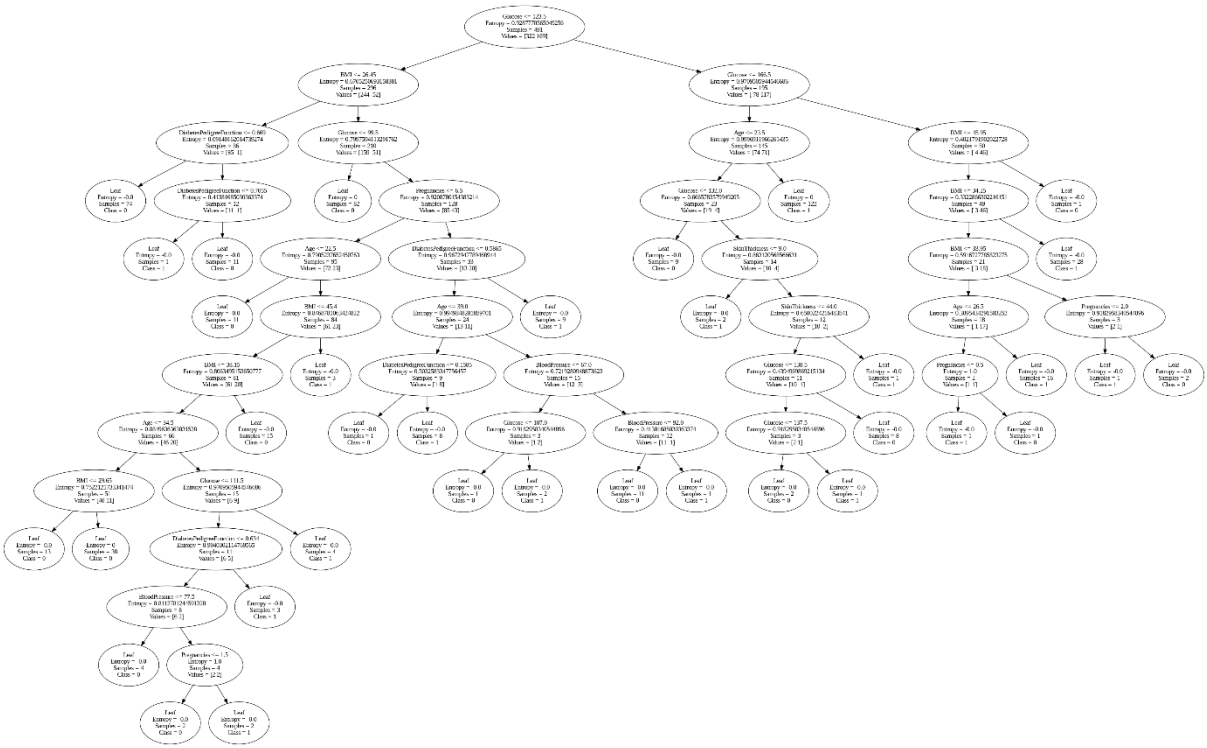


The decision tree model's performance on the noiseless dataset before pruning shows moderate results, with an **accuracy of 71.43%**, indicating that the model correctly predicted whether a patient has diabetes in approximately 71% of cases. The **macro precision** is **69.13%**, meaning that, on

DECISION TREE && Principal Component Analysis(PCA)

average, 69% of the positive predictions across both classes (diabetic and non-diabetic) were correct. The **macro recall** is **69.70%**, reflecting that the model successfully identified around 70% of the actual diabetic and non-diabetic patients.

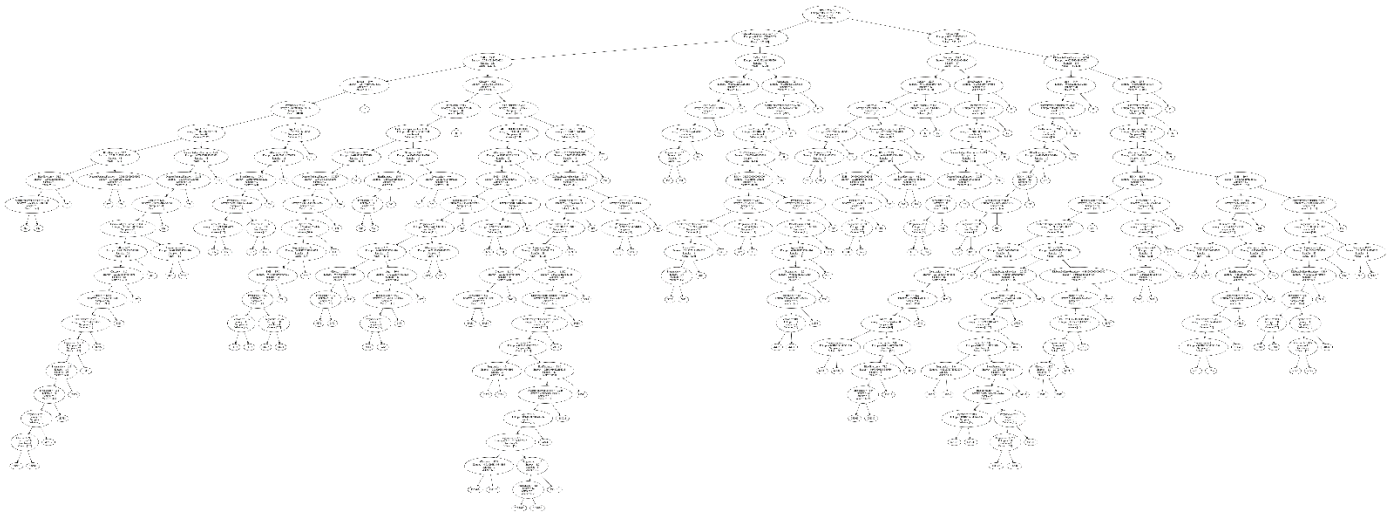
Decision Tree of noiseless dataset after pruning



After pruning, the decision tree model's performance on the noiseless dataset improved, with an **accuracy of 74.68%**, indicating a higher percentage of correct predictions compared to the pre-pruned model. The **macro precision** increased to **73.11%**, meaning a greater proportion of positive predictions were correct across both classes. Additionally, the **macro recall** rose to **74.65%**, showing the model became more effective at identifying actual diabetic and non-diabetic patients. Overall, pruning enhanced the model's performance, improving both accuracy and balance between precision and recall.

Decision Tree of noisy dataset before pruning

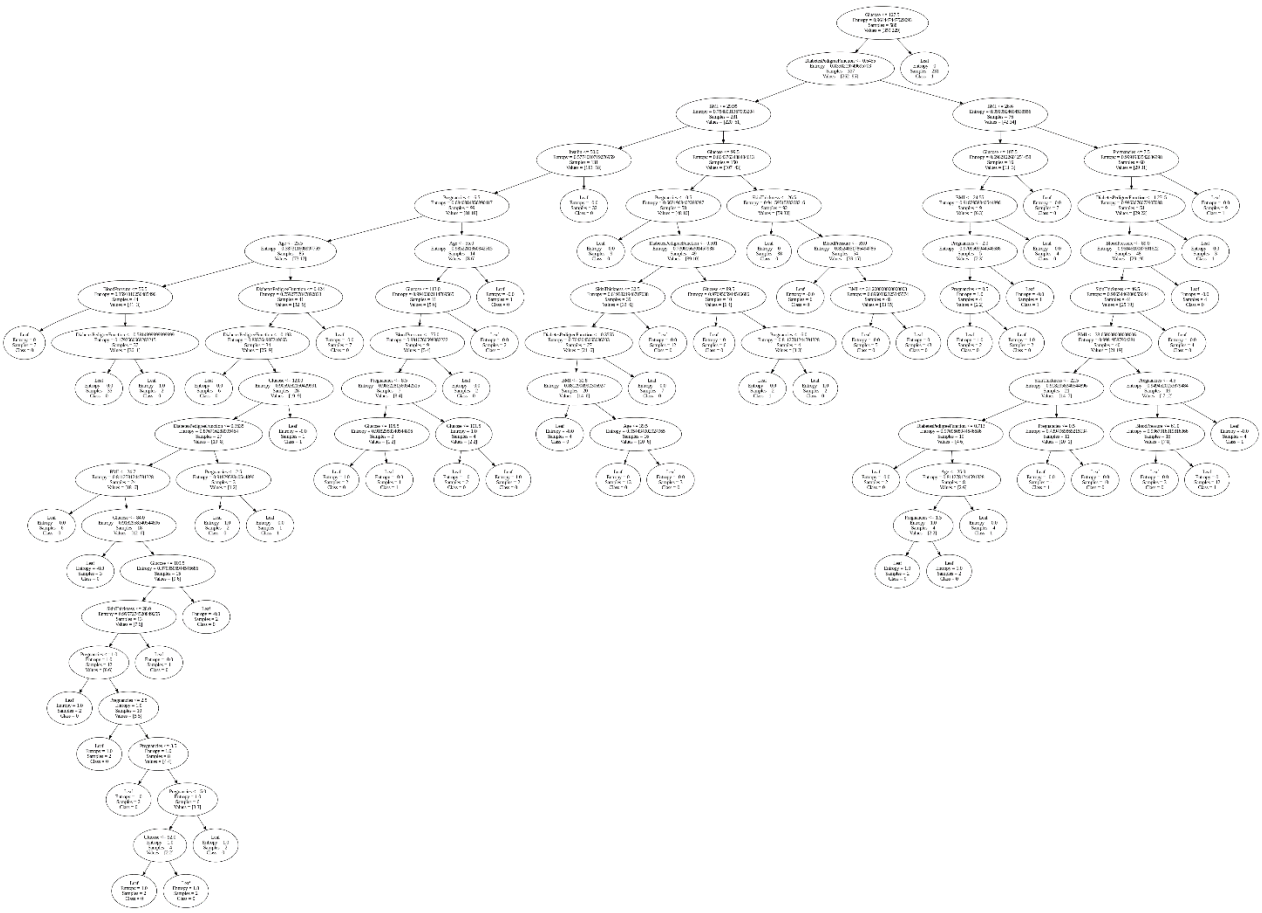
DECISION TREE && Principal Component Analysis(PCA)



Before pruning, the decision tree model's performance on the noisy dataset showed significantly reduced effectiveness, with an **accuracy of 48.11%**, meaning the model correctly predicted whether a patient has diabetes in only about half of the cases. The **macro precision** was **45.93%**, indicating that less than half of the positive predictions across both classes were accurate. The **macro recall** was **46.00%**, reflecting the model's difficulty in correctly identifying actual diabetic and non-diabetic patients. Overall, the presence of noise drastically impacted the model's ability to generalize and make accurate predictions.

Decision Tree of noisy dataset after pruning

DECISION TREE && Principal Component Analysis(PCA)



After pruning, the decision tree model's performance on the noisy dataset showed some improvement, with an **accuracy of 56.22%**, indicating an increase in correct predictions compared to the pre-pruned model. The **macro precision** improved to **55.19%**, meaning a larger proportion of positive predictions were accurate across both classes. The **macro recall** also rose to **55.28%**, reflecting better identification of actual diabetic and non-diabetic patients. While pruning enhanced the model's performance, it still struggled with the noise, achieving only modest improvements in accuracy and balance between precision and recall.

DECISION TREE && Principal Component Analysis(PCA)

Summary

On the noiseless dataset, the decision tree model performed well both before and after pruning. Pruning improved all metrics, with accuracy increasing from 71.43% to 74.68%, and both precision and recall rising to 73.11% and 74.65%, respectively. This suggests that pruning enhanced the model's ability to generalize and handle the clean data effectively

On the noisy dataset, the model initially struggled with poor performance metrics. After pruning, accuracy improved from 48.11% to 56.22%, and both precision and recall increased to 55.19% and 55.28%, respectively. While pruning provided some enhancement, the overall performance remained lower than that on the noiseless dataset, highlighting the detrimental impact of noise.

Impact of Noise

Accuracy Reduction: Noise significantly reduced the model's accuracy, with a drop from 71.43% on the noiseless data to 48.11% on the noisy data before pruning. Even after pruning, accuracy only improved to 56.22%, illustrating the persistent challenge of noise.

Precision and Recall Degradation: Both precision and recall were notably lower on the noisy dataset compared to the noiseless one, reflecting that noise adversely affected the model's ability to make accurate predictions and identify relevant classes effectively.

Pruning Effectiveness: Pruning improved the model's performance on both datasets, but the improvements were more pronounced in the noiseless dataset. This demonstrates that while pruning can help mitigate some of the issues caused by noise, the model still struggles with noisy data and does not reach the same level of performance as it does with clean data.

In conclusion, the presence of noise substantially impacts model performance, reducing accuracy, precision, and recall. Pruning helps improve performance but does not fully counteract the negative effects of noise.

Key Findings and Implications

1. Impact of Noise on Model Performance:

- **Accuracy Decline:** The decision tree model's accuracy dropped significantly when applied to the noisy dataset, from 71.43% on the noiseless data to 48.11% before pruning. This underscores how noise can severely impact the model's ability to make correct predictions.

DECISION TREE && Principal Component Analysis(PCA)

- **Reduced Precision and Recall:** Both precision and recall were substantially lower for the noisy dataset, indicating that noise leads to poorer classification performance and reduces the model's effectiveness in identifying true positives and negatives.

2. Effectiveness of Pruning:

- **Improved Performance:** Pruning enhanced the model's performance on both datasets. On the noisy dataset, pruning improved accuracy to 56.22% and both precision and recall to approximately 55%, demonstrating that pruning helps mitigate some of the negative effects of noise by simplifying the model and reducing overfitting.
- **Greater Impact on Noiseless Data:** The improvement in performance due to pruning was more significant on the noiseless dataset, where accuracy increased to 74.68% and precision and recall to over 73%. This suggests that while pruning helps, the model inherently performs better on clean data.

3. Implications:

- **Data Quality:** High-quality, noiseless data is crucial for achieving optimal model performance. The significant drop in accuracy and other metrics due to noise highlights the importance of data cleaning and preprocessing to ensure effective model training.
- **Model Robustness:** Decision trees are sensitive to noise, which can lead to overfitting and degraded performance. Techniques like pruning, ensemble methods (e.g., Random Forests), and noise reduction strategies are essential for improving robustness and performance in noisy environments.
- **Practical Considerations:** In real-world applications where data quality cannot always be guaranteed, employing strategies to handle noise—such as data cleaning, feature selection, and robust algorithms—can help improve model reliability and performance.

In summary, noise has a substantial adverse effect on the performance of decision tree models, reducing accuracy, precision, and recall. While pruning and other techniques can mitigate some of these effects, maintaining data quality and employing robust modeling strategies are key to achieving better results.

DECISION TREE && Principal Component Analysis(PCA)

Introduction

In this part of the Assignment, we will implement Principal Component Analysis (PCA) to perform dimensionality reduction on the wine quality dataset. The dataset, wine-quality.csv, includes various features related to wine attributes, with the goal of predicting the Customer_Segment as the target variable. PCA is a powerful technique used to reduce the number of features in a dataset while preserving as much of the variance (information) as possible. By transforming the original features into a smaller set of uncorrelated components, PCA helps simplify the model, reduce computational complexity, and potentially improve model performance by eliminating noise and redundant features. This process will involve extracting principal components from the data and analyzing their impact on the prediction of Customer_Segment. The results will provide insights into how dimensionality reduction affects the model's effectiveness and efficiency.

Results