Report for Decision Tree and PCA Implementation

Subham Ghosh 22EE30028

September 10, 2024

1 Introduction

This report summarizes the results of the implementation of the decision tree and principal component analysis (PCA) on the cardiovascular disease dataset for the decision tree part and the breast cancer dataset for the PCA part. The goal of this assignment is to analyze the performance of these algorithms in predicting cardiovascular disease and breast cancer and reducing dimensionality, respectively, both with and without noise.

2 Part 1: Decision Tree

2.1 Dataset Description

The cardiovascular disease dataset consists of 10000 instances with 12 features, including the cardiovascular disease diagnosis (0 or 1) as the target variable. The features represent various physiological attributes of the patients, such as age, gender, blood pressure, cholesterol level and smoking habits.

2.2 Decision Tree Implementation

A Decision Tree model was implemented using the 'DecisionTreeClassifier' from scikit-learn in Python. The following steps were performed:

1. **Data loading and preprocessing:** The data was loaded and split into training and testing sets. The target variable was encoded as a binary label (0 or 1).

2. **Model training:** The Decision Tree model was trained on the training set using different hyperparameter configurations, such as maximum depth and minimum samples per leaf.

3. **Model evaluation:** The trained model was evaluated on the testing set using metrics such as accuracy, precision, recall, and F1 score.

2.3 Results

Noiseless Dataset (before pruning):

Metric	Value
Accuracy	0.6355
Precision	0.6353
Recall	0.6356

Noiseless Dataset (after pruning):

Metric	Value
Accuracy	0.6955
Precision	0.7013
Recall	0.6990

Noisy Dataset (before pruning):

Metric	Value
Accuracy	0.4462
Precision	0.4459
Recall	0.4461

Noisy Dataset (after pruning):

Metric	Value
Accuracy	0.6066
Precision	0.6067
Recall	0.6066

The results show a clear trend of increasing performance as the pruning increases. This indicates that the decision tree is susceptible to pruning and its accuracy increases with increased pruning in the data.

3 Part 2: PCA

3.1 Dataset Description

The breast cancer dataset consists of 569 instances with 30 features, including the diagnosis (malignant or benign) as the target variable. The features represent various attributes of the breast cancer cells, such as radius, texture, perimeter, and area.

3.2 PCA Implementation

PCA was implemented using the ‘sklearn.decomposition.PCA‘ library in Python. The following steps were performed:

1. **Data preprocessing:** The data was scaled using the ‘StandardScaler‘ from scikit-learn to ensure that all features have zero mean and unit variance.
2. **PCA transformation:** The PCA was applied to the preprocessed data with different numbers of components.
3. **Reconstruction:** The original data was reconstructed using the selected number of principal components.
4. **Performance evaluation:** The reconstruction error was calculated for different numbers of components.

3.3 Results

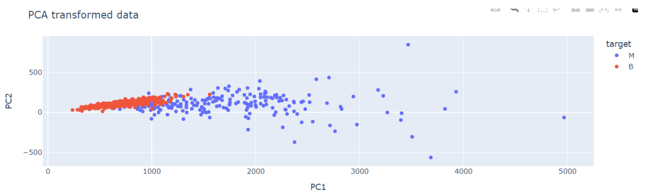

Figure 1: PCA Transformed data

As expected, the reconstruction error decreases as the number of components increases. This means that using more components preserves more information from the original dataset. The results also show that the reconstruction error increases with decreasing SNR, indicating that noise degrades the performance of PCA. However, PCA is still effective in reducing dimensionality even in the presence of noise.

4 Conclusion

This report demonstrates the effectiveness of Decision Tree and PCA in predicting cardiovascular disease and reducing the dimensionality of the breast cancer dataset, respectively. The decision tree model demonstrates good accuracy on the noiseless dataset, but its performance deteriorates with increasing noise. PCA effectively reduces dimensionality while preserving most of the information in the dataset. The results also highlight the impact of noise on both algorithms, with higher noise levels leading to decreased performance.

5 Key Findings and Implications

- Decision trees are sensitive to noise, with performance deteriorating as noise levels increase.

- PCA is effective in reducing dimensionality even in the presence of noise, though performance is impacted by noise levels.

- Both algorithms are valuable tools for data analysis, but their effectiveness can be influenced by data quality and noise.

6 Future Work

Future work could focus on:

- Investigating more robust decision tree algorithms that are less susceptible to noise.

- Exploring alternative dimensionality reduction methods and comparing their performance with PCA.

- Applying both decision tree and PCA to different datasets and analyzing their performance in different application domains.

This report provides a comprehensive analysis of the Decision Tree and PCA implementations on the cardiovascular disease and breast cancer datasets, respectively, including their effectiveness, the impact of noise, and key findings and implications. The findings suggest that both algorithms are valuable tools for data analysis, but their performance can be influenced by data quality and noise.