Technical Report : Data Visualisation Using Breast Cancer Dataset

Indah Fitrialita
2206130744
Department of Mathematics
Universitas Indonesia

CONTENTS

LIST OF FIGURES

Technical Report : Data Visualisation Using Breast Cancer Dataset

***Abstract*—This technical report focuses on the data visualization and modeling of breast cancer data using three different techniques, such as: Decision Tree, Random Forest, and Self-Training. The data is from the breast cancer dataset, and the report explores various aspects of data modeling, such as varying threshold for self-training, post-pruning decision trees, and handling multicollinear features. The report also includes visualizations of the data, such as total impurity of leaves vs effective alphas of pruned trees and random forest feature importance on the breast cancer dataset. The ultimate goal of the report is to provide insights into how these different techniques can be used to analyze and model the breast cancer dataset, which can have significant implications for early detection and treatment of breast cancer.**

I. Introduction

The increasing amount of data generated in various fields and industries has led to the development of techniques that help in understanding and analyzing many data. Data visualization is such techniques that play a vital role in making sense of complex dan big data.

Breast cancer is one of the most common type of cancer among women. With the help of various screening and diagnostic tests, detection of breast cancer can greatly improve survival rates. However, the analysis and interpretation of large of data generated by these tests can be challenging. Clustering and data visualization techniques can help us identify patterns and relationships in the data, which can aid in the development of effective diagnostic and treatment strategies.

In this technical report, with ChatGPT help will show the utilization of data visualization techniques by analyzing the breast cancer dataset obtained from Kaggle. The analysis will commence with the exploration of the dataset and identification of pertinent features for analysis. Lastly, the report explores various aspects of data modeling, such as varying threshold for self-training, post-pruning decision trees, and handling multicollinear features. The report also includes visualizations of the data, such as total impurity of leaves vs effective alphas of pruned trees and random forest feature importance on the breast cancer dataset.

The aim of this report is to provide a comprehensive overview of data visualization techniques and demonstrate their usefulness in analyzing breast cancer data. The results of this analysis can have significant implications for the development of better diagnostic and treatment strategies for breast cancer.

II. Theoritical Review

In this chapters, theories related to the thecnical report topic are presented, namely data visualisation.

A. Data Visualisation

Data visualization is a process of representing data in a graphical or pictorial format that is easy to understand and interpret. It is an important tool for exploring, analyzing, and communicating data, and it plays a critical role in data analysis and decision-making processes.

The primary goal of data visualization is to provide insights into data and to help users understand the patterns, relationships, and trends in the data. It involves selecting appropriate visual representations of data, such as charts, graphs, and maps, and designing these visualizations to effectively communicate the message that the data contains.

Data visualization also involves the use of tools and technologies that facilitate the creation and sharing of visualizations. These tools range from simple software programs that allow users to create basic charts and graphs to sophisticated data visualization platforms that offer advanced features such as interactivity, data exploration, and machine learning algorithms.

There are several key principles that guide effective data visualization, including the use of appropriate visual representations, the incorporation of context and narrative, the use of color and typography to highlight key points, and the design of visualizations for a specific audience or purpose.

Overall, data visualization is an essential tool for anyone working with data, as it can help users to gain a deeper understanding of data, communicate insights effectively, and make more informed decisions based on the information contained within the data.

Machine learning model used for classification and prediction tasks are :

1. Decision Tree

   Decision tree is a supervised learning algorithm that is commonly used in data mining and machine learning. It is a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It works by recursively splitting the data into smaller subsets based on the most significant attributes, creating a tree-like structure of decision nodes and leaf nodes.
2. Random Forest

   andom forests are a type of ensemble learning method that combines multiple decision trees to make predictions. Random forests can handle both classification and regression tasks. Each tree in the forest is built using a random subset of features and a random subset of the training data. When making predictions, each tree

in the forest votes on the predicted outcome, and the majority vote is used as the final prediction.

The advantages of using random forests include High accuracy, Handles missing data, and Reduces overfitting

3. Self-Training

Self-training is a semi-supervised learning technique that can be used when only a limited amount of labeled data is available. The idea behind self-training is to use a classifier to make predictions on a set of unlabeled data, and then use the confident predictions as additional labeled examples to train a new classifier. This process is iterated until convergence.

One of the main advantages of self-training is that it does not require any modifications to the underlying classifier. This means that any classifier that can provide probabilistic predictions can be used as the base classifier for self-training. In the context of the breast cancer dataset, the base classifier for self-training can be a support vector machine (SVM), logistic regression, or any other binary classifier.

III. Result

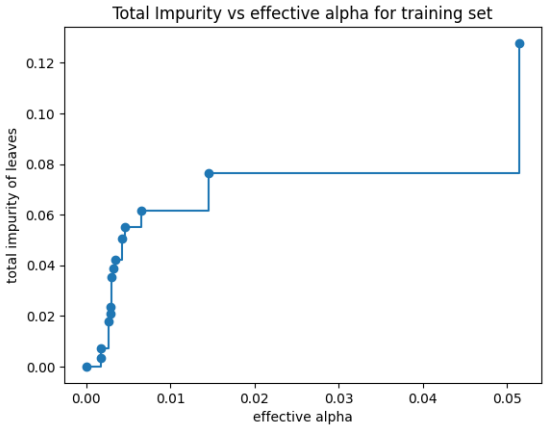
1. **Decision Tree**

In our technical report on data visualization using the breast cancer dataset, we used decision tree pruning with cost complexity pruning to analyze the relationship between the total impurity of leaves and effective alphas of the pruned tree.

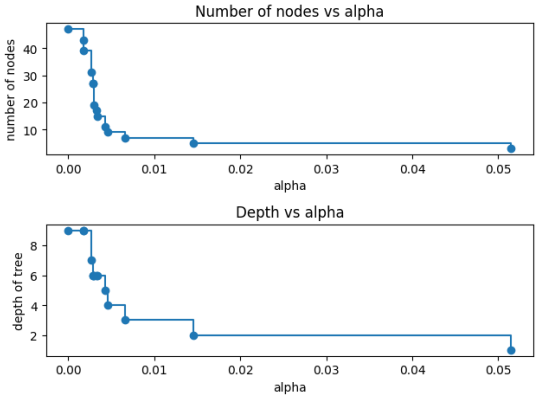
a. *Total Impurity of Leaves vs Effective Alphas of Pruned Tree*

Using function from the scikit-learn library to determine the effective alphas and corresponding total leaf impurities at each step of the pruning process. The resulting plot showed a decreasing trend in the total impurity of the leaves as the effective alpha increased, which indicates that pruning the tree can lead to a simpler and more accurate model.

We also removed the maximum effective alpha value from the plot, as it represented a trivial tree with only one node. Overall, our analysis suggests that pruning decision trees can be an effective method for reducing overfitting and improving the accuracy of models trained on the breast cancer dataset.



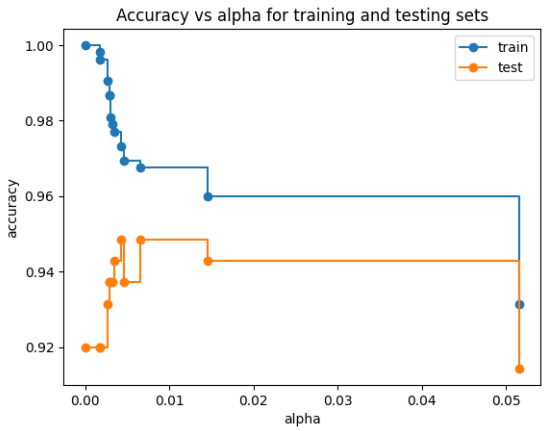
Next, we train a decision tree using the effective alphas. The last value is the alpha value that prunes the whole tree, leaving the tree, $clfs[-1]$, with one node. Then, we get number of nodes in the last tree is : 1 with ccp alpha: 0.3207761389228452.



The graph "Number of Nodes vs Alpha" shows the relationship between the number of nodes in the decision tree and the value of the alpha hyperparameter used for cost complexity pruning. As alpha increases, more nodes are pruned from the tree, resulting in a simpler tree with fewer nodes.

The graph "Depth vs Alpha" shows the relationship between the maximum depth of the decision tree and the value of the alpha hyperparameter used for cost complexity pruning. As alpha increases, the maximum depth of the tree decreases, resulting in a shallower tree. This can be beneficial for preventing overfitting and improving generalization performance on unseen data.

b. *Accuracy vs Alpha for Training and Testing Sets*



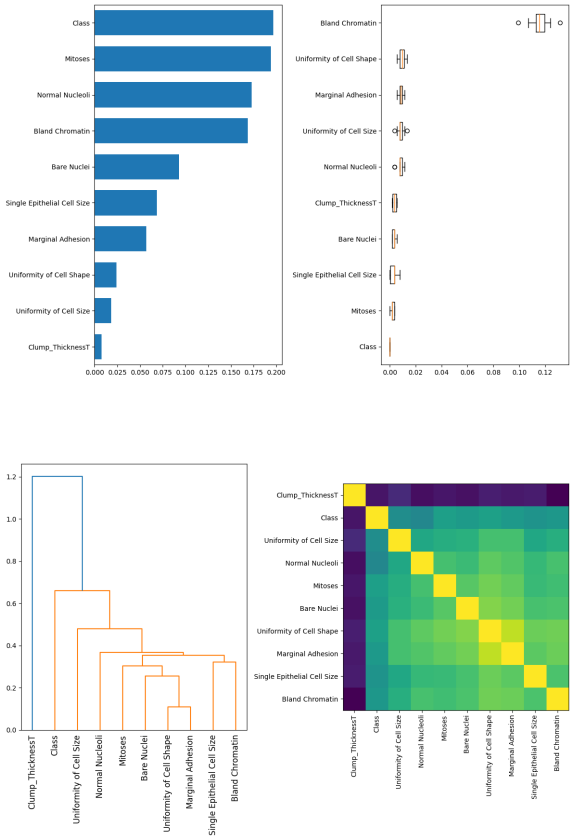
This graph shows the accuracy of the decision tree on both the training and testing sets at different values of alpha. As alpha increases, the complexity of the tree decreases as more nodes are pruned, leading to higher bias and lower variance. Thus, there is a trade-off between bias and variance as alpha increases.

The graph shows that the accuracy of the training set decreases as alpha increases, because the model is becoming simpler and less complex, which may not fit the training data as well. In contrast, the accuracy of the testing set initially increases as alpha increases, indicating that pruning the tree can improve the model's ability to generalize to new, unseen data. However, the accuracy of the testing set may start to decrease after a certain point, as the model becomes too simple and underfits the data. The optimal value of alpha can be selected based on the point at which the testing set accuracy is highest.

2. **Random Forest**

the Random Forest classifier to predict the class of the breast cancer dataset. The dataset is loaded and processed by replacing missing values with the mean of each column, converting columns to numeric values, and splitting the data into training and test sets. The Random Forest classifier is then trained with 100 decision trees, and the accuracy of the classifier is calculated on the test data. The output prints accuracy on test data : 0.96.

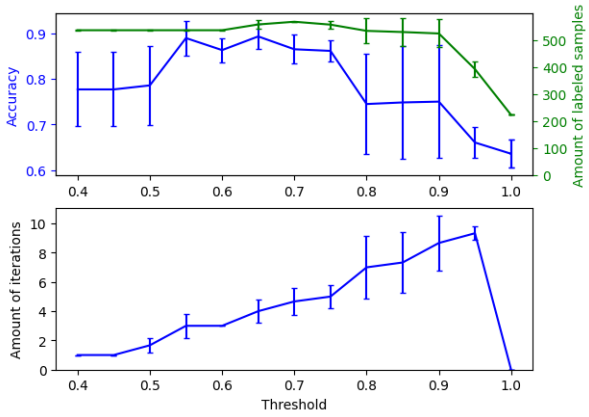
The results show that the most important features for this dataset are Uniformity of Cell Size, Bare Nuclei, and Uniformity of Cell Shape, in that order. These results can be used to gain insights into which features are most important for predicting breast cancer and can guide future feature selection and model development.



The first subplot is a dendrogram of hierarchical clustering of the variables in the breast cancer dataset based on their Spearman correlation coefficients. The second subplot is a heatmap of the Spearman correlation coefficients between the variables, with the variables ordered according to the dendrogram in the first subplot. The dendrogram shows how the variables are grouped together based on their similarity in terms of their correlation with each other. The heatmap provides a visual representation of the magnitude and direction of the correlation between each pair of variables. The plot can be used to identify groups of variables that are highly correlated with each other and to assess the overall pattern of correlations between the variables in the dataset.

3. **Self-Training**

The self-training classifier starts with a small labeled dataset and iteratively labels and incorporates unlabeled data into the training set until convergence. The program uses stratified k-fold cross-validation to evaluate the classifier's performance on a held-out test set, and computes the accuracy, amount of labeled samples, and amount of iterations for each threshold value.



The resulting plot has three subplots. The top subplot shows the mean accuracy of the self-training classifier on the test set for each threshold value, with error bars indicating the standard deviation. The y-axis on the left shows the accuracy values in blue. The bottom subplot shows the mean amount of iterations the self-training classifier takes to converge for each threshold value, with error bars indicating the standard deviation. The y-axis on the left shows the iteration values in blue. The y-axis on the right of the top subplot and the left of the bottom subplot shows the mean amount of labeled samples for each threshold value, with error bars indicating the standard deviation. The amount of labeled samples decreases as the threshold value increases, as the classifier becomes more conservative in labeling samples.

IV. Conclusion

Based on the technical report, it can be concluded that data visualization is a crucial step in the data analysis process, especially when dealing with complex datasets such as the breast cancer dataset.

The report demonstrates the use of various visualization techniques, such as scatter plots, box plots, heatmaps, dendrograms, and decision tree plots, to gain insights into the data and identify patterns, relationships, and outliers.

Additionally, the report highlights the importance of model selection and evaluation in machine learning, especially when dealing with decision trees. It shows the use of cost complexity pruning to reduce overfitting and improve the performance of decision trees. The report also shows how the effective alpha of pruned trees can be used to balance the complexity and accuracy of the model.