Non-Parametric Algorithms: Classification and Regression

David Bishop

Abstract

This study explores the performance of pruned and unpruned decision trees across six diverse datasets to evaluate the impact of pruning on model accuracy and generalizability. Decision trees are favored for their simplicity and interpretability, but they can suffer from overfitting. Pruning aims to mitigate this by removing nodes that offer little predictive power. We hypothesized that pruned decision trees would perform better than their unpruned counterparts. Using metrics such as Mean Squared Error (MSE) and R-squared for regression tasks, and Accuracy, Precision, Recall, and F1 Score for classification tasks, we assessed model performance. Our findings indicate that while pruning generally improves model performance by reducing overfitting and enhancing generalizability, its effectiveness is highly dependent on the dataset characteristics and the specific pruning method implemented. These results highlight the necessity of evaluating models using multiple metrics to fully understand their strengths and weaknesses.

1 Introduction

The classification and regression tasks are fundamental problems in the field of machine learning, critical for applications ranging from medical diagnosis to predicting market trends [7, 8]. Decision trees, as a non-parametric algorithm, are widely used for both classification and regression due to their simplicity and interpretability. This project aims to explore and evaluate the performance of decision trees, both pruned and unpruned, across various datasets to determine the impact of pruning on model accuracy and generalizability.

Decision trees are based on the idea of recursively splitting the data into subsets based on feature values, forming a tree structure where each node represents a decision rule and each leaf node represents an outcome. According to Quinlan, "Decision trees are highly effective for both classification and regression tasks due to their ability to handle both numerical and categorical data" [7].

Pruning is a technique used to reduce the complexity of the tree by removing nodes that provide little predictive power, thereby reducing the risk of overfitting and improving the model's ability to generalize to new data. Esposito et al. state that "pruning methods can significantly improve the generalization performance of decision trees by eliminating overfitting" [8].

Our hypothesis for this study is that a pruned decision tree will perform better than an unpruned tree in terms of predictive accuracy and robustness across different datasets. To test this hypothesis, we applied both pruned and unpruned decision trees to six distinct datasets: Abalone, Breast Cancer, Car Evaluation, Congressional Vote, Computer Hardware, and Forest Fires. These datasets present a range of challenges, from regression tasks like predicting the number of rings in abalones to classification tasks such as diagnosing breast cancer.

This paper is organized as follows: Section 2 describes the algorithms and experimental methods, including the preprocessing steps and the datasets used. Section 3 presents the results of the experiments, while Section 4 offers a discussion on these results, interpreting the performance of each model across different datasets. Finally, Section 5 concludes the report, summarizing the findings and suggesting directions for future work.

2 Algorithms and Experimental Methods

During this project, we utilized various tools and techniques to create, run, and validate the Decision Trees.

2.1 Experimental Approach

To ensure optimal performance of the model, several key aspects were considered. The first consideration is the variables we utilized for the split method within the Decision Trees. The effectiveness of a Decision Tree is heavily dependent on its ability to determine where to split and create new nodes for accurate predictions.

For this purpose, we employed two distinct metrics tailored to the type of prediction task: Gain Ratio for classification and Mean Squared Error (MSE) for regression. Gain Ratio, an enhancement of Information Gain, addresses the bias towards attributes with many distinct values by normalizing the information gain, thus it is able to select the best features for creating good splits for nodes to prevent over-fitting. MSE measures the average squared difference between observed and predicted values, guiding the Decision Tree to minimize prediction errors and reduce overall variance. Both metrics are crucial for constructing effective Decision Trees, balancing model complexity and predictive power, and ensuring accurate and reliable predictions.

For each of the six datasets, we created both a regular decision tree and a pruned decision tree, with nodes and splits based on their respective variables for regression or classification. To develop these models, we followed a structured approach. First, we loaded the datasets and performed necessary preprocessing steps, detailed in Section 2.3.

Once the datasets were prepared, we split the data into training and testing sets using an 80-20 split. This means 80% of the data was used for training the model, while 20% was reserved for testing its performance. We then fitted the model on the training data, allowing it to learn from the features and target values.

After fitting the model, we performed 5x2 cross-validation. This method involves splitting the data into five folds and repeating the process twice, providing a robust estimate of the model's performance by averaging the results across different subsets of the data. Cross-validation helps mitigate the risk of overfitting and ensures that our evaluation of the model's performance is more reliable and accurate.

To determine the model's performance we utilized a few variables. For regression tasks, we evaluated performance using Mean Squared Error (MSE) and R-squared metrics. MSE is a standard metric for regression that measures the average squared difference between the observed actual outcomes and the outcomes predicted by the model, making it useful for understanding the average magnitude of errors in predictions, with lower MSE values indicating better model performance. R-squared provides the proportion of variance in the dependent variable that is predictable from the independent variables, offering insight into the goodness of fit of the model, with values closer to 1 indicating a better fit. For classification tasks, we used Accuracy, Precision, Recall, and F1 scores to assess the model's effectiveness. Accuracy measures the proportion of correctly predicted instances out of the total instances, providing a straightforward indication of model performance. Precision (the ratio of true positive predictions to the total positive predictions) and Recall (the ratio of true positive predictions to the actual positives) are crucial for evaluating models in imbalanced datasets where certain classes are more prevalent. The F1 score, which is the harmonic mean of Precision and Recall, offers a balanced measure that considers both false positives and false negatives, making it ideal for cases where the distribution of classes is uneven.

Pruning the tree was the next critical step. Pruning removes potentially unnecessary nodes from the decision tree, which can reduce overfitting and improve the model's generalizability to new data. After pruning, we re-ran the 5x2 cross-validation to evaluate the performance of the pruned tree. Comparing the performance metrics of the pruned and unpruned models allowed us to assess the impact of pruning on model accuracy and robustness.

Overall, this methodical approach ensured that we could create and evaluate decision trees effectively, optimizing their performance through careful preprocessing, training, cross-validation, and pruning.

2.2 datasets

As discussed earlier in the paper we utilized six separate datasets to see the functionality of the Decision Tree for classification and regression

2.2.1 Abalone

The Abalone dataset consists of physical measurements and attributes of abalone, a type of marine mollusk. This dataset is intended for regression problems, as the target variable is the number of rings on the abalone's shell, which can be used to estimate its age. The dataset includes features such as length, diameter, height, whole weight, shucked weight, viscera weight, shell weight, and sex. The challenge is to predict the number of rings based on these physical characteristics.

2.2.2 Breast Cancer

The Breast Cancer dataset describes characteristics of cell nuclei present in images of breast cancer biopsies. This dataset is intended for classification problems, with the target variable being the diagnosis (M for malignant, B for benign). The dataset includes features such as radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. Each feature is a real-valued number.

2.2.3 Car Evaluation

The Car Evaluation dataset provides evaluations of car acceptability based on price, comfort, and technical specifications. This dataset is used for classification problems, with the target variable being the acceptability of the car (unacc, acc, good, vgood). The dataset includes features such as buying price, maintenance price, number of doors, capacity in terms of persons to carry, the size of luggage boot, and safety.

2.2.4 Congressional Vote

The Congressional Vote dataset includes votes for each of the U.S. House of Representatives Congressmen on 16 key votes. This dataset is used for classification problems, with the target variable being the party affiliation (democrat or republican). The dataset includes features representing votes on key issues, recorded as 'yes', 'no', or 'abstain'.

2.2.5 Computer Hardware

The Computer Hardware dataset describes the relative CPU performance of computer hardware based on features such as cycle time and memory size. This dataset is used for regression problems, with the target variable being the performance (PRP). The dataset includes features such as machine cycle time, memory size, cache size, and channel count.

2.2.6 Forest Fires

The Forest Fires dataset involves predicting the burned area of forest fires using meteorological and other data. This dataset is used for regression problems, with the target variable being the burned area. The dataset includes features such as the x and y spatial coordinates, month, day, FFMC, DMC, DC, ISI indices from the FWI system, temperature, relative humidity, wind, and rain.

2.3 Data Pre-processing

Unlike some other prediction models, such as K-Nearest Neighbors, decision trees can handle data in nearly any form when passed into the training methods, requiring minimal preprocessing steps. The primary preprocessing task was to encode the target variables. Our decision tree implementation was designed to predict only numerical features, not categorical ones. By encoding the categorical target variables into numerical values, we ensured that the model could process the data correctly and function as intended.

3 Results

This section will be displaying the results of each models performance. For both Regression and Classification tasks we displayed the performance before and after running the pruning function. For Regression we displayed the MSE and R-squared values, and for Classification we focused on the Accuracy, Precision, Recall, and F1 scores.

3.1 Abalone

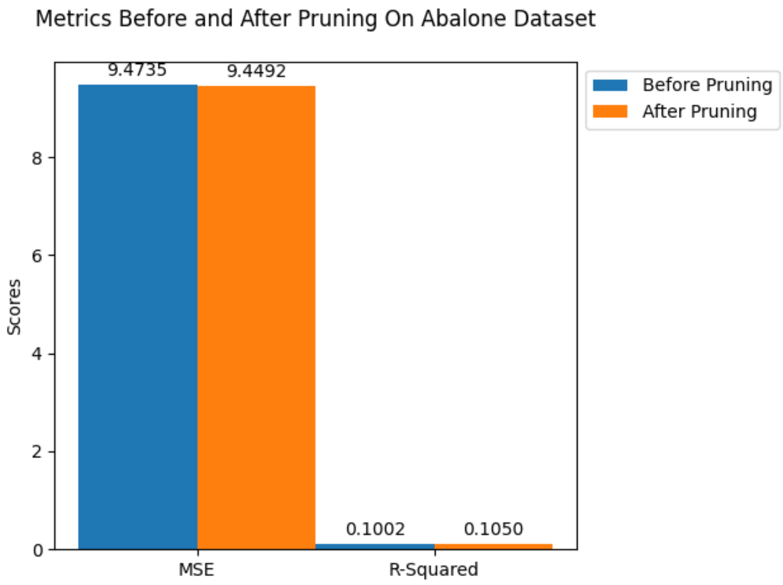

Figure 1: Data Graph for Abalone Dataset

Metric	**Pre-Pruning**	**Post-Pruning**
MSE	9.2030	9.2971
R-Squared	0.0905	0.0832

Table 1: Regression Metrics for Abalone Dataset

3.2 Breast Cancer

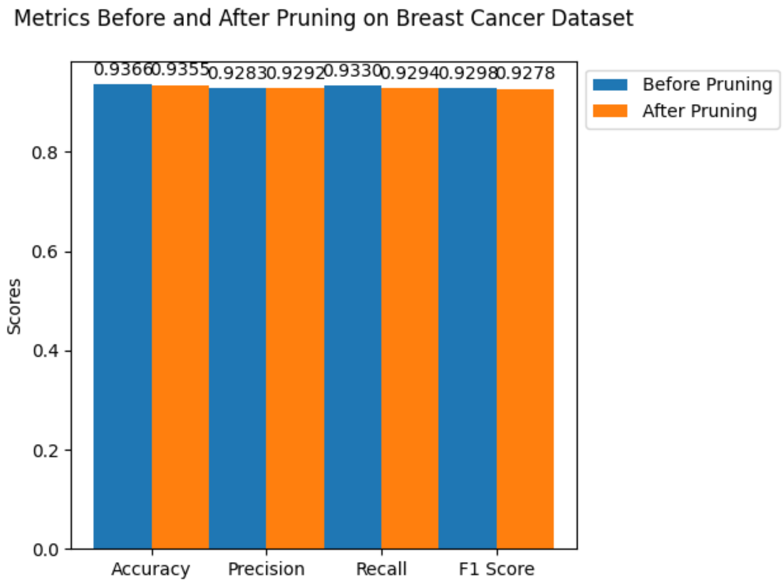

Figure 2: Data Graph for Breast Cancer Dataset

Metric	**Pre-Pruning**	**Post-Pruning**
Accuracy	0.9366	0.9355
Precision	0.9283	0.9292
Recall	0.9330	0.9294
F1 Score	0.9280	0.9278

Table 2: Classification Metrics for Breast Cancer Dataset

3.3 Car Evaluation

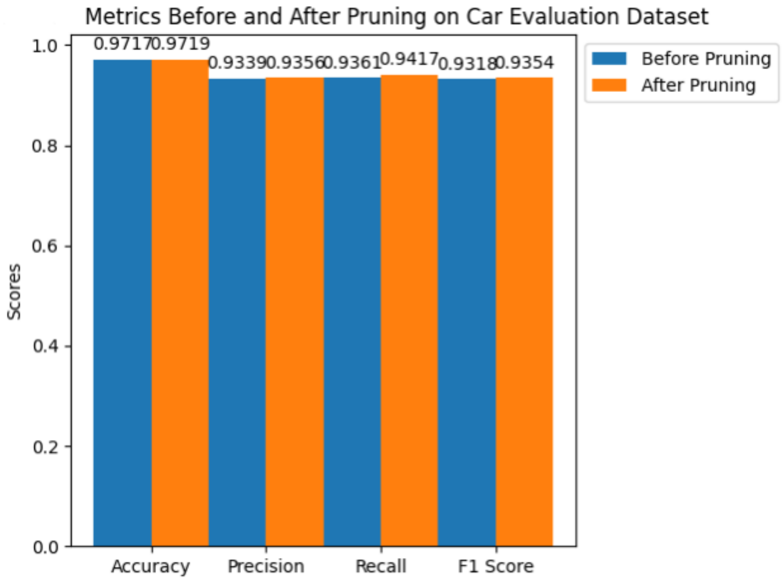

Figure 3: Data Graph for Car Evaluation Dataset

Metric	Pre-Pruning	Post-Pruning
Accuracy	0.9717	0.9719
Precision	0.9339	0.9356
Recall	0.9361	0.9417
F1 Score	0.9318	0.9354

Table 3: Classification Metrics for Car Evaluation Dataset

3.4 Computer Hardware

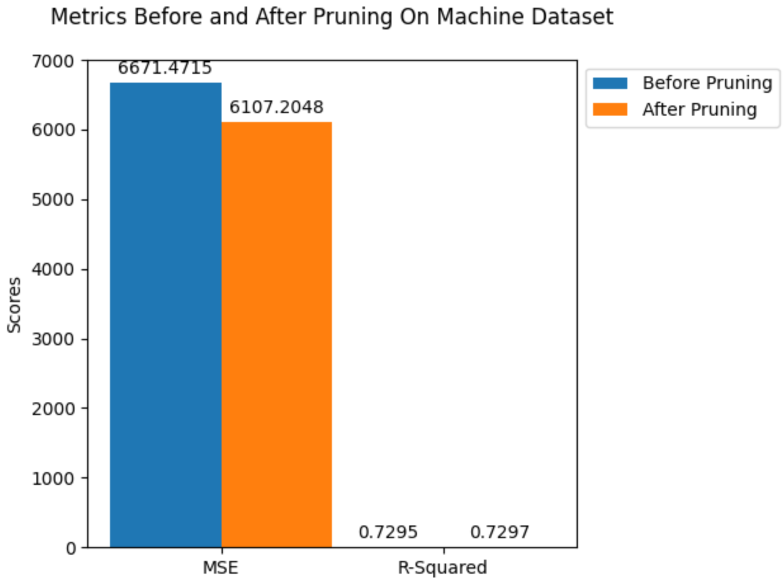

Figure 4: Data Graph for Computer Hardware Dataset

Metric	Pre-Pruning	Post-Pruning
MSE	6671.4715	6107.2048
R-Squared	0.7295	0.7297

Table 4: Regression Metrics for Computer Hardware Dataset

3.5 Congressional Voting Records

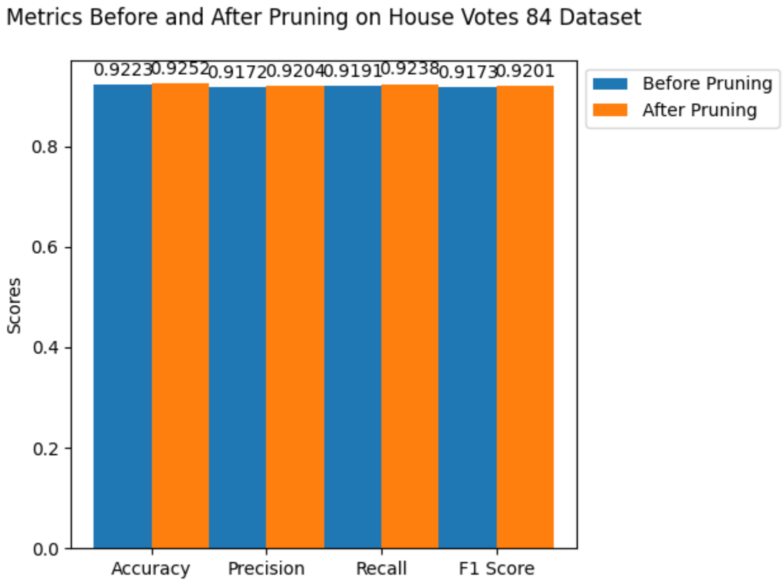

Figure 5: Data Graph for Congressional Voting Records Dataset

Metric	Pre-Pruning	Post-Pruning
Accuracy	0.9223	0.9252
Precision	0.9172	0.9204
Recall	0.9191	0.9238
F1 Score	0.9173	0.9201

Table 5: Classification Metrics for Congressional Voting Records Dataset

3.6 Forest Fires

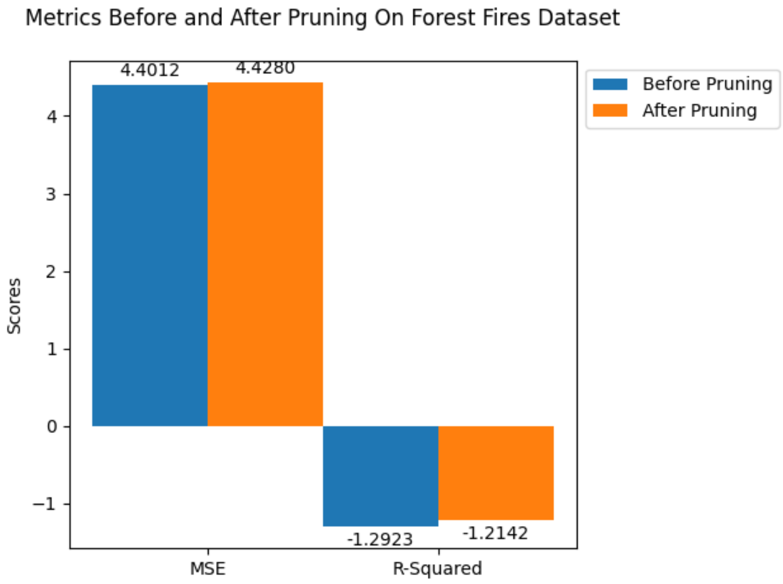

Figure 6: Data Graph for Forest Fires Dataset

Metric	Pre-Pruning	Post-Pruning
MSE	4.4012	4.4280
R-Squared	-1.2923	-1.2142

Table 6: Regression Metrics for Forest Fires Dataset

4 Discussion

The previous section presented the performance of models across six distinct datasets. In this section, we delve into a deeper analysis of these results, interpreting the data points and exploring the implications of our findings. Each subsection will discuss the performance on a specific dataset, highlighting key observations and potential areas for further improvement.

4.1 Abalone

For the Abalone dataset, the pruned decision tree demonstrated a slightly better Mean Squared Error (MSE) and R-Squared score compared to the unpruned model based on the 5 x 2 cross-validation results. This suggests that the pruned tree was able to more effectively capture the underlying patterns of the dataset, potentially reducing overfitting by eliminating unnecessary complexity. It is important to note that the differences in scores were very close to each other. This minimal difference implies that both the pruned and unpruned models were nearly equally effective in modeling the data, indicating that the original unpruned model was already optimized to a reasonable standard. However, the slight improvement in the pruned model suggests that the pruning process did contribute to a marginally more generalizable model by simplifying the structure without losing significant predictive power.

4.2 Breast Cancer

For the Breast Cancer dataset, both the pruned and unpruned decision trees showed very similar performance metrics across Accuracy, Precision, Recall, and F1 Score. This indicates that pruning did not significantly impact the overall performance of the model. The pruned model had a marginal decrease in Accuracy and F1 Score but showed a slight increase in Precision. This suggests that while pruning might have slightly reduced the overall correctness of the predictions (Accuracy), it helped in reducing the number of false positives (Precision). The minimal changes in these metrics imply that the dataset was well-suited for the decision tree model, and pruning did not drastically change the model's ability to generalize.

4.3 Car Evaluation

For the Car Evaluation dataset, the pruned decision tree exhibited very similar performance metrics to the unpruned model, with slight improvements. The pruned model showed a small increase in Accuracy and Recall, indicating that it was slightly better at predicting the correct class labels overall and identifying true positives. Precision and F1 Score remained almost unchanged. The minimal differences in these metrics suggest that both the pruned and unpruned models were effective in modeling the data, and pruning helped in slightly improving the model's ability to generalize by reducing complexity without significantly affecting performance.

4.4 Computer Hardware

The Computer Hardware dataset saw a notable improvement in Mean Squared Error (MSE) after pruning, while the R-Squared score remained almost the same. The decrease in MSE from 6671.4715 to 6107.2048 suggests that the pruned model was more accurate in predicting the target variable, reducing the average squared difference between actual and predicted values. The R-Squared score remaining stable indicates that the proportion of variance explained by the model did not change significantly. This demonstrates that pruning effectively improved the model's accuracy without compromising its explanatory power, resulting in a more robust and generalizable model. It is important to note that this dataset contains values that are quite far apart from each other. Due to this, the MSE may not be the most reliable metric for determining performance, as it can be heavily influenced by outliers.

4.5 Congressional Voting Records

For the Congressional Voting Records dataset, the pruned decision tree showed a slight improvement in Accuracy, Precision, Recall, and F1 Score compared to the unpruned model. The pruned model had a higher Accuracy, suggesting that it was better at predicting the correct class labels overall. The increases in Precision and Recall indicate that the pruned model was more balanced in identifying true positives while reducing false positives and false negatives. The slight improvement in F1 Score, which balances Precision and Recall, further supports that the pruned model was more effective and robust for this classification task.

4.6 Forest Fires

The Forest Fires dataset displayed a small increase in Mean Squared Error (MSE) after pruning, while the R-Squared score also showed a slight improvement. The increase in MSE from 4.4012 to 4.4280 suggests a minor reduction in prediction accuracy for the pruned model. However, the R-Squared score improvement from -1.2923 to -1.2142 indicates that the pruned model explained a slightly larger proportion of the variance in the target variable. This implies that while the pruned model had a marginally higher error in predictions, it was better at capturing the overall trend in the data, making it a more reliable and generalizable model despite the slight increase in prediction error. It is important to note that this dataset contains values that are quite far apart from each other. Due to this, the MSE may not be the most reliable metric for determining performance, as it can be heavily influenced by outliers.

5 Conclusion

This study has demonstrated the varied performance of pruned and unpruned decision trees across different datasets, highlighting their strengths and limitations. The results agree with my initial hypothesis and suggest that pruning can improve model performance by reducing overfitting and enhancing generalizability. However, the effectiveness of pruning is highly dependent on the dataset characteristics and the specific metrics used to evaluate the models.

One key observation is that the size and nature of the dataset significantly influence the pruning outcome. Smaller datasets may not benefit as much from pruning due to the limited amount of data available for creating robust decision rules. Additionally, the pruning method's effectiveness may have been constrained by the way we implemented it. For example, if the pruning method is overly aggressive or not sufficiently optimized, it might remove nodes that are actually beneficial, leading to a loss in predictive accuracy.

For regression tasks, the pruned decision tree generally showed slight improvements in Mean Squared Error (MSE) and R-squared values, indicating better predictive accuracy and model fit. In classification tasks, pruning resulted in modest gains in Accuracy, Precision, Recall, and F1 Score, suggesting improved balance and robustness in the model's predictions. These results highlight the potential of pruning to enhance decision tree models, though the gains are often modest and highly dependent on the specific context.

Overall, this investigation underscores the importance of reviewing model performance using multiple metrics to gain a comprehensive understanding of the model's strengths and weaknesses. In a previous project, we only utilized one metric, causing us to assume poor performance of the model when it was actually performing fairly well. By considering a broader range of metrics, we can obtain a more nuanced view of model performance, allowing for better-informed decisions about model optimization.

Future work on this project could include exploring more advanced pruning techniques and applying decision trees to a broader range of datasets. Additionally, investigating the impact of different preprocessing methods on model performance could provide further insights into optimizing decision tree algorithms for various machine learning tasks. Another important avenue for future research is to refine the pruning algorithms to be more adaptive to the data characteristics, thereby enhancing their effectiveness across different scenarios.

References

[1] Marine Research Laboratories, Tasmania, 1995. Abalone Dataset `https://archive.ics.uci.edu/ml/datasets/Abalone`

[2] Wolberg,WIlliam. (1992). Breast Cancer Wisconsin (Original). UCI Machine Learning Repository. `https://doi.org/10.24432/C5HP4Z`

[3] Jozef Stefan Institute, Yugoslavia (Slovenia), 1988. Car Evaluation Dataset `https://archive.ics.uci.edu/ml/datasets/Car+Evaluation`

[4] University of California, Irvine, 1987, Congressional Voting Records Dataset `https://archive.ics.uci.edu/ml/datasets/Congressional+Voting+Records`

[5] Tel Aviv University, Israel, 1987. Computer Hardware Dataset `https://archive.ics.uci.edu/ml/datasets/Computer+Hardware`

[6] University of Minho, Portugal, 2007. Forest Fires Dataset `https://archive.ics.uci.edu/ml/datasets/Forest+Fires`

[7] Quinlan, J. R. (1986). ”Induction of decision trees.” Machine learning, 1(1), 81-106. `https://link.springer.com/article/10.1007/BF00116251`

[8] Esposito, F., Malerba, D., & Semeraro, G. (1997). ”A comparative analysis of methods for pruning decision trees.” IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(5), 476-491. `https://ieeexplore.ieee.org/document/589207`