*FUDMA Journal of Sciences (FJS)*
*ISSN online: 2616-1370*
*ISSN print: 2645 - 2944*
*Vol. 9 No. 3, March, 2025, pp 259 - 267*
*DOI: https://doi.org/10.33003/fjs-2025-0903-3321*

# AN ENHANCED CLASSIFICATION AND REGRESSION TREE ALGORITHM USING GINI EXPONENTIAL

***[1]Safinatu Bello, [2]Ahmad Abubakar Aliyu, [2]Muhammad Aminu Ahmad, [2]Adamu Abdullahi, [1]Sa'adatu Abdulkadir, [1]Abubakar Muazu Ahmed and [1]Suleiman Dauda**

[1]Department of Informatics, Kaduna State University, Kaduna – Nigeria.
[2]Department of Secure Computing, Kaduna State University, Kaduna – Nigeria.

*Corresponding authors' email: finabe210@gmail.com

## ABSTRACT

Decision tree algorithms, particularly Classification and Regression Trees (CART), are widely used in machine learning for their simplicity, interpretability, and ability to handle both categorical and numerical data. However, traditional decision trees often encounter limitations when dealing with complex, high-dimensional, or imbalanced datasets, as conventional impurity measures such as the Gini Index and Information Gain may fail to capture subtle variations in the data effectively. This study enhances the traditional Classification and Regression Trees (CART) model by introducing the Gini Exponential Criterion, which incorporates an exponential weighting factor into the split point calculation process. This novel approach amplifies the influence of highly discriminative features, resulting in more refined splits and improved decision boundaries. The enhanced CART model was evaluated on two benchmark datasets: the wine quality dataset and the hypothyroid dataset, with preprocessing steps like feature scaling and SMOTE for class imbalance, and hyperparameter tuning via Bayesian Optimization. On the wine quality dataset, the enhanced model improved accuracy from 57% (traditional CART) to 86%, while on the hypothyroid dataset, it achieved an impressive accuracy of 98%. These results highlight the model's ability to handle complex and imbalanced data effectively. Feature importance analysis and decision tree visualization further demonstrated the model's interpretability. The study concludes that the Gini Exponential Criterion significantly improves CART's performance, offering better generalization and clearer decision boundaries. This advancement is particularly valuable for applications requiring precise and interpretable predictions, such as healthcare diagnostics and quality assessment. Future work could explore integrating this criterion into ensemble methods and testing its scalability on larger datasets.

**Keywords**: Gini index, Information gain, Decision Tree, Classification, Regression Tree

## INTRODUCTION

Classification and Regression Trees (CART) are foundational predictive models in machine learning, widely used for both classification and regression tasks. These models employ a binary tree structure, recursively partitioning the input space to predict the target variable with high accuracy while maintaining interpretability. Their versatility has led to applications in diverse fields such as healthcare, finance, cybersecurity, and environmental sciences. Despite their popularity, traditional decision trees face challenges, including overfitting, computational inefficiency, and suboptimal performance in high-dimensional or imbalanced datasets. These limitations are often tied to the choice of impurity measures, such as the Gini index and Information Gain, which guide split decisions during tree construction.

While decision trees as shown in Figure 1, are valued for their interpretability and robustness, their performance is heavily influenced by the impurity measures used to determine optimal splits. Previous research, such as Tangirala (2020), has compared the Gini index and Information Gain, found comparable performance but highlighting limitations in handling complex datasets. Existing studies have primarily focused on conventional impurity measures without exploring algorithmic modifications that could enhance split decisions. This gap underscores the need for innovative approaches to improve the discriminative power of decision boundaries, particularly in scenarios involving noisy data or complex decision spaces.

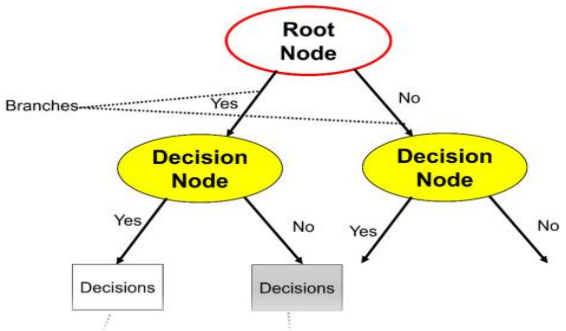


Figure 1: Visual representation of decision tree (Hasija, 2023)

To address these challenges, this study proposes an enhanced CART model that introduces a novel modification to the split point calculation. The proposed "Gini Exponential" approach integrates an exponential weighting factor into the split criterion, amplifying the influence of highly discriminative features. This modification aims to capture subtle data variations that traditional methods might overlook, leading to more refined decision boundaries and improved accuracy. The effectiveness of the Gini Exponential-based decision tree will be validated through comprehensive experiments on benchmark datasets for classification and regression tasks. Performance will be evaluated using metrics such as accuracy, precision, recall, F1-score, and mean squared error, alongside an analysis of tree depth, model complexity, and interpretability. This employed primarily in the CART (Classification and Regression Tree) algorithm, Gini impurity assesses the degree of disorder or impurity in a dataset (Bouke et al., 2023; Mustafa et al., 2024). It calculates the probability of incorrectly classifying a randomly chosen element if it were labeled according to the distribution of labels in the dataset (Northcutt et al., 2021). The Gini impurity for a node is defined as:

$$Gini\ Impurity = \ 1 - \sum_{i=1}^{n} P_i^2 \qquad (1)$$

where $p_i$ represents the proportion of instances belonging to class iii in the node. A Gini impurity of 0 indicates perfect purity (all instances belong to a single class), while higher values signify greater impurity. When constructing a decision tree, the algorithm evaluates potential splits by calculating the Gini impurity of the child nodes and selects the split that results in the lowest weighted average impurity, thereby increasing the overall purity of the subsets

However, utilizing information gain in algorithms like ID3 and C4.5, is based on the concept of entropy from information theory (Reddy and Chittineni, 2021). Entropy measures the amount of uncertainty or disorder within a set of data (Juszczuk et al., 2021). The entropy E of a node is calculated as:

$$E \ = \ - \sum_{i=1}^{n} p_i \ \log_2(P_i) \qquad (2)$$

where $p_i$ is the proportion of instances of class i in the node. Information Gain evaluates the reduction in entropy achieved by partitioning the dataset based on a particular attribute. The split that provides the highest Information Gain is chosen, as it most effectively reduces uncertainty in the dataset.

These traditional splitting criteria have been instrumental in the development of decision tree algorithms, providing systematic methods for data partitioning (Bittencourt et al., 2024). However, they also have limitations. For instance, Information Gain can be biased towards attributes with a large number of distinct values, potentially leading to overfitting. To mitigate this, the Gain Ratio was introduced in the C4.5 algorithm, which adjusts Information Gain by considering the intrinsic information of a split (Lestari, 2020). Despite these advancements, challenges such as handling continuous attributes, managing missing values, and preventing overfitting persist, prompting ongoing research into alternative splitting criteria and methods to enhance decision tree performance (Sharief et al., 2024).

The proposed Gini Exponential approach represents a significant advancement in decision tree algorithms, offering a balance between accuracy and interpretability. By enhancing the discriminative power of decision boundaries, this method addresses the limitations of traditional models and provides a promising solution for real-world applications requiring transparency, such as medical diagnosis, fraud detection, and financial risk assessment. Furthermore, this research contributes to the broader machine learning landscape by improving the foundational building blocks of ensemble methods like random forests and gradient boosting. From an ethical standpoint, the enhanced algorithm has the potential to deliver fairer and more reliable predictions in sensitive domains, ultimately supporting better decision-making and reducing risks such as misdiagnosis or false positives. Through rigorous experimentation, this study aims to demonstrate the superior predictive performance and practical applicability of the Gini Exponential approach.

## MATERIALS AND METHODS

The proposed methodology introduces an enhanced version of the Classification and Regression Trees (CART) algorithm by incorporating an exponential weighting factor into the split point calculation. Traditional CART models rely on impurity measures like the Gini index or Information Gain, which often yield similar performance levels regardless of dataset balance. To address this limitation, the study proposes the Gini Exponential criterion, which modifies the traditional Gini Index with an exponential function. This adaptive weighting assigns higher penalties to high-impurity nodes, prioritizing splits with greater class separation and improving classification accuracy, particularly in complex, high-dimensional, or noisy datasets.

The methodology is applied to a wine quality classification task, leveraging advanced techniques such as Bayesian Optimization, Synthetic Minority Over-sampling Technique (SMOTE) for handling class imbalance, and Feature Importance Analysis for interpretability. The dataset, consisting of physicochemical attributes of red wine, undergoes preprocessing steps including feature scaling using StandardScaler to normalize values and SMOTE to address class imbalance by generating synthetic data points for underrepresented classes. This ensures a balanced dataset, enabling the model to learn patterns from all quality levels without bias toward the majority class.

The core innovation lies in integrating the Gini Exponential criterion into the Gradient Boosting Classifier (GBC) for node splitting. Unlike the traditional Gini Index, which measures node impurity, the Gini Exponential emphasizes larger class separations through an exponential weighting factor. This modification enhances feature selection, sharpens decision boundaries, and improves the model's ability to distinguish between wine quality levels. By combining these advancements, the proposed methodology aims to achieve better generalization and classification accuracy, demonstrating the potential of the Gini Exponential criterion in enhancing decision tree algorithms for real-world applications.

**Algorithm Implementation**

```
1.  FUNCTION Main():
2.      CALL LoadDataset()
3.      CALL PreprocessData()
4.      CALL HandleClassImbalance()
5.      CALL SplitData()
6.      best_params ← CALL OptimizeHyperparameters()
7.      best_model ← TRAIN_MODEL(best_params)
8.      CALL EvaluateModel(best_model)
9.      CALL VisualizeResults(best_model)
10.     END FUNCTION
11. FUNCTION LoadDataset():
12.     data_path ← "C:/Users/CSE/Desktop/All_Thesis/Safeenahb_KASU_Thesis/winequality-red.csv"
13.     wine  df ← READ  CSV(data  path, delimiter=',')
14.     wine_df.columns ← STRIP_WHITESPACE(wine_df.columns)
15.     X ← DROP_COLUMN(wine_df, 'quality')  // Features
16.     y ← SELECT  COLUMN(wine  df, 'quality')  // Target variable
17.     PRINT("Class Distribution:\n", COUNT_VALUES(y))
18.     RETURN X, y
19. FUNCTION PreprocessData(X):
20.     scaler ← NEW StandardScaler()
21.     X_scaled ← scaler.FIT_TRANSFORM(X)
22.     RETURN X_scaled
23. FUNCTION HandleClassImbalance(X, y):
24.     smote ← NEW SMOTE(random_state=42)
25.     X_resampled, y_resampled ← smote.FIT_RESAMPLE(X, y)
26.     RETURN X_resampled, y_resampled
27. FUNCTION SplitData(X, y):
28.     (X_train, X_test, y_train, y_test) ← train_test_split(X, y, test_size=0.3, random_state=42)
29.     RETURN X_train, X_test, y_train, y_test
30. FUNCTION OptimizeHyperparameters():
31.     study ← NEW OptunaStudy(direction='maximize')
32.     FOR trial IN range(50):  // Perform 50 optimization trials
33.         params ← SUGGEST_HYPERPARAMETERS(trial)
34.         model ← NEW GradientBoostingClassifier(params, random_state=42)
35.         cv ← NEW StratifiedKFold(n  splits=5, shuffle=True, random  state=42)
36.         scores ← CROSS_VAL_SCORE(model, X_train, y_train, cv=cv, scoring='accuracy', n_jobs=-1)
37.         mean_score ← MEAN(scores)
38.         study.REPORT(mean_score, trial)
39.     best_params ← study.BEST_PARAMS
40.     PRINT("Best Parameters:", best_params)
41.     RETURN best_params
42. FUNCTION TRAIN_MODEL(best_params):
43.     best_model ← NEW GradientBoostingClassifier(best_params, random_state=42)
44.     best_model.FIT(X_train, y_train)
45.     RETURN best_model
46. FUNCTION EvaluateModel(model):
47.     y_pred ← model.PREDICT(X_test)
48.     accuracy ← CALCULATE_ACCURACY(y_test, y_pred)
49.     PRINT(f"Accuracy: {accuracy:.2f}\n")
50.     PRINT("Classification Report:")
51.     PRINT(GENERATE_CLASSIFICATION_REPORT(y_test, y_pred))
52.     cm ← GENERATE_CONFUSION_MATRIX(y_test, y_pred)
53.     CALL PlotConfusionMatrix(cm)
54. FUNCTION VisualizeResults(model):
55.     feature  importance ← EXTRACT  FEATURE  IMPORTANCE(model)
56.     CALL PlotFeatureImportance(feature_importance)
57.     CALL PlotDecisionTree(model)
58. FUNCTION PlotConfusionMatrix(cm):
59.                 PLOT_HEATMAP(cm,  annot=True,  fmt='d',  cmap='Blues',  xticklabels=SORTED_CLASSES(y),
yticklabels=SORTED_CLASSES(y))
60.     SET_TITLE('Confusion Matrix')
61.     LABEL_AXES('Predicted', 'Actual')
62. FUNCTION PlotFeatureImportance(importance):
63.     PLOT_BAR_CHART(importance, X.columns)
64.     SET_TITLE('Feature Importance')
65. FUNCTION PlotDecisionTree(model):
66.     first  tree ← model.ESTIMATORS  [0, 0]
67.    PLOT_TREE(first_tree, filled=True, feature_names=X.columns, class_names=SORTED_CLASSES(y), max_depth=2,
rounded=True)
68.     SET_TITLE('Decision Tree Structure (First Tree in Gradient Boosting Ensemble)')
```

**Dataset Source**

This work embraced two datasets for the evaluation of the Enhanced CART model. The Red Wine Quality dataset, obtained from the UC Irvine Machine Learning Repository, is a widely used benchmark for classification tasks in machine learning research. It consists of physicochemical attributes of red wines, such as fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, and alcohol content, along with a quality rating assigned by wine experts. The quality ratings range from 0 to 10, though the dataset primarily includes ratings between 3 and 8, representing the ordered classes of wine quality (Awujoola et al.,2024).

A notable characteristic of this dataset is its class imbalance, where the majority of samples belong to the "normal" quality category, while excellent and poor quality wines are underrepresented. This imbalance poses challenges for classification models, as they may struggle to accurately predict minority classes. To address this issue, techniques such as SMOTE (Synthetic Minority Over-sampling Technique) or outlier detection algorithms can be employed to enhance the representation of rare classes, enabling more robust learning and improved model performance.

The dataset is freely available for download at https://archive.ics.uci.edu/dataset/186/wine+quality and serves as a valuable resource for exploring predictive modeling, feature importance analysis, and handling imbalanced data in real-world applications. Its rich set of features and ordered class structure make it an ideal choice for evaluating machine learning algorithms in regression and classification tasks.

**Performance Evaluation Metrics**

Evaluating the performance of the enhanced Classification and Regression Tree (CART) model, which incorporates an exponential weighting factor into the split point calculation, requires the application of specific metrics tailored to both classification and regression tasks.

For classification tasks, accuracy is determined by the ratio of correctly predicted instances to the total number of instances (Awujoola et al., 2021).

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (3)$$

Precision assesses the proportion of true positive predictions among all positive predictions made by the model, calculated as the number of true positives divided by the sum of true positives and false positives.

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

Recall, or sensitivity, measures the proportion of actual positives correctly identified by the model, computed as the number of true positives divided by the sum of true positives and false negatives.

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

The F1 score provides a harmonic mean of precision and recall, offering a balance between the two metrics, and is calculated as twice the product of precision and recall divided by their sum.

$$F1 = 2\,x\,\frac{Precision\ x\ Recall}{Precision + Recall} \quad (6)$$

**RESULTS AND DISCUSSION**

The results, derived from experiments on benchmark datasets compared the performance of the enhanced model against traditional CART model. Initial experiments focused on the wine quality dataset, followed by additional tests on the hypothyroid dataset to evaluate the model's generalization capability across diverse data characteristics.

**Traditional CART on Wine Dataset**

A detailed classification report (Table 1) illustrates key performance metrics, including accuracy, precision, recall, and F1-score. The confusion matrix (Figure 2) provides insights into the model's classification performance across different classes. These results serve as a baseline for comparing the enhanced CART model, highlighting the strengths and limitations of the traditional approach in handling the complexities of the wine quality dataset.

**Table 1: Classification Report for the Traditional CART**

| Class | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| 0 | 0.66 | 0.65 | 0.65 | 195 |
| 1 | 0.57 | 0.6 | 0.59 | 200 |
| 2 | 0.46 | 0.43 | 0.44 | 61 |
| 3 | 0.07 | 0.06 | 0.06 | 17 |
| 4 | 0.33 | 0.17 | 0.22 | 6 |
| 5 | 0 | 0 | 0 | 1 |
| Accuracy | - | - | 0.57 | 480 |
| Macro Avg | 0.35 | 0.32 | 0.33 | 480 |
| Weighted Avg | 0.57 | 0.57 | 0.57 | 480 |

The evaluation of the traditional CART model on the wine quality dataset reveals moderate overall accuracy (0.57), but significant performance variations across classes. The model performs well for majority classes (Class 0 and Class 1), achieving precision, recall, and F1-scores of 0.66/0.65/0.65 and 0.57/0.60/0.59, respectively, likely due to their larger sample sizes. However, performance declines for smaller classes: Class 2 shows lower metrics (0.46/0.43/0.44), while Class 3 performs poorly (0.07/0.06/0.06) due to its extremely small sample size (17 instances). Classes 4 and 5, the smallest subsets, exhibit further deterioration, with Class 4 achieving an F1-score of 0.22 and Class 5 scoring zero across all metrics, reflecting the model's inability to predict rare classes. The macro averages (precision: 0.35, recall: 0.32, F1-score: 0.33) highlight suboptimal performance when considering all classes equally, while the weighted averages align with the overall accuracy (0.57), indicating the model's reliance on majority classes. In summary, the traditional CART model performs reasonably for dominant classes but struggles with imbalanced datasets and rare classes. This underscores the need for enhancements, such as modified split criteria or resampling techniques, to improve predictive accuracy for minority classes and address dataset imbalance.

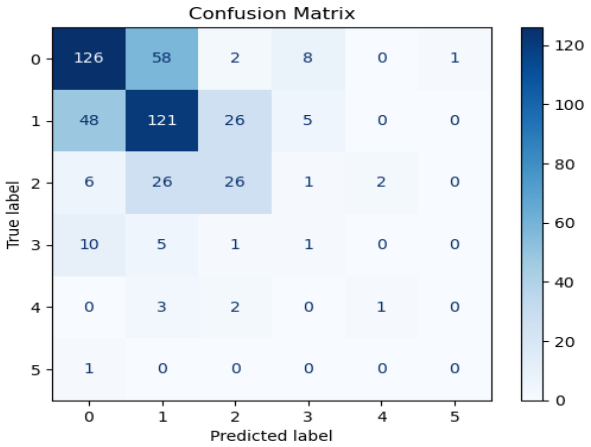


Figure 2: Confusion Matrix from the Traditional CART

The confusion matrix (Figure 2) provides a detailed breakdown of the traditional CART model's performance on the wine quality dataset. For Class 0, the model correctly classifies 126 instances, but misclassifies 58 as Class 1, 2 as Class 2, 8 as Class 3, and 1 as Class 5. For Class 1, it correctly identifies 121 instances, with misclassifications of 48 as Class 0, 26 as Class 2, and 5 as Class 3. Class 2 sees 26 correct classifications, but 6 are misclassified as Class 0, 26 as Class 1, 1 as Class 3, and 2 as Class 4.

Class 3 performs poorly, with only 1 correct classification and misclassifications of 10 as Class 0, 5 as Class 1, and 1 as Class 2. Class 4 achieves 1 correct classification, with 3 misclassified as Class 1 and 2 as Class 2. Class 5 has no correct classifications, with its single instance misclassified as Class 0.

These results highlight the model's strength in handling larger classes (e.g., Class 0 and Class 1) but significant struggles with smaller or rare classes (e.g., Class 3, Class 4, and Class 5). The confusion matrix underscores the need for techniques like resampling, cost-sensitive learning, or modified split criteria to address class imbalance and improve classification accuracy for minority classes.

**Enhanced CART on Wine Dataset**

This section evaluates the enhanced CART model on the wine quality dataset, highlighting its improved performance over the traditional approach. Table 2 provides a classification report summarizing key metrics (accuracy, precision, recall, and F1-score), demonstrating the model's enhanced effectiveness. Figure 3 visualizes the confusion matrix, revealing better classification performance across classes, particularly for smaller or less frequent ones.

The results indicate that the enhanced CART model addresses the limitations of the traditional approach by improving split point calculations, leading to better performance for minority classes while maintaining or enhancing overall accuracy. These findings underscore the value of the proposed enhancements, setting the stage for a detailed comparison between the two models and highlighting their potential in real-world applications.

**Table 2: Classification Report for the Enhanced CART**

| Class | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| 3 | 0.98 | 0.98 | 0.98 | 200 |
| 4 | 0.91 | 0.97 | 0.94 | 197 |
| 5 | 0.7 | 0.76 | 0.73 | 218 |
| 6 | 0.72 | 0.54 | 0.62 | 220 |
| 7 | 0.86 | 0.95 | 0.91 | 188 |
| 8 | 0.98 | 1 | 0.99 | 203 |
| Accuracy | - | - | 0.86 | 1226 |
| Macro Avg | 0.86 | 0.87 | 0.86 | 1226 |
| Weighted Avg | 0.86 | 0.86 | 0.86 | 1226 |

The evaluation of the enhanced CART model on the wine quality dataset shows significant improvements, achieving an overall accuracy of 0.86. Key performance metrics (precision, recall, F1-score) reveal strong results for most classes, with Classes 3 and 8 performing exceptionally well: Class 3 achieves precision: 0.98, recall: 0.98, F1-score: 0.98, while Class 8 attains precision: 0.98, recall: 1.00, F1-score: 0.99. Class 4 also performs well, with precision: 0.91, recall: 0.97, F1-score: 0.94, and Class 7 shows strong results with precision: 0.86, recall: 0.95, F1-score: 0.91.

For Classes 5 and 6, performance is slightly lower but still reasonable: Class 5 achieves precision: 0.70, recall: 0.76, F1-score: 0.73, while Class 6 attains precision: 0.72, recall: 0.54, F1-score: 0.62, reflecting challenges with overlapping or less distinct features.

The macro averages (precision: 0.86, recall: 0.87, F1-score: 0.86) and weighted averages (aligned with overall accuracy at 0.86) indicate consistent performance across all classes, with effective handling of both majority and smaller classes.

In summary, the enhanced CART model demonstrates robust performance, particularly for Classes 3, 4, 7, and 8, highlighting the effectiveness of the exponential weighting factor in refining split point calculations. While improvements are needed for some challenging classes (e.g., Class 6), the results underscore the model's potential to address dataset complexities and pave the way for further advancements in decision tree algorithms.

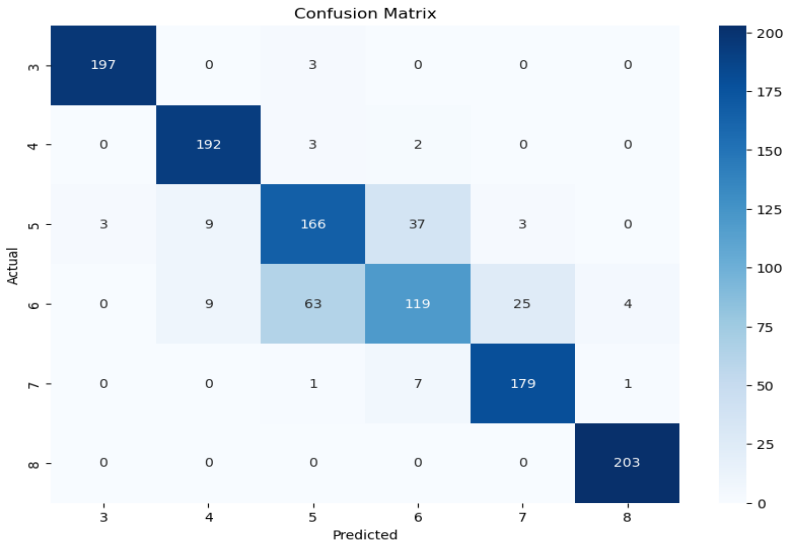


Figure 3: Confusion matrix obtained from the enhanced

The confusion matrix for the enhanced model on the Wine Quality dataset (Figure 4.4) reveals strong classification performance, with some misclassifications between adjacent classes. For Class 3, 197 instances were correctly classified, with 3 misclassified as Class 5. Class 4 saw 192 correct classifications, with 3 misclassified as Class 5 and 2 as Class 6. Class 5 had 166 correct classifications, but 9 were misclassified as Class 4, 3 as Class 3, and 37 as Class 6.

Class 6 experienced more misclassifications, with only 119 correct classifications, and errors including 9 as Class 4, 63 as Class 5, 25 as Class 7, and 4 as Class 8. Class 7 performed well, with 179 correct classifications and only 1 misclassified as Class 5 and 7 as Class 6. Class 8 achieved perfect accuracy, with all 203 instances correctly classified and no misclassifications.

The matrix highlights the model's strength in classifying most classes accurately but reveals challenges in distinguishing between Classes 5 and 6, likely due to overlapping feature distributions. Overall, the model demonstrates robust performance, with room for improvement in handling closely related classes

**The Hypothyroid Dataset for Traditional CART**

This section discusses the results obtained from the evaluation of the traditional CART model on the hypothyroid dataset. Table 3 presents the classification report from the experiment, while Figure 4 visualize the confusion matrix.

**Table 3: Classification report of hypothyroid Dataset on traditional CART**

| Class | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| 0 | 0.88 | 0.88 | 0.88 | 81 |
| 1 | 0.99 | 0.99 | 0.99 | 1051 |
| Accuracy | - | - | 0.98 | 1132 |
| Macro Avg | 0.93 | 0.93 | 0.93 | 1132 |
| Weighted Avg | 0.98 | 0.98 | 0.98 | 1132 |

The traditional CART model achieves strong performance on the hypothyroid dataset, with an overall accuracy of 0.98. For the minority class (Class 0, 81 instances), the model attains precision: 0.88, recall: 0.88, and F1-score: 0.88, indicating slight challenges in predicting this less frequent class due to dataset imbalance. In contrast, for the majority class (Class 1, 1051 instances), the model performs exceptionally well, achieving precision: 0.99, recall: 0.99, and F1-score: 0.99, reflecting near-perfect reliability in predicting this dominant class.

The macro averages (precision: 0.93, recall: 0.93, F1-score: 0.93) show balanced performance across both classes, while the weighted averages align with the overall accuracy at 0.98, emphasizing the model's strength in handling the majority class. Despite robust performance overall, the model's slightly lower effectiveness for the minority class highlights a common limitation of decision trees with imbalanced datasets. Future improvements could involve techniques like resampling or cost-sensitive learning to enhance prediction accuracy for rare classes. These results underscore the model's effectiveness in medical diagnostics while identifying areas for refinement.

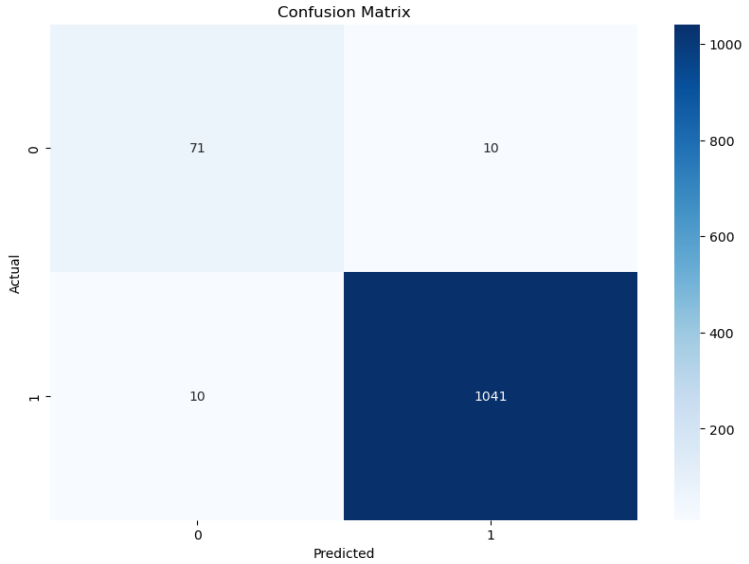


Figure 4: Confusion matrix for the hypothyroid with Traditional CART

The confusion matrix for the traditional CART model on the hypothyroid dataset reveals strong performance for the majority class (Class 1, 1041 instances), with 1041 correct classifications and only 1 misclassification as Class 0. However, for the minority class (Class 0, 71 instances), the model correctly identifies 61 instances but misclassifies 10 as Class 1, reflecting challenges due to dataset imbalance.

Overall, the model excels in predicting the majority class but struggles slightly with the minority class, highlighting a common issue with imbalanced datasets. To address this, techniques like resampling (e.g., SMOTE) or adjusting class weights could improve the model's ability to classify the minority class accurately. These enhancements would strengthen the model's robustness, particularly for critical applications like medical diagnostics.

**Enhanced CART for Hypothyroid Dataset**

This section evaluates the enhanced CART model on the hypothyroid dataset, presenting results through a classification report (Table 4) with key metrics like precision, recall, and F1-score. Figure 5 shows the confusion matrix, detailing correct and incorrect classifications across classes. Together, these results demonstrate the enhanced model's effectiveness in addressing the dataset's challenges.

**Table 4: Enhanced CART Results Discussion**

| Class | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| 0 | 0.98 | 0.99 | 0.98 | 1032 |
| 1 | 0.99 | 0.98 | 0.98 | 1057 |
| Accuracy | - | - | 0.98 | 2089 |
| Macro Avg | 0.98 | 0.98 | 0.98 | 2089 |
| Weighted Avg | 0.98 | 0.98 | 0.98 | 2089 |

The Enhanced CART model achieves an overall accuracy of 0.98 on the hypothyroid dataset, demonstrating excellent performance. For Class 0 (1032 instances), the model attains precision: 0.98, recall: 0.99, and F1-score: 0.98, indicating near-perfect identification with minimal errors. Similarly, for Class 1 (1057 instances), it achieves precision: 0.99, recall: 0.98, and F1-score: 0.98, showing exceptional performance with very few misclassifications.

The macro averages (precision: 0.98, recall: 0.98, F1-score: 0.98) and weighted averages (aligned with overall accuracy at 0.98) reflect consistent and balanced performance across both classes. While the results are highly promising, minor discrepancies in recall and precision suggest room for slight improvements. Overall, the model proves effective for medical diagnostics, particularly in datasets with balanced class distributions.

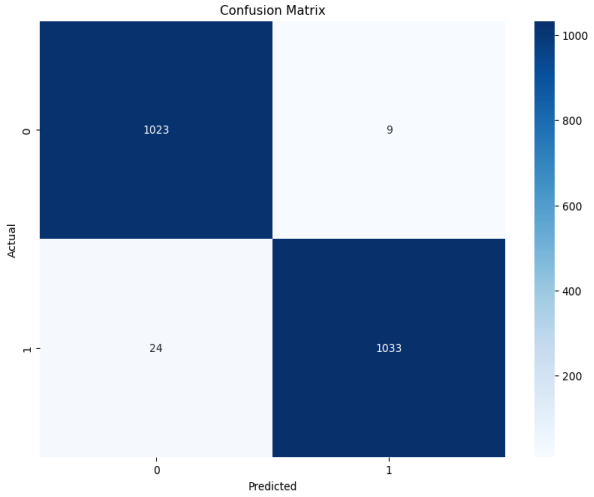


Figure 5: Confusion Matrix for Enhanced CART for hypothyroid

The Enhanced CART model demonstrates highly accurate performance on the hypothyroid dataset, as shown in Figure 4.7. For Class 0 (non-hypothyroid), the model correctly classifies 1,023 instances but misclassifies 9 as Class 1, indicating a strong true negative rate with minimal false positives. For Class 1 (hypothyroid), it correctly identifies 1,033 instances but misclassifies 24 as Class 0, reflecting a high true positive rate with some false negatives, which are critical in medical diagnostics.

The model achieves a well-balanced classification performance with a low error rate, effectively distinguishing between hypothyroid and non-hypothyroid cases. Future improvements could focus on reducing false negatives through feature engineering, hyperparameter tuning, or ensemble techniques to further enhance predictive accuracy.

## CONCLUSION

This study successfully addresses the limitations of traditional decision tree algorithms by introducing the Enhanced CART model, which integrates the Gini Exponential criterion. This innovation incorporates an exponential weighting factor into split point calculations, amplifying the influence of highly discriminative features and resulting in refined splits, sharper decision boundaries, and improved performance on complex, high-dimensional, or imbalanced datasets.

The model's effectiveness was validated on two benchmark datasets: the wine quality dataset, where accuracy improved from 57% (traditional CART) to 86% (Enhanced CART) and the hypothyroid dataset, achieving 98% accuracy and demonstrating robustness in handling class imbalance. The proposed model offers superior predictive performance and adaptability, supported by advanced techniques like Bayesian Optimization and SMOTE. Future research could explore applying the Gini Exponential criterion to ensemble methods like Random Forests or XGBoost, testing it on larger datasets, and refining the exponential weighting factor for broader adaptability. The study highlights the Enhanced CART model as a powerful, interpretable tool for classification tasks, with significant potential for practical applications requiring precise and reliable predictions.

## REFERENCES

Abedinia, A., & Seydi, V. (2024). Building semi-supervised decision trees with semi-cart algorithm. *International Journal of Machine Learning and Cybernetics*, 1-18. https://doi.org/10.1007/s13042-024-02161-z

Ali Fernando, W., Jollyta, D., Priyanto, D., & Oktarina, D. (2024). The Influence of Data Categorization and Attribute Instances Reduction Using The Gini Index On The Accuracy of The Classification Algorithm Model. *Jurna lIlmiah Kursor*, *12*(3), 111-122. https://doi.org/10.21107/kursor.v12i3.372

Ali, M. S. A. M., Zabidi, A., Tahir, N. M., Yassin, I. M., Eskandari, F., Saadon, A., ... & Ridzuan, A. R. (2024). Short-term Gini coefficient estimation using nonlinear autoregressive multilayer perceptron model. *Heliyon*, *10*(4). https://doi.org/10.1016/j.heliyon.2024.e26438

Altaf, I., Butt, M. A., & Zaman, M. (2022, June). Systematic consequence of different splitting indices on the classification performance of random decision forest. In *2022 2nd International Conference on Intelligent Technologies (CONIT)* (pp. 1-5). IEEE. http://dx.doi.org/10.1109/CONIT55038.2022.9848372

Awujoola, O. J., Ogwueleka, F. N., Irhebhude, M. E., & Misra, S. (2021). Wrapper based approach for network intrusion detection model with combination of dual filtering technique of resample and SMOTE. In *Artificial Intelligence for Cyber Security: Methods, Issues and Possible Horizons or Opportunities* (pp. 139-167). Cham: Springer International Publishing.

Awujoola, J. O., Enem, T. A., Ogwueleka, F. N., Abioye, O., & Adelegan, R. O. (2024). Machine Learning—Enabled Predictive Analytics for Quality Assurance in Industry 4.0 and Smart Manufacturing: A Case Study on Red and White Wine Quality Classification. In *Industry 4.0, Smart Manufacturing, and Industrial Engineering* (pp. 65-97). CRC Press.

Bittencourt, J. C. N., Costa, D. G., Portugal, P., &Vasques, F. (2024). Towards lightweight fire detection at the extreme edge based on decision trees. *In 2024 IEEE 22nd Mediterranean Electrotechnical Conference (MELECON) (pp. 873-878). IEEE.* http://doi.org/10.1109/melecon56669.2024.10608598

Bouke, M. A., Abdullah, A., Frnda, J., Cengiz, K., & Salah, B. (2023). BukaGini: A stability-aware Gini index feature selection algorithm for robust model performance. *IEEE*

*Access, 11*, 59386-59396. http://doi.org/10.1109/ACCESS.2023.3284975

Charbuty, B., & Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends, 2(01), 20-28*.

Disha, R. A., & Waheed, S. (2022). Performance analysis of machine learning models for intrusion detection system using Gini impurity-based weighted random forest (GIWRF) feature selection technique. *Cybersecurity, 5(1), 1*. https://doi.org/10.1186/s42400-021-00103-8

Fayaz, S. A., Zaman, M., & Butt, M. A. (2021). An application of logistic model tree (LMT) algorithm to ameliorate Prediction accuracy of meteorological data. *International Journal of Advanced Technology and Engineering Exploration*, *8*(84), 1424-1440. https://doi.org/10.19101/IJATEE.2021.874586.

Hasija, Y. (2023). *All about Bioinformatics: From Beginner to Expert*. Elsevier.

Iorliam, I. B., Ikyo, B. A., Iorliam, A., Okube, E. O., Kwaghtyo, K. D., & Shehu, Y. I. (2021). Application of machine learning techniques for Okra shelf-life prediction. *Journal of Data Analysis and Information Processing*, *9*(3), 136-150. https://doi.org/10.4236/jdaip.2021.93009

Juszczuk, P., Kozak, J., Dziczkowski, G., Głowania, S., Jach, T., &Probierz, B. (2021). Real-world data difficulty estimation with the use of entropy. *Entropy, 23(12), 1621*. https://doi.org/10.3390/e23121621

Lee, S., Lee, C., Mun, K. G., & Kim, D. (2022). Decision tree algorithm considering distances between classes. *IEEE Access*, *10*, 69750-69756. http://dx.doi.org/10.1109/ACCESS.2022.3187172

Lestari, A. (2020). Increasing accuracy of C4.5 algorithm using information gain ratio and AdaBoost for classification of chronic kidney disease. *Journal of Soft Computing Exploration*, *1(1), 32-38*. https://doi.org/10.52465/joscex.v1i1.6

Mustafa, O. M., Ahmed, O. M., & Saeed, V. A. (2024). Comparative analysis of decision tree algorithms using Gini and entropy criteria on the forest cover types dataset. *In The International Conference on Innovations in Computing Research (pp. 185-193). Cham: Springer Nature Switzerland*. https://doi.org/10.1007/978-3-031-65522-7_17

Mienye, I. D., Sun, Y., & Wang, Z. (2019). Prediction performance of improved decision tree-based algorithms: A review. *Procedia Manufacturing, 35, 698-703*. https://doi.org/10.1016/j.promfg.2019.06.011

Northcutt, C., Jiang, L., & Chuang, I. (2021). Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research, 70, 1373-1411*. https://doi.org/10.1613/jair.1.12125

Priyanka, & Kumar, D. (2020). Decision tree classifier: A detailed survey. *International Journal of Information and Decision Sciences, 12(3), 246-269*.

Rahmati, O., Avand, M., Yariyan, P., Tiefenbacher, J. P., Azareh, A., & Bui, D. T. (2022). Assessment of Gini-, entropy-and ratio-based classification trees for groundwater potential modelling and prediction. Geocarto International, 37(12), 3397-3415. http://dx.doi.org/10.1080/10106049.2020.1861664

Reddy, G. S., & Chittineni, S. (2021). Entropy-based C4.5-SHO algorithm with information gain optimization in data mining. *PeerJ Computer Science, 7, e424*. https://doi.org/10.7717/peerj-cs.424

Sharief, F., Ijaz, H., Shojafar, M., & Naeem, M. A. (2024). Multi-class imbalanced data handling with concept drift in fog computing: A taxonomy, review, and future directions. *ACM Computing Surveys, 57(1), 1-48*. http://dx.doi.org/10.1145/3689627

Tangirala, S. (2020). Evaluating the impact of Gini index and information gain on classification using decision tree classifier algorithm. *International Journal of Advanced Computer Science and Applications, 11(2), 612-619. http://dx.doi.org/10.14569/IJACSA.2020.0110277*

Yang, S., Li, N., Sun, D., Du, Q., & Liu, W. (2021). A differential privacy preserving algorithm for greedy decision tree. In *2021 2nd International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE)* (pp. 229-237). IEEE. http://dx.doi.org/10.1109/ICBASE53849.2021.00050