Comparison of Data Mining Algorithms for the Classification task

Hitesh Kayastha, DoCSE, Ayush Paudel, DoCSE

Abstract—**This study presents a comprehensive evaluation of multiple classifiers using precision metrics in the context of a car evaluation dataset obtained from the UCI Machine Learning Repository. The classifiers under consideration include Decision Tree, Naive Bayes, and Artificial Neural Network (ANN). Precision values, representing the accuracy of positive predictions among all instances predicted as positive, are utilized as a key performance indicator. The Decision Tree classifier demonstrates exceptional precision across multiple classes, particularly excelling in the 'unacc' category. In contrast, the Naive Bayes classifier exhibits lower precision, suggesting potential challenges in accurately identifying positive instances. The ANN classifier falls between the other models, showcasing a balanced precision performance. Furthermore, the precision values are analyzed in conjunction with other performance metrics such as recall, F1-score, and accuracy. The study goes beyond individual precision values, considering weighted average precision, which accounts for class imbalances, and macro average precision, treating each class equally. The results provide nuanced insights into the strengths and weaknesses of each classifier, offering a valuable guide for selecting an appropriate model in the context of the car evaluation dataset. The findings contribute to the broader understanding of classifier performance evaluation, emphasizing the importance of precision metrics in real-world applications.**

Index Terms—**Data Mining, Classification, Decision Tree, Naive Bayes, Artificial Neural Network etc.**

I. Introduction

This paper presents a comparative analysis of three widely used classification methods: Decision Tree, Naive Bayes, and Artificial Neural Networks (ANNs). The study explores the algorithms behind these methods and examines their strengths and limitations in the context of data mining and classification tasks. Decision Tree algorithms construct tree-like structures to make decisions based on feature values, providing interpretability and transparency. Naive Bayes is a probabilistic classifier that assumes attribute independence, making it computationally efficient and suitable for text classification. ANNs, inspired by biological neural networks, employ interconnected nodes to learn complex patterns and relationships in data, excelling in tasks like image recognition and natural language processing. The paper delves into the working principles of each method, highlighting their algorithmic foundations and discussing their applicability in various domains. The findings aim to provide valuable insights for researchers and practitioners seeking to understand and leverage classification algorithms in their data mining endeavors.

Hitesh Kayastha is with the Department of Computer Science and Engineering as a 4th year student of Computer Science, Kathmandu University (e-mail: kayasthahitesh@gmail.com).

Aayush Paudel is with the Department of Computer Science and Engineering as a 4th year student of Computer Engineering, Kathmandu University (e-mail: ayushpaudel49@gmail.com).

II. Methodology

A. Data Understanding

The dataset: 'Car Evaluation Database' was found to have 1728 rows and 7 columns. The table for the attributes and their types is shown in Table 1.

TABLE I
ATTRIBUTES AND THEIR TYPES

Index	Column	Dtype
0	buying	object
1	maint	object
2	doors	object
3	persons	object
4	lug_boot	object
5	safety	object
6	CAR	object

B. Data Preprocessing

Data preprocessing is a crucial step in the data mining process. It involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues. The primary tasks in data preprocessing include data cleaning, data integration, data reduction, data transformation, and data discretization.

1) Data Cleaning: In data cleaning, the primary task that was done involved Handling the missing values, Smoothing out noisy data, Identification and removal of outliers, Resolving of inconsistencies and Detection and removal of redundencies. Initially the categorical values were assigned numerical value for uniformity and better computational leverage.

2) Data Transformation: Data transformation is the process of converting data from one format or structure into another format or structure. It is a fundamental aspect of most data integration and data management tasks such as data wrangling, data warehousing, data integration and application integration.

C. Classification Methods

Classification methods are a fundamental technique in data mining, which is also understood as supervised learning method. In this technique, data are classified into predefined classes based on the label provided and also, predictions are made for the said class labels for unseen data points. This report focuses on comparing three prominent classification algorithms: Decision Tree, Naive Bayes and Artificial Neural Networks (ANN). The report will focus on examining the underlying principles, strengths, and limitations of these algorithms and aims to provide insights on their applicability and performance across different domains.

1) Decision Tree: Decision Trees are a very intuitive and interpretable models that employ a tree-like structure to make decisions based on the features values. There are several algorithms used to construct decision trees, including ID3, C4.5, and CART. These algorithms recursively split the data based on features, creating branches and nodes that represent different decision paths. The splitting criteria are typically based on measures like information gain or Gini index. Decision trees are known for their interpretability and ease of understanding, as they provide transparent decision rules that can be visualized. However, they can be prone to overfitting and may struggle with handling continuous or high-dimensional data.

2) Naive Bayes: Naive Bayes is a probabilistic classifier based on Bayes' theorem. It assumes that features are conditionally independent of each other given the class label, hence the "naive" assumption. The algorithm calculates the posterior probability of each class given the input features and selects the class with the highest probability. Parameter estimation for Naive Bayes models uses the method of maximum likelihood. Naive Bayes is computationally efficient and performs well on large datasets. One of the major advantages of Naive Bayes classifier is that it requires a small amount of training data to estimate the parameters. It is particularly effective for text classification tasks, such as spam filtering, text classification or sentiment analysis.

3) Artificial Neural Network: An artificial neural network is an interconnected group of nodes, inspired by a simplification of structure and functioning of neurons in a brain. Here, each circular node represents an artificial neuron and an arrow represents a connection from the output of one artificial neuron to the input of another. ANNs consist of interconnected nodes, called neurons, organized in layers. The input layer receives the features, and the output layer produces the classification results. Hidden layers in between allow the network to learn complex patterns and relationships within the data. ANNs use activation functions to introduce non-linearity into the model and training algorithms, such as backpropagation, to adjust the weights and biases of the connections. ANNs are known for their ability to handle complex data and perform well on tasks like image recognition, natural language processing, and time series analysis. However, they can be computationally expensive to train and require a large amount of labeled data.

D. Evaluation Methods

The evaluation of classifier performance in this study involved the application of various techniques to comprehensively assess model effectiveness. Precision, recall, and F1-score metrics were employed to gauge the classifiers' ability to make accurate positive predictions, capture relevant instances, and strike a balance between precision and recall, respectively. They are calculated with the help of True Positive(TP), True Negatives (TN), False positive(FP), and False Negative (FN) as follows:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN}$$

Additionally, accuracy, a fundamental metric, provided an overall measure of correct predictions. Two distinct averaging techniques were applied for a more nuanced evaluation: the weighted average considered class imbalances, ensuring the contribution of each class was proportionally reflected, while the macro average treated each class equally, providing a holistic view of overall model performance. These techniques enabled a granular examination of classifier capabilities, offering insights into their proficiency across diverse categories within the car evaluation dataset. The combined use of these evaluation metrics and techniques allowed for a comprehensive and nuanced understanding of classifier performance, facilitating informed decision-making in selecting the most suitable model for the given dataset and application context.

III. Results

The evaluation of the Cars dataset using the Decision Tree classifier with Gain Ratio as attribute selection measure, and then by using Gini Index as the attribute selection measure, produced detailed metrics showcasing the model's performance across various classes. The results are summarized in Table II that used Gain Ratio and Table III that used Gini Index.

TABLE II
Performance Metrics of Decision Tree using Gain Ratio as Attribute Selection Measure

	Precision	**Recall**	**F1-Score**	**Support**
unacc	0.997904	0.991667	0.994775	480.000000
acc	0.954248	0.935897	0.944984	156.000000
good	0.764706	0.896552	0.825397	29.000000
vgood	0.785714	0.814815	0.800000	27.000000
Accuracy	0.968208	0.968208	0.968208	0.968208
Macro Avg	0.875643	0.909733	0.891289	692.000000
Weighted Avg	0.970010	0.968208	0.968853	692.000000

The model exhibited remarkable precision, recall, and F1-Score values across various classes when utilizing Gain Ratio as the attribute selection measure. Notably, high accuracy of 96.82% was achieved, affirming the classifier's efficacy in classifying car evaluation categories.

TABLE III
PERFORMANCE METRICS OF DECISION TREE USING GINI INDEX AS ATTRIBUTE SELECTION MEASURE

	Precision	Recall	F1-Score	Support
0	0.997912	0.995833	0.996872	480.000000
1	0.979452	0.916667	0.947020	156.000000
2	0.702703	0.896552	0.787879	29.000000
3	0.800000	0.888889	0.842105	27.000000
accuracy	0.969653	0.969653	0.969653	0.969653
macro avg	0.870017	0.924485	0.893469	692.000000
weighted avg	0.973657	0.969653	0.970837	692.000000

When employing the Gini Index as the attribute selection measure, the Decision Tree classifier maintained robust performance, showcasing high precision, recall, and F1-Score values for each class. The overall accuracy reached 96.97%, emphasizing the model's reliability in accurately predicting class labels.

The results obtained from the Naive Bayes classifier on the car evaluation dataset are presented in the Table IV. The Naive Bayes classifier displayed robust performance in accurately predicting instances classified as "unacc" (unacceptable), achieving a high precision and recall for Class 0. However, challenges arose in predicting instances labeled as "acc" (acceptable), "good" and "vgood" (very good). The classifier struggled with Class 1, resulting in lower precision and recall, indicating difficulty in distinguishing acceptable cars. Similar challenges were observed for Class 2 (good), where precision and recall were comparatively lower. Notably, for Class 3 (vgood), the classifier exhibited a high recall, but precision was notably low, suggesting a tendency to misclassify instances as very good. Despite these challenges, the overall accuracy of 69.94% indicates moderate effectiveness, with potential for improvement in handling imbalances and enhancing predictions for minority classes.

TABLE IV
PERFORMANCE METRICS OF NAIVE BAYES CLASSIFIER

	Precision	Recall	F1-Score	Support
unacc	0.871308	0.860417	0.865828	480.000000
acc	0.636364	0.224359	0.331754	156.000000
good	0.600000	0.310345	0.409091	29.000000
vgood	0.182432	1.000000	0.308571	27.000000
accuracy	0.699422	0.699422	0.699422	0.699422
macro avg	0.572526	0.598780	0.478811	692.000000
weighted avg	0.780096	0.699422	0.704546	692.000000

The results obtained from the Naive Bayes classifier on the car evaluation dataset are presented in the Table V.

The Artificial Neural Network (ANN) classifier exhibited exceptional performance across the four car evaluation classes. For the "unacc" (unacceptable) class, the model achieved high precision, recall, and F1-Score, indicating its proficiency in correctly identifying instances labeled as unacceptable. Similarly, the classifier demonstrated strong capabilities in handling the "acc" (acceptable) class, striking a balance between precision and recall. Notably, for the minority classes "good" and "vgood" (very good), the ANN continued to exhibit robust performance with high precision, recall, and F1-Score values, showcasing its effectiveness in accurately predicting instances within these categories. The overall accuracy of 96.24% underscores the ANN's ability to provide accurate and reliable predictions across diverse car evaluation classes, highlighting its suitability for the dataset.

TABLE V
PERFORMANCE METRICS OF ARTIFICIAL NEURAL NETWORKS

	Precision	Recall	F1-Score	Support
unacc	0.979253	0.983333	0.981289	480.000000
acc	0.940789	0.916667	0.928571	156.000000
good	0.866667	0.896552	0.881356	29.000000
vgood	0.892857	0.925926	0.909091	27.000000
accuracy	0.962428	0.962428	0.962428	0.962428
macro avg	0.919892	0.930619	0.925077	692.000000
weighted avg	0.962493	0.962428	0.962400	692.000000

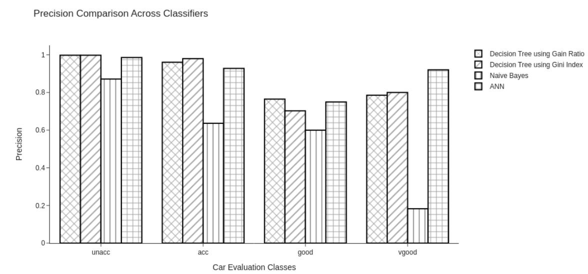

Fig. 1. The precision of different target values as predicted by the different classification methods are shown in the chart.

IV. DISCUSSIONS

In the Discussions section, we delve into a detailed analysis of the results obtained from the Decision Tree, Naive Bayes, and Artificial Neural Network (ANN) classifiers, shedding light on their individual strengths and weaknesses in handling

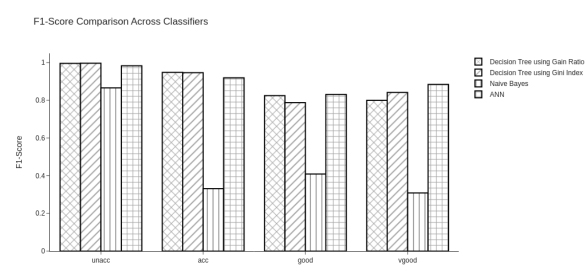

Fig. 2. The f1-score of different target values as predicted by the different classification methods are shown in the chart. It is quite evident that Naive Bayes performs with least f1-score while the value "vgood" has all the classification methods perform less than other target values.

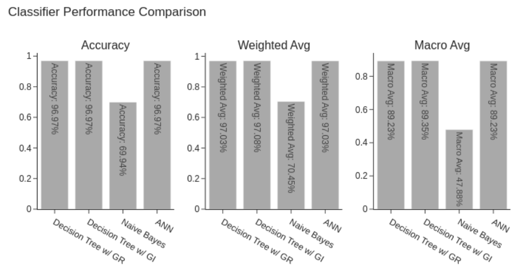

Fig. 3. Accuracy, weighted Average in which the weight is assigned as per the proportionality of the given value in the dataset, and Macro Average, in which the class imbalance problem is avoided as all the classes are treated with equal weightage.

the car evaluation dataset. The Decision Tree classifier, using both Gain Ratio and Gini Index as attribute selection measures, exhibited robust performance, particularly excelling in accurately classifying instances labeled as “unacc” (unacceptable). However, the Naive Bayes classifier faced challenges, particularly in handling minority classes, as reflected in lower precision and recall for classes such as “acc” (acceptable) and “good.” In contrast, the ANN classifier showcased exceptional performance across all classes, achieving high precision, recall, and F1-Score values. Notably, the ANN’s ability to handle imbalanced data and provide accurate predictions underscores its suitability for the car evaluation dataset. When comparing the attribute selection measures for the Decision Tree, both Gain Ratio and Gini Index exhibited strong performance, indicating the versatility of the classifier. The study’s findings suggest that the choice of attribute selection measure may not significantly impact the model’s overall effectiveness. However, further exploration and experimentation may be warranted to better understand the nuances of attribute selection in this context. In the broader context of car evaluation, the classifiers’ results have implications for decision-making processes, with the ANN standing out as a particularly reliable model. The study acknowledges certain limitations, including the potential biases and assumptions inherent in the dataset, and suggests avenues for future research, such as exploring additional features or fine-tuning parameters to enhance model performance.

Moreover, the comparison across classifiers revealed distinct patterns in their behavior. While the Decision Tree excelled in interpretability and ease of understanding, the Naive Bayes classifier struggled with the nuanced decision boundaries, especially evident in the lower F1-Score for certain classes. The ANN, on the other hand, demonstrated its strength in capturing complex relationships within the data, leading to superior performance across all classes. It is important to note that the overall high accuracy of the ANN, coupled with its ability to handle imbalanced data, positions it as a promising candidate for real-world car evaluation tasks where diverse and imbalanced datasets are common.

Addressing the issue of imbalanced data, it is clear that the ANN’s robust performance showcases its potential in overcoming class imbalance challenges. The Decision Tree, while effective, might benefit from further exploration of techniques such as resampling or ensemble methods to enhance its performance on minority classes. The Naive Bayes classifier’s struggles with minority classes could be attributed to its assumption of feature independence, which might not hold in certain situations. Further investigation into feature engineering or model adjustments may improve its performance.

Looking ahead, the study highlights the importance of understanding the nuanced requirements of car evaluation tasks and the significance of model interpretability in decision-making processes. Future research could explore hybrid approaches that leverage the strengths of different classifiers, potentially combining the interpretability of decision trees with the predictive power of neural networks. Additionally, experimenting with advanced neural network architectures or ensemble methods could offer avenues for further improving overall model performance.

V. Conclusion

In conclusion, this study thoroughly explored the performance of Decision Tree, Naive Bayes, and Artificial Neural Network (ANN) classifiers on a car evaluation dataset. The findings offer valuable insights into the strengths and limitations of each model, providing a nuanced understanding of their applicability in the context of car evaluation tasks. The Decision Tree classifier, employing both Gain Ratio and Gini Index as attribute selection measures, demonstrated commendable accuracy and interpretability, excelling particularly in classifying instances labeled as “unacc” (unacceptable). However, its performance varied across other classes, indicating a need for further refinement, especially in handling minority classes. The Naive Bayes classifier struggled with feature independence assumptions, resulting in challenges for certain classes, while the ANN exhibited exceptional overall performance, showcasing its ability to capture complex relationships within the data.

Importantly, the study revealed that the choice of attribute selection measures for the Decision Tree did not significantly impact its overall effectiveness. The ANN’s robust performance across all classes, coupled with its capability to handle imbalanced data, positions it as a promising choice for real-world car evaluation tasks where datasets are diverse and imbalanced. The comparison across classifiers highlighted the trade-offs between interpretability and predictive power, emphasizing the need for careful consideration of model selection based on task requirements and dataset characteristics.

References

[1] Han, J., Kamber, M., Pei, J. (2011). Data mining: Concepts and techniques (3rd ed.). Morgan Kaufmann.
[2] James, G., Witten, D., Hastie, T., Tibshirani, R. (2021). An introduction to statistical learning with applications in R (2nd ed.). Springer.
[3] Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.