Performance Comparison of Data Mining Algorithms: A Case Study on Car Evaluation Dataset

Jamilu Awwalu[#1], Anahita Ghazvini[#2], and Azuraliza Abu Bakar [*3]

[#1,2]Postgraduate Students at Centre for Artificial Intelligence and Technology (CAIT)

[*3]Professor at Centre for Artificial Intelligence and Technology (CAIT)

Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia (UKM) 43600, Bangi Selangor, MALAYSIA.

Abstract*— Cars are essentially part of our everyday lives. There are different types of cars as produced by different manufacturers; therefore the buyer has a choice to make. The choice buyers or drivers have mostly depends on the price, safety, and how luxurious or spacious the car is. Data mining tasks in terms of classification or prediction are applied in a variety of domains which includes manufacturing and business. But the choice of algorithm can be confusing because some algorithms are argued to have better performance record than others, depending on the associated task and nature of dataset. This study analyzes the performance of three data mining algorithms in terms of speed and accuracy on the car evaluation dataset obtained from the University of California Irvine (UCI) dataset.

***Keywords*— Data Mining, Decision Tree, Neural Network, Naive Bayesian**

I. INTRODUCTION

Safety, cost, and luxury are important factors to consider in buying cars. These factors vary based on type, model, and manufacturer of the car. However, these factors are so crucial in aspect like accident number reduction. Standard equipments are part of the factors to consider when buying a car. Standard equipments include conveniences, performance enhancers, and safety equipment. Safety as mentioned in the factors, is really indispensible, also as much as conveniences which in the case of this study falls under the attributes; door, maintenance, and luggage boot.

Cost consideration as stated by [1] is crucial to ensure the car bought is worth what it costs, because buying a car is a huge step towards independence, but independence comes with responsibilities. To succeed it is important to understand the true financial responsibility that comes with owning a car. The study uses the attribute 'buying', which means the buying cost of a car to determine its acceptability or not based on its cost in relation the other important attributes which are; maintenance, doors, persons, lug_boot, and safety.

Data mining is a branch of Artificial Intelligence that is applied in a variety of domains nowadays. These domains includes but are not limited to Medical, Manufacturing, Education, and Business. The application of data mining techniques in any domain mainly employs algorithms such as Artificial Neural Network, Naive Bayes, Support Vector Machines, and other Machine Learning algorithms that are linked to data mining in either classification, clustering, association rules mining, sequence and pattern mining, or prediction tasks.

II. RELATED WORK

A study conducted by [2] on employing neural network and naive Bayesian classifier in data mining for car evaluation to investigate the performance of Bayesian Neural Network and Naive Bayesian classification methods using the car evaluation dataset. Findings from the study proved the researchers assumption that Bayesian Neural Network (BNN) is slower, ambiguous, and more difficult to manipulate than naive Bayesian (NB). However, BNN shows an amazing percentage of accuracy on the dataset.

Artificial Neural Networks (ANN) an a classification algorithm that is widely used in data mining was used in a study conducted by [3] to compare the performance of Decision Tree and ANN to develop prediction models; and the comparative study of Bayesian and ANN classifiers on motion picture [4]. Also, [5] conducted a study on evaluation of an on-vehicle adaptive tourist service. In the study they described the methodology and results obtained in evaluation of a system that provides personalised tourist information onboard cars. With a simulator and using layered sampling strategy and statistics metrics to compare the system suggestions to the user's answers. Also, they analysed several dimensions of adaptation. The car dataset used for this study as obtained from the University of California Irvine (UCI) dataset repository was used by [6] on modelling performance of different classification methods.

III. DATASET DESCRIPTION

The dataset used in this study which is a collection of the records on specific attributes on cars donated by Marco Bohanec in 1997 was obtained from the UCI dataset repository. The car evaluation dataset as described in the UCI dataset repository was derived from simple hierarchical decision, and is categorized descriptively in table 1.

TABLE 1 CAR EVALUATION DATASET

Data Set Characteristics:	Multivariate	Number of Instances:	1728
Attribute Characteristics:	Categorical	Number of Attributes:	6
Associated Tasks:	Classification	Missing Values?	No

The class attributes in the Car evaluation dataset are:

- Acceptable: This is denoted as ‘acc’
- Good: This is denoted as ‘good’
- Unacceptable: This is denoted as ‘unacc’
- Very Good: This denoted as ‘vgood’

A standard data analysis was done on the dataset to identify some patterns in the data and also present the data in tables based on attribute range and their frequencies. The output from the data analysis shown in Table 2 and Figure 1 describes the distribution of the four class attributes in the dataset.

TABLE 2 FREQUENCY OF CLASS OUTPUT FROM THE DATASET

Class	**Frequency**	**Relative Frequency in %**
Acc	385	22.28009259
Good	70	4.050925926
Unacc	1207	69.84953704
Vgood	66	3.819444444
Total	1728	100

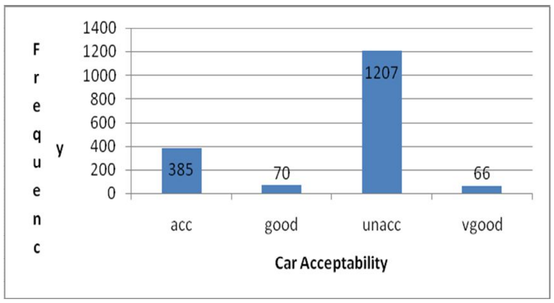

Fig. 1 Frequency of Class Output from the Dataset

Table 2 and Figure 1 show the frequency of the class output which is the final outcome from the dataset. It shows that out of the total 1728 cars in the dataset, 385(22.28 %) were acceptable, 70 (4.05 %) were good, 1207 (69.85 %) were unacceptable, and 66 cars (3.82%) were very good. From the above we can conclude that more than half of the cars evaluated were not of acceptable.

IV. CLASSIFICATION METHODS

The Naive Bayesian algorithm, named after Thomas Bayes (1702 – 1761) is a learning algorithm as well as a statistical method for classification. It captures uncertainty in a principled way by using probabilistic approach. Naive Bayesian classification provides practical learning algorithms and prior knowledge and observed data can be combined.

The Artificial Neural Network (ANN) algorithm takes data as input then process and generalizes output using biological brain patterns of humans or animals. It is designed to learn in a non linear mapping between input and output data.

Decision tree builds classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes.

V. EXPERIMENT

The experiment was carried out using three classifier models, namely; decision tree, neural network, and naive Bayesian classifiers. This is in view to finding out which of the classifier best suits the dataset in terms of classifying the pre-processed data, trained data, testing, and making prediction using the model obtained from the training process. The detailed procedure of the experimentation is as follows:

A. Data Cleaning

The data as obtained from the UCI dataset repository have to be cleaned and to ensure that it is in the standard quality before the model creation is initiated. The data cleaning conducted on the dataset as shown in Table 3 is the conversion of nominal attributes to numeric attributes. The nominal to numeric conversion process was conducted in order to make the process of normalization possible.

TABLE 3 NOMINAL TO NUMERIC CONVERSION

Attribute	**Nominal**	**New Numeric Value**
Buying	vhigh	4
	High	3
	Med	2
	Low	1
Maintenance	vhigh	4
	High	3
	Med	2
	Low	1
Luggage Boot	Small	3
	Med	2
	Big	1
Safety	Low	1
	Med	2
	High	3

B. Data Transformation

Data transformation is a very crucial process in data pre-processing. It involves normalization and aggregation. Normalization is a process of scaling the value of data to specific rate. Normalization can be done using the min-max or the z-score methodology. For this study, the min-max normalization technique is used to normalize the dataset. As a principle, the min-max normalization result always ranges between 0 and 1.

C. Data Set Split

The pre-processed dataset was split into two halves of varying sizes at different times for use as training and testing data set across the different data mining classification algorithms for model creation and observation of which of the models performs best.

1) Training and Testing

The data set used for training is mainly a portion from the dataset from which the classifying algorithm used learns the class/result of the model created from each model, and the four splits used in this study are shown in table 4. The learning method is based on the attributes or features of the dataset in comparison the result/class. And finally the output is a model used to compare against the other half of the dataset, which is the testing data.

TABLE 4 CAR DATASET SPLIT FOR MODEL CREATION

Training and Testing % Split
90% 10%
66% 44%
50% 50%
10 Folds

2) Classification

The classification and the model creation were done using the following three data mining classifiers from WEKA:

- J48: This is a type of decision tree classifier.
- Multilayer Perceptron: This is a type of Artificial Neural network classifier
- Naive Bayesian
- 10-Folds Cross Validation

3) Application of Class Association Rules (CAR)

The association rule and model creation was done using the Apriori type algorithm. This was done in order to get the best attributes association rules for each class in the car dataset. The experiment on this was conducted from two perspectives in order to compare the results with a view to analysing the conditions where the number of the best rules is high based.

VI. RESULTS AND DISCUSSION

The result of the experiment is presented in this section in the following order:

The presentation of the results from the experiment is based on the following experiments:

A. Classification

Training model using all attributes including the class attribute. This is considered to be a supervised model creation, because the model is built based on the class values in correspondence to the values of attributes respectively

The accuracy achieved under different experiment conditions or setting by Decision Tree, Naive Bayesian, and Artificial Neural Network (ANN) are presented in Tables 5, 6, and 7 respectively.

TABLE 5 CLASSIFICATION ACCURACY OF DECISION TREE

Percentage Split		**Time in Seconds**		**Decision Tree**	
Training %	**Testing %**	**Build**	**Test**	**Correct %**	**Incorrect %**
90	10	0.07	0.01	93.06	6.93
66	44	0.01	0.01	90.81	9.18
50	50	0.01	0.02	92.7	7.29
10 Folds		0.01	0.01	93.22	6.77

TABLE 6 CLASSIFICATION ACCURACY OF NAIVE BAYESIAN

Percentage Split		**Time in Seconds**		**Naive Bayesian**	
Training %	**Testing %**	**Build**	**Test**	**Correct %**	**Incorrect %**
0	10	0.02	0.05	93.06	6.93
66	44	0	0.03	92.51	7.48
50	50	0	0.04	92.7	7.29
10 Folds		0	0.25	93.51	6.48

TABLE 7 CLASSIFICATION ACCURACY OF ANN

Percentage Split		**Time in Seconds**		**ANN**	
Training %	**Testing %**	**Build**	**Test**	**Correct %**	**Incorrect %**
90	10	7.1	0	93.06	6.93
66	44	7.19	0.01	90.81	9.18
50	50	6.98	0.02	92.7	7.29
10 Folds		7	0.03	93.51	6.48

B. Clustering

Training model without class attributes. This is considered unsupervised because before the model is created, the values of the dataset are clustered; then the model is created for training and tested based on the cluster created.

The accuracy achieved under different experiment conditions or setting by Decision Tree, Naive Bayesian, and Artificial Neural Network (ANN) are presented in Tables 8, 9, and 10 respectively.

TABLE 8 CLUSTERING ACCURACY OF DECISION TREE

Percentage Split		**Time in Seconds**		**Decision Tree**	
Training %	**Testing %**	**Build**	**Test**	**Correct %**	**Incorrect %**
90	10	0.17	0.02	100	0
66	44	0.01	0.02	100	0
50	50	0	0.02	100	0
10 Folds		0.01	0.01	100	0

TABLE 9 CLUSTERING ACCURACY OF NAIVE BAYESIAN

Percentage Split		Time in Seconds		Naive Bayesian	
Training %	**Testing %**	**Build**	**Test**	**Correct %**	**Incorrect %**
90	10	0.01	0	100	0
66	44	0	0.01	100	0
50	50	0	0.01	100	0
10 Folds		0.01	0.01	100	0

TABLE 10 CLUSTERING ACCURACY OF ANN

Percentage Split		Time in Seconds		ANN	
Training %	**Testing %**	**Build**	**Test**	**Correct %**	**Incorrect %**
90	10	7.12	0.01	100	0
66	44	7.12	0.01	100	0
50	50	7.12	0.01	100	0
10 Folds		7.51	0.03	100	0

C. *Class Association Rules (CAR)*

This algorithm produces the association rules of the relevant values of each attribute to the class attribute value. Apriori was selected as the algorithm for the class association rules in this section. Table 11 and 12 shows the result of the individual classifiers presented under different experiment setting.

TABLE 11 APRIORI CLASS ASSOCIATION RULES ON COMPLETE CAR DATASET – CAR (FALSE)

No. Cycles	15
Min Support	0.25 (432 instances)
Minimum metric <confidence>	0.9
No. Of Rules Used	10
No. of Best rules found <conf. Level 0.9>	10

TABLE 12 APRIORI CLASS ASSOCIATION RULES ON COMPLETE CAR DATASET – CAR (TRUE)

No. Cycles	17
Min Support	0.15 (259 instances)
Minimum metric <confidence>	0.9
No. Of Rules Used	20
No. of Best rules found <conf. Level 0.9>	10

VII. DISCUSSIONS

A. *Accuracy*

- The classified dataset result from the comparison between the three classifiers shows that Decision Tree and ANN have exactly the same accuracy across the three (90:10, 66: 44, 50: 50) settings.
- The clustered dataset result from the comparison across the four models was 100% accurate across all model with the four experiment setting (90:10, 66: 44, 50: 50, 10-Folds).
- Comparing the result of the three classifiers on the dataset (with class attributes) as shown from the results in Tables 5, 6, and 7 under classification; it is observed that using 10 folds on the models produces result which completely differs from the result from the percentage split. The result shows Naive Bayesian and ANN to be the best models on the dataset with both models having the same accuracy percentage. But, the 10 folds cross validation achieved higher accuracy in all algorithms used in Tables 5, 6, and 7.
- However, to provide a distinction between the performance of the two best hold-out models (decision tree and ANN) from classification and clustering results showed in tables 5, 6, and 7; tables 8, 9, and 10; time can be considered as a factor; because it takes decision tree less than it takes ANN to build the model. Also, it can be concluded that the Naive Bayesian has the lowest accuracy on the dataset compared to Decision Tree and ANN.
- The Apriori Class Association Rule used on the dataset achieved the same accuracy which was 10 best rules. These rules were maintained consistency in the outcome of the two experiments despite the fact that the experiments were under different settings.

A general observation on the dataset with regards to accuracy is the dimensionality of the class attribute. This means, the smaller the dimension or attribute values for the class variable; the higher the accuracy of the model. This was observed from the 'classified' and the 'clustered' dataset. The classified dataset has a class with four attribute values (i.e acc, unacc, good, vgood), thus; having a model with the highest accuracy to be 93%. This accuracy is low compared to the clustered car dataset which has only two clusters (i.e. cluster1, and cluster2) as values for the class attribute, and the accuracy obtained from using the clustered dataset to build a model was 100% across all algorithms used under different experiment settings.

To prove this further, the dataset was clustered into four clusters and the same test specifications which yielded 100% accuracy on the classification experiment was used on the clustering experiment on the four clustered outcomes; but the highest accuracy was 30%. This means that the clustering experiment achieved 70% less accuracy compared to the classification experiment outcome.

B. Speed

In terms of the time taken to build and test the model, the result shows Naive Bayesian to be the fastest. Followed by Decision Tree with a very little difference, and ANN at last taking the most time to build and test the model. However, the three models were observed to have a varying duration for model building and testing in proportion to the percentage split; where a smaller training test implies a longer time testing the mode, and vice versa. Also, the 10 Fold was observed to be almost the same in duration of training and testing as the percentage split.

C. Interpretability

The computation process in WEKA for Decision Tree and Naive Bayesian are readable and understandable. But ANN is obviously hard to understand because it is a Black-Box algorithm. But in general the results are readable and understandable.

VIII. Conclusion

The comparative analysis of the models used in this study shows that Multilayer Perceptron of Artificial Neural Network (ANN) takes longer to build and test a model compared to Decision Tree, Naive Bayesian, and the 10-Folds Cross Validation. However, in terms of accuracy, the Multilayer Perceptron seem be the best to cut across dataset percentage split and cross validation algorithms. Also, it was observed in this study that the smaller the number of the dimension of class of a dataset, the higher the accuracy of the model would be.

Acknowledgment

Much gratitude and credit to the University of California Irvine (UCI) data repository and Marco Bohanec for making the car evaluation dataset available, also to the timeless support and advice given by Professor Azuraliza Abu Bakar.

Reference

[1] C. M. Standards and M. Practices, "Lesson Plan The True Cost of Owning a Car," pp. 1–5.

[2] S. Makki, A. Mustapha, J. M. Kassim, E. H. Gharayebeh, and M. Alhazmi, "Employing Neural Network and Naive Bayesian Classifier in Mining Data for Car Evaluation," no. April, pp. 12–14, 2011.

[3] D. Delen, G. Walker, and A. Kadam, "Predicting breast cancer survivability: a comparison of three data mining methods," *Artif. Intell. Med.*, vol. 34, no. 2, pp. 113–127, Jul. 2014.

[4] R. Russo, "Bayesian and Neural Networks for Motion Picture Recommendation," 2006.

[5] L. Console, C. Gena, I. Torre, D. Informatica, and U. Torino, "Evaluation of an on-vehicle adaptive tourist service."

[6] S. Singh, "Modeling Performance of Different Classification Methods: Deviation from the Power Law," no. April, 2005.