ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 8, Issue 7 , July 2021

Exploration of Tree Induction Algorithms in Classification of Vehicle Tyres

NwitoAmos Dornubari, Gboraloo AngelaWaale

Engr. and P.G. Student, Department of Computer Science, University of Lagos, Lagos State, Nigeria
Technologist, Department of Computer Science, KenuleSaroWiwa Polytechnic, Bori, Rivers State, Nigeria

ABSTRACT - Classification is one of the data mining techniques used to group records into various categories established on the prediction of a dependent attribute called class. Classification techniques such as tree induction have been employed in numerous applications where decisions are required on large dataset because of its simplicity of understanding and interpretation. Vehicle tyres plays vital roles in the country and world at large, they are used by automobiles such as baby carriage, shopping carts, motorcycles, bicycles, forklifts, farm equipment, buses, cars, trucks, aircraft landing gears. In this research, tree induction-based classification is explored in a comparative approach to mine Vehicle tyres dataset using decision trees such as C4.5, Random Tree and Random Forest algorithms to explicitly and visually represent decisions. The algorithms were implemented in java using WEKA and results compared.

KEYWORDS: Classification, Datamining, Vehicletyres, C4.5, Random Tree and Random Forest.

I. INTRODUCTION

The essential roles played by vehicles tyres in the country and world at large cannot be over emphasized as this can be seen in different automobiles such as baby carriage, shopping carts, motorcycles, bicycles, forklifts, farm equipment, buses, cars, trucks, aircraft landing gears and so on. The production of these tyres generates huge volume of data such that the classes of these tyres are time consuming to be identified for decision taking as each class is determined by several attributes. Classification is one of the techniques in data mining that can assist in identifying and categorizing these data into their respective groups for management decision makings.
Data mining is a procedure for learning interesting patterns and knowledge from huge dataset. The data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically [5].

According to [4]data mining is process of extracting valid, previously unknown, comprehensible, and actionable information from large databases and using it to make crucial business decisions.

Classification is a data mining technique use to establish a specific predetermined class for each record in a database from a finite set of possible class values [4].

Tree induction commonly referred to as decision tree induction in data mining, are used for classification and predictive operations. Tree induction operation can be represented in the form of a model such as classifier or regression model. Tree Induction is referred to as classification tree when employ for classification operations, and regression tree when employ for regression operations. Tree induction are widely used in data mining because of its self-explanatory nature. You do not need to be an expert in data mining before you can understand decision trees [17].

This research is motivated by the essential roles of vehicle tyres in every country such as; carrying of the vehicle load, generate and transmit the forces to the vehicle, guide(Steer) the vehicle, accelerate or braking, beautify the Automobile, absorb noise and mechanical vibration, the production properties such as grip, rolling resistance and fuel consumption, road handling, noise and comfort, wear and endurance and the benefit of decision tree algorithms to obtain a better understanding of the characteristics of these data set to quickly decide which characteristic determine which class of tyre. Vehicle dataset contain enormous information about tyres such that it is difficult to know which properties determine the class of tyre.

The researchers carried out a comparative analysis of Tree induction algorithms to classify vehicle tyre dataset according to their known Global Dimension (GD) codes using, **C4.5, Random Tree** and **Random Forest** algorithms to determine some characteristic which will be useful in production and management decision making for a comparative exploration.

II. THE SIGNIFICANT OF THE PROPOSED SYSTEM

1. The decision tree induction will assist information miners to identify interested groups in the data and as well understand the association between the groups.
2. The system serves as a decision support system for the management.

III. RELATED WORK ON TREE INDUCTION ALGORITHMS

In the comparative analysis of Tree induction algorithms of [21] using ID3, C4.5 and CART to predict the performance of student. The study was conducted on 48 students and result indicated that CART algorithm had better accuracy compared to the ID3 and C4.5. algorithms.

In the comparative study of ID3 and C4.5 decision tree [6] using three different sample size of weather repository data set, the study indicated that C4.5 perform better than ID3 in terms of accuracy percentage of 94.15% to 96.2% respectively, and also in terms of execution time, C4.5 takes less time in execution than ID3.

In a comparative analyis by[16] C4.5 and C5.0 was compared on crop pest data, result indicated that C5.0 perfomerd better than C4.5 in terms of correctly classified instances of 197 and 195, accuracy of prediction of 99.49% and 98.48% and time taken 0.01 and 0.02 and error rate of 0.52% and 1.52% resptectly[16].

In the study of [20]ID3, CART, and C4.5 was compared using weather repository dataset, on different properties of the algorithms and indicated that these algorithms “attain their maximum value all records belong to the same class’, ID3 can handle categorical data as well as missing values while CART and C4.5 can handle both categorical and numerical values and missing values and that better result can be obtained with these algorithms knowing which algorithm is suitable for a specific category of data set.

In the a comparative study of tree induction algorithm by [13] using J48(C4.5), Multilayer Perceptron and Naives Bayes on Haematological data, result shows that J48(C4.5), has a greater classifier accuracy value of 97.16% compare to Multilayer perceptron of 86.55% and Naïve bayes of 70.28% for correctly classified instances.

Kumar &Kiruthika,2015; compared different classification algorithms such as ID3, C4.5, C5 and hunt algorithm for different properties and concluded that C4.5, C5 and CART can handle continuous(numerical) and categorical data while ID3 handles categorical only, ID3 has low speed among all, and terms of formala, ID3 uses infromation gain, C4.5 and C5 uses split Info and gain ration while CART uses Gini diversity Index(simply refered to as Gini index).

Yamuna & Venkatesan, 2014, compared ID3, C4.5, and CART using Boosting method an ensemble technique in terms of sensitivity and specificity andindicated that C4.5 has the heighest value of 71.7% in sensitivity, CART has the heighest value of 77.3% in specificity while ID3 is the lowest amongst them with C4.5 having the highest classification rate of 73.5%. CART performs better than all with ensemble method of Bagging (Boostrap aggregation).

In the comparative study of [14] using J48 (C4.5), Random Forest, Naïve Bayesian, Random Tree and Decision Stump decision tree induction algorithms on airline, shows diverse results, where Naïve Bayes and J48 produced almost same value in performance measure in accuracy but Random Forest gives the best overall performance followed bt Naïve Bayes and J48.

A comparative study of tree inductions conducted by [22] using CART and C4.5 with Bagging and Boosting, and Random Forest. Result shows that Random Forest gives a good result but CART with boosting technique and C4.5 with bagging technique give the best overall result.

Random Forest developed by [3] is another decision tree induction algorithm based on ensemble technique.It build many decision trees called forest using Bagging technique. Randome forest is used for both classification and regression just like CART.

IV. METHODOLOGY

Considering the role of vehicle tyres in the world and the importance of tree induction algorithms in data analysis. We explore C4.5, Random tree and Random forest on Michelin tyre dataset for a comparative exploration. The major objective of this research is the classification of Michelin vehicle tyres dataset using tree induction algorithms.

Tree Induction Algorithms

Decision Tree Induction algorithms proposed for classification includes:CART (Classification and Regression Trees) by [12]ID3 (Iterative Dichotomizer 3) by [15], C4.5 as an upgrade of ID3, C5 by [18][19] respectively, Explore by [23] and so on.

Random Forest: Random Forest developed by[3] is another decision tree induction algorithm based on ensemble technique.It build many decision trees called forest using Bagging technique. Randome forest is used for both classification and regression just like CART.

Growing a Decision Trees

Trees are grown in a top-down recursive divide-and-conquer manner, in growing and splitting a tree, first the predictor attribute is tested against the target or class attribute using **attribute selection measures.**

Attribute Selection Measures

Attribute selection measures, also referred to as **splitting rules** are used to decide how the records in a given node, (test attribute) are split. An **attribute selection measure** is used to choose the splitting criterion that “best” partitions the input data set into several fragmented partitions of the class-labelled training tuples into individual classes.

It measures the homogeneity or heterogeneity of the table base on the classes. A table is said to be homogeneous or pure if it contains only a single class and impure or heterogeneous if a data table contains several classes.

There are several indices for quantitatively measuring the degree of impurity. Gini index or gini purity, Information gain, GainRatio and classification error are the most well-known indices for measuring degree of impurity. It splits the dataset by the values in test node (split attribute) to form the branches and recursively perform the process on each branch. The branch with entropy = 0 forms a leaf node in the tree.

1. **C4.5:** Attribute selection measures used by C4.5, Random Tree and Random Forest include:

   1. **Gain Ratio**

      C4.5 uses the following formulas:

      a. Information **Entropy** of the dataset is given as:

$$Entropy(D) = -\sum_{i-1}^{n} P_i log_2 P_i$$

      b. **Information Gain** is defined as:

$$Gain(D,v) = Ent(D) - \sum_{i-1}^{n} \frac{|D_i|}{|D|} Ent(D_i)$$

      c. **SplitInfo** is defined as:

$$SplitInfo(p,A) = -\sum_{i=1}^{n} \frac{|D_i|}{|D|} \log_2 \frac{|D_i|}{|D|}$$

d. Gain ratio is defined as:

$$GainRatio = \frac{Gain(p,T)}{SplitInfo(p,T)}$$

The attribute that has the maximum gain ratio becomes the Splitting attribute

2. Random Forest: Attribute selection measure used by random forest is either gini index also referred as gini purity and gini split which is for CART or entropy and information gain for C4.5 [2].

Gini Index is defined as:

$$Gini(D) = 1 - \sum_{i}^{n} p_i^2$$

2. Gini Split is defined as:

$$G_{split} = 1 - \sum_{i}^{k} \frac{n_i}{n} G_i$$

3. **Random Tree:** Random Tree also uses gini index.

V. DESIGN AND IMPLEMENTATION

The proposed system Vehicle Tyre Classification System (VTCS) is a desktop-based program implemented in Java Programming Language using Wekafor the classification of vehicle tyres using Global Dimension Code (GD) as the Predictor is shown in figure 1 below which consists of two basic objects; primary and secondary objects. The primary objects in the system include:

a. **The dataset**: The dataset use as input to the system is a mixed attribute dataset stored on the system as seen in 2 below. The user is expected to prepare the data before applying the tree induction algorithms.
b. Tree induction algorithm: the data miner selects the algorithm for the classification task.
c. The output is displayed as decision tree which can be viewed in a graphical form.

The secondary object of the system is the data miner (the user).
The architecture of the proposed system as shown in figure 1 below consists of two main stages
i. The pre-process stage
ii. The data mining stage

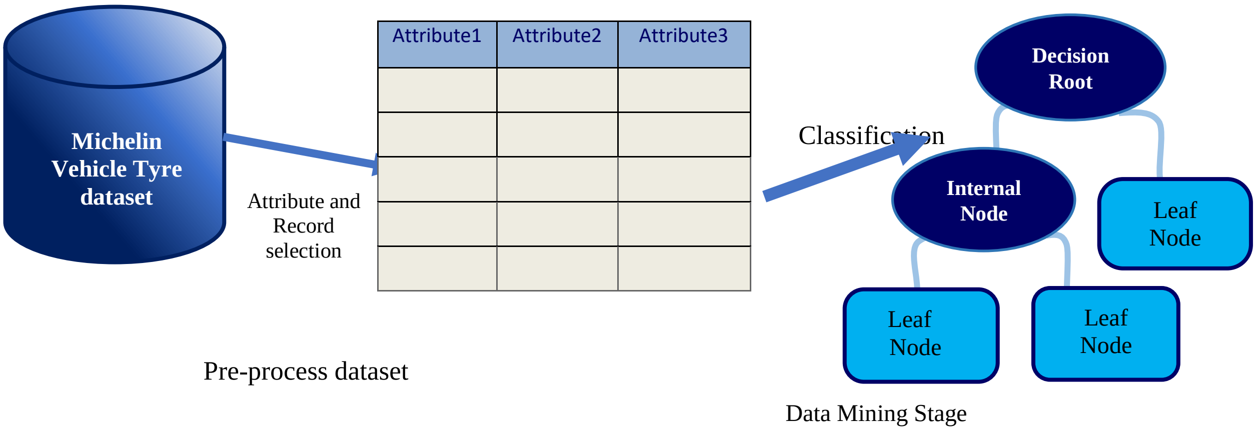


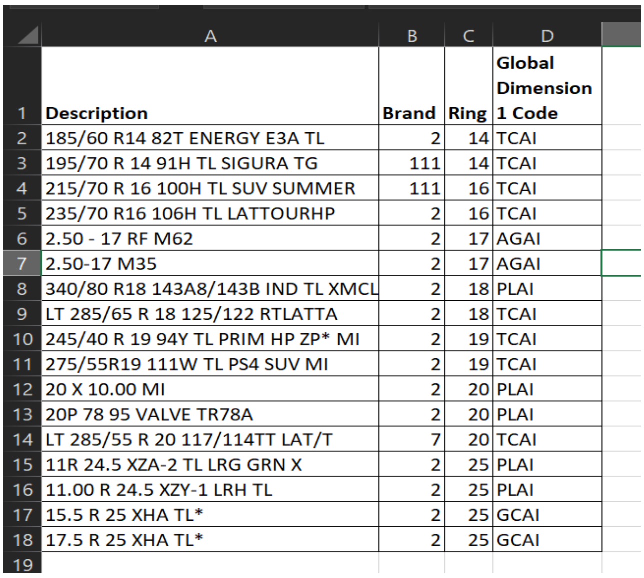
Figure 1: Architecture of the System

1. **The Pre-process Stage:** In the pre-process stage, the data miner prepares the data set to be use for the data mining task from the original dataset where the data is selected and save with a CSV or ARFF file extension using Microsoft Excel.

2. **The data mining Stage:** In the data mining stage, the data miner applies any algorithm such as C4.5, Random Tree or Random Forest, the input dataset is categorized into their respective classes which can be displayed in graphical form for easily understanding and interpretation.

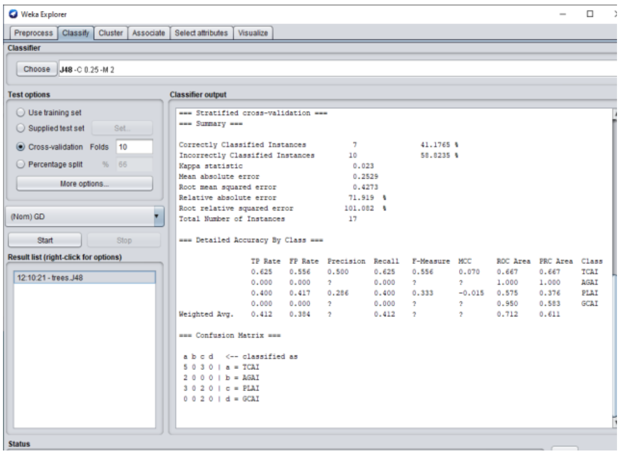
VI. IMPLEMENTATION

The data set used in the exploration of various tree inducers is a sample data collected from Michelin vehicle tyres, which consist of several attributes such as Description, Brand, Ring and Global Dimension (GD) as shown in Figure 2 below:



	A	B	C	D
1	Description	Brand	Ring	Global Dimension 1 Code
2	185/60 R14 82T ENERGY E3A TL	2	14	TCAI
3	195/70 R 14 91H TL SIGURA TG	111	14	TCAI
4	215/70 R 16 100H TL SUV SUMMER	111	16	TCAI
5	235/70 R16 106H TL LATTOURHP	2	16	TCAI
6	2.50 - 17 RF M62	2	17	AGAI
7	2.50-17 M35	2	17	AGAI
8	340/80 R18 143A8/143B IND TL XMCL	2	18	PLAI
9	LT 285/65 R 18 125/122 RTLATTA	2	18	TCAI
10	245/40 R 19 94Y TL PRIM HP ZP* MI	2	19	TCAI
11	275/55R19 111W TL PS4 SUV MI	2	19	TCAI
12	20 X 10.00 MI	2	20	PLAI
13	20P 78 95 VALVE TR78A	2	20	PLAI
14	LT 285/55 R 20 117/114TT LAT/T	7	20	TCAI
15	11R 24.5 XZA-2 TL LRG GRN X	2	25	PLAI
16	11.00 R 24.5 XZY-1 LRH TL	2	25	PLAI
17	15.5 R 25 XHA TL*	2	25	GCAI
18	17.5 R 25 XHA TL*	2	25	GCAI
19				

Figure 2: Vehicle Tyre dataset



Weka Explorer

Classifier: J48 -C 0.25 -M 2

Test options: ● Cross-validation Folds 10

(Nom) GD

Result list (right-click for options): 12:10:21 - trees.J48

Classifier output

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances           7               41.1765 %
Incorrectly Classified Instances        10               58.8235 %
Kappa statistic                          0.023
Mean absolute error                      0.2529
Root mean squared error                  0.4273
Relative absolute error                 71.919  %
Root relative squared error            101.082  %
Total Number of Instances               17

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.625    0.556    0.500      0.625    0.556      0.070    0.667     0.667     TCAI
                 0.000    0.000    ?          0.000    ?          ?        1.000     1.000     AGAI
                 0.400    0.417    0.286      0.400    0.333      -0.015   0.575     0.376     PLAI
                 0.000    0.000    ?          0.000    ?          ?        0.950     0.583     GCAI
Weighted Avg.    0.412    0.384    ?          0.412    ?          ?        0.712     0.611

=== Confusion Matrix ===

 a b c d   <-- classified as
 5 0 3 0 | a = TCAI
 2 0 0 0 | b = AGAI
 3 0 2 0 | c = PLAI
 0 0 2 0 | d = GCAI
```

Figure 3: C4.5 (J48) Classification Output

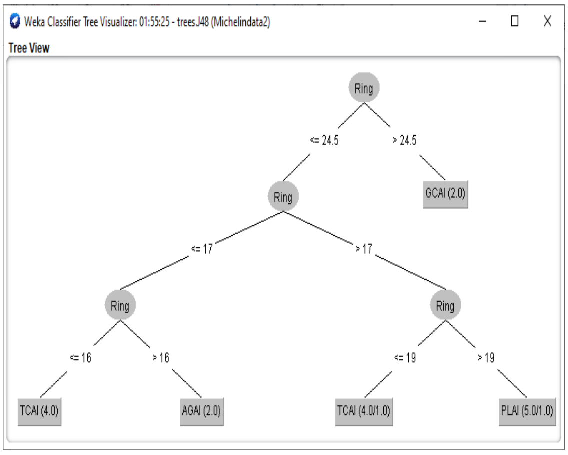
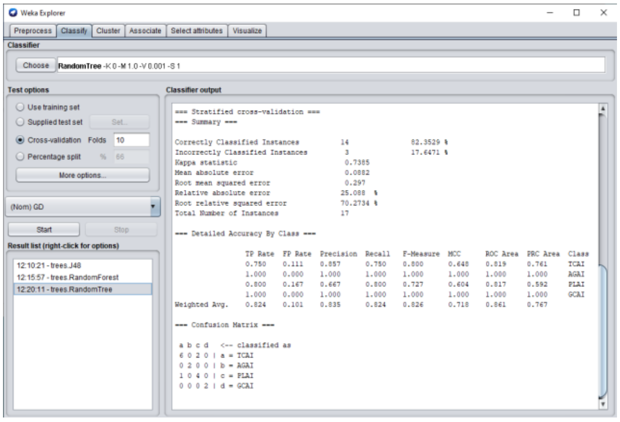


Figure 4: J48 Tree view



Weka Explorer

Classifier: RandomTree -K 0 -M 1.0 -V 0.001 -S 1

Test options: ● Cross-validation Folds 10

(Nom) GD

Result list (right-click for options): 12:10:21 - trees.J48, 12:15:57 - trees.RandomForest, 12:20:11 - trees.RandomTree

Classifier output

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          14               82.3529 %
Incorrectly Classified Instances         3               17.6471 %
Kappa statistic                          0.7385
Mean absolute error                      0.0882
Root mean squared error                  0.297
Relative absolute error                 25.088  %
Root relative squared error             70.2734 %
Total Number of Instances               17

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.750    0.111    0.857      0.750    0.800      0.648    0.819     0.761     TCAI
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     AGAI
                 0.800    0.167    0.667      0.800    0.727      0.604    0.817     0.592     PLAI
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     GCAI
Weighted Avg.    0.824    0.101    0.835      0.824    0.826      0.718    0.861     0.767

=== Confusion Matrix ===

 a b c d   <-- classified as
 6 0 2 0 | a = TCAI
 0 2 0 0 | b = AGAI
 1 0 4 0 | c = PLAI
 0 0 0 2 | d = GCAI
```

Figure 5: Random Tree Classification Output

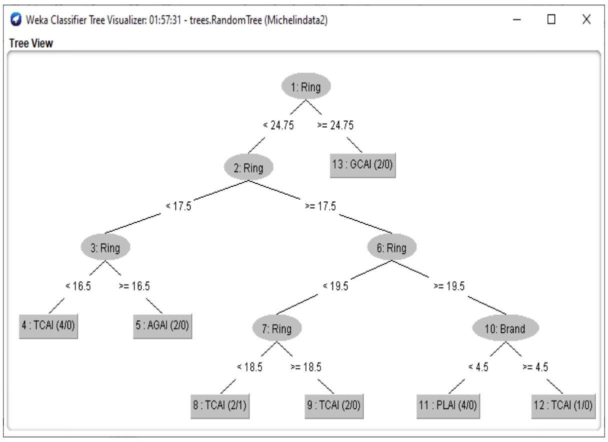
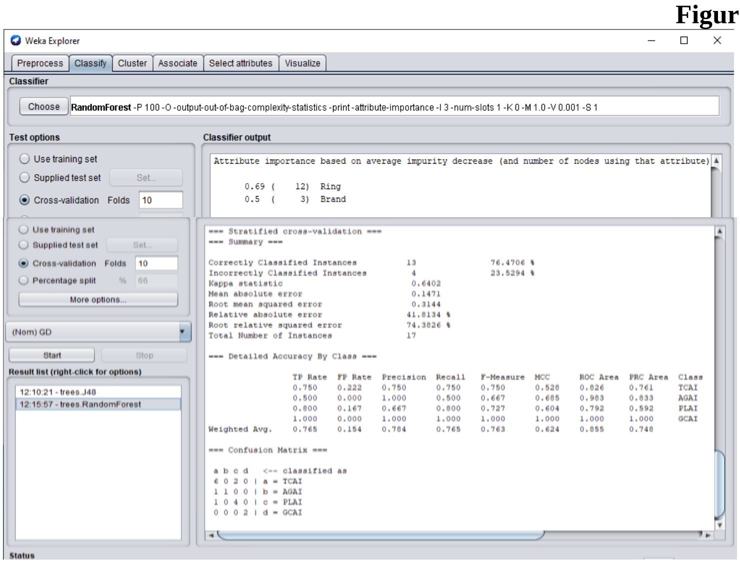


Figure 6: Random Tree, Tree View



Weka Explorer

Classifier: RandomForest -P 100 -O -output-out-of-bag-complexity-statistics -print-attribute-importance -I 3 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1

Test options: ● Cross-validation Folds 10

(Nom) GD

Result list (right-click for options): 12:10:21 - trees.J48, 12:15:57 - trees.RandomForest

Classifier output

```
Attribute importance based on average impurity decrease (and number of nodes using that attribute)

      0.69 (    12)  Ring
      0.5  (     3)  Brand
```

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          13               76.4706 %
Incorrectly Classified Instances         4               23.5294 %
Kappa statistic                          0.6402
Mean absolute error                      0.1471
Root mean squared error                  0.3144
Relative absolute error                 41.8134 %
Root relative squared error             74.3826 %
Total Number of Instances               17

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.750    0.222    0.750      0.750    0.750      0.528    0.826     0.761     TCAI
                 0.500    0.000    1.000      0.500    0.667      0.685    0.983     0.833     AGAI
                 0.800    0.167    0.667      0.800    0.727      0.604    0.792     0.592     PLAI
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     GCAI
Weighted Avg.    0.765    0.154    0.784      0.765    0.763      0.624    0.855     0.748

=== Confusion Matrix ===

 a b c d   <-- classified as
 6 0 2 0 | a = TCAI
 1 1 0 0 | b = AGAI
 1 0 4 0 | c = PLAI
 0 0 0 2 | d = GCAI
```

Figure 7: Random Forest Classification Output

VII. RESULT EVALUATION FOR CLASSIFICATION

In evaluating the tree inducers, the following measures were used:

1. **Evaluation using Stratified Cross-validation** of 10 folds: The classification table 1 shows how the dataset was classified by each tree inducers where Random Tree has the highest number of correctly classified instances.

Table 1 Classified instances of Tree Inducer

Tree Inducers	Correctly Classified Instances	Percentage	Incorrectly Classified Instances	Percentage
C4.5(J48)	7	41.1765 %	10	58.8235 %
Random Tree	14	82.3529%	3	17.6471%
Random Forest	13	76.4706 %	4	23.5294 %

2. Evaluation using Confusion Matrix

In machine learning, a Confusion matrix is the summary of prediction result in a tabular form used to describe the performance of an algorithm in classification technique of data mining operation. It shows how confused the classification model is when making predictions [6], including the type of errors made by the classifier and the correct and incorrect instance of predictions are summarized with count values and broken down by each classes. The table usually contains information about actual and predicted classifications. The basic measures used in Confusion matrix includes:

i. **Accuracy**: is used to measure the proportion of correctly classified instances in the dataset:

$$\text{Accr} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

ii**. Misclassification Error** is used to measure the proportion of incorrectly classified instances in the dataset.

$$\text{MCR} = \frac{\text{FP} + \text{FN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}}$$

iii. **True Positive Rate (Recall or Sensitivity):** True Positive rate also known as **Recall** or Sensitivity is the number of correctly classified instances by the classifier that is actually true. it measures how good the model is good at detecting positive classes. The best sensitivity is 1.0, whereas the worst is 0.0.

Sensitivity is calculated as:$\text{TPR} = \frac{\text{TP}}{\text{TP+FN}}$

iv. **False Positive Rate*:*** is the number of instances incorrectly classified as positive when it is false. The system predicts it to be true when it is false. It is regarded as **type I error**, which occurs when the system rejects the hypothesis and accept an alternative. $\mathbf{TPR} = \frac{\mathbf{FP}}{\mathbf{FP+TN}}$

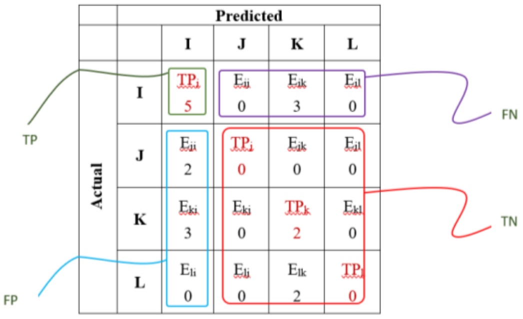
v. False Negative Rate: is the number of instances incorrectly classified by the classifier as negative when it is true. The system predicts it to be false when it is actually true. It is regarded as **type II error**, which occurs when the system accepts a negative hypothesis but the hypothesis is false. $\mathbf{FN} = \frac{\mathbf{FN}}{\mathbf{FN+FP}}$

vi. **Precision (Positive predictive value)**

Precision also known Positive predictive value is a measure of how good the model is at assigning positive events to the positive class. i.e.,when the classifier predicts positive, how correct is the prediction.$\boldsymbol{PRC} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}}$

vii **ROC Area**: is Receiver operating characteristics for TPR and FPR.

Confusion Matrix for C4.5



		Predicted			
		I	J	K	L
Actual	I	TP_i 5	E_{ij} 0	E_{ik} 3	E_{il} 0
	J	E_{ji} 2	TP_j 0	E_{jk} 0	E_{jl} 0
	K	E_{ki} 3	E_{kj} 0	TP_k 2	E_{kl} 0
	L	E_{li} 0	E_{lj} 0	E_{lk} 2	TP_l 0

Legend
TP = True Positive
FN = False Negative
FP = False Positive
TN = True Positive

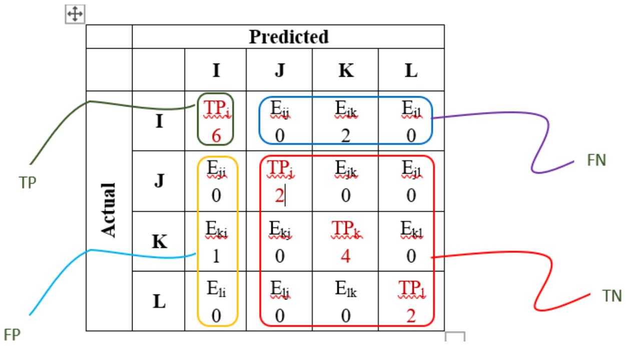
i. Accuracy for C4.5 is calculated as

$$Accr = \frac{TP+TN}{TP+TN+FP+FN} = \frac{9}{17} = 0.529 = \quad 0.53$$

ii. Misclassification rate for C4.5 is calculated as

$$MCR = \frac{FP+FN}{TP+FN+FP+TN} = \frac{8}{17} = 0.471 \quad = 0.47$$

Confusion Matrix for Random Tree



		Predicted			
		I	J	K	L
Actual	I	TP_i 6	E_{ij} 0	E_{ik} 2	E_{il} 0
	J	E_{ji} 0	TP_j 2	E_{jk} 0	E_{jl} 0
	K	E_{ki} 1	E_{kj} 0	TP_k 4	E_{kl} 0
	L	E_{li} 0	E_{lj} 0	E_{lk} 0	TP_l 2

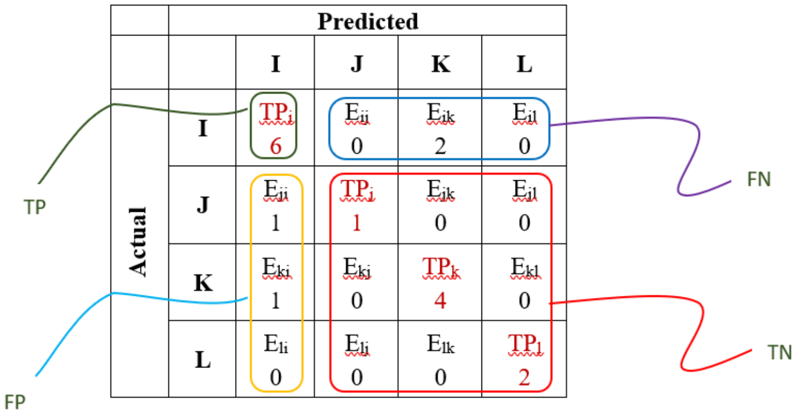
i. Accuracy for Random Tree is calculated as

$$Accr = \frac{TP+TN}{TP+TN+FP+FN} = \frac{16}{19} = \quad 0.842 \quad = 0.84$$

ii. Misclassification rate for Random Tree is calculated as

$$MCR = \frac{FP+FN}{TP+FN+FP+TN} = \frac{3}{19} = \quad 0.158$$

Confusion Matrix for Random Forest



		Predicted			
		I	J	K	L
Actual	I	TP_i 6	E_{ij} 0	E_{ik} 2	E_{il} 0
	J	E_{ji} 1	TP_j 1	E_{jk} 0	E_{jl} 0
	K	E_{ki} 1	E_{kj} 0	TP_k 4	E_{kl} 0
	L	E_{li} 0	E_{lj} 0	E_{lk} 0	TP_l 2

i. Accuracy for Random forest is calculated as

$$Accr = \frac{TP+TN}{TP+TN+FP+FN} = \frac{13}{17} = 0.764 = 0.76$$

ii. **Misclassification rate** for Random forestis calculated as

$$MCR = \frac{TP+TN}{TP+TN+FP+FN} = \frac{4}{17} =0.235= 0.24$$

iii. True Positive Rate **(**Recall or **Sensitivity**) Random forest is = 0.765

The weighted average of the algorithms is summarized in table 2 below: and plotted in the graph in figure 8 below:

Table 2 Weighted Average of performance Measure

	Accuracy	Misclassification	Precision	Recall	ROC Area
C4.5	0.53	0.47	0.000	0.412	0.712
Random Tree	0.842	0.158	0.835	0.824	0.861
Random forest	0.76	0.24	0.784	0.765	0.855

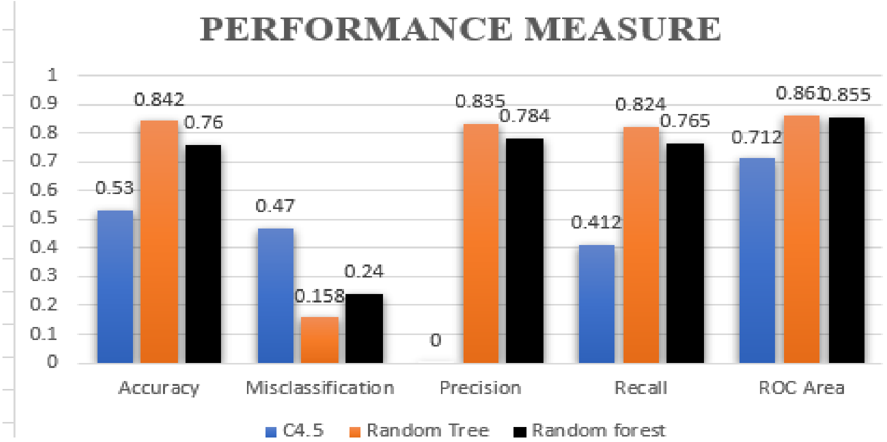


Figure 8: Analysis of the Tree inducers

From the evaluated parameters of the graph, Random tree has the highest percentage in accuracy, precision, Recall, and ROC area and the least misclassification rate

VII. CONCLUSION

This research comparatively explored three tree induction algorithms; C4.5, Random tree and Random Forest to classify the classes of vehicle tyres base on certain attributes. From these study, one can easily categorize a tyre as TCAI, GCAI, PLAI or AGAI by looking at the Ring size and brand and identify the most importance variable in the dataset as Ring and Brand, which is indicated by random forest. The performance measure use in evaluating the tree inducers include Accuracy, misclassification rate, recall, and precision from Confusion Matrix, it was observed that Random Tree performs better than C4.5 and random forest with 0.84%, 0.76%, 0.53% for Accuracy, 0.16%, 0.24%, 0.47% for Misclassification, 0.84%, 0.78%, 0.00% for precision and 0.82%, 0.77%, 0.41% for Recall or Sensitivity for Random forest and C4.5 respectively From this research studies, Random tree is proven to be better than C4.5 and random forest,the tree inducers perform well on multiclass dataset. This research can be extended to Naïve Bayes, and anomaly detection in future explorations.

REFERENCES

[1]. Abraham, S., Henry, F. K., & Sudershan, S. (2011). *Database System Concepts* (6th ed.). McGrawHill.

[2]. Akash Dubey (2018) Feature selection using Random Forest, retrieved from https://towardsdatascience.com/feature selection using Random Forest, 26/7/2021.

[3]. Breiman, L., & Cutler, A. (2001). Random Forest. Berkeley: https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf.

[4]. Connolly, T., & Beggs, C. (2015). *Database Systems, A Practical Approach to Design, Implementation and Management.* (G. E. 6 Edition, Ed.) New York: Pearson EducationLimited.

[5]. Han, J., Kamber, M., & Pei, J. (2012). *Data Mining, Concepts and Techniques.* (3. Edition, Ed.) USA: Morgan Kaufman Publishers by Elsevier.

[6]. Hssina, B., Merbouha, A., Ezzikouri, H., &Erritali, M. (2014). A Comparative Study of Decision Tree ID3 and C4.5. *International Journal of Advanced Computer Science and Applications (IJACSA), 1*, 14-19.

[7]. Jason, B. (2016, November 18). *What is Confusion Matrix in machine Learning.* Retrieved from Machinelearningmastery.com: https://machinelearningmastery.com/confusion-matrix-machine-learning/

[8]. Jason, B. (2016, 2019). *https://machinelearningmastery.com/bagging-and-random-forest-ensemble-algorithms-for-machine-learning/.* Retrieved 10 30, 2019, from https://machinelearningmastery.com

[9]. Jiawei, H., Micheline, K., & Pei, J. (2012). *Data Mining Concepts and Techniques* (3rd ed.). Morgan Kaufmann Publishers.

[10]. Josh, S. (2018, Febuary 26). StatQuest: Random Forests Part 1 - Building, Using and Evaluating. Y. Retrievedfrom https://www.youtube.com: https://www.youtube.com/watch?v=J4Wdy0Wc_xQ 26/10/2019

[11]. Kumar, K. V., & P., Kiruthika. (2015). An Overview of Classification Algorithm in Data mining . *International Journal of Advanced Research in Computer and Communication EngineeRim, 4*(12), 255-257.

[12]. Leo Breiman, J. F. (1984). *Classification and Regression Trees.* (1. Edition, Ed.) New York : Brooks/Cole Publishing, Monterey.

[13]. Nurul, A., & Ahsan, H. (2015). Comparison of Different Classification Techniques Using WEKA for Hematological Data. *American Journal of EngineeRim Research (AJER) , 4*(3), 55-61.

[14]. Prasad, A. J., Deepa, A., & Rejeswari, K. (2018). A Comparative Study on Different Classification Algorithms Using Airline Data set. *International Journal on Recent and Innovation Trends in Computing and Communication, VI*(1), 142-145.

[15]. Quinlan, J. R. (1975). *ID3* (Vol. 1). University of Sydney: Kaufman Serie.

[16]. Revathy, R., & Lawrence, R. (2017). Comparative Analysis of C4.5 and C5.0 Algorithms on Crop Pest Data. *International Journal of Innovated Research in Computer and Communication EngineeRim, 5*(1), 50-58. Retrieved from www.ijircee.com

[17]. Rockach, L., & Maimon, O. (2008). *Data Mining with Decision Trees, Theory and Applications.* Singapore: World Scientific Publishing Co Pte. Ltd.

[18]. Sonia, S., & Manoj, G. (2014). Comparative Study ID3, CART and C4.5 Decision Tree Algorithm: A survey. *International Journal of Advanced Information Science and Technology, 3*(7), 47-52.

[19]. Surjeet, K. Y., Brijesh, B., & Saurabh, P. (2012). Data Mining Applications: A comparative Study for Predicting Student's performance. *International Journal Of Innovative Technology & Creative Engineerim , 1*, 13-19.

[20]. Yamuna, N. R., & Venkatesan, P. (2014). A Comparative Analysis of Decision Tree Methods to Predict Kidney Transplant Survival. *International Journal of Advanced Research in Computer Science, 5*(3), 225-229.

[21]. Zahidul, I. M. (2010, january 12 August,2019). *EXPLORE, a novel decision tree algorithm.* Retrieved from researchgate: https://www.researchgate.net/publication/220862865