Assignment 4: a mathematical essay on decision tree

Krishna Sah Teli
Cyber-Physical Systems (I2MP)
Indian Institute of Technology (IITM)
Madras - 600036, India
ge22m018@smail.iitm.ac.in

***Abstract*—In this paper, decision tree has been studied intensively with an application exploring evaluation of car based on its safety measures. It explains about the different parameters related to car which play crucial role in understanding its safety. The correlation between the safety and other parameters has been depicted graphically with the help of data visualization tools. Even though the price of car is high if its safety feature is not satisfactory then it will not be categorized as of good quality. Data analysis shows that the maintenance cost increases as the surge of buying price of car and the capacity of car also does contribute in its grading. It also shows that the number of doors increase with the increase in capacity of car and decrease with that of its price. It also allows us to visualize the price of car does not make it worthy if safety parameter is ignored and the reason of for low price car becoming more popular and accessible. With the help of a classification decision tree machine learning model, the car evaluation dataset has been implemented to build a classification model and has been explored its accuracy, precision and recall properties. Using two different splitting indices gini index and information gain, two different decision tree models have been trained and study shows that regardless of whether dataset is balanced or not, these classification models give same accuracy. In this revised paper, I have made improvement in modeling, and conclusion sections. I have also made some correction in Introduction section.**

***Index Terms*—Decision Tree, Data Visualization, Correlation, Gini, Information Gain, Precision**

I. Introduction

The technology has been growing rapidly and its applications have great impact on our day-to-day work. It has been playing great role in modernizing every aspects of our life from industrial to personal point of view. It has been possible because of data. Data is the fuel for the technology. Many industries and companies came to know the importance of data during the covid-19 time as they were not having tangible access other than online to reach out to customers. By preprocessing and feature engineering techniques, raw data can be converted into meaningful data which later used for training different machine learning models or developing artificial intelligence based device. ML and data science has geared up the revolutionary changes in the field of technology. It has not only contributed in the technical field but we can see its impact on every field from agriculture to astronomy. Thus, it has become the most trending topic in current time period. Ml-model can predict, estimate, recognize, detect, [3], [6], [7], [9] etc. based on the data and type of training model. In this paper, car evaluation data has been used to classify the car quality based on its safety measures. It also explains the correlation between the data features such as number of doors, capacity of car, luggage size, cost price, maintenance price, safety with respect to its grading index. With the help of data analysis and visualization, ground truth has been explored followed by development of classification model.

To build the classification model, decision tree, a supervised machine learning model has been implemented. [1], [2] A decision tree is composed of a root node, branches and leaf nodes. Each internal node denotes a test on an attribute, each branch shows the result of test and each leaf represents a class label. It is also called top-down approach as root node is at top and remaining branches and leaf nodes are toward bottom. There are two different splitting indices such as GINI index and Entropy index which help in splitting the parent node into child nodes. Decision tree is known by CART as it is capable of both classification as well as regression problems. It has contributed in foundation of different algorithms such as random forest, bagged decision trees and boosted decision trees. It has several application in industries [2], [3] from small scale company to big company. It has wide range of applications such as intrusion detection [7]–[9], medical checkup [10]–[13], weather forecasting, developing AI-based devices, e-commerce stack fall down prediction, banking, scientific studies, etc.

The main focus in this paper is to demonstrate how a safety of car contributes in its evaluation. It correlates the car evaluation to its other features such as buying cost, maintenance cost, number of doors, capacity, etc. With the help of visualization tools and techniques, the correlation between the features of car have been illustrated in graphical form. It also illustrates implementation of decision tree over the data to develop a classification model. It has been found out that the accuracy (80.53%)remains same for both techniques i.e decision tree with gini, and entropy.

The rest of the paper has been organized as follows: section II outlines the data set related to our problem, section III explains the mathematics behind the decision tree and tools used for developing our model. section IV Modelling and section V concludes the paper with key contributions made and directions for future works.

II. Data set

The source of the car evaluation data is kaggle which is famous for the data science and storage of data. This data has totally 7 columns namely 'buying', 'maint', 'doors', 'persons', 'lug_boot', 'safety', 'decision'. All the features have been taken into consideration for training the model. 'buying' shows the car price and it has three different values namely low, medium, high, and very high where as 'maint' explains about the maintenance cost of the car and it has also values similar to 'buying'. Similarly, 'persons' shows the capacity of car, and lug_boot tells about the size of luggage which can be accommodated in car. Further, 'safety' attribute represents the safety index of car such as low, medium, high, and very high and based on it, car is evaluated into 'decision' attribute such as acceptable, unacceptable, good, and very good.

With the data analysis and data visualization tools, the correlation among the data features has been studied and correlated to the safety and car evaluation. The fig.1 shows about 70% of cars are categorized as unacceptable and out of remaining 22.2% of cars are classified as just acceptable, only 3% are good, and 4% are considered as very good of quality.

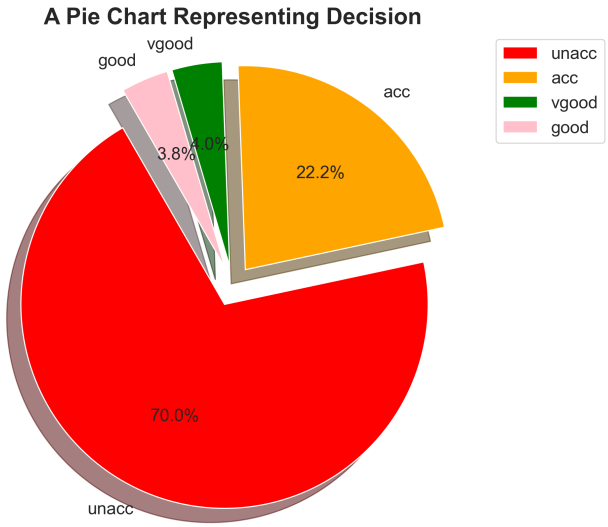

Fig. 1. Distribution of decision features

Similarly, the fig.2 shows about 80% of the cars are classified into unacceptable category regardless of their buying price. However, some of the low price cars fall into good and very good category whereas there is no possibility for high price or very high price cars. It shows that the price of car does not show the quality of car.

The fig. 3 illustrates that the maintenance cost of the car is proportional to its buying cost. It may be because the spare part of costly car is also costly. It does not tell that if the maintenance cost is expensive, then the car should be of good safety quality. Even though the price and maintenance cost is low, one can be good quality of car.

Likewise fig.3, figure 4 also describes that the capacity of car helps in evaluating the quality as if the capacity is 2

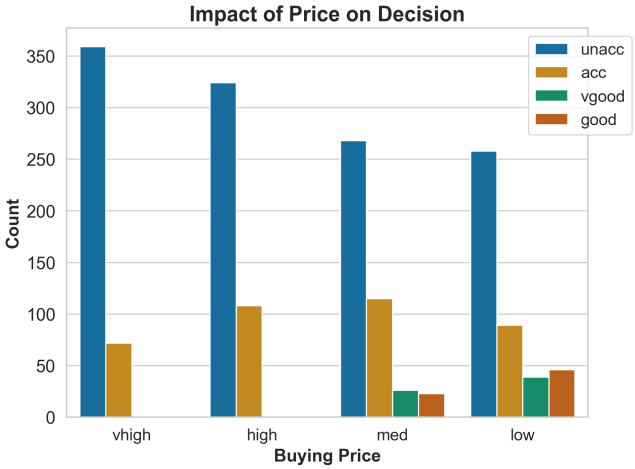

Fig. 2. Showing the prices of cars

persons then it is categorized as unacceptable regardless of other facilities. If the capacity is more than or equal to 4, then there is about 50% of chance to fall into unaccepted and 10% chance to fall into good or very good. It also shows that about 30% of cars are just classified into acceptable.

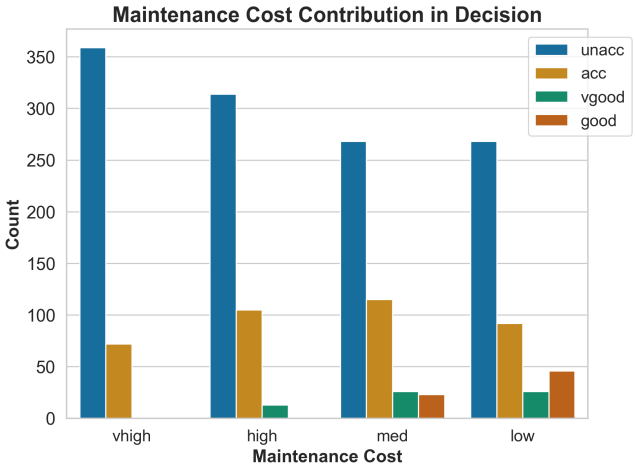

Fig. 3. Maintenance cost of car

The figure 5 depicts that the if the safety is compromised then it will be put into unacceptable category otherwise there is possibility of labelled as good or very good. Similarly, 30% of highly safe cars are discarded because of some other reasons, it may be capacity is less, maintenance cost is high. It shows that if the safety is medium then there is no chance of getting very good labelled but there is more than 50% chance to fall into unacceptable category.

The below figure 6 illustrates that the evaluation of cars based on the number of doors present in it. This figure clearly shows that number of doors is not so important parameter to evaluate the car quality as regardless of door counting, distribution is almost similar. However, if the doors is 2, it shows that the car size is small ans its capacity is less thus

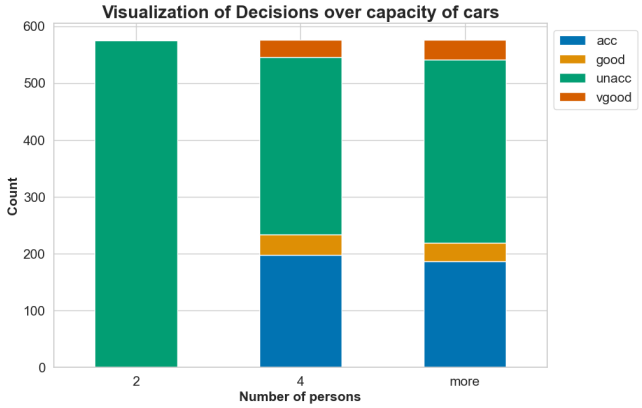

Fig. 4. Distribution of capacity of cars

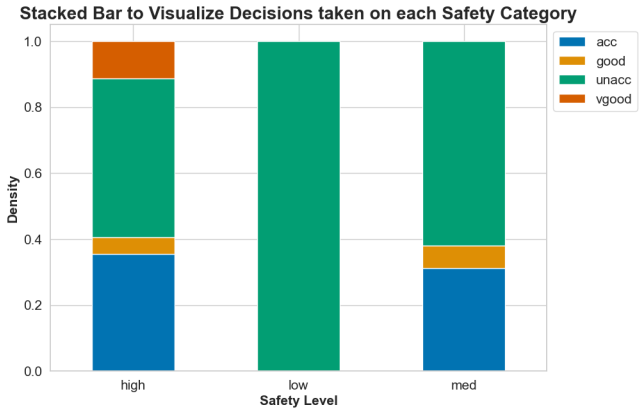

Fig. 5. Safety of cars with respect to decision

there is comparatively more probability of getting unaccepted.

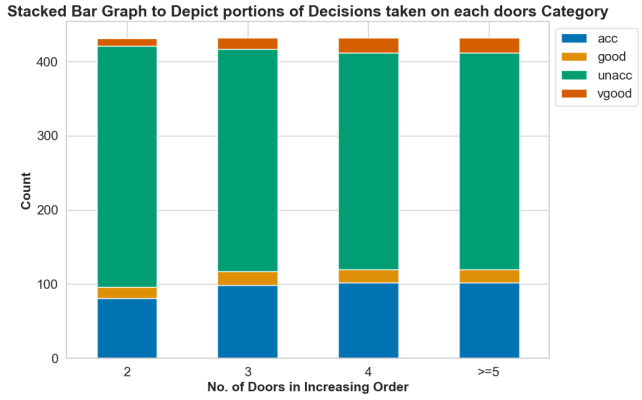

Fig. 6. Car evaluation based on the number of doors

Here the figure 7 depicts the relation between the luggage size and decision taken over car evaluation. It shows that more number of cars has facility to hold big size of luggage and more probability of getting classified into either acceptable, good, or very good. However, there is more probability of falling into unacceptable regardless of luggage size holding capacity.

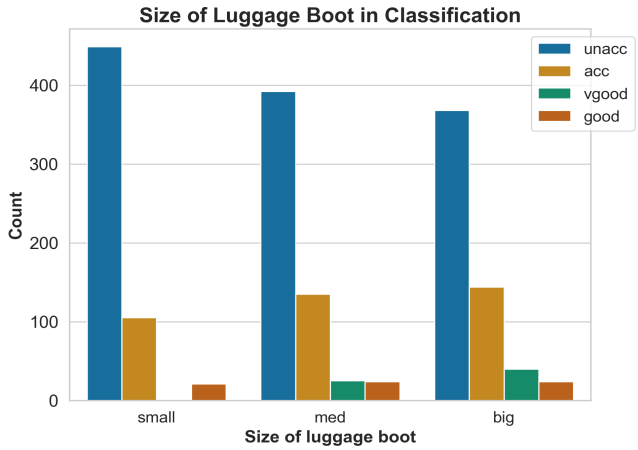

Fig. 7. Car evaluation based on the number of doors

III. Decision Tree

A decision tree is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks [1], [3]. It has a hierarchical, tree structure, which consists of a root node, branches, internal nodes and leaf nodes. A decision tree starts with a root node, which does not have any incoming branches. The outgoing branches from the root node then feed into the internal nodes, also known as decision nodes. Based on the available features, both node types conduct evaluations to form homogenous subsets, which are denoted by leaf nodes, or terminal nodes. The leaf nodes represent all the possible outcomes within the dataset.

Decision tree learning employs a divide and conquer strategy by conducting a greedy search to identify the optimal split points within a tree. This process of splitting is then repeated in a top-down [2], recursive manner until all, or the majority of records have been classified under specific class labels. Whether or not all data points are classified as homogenous sets is largely dependent on the complexity of the decision tree. Smaller trees are more easily able to attain pure leaf nodes—i.e. data points in a single class. However, as a tree grows in size, it becomes increasingly difficult to maintain this purity, and it usually results in too little data falling within a given subtree. When this occurs, it is known as data fragmentation, and it can often lead to overfitting.

A. Attribute Selection

The primary challenge in the Decision Tree implementation is to identify the attributes which we consider as the root node and each level. This process is known as the attributes selection. There are different attributes selection measure to identify the attribute which can be considered as the root node at each level.

There are 2 popular attribute selection measures. They are as follows: 1. Information Gain 2. Gini index

While using Information gain as a criterion, we assume attributes to be categorical and for Gini index attributes

are assumed to be continuous [5]. These attribute selection measures are described below.

B. Information Gain

[4] By using information gain as a criterion, we try to estimate the information contained by each attribute. Information Gain is based on a concept called Entropy.

Entropy measures the impurity in the given dataset. In Physics and Mathematics, entropy is referred to as the randomness or uncertainty of a random variable X. In information theory, it refers to the impurity in a group of examples. Information gain is the decrease in entropy. Information gain computes the difference between entropy before split and average entropy after split of the dataset based on given attribute values.

$$Entropy = \sum_{i=1}^{c}(-P_i * log_2(P_i))$$

Here, c is the number of classes and P_i is the probability associated with the i_{th} class.

The ID3 (Iterative Dichotomiser) Decision Tree algorithm uses entropy to calculate information gain. So, by calculating decrease in entropy measure of each attribute we can calculate their information gain. The attribute with the highest information gain is chosen as the splitting attribute at the node.

$$Informationgain(S, a) =$$

$$Entropy(S) - \sum_{i=1}^{k}(\frac{|S_i|}{|S|} * Entropy(S_i))$$

where a represents a specific attribute or class label, $Entropy(S)$ is the entropy of dataset, S, $|S_i|/|S|$ represents the proportion of the values in S_i to the number of values in dataset, S and $Entropy(S_i)$ is the entropy of dataset, S_i

C. Gini Index

Another attribute selection measure that CART (Categorical and Regression Trees) uses is the Gini index. It uses the Gini method to create split points. It is given by

$$Gini = 1 - \sum_{i=1}^{c}(P_i)^2$$

Here, again c is the number of classes and P_i is the probability associated with the i_{th} class.

Gini index says, if we randomly select two items from a population, they must be of the same class and probability for this is 1 if the population is pure.

It works with the categorical target variable "Success" or "Failure". It performs only binary splits. The higher the value of Gini, higher the homogeneity. CART (Classification and Regression Tree) uses the Gini method to create binary splits.

IV. Modelling

Decision tree is a hierarchical tree structure classification and regression model [1], [2]. It consists of root node which further splits into child nodes and at the end it produces leaf nodes which represent all possible outcomes. There are two different methodologies have been implemented for attribute selection which are gini and information gain. The model gives same accuracy 80.53% with both methodologies. Since the training set accuracy is 80.53% and test set accuracy is 78.48%, hence they are comparable. Thus, it can be concluded that model is not suffering from either overfitting or underfitting. In order to evaluate, confusion matrix has been used and f-score, recall and precision have been found out for decision attribute. Finally, decision tree has been generated to visualize the top-down splitting as shown in below figures.

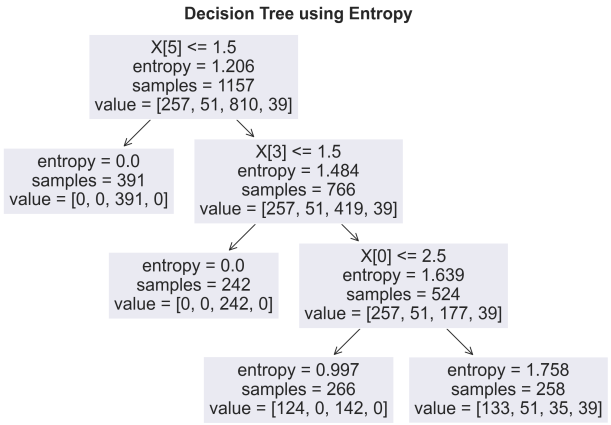

Fig. 8. Decision tree diagram with entropy

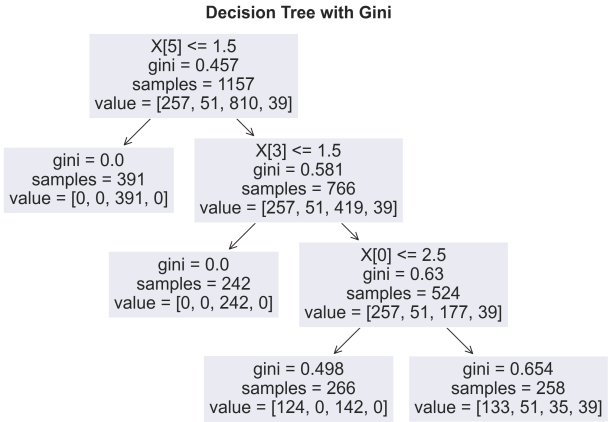

Fig. 9. Decision tree diagram with gini

Acknowledgement

It is my pleasure to express with deep sense of gratitude to Professor Gaurav Raina, for his guidance, continual encouragement, understanding.

I would like to extend my gratitude to IIT Madras, for providing with an environment to work in and for his inspiration during the tenure of the course.

CONCLUSION

In this paper, decision tree has been used to evaluate the car based on safety measures from the given data set. Data analysis and visualization has been implemented to visualize the correlation among the variables and cause-effect. It has been found out that the accuracy of model is 80.53% with information gain as well as gini indices. The fact which has been derived from analysis is that car evaluation more sensitive towards the safety measures as if safety is compromised then car will be categorized as unacceptable. There are some features such as doors, and luggage size do not have so much impact on car evaluation. This paper also helps new researcher to get more information about decision tree algorithm and its related terminologies.

REFERENCES

[1] A. Sagu, "Machine Learning Decision Tree Classifier and Logistics Regression Model," International Journal of Advanced Trends in Computer Science and Engineering, vol. 9, no. 1.4, 2020, doi: 10.30534/ijatcse/2020/2491.42020.

[2] L. Rokach and O. Maimon, "Top-down induction of decision trees classifiers - A survey," IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews, vol. 35, no. 4, 2005, doi: 10.1109/TSMCC.2004.843247.

[3] B. Charbuty and A. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," Journal of Applied Science and Technology Trends, vol. 2, no. 01, 2021, doi: 10.38094/jastt20165.

[4] A. Trabelsi, Z. Elouedi, and E. Lefevre, "Decision tree classifiers for evidential attribute values and class labels," Fuzzy Sets Syst, vol. 366, 2019, doi: 10.1016/j.fss.2018.11.006.

[5] S. Tangirala, "Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm," International Journal of Advanced Computer Science and Applications, no. 2, 2020, doi: 10.14569/ijacsa.2020.0110277.

[6] J. Ababneh, "Application of Naïve Bayes, Decision Tree, and K-Nearest Neighbors for Automated Text Classification," Mod Appl Sci, vol. 13, no. 11, 2019, doi: 10.5539/mas.v13n11p31.

[7] Z. Zuo, "Sentiment Analysis of Steam Review Datasets using Naive Bayes and Decision Tree Classifier," Student Publications and Research - Information Sciences, 2018.

[8] A. Ammar, "A Decision Tree Classifier for Intrusion Detection Priority Tagging," Journal of Computer and Communications, vol. 03, no. 04, 2015, doi: 10.4236/jcc.2015.34006.

[9] S. Mohammadi, H. Mirvaziri, M. Ghazizadeh-Ahsaee, and H. Karimipour, "Cyber intrusion detection by combined feature selection algorithm," Journal of Information Security and Applications, vol. 44, 2019, doi: 10.1016/j.jisa.2018.11.007.

[10] H. Alqahtani, I. H. Sarker, A. Kalim, S. M. Minhaz Hossain, S. Ikhlaq, and S. Hossain, "Cyber intrusion detection using machine learning classification techniques," in Communications in Computer and Information Science, 2020, vol. 1235 CCIS. doi: 10.1007/978-981-15-6648-6_10.

[11] M. R. Hasan, N. A. A. Bakar, F. Siraj, M. S. Sainin, and M. S. Hasan, "Single decision tree classifiers ' accuracy on medical data," Proceedings of the 5th International Conference on Computing and Informatics, ICOCI 2015, no. 188, 2015.

[12] I. Polaka, I. Tom, and A. Borisov, "Decision Tree Classifiers in Bioinformatics," Scientific Journal of Riga Technical University. Computer Sciences, vol. 42, no. 1, 2011, doi: 10.2478/v10143-010-0052-4.

[13] D. Lavanya and K. U. Rani, "Performance Evaluation of Decision Tree Classifiers on Medical Datasets," Int J Comput Appl, vol. 26, no. 4, 2011, doi: 10.5120/3095-4247

[14] J. Chopda, H. Raveshiya, S. Nakum, and V. Nakrani, "Cotton Crop Disease Detection using Decision Tree Classifier," 2018. doi: 10.1109/ICSCET.2018.8537336.

Source code link: https://drive.google.com/drive/folders/1Vt3-DpgtDe_Ooc20POwOybEbN4Wfntsh?usp=share_link