Assignment 4: A Mathematical Essay on a Decision Tree.

Mustefa Abrahim
Deprt. Data Science
Indian Institute of Technology Madras (IITM)
ge22m014@smail.iitm.ac.in

***Abstract*—The main goal of this paper is to predict car safety using decision tree algorithms. The decision tree algorithm works by making a decision based on the conditions of the features. Nodes represent conditions or tests on an attribute, branches represent the results of the tests, and leaf nodes represent decisions based on the conditions. [3] We predict the safety of cars based on a variety of factors such as population, cost of purchase, number of doors, and maintenance costs, among others. We concluded with 95% accuracy in predicting car safety.**

Keywords: Decision tree, supervised learning, modeling, Encoding.

I. Introduction

A Decision Tree is a Supervised learning technique that can be applied to both classification and regression problems, but it is most commonly applied to classification. [1] It's a tree-structured classifier, with internal nodes representing dataset features, branches representing decision rules, and each leaf node representing the outcome. The Decision Node and the Leaf Node are the two nodes in a Decision tree. Decision nodes are used to make decisions and have multiple branches, whereas Leaf nodes are the outcomes of those decisions and do not. The characteristics of the given dataset are used to make decisions or run tests. [2]

A. *Technical Aspects of Decision Tree*

What is the Decision Tree algorithm? In a decision tree, the algorithm begins at the root node and works its way up to predict the class of a given dataset. This algorithm compares the values of the root attribute with the values of the record (real dataset) attribute and then follows the branch and jumps to the next node based on the comparison. The algorithm compares the attribute value with the other sub-nodes and moves on to the next node. It repeats the process until it reaches the tree's leaf node. [2] We can use the following steps to better understand the technical aspects of the Decision tree algorithm.

Step 1: Begin the tree with the root node, which contains the entire dataset, says S.

Step 2: Using the Attribute Selection Measure, find the best attribute in the dataset.

Step 3: Subdivide the S into subsets containing potential values for the best attributes.

Step 4: Create the decision tree node with the best attribute.

Step 5: Create new decision trees recursively using the subsets of the dataset created in step 3. Continue this process until you reach a point where you can no longer classify the nodes and refer to the final node as a leaf node. [3]

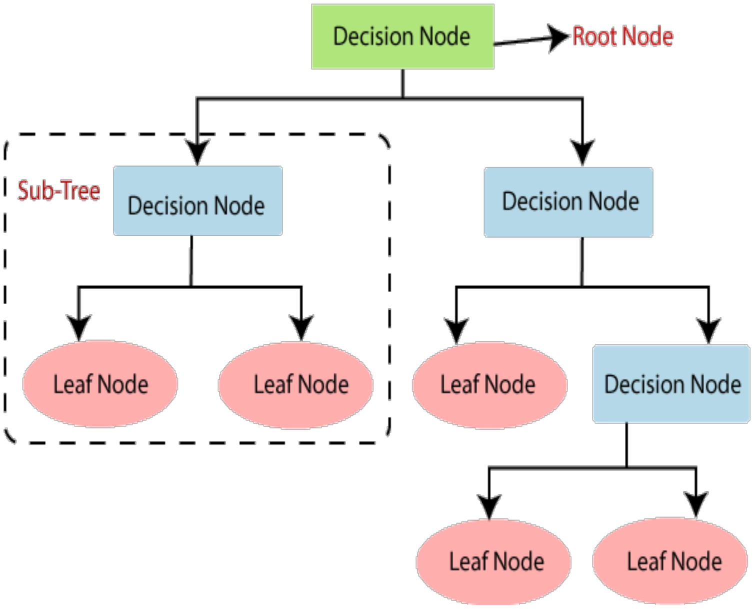

Figure 1: Decision Tree Model

The root node: is the point at which the decision tree begins. It represents the entire dataset, which is divided further into two or more homogeneous sets.

Leaf Node: Leaf nodes are the tree's final output node, and the tree cannot be further separated after obtaining a leaf node.

Splitting: Splitting is the process of dividing the decision node/root node into sub-nodes based on the conditions specified.

A branch/subtree: is a tree that is formed by splitting the tree.

B. Predicting Safety of Car

Our main goal is to evaluate car safety based on various decision nodes such as maintenance, doors, people, buying, safety, and predicting the target. [1] First, we checked our given dataset by using the info method, description, and isna.count() function to see if it had a null value or not. Then, because our given data was categorical, we replaced it with encoded data in places where there was no encoded data, such as low replace by 0 and medium replace by 1, and high replace by 3, then classified it into the decision and predict parts, and we classified in to train and test data by an 80:20 ratio and predicted the safety of the car. [3]

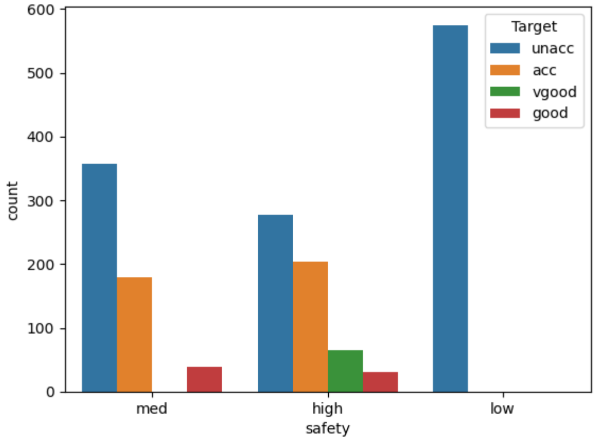

Figure 2: Visualisation before Encoded:
This is a visualization of our data before encoding, and it contains categorical data such as good, very good, accurate, and inaccurate, which we replaced with encoded data and predicted the car's safety. As we can see from the graph, it had some values before it was encoded, such as a number, and categorical data such as accurate, inaccurate, very good, and good. [1]

In the Car evaluation data, we have many decision nodes that are used to predict the leaf node or output such as the number of people it can carry, the number of doors it has, the cost of maintenance, and the safety of the car.

II. Data

The Car Evaluation Database was derived from a simple hierarchical decision model. The prediction task is to classify a car based on its safety. [3] The car dataset is divided into two parts. The first is Car Acceptability, and the second is Technical Characteristics. The overall price of buying and the cost of maintenance are two aspects of car acceptability. The number of doors (doors), the capacity in terms of people carried, the size of the luggage boot, and the car's estimated safety. [2]

- Number of occurrences: 1727
- The number of attributes: 7.
- Null Value: no.

The table below shows some attributes of our dataset in the Car evaluation we used to predict car safety. As we can see in this table, we have some integer values and some character values with different data types, so to do the decision tree we have to make the same type otherwise it will be difficult to give the decision on the different data types, so we encoded to the number and made the decision on it. [2]

Attribute	Attribute values
Buying	high,high,med,low
maint	high,high,med,low
doors	2,3,5more
persons	2,4,,more
lug_boot	2,3,4,more
safety	small,med,med,big
Target	uacc,acc,good,vgood

Table 1: Attribute and its values in Car Evaluation

III. Modeling and Visualization

We predict the safety of car evaluation based on some of the decision conditions given our data like people, doors, maintenance, how many bags it carries, and cost of buying, and we observed before encoding the given dataset have some negative correlations which mean when one variable increasing others is decreasing such things can affect our decision we encoded that and after encoding there is no negative correlation, so the problem of negative correlation was fixed and it became good for the decision to made and finally, we got the accuracy 95% we predict the car safety. [1]After encoded and classified into train and test data and done Decision tree analysis on it and we got the below decision tree

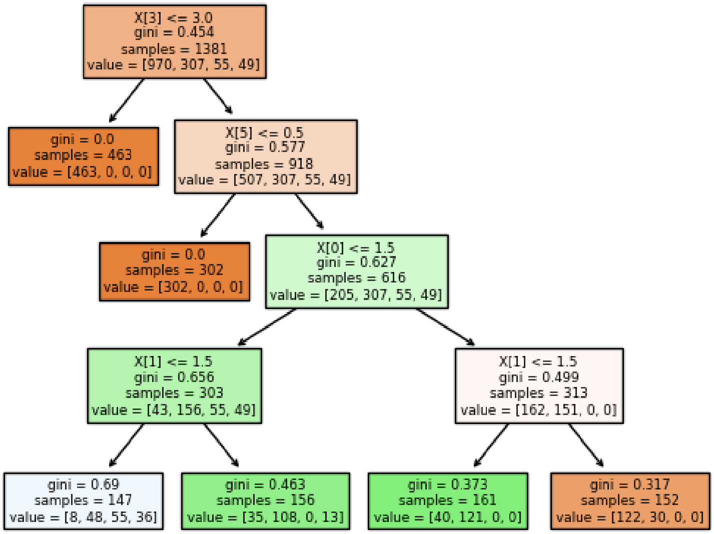

Figure 3: Our Decision Tree of Encoded datasets
In addition, we observed that Decision trees are less suitable for estimation tasks that require predicting the value of a continuous attribute. In classification problems with many classes and a small number of training illustrations, decision trees are prone to errors.

IV. Conclusions

In this paper, we did the prediction of Car safety based on the different conditions of the decision by the Decision tree algorithm method and finally, we got 95% accuracy for the Cars safety, we used the train and test classification by the ratio of 80 to 20.

References

[1] . Sorower MS. A literature survey on algorithms for multi-label learning. Oregon State University, Corvallis. 2010 Dec;18.

[2] . Utku A, Hacer (Uke) Karacan, Yildiz O, Akcayol MA. Implementation of a New Recommendation System Based on Decision Tree Using Implicit Relevance Feedback. JSW. 2015 Dec 1;10(12):1367-74.

[3] . Gershman A, Meisels A, Lüke KH, Rokach L, Schclar A, Sturm A. A Decision Tree Based Recommender System. InIICS 2010 Jun 3 (pp. 170-179).