# DAL 2023 Assignment 4: A Mathematical Essay on Decision Trees

Arun Palaniappan
*Department of Mechanical Engineering*
*Indian Institute of Technology Madras*
Chennai, India
me20b036@smail.iit.ac.in

***Abstract*—This paper provides an overview of the mathematical formulations and applications of decision trees. First, the fundamentals of decision trees are explored, as well as the specifics necessary to comprehend the research. The problem statement and data-set are introduced in a detailed way. The data-set under consideration is one of the evaluation of cars, to explore which feature of the car is more contributing to its acceptability. To accomplish this efficiently, the data has been thoroughly evaluated, and both visual and quantitative insights have been provided in this paper. Finally, the results of the decision tree model are reported to better understand which features are essential when associating maintenance cost, safety and buying price with the acceptability of the car. The "Data" section has been improved in this paper by doing extensive and exhaustive visual analysis.**

***Index Terms*—decision trees, precision, recall, car, acceptability, safety, visualization**

## I. Introduction

*Decision tree* is a statistical and non-parametric technique employed to model and predict the association between various factors or variables. It falls under the category of supervised learning algorithms, used for both classification and regression tasks, such as spam detection. The objective of decision trees is to determine the set of conditional control statements to represent the training data in a tree like structure.

This is accomplished by an intelligent divide and conquer algorithm. We split the data-set based on the value of a feature to attain maximum purity on the sub-sets. We define this purity mathematically by using different functions like *GINI-Index* and *Entropy*. To test how our model will perform on unforeseen data-sets in the future, we build a train-validation split, then train the model (minimize the impurity) on the train split, and then check the prediction accuracy on the validation split.

The "Car Evaluation" data-set contains information about various attributes of cars, such as their maintenance costs, the number of doors, the number of passengers that can be seated, luggage capacity, safety rating, and the overall acceptability of the car. This data-set is useful, particularly in the context of car quality assessment.

In this paper, a comprehensive understanding of decision tree fundamentals are provided. The paper also undertakes a detailed exploration of the targeted problem and elucidates how the resulting insights are supported by a combination of visual representations and quantitative evaluation.

## II. decision trees

In this section, we will describe the mathematical details of decision trees. It is a supervised learning algorithm used for modeling the relationship between two types of variables, the target(dependent) and features(independent). The following is an overview of the jargons and mathematics utilised in the decision trees approach:

- *Root Node:* The top node of the decision tree, from which all other nodes are born. It represents the entire training data-set.
- *Node:* It represents a decision point in the tree, where the data is split into subsets based on the outcome of the purity test.
- *Leaf Node:* It is the node at the end of a branch, where a final prediction is made.
- *Parent and child node:* Child nodes are born from the parent node by a conditional control statement.
- *Purity:* It denotes the variance of the data-set at that particular node. If the purity at a node is poor, we need to split the node to make it more pure.
- *Information gain:* It is a mathematical quantity for representing how mach variance is being explained or information is being gained, by making a particular split.
- *Max depth:* It is the maximum depth to which the splitting of nodes occurs.

### A. Assumptions in decision trees

Decision tree is a powerful statistical technique, but its validity and reliability rests upon certain assumptions that must be met for accurate and meaningful results. Firstly, the entire training data-set is considered as the root node. Secondly, if the feature values are continuous, it needs to be discretized. A statistical method is used to decide which attributes will be the leaf or internal nodes of the tree. This strategy aims to choose attributes that increase information gain.

### B. Types of purity metrics

- *GINI-Index:* It is a measure of impurity due to a split in the intermediate nodes. It is calculated as,

$$\text{GINI-Index} = 1 - \sum_{i=1}^{n} (p_i)^2 \tag{1}$$

A weighted sum of all GINI-indices are calculated for all the features, then the feature with the lowest GINI-impurity is chosen to make the split at that node.

- *Entropy information gain:* It denotes randomness due to a split at the intermediate nodes. It is calculated as,

$$\text{Entropy} = -\sum_{i=1}^{n} p_i * log(p_i) \tag{2}$$

Information Gain is calculated as the difference of the entropy of the parent node and the weighted average of the entropies of the child nodes:

$$\text{Gain} = Entropy(P) - (Entropy(C))_{avg} \tag{3}$$

The information gain for all the features are calculated and the feature with the highest gain is chosen to make the split at that node.

### C. Algorithm

We will now explain the steps involved in modeling a decision tree:

- We begin with the entire data-set in the root node of the tree.
- We use one of the two purity measures defined above to find the best feature to split the data-set at the current node.
- Generate the child nodes from the parent node using the above criterion.
- We repeat step-2 and step-3 to continue splitting the nodes until a stopping criterion like the max depth of the tree, minimum number of samples per leaf is met or there is no further significant improvement in the information gain.

### D. Evaluation Metrics

We now define two quantities which can be used to evaluate the model.

- *Weighted average of $F_1$-scores ($F_{1,avg}$):* It is defined as the weighted average of $F_1$ scores of all the target classes, where the weights are the support ($S_i$) of each class.

$$F_{1,i} = \frac{2 * precision_i * recall_i}{precision_i + recall_i} \tag{4}$$

$$F_{1,avg} = \frac{\sum_{i=1}^{i=n} S_i * F_{1,i}}{\sum_{i=1}^{i=n} S_i} \tag{5}$$

- *Accuracy:* It is the total number of correct predictions divided by total number of validation points,

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

## III. The Data

A data-set with information about various attributes of cars, such as their maintenance costs, the number of doors, the number of passengers that can be seated, luggage capacity, safety rating, and the overall acceptability of the car are provided to us. The ultimate aim of the study is to identify key relationships between a car's acceptability and its attributes.

### A. Data description

In this section, we describe the structure of the data-set. The data-set has 1728 entries. It has been split into train(80%)-validation(20%) data-set. The columns from the data-set are:

- *buying:* The buying price of the car {vhigh, high, med, low}.
- *maintenance:* The cost spent in maintenance of the car {vhigh, high, med, low}.
- *doors:* The number of doors in the car {2, 3, 4, 5, more}.
- *persons:* The capacity of the car in terms of the number of people that can be seated {2, 4, more}.
- *lug_boot:* The size of the luggage boot of the car {small, med, big}.
- *safety:* This contains the safety rating information of the car {low, med, high}.
- *target:* This is the target variable that we have to predict which is the acceptability of the car among people {unacc, acc, good, vgood}.

### B. Data cleaning

While cleaning the data, we had to handle categorical classes. Categorical classes were encoded by integers using *LabelEncoder* from scikit-learn. For instance, safety column was encoded using ['low':0, 'med':1, 'high':2].

### C. Data Visualization

*1) Correlation plot:* Fig. 1 shows the heat map of the correlation values between variables in the data-set:

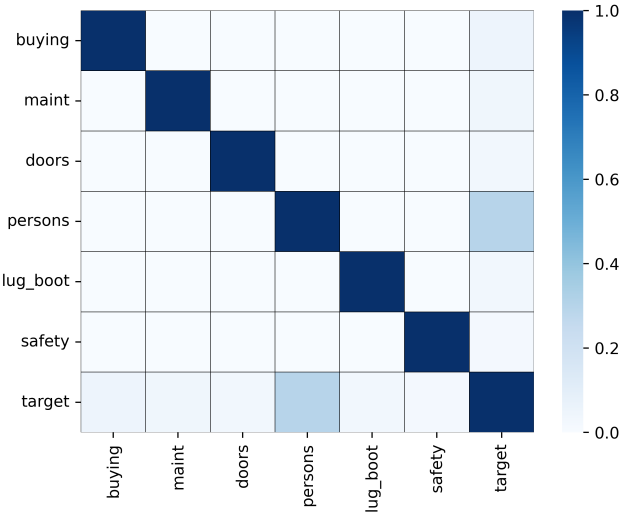


Fig. 1. Correlation heat-map within the variables

From the correlation plot we can conclude that:

- One obvious correlation is the *target* and the *number of passengers*. Higher the capacity, higher is the acceptability of the car.
- There are no other noticeable correlations among the variables.

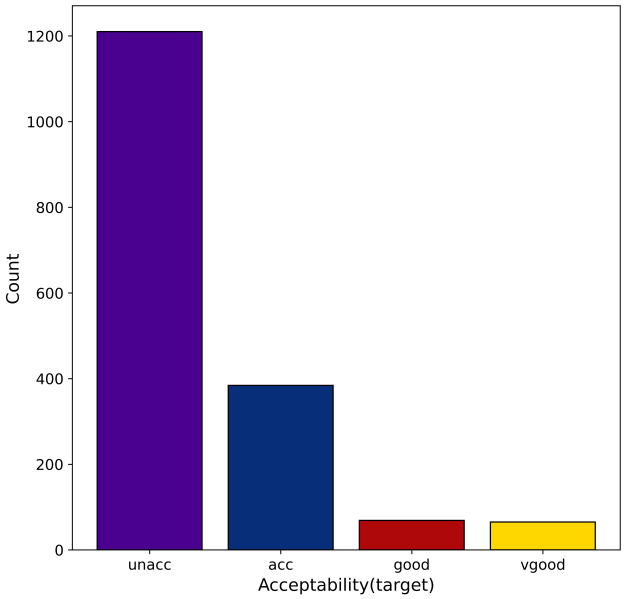


Fig. 2. Acceptability distribution

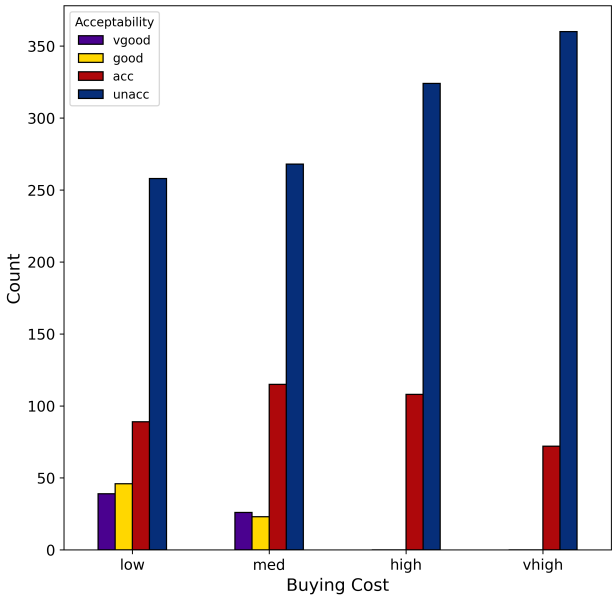


Fig. 4. Buying cost distribution vs Acceptability

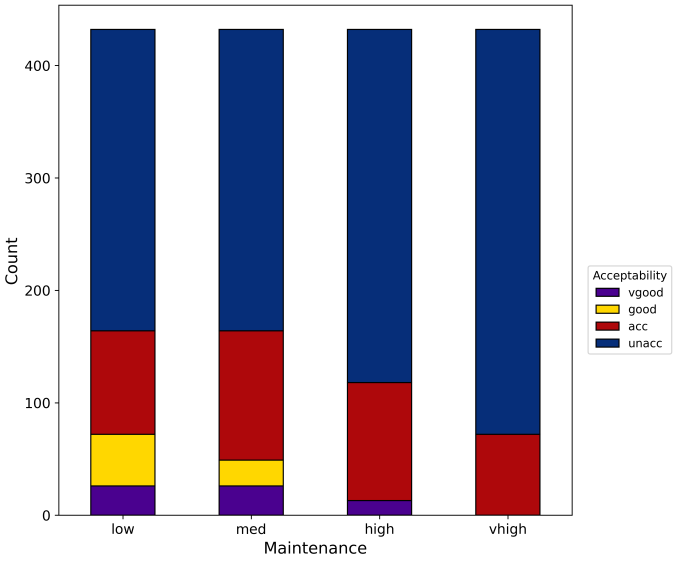


Fig. 3. Maintenance cost distribution vs Acceptability

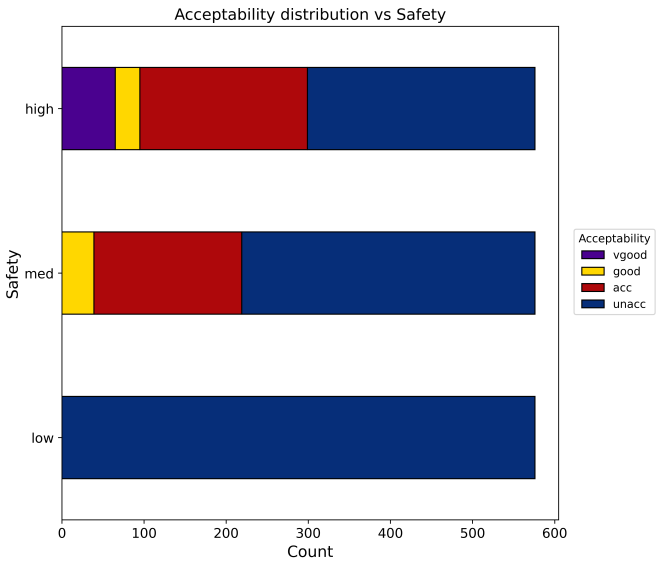


Fig. 5. Acceptability distribution and Safety

*2) Acceptability distribution:* From fig. 2, we can observe significant class imbalance because around 70% of the cars in the data are unacceptable, around 20% are acceptable and the other 10% are either good or very good.

*3) Maintenance cost distribution and Acceptability:* We can observe from fig. 3 that the majority of cars that are deemed as good or very good have low to medium maintenance costs. Hence, vehicles with high maintenance expenses are often viewed as unacceptable.

*4) Buying cost distribution and Acceptability:* Fig. 4 shows that all vehicles that are rated as good or very good have low or medium purchase costs. Thus, expensive cars are typically considered unacceptable.

*5) Safety distribution and Acceptability:* We can observe from fig. 5 that most good or very good cars have medium to high safety. Hence, vehicles with low safety are often viewed as unacceptable.

*6) Number of People and Acceptability distribution:* Fig. 6 shows that all vehicles that are rated as good or very good have a capacity of 4 or more. Thus, 2-seater cars are typically considered unacceptable.

## IV. The Problem

### A. Outline

We have done the analysing, cleaning and visualizing the data. A train-set and a validation-set are created from the data-

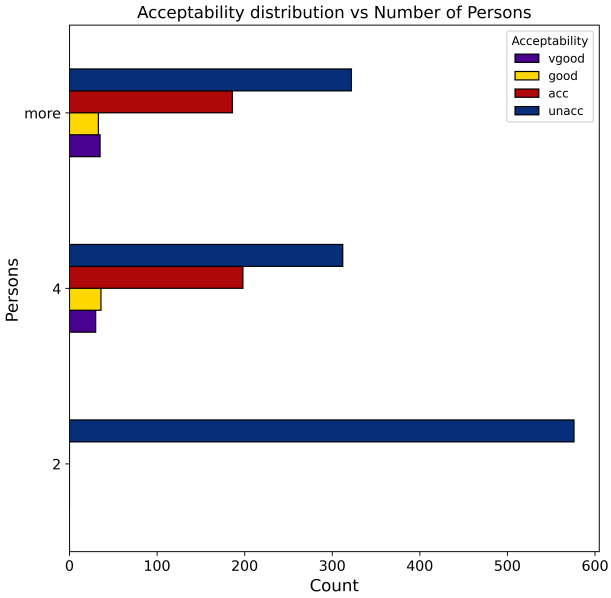


Fig. 6. Acceptability distribution vs Number of People

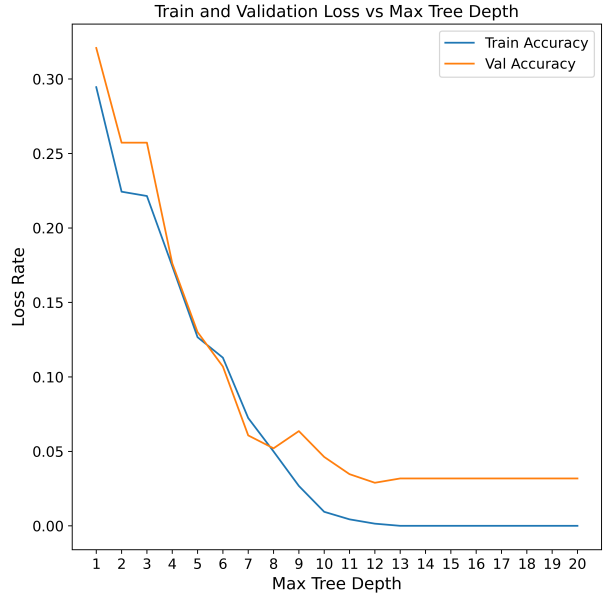


Fig. 7. Train and Validation Loss vs Max Tree Depth

set. The validation-set is used to make predictions after the Decision Tree model has been trained using the train-set. The validation-set predictions are used to calculate two different metrics to assess model performance.

### B. Decision tree modelling

Before fitting the decision tree to the training data by using *scikit-learn* package in python, we need to specify the *max_length* hyper-parameter of the tree, which denotes the maximum depth to which the tree branches. In order to accomplish this, we plot the train loss and validation loss against *max_length* as a variable. From fig. 7, we can observe an elbow point at *max_length=12*, achieving a very low loss and having no further improvement in accuracy with the increase in *max_length*.

We now instantiate a Decision Tree model, with hyper-parameter (*max_length=12*), for predicting the acceptability. The predictor variables that we use are *'buying', 'maint', 'doors', 'persons', 'lug_boot' & 'safety'*. We use weighted average of $F_1$-scores ($F_{1,avg}$) and *Accuracy* score as our evaluation criteria on the validation data-set. After testing on the Validation Split, the model obtains a $F_{1,avg}$-score of $0.97$ and also an *Accuracy* score of $0.97$.

### C. Confusion matrix

Fig. 8 represents the confusion matrix generated and gives the count to all the possible outcomes in the validation set(20% of the data).

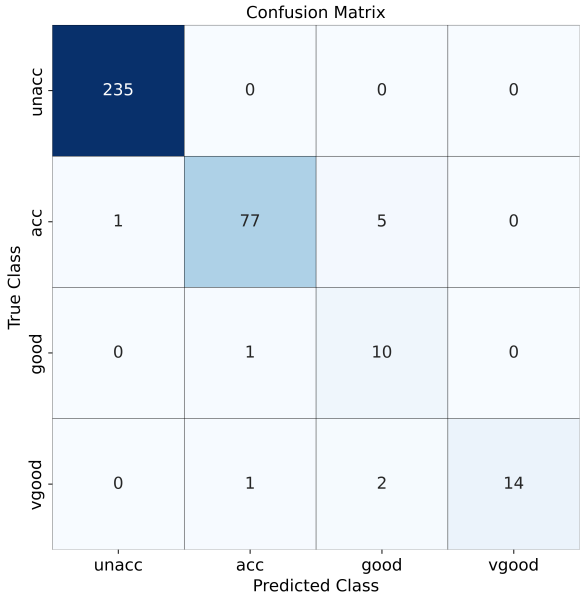


Fig. 8. Confusion Matrix

## V. Conclusion

The conclusions that can be drawn from all the above analysis are:

- Irrespective of which purity parameter that is used (GINI-Index or Entropy), we almost had the same accuracy in both the cases.
- The number of doors does not affect the acceptability of cars in general.
- Two seater cars are generally considered unacceptable because the majority of the population have families.
- The acceptability of a car is greatly influenced by its safety rating. All vehicles categorized as very good have a high safety rating, whereas all vehicles with low safety ratings are unacceptable.
- Better levels of acceptability are typically found in vehicles with more luggage space.

Future research could use *Random-Forrest* along with boosting and bagging, which is an ensemble of a large number of decision trees, each trained on a different subset of the data-set, so as to not overfit it. We can also use *One-Hot Encoding* for categorical variables, that is creating dummy columns for denoting the presence of a class to avoid any unnecessary bias. We have observed that, there is class imbalance in the acceptability column. This can create inherent bias in the model, hence we can adopt class-balancing techniques to overcome this bias. We could also explore on using regularised techniques such as adding a penalty for exceeding the max tree depth limit, because deeper trees tend to overfit the data-set.

## References


[1] "Decision Tree." Wikipedia, Wikimedia Foundation, 28 Sept. 2023, *https://en.wikipedia.org/wiki/Logistic_regression* (accessed 30 Sept. 2023)

[2] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. *An Introduction to Statistical Learning : with Applications in R*. New York :Springer, 2013.