# A mathematical essay of Decision Trees

Devansh Sanghvi
*Department Of Biotechnology*
*Indian Institute of Technology, Madras*
Chennai, India
be19b002@smail.iitm.ac.in

***Abstract*—This paper aims to develop an understanding of the mathematical properties of decision trees and showcase its application in classifying a car based on its safety.**

## I. Introduction

**Decision Trees** is one of the most important tools used for classification and prediction. Once we receive data with some independent variables and a target variable which we want to determine for a given data point, we use statistical tools to develop a relationship between the target and the independent variables to decide on mathematical rules to assign a given data point to a class. The decision tree classifier uses a tree like structure to perform this classification.

Decision trees have applications spanning multiple fields . In general, they are constructed using an algorithmic approach that identifies ways to split the data set. There are two common attribute selection measures, namely: entropy and Gini Index. The formula and their mathematical interpretation is given below.

$$Gini = 1 - \sum_{i=1}^{n} p^2(c_i)$$

$$Entropy = -\sum_{i=1}^{n} p(c_i) log_2(p(c_i))$$

Where $p(c_i)$ is the probability of class i being in a node, and n is the number of classes in the target variable. Gini index and entropy are the criterion for calculating information gain. Decision tree algorithms use information gain to split a node. Entropy in statistics is analogous to entropy in thermodynamics and it signifies disorder. If there are multiple classes in a particular node, then that node is disordered. Information gain is the entropy of parent node minus sum of weighted entropies of child nodes. Weight of the entropies are number of samples in the node divided by total samples in the child nodes. Similarly, information gain can be calculated using Gini Index.
The gini impurity measures the frequency at which any element of the data set will be mislabelled when it is randomly labelled.

The attribute selection measures are used to select the splitting attribute used at the node. The structure of a decision tree is shown below.

To demonstrate the effectiveness of the Decision tree classifier, we use the standard features of a car like the buying price, maintenance cost and luggage boot size to classify the safety level of cars as unacceptable, acceptable, good or very good. The pertinent data has been collected by Marko Bohanec and Blaz Zupan [1]. In the following sections we will dive a little deeper into decision tree classifiers, visualize the data collected, analyze it and then perform our model to classify new data points.

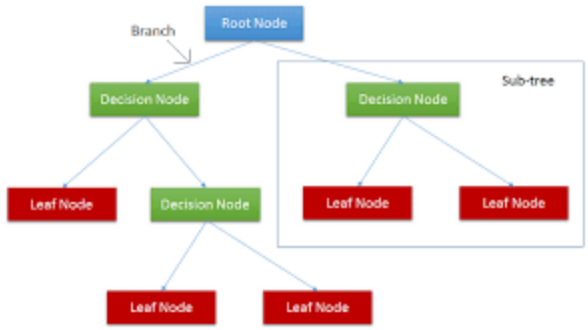


Fig. 1. The structure of a decision tree

This paper will serve as a part of the requirements to be fulfilled for receiving credits in the course **EE4708: Data Analytics Laboratory**.

## II. Decision Trees

As mentioned in the section above, decision trees use the concepts of Gini Index and Entropy to classify between the multiple classes and designate mathematical rules for each class. There are two main differences between the Gini Index and entropy:

- The range of gini index is $[0, 0.5]$, whereas the range of entropy is $[0, 1]$. Maximum entropy and gini will be when both the classes in a binary classification have a probability of 0.5 for being in a particular node. Minimum gini and entropy are when one node contains only one class, i.e. the node is pure. For the best classification in a node, we would ideally want the minimum gini/ entropy depending on the parameter we use. The distribution of gini, entropy with probability of a class being in a node is shown below.
- Computationally, since the entropy requires algorithms, it is more expensive than calculating gini index. Calculating gini index is faster.

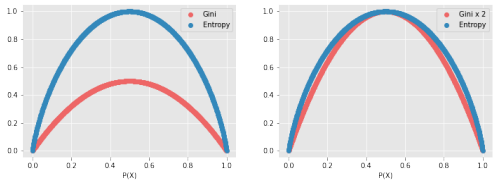


Fig. 2. The distribution of gini/ entropy with probability of a class

We will now look at some important advantages, disadvantages and metrics of evaluation for decision trees over other alternate machine learning algorithms that can be used for classification/ regression:

- Advantages
  1) Decision trees require much lesser data preparation than other supervised learning methods.
  2) Decision trees do not require normalization of data
  3) Decision trees do not require scaling of data too.
  4) Missing values in a decision tree do not affect the process of building a tree too.
  5) Decision trees are easier to explain and visualize than most machine learning algorithms.
- Disadvantages:
  1) A small change in the data can change the entire structure of a tree. Very volatile, and high variance in expectations.
  2) Calculations can turn out to be much more complex than other algorithms.
  3) Takes higher time to train the model.
  4) Decision tree algorithm is inadequate for regression and in determining continuous values.
- Metrics of evaluation: As for many other machine learning algorithms, we can extract the classification report for our classification. The main metrics from the classification report that are of our interest are:
  1) Precision: How many values predicted to be in a certain class are in that class
  2) Recall: How many values in each class were given the correct label
  3) F1-score: weighted average of precision and recall

  The metric which you give the most importance to depends on your interest in false positives/ false negatives.

## III. Pre-processing and Exploratory Analysis on the Data

In this section, we will first pre-process the data to extract the important features in the data provided, then we will visualize the data to understand the effects of certain features and modify the data-set to optimize the results from our decision trees model.

### A. Pre-processing the data

The initial data-set has 1727 samples and 6 independent features, namely: the cost of buying the car, the cost for maintenance, number of doors in the car, number of persons that can sit in the car, how large is the luggage boot and predicted safety of the car. . Pre-processing the data-set includes (Note that the changes are made in both the training and testing data-sets):

1) Missing value assessment: No features in our data set have missing values, hence we can skip this part of pre-processing.
2) Additional Variables: No features seem to depend on each other and hence, it is safe to include every feature in our model.

### B. Visualization

After preprocessing the data, we visualize the data, to better understand the correlation between different features:

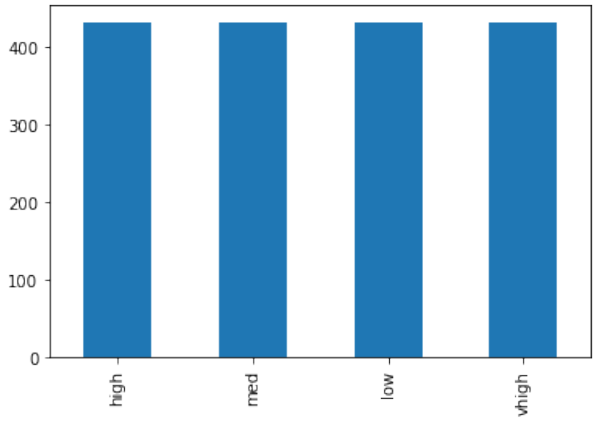


Fig. 3. Count plot of the 'buying' column in the data set

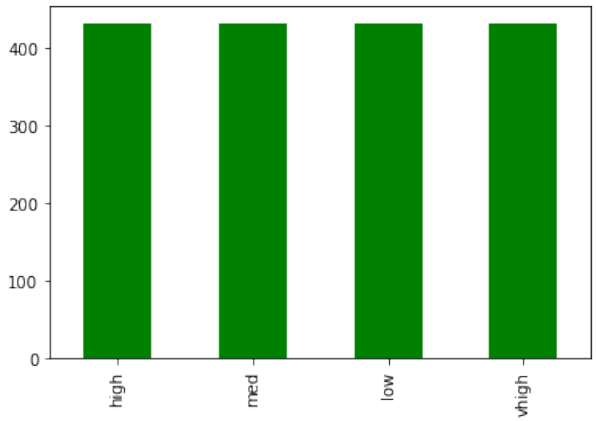


Fig. 4. Count plot of the 'maintenance' column in the data set

1) Exploration of the countplots: The independent variables are almost unskewed, with an almost equal distribution of the counts of all the independent variables. The dependent variable seems to show a large number of 'unacc' values, meaning that most of the cars are unacceptable with respect to their features.
2) Countplot of the luggage boot size per target: As the luggage boot size increases, the acceptance rate for the

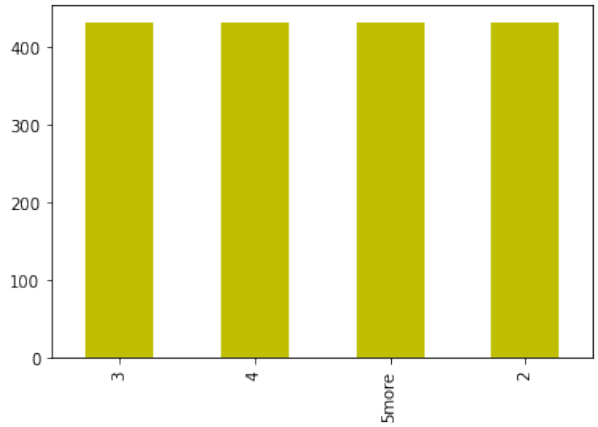


Fig. 5. Count plot of the 'doors' column in the data set

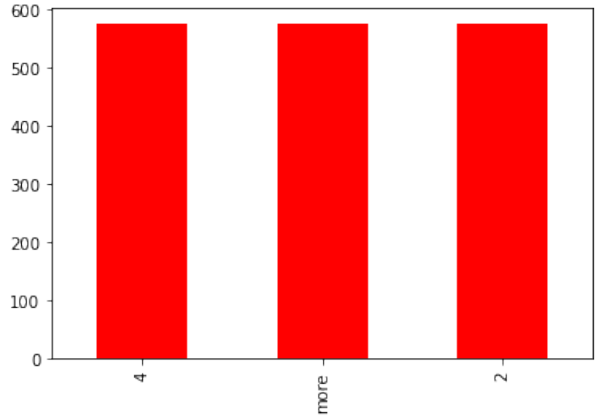


Fig. 6. Count plot of the 'persons' column in the data set

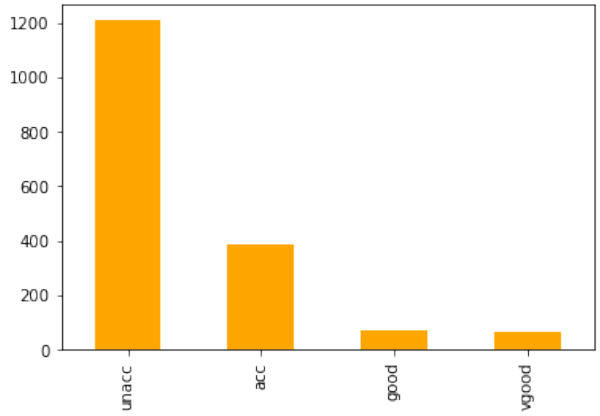


Fig. 7. Count plot of the 'target column in the data set

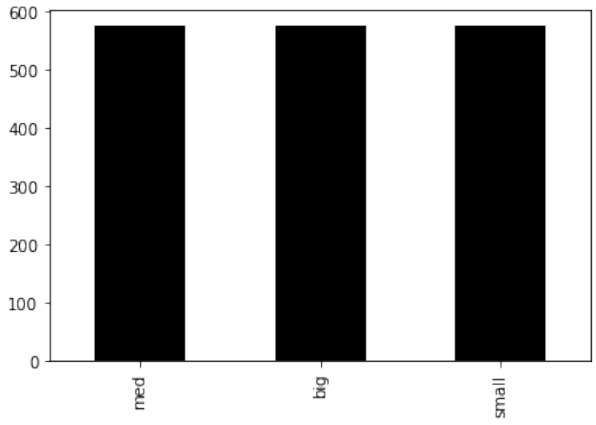


Fig. 8. Count plot of the 'luggage boot' column in the data set

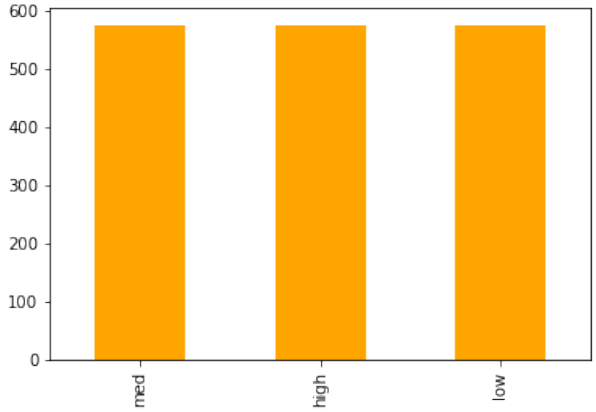


Fig. 9. Count plot of the 'safety' column in the data set

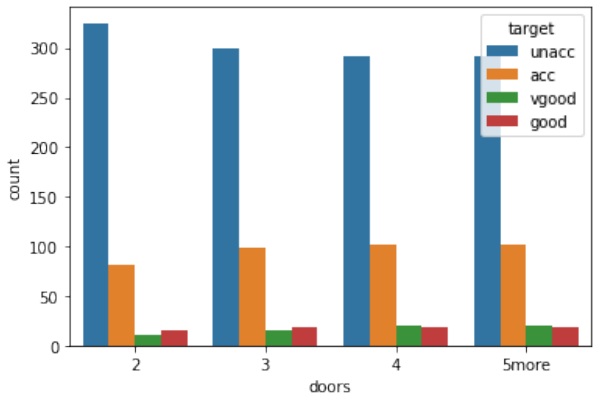


Fig. 10. Count plot of the number of doors per target

cars increases. The number of cars that are unacceptable are similar in number, but decrease as the luggage boot size increases.

3) Countplot of the luggage boot size per target: All cars with a low predicted safety rating are unacceptable and should be rejected. More the predicted safety, higher the acceptance rate amongst those cars.
4) Trend amongst the cost for unacceptable cars: Surprisingly, most cars which are unacceptable cost really large amounts of money, and the acceptance rate increases as the cost decreases.

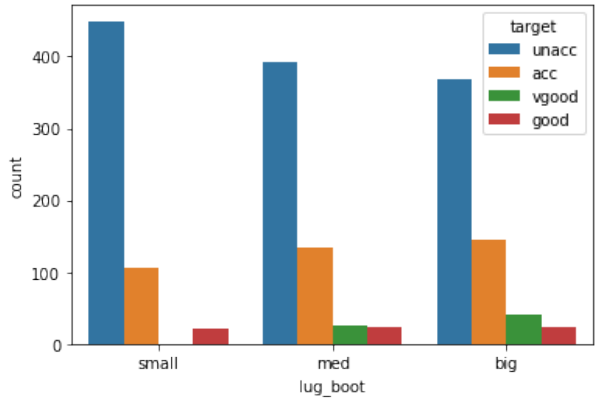


Fig. 11. Count plot of the luggage boot size per target

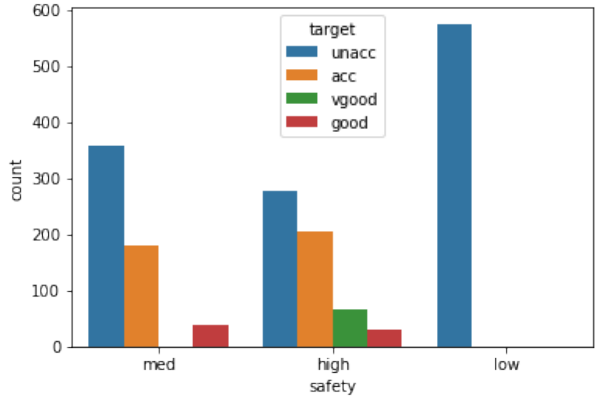


Fig. 12. Plot to compare the predicted safety levels and the target values

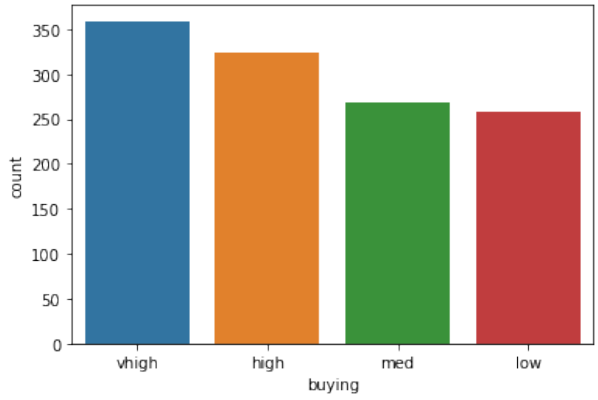


Fig. 13. Count plot of the buying column, when the target column is set to 'unacc'

## IV. The Decision Trees Model

After preprocessing the data, we move to the heart of the problem: the prediction of whether an unseen car (not in the training data-set) should be accepted. Here are the steps for the same:

- Approopriate Data type: Before applying a decision trees model to our data, we will analyze the data types for all variables, and change them, when necessary. Initially, all variables are of the categorical category. The scikit-learn package used to build the model cannot use categorical variables, and hence we will have to undergo a conversion process. Each variable follow a rank and order and thus, we perform an ordinal encoding on the data to maintain the order. For instance, the size of the luggage boot has three categories: small, medium, and large which has an order.
- Decision model and results After the conversion of the data types, we run the decision trees model using the simple test/ train model, using the two criterion for classification: entropy and gini index and evaluate our results using the classification report. The decision tree we predict can also be visualized:

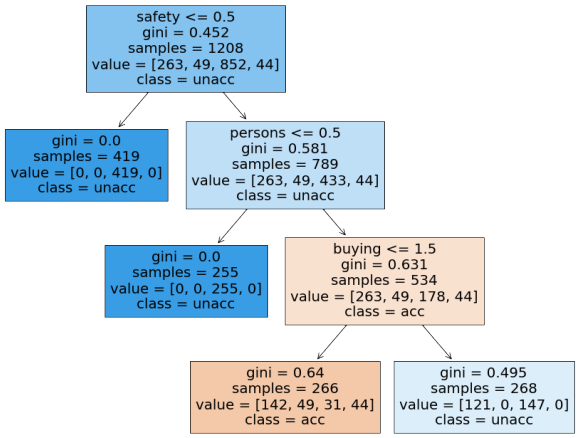


Fig. 14. Predicted decision tree visualization

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| acc | 0.53 | 0.51 | 0.52 | 121 |
| good | 0.00 | 0.00 | 0.00 | 20 |
| unacc | 0.85 | 0.96 | 0.90 | 357 |
| vgood | 0.00 | 0.00 | 0.00 | 21 |
| accuracy | | | 0.78 | 519 |
| macro avg | 0.34 | 0.37 | 0.36 | 519 |
| weighted avg | 0.71 | 0.78 | 0.74 | 519 |

Fig. 15. Classification report for the predicted decision tree using gini index

The results are the same for the two classification criterion: gini index and entropy. The precision, recall, f1 scores are 53%, 51%, and 52%, respectively, where the number of acceptable conditioned cars samples (support) is 121. The precision, recall, f1 scores are 0%, 0%, and 0%, respectively, where the number of good conditioned cars samples (support) is 20. The precision, recall, f1 scores are 85%, 96%, and 90%, respectively, where the number of unacceptable conditioned cars samples (support) is 357. The precision, recall, f1 scores are 0%, 0%, and 0%, respectively, where the number of very

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| acc | 0.53 | 0.51 | 0.52 | 121 |
| good | 0.00 | 0.00 | 0.00 | 20 |
| unacc | 0.85 | 0.96 | 0.90 | 357 |
| vgood | 0.00 | 0.00 | 0.00 | 21 |
| accuracy | | | 0.78 | 519 |
| macro avg | 0.34 | 0.37 | 0.36 | 519 |
| weighted avg | 0.71 | 0.78 | 0.74 | 519 |

Fig. 16. Classification report for the predicted decision tree using entropy

good conditioned cars samples (support) is 21. From this report, we can see that the model performs better for the unacceptable car samples than the acceptable cars, which is an interesting finding. Moreover, the good and very good sampled cars have zero precision, recall and f1-score. This means that the true positives, and the false negatives for the two conditioned cars are both zero. The two classes could have been avoided for better classification results.

The overall accuracy is 78%.

## V. Conclusion

The decision trees model was succesfully demonstrated, both visually and numerically. While the accuracy of the model is decent, the classification would have been better without the unneccesary classes, 'vgood' and 'good'. The results were the same for both the classification criteria: gini index and entropy.

## References

[1] "Car Evaluation Data Set", archive.ics.uci.edu, 1990 [Online]. Available: **Link**.