A Mathematical Essay on Decision Trees

Aryan Pandey
Department of Ocean Engineering
Indian Institute of Technology Madras
Chennai, India
na19b030@smail.iitm.ac.in

***Abstract*—This document is an overview of the mathematical aspects of Decision Tree as well as its application on a sample data set. The algorithm has been applied on a data set of cars and the prediction task is to classify a car based on its safety**

I. Introduction

Classification is a two-step process, learning step and prediction step, in machine learning. In the learning step, the model is developed based on given training data. In the prediction step, the model is used to predict the response for given data. Decision Tree is one of the easiest and popular classification algorithms to understand and interpret.

Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving regression and classification problems too. The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data(training data).

In Decision Trees, for predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with the record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node.

In this paper, the aim is to apply Decision Tree to classify a car based on its safety. The paper systematically goes through first the mathematical details of Decision Tree, the nature of the data set which has been given to us, then the problem we have in hand and how it has been solved, and finally the conclusions which were drawn. Useful insights and figures have been presented whenever necessary.

II. Datasets

The dataset given has the following details of multiple cars with the aim of classifying it based on its safety. It has the details of buying price, price of maintenance, number of doors, seating capacity, size of luggage boot and the estimated safety of the car.

The dataset consists of purely categorical columns in which the columns related to buying price, price of maintenance, and estimated safety of the car are categorised into very high, high, medium and low. The target condition of the car has 4 categories which are very good, good, acceptable and unacceptable.

Fig.1 shows that all these columns have a similar split across the categories. This similar observation can be made across all columns except for the target column where it is unbalanced.

		Count
buying	maint	
high	high	108
	low	108
	med	108
	vhigh	108
low	high	108
	low	108
	med	108
	vhigh	108
med	high	108
	low	108
	med	108
	vhigh	108
vhigh	high	108
	low	108
	med	108
	vhigh	108

Fig. 1. Dataset has similar splits

Fig.2 shows us the correlation of the features with each other as well as the correlation of the features with the target variable. We can see that the features have no correlation with each other whereas some features like safety and seating capacity have a good correlation to the target.

III. Models

This section discusses the mathematical and conceptual aspects of the Decision Tree algorithm.

A. Intuition

The Decision tree algorithm is a simple yet efficient supervised learning algorithm wherein the data points are continuously split according to certain parameters and/or the problem that the algorithm is trying to solve.

Every decision tree includes a root node, some branches, and leaf nodes. The internal nodes present within the tree describe the various test cases. Decision Trees can be used to solve both classification and regression problems. The algorithm can be thought of as a graphical tree-like structure that uses various tuned parameters to predict the results. The decision trees apply a top-down approach to the dataset that is fed during training.

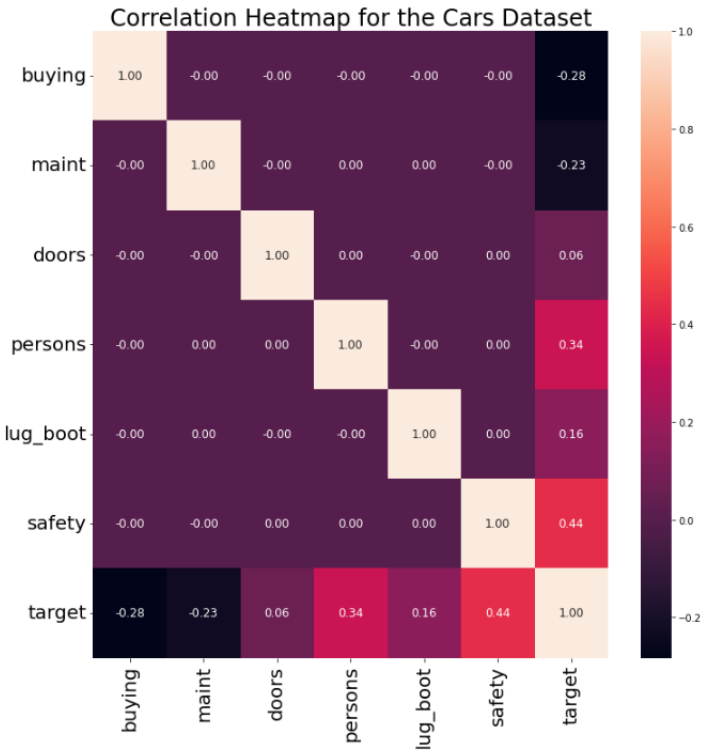

Fig. 2. Correlation Heatmap of features

B. The Algorithm

Entropy: Entropy is the amount of information needed to accurately describe the data. If the data is homogeneous, then the entropy is 0. Mathematically, entropy is written as:

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

Gini Index: It measures the impurities in the node. It has a value between 0 and 1. It is the sum of square if the probabilities of each class. It is formulated as:

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

- Root Node: It represents the entire population or sample and this further gets divided into two or more homogeneous sets.
- Splitting: It is a process of dividing a node into two or more sub-nodes.
- Decision Node: When a sub-node splits into further sub-nodes, then it is called the decision node. Leaf / Terminal Node: Nodes do not split is called Leaf or Terminal node.
- Pruning: When we remove sub-nodes of a decision node, this process is called pruning. You can say the opposite process of splitting.
- Branch / Sub-Tree: A subsection of the entire tree is called branch or sub-tree.
- Parent and Child Node: A node, which is divided into sub-nodes is called a parent node of sub-nodes whereas sub-nodes are the child of a parent node.

C. Types of Decision Tree Algorithms

There are 2 types of Decision tree algorithm. The 2 types are listed below:-

- Categorical Variable Decision Tree: Decision Tree which has a categorical target variable then it called a Categorical variable decision tree.
- Continuous Variable Decision Tree: Decision Tree has a continuous target variable then it is called Continuous Variable Decision Tree.

D. Assumptions

Below are some of the assumptions we make while using Decision tree:

- In the beginning, the whole training set is considered as the root.
- Feature values are preferred to be categorical. If the values are continuous then they are discretized prior to building the model.
- Records are distributed recursively on the basis of attribute values.
- Order to placing attributes as root or internal node of the tree is done by using some statistical approach.

E. Disadvantages

- They are not well suited to continuous variables.
- Usually, they provide a lower prediction accuracy than predictive algorithms.
- Over-fitting is a problem if the design of the tree is too complex.

IV. Modelling

In this section, we will explore how the model was applied to the dataset at hand and what we can infer from it. The categorical features were first encoded ordinally to ensure that the ranking of the categories was maintained in the encodings. We then split the dataset into a train and validation split of an 85:15 ratio and train the Decision Tree Classifier.

We obtain an accuracy of 97% and an F1 score of 0.9 on the dataset. The confusion matrix on the validation set is given in Fig. 3.

We can also find out the first few layers of the decision tree. This has been represented in Fig.4. As you can see the safety is the primary feature that is taken into account when plotting the tree.

V. Conclusions

Safety was one of the key factors in predicting the target variable for the car. The model is doing a good job of segregating all the categorical features into the different classes. This shows that the dataset has a good amount of separability and the misclassification rate is very low.

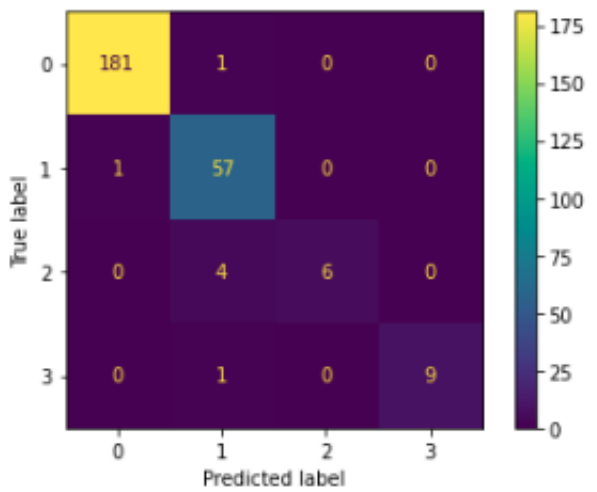

Fig. 3. Confusion Matrix of the Model Predictions

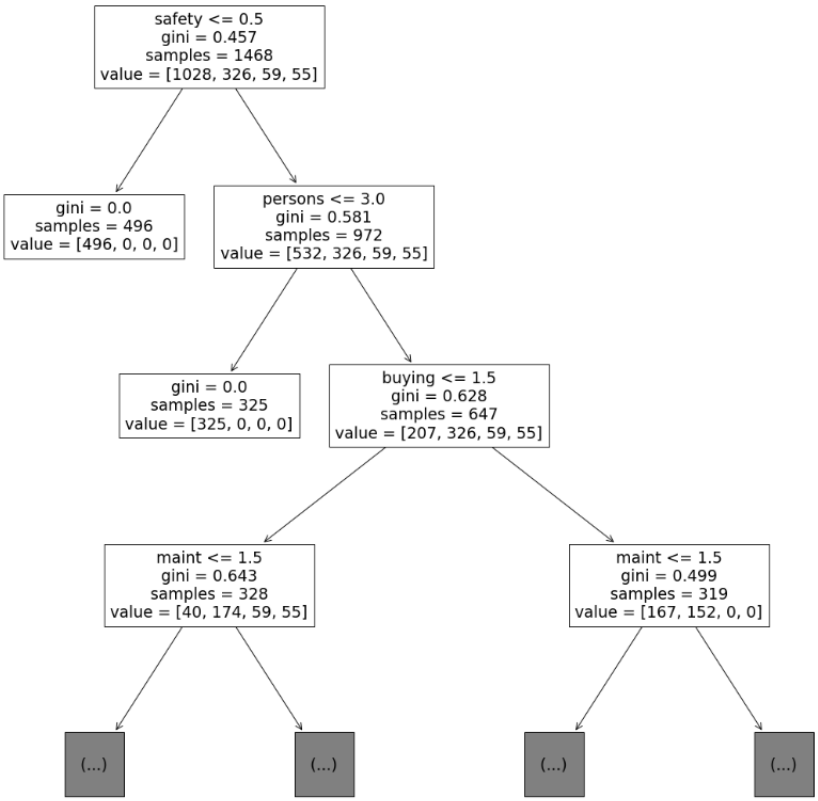

Fig. 4. Plotting the first few layers of the tree