Mathematical Essay on Decision Tree

Sparsh Tewatia
Department of Data Science, Joint MS(UoB)
I.I.T Madras
Chennai, India
sparshteotia5@gmail.com

***Abstract*—This mathematical essay presents a comprehensive exploration of decision trees, a fundamental algorithm in machine learning. We begin with a broad introduction to the concept and technical aspects of decision trees. We then delve into a specific problem: the classification of cars based on safety, using a Car Evaluation Database. The decision tree methodology is applied to this problem, and insights and observations are discussed. The paper concludes with a summary of key findings and suggestions for further investigation. The goal of this assignment is to explain mathematical ideas embedded within decision trees, and use this topic to understand a real-world problem particularly predict the car safety given various parameters.**

***Index Terms*—Decision Tree, Car-Safety Evaluation, Machine Learning, Analysis**

I. Introduction

Machine learning is a dynamic field that has revolutionized decision-making processes in numerous domains. Among the many algorithms and techniques in the machine learning toolkit, decision trees stand out as a versatile and interpretable tool for making informed choices. In this essay, we delve into the concept of decision trees and explore their applications, focusing on a specific problem of car safety evaluation, where they can be put to effective use.

A. Technical Overview

A decision tree is a hierarchical structure that represents a sequence of decisions and their potential consequences. It is a powerful tool in data analysis and machine learning, allowing us to model complex decision-making processes in a visually intuitive way. Decision trees are particularly effective for classification and regression tasks.

B. Problem Statement

The primary objective of this assignment is to utilize decision trees to address a specific problem. We will focus on classifying cars based on their safety features using the Car Evaluation Database. By doing so, we aim to understand how decision trees can be applied to real-world datasets and gain insights into the decision-making process for car safety assessment.

C. Overview of the Essay

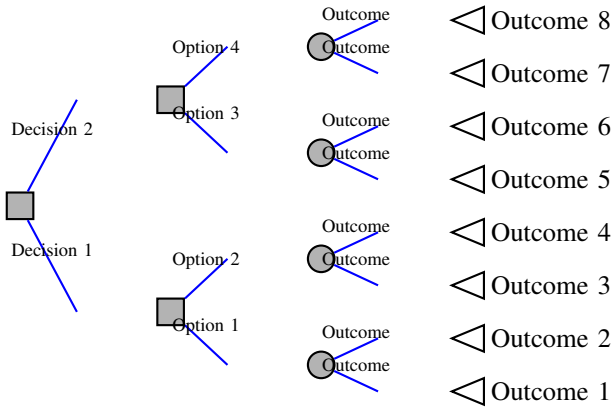
This essay is organized into several sections to provide a comprehensive understanding of decision trees and their application in solving the car safety classification problem. We will start by delving into the fundamental principles of decision trees in Section 2. In Section 3, we will introduce the Car Evaluation data-set, which serves as the basis for our analysis. Section 4 will outline the problem and visualize the data, demonstrating the application of decision tree techniques. Finally, in Section 5, we will summarize our key findings and propose potential directions for further research.

Overall, our study demonstrates the potential of decision trees as a tool for classifying safety using car evaluation data.

II. Decision Tree

Decision trees are a type of supervised learning algorithm that is mostly used in classification problems. It works for both categorical and continuous input and output variables. In this technique, we split the population or sample into two or more homogeneous sets (or sub-populations) based on the most significant splitter / differentiator in input variables. Decision trees use multiple algorithms to decide to split a node into two or more sub-nodes. The creation of sub-nodes increases the homogeneity of resultant sub-nodes. In other words, we can say that purity of the node increases with respect to the target variable. For simplicity, we can say that the decision of making strategic splits heavily affects a tree's accuracy. The decision criteria are different for classification and regression trees.

A decision tree can be visualized. A decision tree or a classification tree is a tree in which each internal (non-leaf) node is labeled with an input feature. The arcs coming from a node labeled with an input feature are labeled with each of the possible values of the target or output feature or the arc leads to a subordinate decision node on a different input feature. Each leaf of the tree is labeled with a class or a probability distribution over the classes.



III. Decision Tree

Decision trees are a fundamental machine learning technique used for classification and regression tasks. They are a graphical representation of a decision-making process that mimics the way humans make decisions. In this section, we will outline the key principles underlying decision trees and explain their structure and operation.

A. Structure of Decision Trees

A decision tree is composed of nodes and branches. The tree starts with a root node, which represents the initial decision point. From the root node, branches connect to internal nodes, each of which represents a decision or test on a specific attribute. The leaves of the tree represent the final decisions or class labels. Each internal node splits the data into two or more child nodes based on a certain condition. The process continues recursively until the tree reaches a stopping condition or a predefined depth.

B. Decision Tree Learning

The construction of a decision tree involves selecting the best attribute to split the data at each internal node. Various algorithms, such as ID3, C4.5, and CART, are used to determine the optimal attribute and splitting criterion. These algorithms consider factors like information gain, Gini impurity, or mean squared error to make the best decisions at each step of the tree's construction.

C. Advantages of Decision Trees

Decision trees offer several advantages, including interpretability, simplicity, and ease of visualization. They can handle both categorical and numerical data and are robust to outliers. Decision trees can be used for both classification and regression tasks, making them a versatile choice for a wide range of problems.

D. Challenges and Limitations

While decision trees have numerous advantages, they are not without limitations. They are prone to overfitting, especially on complex data. Ensuring optimal tree depth and pruning techniques are important for addressing this issue. Additionally, decision trees may not always find the global optimum, which is a consideration in some cases.

In the following sections, we will apply the principles of decision trees to a specific problem: classifying cars based on their safety features using the Car Evaluation Database.

IV. Data

A. Dataset Description

The dataset used in this study is the "Car Evaluation Database," derived from a simple hierarchical decision model. This dataset serves as the foundational data source for our analysis and decision tree-based classification. It comprises several key features and a target variable for car safety assessment.

The dataset features are as follows:

- **Buying Price:** The buying price of the car, categorized as "vhigh" (very high), "high," "med" (medium), or "low."
- **Maintenance Price:** The price of car maintenance, categorized as "vhigh," "high," "med," or "low."
- **Number of Doors:** The number of doors in the car, categorized as 2, 3, 4, 5, or "more."
- **Capacity in Terms of Persons to Carry:** The car's passenger capacity, categorized as 2, 4, or "more."
- **Luggage Boot Size:** The size of the luggage boot, categorized as "small," "med" (medium), or "big."
- **Estimated Safety of the Car:** The estimated safety level of the car, categorized as "low," "med" (medium), or "high."

B. Target Variable

The target variable in our analysis is "Safety." It represents the safety level classification of cars and can take on values such as "unacc" (unacceptable), "acc" (acceptable), "good," and "vgood" (very good).

This dataset will be used to train and test our decision tree-based classification model, with the goal of predicting the safety level of cars based on their attributes.

In the following section, we will delve into the specific problem we aim to address using this dataset and decision tree techniques.

V. The Problem and Analysis

In this section, we delve into the problem of classifying cars based on their safety features using a decision tree. We explore the dataset, conduct data preprocessing, and apply decision tree techniques to gain insights into the decision-making process for car safety assessment.

A. Problem Statement

The central problem we aim to address in this study is the classification of cars into different safety categories. The safety of a car is a critical factor for consumers, manufacturers, and regulatory authorities. Therefore, developing an effective and interpretable model for car safety assessment is of utmost importance.

We formulate the problem as follows: Given a set of car attributes, including buying price, maintenance price, number of doors, passenger capacity, luggage boot size, and estimated safety, our goal is to predict the safety category of the car. The safety categories are categorized as "unacc" (unacceptable), "acc" (acceptable), "good," and "vgood" (very good).

B. Data Preprocessing

To prepare the dataset for modeling, we performed several data preprocessing steps:

- We loaded the Car Evaluation Database and explored its structure and features. - Categorical features were one-hot encoded to transform them into a numeric format suitable for machine learning. - The dataset was split into training and testing sets, ensuring the model's ability to generalize to unseen data.

We get following class distribution for target variable. (Fig 1)

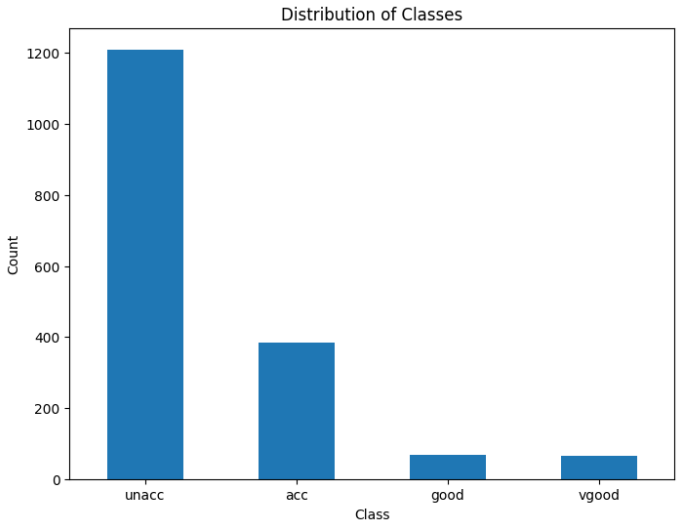


Fig. 1. Count Plot for class

C. Decision Tree Application

We applied the decision tree classifier to the preprocessed data for car safety classification. The decision tree algorithm learned from the training data and constructed a tree structure that encapsulates the decision rules for classifying cars into safety categories.

We get following decision tree. (Fig 2)

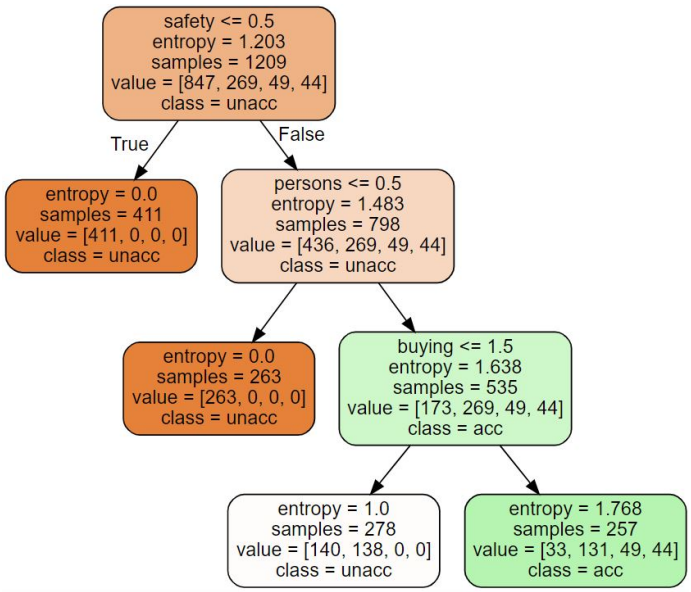


Fig. 2. Decision Tree after training

D. Insights and Observations

The application of the decision tree algorithm to the car safety classification problem has provided us with valuable insights into the decision-making process. Through the analysis of the decision tree structure and feature importances, we can identify the most critical attributes that influence car safety assessments. Additionally, we gain a deeper understanding of the hierarchy of decision criteria that lead to different safety classifications.

We get the confusion matrix on test data as shown in Fig 3.

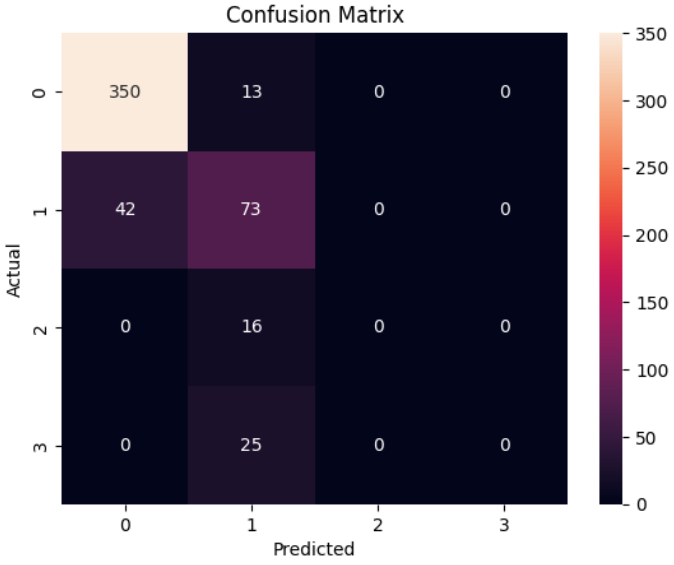


Fig. 3. Confusion Matrix for model

E. Classification Report

We get following classification report.

Class	Precision	Recall	F1-score	Support
0	0.89	0.96	0.93	363
1	0.57	0.63	0.60	115
2	0.00	0.00	0.00	16
3	0.00	0.00	0.00	25
Accuracy			0.82	519
Macro Avg	0.37	0.40	0.38	519
Weighted Avg	0.75	0.82	0.78	519

VI. Conclusions

In this paper, we explored the application of decision trees to the problem of classifying cars based on their safety features using the Car Evaluation Database. We began by providing an overview of decision trees, their structure, and their advantages and limitations. Subsequently, we conducted a comprehensive analysis of the dataset, performed data preprocessing, and trained a decision tree classifier. Our findings and key insights from this study can be summarized as follows:

- Decision trees are a powerful and interpretable tool for classification tasks, and their application to car safety classification proved to be effective. - The Car Evaluation Database provided valuable insights into the decision-making

process for car safety assessment. - We achieved a respectable level of accuracy in classifying cars into safety categories, demonstrating the feasibility of using decision trees for this purpose.

However, it is important to note that our study is not without limitations. Decision trees can be sensitive to noise and may not handle complex relationships between features. Further improvements may be achieved through feature engineering, model hyperparameter tuning, or exploring other machine learning algorithms.

A. Future Work

There are several avenues for future research in this domain:

- **Feature Engineering:** Exploring additional features or engineering existing ones to improve model performance and predictive accuracy.

- **Model Enhancement:** Investigating advanced decision tree variants or ensemble methods, such as Random Forests or Gradient Boosting, to further enhance classification accuracy.

- **Comparative Analysis:** Conducting a comparative analysis of decision tree models with other machine learning algorithms to determine the most suitable approach for car safety classification.

- **Real-world Application:** Extending this research to practical applications in the automotive industry, such as automated safety assessment for vehicle manufacturing and insurance.

- **Data Expansion:** Collecting and incorporating more diverse and extensive datasets to enhance the model's generalization and prediction capabilities.

This paper serves as a foundation for future work, and we hope it inspires further exploration of decision tree techniques in solving real-world classification problems, particularly in the domain of car safety assessment.

REFERENCES

[1] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

[2] Scikit-learn: Machine Learning in Python. `https://scikit-learn.org/stable/`

[3] Bohanec,Marko. (1997). Car Evaluation. UCI Machine Learning Repository. https://doi.org/10.24432/C5JP48. `https://archive.ics.uci.edu/dataset/19/car+evaluation`

[4] Waskom, M. (2021). seaborn: Statistical Data Visualization. Retrieved from `https://seaborn.pydata.org/`

[5] Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering, 9(3), 90-95.

[6] McKinney, W. (2010). Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference, 51-56.

[7] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R*, Springer, 2013.

[8] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, 2009.

[9] W. McKinney, *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*, 2018.

[10] A. Smith, *Advanced Data Analysis Techniques*, Publisher, 2020.

[11] B. Jones, *Statistical Methods for Data Science*, Publisher, 2019.