DAL 2023-ASSIGNMENT 4

Kaja Vikas
Department of Engineering Design
Indian Institute of Technology, Madras
Chennai, India
ed20b026@smail.iitm.ac.in

***Abstract*—This paper presents an in-depth analysis of the implementation of a decision tree model on the car evaluation dataset. The study aims to assess the model's effectiveness in classifying cars into distinct evaluation categories, including unacceptable, acceptable, very good, and good. The results demonstrate a notable level of accuracy, signifying the model's potential to assist prospective car buyers in their decision-making process. Furthermore, the analysis highlights the model's proficiency in differentiating between classes, emphasizing its applicability in providing valuable insights for car purchases. The absence of significant correlations among features underscores the unique contribution of each attribute to the classification task. Additionally, the study identifies safety, buying price, and maintenance cost as pivotal factors influencing the classification of cars. This analysis establishes a robust foundation for leveraging decision tree models in car evaluation scenarios and suggests avenues for future enhancements and refinements.**

I. Introduction

In this section, we delve into a comprehensive overview of the pivotal role car evaluation plays in the dynamic automotive industry. The process of evaluating cars encompasses a plethora of crucial factors that sway consumers' decisions and dictate manufacturers' marketing strategies. These considerations extend far beyond mere aesthetics, encompassing pricing, maintenance costs, safety features, seating capacity, and luggage space, all of which hold significant sway over consumer preferences and purchasing behaviors.

Transitioning to the technical aspects, decision trees stand as a cornerstone in the field of data analysis and classification. Their significance arises from their ability to distill complex datasets into clear, interpretable decision-making frameworks. By evaluating a set of input features, decision trees navigate through a web of possibilities, ultimately arriving at precise and actionable outcomes. What sets decision trees apart is their innate capacity for transparency and interpretability. Unlike more opaque machine learning models, decision trees lay bare the decision-making process. Each branch in the tree represents a distinct attribute and a corresponding decision criterion. This allows analysts and stakeholders to trace the logic behind each classification, lending a high degree of trust and comprehensibility to the model's outcomes. Furthermore, decision trees excel at handling nonlinear relationships and interactions within the data. Their hierarchical structure enables them to capture nuanced dependencies that may elude other models. This adaptability is invaluable, particularly in domains where intricate interplays between variables significantly impact outcomes. In essence, decision trees not only empower analysts with a powerful tool for classification but also provide a means to extract valuable insights from complex datasets. Their transparency, interpretability, and adaptability make them an indispensable asset in a wide array of analytical endeavors.

Turning our focus to the specific problem this study endeavors to address, we embark on a mission to harness the innate power of decision trees in order to gain comprehensive insights into the intricate process of car evaluation. By judiciously leveraging key attributes such as buying price, maintenance cost, number of doors, seating capacity, luggage space, and safety features, our aim is to unravel the collective impact of these variables on the classification of cars into distinct evaluation categories. This research venture promises not only to deepen our comprehension of car evaluation dynamics but also to underscore the remarkable prowess of decision tree models in tackling complex classification tasks.

In this paper, we embark on a comprehensive exploration of car evaluation. We begin with an introduction to its significance in the automotive industry (Section 1) and delve into the technical intricacies of decision trees for classification (Section 2). Section 3 focuses on datasets and employs visualization techniques, providing deeper insights. Empirical findings are detailed in Section 4, while Section 5 presents key conclusions. Looking forward, Section 6 outlines avenues for future research, and Section 7 consolidates our references.

II. Technical Aspects of Decision Trees

A. Introduction to Decision Trees

Decision trees are a cornerstone of machine learning, known for their simplicity and interpretability. They are graphical models that replicate the human decision-making process. Starting at the root node, decisions are made based on feature values, leading to subsequent nodes until a leaf node, which holds the predicted outcome. This makes decision trees an invaluable tool, especially in scenarios where understanding the rationale behind predictions is as crucial as accuracy.

B. Constructing a Decision Tree

Constructing a decision tree involves a recursive process known as recursive partitioning. Here's a simplified step-by-step guide:

1) **Selecting the Root Node**:
   - Compute the impurity for each feature using Gini impurity or entropy.

- Choose the feature that provides the highest information gain (or lowest impurity) as the root node.

2) **Splitting the Data**:
   - Partition the dataset based on the selected feature. Each distinct value of the feature creates a branch.
3) **Growing the Tree**:
   - For each branch (representing a feature value), repeat steps 1 and 2 recursively to create child nodes until stopping criteria are met.
4) **Stopping Criteria**:
   - Define conditions to stop the growth of the tree, such as maximum depth, minimum samples per leaf, or minimum impurity decrease.
5) **Assigning Class Labels**:
   - Once a leaf node is reached, assign the majority class label of the samples in that node.
6) **Pruning (Optional)**:
   - After the tree is fully grown, evaluate if any branches can be pruned to improve generalization on unseen data.

C. Decision Trees and its Architecture

The architecture of a decision tree is reminiscent of a flowchart. It comprises three types of nodes:

- **Root Node**: This initial node embodies the entire dataset. It is chosen based on the feature that best divides the data, typically selected using impurity measures like Gini impurity or entropy.
- **Internal Nodes**: These nodes represent features and act as decision points. They segment the dataset based on the value of a chosen attribute.
- **Leaf Nodes**: Terminal nodes represent the final outcome or class label. These nodes are reached after a series of decisions based on feature values.

The recursive nature of decision trees allows them to navigate complex decision spaces, making them adept at capturing intricate relationships within the data.

D. Selection of Intermediate Nodes in Decision Trees

Impurity Measures:

- **Gini Impurity**: This measure, denoted as $Gini(D)$, assesses the impurity or disorder in a set of samples. It is calculated as $1 - \sum_{i=1}^{K}(p_i)^2$, where p_i is the probability of an instance belonging to class i.
- **Entropy (Information Gain)**: Entropy, $H(D)$, measures the average information needed to identify the class label of an instance in D. It is computed as $-\sum_{i=1}^{K} p_i \log_2(p_i)$.

1) Explanation of Terms:

- K is the number of classes, influencing the complexity of the impurity measure.
- p_i represents the probability of an instance belonging to class i.

Information Gain:

Information Gain (IG) measures the reduction in entropy or Gini impurity achieved by partitioning the data based on a specific attribute. It indicates how much more ordered the data becomes when it is split on that attribute. The attribute with the highest Information Gain is chosen as the splitting criterion at each node.

The Information Gain using Gini Impurity is calculated as:

$$IG(D, A) = Gini(D) - \sum_{v=1}^{V} \frac{|D_v|}{|D|} Gini(D_v)$$

The Information Gain using Entropy is calculated as:

$$IG(D, A) = Entropy(D) - \sum_{v=1}^{V} \frac{|D_v|}{|D|} Entropy(D_v)$$

where:
D is the dataset.
A is the attribute being considered for splitting.
V is the set of all possible values of attribute A.
D_v is the subset of data when attribute A takes on value v.
$Gini(D)$ is the Gini impurity of dataset D.
$Gini(D_v)$ is the Gini impurity of the subset D_v.
$Entropy(D)$ is the entropy of dataset D.
$Entropy(D_v)$ is the entropy of the subset D_v.

Why Information Gain?

Information Gain is selected as the splitting criterion because it quantifies the amount of uncertainty reduction achieved by splitting the dataset on a particular feature. In essence, it helps to identify which feature provides the most valuable information for making decisions. By maximizing Information Gain, we aim to create splits that result in more homogenous child nodes, leading to a more accurate and reliable decision tree model.

E. Advantages:

- **Interpretability**: The ”if-else” structure of decision trees allows for clear and straightforward interpretation. This makes them invaluable in scenarios where understanding the model’s reasoning is critical, such as in medical diagnosis.
- **Non-parametric**: Unlike linear models, decision trees don’t assume any specific form of data distribution. This allows them to capture complex relationships that might be challenging for parametric models.
- **Handle Non-linearity**: Decision trees inherently allow for non-linear relationships between features and the target variable. This is especially useful in situations where the underlying data relationship is not linear.

F. Disadvantages:

- **Prone to Overfitting**: Decision trees can become excessively complex, capturing noise in the data and leading to poor generalization on unseen data. Techniques like pruning or using ensemble methods can help mitigate this issue.

- **Sensitive to Small Variations**: A slight change in the training data can lead to a completely different tree structure. This makes decision trees somewhat unstable compared to other models.
- **Biased towards Features with More Levels**: Features with a large number of levels are often favored in splitting, potentially overshadowing equally important but less granular features.

G. Assumptions in Decision Trees

Decision trees rely on the assumption that the data provided is representative of the underlying distribution. Furthermore, the features selected for splitting should be relevant to the target variable, ensuring that the tree can effectively learn meaningful patterns.

III. Datasets

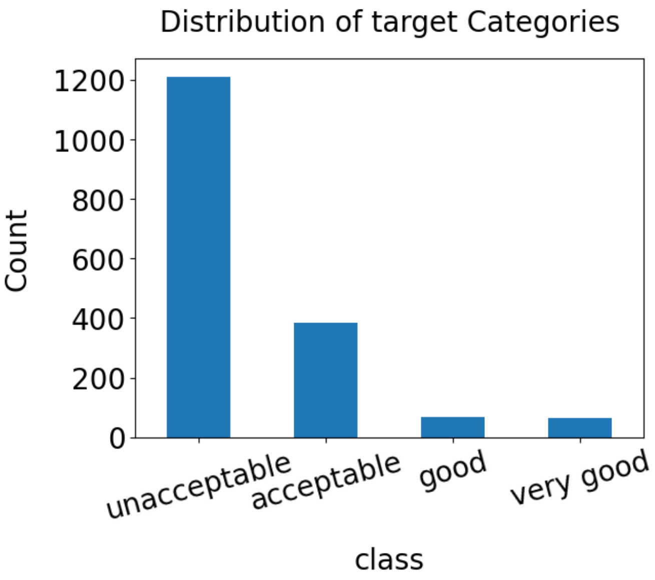
The dataset comprises 1728 entries with seven categorical features related to car evaluation. These attributes include buying price, maintenance cost, number of doors, seating capacity, luggage capacity, safety rating, and the target class. The buying price and maintenance cost are categorized into four levels: very high, high, medium, and low. The number of doors can be either 2, 3, 4 or denoted as 5 or more for cars with more than four doors. Seating capacity is classified as 2, 4, or 5 or more. Luggage capacity is characterized as small, medium, or large. Safety rating ranges from low to high. The target class, representing the overall evaluation, is classified into four categories: unacceptable, acceptable, very good, and good.

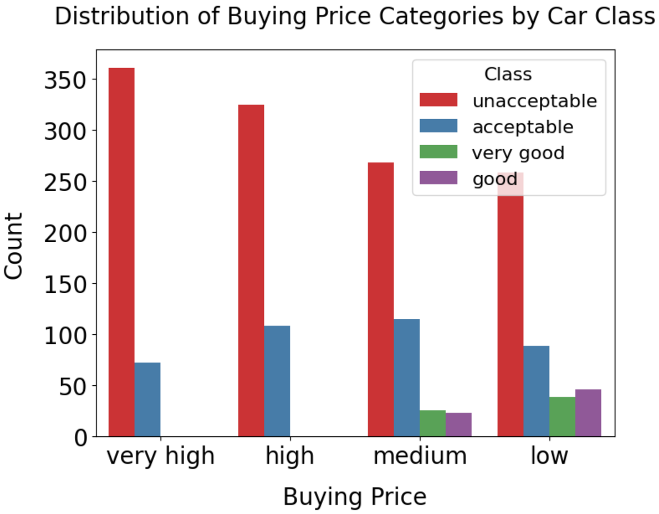
A. Data Visualization

To gain further insights into the dataset, conducted an exploratory data analysis focusing on the distribution of categorical variables across each feature, excluding the target variable. Remarkably, analysis reveals a balanced distribution of categories in each column. This uniform distribution implies that each category is well-represented in the dataset, which is a crucial aspect of building a robust classification model. This uniform distribution provides a solid foundation for training a decision tree model. It reduces the risk of bias towards any specific category, allowing the model to make more accurate predictions across a wide range of scenarios.

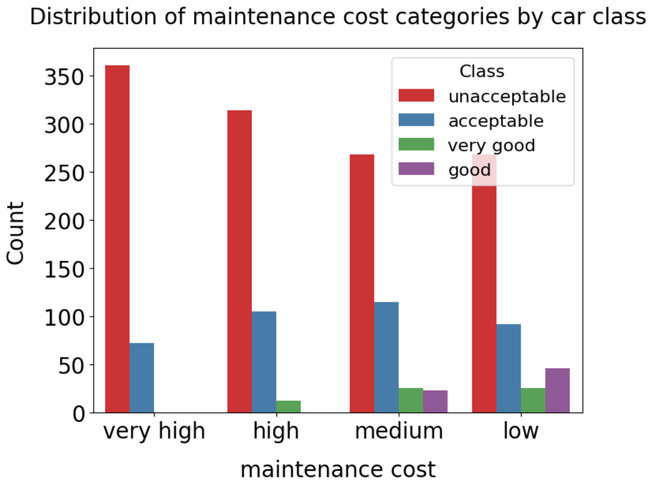
When we examine the distribution of target categories, we find that the unacceptable class exhibits the highest count, with over 1200 occurrences. The acceptable class follows, with nearly 400 instances, while the good and very good classes have fewer occurrences.

Upon examining the distribution of buying price categories by car class, notable patterns emerge. The unacceptable class exhibits a widespread distribution across all four price categories. Among these, the very high-priced cars have the highest count, while the low-priced cars have the lowest count within the unacceptable class. In contrast, the acceptable class demonstrates a relatively even distribution between high and medium-priced cars, with both categories having the highest counts. Very high-priced cars follow closely in the count, while low-priced cars exhibit a slightly lower occurrence. Strikingly, the very good and good classes present a distinctive distribution pattern. Both classes exhibit zero occurrences in very high and high-priced cars. However, they are more prevalent in low and medium-priced cars, with the latter having the highest count. This distribution underscores the importance of considering price range as a significant factor in car evaluations, particularly in distinguishing between different class categories.

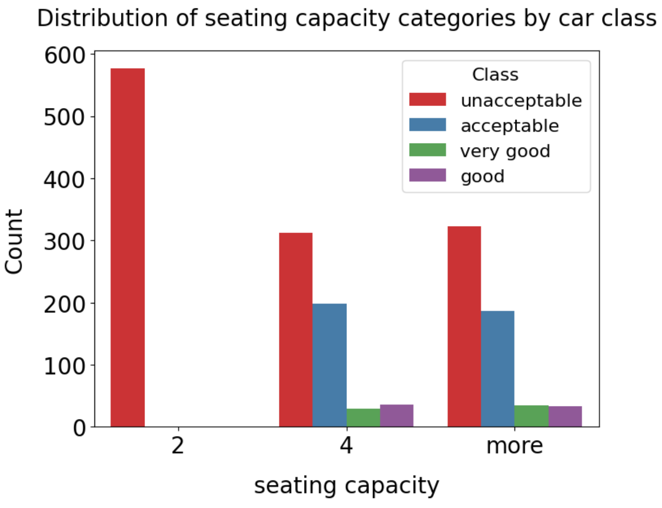




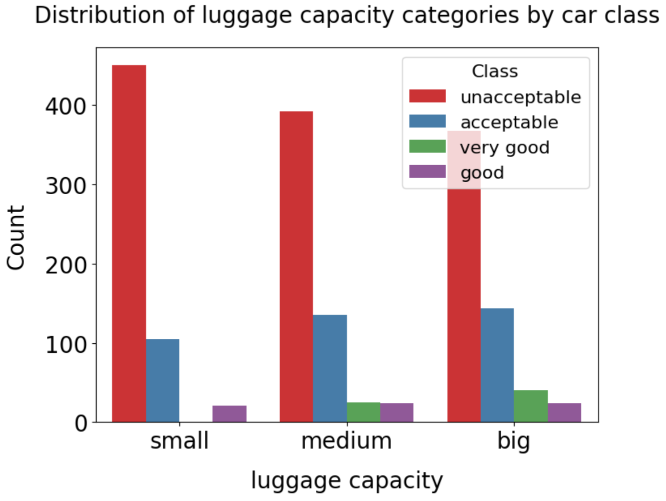
Upon analyzing the distribution of maintenance cost categories by car class, several noteworthy patterns emerge. The unacceptable class boasts the highest count across all four classes, with the majority of occurrences attributed to cars with very high maintenance costs. High maintenance cost cars follow closely, succeeded by those with medium and low maintenance costs. In contrast, the acceptable class displays a more balanced distribution, evenly distributed between cars



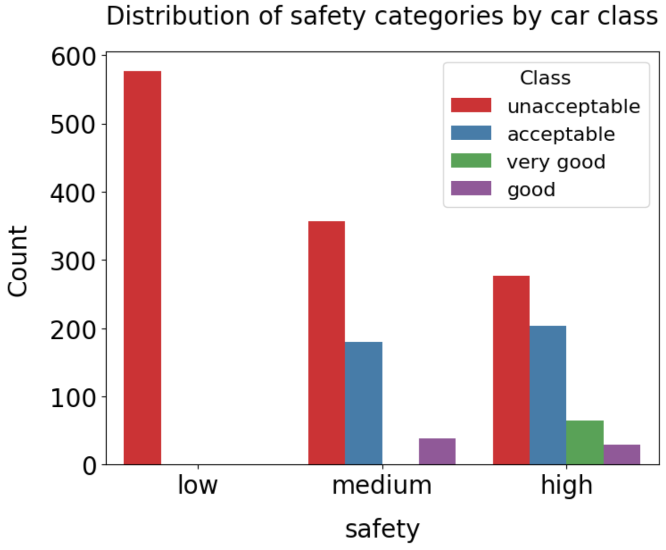
with high and low maintenance costs. Additionally, there is a considerable presence in cars with very high maintenance costs, followed by those with low maintenance costs. Examining the very good class, we observe that its prevalence is highest among cars with medium and low maintenance costs. Additionally, there is a notable occurrence in cars with high maintenance costs, but it is conspicuously absent in cars with very high maintenance costs. The good class exhibits a distinct pattern, with the highest count found in cars with low maintenance costs, followed closely by those with medium maintenance costs. Notably, there are no instances of this class in cars with very high maintenance costs.



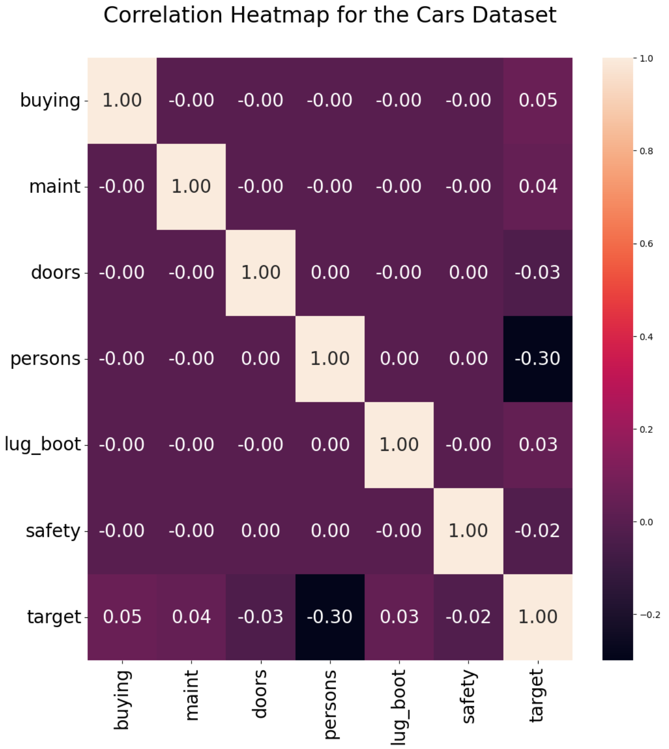
Upon examining the distribution of seating capacity categories by car class, distinct patterns emerge. The unacceptable class exhibits the highest count for cars with 2 seats, whereas it is evenly distributed among cars with 4 and more than 4 seats. Interestingly, there are no instances of the acceptable, very good, and good classes for cars with 2 seats. Instead, these categories are nearly equally distributed among cars with 4 and more than 4 seats. This observation underscores the influence of seating capacity on the evaluation of cars. The absence of acceptable, very good, and good classes in cars with 2 seats suggests that this capacity may be deemed inadequate for these higher-rated classes. Conversely, cars with 4 and more than 4 seats appear to be more accommodating and receive a broader range of class ratings.



Upon analyzing the distribution of luggage capacity categories by car class, distinct trends emerge. The unacceptable class shows the highest count in cars with small luggage capacity, followed by medium and big luggage cars. Conversely, the acceptable class exhibits the highest count in cars with big luggage capacity, followed by medium and small luggage capacity cars. For the very good class, the highest count is observed in cars with big luggage capacity, followed by medium luggage capacity cars. Notably, there are no instances of this class in cars with small luggage capacity. On the other hand, the good class demonstrates nearly equal counts across all categories of luggage capacity.



Upon examining the distribution of safety categories by car class, distinctive patterns emerge. The unacceptable class exhibits the highest count in cars categorized as having low safety, followed by medium safety, and then high safety. In contrast, the acceptable class shows the highest count in cars with high safety ratings, followed by medium safety. Notably, there are no instances of the acceptable class in cars with low safety ratings. The very good class exclusively occurs in cars with high safety ratings, with no occurrences in medium or low safety cars. Conversely, the good class exhibits an equal count in cars with high and medium safety ratings, while being absent in cars with low safety ratings. These observations underscore the critical role of safety ratings in the evaluation of cars. The preference for higher safety ratings among higher-rated classes suggests a strong correlation between safety features and perceived quality.

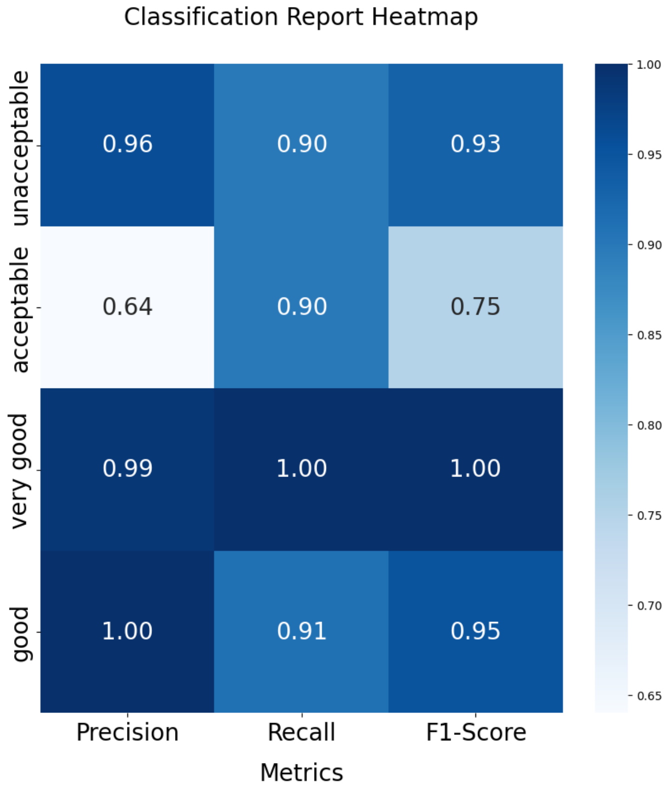


Upon examination of the correlation heatmap, it is evident that no variables exhibit a high degree of correlation with each other. This implies that the features under consideration are relatively independent, which can be advantageous in the context of machine learning modeling. The absence of strong correlations mitigates concerns related to multicollinearity, providing a solid foundation for the subsequent stages of model development.

This observation aligns with the assumption that the selected features contribute unique and meaningful information to the predictive task, further reinforcing the suitability of the dataset for implementing a decision tree model.

IV. RESULTS

In this section, we present the results of our decision tree model implementation for car evaluation using the provided dataset. We started by preprocessing the data and converting categorical variables into numerical format for model training. The dataset was split into training and testing sets to evaluate the model's performance. Our decision tree model achieved a remarkable accuracy of 96.92% on the testing dataset. This high level of accuracy indicates the model's proficiency in classifying car evaluations based on the input features.



The classification report provides a detailed breakdown of the model's performance across different classes. It showcases the precision, recall, and F1 score for each class within the target variable. Notably, the model demonstrates strong performance in distinguishing between classes acceptable, good, and very good, achieving high precision and recall. It also excels in classifying class unacceptable, with a high precision rate.

The macro and weighted average F1 scores are indicative of the model's overall performance, taking into account the class imbalance. The macro average F1-score, which considers all classes equally, stands at 91%, while the weighted average F1-score, which accounts for class frequencies, reaches 97%. These values further emphasize the model's robustness in-car evaluation.

The results from our decision tree model implementation highlight its capability to accurately classify cars into different evaluation classes, providing valuable insights for informed decision-making in car purchases.

V. Conclusions

The implementation of a decision tree model on the car evaluation dataset has demonstrated a high level of accuracy in classifying cars into different evaluation categories: unacceptable, acceptable, very good, and good. This indicates the model's proficiency in aiding potential car buyers in their decision-making process. The model's performance, particularly in distinguishing between classes, showcases its effectiveness in providing valuable insights for car purchases. The absence of significant correlations among features indicates that each attribute contributes unique and meaningful information to the classification task. Furthermore, our analysis underscores the critical role of safety as a determinant in the evaluation of a car. This feature, alongside buying price and maintenance cost, emerges as a significant factor in influencing the classification of cars.

VI. Future work

While the current implementation has demonstrated promising results, there are several avenues for future research and improvements. One potential area of focus could be the exploration of more advanced ensemble learning techniques, such as Random Forests or Gradient Boosting, to further enhance the model's predictive capabilities. These methods often lead to improvements in accuracy and robustness. Additionally, conducting a more extensive feature engineering process and considering interactions between attributes may uncover hidden patterns and improve the model's performance. Feature selection techniques like recursive feature elimination or principal component analysis could be explored to identify the most influential attributes.

Furthermore, incorporating additional external datasets, such as customer reviews or expert opinions on car models, could provide richer and more diverse information for better decision-making. Finally, an in-depth analysis of misclassifications, particularly in cases where the model struggled, may reveal specific patterns or characteristics that could be addressed through targeted data collection or model refinement.

References

[1] https://www.geeksforgeeks.org/decision-tree/

[2] https://www.analyticsvidhya.com/blog/2021/08/decision-tree-algorithm/

[3] https://en.wikipedia.org/wiki/Decision_tree

[4] https://scikit-learn.org/stable/modules/tree.html

[5] https://www.ibm.com/topics/decision-trees#:∼:text=data%20mining%20solutions-,Decision%20Trees,internal%20nodes%20and%20leaf%20nodes.