# DAL 2023-ASSIGNMENT 5

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 Random Forest 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 Random Forest 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, Random Forests represent a pivotal advancement in data analysis and classification. Their significance lies in their ability to aggregate multiple decision trees, harnessing their collective wisdom to arrive at accurate and reliable predictions. This ensemble approach imparts Random Forests with a robustness and generalization capacity that sets them apart in the realm of machine learning.Unlike individual decision trees, Random Forests combine the strengths of numerous models to form a powerful predictive tool. By leveraging the wisdom of the crowd, they achieve higher accuracy and are less susceptible to overfitting. This amalgamation of diverse perspectives enhances the model's trustworthiness and performance.

Moreover, Random Forests maintain an element of interpretability, albeit to a lesser extent compared to individual decision trees. While the collective decision-making process may not be as transparent, the feature importance scores can still provide insights into which attributes are influential in making predictions.Additionally, Random Forests inherit the adaptability of decision trees in handling complex, nonlinear relationships and interactions within the data. The ensemble's collective intelligence allows it to capture subtle dependencies that might be elusive to singular models. This versatility proves invaluable in domains where intricate interplays between variables wield substantial influence over outcomes.Random Forests not only serve as a potent tool for classification but also provide a means to distill valuable insights from intricate datasets. Their ensemble nature, combined with adaptability and a degree of interpretability, renders them an indispensable asset in a diverse range of analytical endeavors.

Turning our focus to the specific problem this study endeavors to address, we embark on a mission to harness the collective power of Random Forests 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, we aim 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 Random Forest 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 Random fores 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 Random Forest

### A. Introduction

Random Forest (RF) is an ensemble learning method that combines multiple decision trees to improve predictive accuracy and generalization. It has gained popularity due to its robustness, versatility, and ability to handle a wide range of data types. This section provides an overview of the RF algorithm and its key components.

### B. Decision Trees

RF builds upon the foundation of decision trees, which are hierarchical structures used for classification and regression tasks. Each tree in the forest is constructed through a recursive

process that partitions the feature space based on attribute values, ultimately leading to leaf nodes containing class labels (in classification tasks) or numerical values (in regression tasks).

### C. Ensemble Learning

RF is classified as an ensemble learning algorithm, which means it leverages the aggregation of multiple models to improve overall predictive performance. This section discusses the ensemble learning paradigm, highlighting concepts like bagging (Bootstrap Aggregating) and the role of diversity among base learners.

### D. Randomness and Bootstrap Aggregatin

Random Forest introduces randomness in two key ways:

**Bootstrap Sampling** Given a training set $D$ of size $N$, $D$ is sampled with replacement to create a bootstrap sample $D_i$ of size $N$, where $i$ ranges from 1 to $B$, the number of trees in the forest. This can be represented as:

$$D_i \sim D, \quad i = 1, 2, \ldots, B$$

**Random Subset of Features:** At each node split during tree construction, only a random subset of features $m$ is considered. This subset is chosen without replacement from the full set of $M$ features, where $m << M$. The selection of $m$ can be controlled by the user or set as a hyperparameter.

### E. Splitting Criteria:

The splitting criterion at node $t$ is determined by a measure of impurity, which can be the Gini impurity ($Gini(D_t)$) or the Information Gain ($IG(D_t)$). For a node $t$ with $N_t$ samples, partitioned into $K$ classes, the impurity measures are defined as follows:

**Gini Impurity:**

$$Gn(D_t) = 1 - \sum_{k=1}^{K} (p_{t,k})^2$$

where $p_{t,k}$ is the proportion of class $k$ samples in node $t$.

**Information Gain:**

$$IG(D_t) = H(D_t) - \sum_{i=1}^{m} \frac{N_{t,i}}{N_t} H(D_{t,i})$$

where $H(D_t)$ is the entropy of node $t$, $m$ is the number of child nodes after the split, $N_{t,i}$ is the number of samples in child node $i$, and $D_{t,i}$ is the subset of data in child node $i$.

### F. Out-of-Bag Error Estimation

In Random Forest, during the process of building each tree, some data points are left out due to the bootstrap sampling technique. These data points are referred to as Out-of-Bag (OOB) samples. These samples were not used for training the tree, they provide a natural and unbiased assessment of how well the tree generalizes to unseen data. For each tree $i$, the out-of-bag (OOB) error is calculated using the samples not included in the bootstrap sample for that tree. The OOB error rate ($Err_i$) is defined as:

$$Err_i = \frac{1}{|D \backslash D_i|} \sum_{(x_j, y_j) \in (D \backslash D_i)} \mathbb{I}(h_i(x_j) \neq y_j)$$

where $h_i$ is the prediction made by tree $i$, $x_j$ is a sample, and $y_j$ is its true label.The OOB error estimation allows you to assess the model's performance without the need for a separate validation set. This is especially useful when you have limited data and want to make the most out of it.

### G. Hyperparameter Tuning

Hyperparameters are configuration settings that dictate the behavior of the Random Forest algorithm.

1) **Number of Trees (B):** This determines how many trees are in the forest. A larger number of trees can lead to a more robust model but may increase computational cost.
2) **Maximum Depth (max depth):** This limits the depth of each individual tree. It controls the complexity of the trees and helps prevent overfitting.
3) **Minimum Samples per Leaf (min samples per leaf):** This sets the minimum number of samples required to be at a leaf node. It can help prevent the model from learning noise.
4) **Number of Features Considered at Each Split (m):** This controls the number of features randomly selected at each split. It can add diversity to the trees.
5) **Class Weights (for imbalanced data):** Assigning different weights to classes during tree construction can help balance the influence of each class.

The process of hyperparameter tuning involves:

1) **Grid Search or Random Search:**
   - Define a grid of hyperparameters or randomly sample hyperparameters.
   - Train multiple Random Forest models with different combinations of hyperparameters.
2) **Cross-Validation:** Evaluate the models using cross-validation to get an unbiased estimate of performance.
3) **Select Optimal Hyperparameters:** Choose the hyperparameters that yield the best performance on the validation set.

Hyperparameter tuning helps optimize the model for the specific dataset and problem at hand, improving its predictive capabilities.

Remember, the choice of hyperparameters can significantly impact the model's effectiveness, so it's crucial to carefully select and fine-tune them based on the characteristics of the data.

### H. Feature Importance

Feature importance in Random Forest can be assessed using measures like Gini Importance ($GI$), Permutation Importance ($PI$), or Mean Decrease in Impurity ($MDI$).

**Gini Importance** The Gini importance of feature $X_j$ is calculated as the sum of the Gini impurity reductions over all nodes where $X_j$ is used for splitting:

$$GI(X_j) = \sum_t \left(p_t \cdot Gn(D_t) - p_t \cdot Gn(D_t^L) - p_t \cdot Gn(D_t^R)\right)$$

where:

$p_t$ is the proportion of samples in node $t$,
$D_t$ is the set of samples in node $t$,
$D_t^L$ is the set of samples in the left child node of $t$,
$D_t^R$ is the set of samples in the right child node of $t$.

**Permutation Importance:** The permutation importance of feature $X_j$ is computed by shuffling the values of $X_j$ in the validation set and measuring the change in prediction accuracy:

$$PI(X_j) = \frac{1}{|D_{\text{val}}|} \sum_{(x_j, y_j) \in D_{\text{val}}} \mathbb{I}(h(x_j) \neq y_j)$$

where $D_{\text{val}}$ is the validation set, $h$ is the RF model, $x_j$ is the feature $X_j$, and $y_j$ is its true label.

### I. Handling Imbalanced Data

Various strategies can be applied to address class imbalance in Random Forest, including:

**Class Weighting:** Assigning different weights to classes during tree construction to give more importance to minority classes.

**Resampling Techniques:** Applying techniques like over-sampling (duplicating samples of the minority class), under-sampling (removing samples of the majority class), or using more advanced methods like SMOTE (Synthetic Minority Over-sampling Technique) to balance the class distribution.

**Advantages of Random Forest:**

- High predictive accuracy.
- Robust to overfitting.
- Handles non-linear relationships.
- Provides feature importance.
- Resilient to outliers.
- Can handle missing values.
- Parallelizable and scalable.

**Disadvantages of Random Forest:**

- Complexity and interpretability.
- Resource-intensive.
- Potential overfitting with default hyperparameters.
- Less effective with small datasets.

**Assumptions of Random Forest:**

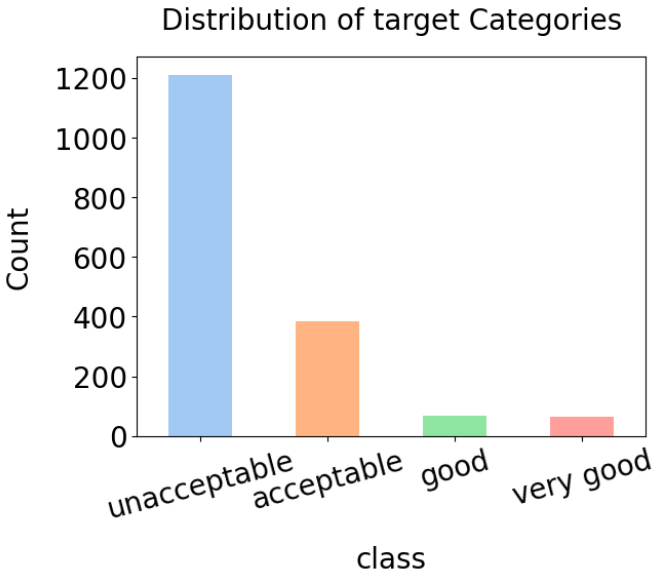
- Independent trees.
- Homogeneous feature importance.
- Input data of sufficient quality.
- Randomness in feature selection.

## 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.

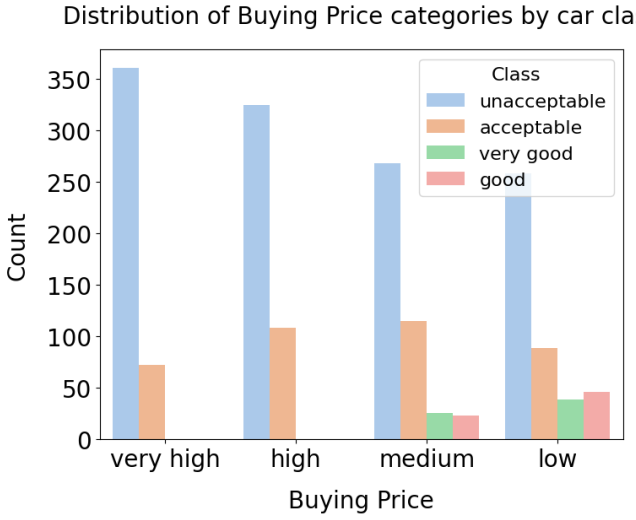
### A. Data Visualization

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.

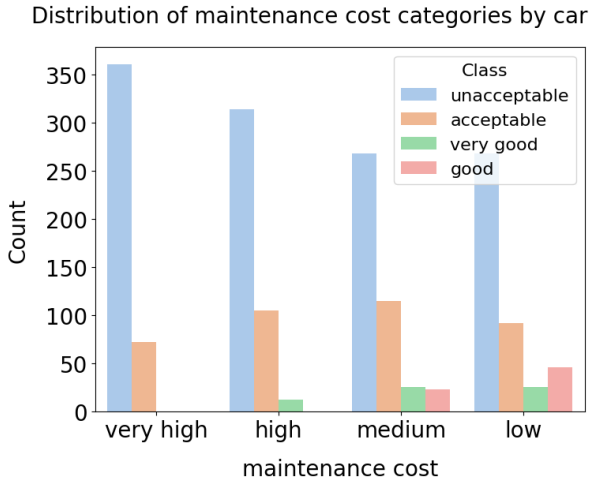



When we examine the distribution of target categories, we find that the unacceptable class exhibits the highest count, with over 1200 occurrences. The acceptable class follows, with nearly 400 instances, while the good and very good classes have fewer occurrences.

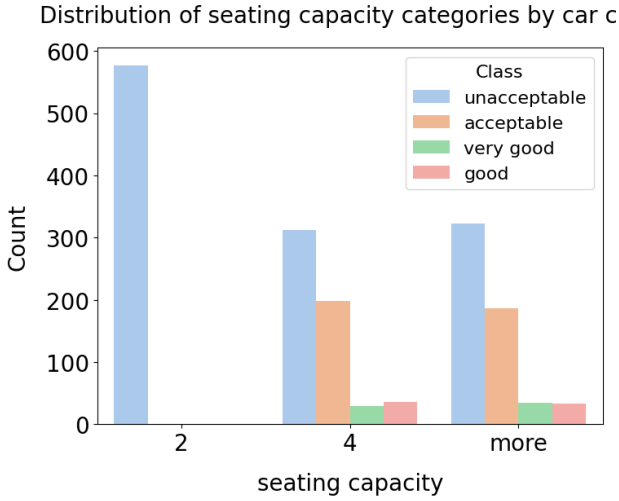
Upon examining the distribution of buying price categories by car class, notable patterns emerge. The unacceptable class exhibits a widespread distribution across all four price categories. Among these, the very high-priced cars have the




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.

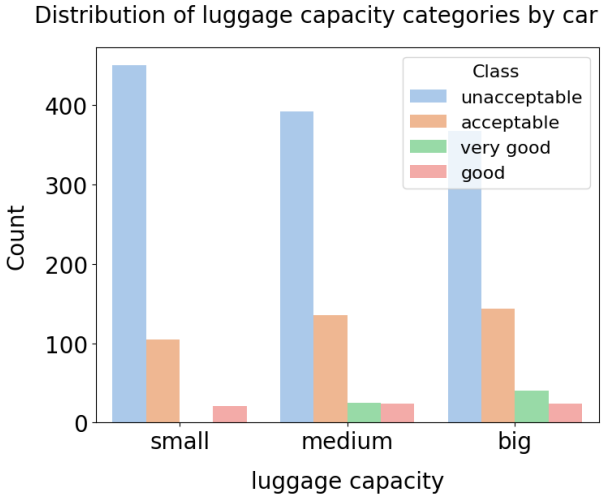



Upon analyzing the distribution of maintenance cost categories by car class, several noteworthy patterns emerge. The unacceptable class boasts the highest count across all four classes, with the majority of occurrences attributed to cars with very high maintenance costs. High maintenance cost cars follow 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.

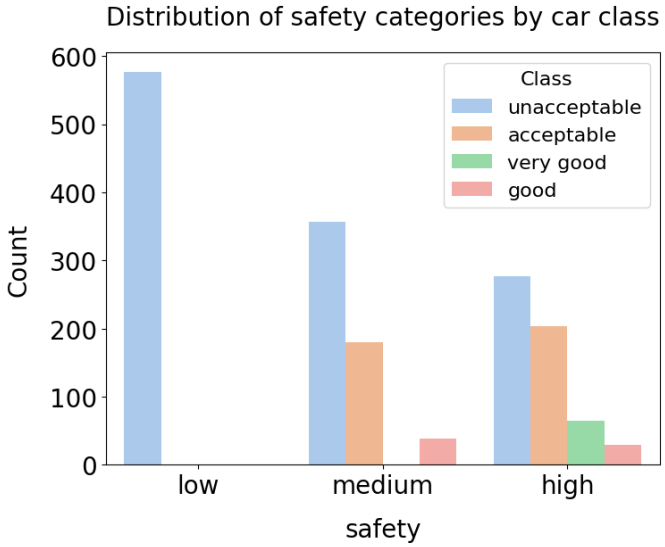



Upon examining the distribution of seating capacity categories by car class, distinct patterns emerge. The unacceptable class exhibits the highest count for cars with 2 seats, whereas it is evenly distributed among cars with 4 and more than 4 seats. Interestingly, there are no instances of the acceptable, 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.

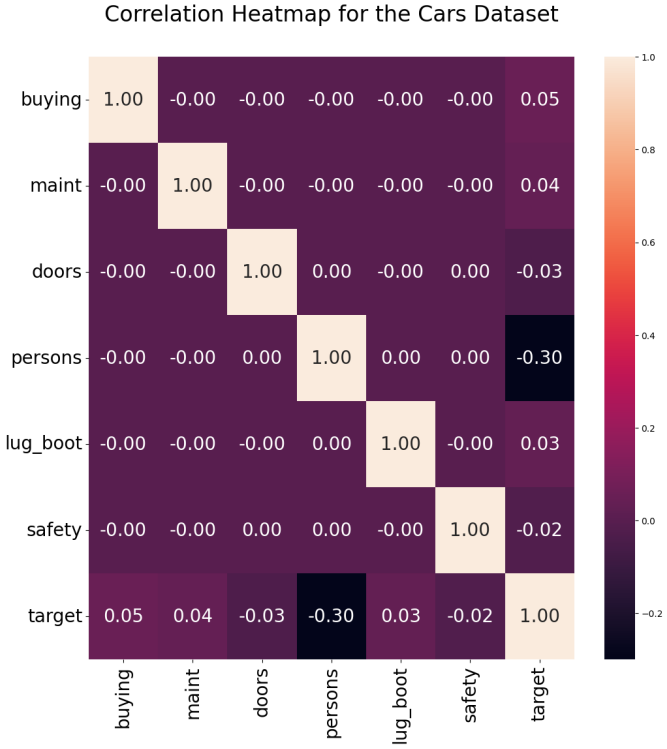
Upon analyzing the distribution of luggage capacity categories by car class, distinct trends emerge. The unacceptable class shows the highest count in cars with small luggage capacity, followed by medium and big luggage cars. Conversely, the acceptable class exhibits the highest count in cars with big 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.



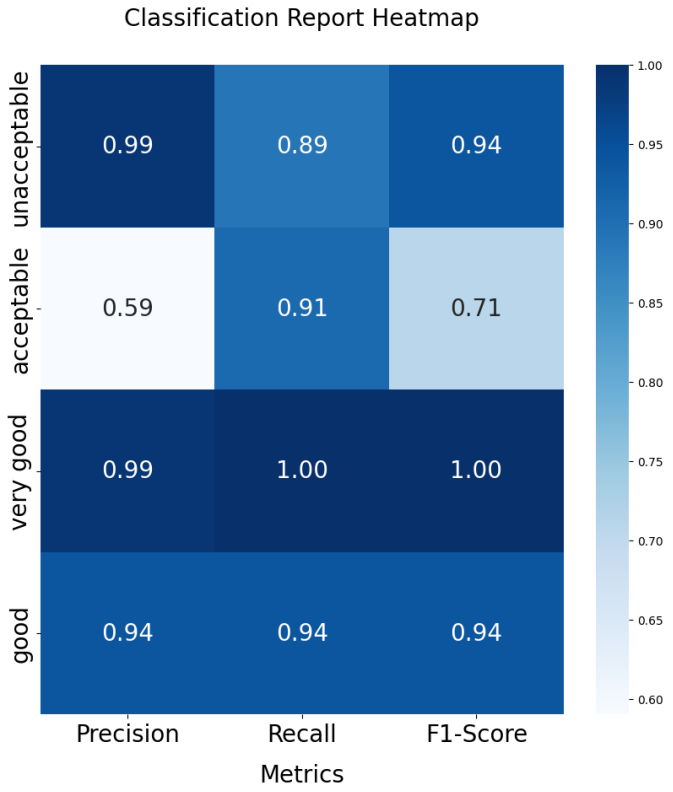

Upon examination of the correlation heatmap, it is evident that no variables exhibit a high degree of correlation with each other. This implies that the features under consideration are relatively independent, which can be advantageous in the context of machine learning modeling. The absence of strong 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 Random Forest model implementation for car evaluation using the provided dataset. Similar to the decision tree model, we began by preprocessing the data, converting categorical variables into numerical format for model training. The dataset was then split into training and testing sets to evaluate the model's performance.Our Random Forest model achieved an impressive accuracy of 97.11% 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 further emphasize the model's robustness in-car evaluation. The macro average F1-score, which considers all classes equally, stands at 92%, while the weighted average F1-score, which accounts for class frequencies, reaches 97%. These values highlight the model's overall effectiveness in accurately classifying car evaluations.

The results from our random forest 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 Random Forest model on the car evaluation dataset has demonstrated an exceptional level of accuracy in classifying cars into different evaluation categories: unacceptable, acceptable, very good, and good. This indicates the model's proficiency in assisting 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 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.analyticsvidhya.com/blog/2021/06/understanding-random-forest/
[2] https://www.javatpoint.com/machine-learning-random-forest-algorithm
[3] https://builtin.com/data-science/random-forest-algorithm
[4] https://www.ibm.com/topics/random-forest
[5] https://www.simplilearn.com/tutorials/machine-learning-tutorial/random-forest-algorithm