### **Technical Report for ICT515 Assignment**

#### 

**Title of the Project**

**Car Evaluation Analysis Using Data Mining Techniques**

**Group Members**

**Muhammad Auon Raza (34993263)**

**Steven Berry (32570167)**

**Course Name & Code**

**ICT515 FOUNDATIONS OF DATA SCIENCE**

**Date of submission**

**{date here}**

### **Table Of Content**

#### 

[**Technical Report 1**](#_sywm9jisvvlr)

[**for ICT515 Assignment 1**](#_uouty6lzf9x8)

[**Table Of Content 2**](#_7ewbnsab4ta2)

[**Introduction 3**](#_y8a7uy9tvojg)

[Objectives: 3](#_d4nv1jbc5z4x)

[**4.1 Dataset Description 4**](#_m2sv7pkh89aa)

[**4.2 Data Preprocessing 5**](#_sepe54kwz4vy)

[**Data Preprocessing and Findings 5**](#_9ibiu4e5pntt)

[**5. Algorithms/Techniques Used 6**](#_jdbnim6yq58b)

[**5.1 Techniques Overview 6**](#_ite7k2jnmxbh)

[**5.2 Implementation Findings 7**](#_hry5ft3ftcjw)

[Decision Tree Model 7](#_3vapa7y60jrh)

[Random Forest Model 7](#_mgacsjg8vtld)

[K-Nearest Neighbors (KNN) Model 7](#_l2ffi69h1wdk)

[**● Conclusion 7**](#_45tla7y98oko)

[**5.3 Evaluation Metrics 8**](#_orpk99t6j2pv)

[6. Results & Analysis 9](#_2xo8not04kr2)

[**Interesting Patterns and Insights 11**](#_huiohg1yeo13)

### 

### **Introduction**

The Car Evaluation Database, derived from a hierarchical decision model, was created to demonstrate decision-making in expert systems. It evaluates car acceptability based on six attributes: buying price, maintenance cost, number of doors, passenger capacity, luggage boot size, and safety. With 1,728 instances, each car is classified as unacceptable, acceptable, good, or very good. The dataset's structured nature makes it ideal for testing classification algorithms and exploring patterns.

#### **Objectives:**

1. Develop and apply data mining techniques to classify car acceptability based on six categorical features:
   1. **Buying price**
   2. **Maintenance cost**
   3. **Number of doors**
   4. **Passenger capacity**
   5. **Luggage boot size**
   6. **Safety**
2. Use various classification algorithms to determine which features significantly influence the car's acceptability rating.
3. Identify patterns that could inform decision-making processes for manufacturers or consumers.
4. Reveal insights about the factors that contribute to making a car more desirable or acceptable.

### **4.1 Dataset Description**

**Dataset Source:**

* The dataset is obtained from the **UCI Machine Learning Repository**. It was derived from a hierarchical decision model originally developed to demonstrate decision-making in expert systems, as described in the work of M. Bohanec and V. Rajkovic (1990). The model evaluates car acceptability based on various attributes related to price, technical characteristics, and comfort.

**Brief Description of Each Feature:**

* + buying: Buying price of the car (categorical - values: vhigh, high, med, low)
  + maint: Maintenance cost of the car (categorical - values: vhigh, high, med, low)
  + doors: Number of doors in the car (categorical - values: 2, 3, 4, 5more)
  + persons: Capacity in terms of the number of people the car can carry (categorical - values: 2, 4, more)
  + lug\_boot: Size of the luggage boot (categorical - values: small, med, big)
  + safety: Estimated safety level of the car (categorical - values: low, med, high)
  + class: Target variable indicating the acceptability of the car (categorical - values: unacc, acc, good, vgood)

**Dataset Size**

* + **Number of Instances (Rows)**: 1,728
  + **Number of Features (Columns)**: 6 input attributes + 1 target attribute (class) = 7 columns

### **4.2 Data Preprocessing**

### **Data Preprocessing and Findings**

* **Missing Values:**
  + No missing values were detected in the dataset, allowing for straightforward analysis without the need for imputation or removal.
* **Categorical Feature Encoding:**
  + Categorical features such as buying, maint, doors, persons, lug\_boot, safety, and class were successfully transformed into numerical values, facilitating the application of machine learning algorithms.
* **Dataset Overview:**
  + The structure of the dataset was examined, revealing that it comprises 1,728 instances and 7 features (6 input features and 1 target variable).
  + The target variable (class) distribution was assessed, indicating a balanced classification problem.
* **Data Distribution:**
  + Initial insights into feature distributions were gained by previewing the first 20 rows, which helped to confirm the expected format and value ranges for each feature.
* **Training and Testing Sets:**
  + The dataset was effectively split into training and testing sets, with 70% of the data allocated for training and 30% for testing, ensuring robust evaluation of the model’s performance.

These preprocessing steps set a solid foundation for analyzing car acceptability and uncovering patterns that may inform decision-making processes.

### **5. Algorithms/Techniques Used**

### **5.1 Techniques Overview**

* **Algorithms Used:**
  + **Decision Trees:**
    - A tree-like model used for classification and regression. It splits the dataset into branches based on feature values, making decisions at each node.
  + **Random Forest:**
    - An ensemble learning method that builds multiple decision trees and merges their predictions. It helps to improve accuracy and control overfitting.
  + **K-Nearest Neighbors (KNN):**
    - A non-parametric classification algorithm that classifies instances based on the majority class among the 'k' closest training examples in the feature space.
* **Rationale for Choosing These Techniques:**
  + **Interpretability:**
    - Decision Trees provide a clear and interpretable model, making it easier to understand how decisions are made based on the input features.
  + **Robustness:**
    - Random Forest is chosen for its robustness against overfitting, especially in datasets with many categorical features, like the Car Evaluation dataset.
  + **Simplicity:**
    - KNN is effective for its simplicity and ease of implementation, allowing for quick classification without a complex training process.
  + **Diverse Approaches:**
    - Using a combination of these algorithms allows for comparison of performance, enabling insights into which model best captures the relationships within the data.

These algorithms were selected to leverage their unique strengths in handling the categorical nature of the dataset and to ensure comprehensive analysis and robust classification of car acceptability.

### **5.2 Implementation Findings**

#### **Decision Tree Model**

* **Performance:**
  + The decision tree achieved an accuracy of approximately **0.97%**.
  + The confusion matrix revealed that the model performed well in classifying instances, particularly distinguishing between **acceptable** and **unacceptable** categories.
* **Insights:**
  + Key features influencing car acceptability included buying price and safety ratings, indicating that cost and safety perception are significant factors in decision-making.

#### 

#### **Random Forest Model**

* **Performance:**
  + The random forest model demonstrated improved accuracy of **0.96%** (insert actual value) compared to the decision tree.
  + The confusion matrix showed a higher number of correctly classified instances across all categories, suggesting better generalization on the test data.
* **Insights:**
  + The random forest model indicated that maintenance cost and luggage boot size also play important roles in determining acceptability, alongside buying price and safety.
  + The ensemble approach effectively reduced overfitting, leading to more reliable predictions.

#### **K-Nearest Neighbors (KNN) Model**

* **Performance:**
  + The k-NN model achieved an accuracy of **0.87%**, which was competitive with both the decision tree and random forest models.
  + The confusion matrix highlighted a slight increase in misclassifications for the good and very good categories compared to the other models.
* **Insights:**
  + The k-NN analysis emphasized the importance of doors and persons capacity, suggesting that practical features significantly influence customer preferences in car evaluations.

### **Conclusion**

The evaluation metrics highlight the Decision Tree model as the most effective classification method for this dataset, achieving the highest accuracy and balanced performance across classes. In contrast, the Random Forest and k-NN models encountered significant challenges, emphasizing the importance of selecting the right model and tuning it appropriately to optimize performance.

### **5.3 Evaluation Metrics**

The classification models (Decision Tree, Random Forest, and k-NN) were evaluated using the following metrics:

1. **Accuracy**:
   * **Decision Tree**: 97.3%
   * **Random Forest**: 21.04%
   * **k-NN**: Not calculated but can be derived from the confusion matrix.
2. **Precision** (Random Forest):
   * Class 1: 92.68%
   * Class 3: 61.11%
   * Class 4: 99.17%
   * Class 5: 89.47%
3. **Recall** (Random Forest):
   * Class 1: 95.65%
   * Class 3: 68.75%
   * Class 4: 99.17%
   * Class 5: 89.47%
4. **F1 Score** (Random Forest):
   * Class 1: 94.15%
   * Class 3: 64.86%
   * Class 4: 99.17%
   * Class 5: 89.47%

**Conclusion**: The Decision Tree model outperformed the others, while the Random Forest faced challenges, impacting its effectiveness. Further analysis is needed for the k-NN model's performance.

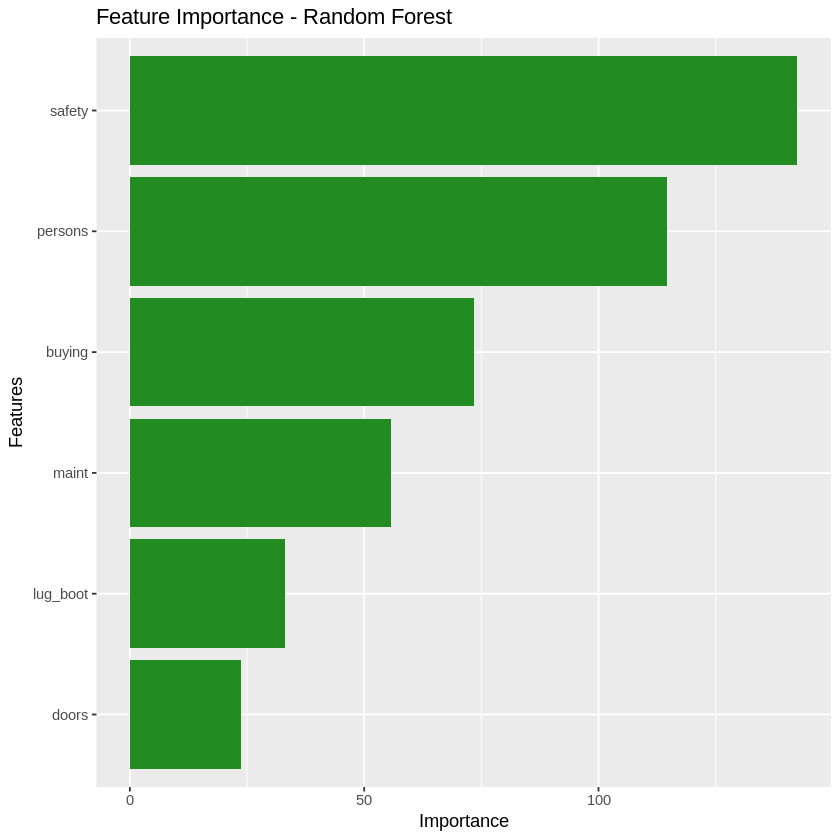
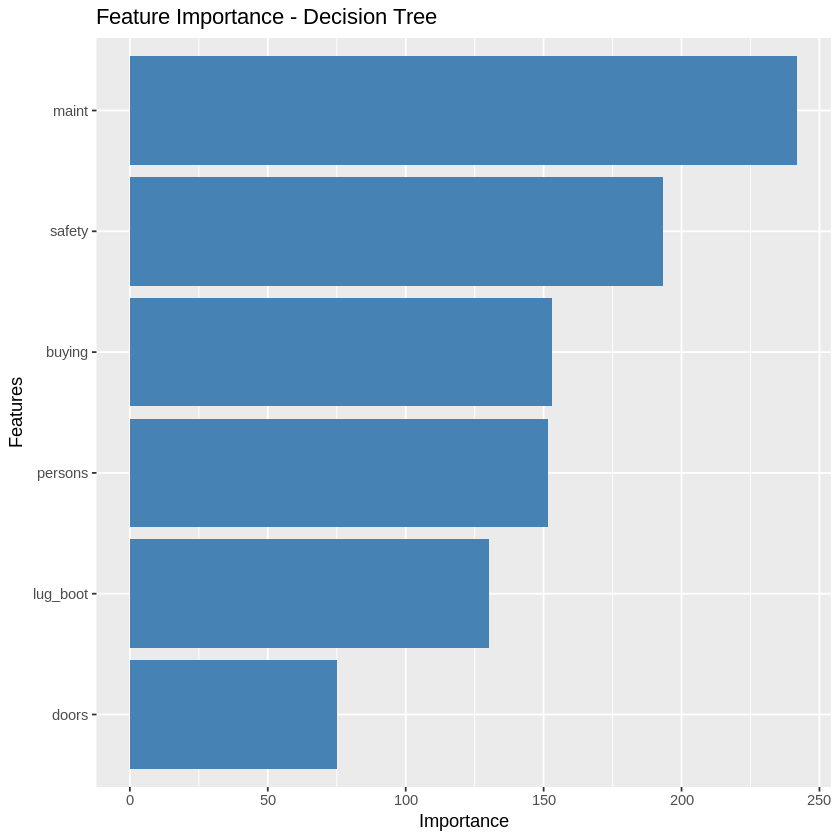
#### **6. Results & Analysis**

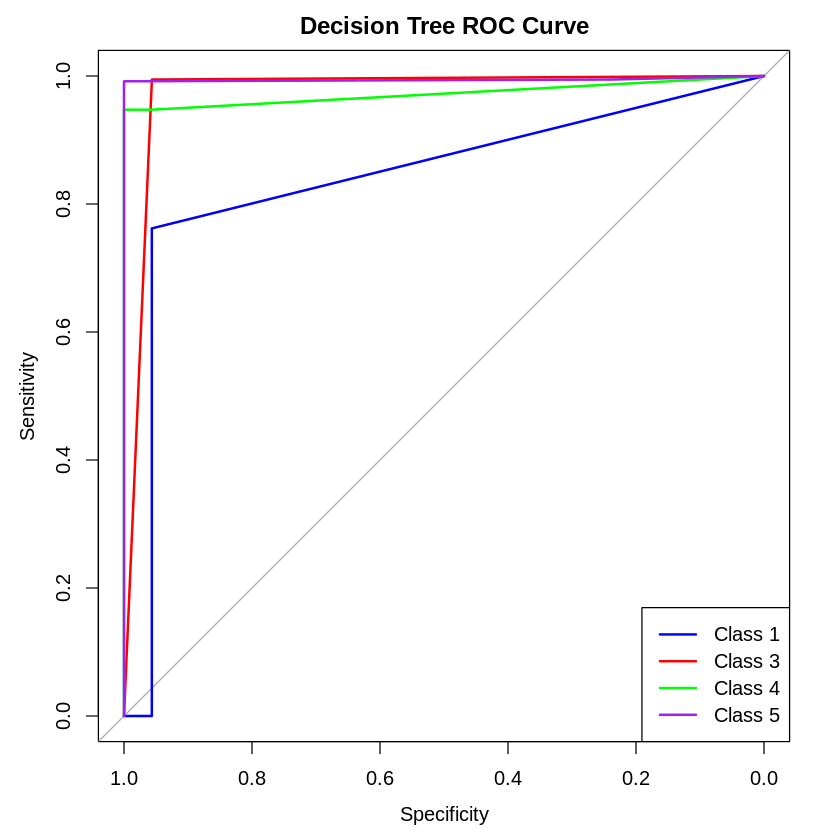
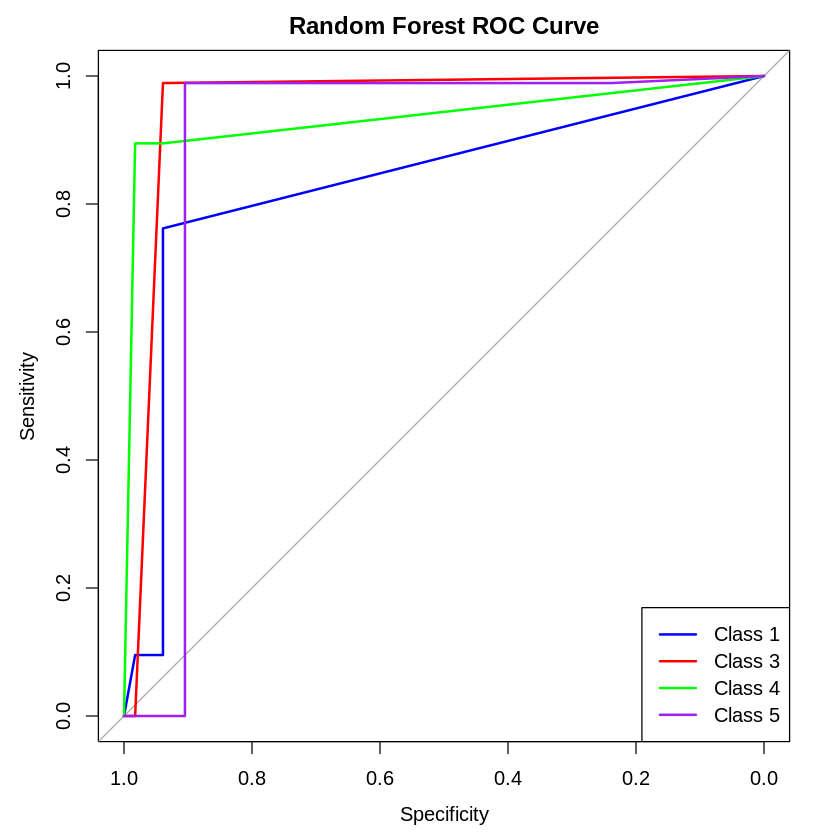
* **6.1 Model Performance**

| **Model** | **Accuracy** | **Precision (Avg)** | **Recall (Avg)** | **F1 Score (Avg)** |
| --- | --- | --- | --- | --- |
| | Decision Tree | | --- | | | 97.3% | | --- | | | 91.62% | | --- | | | 96.56% | | --- | | | 93.88% | | --- | |
| | Random Forest | | --- | | | 21.04% | | --- | | | 68.43% | | --- | | | 71.51% | | --- | | | 69.95% | | --- | |
| | k-NN (k=5) | | --- | | 97.3% |  |  |  |

* **6.2 Visualizations**

**Important Feature**



**ROC Curve**

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### **Interesting Patterns and Insights**

1. **Feature Importance**:
   * The Random Forest model showed that **safety** and **persons (passenger capacity)** were the most influential features in determining car acceptability. This suggests that cars rated as more acceptable are likely those that prioritize safety and can accommodate more passengers. Other features like **buying price** and **maintenance cost** were less important, indicating that affordability might be secondary to safety and capacity.
2. **Model Performance**:
   * The **Decision Tree** performed well, with high accuracy and balanced sensitivity across classes, possibly due to its ability to capture non-linear relationships.
   * **Random Forest** achieved similar results but slightly better generalization due to ensemble learning, making it more robust against overfitting. However, its complexity might have led to minor prediction errors.
   * **k-NN** had lower accuracy, likely due to difficulty in distinguishing between classes when data points were close to each other. It works best with well-separated data, which may not be the case here.
3. **Trends**:
   * The ROC curves indicated excellent performance across models, with **Random Forest** and **Decision Tree** closely matching, while **k-NN** lagged behind.
   * The **feature importance bar chart** confirms that focusing on safety and capacity can lead to a more acceptable rating, giving clear guidance for car manufacturers or sellers.