# **Comprehensive Data Exploration and Predictive Modeling for Car Acceptability Using Decision Trees**

# **Comprehensive Exploration of the Training Dataset**

The process of building a robust and effective model begins with a comprehensive exploration of the dataset used for training. This pivotal phase involves delving into the intricacies of the data, uncovering patterns, identifying outliers, and gaining insights that lay the groundwork for informed decision-making throughout the model development process (Idreos et al.).

## **Summary Statistics**

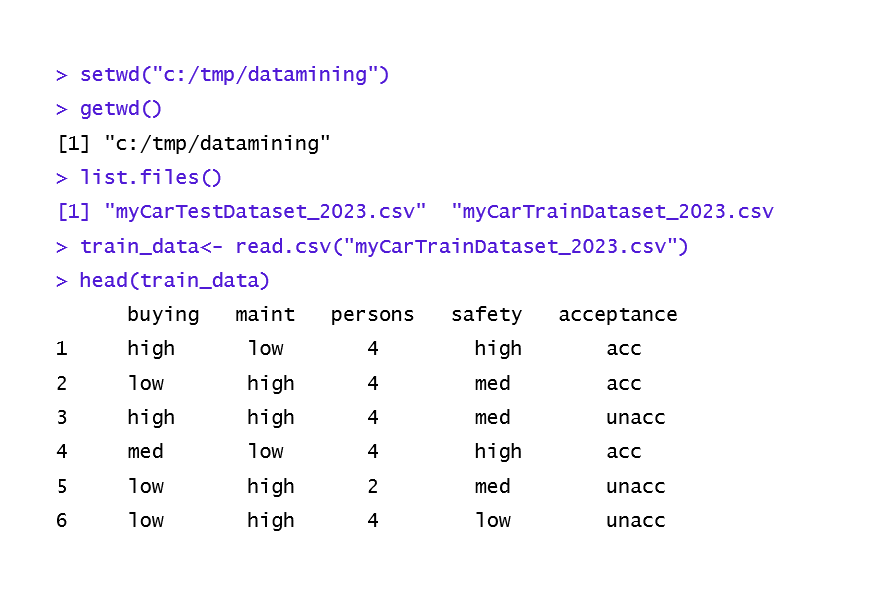
In order to gain a comprehensive understanding of the training dataset and extract meaningful insights, a thorough data exploration process was conducted using R Studio environment as shown in Figure 1.

Figure 1. Screenshot image from Rstudio

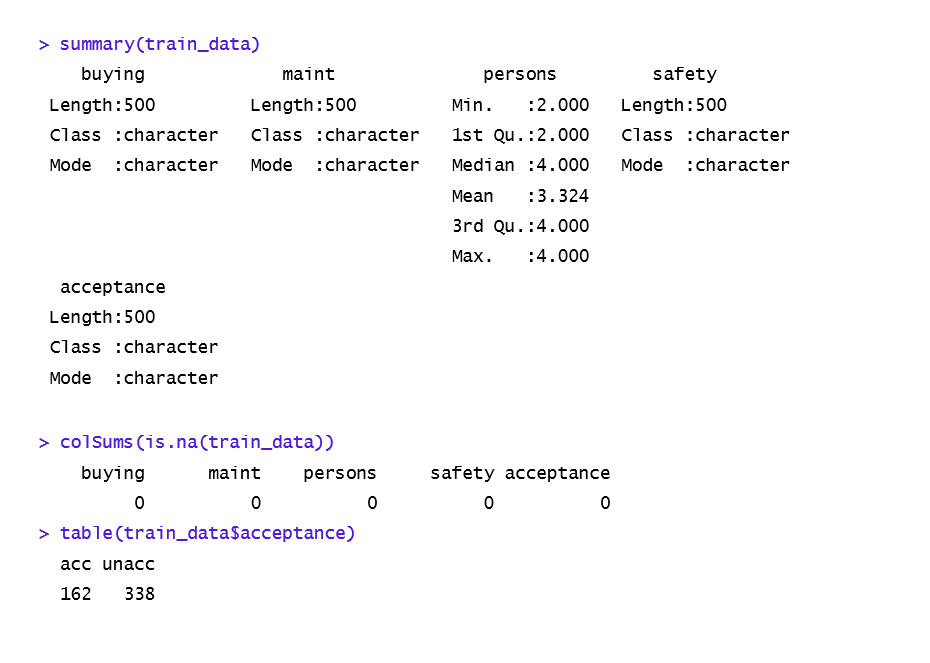
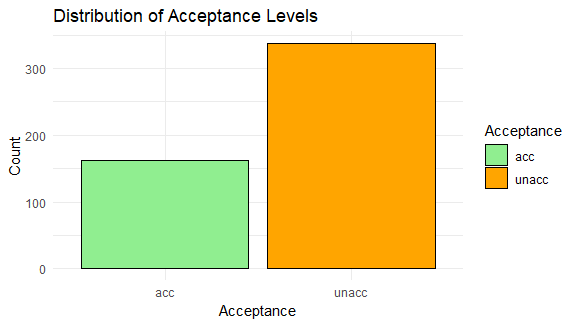
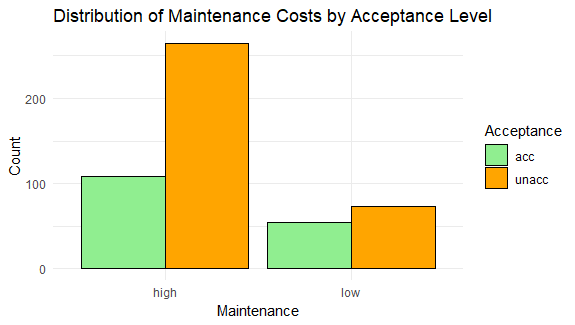
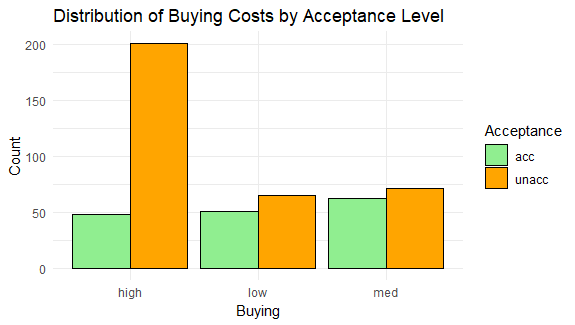
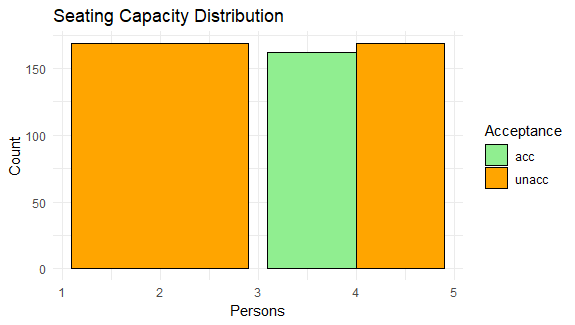


Figure 1. Screenshot image from Rstudio (contd.,)

In examining the training dataset for car acceptability, a comprehensive set of attributes, including "buying," "maint," "persons," "safety," and the target variable "acceptance" were observed. The dataset comprised of 500 rows with no missing values and this reassures that none were present across the columns which indicated a clean dataset ready for analysis.

A notable feature of the dataset is the imbalance in the target variable, "acceptance," with 338 instances labeled as "unacc" and 162 as "acc." This skew suggests a prevalence of cars being deemed "unacceptable," which could influence the performance of predictive models. Addressing this class imbalance is crucial to ensure balanced predictions.

The dataset visualization uncovers crucial insights into buying costs, maintenance costs, seating capacity, and safety ratings as pivotal factors in determining car acceptability. Addressing class imbalance is recommended for robust predictive modeling, providing a solid foundation for further analysis and model development (Ieno and Zuur, 2015).



(a)

(e)

(d)

(c)

(b)

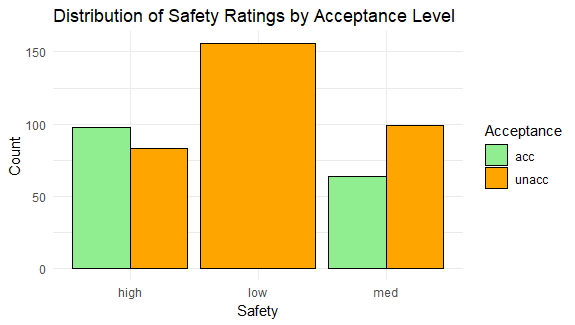


Figure 2. Data Visualization. (a) Distribution of Acceptance levels, (b) Distribution of Buying cost by Acceptance levels, (c) Distribution of Maintenance cost by Acceptance levels, (d) Seating capacity distribution and (e) Distribution of Safety ratings by Acceptance levels

## **Visualization of Acceptance Distribution**

The dataset analysis illuminates the prevalence of unacceptable cars, with "unacc" (Unacceptable) dominating as shown in Figure 2(a). The higher instances of unacceptable cars suggest potential shortcomings or non-compliance with specific criteria.

## **Visualization of Buying and Maintenance Costs**

In examining the relationship between buying and maintenance costs and the acceptability of cars, distinctive patterns emerge, revealing key insights into the factors influencing acceptability. High buying costs demonstrate a significant correlation with cars falling into the "unacc" category, signaling a potential association between elevated expenses and unacceptability as shown in Figure 2(b). Conversely, the occurrence of "acc" instances is notably lower in comparison. The trend continues with low buying costs, where a higher proportion of cars are labeled as "unacc," despite the presence of acceptable cars with lower costs, albeit in smaller numbers. In the context of medium buying costs, a balanced distribution is observed between acceptable and unacceptable cars.

As Figure 2(c) signifies, cars with high maintenance costs show a pronounced inclination towards unacceptability, with a notable disparity between "unacc" and "acc" instances. Similarly, for low maintenance costs, a higher number of cars are labeled as "unacc," even though acceptable instances with lower costs exist. This analysis provides a nuanced understanding of how buying and maintenance costs contribute to the categorization of cars as acceptable or unacceptable, emphasizing the significance of these financial considerations in determining overall acceptability.

## **Visualization of Seating Capacity Distribution**

The analysis shown in Figure 2(d) suggests, a correlation exists between seating capacity and car acceptability. For persons < 3, cars are more likely to be labeled as unacceptable. On the other hand, for persons > = 3, cars are labeled as acceptable and unacceptable.

## **Visualization of Safety Ratings**

The examination of safety ratings in determining car acceptability reveals distinct patterns. Cars with high safety ratings are predominantly labeled as acceptable, showcasing a positive correlation as shown in Figure 2(e). Conversely, the scarcity of acceptable cars with low safety ratings and a high count of unacceptable cars suggests a strong link between low safety ratings and unacceptability. In the medium safety category, a balanced distribution leans slightly towards unacceptability. These insights provide valuable considerations for predictive modeling, emphasizing the importance of safety ratings in determining overall car acceptability.

# **Building a Full Decision Tree for Car Acceptance Prediction**

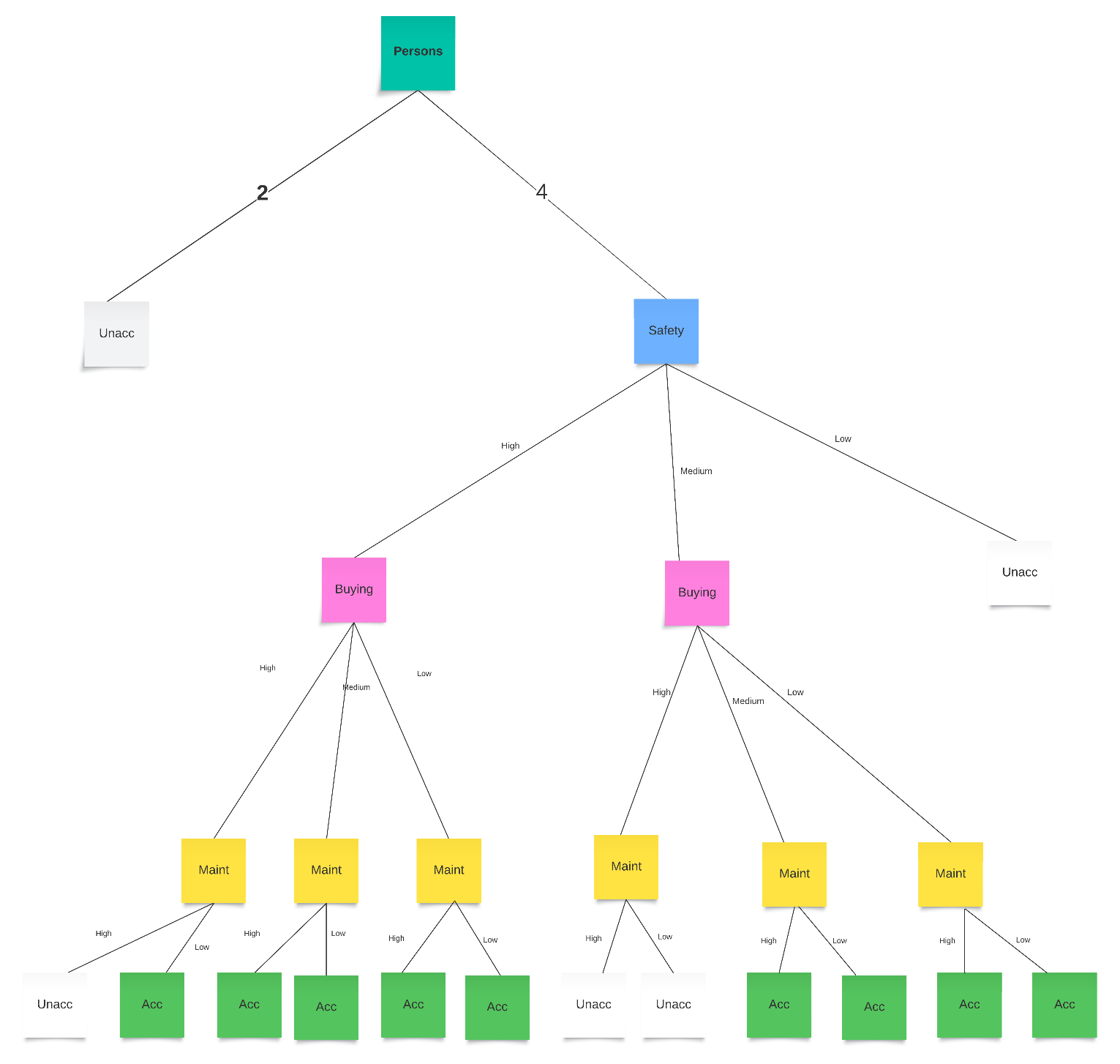
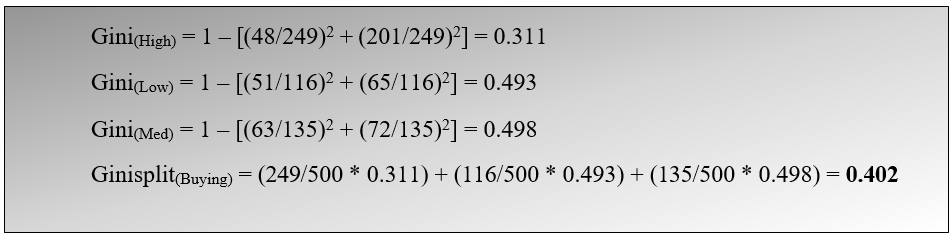
This section details the steps involved in building a fully grown decision tree. Data mining and machine learning methods were employed to analyze a dataset. The Hunt Algorithm combined with Greedy Strategy using Gini impurity was applied and facilitated the construction of a decision tree incorporating a multiway split as shown in Figure 3 (Myles et al., 2004), (Barros et al., 2015).

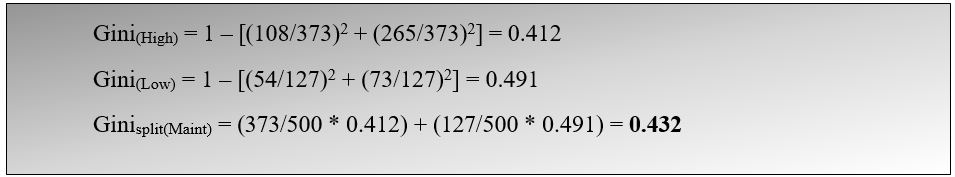
Figure 3. Fully grown decision tree (self-created)

The fully grown decision tree was constructed using the Gini impurity measure and the sample calculation are as follows.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **High (249)** | | **Low(116)** | | **Med(135)** | |
| **acc** | **48** | **acc** | **51** | **acc** | **63** |
| **unacc** | **201** | **unacc** | **65** | **unacc** | **72** |

**Buying**

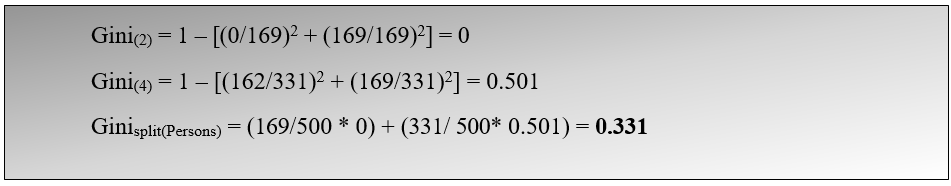


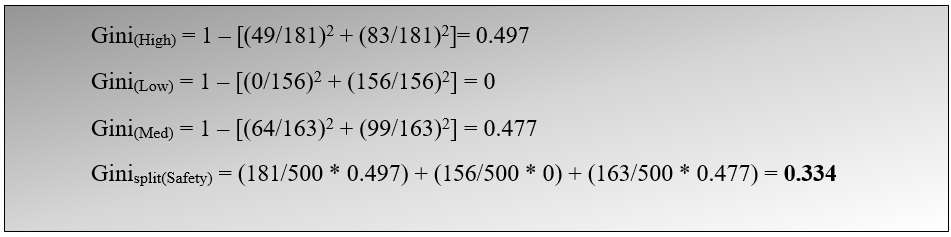
**Maintenance**

|  |  |  |  |
| --- | --- | --- | --- |
| **High (373)** | | **Low(127)** | |
| **acc** | **108** | **acc** | **54** |
| **unacc** | **265** | **unacc** | **73** |

|  |  |  |  |
| --- | --- | --- | --- |
| **2 (169)** | | **4(331)** | |
| **acc** | **0** | **acc** | **162** |
| **unacc** | **169** | **unacc** | **160** |

**Persons**

****

**Safety**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **High (181)** | | **Low(156)** | | **Med(163)** | |
| **acc** | **98** | **acc** | **0** | **acc** | **64** |
| **unacc** | **83** | **unacc** | **156** | **unacc** | **99** |

Gini impurity split value is the specific value of the attribute at which the dataset is divided into subsets (Padula, 2023). By following Greedy strategy, ‘**Persons**’ was selected as the root node of the decision tree as it had the lowest Ginisplit value of **0.331**. Following this, Gini split was calculated for other attributes such as buying, maintenance and safety by keeping Persons= 4. Then, the best split was spotted using the least Ginisplit value. The calculation was then performed iteratively (as shown in Appendix I). The algorithm systematically evaluates different split points for each attribute, computing the Gini impurity at each step. The iteration continues until an optimal point is reached. This optimal point corresponds to the split value that minimizes the Gini impurity, indicating the best way to partition the data based on the selected attribute.

# **Pruned Fully Grown Decision Tree and Confusion Matrix**

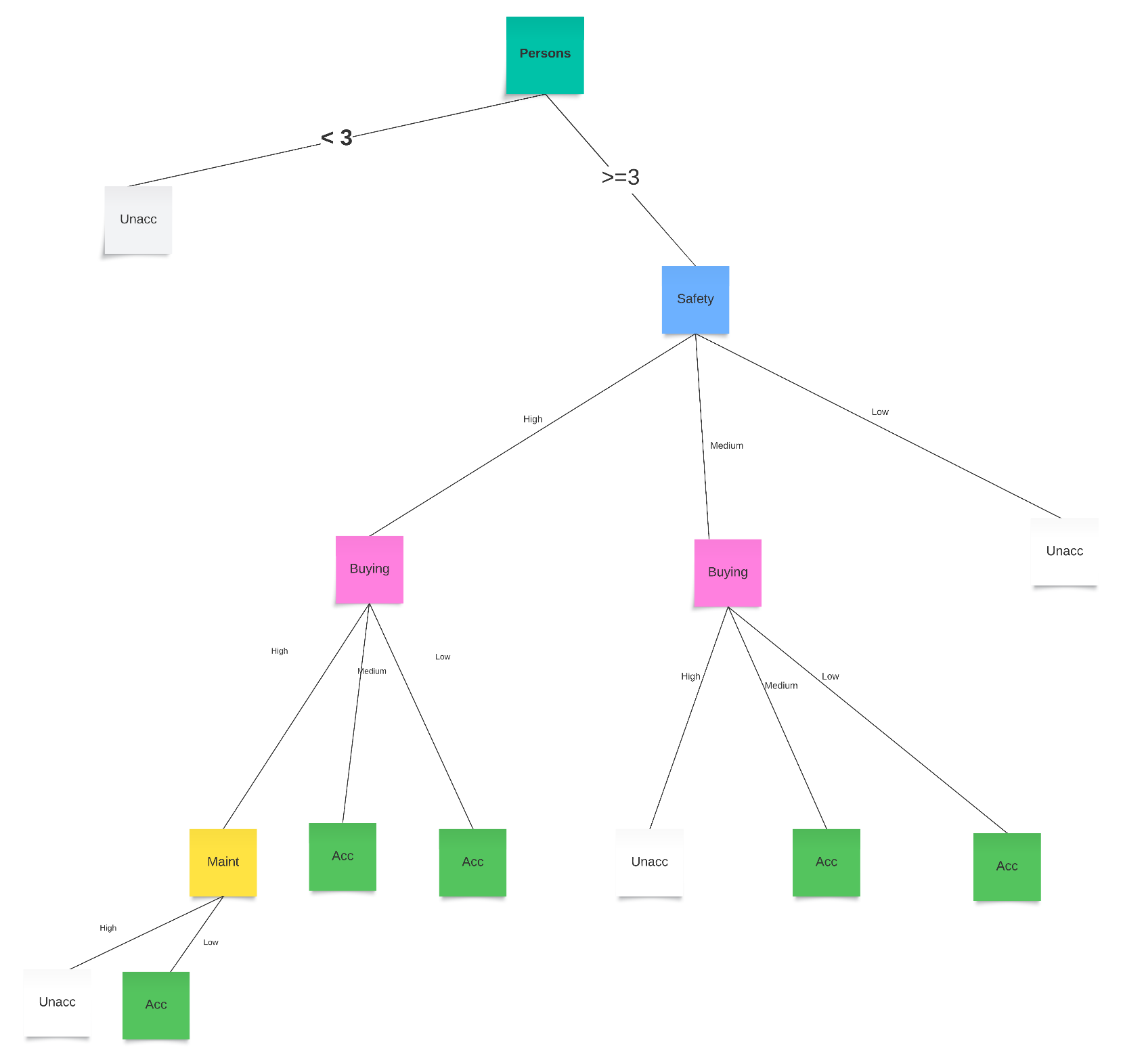
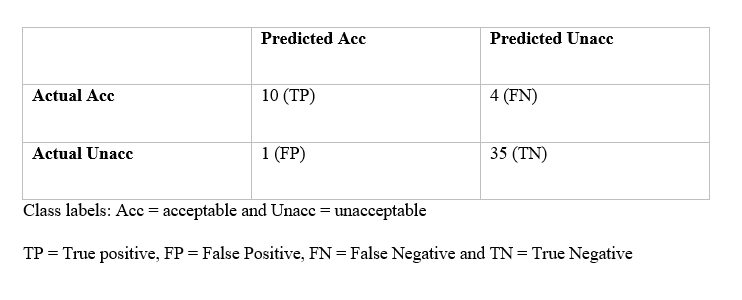
The previous sections entailed the meticulous construction of a fully-grown decision tree using Data mining and machine learning methods. Subsequent to this, a crucial post-pruning procedure (as shown in Figure 3) was implemented on sub-trees where every leaf node shares a uniform class label (Bonaccorso, 2018). Post-pruning involves growing the tree to its maximum size and then removing or collapsing nodes that do not contribute significantly to the overall predictive power of the tree. The primary aim is to simplify the structure of the tree to improve its generalization to new data as shown in Figure 4. By removing unnecessary details captured during training, pruning helps to reduce overfitting and promotes better performance on unseen data (Yeom et al., 2021).

Figure 3. Pruned fully grown decision tree (self-created)

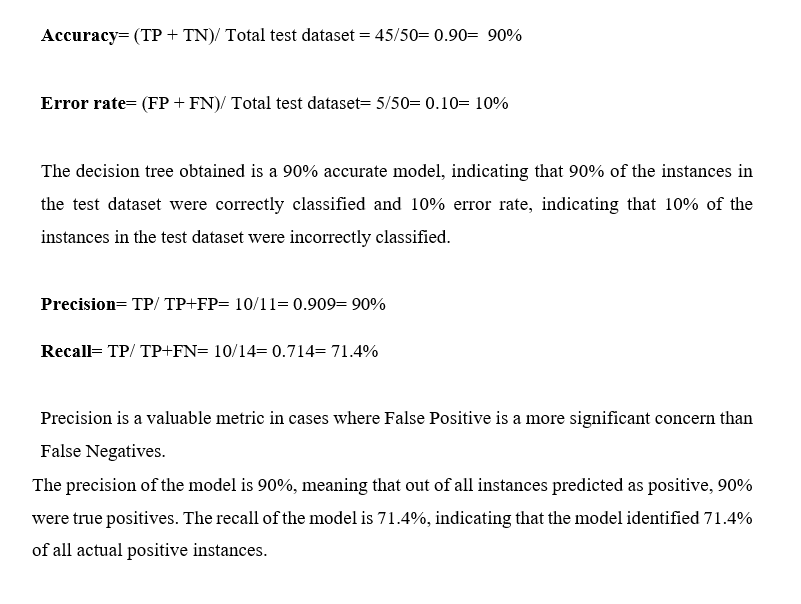
# **Confusion Matrix**

The pruned decision tree was then subjected to testing using a dedicated test dataset. Predictions were generated through the utilization of the predict function in the R Studio environment. A pivotal tool in the evaluation of the model's performance, the confusion matrix encapsulates a comprehensive summary of predicted and actual class labels. It is widely used in machine learning for supervised classification or determination of the behaviour of classification models (Wang et al., 2022). It elucidates key metrics such as true positives, true negatives, false positives, and false negatives, forming the basis for assessing the model's accuracy, precision and recall.

Interpreting the results within the contextual framework of the case involves a nuanced consideration of the implications for predicting car acceptability. The confusion matrix provides valuable insights into the model's strengths and weaknesses, elucidating areas of proficiency and potential avenues for improvement. The context-specific interpretation delves into the significance of correct predictions (true positives and true negatives) and the ramifications of misclassifications (false positives and false negatives) in the context of predicting car acceptability (Hasnain et al., 2020).

Using the pruned decision tree values as the actual class and the test dataset values as the predicted class, the confusion matrix was made and presented below.

From the above confusion matrix the following result was obtained:



**Data Mining and Machine Learning Approaches in Clinical Medicine and Sports**

Decision tree algorithms and pruning techniques have extensive applications across various industries due to their versatile nature and effectiveness in solving complex problems. This section provides the insights of DM and ML approaches in both the sports and clinical medicine domains based on insights gained from reviewing relevant journal articles.

According to Obermeyer and Emanuel, 2016, the study highlights the transformative impact of machine learning (ML) on clinical medicine by drawing parallels between ML algorithms and medical professionals in their problem-solving approaches. ML, resembling a doctor in residency, learns rules from patient-level observations by adeptly handling numerous predictors in nonlinear ways. This enables the analysis of diverse medical data from radiographs to insurance claims expanding the understanding of complex medical information (Obermeyer and Emanuel, 2016).

Challenges associated with ML in clinical contexts are emphasized including the need for careful consideration of data quality, potential biases and the risk of overfitting. Addressing these challenges through independent validation sets is crucial for ensuring the reliability and generalizability of ML models in healthcare applications.

Furthermore, the study predicts disruptions in three key medical areas. ML is expected to enhance prognostic models by leveraging rich predictor variables from electronic health records specifically benefiting patients with serious illnesses. Automation of image interpretation and surpassing human accuracy is anticipated to transform the roles of radiologists and pathologists. ML is also expected to improve diagnostic accuracy by generating differential diagnoses and suggesting high-value tests by reducing errors and overuse.

Despite the optimistic outlook on ML's potential benefits, the text also acknowledges ongoing challenges including the need for clear diagnostic standards, accessibility to high-value electronic health record data and individualized model development for each diagnosis. The concluding sentiment is optimistic with the belief that patients will be the primary beneficiaries as ML continues to reshape and augment clinical medicine.

The research paper by Gifford and Bayrak, 2023 underscores the efficacy of machine learning (ML), particularly decision trees in extracting valuable insights from extensive sports data. Through the implementation of decision tree and logistic regression models, the study effectively establishes measurable correlations between diverse team statistics and NFL game outcomes. Notably, the decision tree model exhibits a 79% accuracy in predicting wins and losses while the logistic regression model achieves an 83% accuracy, showcasing the robust predictive capabilities of ML in sports analytics (Gifford and Bayrak, 2023).

Nevertheless, the study introduces a crucial consideration for future ML applications in sports analytics. While the models reveal correlations, the potential for reverse causation wherein winning impacts team statistics, necessitates further exploration. The research also advocates for an examination of the temporal dynamics of team statistics, exploring trends over different game quarters and evolving patterns over time. Additionally, the study encourages the development of predictive models for anticipating future game outcomes. This nuanced comprehension can enhance the application of ML in sports analytics, providing coaches and decision-makers with more refined insights in the competitive sports landscape.

# **References**

BARROS, R. C., CARVALHO, A. C. P. D. L. F. & FREITAS, A. A. 2015. *Automatic Design of Decision-Tree Induction Algorithms Rodrigo C. Barros, André C.P.L.F. de Carvalho, Alex A. Freitas,* Cham [Switzerland, Springer.

BONACCORSO, G. 2018. *Machine Learning Algorithms Popular Algorithms for Data Science and Machine Learning, 2nd Edition,* Birmingham, Packt Publishing Ltd.

GIFFORD, M. & BAYRAK, T. 2023. A predictive analytics model for forecasting outcomes in the National Football League games using decision tree and logistic regression. *Decision analytics journal,* 8**,** 100296.

HASNAIN, M., PASHA, M. F., GHANI, I., IMRAN, M., ALZAHRANI, M. Y. & BUDIARTO, R. 2020. Evaluating Trust Prediction and Confusion Matrix Measures for Web Services Ranking. *IEEE access,* 8**,** 90847-90861.

IDREOS, S., PAPAEMMANOUIL, O. & CHAUDHURI, S. Overview of Data Exploration Techniques. International Conference on Management of Data, 2015. ACM, 277-281.

IENO, E. N. & ZUUR, A. F. 2015. *A beginner's guide to data exploration and visualisation with R / Elena N. Ieno, Alain F. Zuur,* Newburgh, United Kingdom, Highland Statistics Ltd.

MYLES, A. J., FEUDALE, R. N., LIU, Y., WOODY, N. A. & BROWN, S. D. 2004. An introduction to decision tree modeling. *Journal of chemometrics,* 18**,** 275-285.

OBERMEYER, Z. & EMANUEL, E. J. 2016. Predicting the Future — Big Data, Machine Learning, and Clinical Medicine. *The New England journal of medicine,* 375**,** 1216-1219.

PADULA, W. V. 2023. Introduction to decision tree modeling. Oxford: Oxford University Press.

WANG, Y., JIA, Y., TIAN, Y. & XIAO, J. 2022. Deep reinforcement learning with the confusion-matrix-based dynamic reward function for customer credit scoring. *Expert Systems with Applications,* 200**,** 117013.

YEOM, S.-K., SEEGERER, P., LAPUSCHKIN, S., BINDER, A., WIEDEMANN, S., MÜLLER, K.-R. & SAMEK, W. 2021. Pruning by explaining: A novel criterion for deep neural network pruning. *Pattern Recognition,* 115**,** 107899.