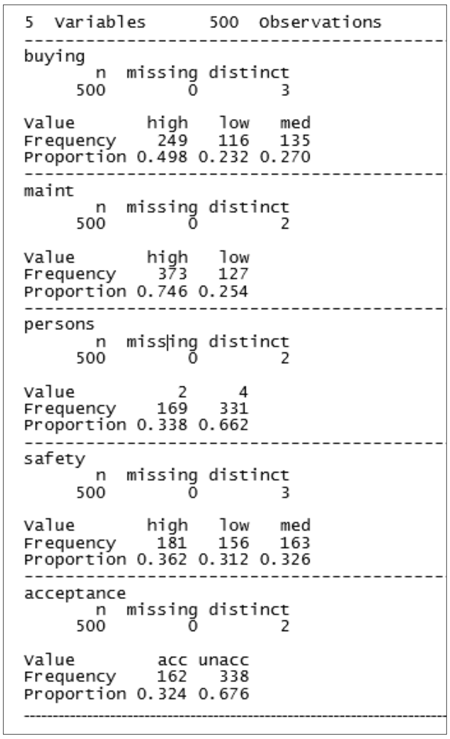
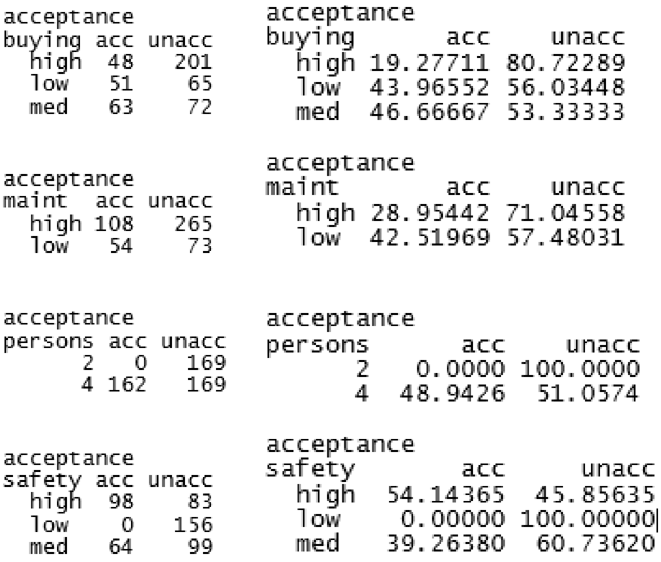


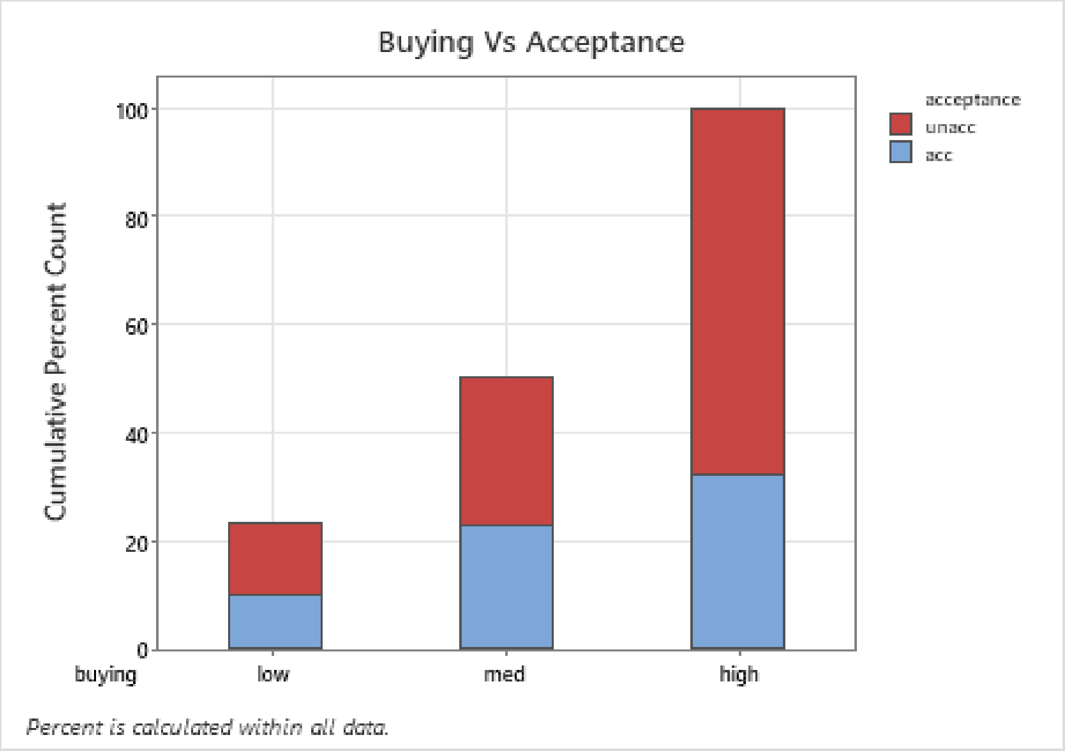
```
5  Variables      500  Observations
------------------------------------------
buying
       n  missing distinct
     500        0        3

Value        high   low   med
Frequency     249   116   135
Proportion 0.498 0.232 0.270
------------------------------------------
maint
       n  missing distinct
     500        0        2

Value        high   low
Frequency     373   127
Proportion 0.746 0.254
------------------------------------------
persons
       n  missing distinct
     500        0        2

Value           2     4
Frequency     169   331
Proportion 0.338 0.662
------------------------------------------
safety
       n  missing distinct
     500        0        3

Value        high   low   med
Frequency     181   156   163
Proportion 0.362 0.312 0.326
------------------------------------------
acceptance
       n  missing distinct
     500        0        2

Value         acc unacc
Frequency     162   338
Proportion 0.324 0.676
------------------------------------------
```



```
      acceptance                acceptance
buying acc unacc            buying        acc     unacc
  high  48   201              high 19.27711 80.72289
  low   51    65              low  43.96552 56.03448
  med   63    72              med  46.66667 53.33333

      acceptance                acceptance
maint  acc unacc            maint         acc     unacc
  high 108   265              high 28.95442 71.04558
  low   54    73              low  42.51969 57.48031

       acceptance               acceptance
persons acc unacc           persons       acc    unacc
      2   0   169                 2    0.0000 100.0000
      4 162   169                 4   48.9426  51.0574

      acceptance                acceptance
safety acc unacc            safety        acc      unacc
  high  98    83              high   54.14365  45.85635
  low    0   156              low     0.00000 100.00000
  med   64    99              med    39.26380  60.73620
```

Figure 1: Statistics using Chi-Square function

Figure 2: Attributes Vs Acceptance



Percent is calculated within all data.

Figure 3: Buying Vs Acceptance

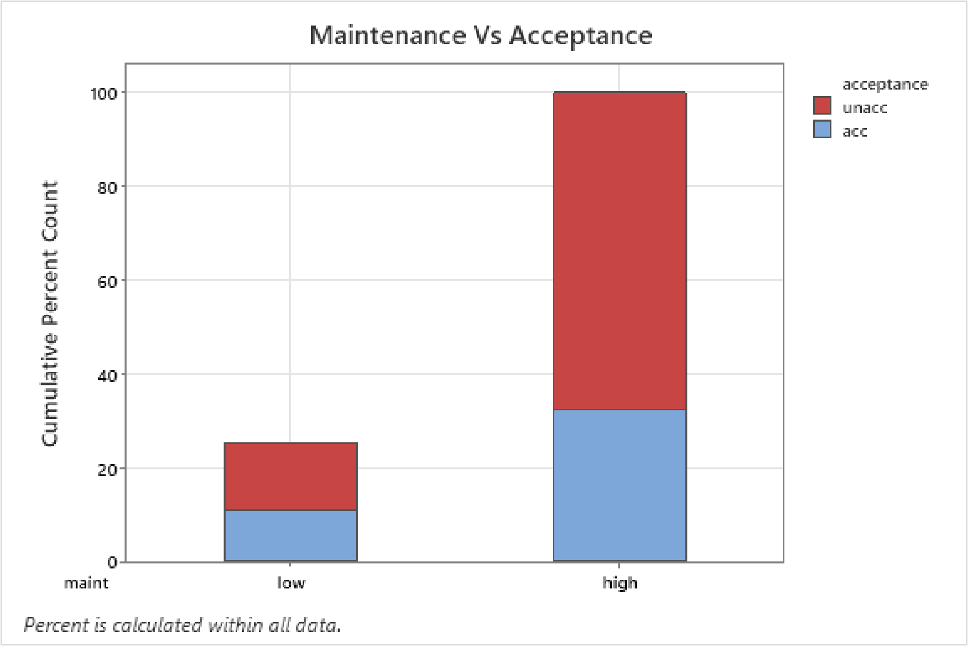


Figure 4: Maintenance Vs Acceptance

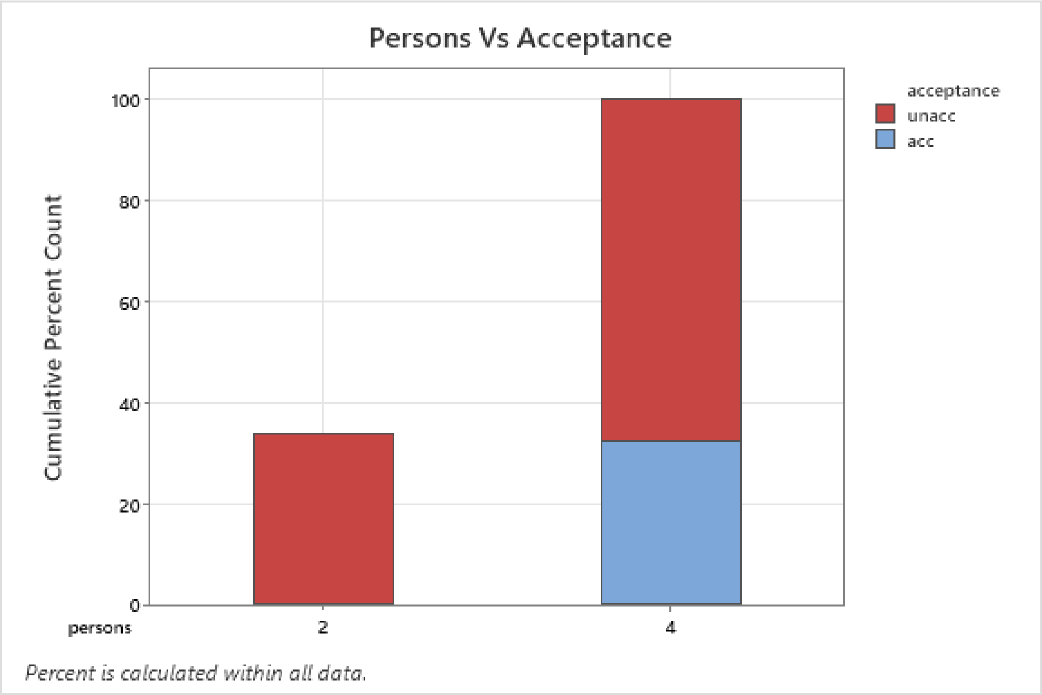


Figure 5: Persons Vs Acceptance

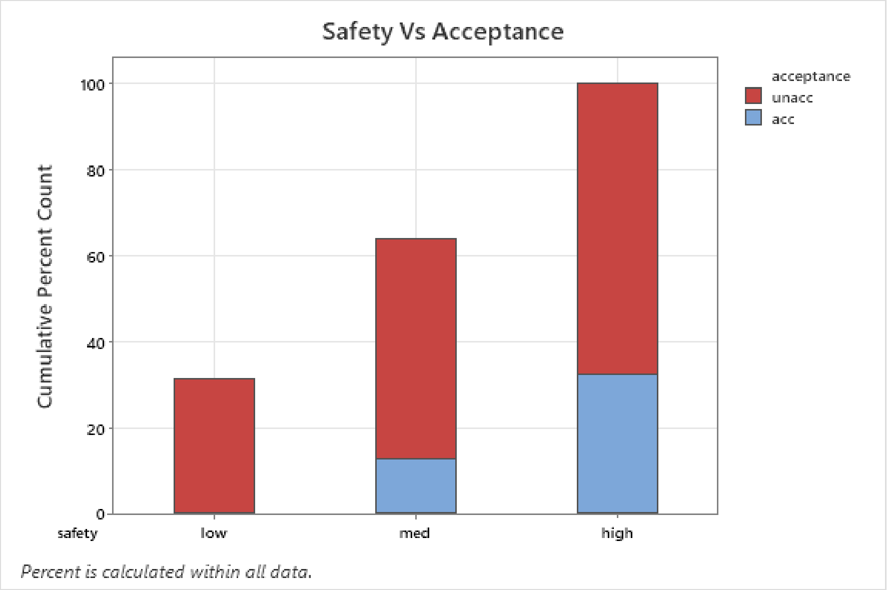


Figure 6: Safety Vs Acceptance

2. DECISION TREE

Decision tree (DT) algorithms are commonly employed for classification, particularly suited for categorical data. The objective is to classify data into a finite number of classes based on the values of input variables. The greedy strategy grows a DT by making a series of locally optimum decisions regarding which attribute to use for partitioning the data. Hunt's algorithm, ID3, C4.5, CART are greedy DT induction algorithms (Mukhopadhyay and Koturwar, 2014).

Hunt's algorithm constructs a DT recursively by partitioning a dataset into smaller, more homogenous subsets. The recursive process terminates when every record in a node belongs to the same class. When the training set compromises records of a single class, a corresponding leaf node (finalized decision) is assigned. If a node contains records from various classes, an attribute test condition is used to split the data into more purer subsets. Child nodes are then generated for each subset and this recursive process persists until all leaf nodes are established (Mohanapriya and Lekha, 2018). The split condition's attribute is usually one that maximises information gain or minimises impurity, such as the GINI index used in this report. It assesses the purity of a specific class by splitting along a particular attribute, and the split condition with the lowest GINI split is chosen. The goal is to find the optimal split that increases the purity of the resulting subsets, aiming for greater certainty, precision and accuracy in the algorithm's outcomes (Tangirala, 2020).

RESULTS

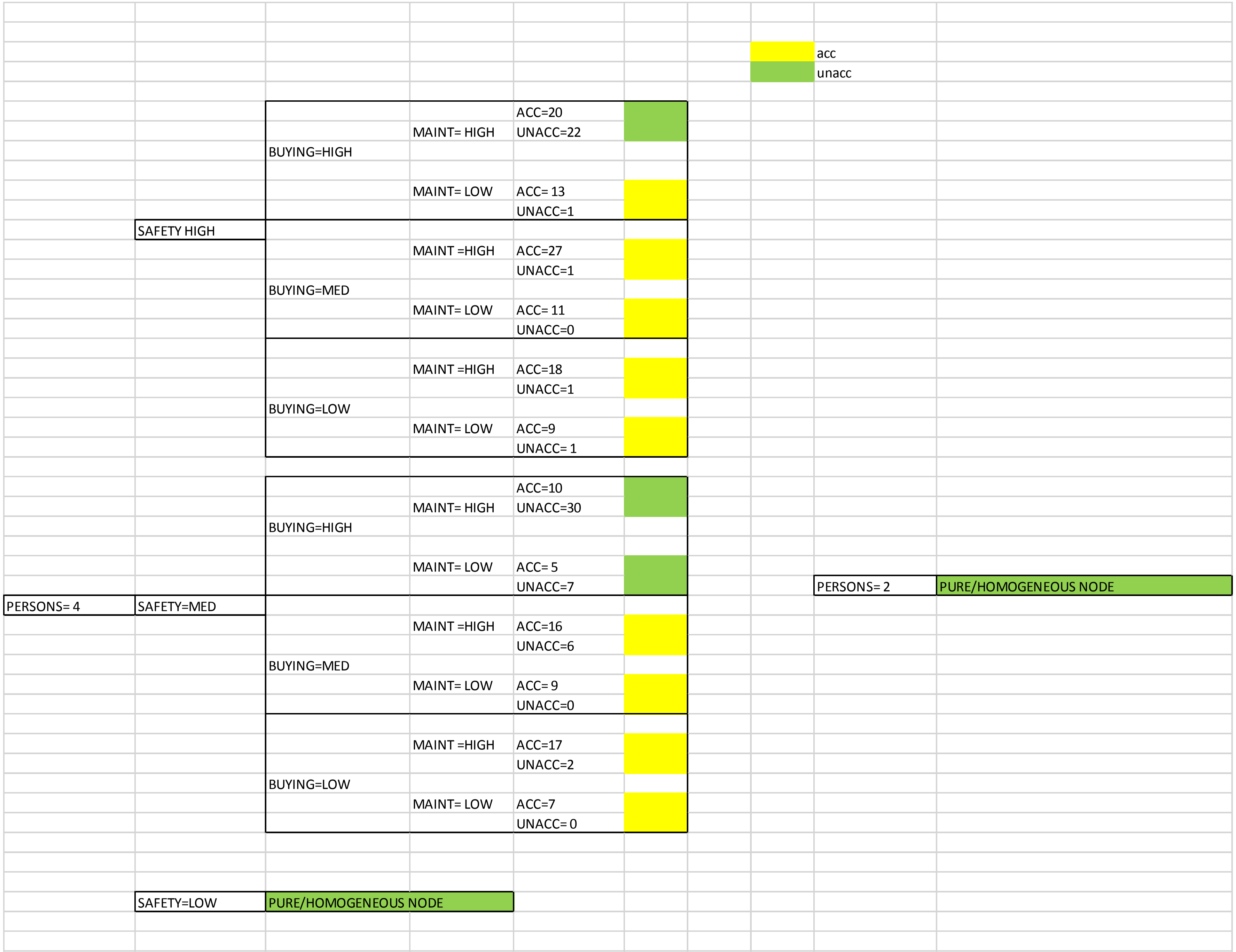


Figure 8: DT breakdown and acceptance/unacceptance instances to maintenance

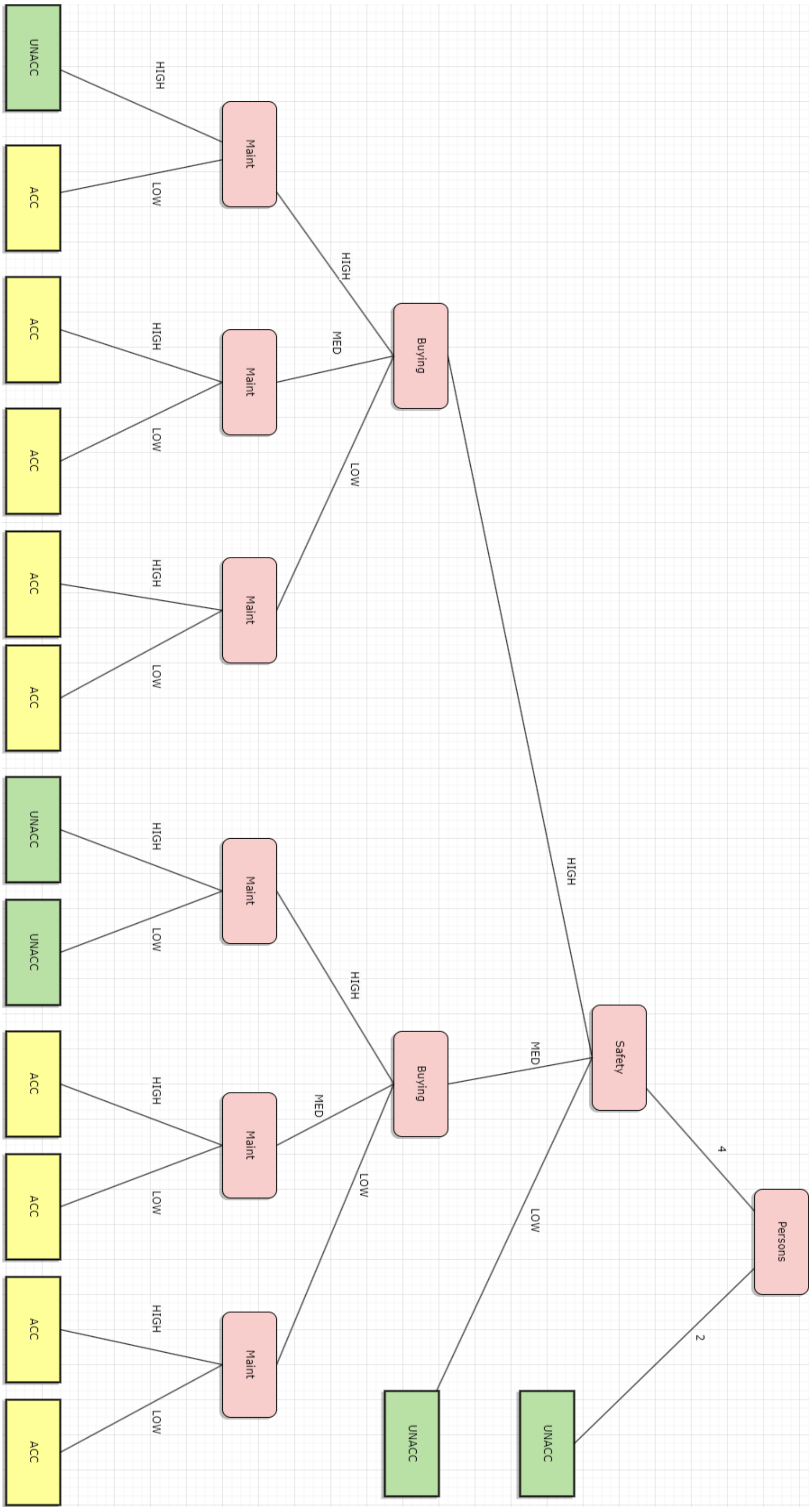


Figure 9: fully-grown DT

3. POST-PRUNING AND CONFUSION MATRIX

The pruning stage aims to reduce branches lacking statistical validity within DTs. Pruning is employed to enhance tree comprehensibility by reducing its size and maintaining/improving accuracy (Barros, De Carvalho and Freitas, 2015). There are two pruning methods: pre-pruning and post-pruning. Post-pruning follows a bottom-up approach by first generating a fully-grown DT. It then identifies and removes insignificant branches, replacing them with leaf nodes that represent values associated with the most prevalent instances of each internal node (Mehedi Shamrat *et al.*, 2021). The criteria for post-pruning the DT involves the application of a greedy strategy, which makes locally optimum decisions. Pruning is essential to prevent data overfitting by removing overfitted fragments and noisy/erroneous data, enhancing DTs' ability to provide efficient predictions.

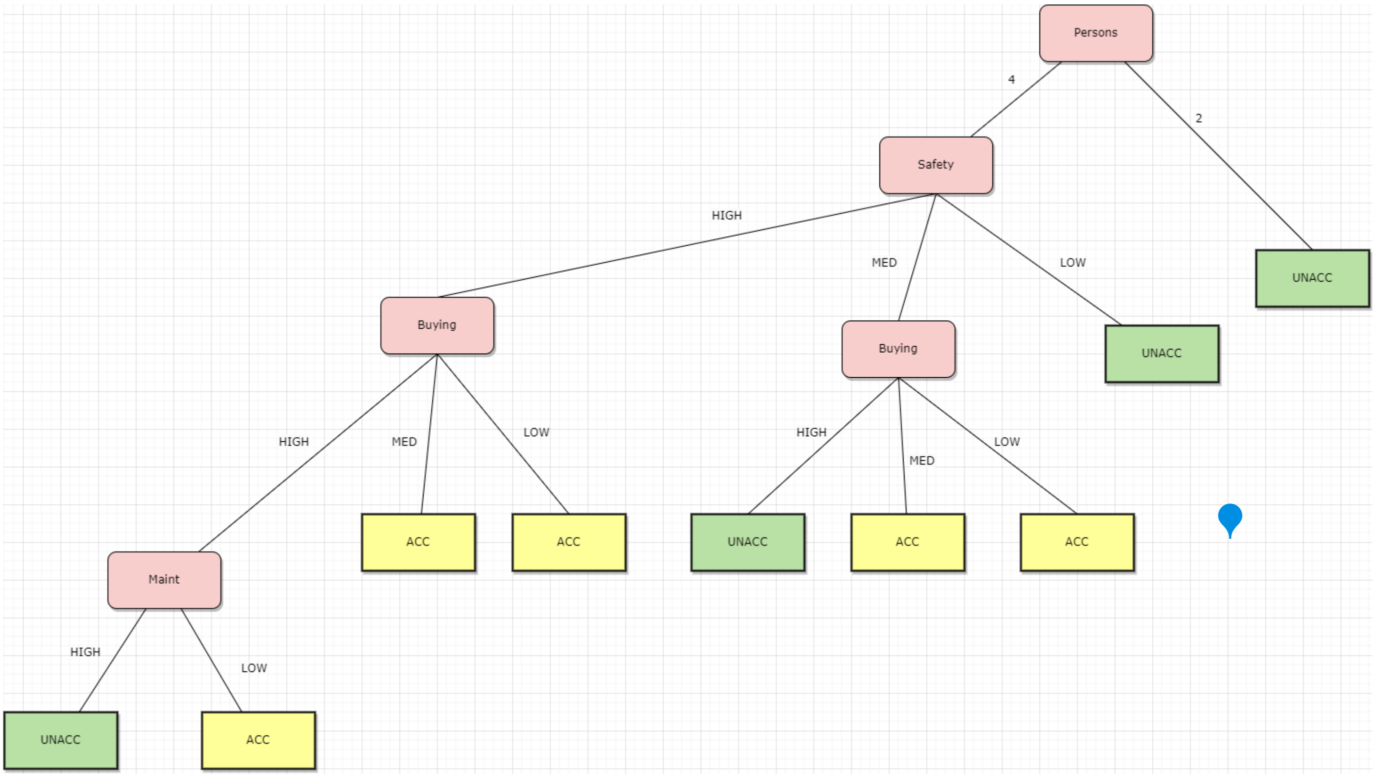


Figure 10: Post-pruned DT

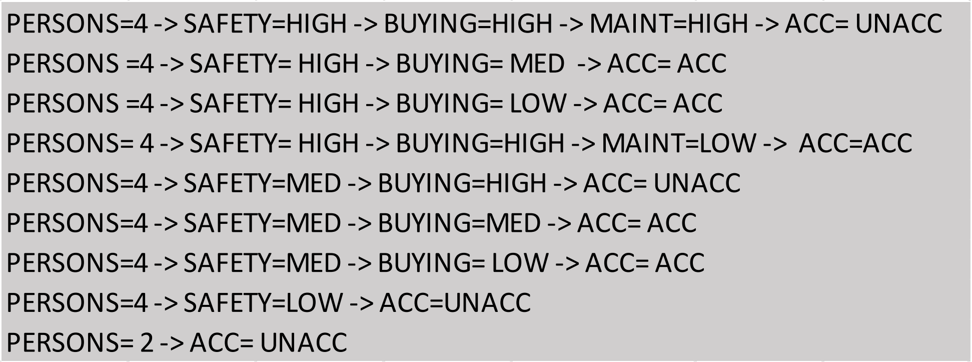
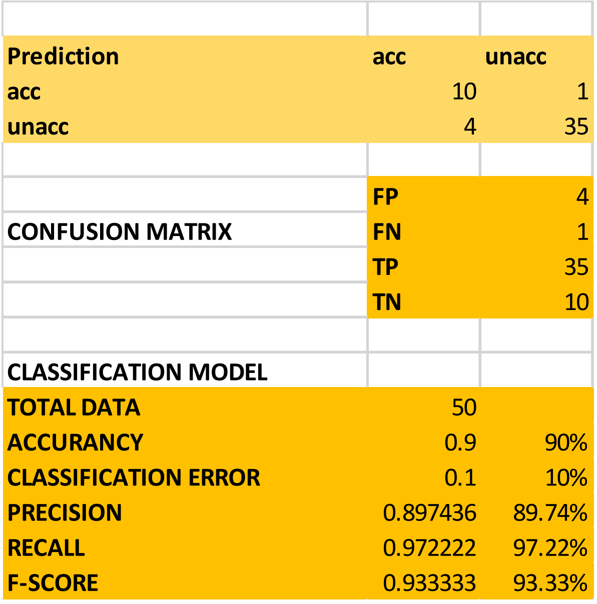

PERSONS=4 -> SAFETY=HIGH -> BUYING=HIGH -> MAINT=HIGH -> ACC= UNACC
PERSONS =4 -> SAFETY= HIGH -> BUYING= MED -> ACC= ACC
PERSONS =4 -> SAFETY= HIGH -> BUYING= LOW -> ACC= ACC
PERSONS= 4 -> SAFETY= HIGH -> BUYING=HIGH -> MAINT=LOW -> ACC=ACC
PERSONS=4 -> SAFETY=MED -> BUYING=HIGH -> ACC= UNACC
PERSONS=4 -> SAFETY=MED -> BUYING=MED -> ACC= ACC
PERSONS=4 -> SAFETY=MED -> BUYING= LOW -> ACC= ACC
PERSONS=4 -> SAFETY=LOW -> ACC=UNACC
PERSONS= 2 -> ACC= UNACC

Figure 11: Resulting classifiers from the DT



Prediction	**acc**	**unacc**
acc	10	1
unacc	4	35
	FP	4
CONFUSION MATRIX	**FN**	1
	TP	35
	TN	10
CLASSIFICATION MODEL		
TOTAL DATA	50	
ACCURANCY	0.9	90%
CLASSIFICATION ERROR	0.1	10%
PRECISION	0.897436	89.74%
RECALL	0.972222	97.22%
F-SCORE	0.933333	93.33%

Figure 12: Confusion Matrix and Classification Model

The model is tested against the test dataset. The confusion matrix is a breakdown of correct and incorrect predictions for each class, comparing predicted outcomes with actual values (Barros, De Carvalho and Freitas, 2015). This model achieves 95% accuracy, with a 10% classification error, indicating high predictive accuracy and high performance. The model demonstrates a high recall of 97% and a high precision of 89%, thus, signifying the model's effectiveness in capturing most of the actual positive (acceptable) instances, while also maintaining a high precision in its positive predictions. F-score considers the balance between these two metrics. An F-score of 93% signifies a strong balance, indicating the model's overall effective performance in classifying records. Discussions around the significance of FP and FN can guide decisions on optimizing recall/precision metrics based on specific priorities/consequences.

4. IMPLICATIONS

While the report centers on a hypothetical scenario regarding customer car acceptance within domains of market research and customer behaviour analysis, DTs showcase versatility in day-to-day applications. Real-life examples of classification span domains such as medical diagnosis, climate forecasting, credit approval, customer segmentation and fraud detection (Alraddadi, 2023). With medical diagnosis, DT classifiers play a role in breast cancer detection, ovarian cancer and the interpretation of heart sounds (Figure 13). Their most ground-breaking usage involves their applications in diagnosing the condition (severe, mild or moderate) of COVID-19 patients (Giotta *et al.*, 2022). This showcases DT's adaptability in addressing decision-making processes in many domains.

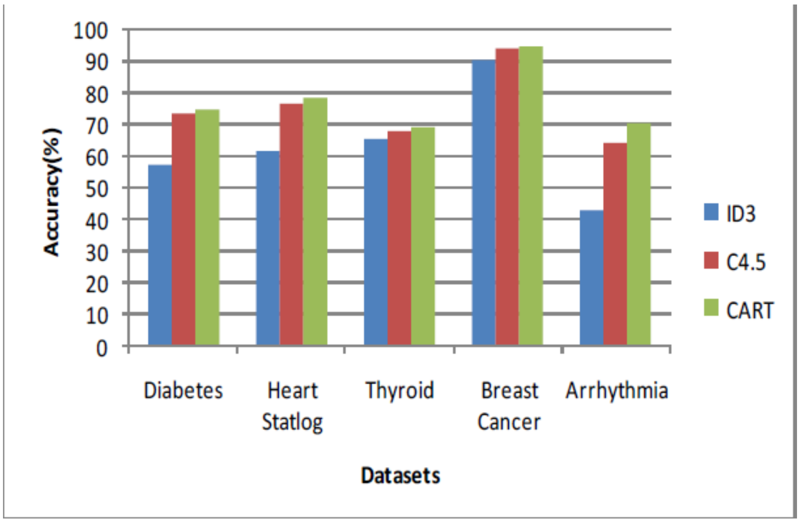


Figure 13: Use of DT algorithms to predict medical diagnosis (Lavanya and Rani, 2011)

DTs offer a systematic approach to navigating complex relationships between input and target variables by partitioning the original input variables into insightful subgroups. Their simplicity enhances accessibility for comprehension and interpretation. DTs also effectively handle missing data, treating it as a separate category, analysed along other groups categories, or constructing a DT model with the variable containing missing values as the target for prediction. These benefits of DTs are commonly exploited in medical diagnosis (Song and Lu, 2015).

However, DTs often face challenges such as overfitting with large complex data, underfitting with smaller datasets, and potential bias toward variables with numerous possible splits. The greedy nature of the DT construction, focusing on locally optimal decisions at each step, does not guarantee a global optimal structure. DT's predictive power is often poor compared to other classification techniques like random forest algorithms (RF/RFA) (Williams, 2011).

To boost the model's accuracy in this report, the usage of RFA is recommended. RFA, with its ensemble (collection) of unpruned DTs, addresses the issues associated with single DT instability. RFA employs multiple trees that use different variables and overfit the data in diverse ways. This 'randomness' enhances robustness against outliers, noise and overfitting. Unlike single DTs, small changes in the training dataset may minimally impact the final decisions of the resulting RF model (Williams, 2011). The independence and autonomous construction of each DT within RFs make it a promising technique worth exploring for machine learning for customer behaviour analysis beyond the applied method in this report.