Department of Information Engineering & Mathematics

Management Engineering

Business Intelligence Project Report

LEVENT ERGUL

095833

1.ABSTRACT

Cars are vital in everyday life. It plays an important role as it is a comfortable means of transportations. Every car has a distinct favour in terms of price, feature, safety and the level of luxury it provides. People tend to make clear choices when they decide to buy a car for themselves. They evaluate different cars on various parameters. Manufacturing and business are interested to know the popular features on which buyers make their choice as it can enhance their business value. Data mining algorithms can be employed in this respect. Various data mining algorithms perform differently.

This project aims to understand the decision making process people go through when buying a car and apply it to an entire group of models as a classifier from an acceptable car to unacceptable ones. It will also be examined how features affect the process. There are more than two different evaluation classes that are why it is a multi-class classification problem.

Accordingly, a categorical attribute was determined as the output varies according to the original data set. For this project, will be using two methods. The first one is the Decision Tree and the second one is Naïve Bayes. Also in this research work will compare these two algorithms in terms of the accuracy they offer.

2. INTRODUCTION & BACKGROUND

Cars offer diverse characteristics in terms of model and manufacturer preferences. Cost, safety and luxury are three imperative factors which are considered when buyers make their choice. These factors significantly contribute towards the reduction of accidents occurring. Some standard equipment is also vital to consider when buying cars. Which includes performance enhancers, conveniences and safety tools in cars. Safety as already mentioned is one of the imperative factors for car buying decision. Same is the case for convenience which has attributes such as maintenance, door and luggage boot. Cost deliberation is also crucial to make sure that car which is bought is worth what it has cost to the owner. Financial responsibility also comes with owning a car as it need to be maintained for convenience. This particular research work utilizes attribute “buying” for assessing acceptability of car cost in comparison to the other attributes it is offering such as doors, lug boot, person and safety.

Classifying a good car from a decent to a bad one are usually being done manually with the help of our friendly mechanics who tells us to buy this because of this or from the opinion of our family and friends who had previous experience with car troubles. It would have been nice to have a gadget that can scan car features and tell that it’s an A car or a D car. If there’s such a thing, then there should be no worries in achieving a particular car. In the present time, it is always the car salesman personality that encourages us to buy this car or not. We might or might not know it consciously but we are simply ignoring the factors that would help us financially, comfortably, and safely in a long run.

The hierarchical decision model, from which this dataset is derived, was first presented in M. Bohanec and V. Rajkovic: Knowledge acquisition and explanation for multi-attribute decision making. In 8th Intl Workshop on Expert Systems and their Applications, Avignon, France. Pages 59-78, 1988. Within machine-learning, this dataset was used for the evaluation of HINT (Hierarchy Induction Tool), which was proved to be able to completely reconstruct the original hierarchical model. This, together with a comparison with C4.5, is presented in B. Zupan, M. Bohanec, I. Bratko, J. Demsar: Machine learning by function decomposition. ICML-97, Nashville, TN. 1997 (to appear).

Table 1. The model evaluates cars according to the following concept structure

<table>
<tr><td colspan="8">1.CAR</td></tr>
<tr><td colspan="4">1.1.Price</td><td colspan="4">1.2.Techinal</td></tr>
<tr><td colspan="2">a)buying</td><td colspan="2">b)maint</td><td colspan="4">1.2.1.Comfort</td></tr>
<tr><td></td><td></td><td></td><td></td><td>a)doors</td><td>b)persons</td><td>c)lug_boot</td><td>d)safety</td></tr>
</table>

Input attributes are printed in lowercase. Besides the target concept (CAR), the model includes three intermediate concepts: PRICE, TECH, COMFORT. Every concept is in the original model related to its lower level descendants by a set of examples.

Price concepts show the overall price of the car. As you can see above table, it has two sub-branch, buying price and maintenance cost. Technical concepts show the technical characteristics of the car and also included comfort. Comforts show the how is comfort level and it has four sub-branch. These are doors, persons, lug_boot and safety.

The car evaluation dataset contains examples with the structural information removed, i.e., directly relates target feature (Car evaluation) to the six input features: buying, maint, doors, persons, lug_boot ad safety. These seven features will describe in the next section.

3. DATASET

As said before, the dataset used in this study which is a collection of the records on specific attributes on cars donated by Marco Bohanec in 1997 was obtained from the UCI dataset repository. The car evaluation dataset as described in the UCI dataset repository was derived from simple hierarchical decision, and is categorized descriptively below.

Table 2. Car evaluation dataset

Data Set Characteristic:	Multivariate
Attribute Characteristic:	Categorical
Associated Tasks:	Classification
Number of Instances:	1728
Number of Attributes:	6
Missing Values:	0

There are seven features in the dataset. The relationship between them is shown in Table 1. Now, can describe data features to what they refer to normally. In this way, it will be easier to understand the dataset.

Table 3. Description of the features

Feature	Description
Car	This is target feature and it shows the acceptability situation of cars. It has a four variables and these are: Acceptable, Good, Very Good and Unacceptable.
Buying	It shows the buying price of cars. It has four variables and these are: Very High, High, Medium and Low.
Maint	It shows the maintenance cost of cars. It has four variables and these are: Very High, High, Medium and Low.
Doors	It shows how many doors the cars have. It has four variables and these are: 2,3,4, more than 4.
Person	It shows the car capacity in terms of persons to carry. It has three variables and these are: 2,4 and more than 4.
Lug_boot	It shows the size of car's luggage boot. It has three variables and these are: Small, Medium and Big.
Safety	It shows the estimated safety of the car. It has three variables and these are: Low, Medium and High.

The class features in the Car evaluation dataset are:

- Acceptable: This is denoted as "acc"
- Good: This is denoted as "good"
- Unacceptable: This is denoted as "unacc"
- Very Good: This is denoted as "vgood"

Also, input features variables names different in the dataset. That is why features variables names in the dataset are shown below:

Table 4. Features variables name in dataset

Feature	Variables Name	Denoted	Feature	Variables Name	Denoted
Buying			**Person**		
	High	**high**		**2**	**2**
	Low	**low**		**4**	**4**
	Medium	**med**		**More than 4**	**more**
	Very High	**vhigh**			
Maint			**LugBoot**		
	High	**high**		**Big**	**big**
	Low	**low**		**Medium**	**med**
	Medium	**med**		**Small**	**small**
	Very High	**vhigh**			
Doors			**Safety**		
	2	**2**		**High**	**high**
	3	**3**		**Low**	**low**
	4	**4**		**Medium**	**med**
	More than 4	**5more**			

Features description was done on the dataset to better understanding. As mention before in Table 1 the dataset features are categorical and according to the coding process in R, they can be converted to numeric values for the machine learning model. Also, in the next steps, will be looking at the target and input features frequencies and if is there any correlation between the input features.

4. METHODS

This step will be explained and gives details about the which methods used in the project like algorithms. Before that, need to analyse the dataset more like frequencies. For this reason, preprocessing and data preparation phase and preliminary statistical analysis have been performed.

4.1. DATA VISUALIZATION

A standard data analysis was done on the dataset to identify some patterns in the data and also present the table based on the attribute range and their frequencies. Firstly, the output from the data analysis show in above Table 5 and Figure 1 describes the distribution of the four class attributes in the dataset.

Table 5. Frequency of class output from the dataset

Class	Frequency	Relative Frequency in %
Acc	384	22,23508975
Good	69	3,995367689
Unacc	1210	70,01579038
Vgood	65	3,762752181
Total	**1728**	**100**

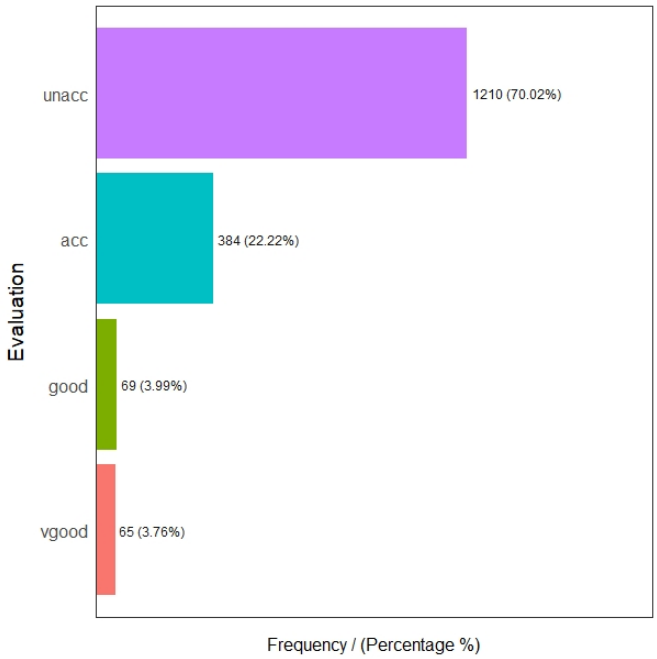

Figure 1. Frequency of class output from the dataset

Table 5 and Figure 1 show the frequency of the class output which is the final outcome from the dataset. It shows that out of the total 1728 cars in the dataset, 384(22.22%) were acceptable, 69 (3.99%) were good, 1210 (70.02 %) were unacceptable, and 65 cars (3.76 %) were very good. From the above, can be concluded that more than half of the cars evaluated were not acceptable.

After the target variable can be more focused on input variables. As you can see their result in Figure 2. Firstly, buying attributes have 4 variables, 432 (25%) were high, 432 (25%) were low, 432 (25. %) were med, and 432 (25%) were vhigh and the maint attribute has exactly the same distribution as buying. Doors attribute have four different variables, 432 (25%) were 2, 432 (25%) were 3, 432 (25%) were 4, and 432(25%) were 5 or more than 5. Person features have three variables, 576 (33.35%) were 2, 576 (33.35%) were 4, 576 (33.33%) were more than 4. LugBoot features have three variables, 576 (33.33 %) were big, 576 (33.33%) were med, 575 (33.33%) were small. Safety has three variables, 576 (33.33%) were high, 576 (33.33%) were low, 575 (33.33%) were med. All variables in the dataset are of characteristic type

As you can see in Figure 2, according to the number of feature variables, all of them distributed equally. This is will be useful when we are looking at the features versus evaluation (output feature). We can comment on which variable has more effect on the evaluation process.

	Buying	frequency	percentage	cumulative_perc
1	high	432	25	25
2	low	432	25	50
3	med	432	25	75
4	vhigh	432	25	100

	Maint	frequency	percentage	cumulative_perc
1	high	432	25	25
2	low	432	25	50
3	med	432	25	75
4	vhigh	432	25	100

	Doors	frequency	percentage	cumulative_perc
1	2	432	25	25
2	3	432	25	50
3	4	432	25	75
4	5more	432	25	100

	Person	frequency	percentage	cumulative_perc
1	2	576	33.33	33.33
2	4	576	33.33	66.66
3	more	576	33.33	100.00

	LugBoot	frequency	percentage	cumulative_perc
1	big	576	33.33	33.33
2	med	576	33.33	66.66
3	small	576	33.33	100.00

	Safety	frequency	percentage	cumulative_perc
1	high	576	33.33	33.33
2	low	576	33.33	66.66
3	med	576	33.33	100.00

Figure 2. Frequency of input features

4.2. DATA EXPLORATION

After the data visualization, this step will be shown more clearly input features and target variable relation. First of all, will look at each input feature number of acceptability numbers below.

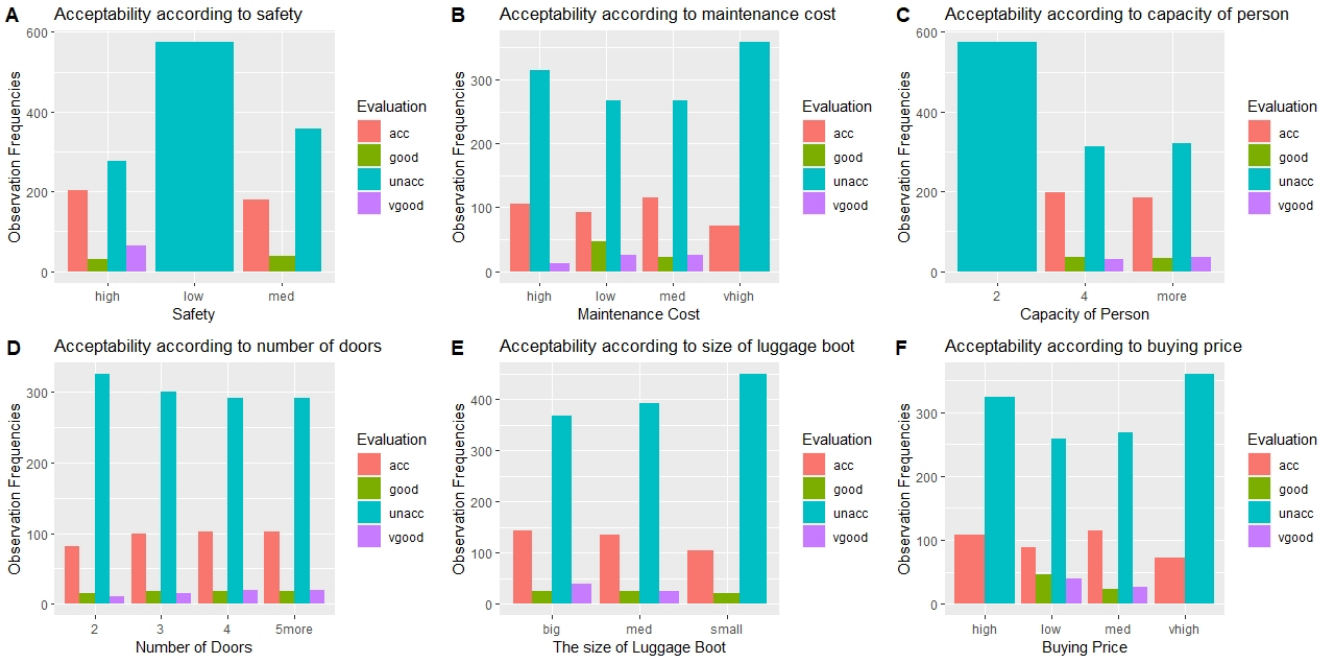

Figure 3. Acceptability according to the input features

Figure 3 tell us more about the dataset. As we knew before in the dataset there are 70% of the unacceptable car evaluation. This means in the feature's variable unacceptable situation is almost higher than others and we can see it above figure. Some conditions are more important for the target variable; we can see clearly. For example, Figure 3-A shows if the car safety low then it is unacceptable for sure or Figure 3-C shows if the car capacity's 2 this is also unacceptable. We do not have the same examples for other features or other evaluation variables. On the other conditions are changeable.

As you can see on the right side Figure 4 shows the correlation between the features. There is no correlation between the input feature. They just affect the output. As we see in figure 3, person and safety are having much effect on it. We will be looking this which input features are more important for target value in the next steps and analyze them

Figure 4. Correlation between the features

4.3. CLASSIFICATION

Classification is the process of predicting the class of given data points. Classes are sometimes called targets, labels or categories. Classification predictive modelling is the task of approximating a mapping function (f) from input variables (X) to discrete output variables (y). Our classification model is a multiclass classification problem. For this prediction part, we have four classes (Acceptable, very good, good and unacceptable) which are called target features. However, before the start of this part, need to be prepared features and target values. In these steps, all variables must be numerical and they need to the similar range. If it is not a similar range, cannot be reached well result and its affect the model in a negative way.

After prepared datasets, in both the classification approaches treated in this work, there is the need of distinguishing the original dataset into two different subsets, called train set and test set. The train-test split procedure is used to estimate the performance of machine learning algorithms (which are Naïve Bayesian, Decision Tree and Random Forest.) when they are used to make predictions on data not used to train the model. It is a fast and easy procedure to perform, the results of which allow us to compare the performance of these algorithms for our predictive modelling problem.

Table 6. Dataset split for model creation

Training & Testing %Split	
90%	10%
70%	30%
50%	50%
10 Folds	

The data set used for training is mainly a portion from the dataset from which the classifying algorithm used learns the class/result of the model created from each model, and the four splits used in this study are shown in table 6. The learning method is based on the attributes or features of the dataset in comparison to the result/class. And finally, the output is a model used to compare against the other half of the dataset, which is the testing data.

4.3.1. NAÏVE BAYES

Naïve Bayes is a basic strategy for building classifiers models that dole out class names to issue examples, spoken to as vectors of highlight esteems, where the class marks are drawn from some limited set. Guileless Bayes classifiers are immediate classifiers that are known for being direct yet particularly beneficial. The probabilistic model of Naive Bayes classifiers relies on Bayes' speculation and the elucidating word Naive begins from the supposition that the segments in a dataset are ordinarily free. Before long, the flexibility supposition is consistently harmed, however, Naive Bayes classifiers still tend to perform uncommonly well under this farfetched doubt. Especially for little example sizes, Naive Bayes classifiers can beat the all the more extreme choice.

Naïve Bayes classifier is a straightforward probabilistic classifier in view of applying Bayes hypothesis (from Bayesian insights) with solid (innocent) freedom suspicions. An innocent Bayes classifier expects that the nearness (or nonattendance) of a specific component of a class is random to the nearness (or non-appearance) of some other element. Innocent Bayes classifiers can deal with a subjective number of autonomous factors, regardless of whether consistent or clear cut. Given an arrangement of factors, $X= \{x_1, x_2, x_3 \ldots x_n\}$. We need to develop the back likelihood for the occasion C_k among an arrangement of conceivable results $C = \{c_1, c_2, c_3 \ldots c_k\}$. In a more commonplace dialect, X is the indicators and C is the arrangement of straight out levels display in the reliant variable utilizing Bayes' rule;

$$P(C|X) = P(X|C)\, P(C)/P(X)$$

$$P(C|X) = P(X1|C) \times P(X2|C) \times \ldots\ldots\ldots\ldots \times P(X_n|C) \times P(C)$$

Where to show P (C|X) is Posterior probability and P (X|C) is a likelihood and P(C) class prior probability and P(X) predictor prior probability.

4.3.2. DECISION TREE

A decision tree is a tree-like graph or model. It is more similar to the rearranged tree since it has its root at the best and it develops downwards. This representation of the data has the advantage compared with other approaches of being important and easy to interpret. The objective is to make a characterization show that predicts the value of an objective attribute (often called class or label) Based on several input attributes of the Example Set. Every inside node of the tree compares to one of the input attributes. The quantity of edges of a nominal inside node is equivalent to the number of possible values of the corresponding input attribute. Outgoing edges of numerical attributes are labelled with disjoint ranges. Each leaf node represents a value of the estimation attribute given the estimation of the input attributes represented by the way from the root to the leaf. Decision Trees are produced by recursive partitioning.

5. RESULT

This section presents the results of the experimentation setup. The process is as follows; it is a supervised learning method. We have trained the model utilizing attributes inclusive of class attributes. As it is a supervised model, the model is built basing on the class values in correspondence to the values of attributes individually. R is used for simulation purposes. The results achieved by various experimentation setups in Decision Tree and Naïve Bayesian. Table 6 show the percentage splits employed which are; 90:10, 70:30, 50:50 and 10-folds cross-validation.

1st Case: Data split is 90% Training and 10% Testing

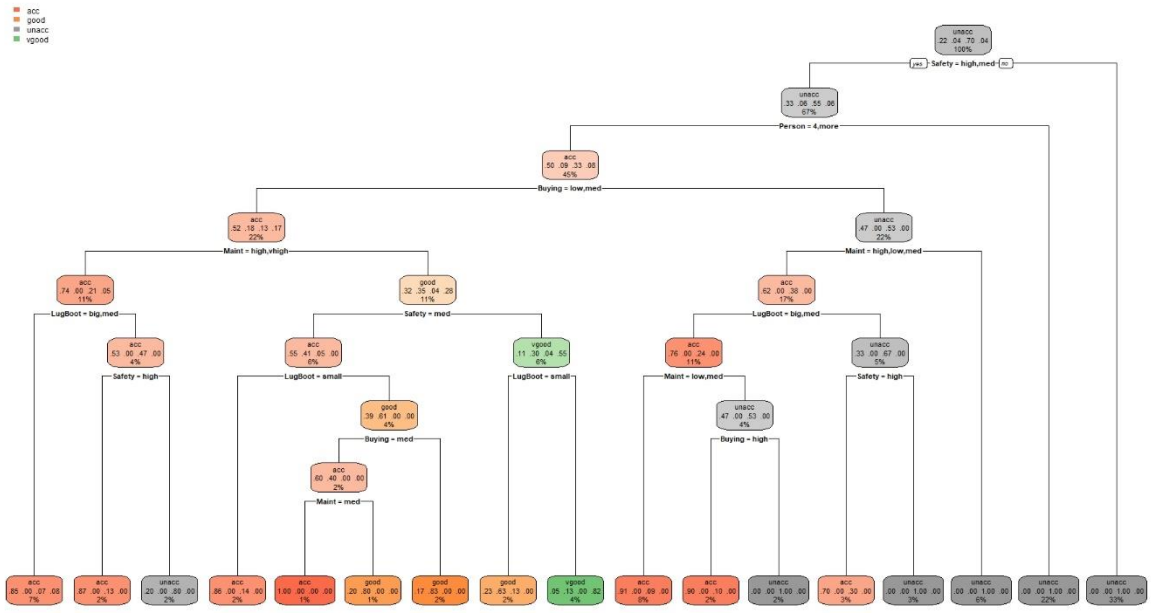

Figure 5. Decision tree with 90% train and 10% test dataset splitting criterion

```
[1] "The Decision Trees approach is more accurate than the Naive Bayes one"

-------------- --------
 **acc_tree**   0.9357

  **acc_nb**    0.8596
-------------- --------
```

Figure 6. Classification accuracy of Decision Tree and Naïve Bayes

2nd Case: Data split is 70% Training and 30% Testing

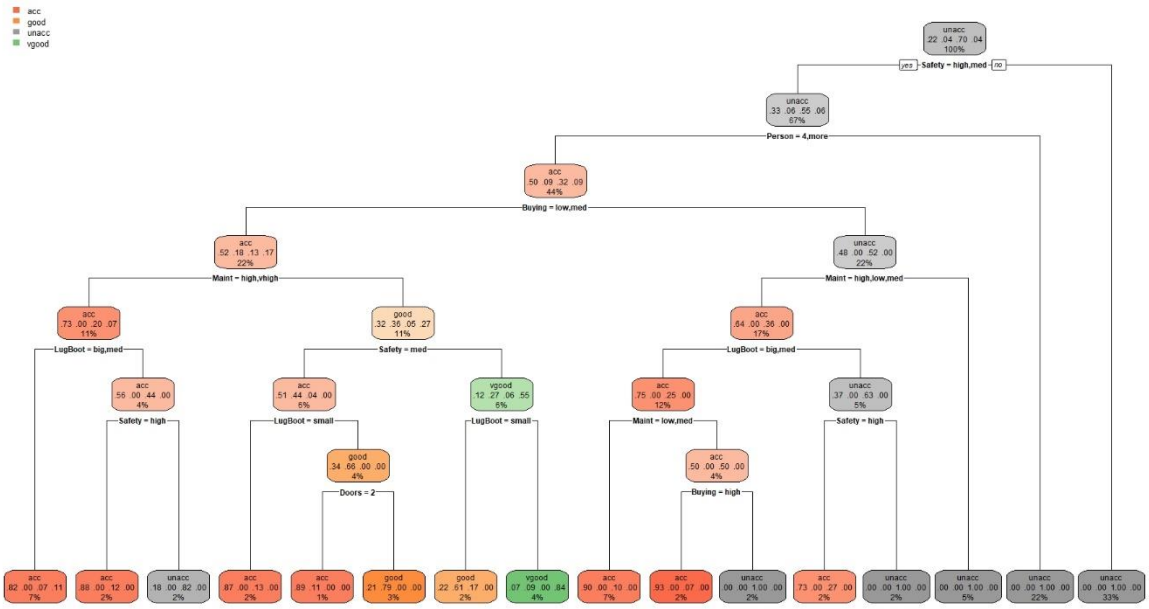

Figure 7. Decision tree with 70% train and 30% test dataset splitting criterion

```
[1] "The Decision Trees approach is more accurate than the Naive Bayes one"

-------------- --------
 **acc_tree**   0.9284

  **acc_nb**    0.853
-------------- --------
```

Figure 8. Classification accuracy of Decision Tree and Naïve Bayes

3rd Case: Data split is 50% Training and 50% Testing

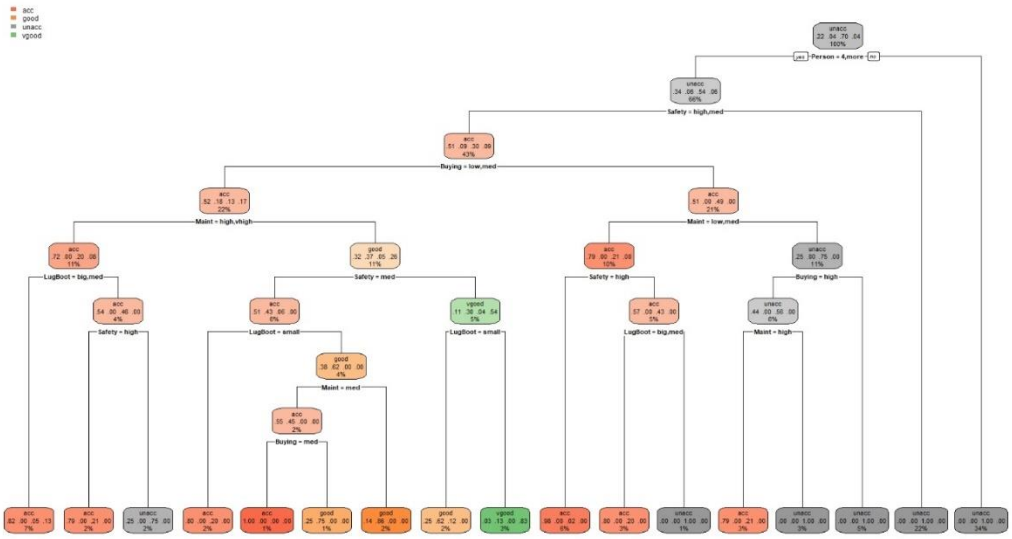

Figure 9. Decision tree with 50% train and 50% test dataset splitting criterion

```
[1] "The Decision Trees approach is more accurate than the Naive Bayes one"

-------------- --------
 **acc_tree**   0.9258

  **acc_nb**    0.8355
-------------- --------
```

Figure 10. Classification accuracy of Decision Tree and Naïve Bayes

4th Case: 10- Fold Cross Validation

```
Cross-Validated (10 fold, repeated 1 times) Confusion Matrix

(entries are percentual average cell counts across resamples)

          Reference
Prediction  acc good unacc vgood
     acc   18.9  2.3   4.6   0.6
     good   1.0  1.1   0.1   0.5
     unacc  1.6  0.0  65.3   0.0
     vgood  0.8  0.6   0.0   2.6

 Accuracy (average) : 0.8785
```

Figure 11. Confusion Matrix of Decision Tree

```
Cross-Validated (10 fold) Confusion Matrix

(entries are percentual average cell counts across resamples)

          Reference
Prediction  acc good unacc vgood
     acc   15.6  2.7   2.5   1.6
     good   0.6  1.2   0.1   0.1
     unacc  6.1  0.0  67.4   0.0
     vgood  0.0  0.2   0.0   2.1

 Accuracy (average) : 0.8617
```

Figure 12. Confusion Matrix of Naïve Bayes

Also in this model, decision trees have higher accuracy than the Naïve Bayes.

The above cases show us, our model accuracies are good enough and results are close to each other. Figure 13 show us the feature importance Figure 14 show us the confusion matrix of the random forest model. For this step, the random forest model used, because it is an easy way to find it.

```
             acc       good      unacc     vgood MeanDecreaseAccuracy MeanDecreaseGini
Buying   92.07572 49.9700560  68.370742 48.685437           106.261091         96.24357
Maint    80.83977 47.6670609  61.540460 31.050618            92.688951         95.27350
Doors    -1.23311 -0.8336098   4.216158  1.178601             1.574718         27.92901
Person  112.91356 33.9253818 147.272570 37.076386           156.582525        178.06977
LugBoot  43.94621 31.3593827  41.335424 37.383764            63.714521         50.78449
Safety  126.40356 52.4380372 153.152246 62.088430           174.485782        215.77449
```

Figure 13. Features importance

```
Call:
 randomForest(x = x, y = y, importance = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 2

        OOB estimate of  error rate: 2.31%
Confusion matrix:
      acc good unacc vgood class.error
acc   378    4     1     1  0.01562500
good    3   62     0     4  0.10144928
unacc  18    1  1191     0  0.01570248
vgood   8    0     0    57  0.12307692
```

Figure 14. Confusion Matrix of Random Forest

First of all, as we mentioned above our model's accuracies are close to each one for different cases and all of them gives good results. You can see the result in Table 6. Also, decision trees help us better understand the process of choosing cars. These trees are so similar each one in the 1st case and the 2nd case. The 3rd one is a little different from them. For the first two cases, the root node is Safety, and a second node is a person. The third case is vice versa, a root node is a person and the second node is safety.

Table 6. Dataset split and Accuracy for models

Percentage Split		Decision Tree	Naïve Bayes
Training %	Testing %	Accuracy %	Accuracy %
90%	10%	93.57%	85.96%
70%	30%	92.84%	85.3%
50%	50%	92.58%	83.55%
10	Folds	87.85%	86.17%

These cases are proved by looking at the feature importance results (Figure 13). You can see the features of safety and person are the highest values for this model in the Mean Decrease Accuracy part. This means these are the most important features of this model. It was expected that such a result could come out after seeing the Figure 4 Correlation between the features. This step proved the expectation.

In addition to these, which feature is more important can be seen in Figure 13. The importance function shows the importance level of the variables. The most important variable for the unacceptable class: **Safety.** Most important variable for acceptable class: **Safety.** The most important variable for the good class: **Safety.** The most important variable for the very good class: **Safety.** When looking at the overall model, the three most important variables are Safety, Carrying capacity and Sales price. When looking at the confusion matrix of Random Forest (Figure 14), the most accurate prediction is made in the "unacceptable class, and acceptable class" while the "very good" class has the highest error rate.

6.REFERENCES

[1] Marko Bohanec, Vladislav Rajkovic. Knowledge acquisition and explanation for multi-attribute decision making

[2] Bohanec, M., Bratko, I., Rajkoviˇc, V.: AN EXPERT SYSTEM FOR DECISION MAKING, In Sol, H.G. (ed.): Processes and Tools for Decision Support, North-Holland, 1983.

[3] Blaz Zupan, Marko Bohanec, Ivan Bratko, Janez Demsar. To appear in Proc. ICML-97. Machine Learning by Function Decomposition

[4] Dataset is available at these links: https://archive.ics.uci.edu/ml/datasets/car+evaluation and https://www.kaggle.com/elikplim/car-evaluation-data-set

[5] Useful websites which help this project on the R part: https://rpubs.com/ , https://stackoverflow.com/