Index

1. Project Objective

The requirement of the project is to use Bayes classification rule to classify a dataset, which are discrete-valued data instances. The software used to analyze the data for the prediction on classification is Weka Software. The data instances will be tested on two algorithms namely Naïve Bayes (NB) and Tree Augmented Naïve Bayes (TAN). The prediction of both the algorithms on the data instances will be studied in comparison based on their results on classification accuracy, per class classification accuracy and confusion matrix. The algorithms will be further trained and tested on different data instances and accuracy deviation will be studied. The selection of algorithm for given dataset will be based on their classification results.

2. Data Preprocessing

2.1 Data Details

Sr. No	Name	Description
1.	car.names	Describes the data.
2.	car.data	Contains the data.
3.	car.c45-names	Describes the class and feature values.

Table 2.1 Data Description

Total number of data instances	1728.
Number of attributes	06
Class Values	unacc (1210), acc (384), good (69), vgood(65).
Attribute Values	buying : v-high, high, med, low maint : v-high, high, med, low doors : 2, 3, 4, 5-more persons : 2, 4, more lug_boot : small, med, big safety : low, med, high
Missing Attribute Values	00

Table 2.2 Data Details

2.2 Data Conversion

Since the data was provided in .data format it was converted into Comma Separated values (CSV) and later converted to Attribute-Relation File Format (ARFF) format in Weka Software.

2.3 Data Resampling

The data provided needed to be resampled into training and testing data instances to train and test for Naïve Bayes and Tree Augmented Naïve Bayes. The table below provides description of division of data instances into training and testing using Weka Software Filters.

DATA	TOTAL INSTANCES	INSTANCES	%
Training Data	1728	1728	100
Training Data	1728	1500	86.80555556
Training Data	1728	500	28.93518519
Training Data	1728	750	43.40277778
Training Data	1728	1250	72.33796296
Testing Data	1728	228	13.19444444
Testing Data	1728	200	11.57407407

Table 2.3 Data Resample into Training and Test Set

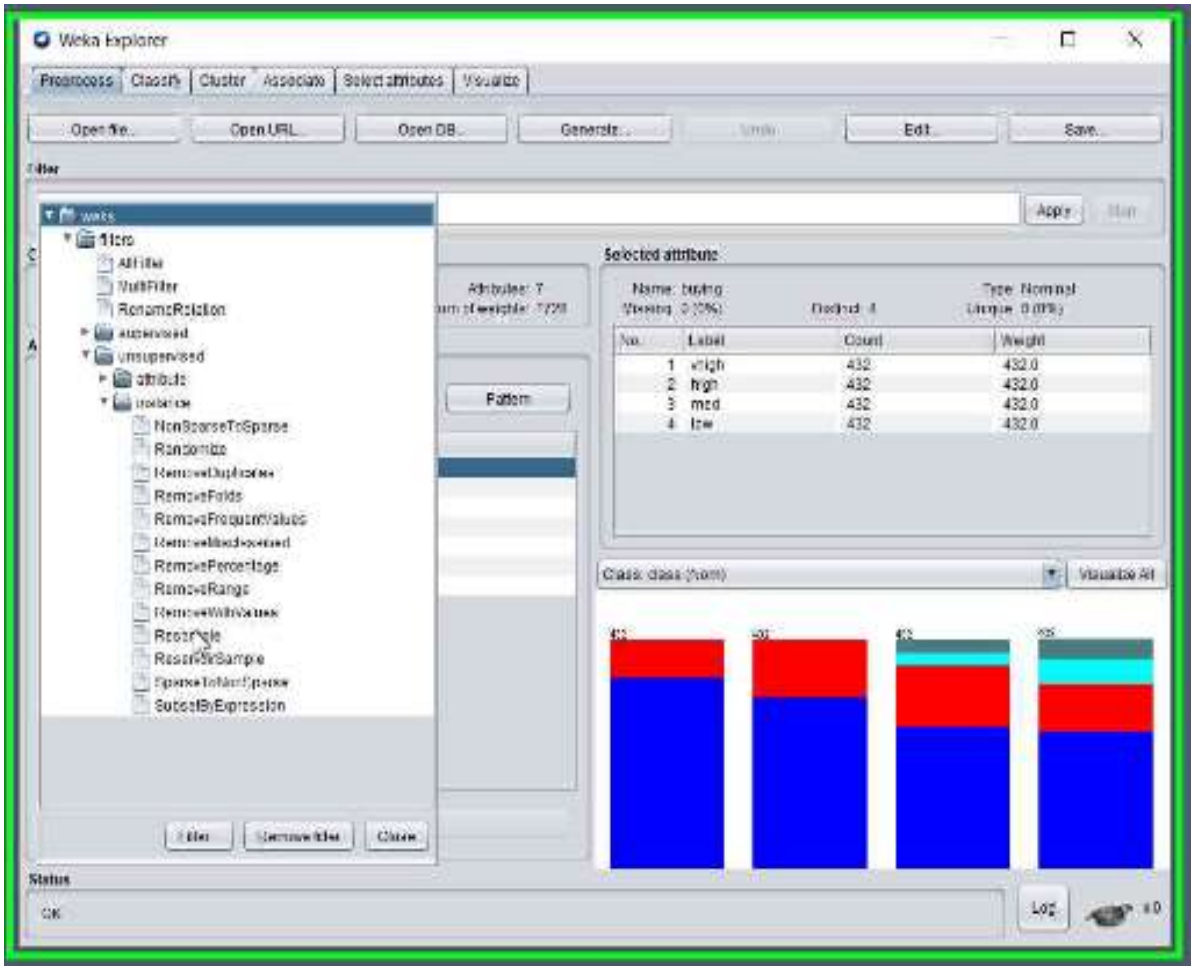

Figure 2.1 Data Resample Option in Weka

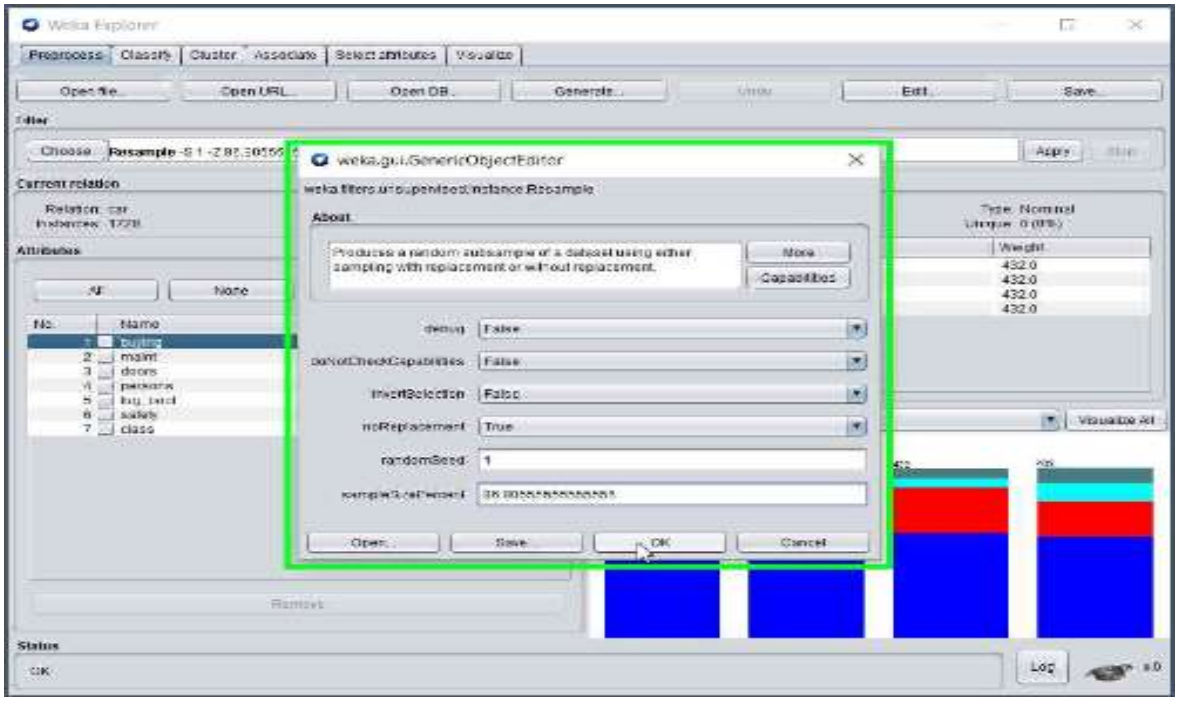

Figure 2.2 Data Resample Percentage

3. Classification Prediction

3.1 Naïve Bayes (NB)

The data instance considered is 1500 for training set and 228 as testing.

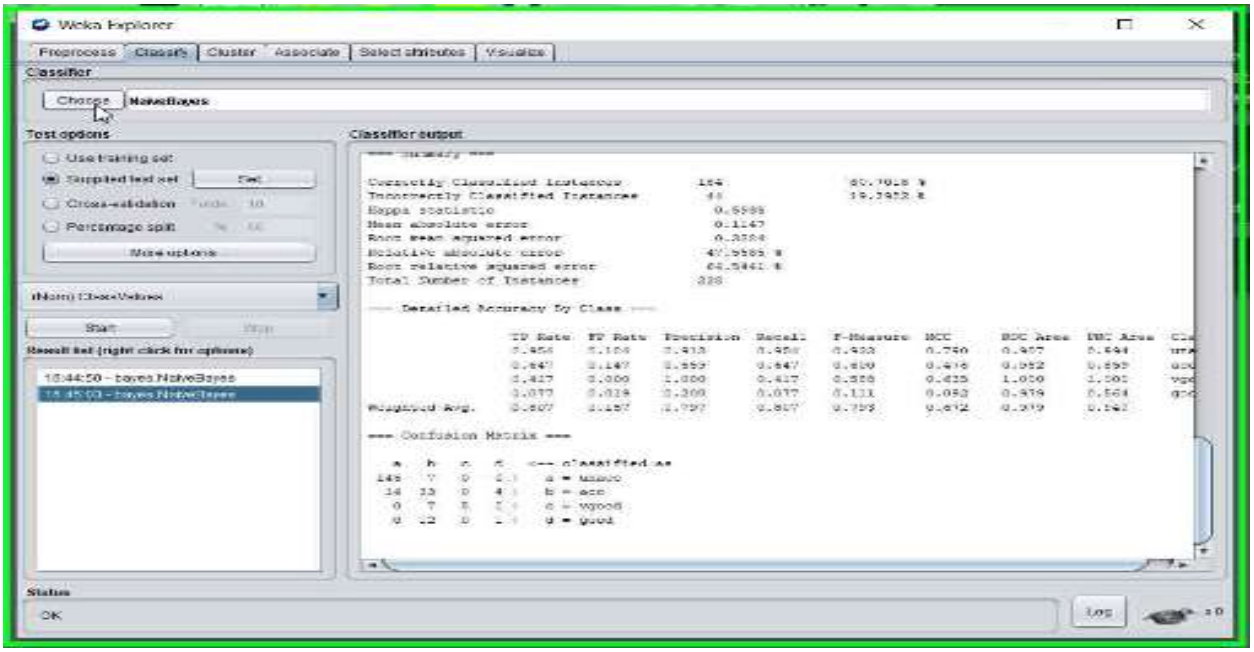

Figure 3.1 Weka Output for Naïve Bayes – Trained in 1500 data instance

3.1.1 Classification Accuracy

```
Correctly Classified Instances         184               80.7018 %
Incorrectly Classified Instances        44               19.2982 %
```

3.1.2 Per Class Classification Accuracy

```
=== Detailed Accuracy By Class ===
```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.954	0.184	0.912	0.954	0.932	0.790	0.987	0.994	unacc
	0.647	0.147	0.559	0.647	0.600	0.476	0.952	0.859	acc
	0.417	0.000	1.000	0.417	0.588	0.635	1.000	1.000	vgood
	0.077	0.019	0.200	0.077	0.111	0.092	0.979	0.564	good
Weighted Avg.	0.807	0.157	0.797	0.807	0.793	0.672	0.979	0.940	

3.1.3 Confusion Matrix

```
=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 145   7   0   0 |   a = unacc
  14  33   0   4 |   b = acc
   0   7   5   0 |   c = vgood
   0  12   0   1 |   d = good
```

3.2 Tree Augmented Naïve Bayes (TAN)

Similarly, for TAN same training and test data instances are taken as NB.

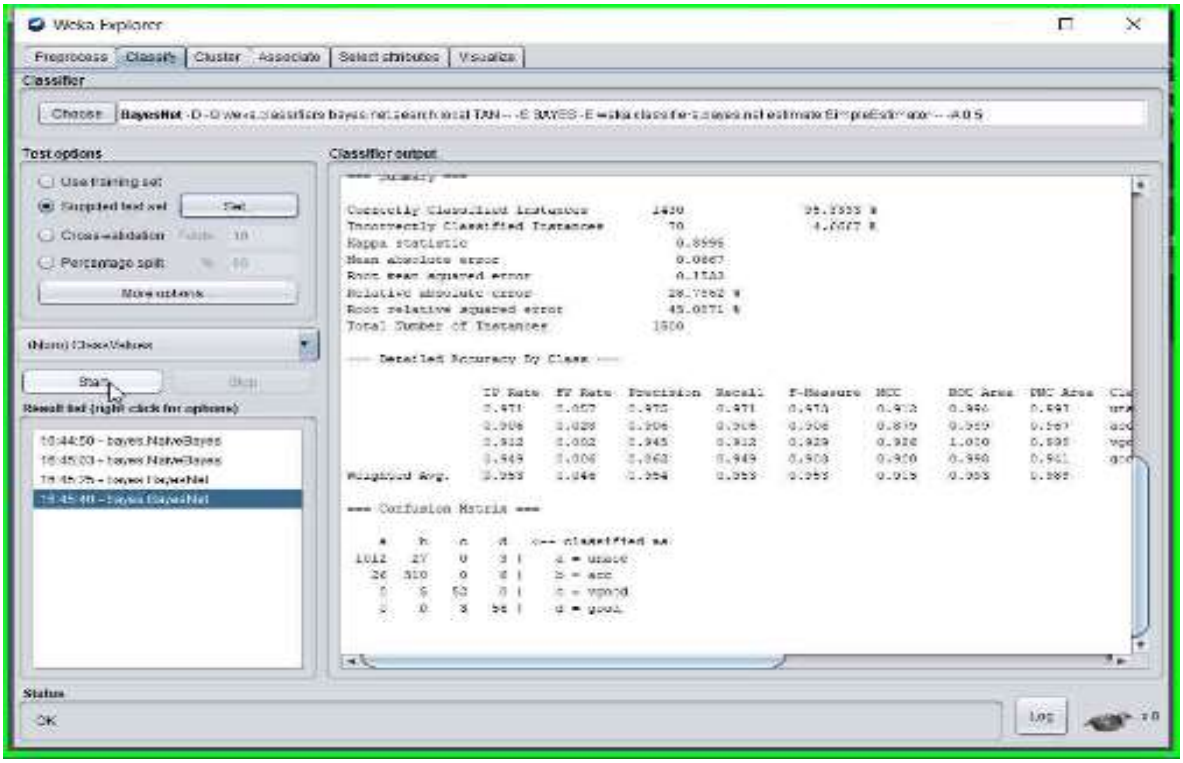

Figure 3.2 Weka Output for TAN – Trained in 1500 data instance

3.2.1 Classification Accuracy

```
Correctly Classified Instances         221               96.9298 %
Incorrectly Classified Instances         7                3.0702 %
```

3.2.2 Per Class Classification Accuracy

```
=== Detailed Accuracy By Class ===
```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.980	0.053	0.974	0.980	0.977	0.931	0.996	0.998	unacc
	0.922	0.017	0.940	0.922	0.931	0.911	0.991	0.972	acc
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	vgood
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	good
Weighted Avg.	0.969	0.039	0.969	0.969	0.969	0.934	0.995	0.992	

3.2.3 Confusion Matrix

```
=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 149   3   0   0 |   a = unacc
   4  47   0   0 |   b = acc
   0   0  12   0 |   c = vgood
   0   0   0  13 |   d = good
```

3.3 ROC Curve

For Analysis we are using ROC to demonstrate that TAN has better curve.

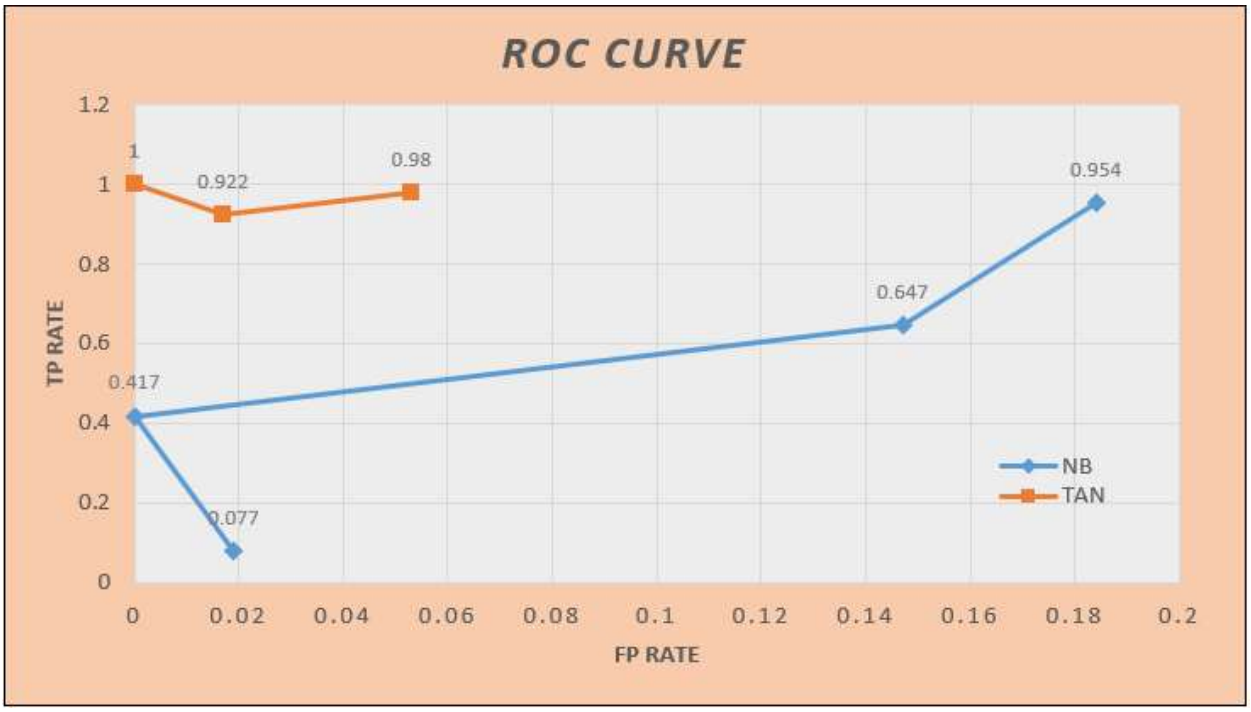

Figure 3.3 ROC Curve – NB and TAN

3.3 Comparative Analysis

Data Instances	Naïve Bayes (NB)		Tree Augmented Naïve Bayes (TAN)	
	CCI (%)	ICI (%)	CCI (%)	ICI (%)
500	76	24	89.5	10.5
750	79.5	20.5	93.5	6.5
1250	81	19	97	3

CCI - Correctly Classified Instances
ICI - Incorrectly Classified Instances

Testing - Last 200 Data Instances

Table 3.1 Accuracy for different numbers of data instances

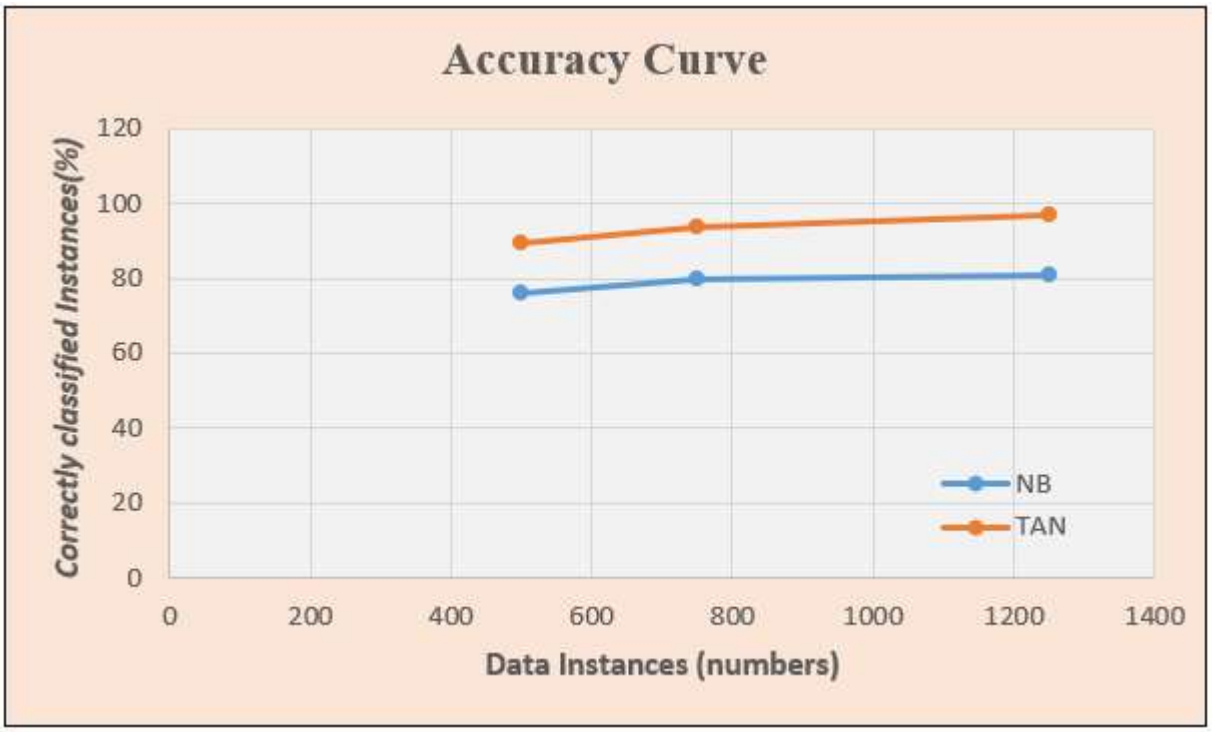

Figure 3.4 Accuracy curve of NB and TAN at different data instances

We can clearly conclude given from the Figure 3.4 that the accuracy for both NB and TAN has increased when the training data instances has increased.

Weka Output – NB – 500 Data Instance

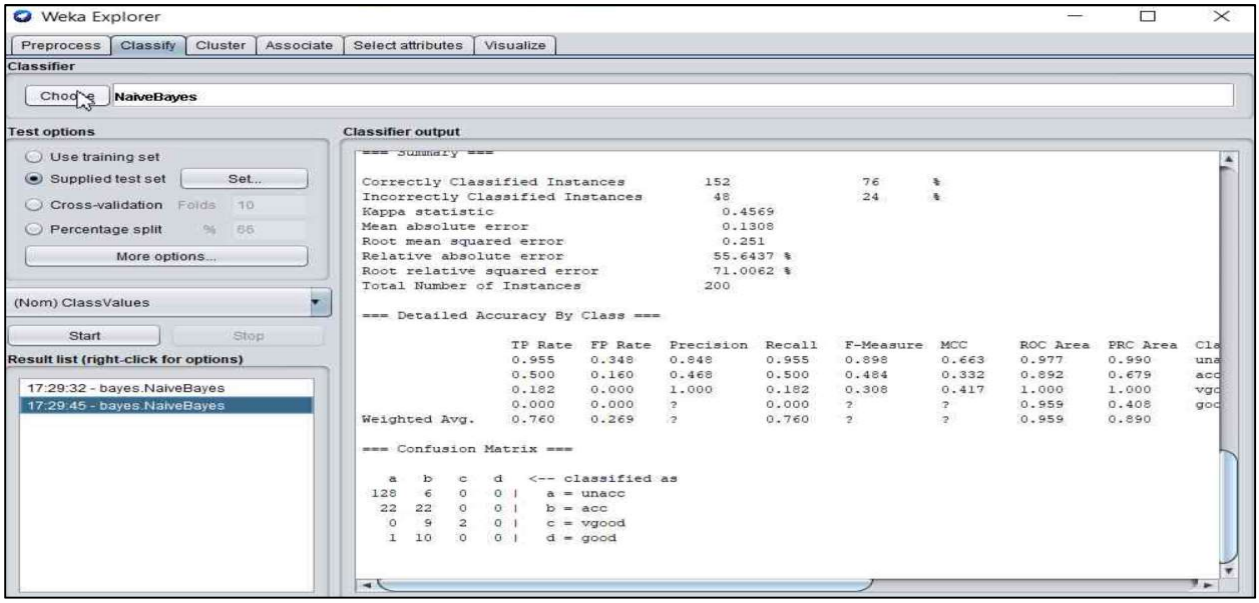

Figure 3.5 Weka NB output trained in 500 data instance

Weka Output – TAN – 500 Data Instance

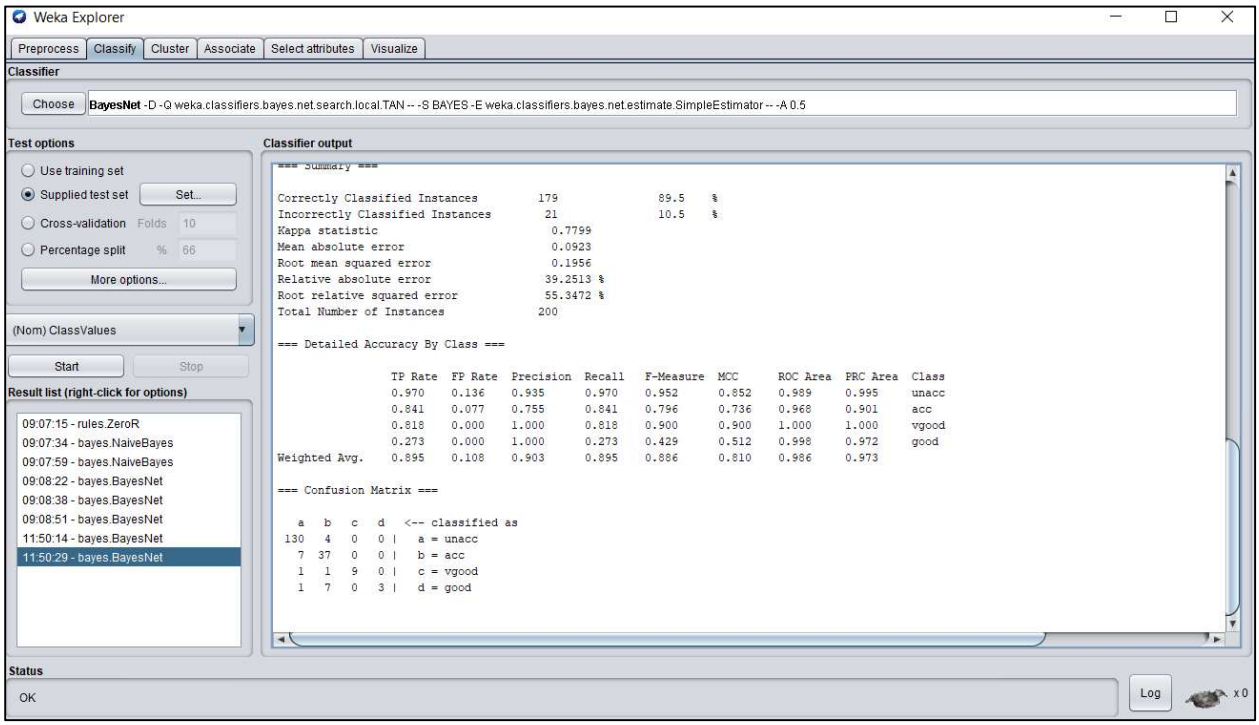

Figure 3.6 Weka TAN output trained in 500 data instance

Weka Output – NB – 750 Data Instance

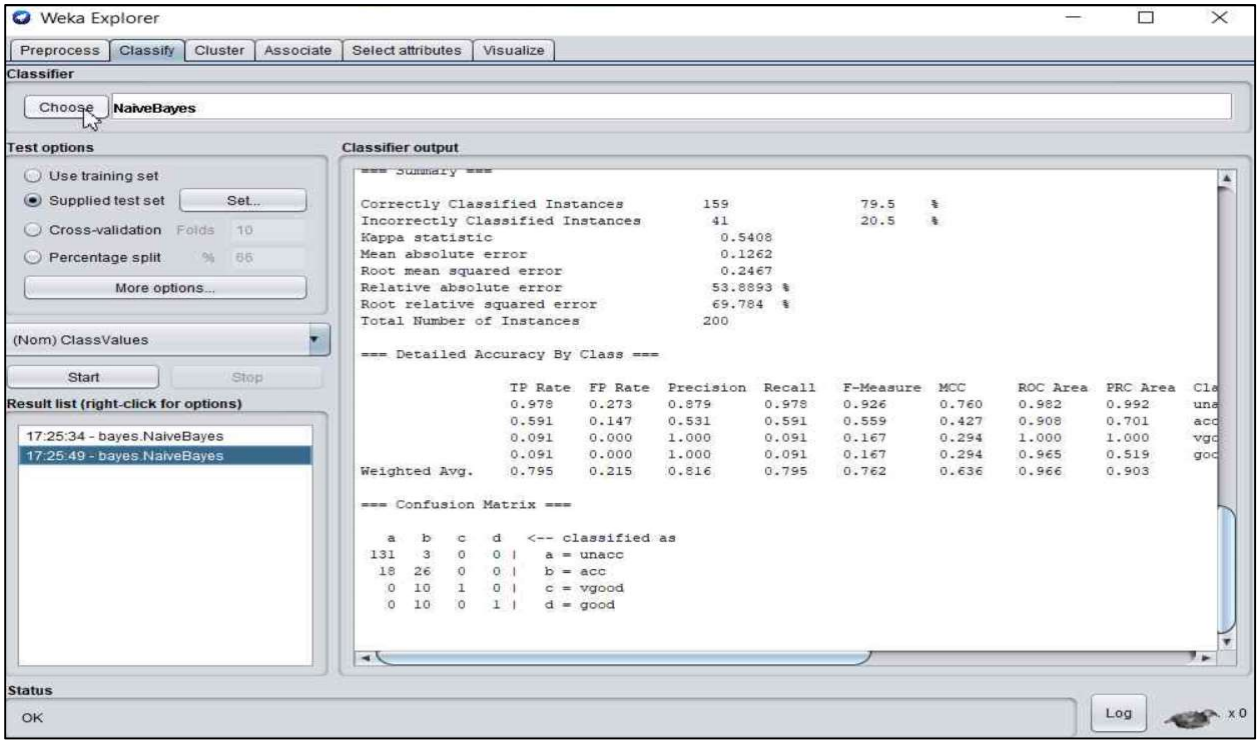

Figure 3.7 Weka NB output trained in 750 data instance

Weka Output – TAN – 750 Data Instance

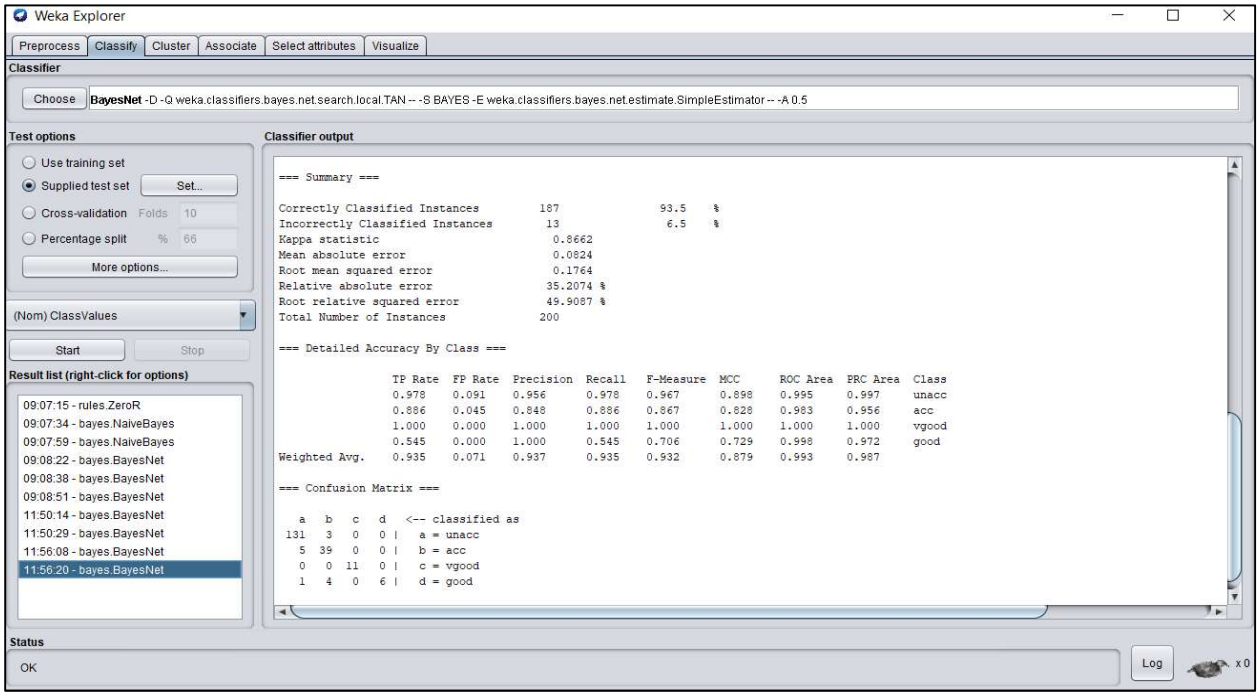

Figure 3.8 Weka TAN output trained in 750 data instance

Weka Output – NB – 1250 Data Instance

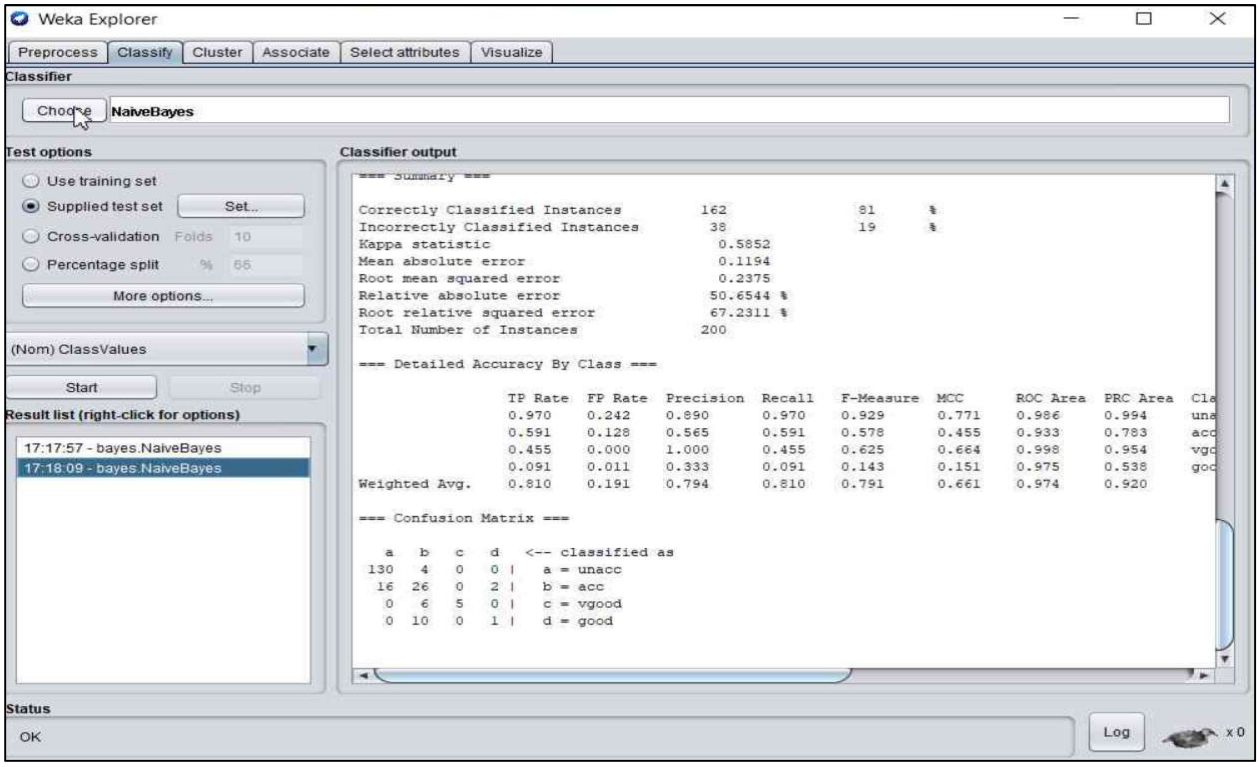

Figure 3.9 Weka NB output trained in 1250 data instance

Weka Output – TAN – 1250 Data Instance

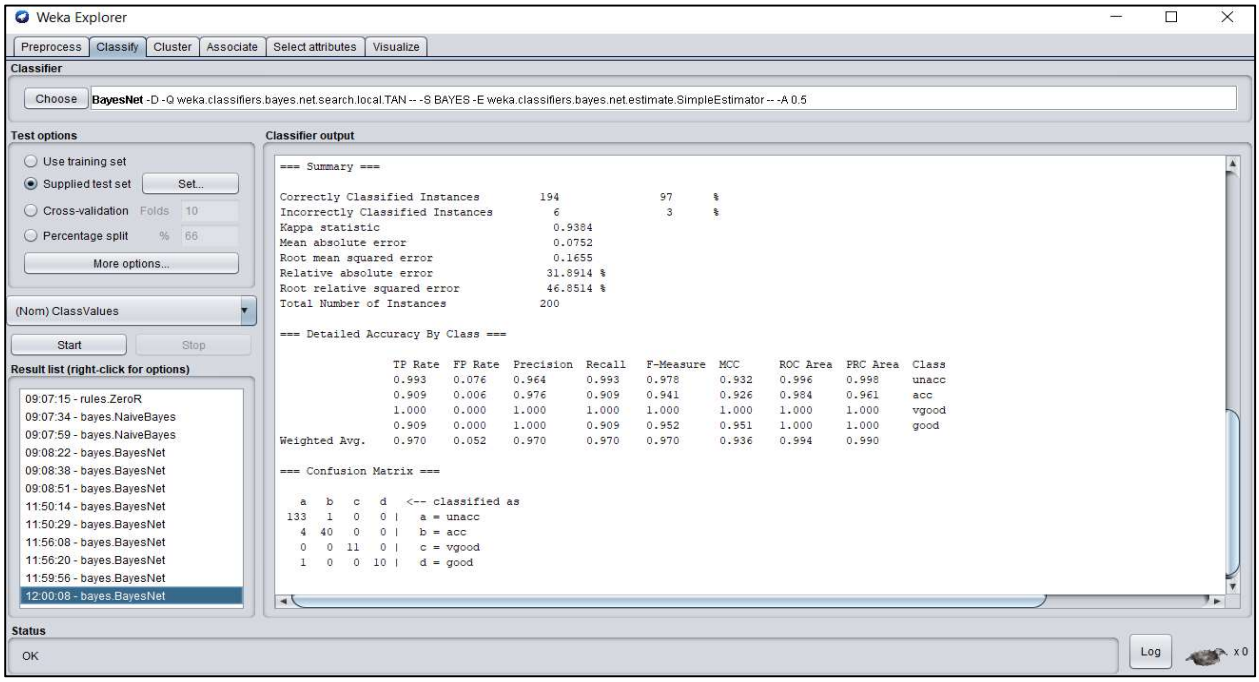

Figure 3.10 Weka TAN output trained in 1250 data instance

4. Conclusion

On comparative study taking in account classification accuracy, per class classification accuracy confusion matrix of both Naïve Bayes (NB) and Tree Augmented Naïve Bayes (TAN) , we can confidently conclude that TAN is preferred classifier then NB in the given data instance (car.data).

5. Error Learning

The attributes doors and persons displayed missing values and data type was string after processing the data.

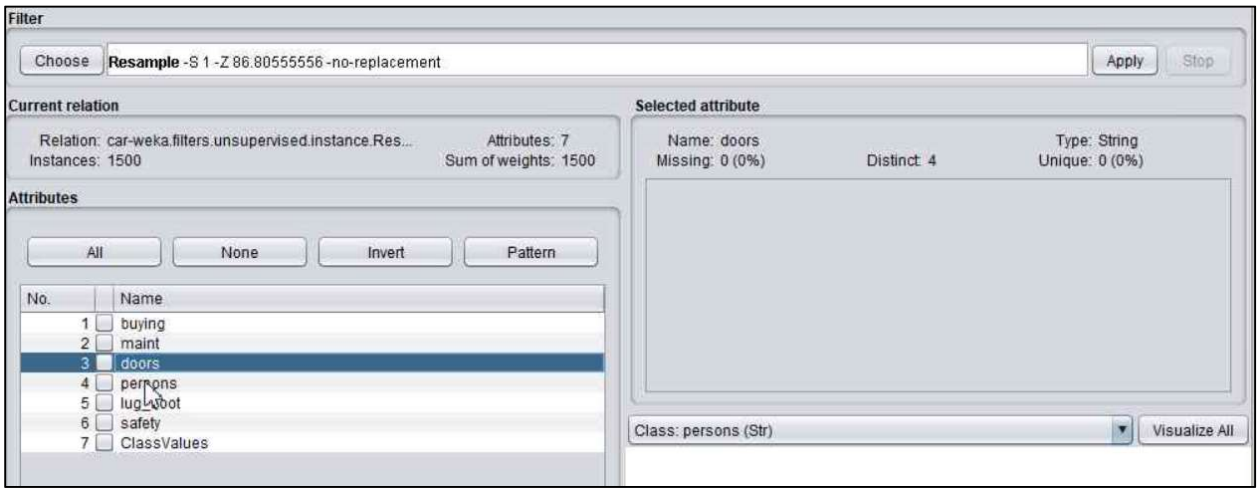

Figure 5.1 doors attribute not numeric or nominal

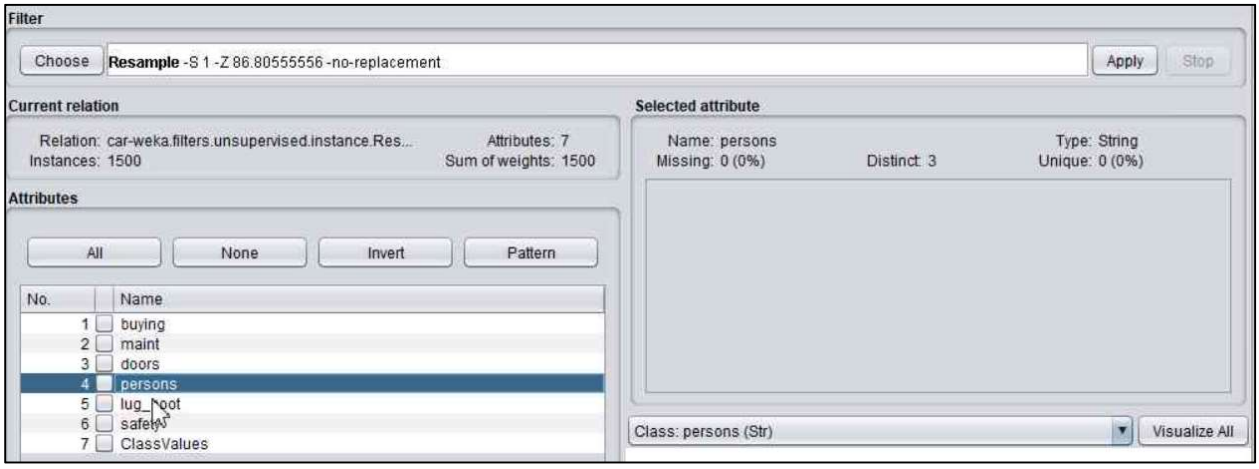

Figure 5.2 persons attribute not numeric or nominal

To process the data correctly and apply NB or TAN algorithm to data instances all attributes should either numeric or nominal.

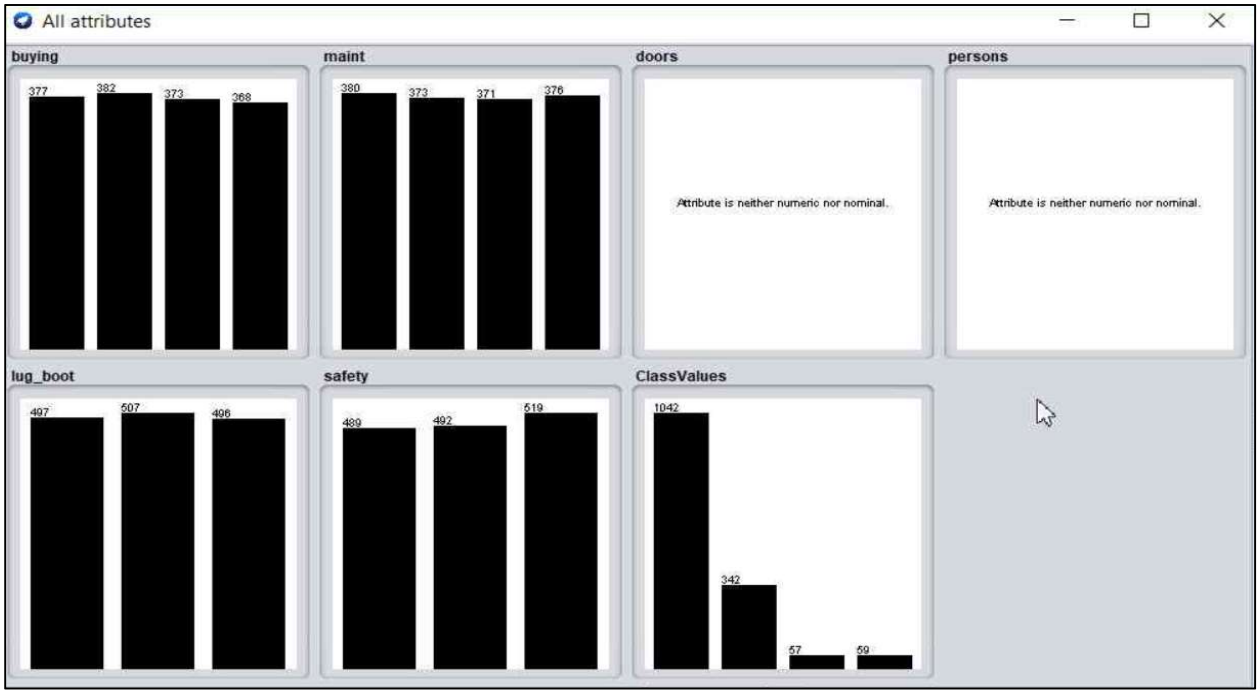

Figure 5.3 Attributes type error

The given attributes(doors, persons) was processed to nominal.

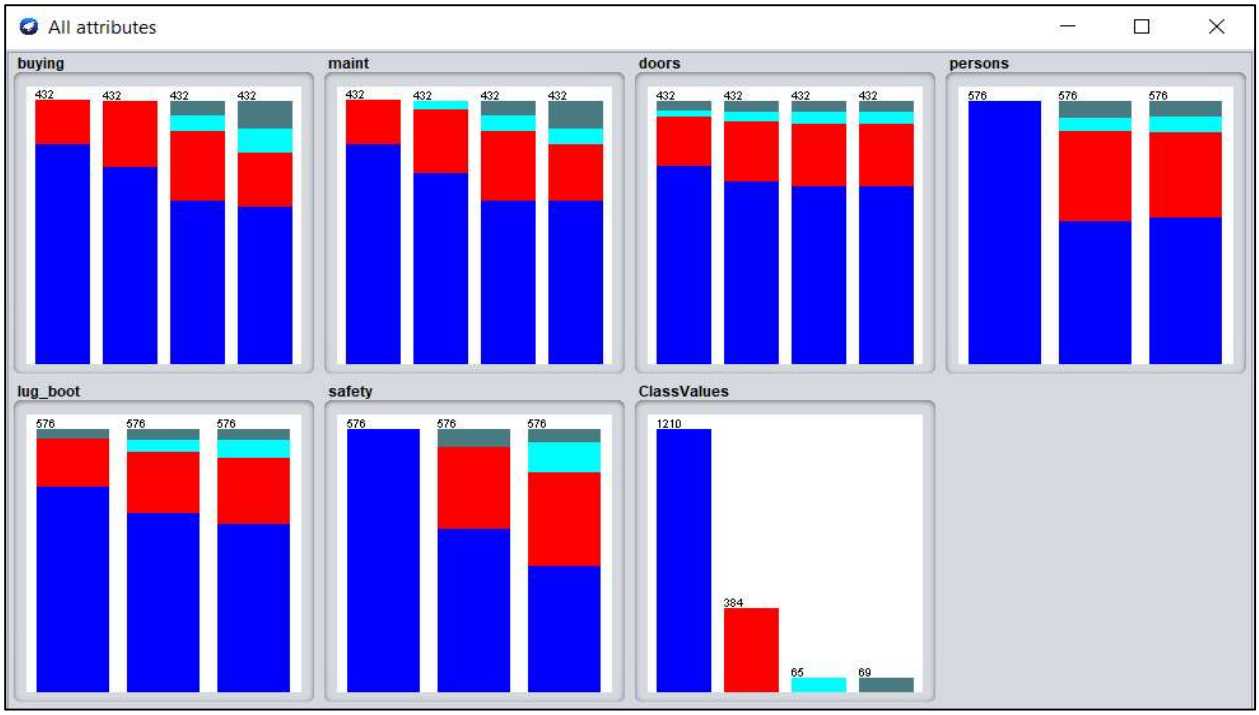

Figure 5.4 Different Attributes