Classification & Analysis of Car Evaluation

Introduction:-

There are various features associated when buying a used car. Those features decide the final condition of the car. However, the importance of one feature may out-weigh the importance of another feature and it may become difficult to decide as a whole and to evaluate a car.

Problem Statement:-

The aim of this project is to assist customers with buying used cars as per their requirements by identifying key features and perform classification analysis on the Car Evaluation Dataset to build a machine learning model using the training data to effectively classify the test instances based on the classifier model.

The different classification algorithms used are-

- Logistic Regression
- K-Nearest Neighbors
- Decision Tree

The cars are classified into one of the categories of- unacceptable, acceptable, good, and very good. We achieve different levels of accuracy for each of the algorithms employed, and we can compare their results and choose the best model as per the requirements.
Using Python in the Jupyter environment, we will analyse and pre-process the dataset, and then visualize it. Further, we will split the data into training and testing sets, and train the classifier models. Finally, the models will be tested and we will calculate their accuracies.

Technologies:-

Programming Language: Python (Anaconda Distribution)
IDE: Jupyter Notebook
Libraries: Numpy, Pandas, Matplotlib, Seaborn, Scikit Learn

Dataset Description:-

The dataset used in this project is the Car Evaluation Dataset created by Marko Bohanec and Blaz Zupan, obtained from the UC Irvine Machine Learning Repository.

Dataset Link: http://archive.ics.uci.edu/ml/datasets/Car+Evaluation

The database contains 1728 instances with 6 attributes and a class/category attribute. The model evaluates a car's acceptability based on the 6 attributes of 'buying', 'maint', 'doors', 'persons', 'lug_boot', 'safety'.

Attributes defining the status of the car:

- buying {vhigh, high, med, low} – buying price of the car in terms of very high, high, medium and low.
- maint {vhigh, high, med, low} – price of maintenance of car in terms of very high, high, medium & low.

- doors {2, 3, 4, 5more} – number of doors in the car i.e. 2 , 3, 4, and 5 or more.
- persons {2, 4, more} – capacity in terms of persons that the car in carry i.e. 2, 4, & more.
- lug_boot {small, med, big} – size of the luggage boot in terms of small, medium and big.
- safety {low, med, high} – estimated safety of the car in terms of low, medium and high.

Class attribute values that define acceptability of an evaluated car in terms of different categories are:

- unacc – unacceptable
- acc – acceptable
- good – good
- vgood – very good

Hypothesis:-

We will take into consideration the various features of a car, namely: buying price, maintenance, doors, persons, luggage boot, and safety, and then evaluate it into a class of its acceptability.

Here we assume that the buying price and safety of a car are the most crucial elements when evaluating a car.

A low buying price and high safety are the most favorable features when evaluating a car.

Data Analysis:-

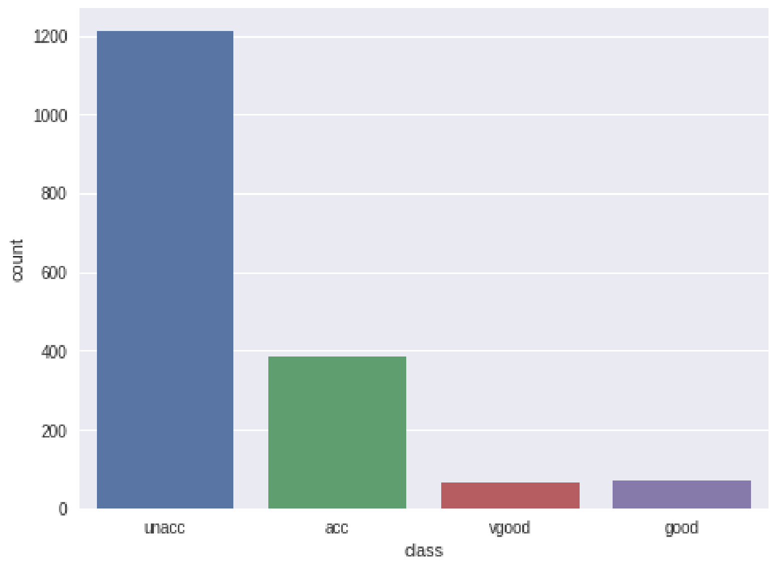


Fig. 1: Overall count plot for the different types of evaluated cars in the dataset.

Here in the graph, we can observe that cars with 'high' and 'very high' buying prices come under the 'unacceptable' and 'acceptable' categories; whereas 'low' and 'medium' car buying prices are considered as 'good' and 'very good' category of cars.

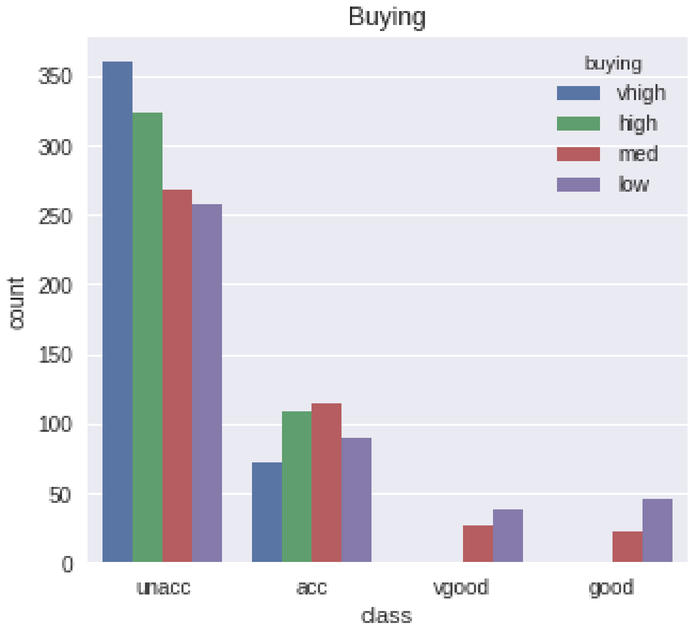
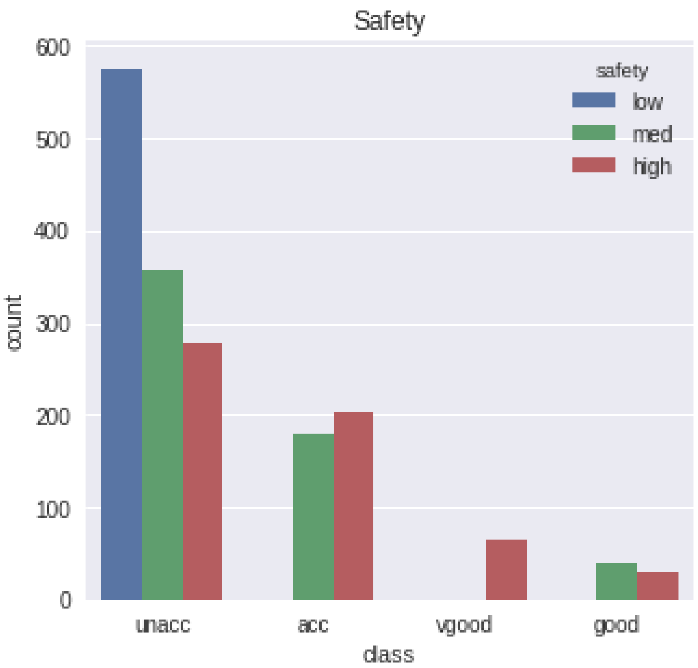


Fig. 2: Plot to compare buying price with classes of evaluated cars



Here in the graph, we can observe that cars with a 'low' safety feature are automatically classified as 'unacceptable' cars; and the ones with 'high' safety are more considered.

From this graph, another interesting observation that we can make is that, cars with a capacity of only 2 people are also automatically classified as unacceptable.

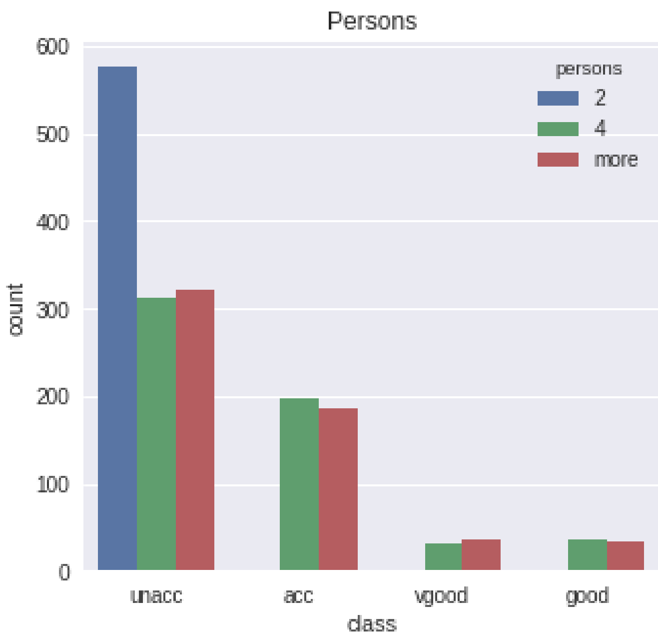
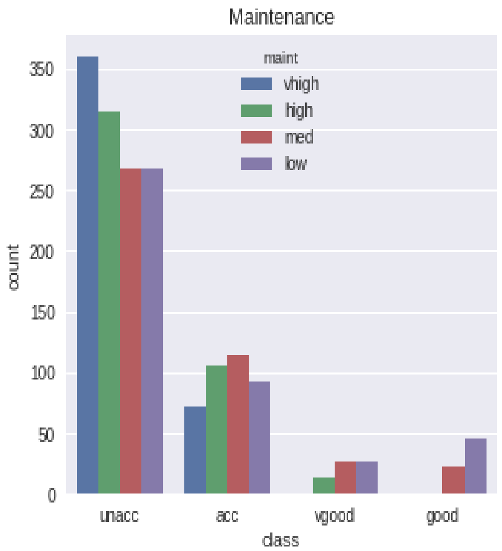
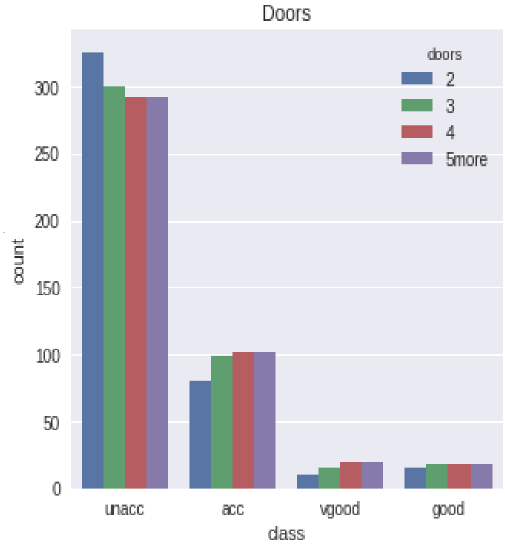
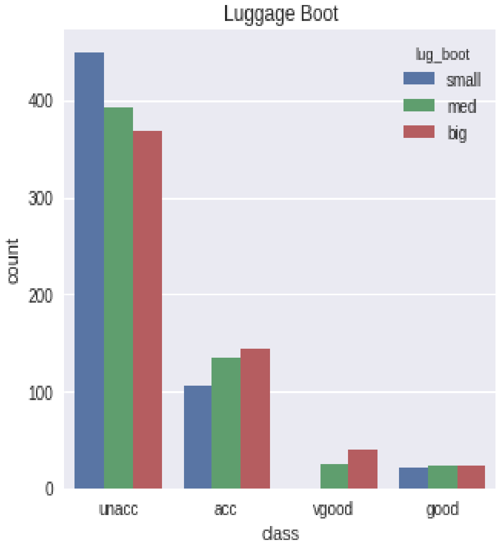


Fig. 4: Plot to compare persons capacity with classes of evaluated cars







From these three graphs, we can infer that the features: maintenance price, number of doors, and luggage boot capacity are not very crucial factors when deciding the acceptability and condition of a used car.

Classification Results:-

Comparing the various Machine Learning models implemented-

- ❖ **Logistic Regression:-**

Accuracy: 0.8410404624277457

Confusion Matrix:

[[59 2 18 0]
[15 0 0 2]
[11 0 229 0]
[7 0 0 3]]

Classification Report:

	precision	recall	f1-score	support
0	0.64	0.75	0.69	79
1	0.00	0.00	0.00	17
2	0.93	0.95	0.94	240
3	0.60	0.30	0.40	10
avg/total	0.81	0.84	0.82	346

Accuracy = 84.10%

- **K-Nearest Neighbors:-**

Accuracy: 0.9479768786127167

Confusion Matrix:

```
[[ 74   1   4   0]
 [  7   9   0   1]
 [  2   0 238   0]
 [  3   0   0   7]]
```

Classification Report:

	precision	recall	f1-score	support
0	0.86	0.94	0.90	79
1	0.90	0.53	0.67	17
2	0.98	0.99	0.99	240
3	0.88	0.70	0.78	10
avg/total	0.95	0.95	0.95	346

Accuracy = 94.79%

- **Decision Tree:-**

Accuracy: 0.9884393063583815

Confusion Matrix:

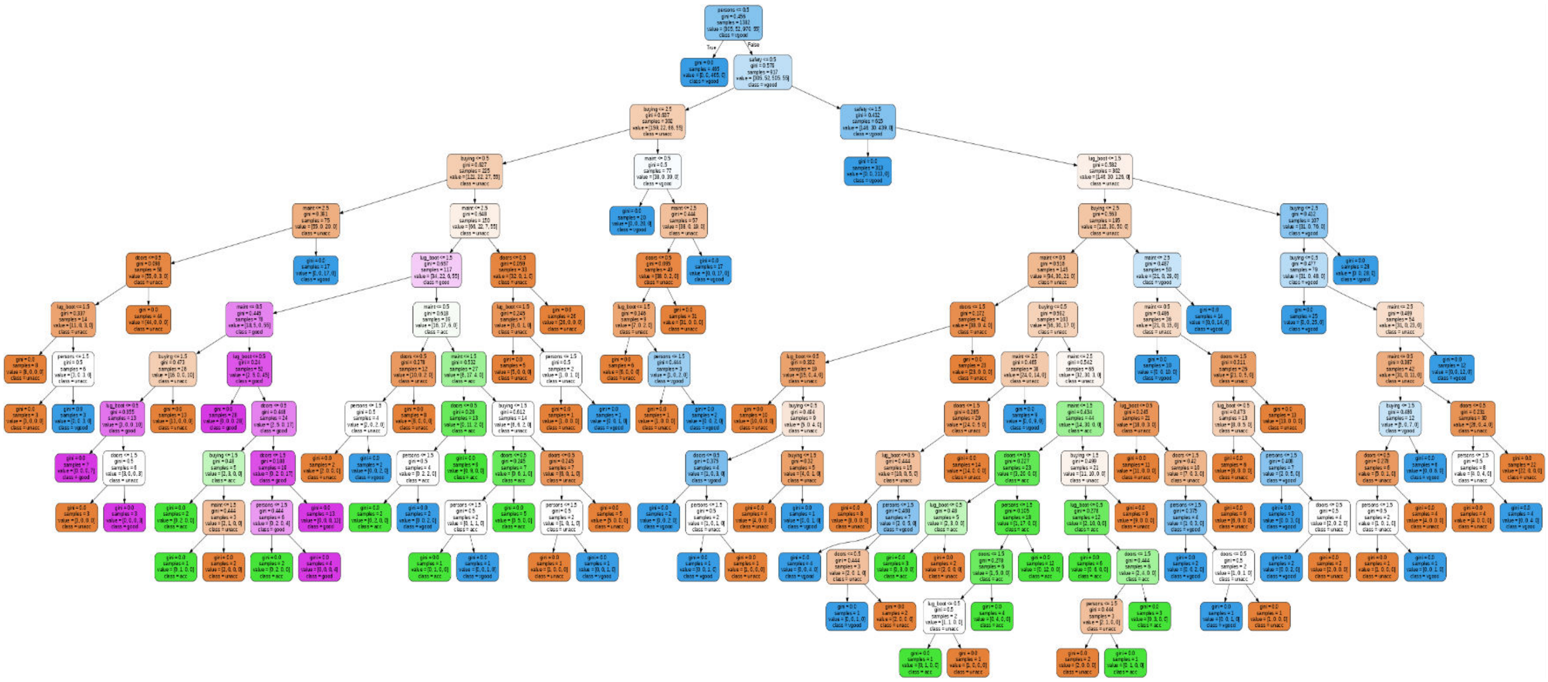
```
[[ 76   2   1   0]
 [  0  17   0   0]
 [  0   0 240   0]
 [  1   0   0   9]]
```

Classification Report:

	precision	recall	f1-score	support
0	0.99	0.96	0.97	79
1	0.89	1.00	0.94	17
2	1.00	1.00	1.00	240
3	1.00	0.90	0.95	10
avg/total	0.99	0.99	0.99	346

Accuracy = 98.84%

As we can see, the Decision Tree algorithm achieved the highest accuracy to best classify the test instances.



The above diagram shows the generated Decision Tree Model

Conclusion:-

- Therefore, we conclude that as per our analysis, the initial hypothesis has been successfully proved.
- The buying price and safety of a car are two of the most important features when evaluating a car; and that a lower price with higher safety implies better acceptability of the used car.
- Out of the three algorithms implemented, the Decision Tree algorithm achieved the highest accuracy, and was the best model to predict the class of the car, given its features.