SUMMARY:

Problem at hand:

Car evaluation dataset from UCI repository consists of 1728 instances and 6 attributes, Based on these 6 attributes(which are buying_price, maintenance, doors, persons, luggage_boot, safety) we predict the class label which is “acceptability”, it is “0” if it’s not acceptable by the customer and “1” if it’s acceptable.

Data Cleaning:

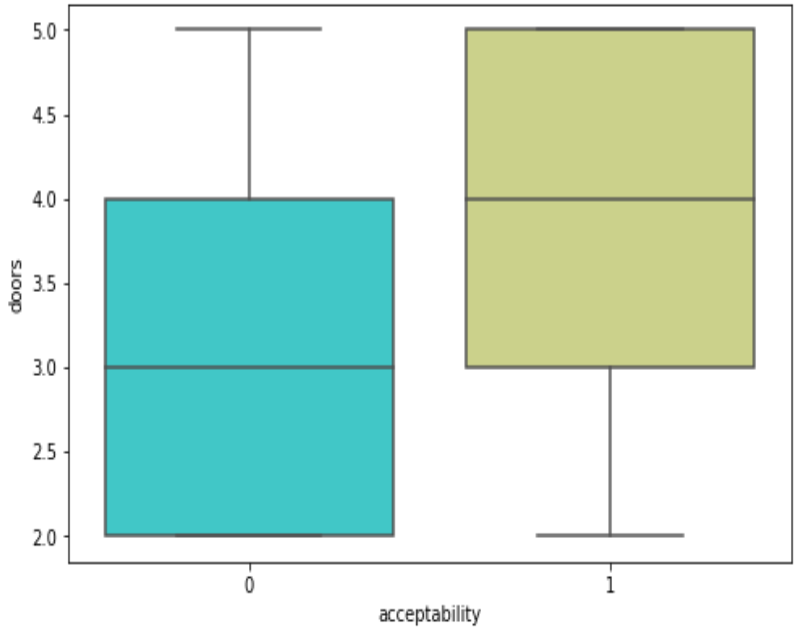
Data cleaning is performed on the data as the data had some ambiguous data type values in columns such as integer and string in one column. So strings like that are replaced with numerical values and then other categorical values are also encoded with the help of label encoder, Label encoder is a better choice here as the data is ordinal, one hot encoder would have been a better choice if our data wasn’t ordinal.

Training and Testing:

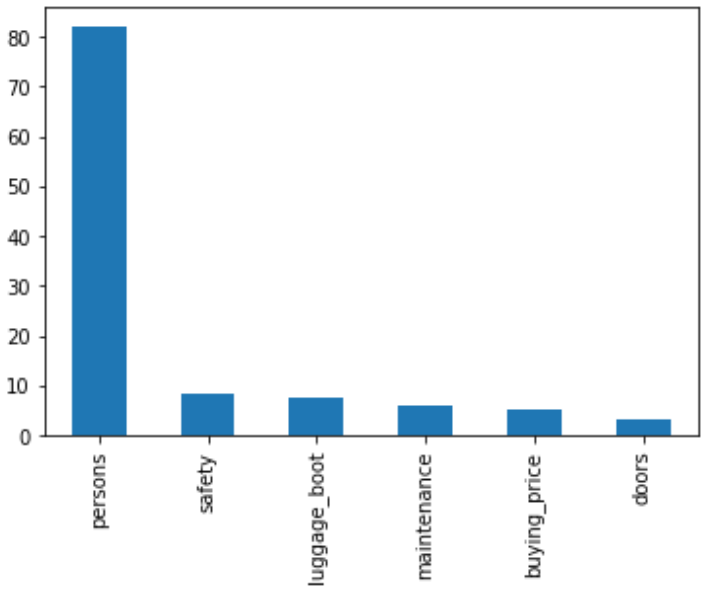
The dataset is then split 50/50 into training and testing sets and is stored in variables like X_train, y_train, X_test, y_test.

Plots:

Box plot is plotted on doors vs acceptability which showed that the cars with more doors are often accepted.



A bar graph is plotted based on chi square scores for feature importance, here we can see that “persons” attribute has a higher effect on the predicted label.



Fitting on different models:

The dataset is applied on different algorithms and confusion matrix values, accuracy is noted in the table below.

Model	TP	FP	TN	FN	accuracy	TPR	TNR
Logistic Regression	105	143	521	95	0.72	0.525	0.78
Decision Tree	247	1	609	7	0.99	0.97	0.99
Random Forest	216	32	584	32	0.92	0.87	0.94
Linear SVM	104	144	529	87	0.72	0.54	0.78
Gaussian SVM	166	82	555	61	0.83	0.73	0.87
Polynomial SVM	162	86	575	41	0.85	0.79	0.86
Naïve Bayesian	131	117	554	62	0.79	0.67	0.82

From above table we can see that Decision tree classifier has highest accuracy in classifying the data correctly.

Algorithms such as KNN and k means clustering are not used as they don't work well with categorical data and only work well with numerical data as they are distance-based algorithms. This is another the reason why scaling is not done on the data as the interpretability will be worse because the data has many categorical attributes.