Classification

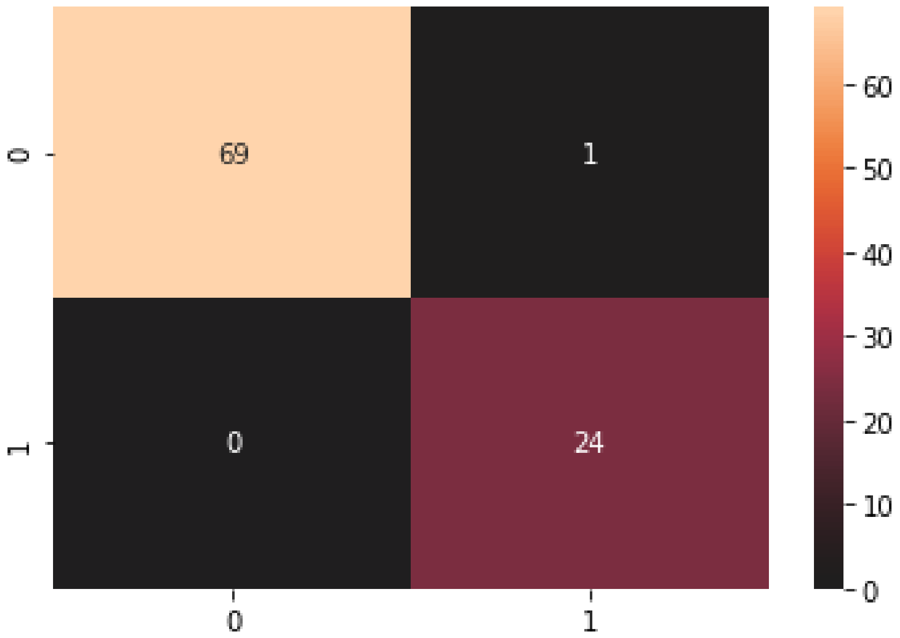
Decision tree is a supervised learning method we used it to classification in our dataset and predict by learning some rules that were inferred by some attributes.

It's simple to understand and visualize, we can visualize tree, not needed dummy variable to be created, able to handle multi-output problems, not needed many data preparation like others method.

But some concepts are hard to learn and it can make biased tree if some attributes dominate.

We used Decision tree Classifier with parameters gini for criterion and best for splitter.

We used Decision tree classification to make a model for predicting which hotel was visited, we transformed columns that had type of string, Category and date to int by label encoder, selected attributes for training, divided dataset into four parts, printed score of training which was 1.0 ,printed score of testing which was 0.9893 and that's look powerful score, concluded important features, predicted hotel attribute, made confusion matrix, used heatmap function to visualize the correlation between test data and predicted data.



We made a good prediction for hotel attribute and assigned it to y_predicted variable and we can see our model predicted 18 record true in the beginning of column compared to our original data (which refers to y_test).

Clustering

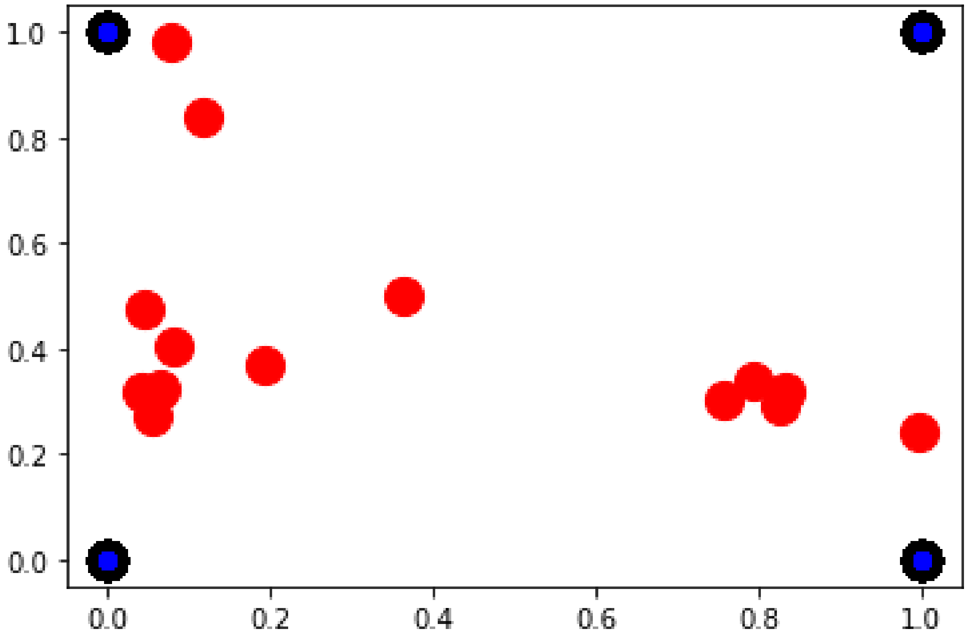
We used KMeans Clustering from Sklearn which is unsupervised learning, it cluster our dataset by dividing data into n groups (we used 15 clusters) with equal variances.

We used KMeans class with parameters k-means++ for init and 10 for n_init.

We transformed data which are had type of string,date and category to int values by label encoder, divided data into four parts but we make dummy variable Y as split function needs two parameters x and y, fitted data, printed score of training 1499117326.343, printed score of testing -107346701.22, printed centers like 3.64298195e-01 5.00000000e-01 3.24373617e+02 2.52417006e+01, printed lables like 1 11 1.

We used 99 iteration, predicted some values like [6 2 6 14 4 11 10 0 9 10 13 2 9 10 6] and visualized original data, test data and centers by scatter.

Red for centers, black for original data and blue for test data


1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0