TASK I

Sai Kalyan Katta

HAPPY CUSTOMERS

saikalyan.katta@gmail.com

DATA:

I have used the ACME-HappinessSurvey2020 dataset for this task.

PRE-PROCESSING:

As I have found no data missing from the dataset, I have skipped the data replacement techniques for this dataset.
Next, I have checked for the correlation between all the columns of the dataset. The correlation matrix is as follows:

0	1	2	3	4	
0	1.000000	0.059797	0.283358	0.087541	0.432772
1	0.059797	1.000000	0.184129	0.114838	0.039996
2	0.283358	0.184129	1.000000	0.302618	0.358397
3	0.087541	0.114838	0.302618	1.000000	0.293115
4	0.432772	0.039996	0.358397	0.293115	1.000000

Looking at the correlation table I have found that the combinations, [X1, X2, X4] and [X1, X4, X5] would be the optimal combinations for the maximum accuracy.

As the algorithms such as Logistic Regression, SVM and KNN require the data to scaled, I have used the StandardScaler to scale the data before I perform the Analysis.

TRAINING AND TESTING DATASETS:

I have split the data into training and testing datasets with 20% of the total dataset being the testing dataset

LOGISTIC REGRESSION:

For the logistic regression, I have used the default 'lbfgs' solver and I have scored an accuracy of 73%.
As the dataset was small, I have also tried using the 'liblinear' solver with 'l1' and 'l2' penalties and the model scored around 60-65% accuracies.

SVM:

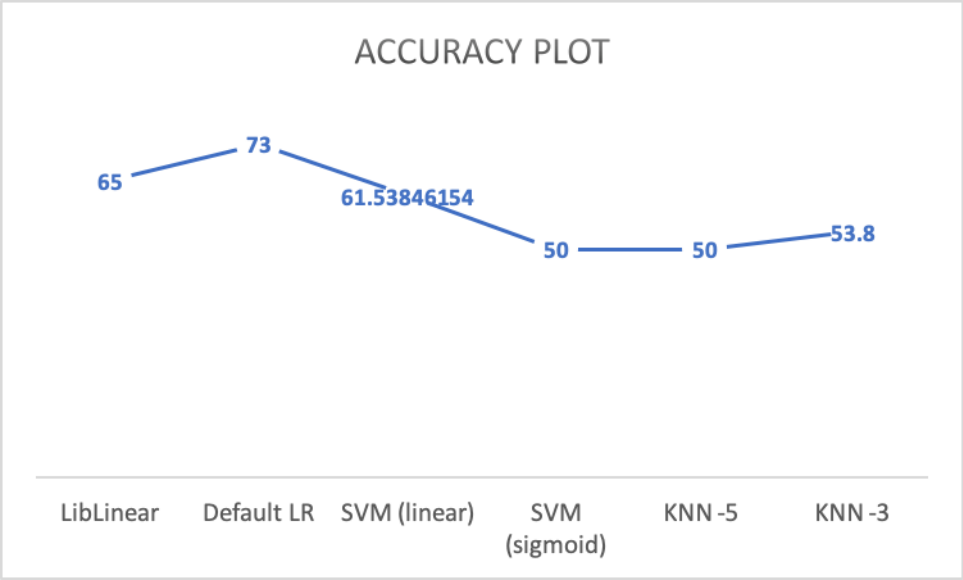
For SVM, I have used the linear and sigmoid kernels.
The accuracies for the linear kernel and the sigmoid kernels were 61.5% and 50% respectively.

KNN:

For KNN, I have used the 3 and 5 nearest neighbors configurations.
The accuracies for the 3 neighbors and 5 neighbors were 53.8% and 50% respectively.

COMPARISON:

The comparison plots for all the different model's accuracies can be seen below.



CONCLUSION:

Looking at the accuracies that I have scored during the analysis, it is good to say that the Logistic regression with the 'lbfgs' solver has performed well when compared to the other models.

NOTE:

I have also tried to perform the PCA (Principle Component Analysis) to reduce the dimensionality of the data before performing the analysis. But I have seen that there was an effect in the accuracy of the algorithm due to the dimensionality reduction. So, I have gone forward with the data without PCA.