Hakan Gündoğdu

INTL 550
HW 3

For this task, I started with the feature selection to gain from computational effort. When I checked the features that have huge impact on the result, score of 12 features attract the attention. After this feature selection I reconstructed the data with a more compact form and split 80% of it to train data and 20% for test data.

Then I used different models, some are supervised, and some are unsupervised, to better prediction. I mostly preferred to use the supervised models, since they are explainable in terms of their aspects and more understandable.

Besides to train and test split, I also used cross valuation to eliminate any incidence and calculated mean and standard deviation of 8 folded results, as shown below.

	Method	y_head	Accuracy	cv_mean	cv_std
2	Random Forest		0.849458	0.856024	0.007478
3	KNN		0.845042	0.850100	0.004153
0	Logistic Regression		0.823364	0.829618	0.005854
1	Gaussian Naive Bayes		0.740666	0.732932	0.013007
4	SVM		0.724609	0.728313	0.201779

Figure 1 Accuracies of different methods and their standard deviation

When we look to table, we see Random forest and KNN performs well, and standard deviation of the folded test is moderately okay. If I did not check the name of the methodology, I would prefer to use the 3rd indexed method, which has near to maximum of the mean of accuracy however lower standard deviation. On the other hand, when we checked the methodology name, we see it is KNN, which is an unsupervised learning method. To eliminate any unexplainable classification, I prefer to use Random Forest.