IST 707 Applied Machine Learning
HW7: kNN, SVM, and Random Forest for handwriting recognition
Name: Jainish Savaliya

In this classification problem we are given the same task as last assignment which is to recognize the digits from 0 to 9 from the given dataset with the given greyscaling of 0 to 255. In this problem set we are supposed to use KNN, SVM and Random Forest prediction models.

For the Preprocessing steps I included all the necessary libraries and packages then I factorized the first column of the dataset which is labels and the rest of the data will be as its because of the number of the pixels is too large and the factorization of them will not work in a proper way. Here in this scenario we already have the train and test dataset so there is no need to create the partitions. Also, I sliced the dataset to 1000 rows so that it will take less time to train and test.

In the KNN algorithm to use the whole dataset I expanded the grid from 5 to 25 with the increment of 2 in sequence. I used the cross-validation method with 3 folds and used the tuning parameters into training the model and predicted that on the test dataset and I got the accuracy of 87.8% with just the training and with the testing it got the accuracy of 89%.

In the SVM Linear algorithm to use the whole dataset I expanded the grid from 0.1 to 2 with the 20 values in sequence for the C . I used the cross-validation method with 3 folds and used the tuning parameters into training the model and predicted that on the test dataset and I got the accuracy of 87% with just the training and with the testing it got the accuracy of 88%.

In the SVM Radial algorithm to use the whole dataset I expanded the grid from 0.1 to 2 with the 20 values in sequence for the C . I used the cross-validation method with 3 folds and used the tuning parameters into training the model and predicted that on the test dataset and I got the accuracy of 9% with just the training and with the testing it got the accuracy of 94% which I ran into python because it gave a very poor accuracy in R.

In the Random Forest algorithm to use the whole dataset I expanded the grid from 1 to 5 to use all the possible value of mtry. I used the cross validation method with 3 folds and used the tuning parameters into training the model and predicted that on the test dataset and I got the accuracy of 89% with just the training and with the testing it got the accuracy of 91%.

After Comparing all three algorithms the ranking of the best algorithms is SVM Radial , Random Forest , SVM Linear and KNN respectively. This is because SVM Radial is more suited for non linear relation ship of the input and output and also I handles well the complex data set . Random Forest handles north the categorical and numerical dataset well but have hard time with noise so that is the reason it is on the 2nd rank. Third is SVM Linear model which is made to deal with the linear dataset but also capable of handling highly corelated features. Finally the last one KNN is suited for small and more structured dataset and it requires the normalization and is sensitive to outliers which is also called the lazy algorithms so may be for those reasons KNN is the worst performing model out of all three.

Now comparing these models to Decision tree and Naive Bayes this three models perform better than Decision Tree and Naive Bayes reason being those models are not suited for handling high dimensional and complex datasets. In contrast SVM, RF and KNN are more powerful and flexible models that can effectively handle these types of datasets, especially for digit recognition task.