AIML Assignment 3 Report
Chris -IS201401015
Santhoshini-IS201401040

The file ppa_tra.dat is shuffled using the file **split.py** and **split2.py** in order this produces the file **train.dat** out of which all training operations are performed.
The testing data is kept in a file called **test.dat.**
For condensed set KNN classifier, the condensed is stored in the file **cset.dat.**
Normal KNN Classifier both training and testing parts are in the file **knnc.py.**
Modified KNN Classifier both training and testing parts are in the file **knncmod.py.**
Condensed set KNN Classifier both training and testing parts are in the file **condknnc.py.**
All codes are written in python and uses the **numpy** python library for the use of matrices and numerical operations. Numpy must be installed prior to running the code.
On average the **training** phase takes **15-20** minutes to run.
Testing phase duration dependents on the number of test queries.

Results for K Nearest Neighbour Classifier using cross validation:
After running the training set for values of k in the range [1,10]:
We get the following results in 3-fold cross validation:
<u>On training set</u>
Mean error for k = 1 is: **190.0/6670 : 2.8%**
Mean error for k = 2 is: **236.0/6670 : 3.5%**
Mean error for k = 3 is: **194.0/6670 : 2.9%**
Mean error for k = 4 is: **206.6/6670 : 3%**
Mean error for k = 5 is: **201.3/6670 : 3%**
Mean error for k = 6 is: **204.0/6670 : 3%**
Mean error for k = 7 is: **197.0/6670 :2.95%**
Mean error for k = 8 is: **199.3/6670 : 2.9%**
Mean error for k = 9 is: **203.3/6670 : 3%**
Mean error for k = 10 is: **211.0/6670 : 3.1%**
<u>On test set</u>
For k value : **1**
Error rate: **11.06% ----------**(166/1500)
For k value : **7**
Error rate: **8.7% ----------**(131/1500)
Results for modified K Nearest Neighbour Classifier using cross validation:
After running the training set for values of k in the range [1,10]:
We get the following results in 3-fold cross validation and distance based voting:
<u>On training set</u>
Mean for test = 1 is: **190/6670 : 2.8%**
Mean for test = 2 is: **190/6670 : 2.8%**
Mean for test = 3 is: **190/6670 : 2.8%**
Mean for test = 4 is: **184/6670 : 2.8%**

Mean for test = 5 is: **179/6670 : 2.6%**
Mean for test = 6 is: **175/6670 : 2.6%**
Mean for test = 7 is: **174/6670 : 2.6%**
Mean for test = 8 is: **173/6670 : 2.6%**
Mean for test = 9 is: **173/6670 : 2.6%**
Mean for test = 9 is: **173/6670 : 2.6%**
Mean for test = 10 is: **173/6670 : 2.6%**
On test set:
For k value : **8**
Error rate: **4.1% ----------**(56/1350)

Results for condensed K Nearest Neighbour:
Building the condensed set based on the training data we get condensed set size to be: **1234** for a training set of size: **6670**
Running 1-NN classifier on the condensed set for the training set, we get **3029/3333** correct classifications which gives error rate as : **9.12%**
Running k-NN classifier on the condensed set with the k = **7** obtained from cross-validation we get **3099/3333** correct classifications which gives error rate as: **7%**
Running k-NN classifier on the condensed set with the k = **8** obtained from cross-validation we get **3098/3333** correct classifications which gives error rate as: **7.05%**