The program is written in Python and needs following packages

To install the packages you need to first install pip by **sudo apt-get install pip** and then just use **pip install <package-name>**

1. sklearn
2. pylab
3. numpy
4. pandas
5. sklearn
6. ggplot
7. pylab
8. csv
9. scipy

We have performed our analysis on LinearSVM and RBF Kernel SVM by running them on two datasets.Although the program flow is same, the data preprocessing varies for the data sets.

Thus, we have submitted two folders for the two data sets

WE are following these convention through both the codes

X=train data

y=train labels

x\_1=test data

y\_1=test labels

***DataSet1. Bank Dataset***

The original dataset consists of 2 csv files-

1.Bank.csv(train data)

2.Bank-full.csv(Test data)

The file to run is "first.py"

You will have to have sklearn,numpy,pandas packages installed

We have divided the code is sections to make it easier to understand.

Section ==>Splitting the data

Here for our ease, we split the data and labels and store them separately in csv files.

Section==>Preprocessing the data

Here we preprocess the data using the steps mentioned in our report

Section==>Running the SVM CLassifier

\* Feature Selection

Here we use c=1 and gamma=auto

After the cross validation section is executed, we can input the values of c and gamma to get accuracy

# The classifier is running on RBF kernel currently.

# TO run the linear classifier , comment the Rbf kernel classifier command and uncomment the linear kernal

\* Precision and Recall

Precision and Recall is needed only when running on test data

Section==>Cross Validtaion with Grid Search

#To run without cross validation, comment the entire section code till the end

# We have initalised the values for C and Gamma

P.S.-This step takes time to execute due to the complexity of the model

\***For the next two datasets, we first preprocess the dataset by splitting into random sets of train and test data using prelim.py Thus, to replicate results, please do not run prelim.py again as it well again generate random datasets.**

***DataSet2. Virus Dataset***

The original dataset consisted dataset.train and Tst.test files.

Since the dataset had a marker of -1, we deleted it manually and converted the files to .txt files before processing

Original Dataset= Dataset.train, we manually converted it to dataset\_train.txt

Then using prelim.py, we generate dtrain.txt and dvalid.txt from dataset.train

The file to run is "virus.py"

This code follows the same procedure for SVM , and GridSearch.

For Cross validation, you will have to uncomment the section till the end of the code

HOwever, since there is no preprocessing required as the data is available in SVMLIb format, we omit that section

Also, we do not have precision and recall in this dataset as we are running this classifier on the train data.

***DataSet 3 . Dorothea Dataset***

We use prelim.py to preprocess dorothra\_train.txt and dorothea\_valid.txt to generate dtrain.data.txt, dvalid.data.txt, dvalid.lables.txt, dtrain.lables.txt

The file to run is "third.py"

This code also follows the same flow as the first two codes.

The only addition is the two algorithms - PCA and Chi2 Feature selection. To use anyone, just uncomment the other algorithms. We can change number of components in PCA and number of features in Chi2 by changing the value in the pipeline at all occurrences.