Machine Learning Assignment 2:

For this assignment, we were tasked with creating a python program that took in an excel file of rows with values of varying degree and needed to take out certain values and create an image with the use of them as well as evaluate the performance of different classifiers for their suitability to classify this dataset.

Task 1

This task simply revolved around properly extracting the data from the CSV file. In order to get proper reading later, we needed to split the labels so that it would not interfere with our classifier predictions. We also needed to just get the number of data relating to a sneaker and a boot as well as parameterize our finds so that it may be used in the next function. Lastly, we needed to display an image using the data. In my case, rather than printing it in the compiler, I found it easier to save the image itself. So, when ran it gets the data and produces the image, it saves it as a PNG file and is saved to whatever folder this program is working out of.

Task 2

We then take in the data read off the excel sheet as “features” and “labels” into the function getFeatures(). Outside of where this function is called is a for loop which loops around different sample sizes to use for the classifiers. This number is then passed into both features and labels. The features variable is a double array whereas the labels are just a single array. We then perform a kFold loop and have it gone through 9 splits

```
cv = KFold(n_splits =9)
    counter = 1
    for train_index, test_index in cv.split(featureToUse):
        trainX = featureToUse.iloc[train_index]
        trainY = labelsToUse.iloc[train_index]
        testX = featureToUse.iloc[test_index]
        testY = labelsToUse.iloc[test_index]
```

These variables are then used in this section and the next. We also time the compiler time to get the linear model prediction. The first score that is printed out is the one which we were to print out for this task. We also get the maximum and minimum value found within each slit when working on the excel file

Task 3

Lastly, we use a Support Vector machine to get our accuracies and record the time it takes similar to the last task. We were required to use both a linear kernel and a radial basis function kernel for this task. For the SVM we now have multiple versions, as each one that uses an RBF kernel has its own gamma value to use when predicting. We used four separate values to get a good understanding of the effect these had on the prediction value. For the last task, we got the time for each score as well as the maximum, minimum and average between all the times. This now will give s a good idea of how these two classifies compare.

Task 4

This task is just our overall comparison of the two classifiers and what we noticed when reading our results. As we started to add greater sample numbers, the time for each prediction began to become much longer. With these samples in mind, SVM always seems to have a significantly lower average time compared to the multi-layered prediction when you have a lower sample. However, once you enter the 10000 range, SVM takes a very long time in comparison to MLP.

Accuracy scores for the MLP and SVM with a linear kernel are relatively close. The RBF however, isn't yet despite using gamma values that differ, with less difference between each gamma there is very little difference in its accuracy score. As gamma value becomes in our case smaller, the difference will start to become more pronounced and will be more indicative of the true accuracy

A somewhat similar case happened with he confusion matrix, in which while MLP had some difference between splits, SVM had much more obvious differences between each gamma and split at lower samples but will begin to underperform when it hits to high a sample.

Ultimately as to which one I would choose it would depend on the sample size being used. It seems from the results of my program that SVM would work better with smaller sample size (around 500) and provides very accurate reading when in this sample. However as it becomes bigger(anything above 5000), It might be ideal to switch to MLP as while it really doesn't get any better, SVM will begin to really struggle.