EEE 591 Project 2 Report

When the code is run where random_seed is kept None for train_test_split, it causes huge difference in the result. For example, accuracy is calculated to be 0.87 for 35 components. However, when the random_seed is assigned an integer value of 3, the accuracy is calculated to be 0.90 for 4 components. This happens because random_seed determines how the dataset is split into training set and test set. If it is set as None, every time the code is run, training set and test set are different and thus, we get different accuracy. So, in my opinion, setting random_seed to a constant is better.

I have chosen the below parameters for constructing the MLPClassifier:

MLPClassifier(hidden_layer_sizes = (100,), activation='logistic', max_iter=2000, alpha=0.00001,solver = 'adam', tol=0.0001,random_state = 5) as the accuracy for this architecture is 0.95.

The above parameters are chosen as they gave the best accuracy so far.

The model predicts that the chance of surviving the real mine field is 95%. Confusion matrix for n_components = 7 and accuracy of 0.95 is :

N=63	Class 1(Rock)-Predicted	Class 2(Mine) - Predicted
Class 1(Rock)-Actual	28	1
Class 2(Mine)-Actual	2	32

This means that total number of data present in test set is 63. The total number of actual data belonging to class 1 i.e. **Rock** is 29 and total number of actual data belonging to class 2 i.e. **Mine** is 34. Out of 29 actual Rock data, 28 are predicted right but 1 is predicted to belong to be mine. This situation is called False positive. Out of 34 actual Mine data, 32 are predicted right but 2 are predicted to belong to class rock. This situation is called False Negative. In short, confusion matrix is used to describe the performance of a classification model.

Setting random_seed to None in MLP classifier is giving different results in different runs. For example, for the above parameters, 5 components gives 0.92 accuracy. In another run, 7 components is giving 0.94 accuracy,10 components is giving 0.94 accuracy and so on. This implies that setting random_seed to None will generate a random number in all runs and thus leading to inconsistent results. However, setting random_seed to 5 is giving 0.95 accuracy for 7 components.

Maximum accuracy is obtained for components=7 which means that those 7 features are the most important features and drive the whole model. They carry the most information in that dataset. Accuracy is getting decreased with inclusion of other features as they do not contain much information and thus do not contribute to the prediction. Thus, the plot of graph is increasing till 7 components and after that, the accuracy starts decreasing.