Cyber Security of Critical Infrastructure
Assignment 3 Report

Gouni Anuraag Kiran(18111017)
Hariom(18111019)
M V Ravi Kumar(18111037)
Shikhar Barve(18111065)
Siddharth Kumar(18111068)
Kamlesh Kumar Biloniya(160317)

In our Assignment, we are provided with HW_TESLA.xls dataset which has :-

- Number of features = 132, which are the readings that are taken from PMUs placed on transmission lines. These readings are voltage angles, Real and Imaginary Current . There are 4150 such records which we would use for Training, Validation and Testing.
- Labels are 0 or 1 depending on whether System is in 'Safe' condition or 'Stressed' Condition.

Data Analysis:-

Given data is real-life data prepared to analysis the stressed and safe load-flow system environment. The Main Aim was being able to predict stressed network in real-time by seeing readings from some or all of the transmission lines.

- Data feature follows Gaussian Distribution.

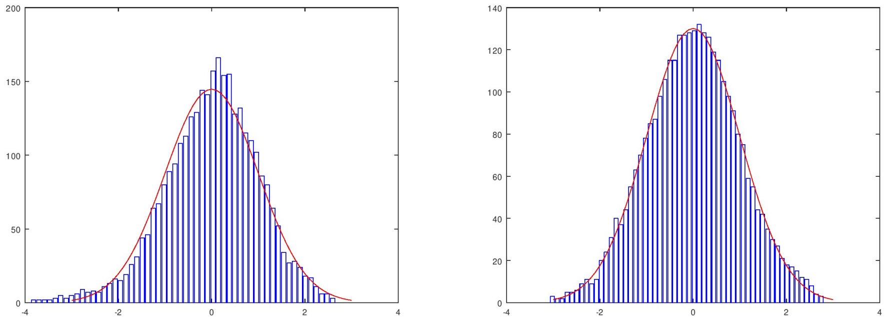

Figure 1: Feature no. 1 and 5

- Many Features have very small variance.

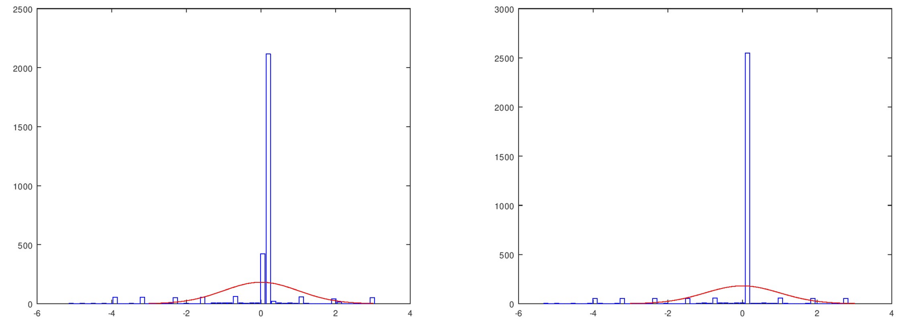

Figure 2: Feature no. 16,28,85

- We observed Some features vary more than others.
- We can use Principal Component Analysis(PCA) for data dimensionality reduction. Using PCA we transform a D-dimensional feature vector into a K-dimesional feature vector.
- We have numerous classifiers namely 1−NN, K−NN, Linear SVM, Decision Tree, Logistic Regression, RBF kernel SVM and many more. We can choose any one which suits our data best.

Model Selection :-

There are several models and each of them have there own pros and cons. We tested accuracy for dataset in three settings.
1.) With PCA. 2.) Using Original full-featured dataset
3.) With feature selection.
Before training the model, we are removing duplicate examples from the dataset then *randomly* dividing whole data into 75% training and 25%testing data.We are using Leave-one-Out Cross-validation(LOOC) on 25% of entire dataset five times with different splits. Therefore, each execution of program will have different training and testing dataset leading to different accuracies.

1.With PCA :- We used PCA and then trained model for Decision Tree, Linear SVM and Logistic Regression. We got result that if we need 99.99% variance to be preserved we need at least 6 principal components. Below, given is the table of % accuracy for different models on dataset for different number of principal components.

# PCA Components	Decision Tree	False Positives	False Negatives
4	98.4570	9	7
5	98.5535	11	4
6	98.8428	5	7

# PCA Components	Linear SVM	False Positives	False Negatives
4	99.2285	2	5
5	99.4214	2	4
6	99.5178	1	4

# PCA Components	Logistic Regression	False Positives	False Negatives
4	99.3249	5	2
5	99.6142	0	4
6	99.6142	0	4

2.Using Original full-featured dataset :- Using PCA we reduced the number of dimensions very much but it completely transformed the training as well as test features according to principal components. So we could not emphasize the individual importance of original features from transformed data points.Here, we trained models like SVM with linear kernel, Decision Tree and Logistic Regression. Below table shows the accuracy

Iteration	Decision Tree	False Positives	False Negatives
1	99.5183	2	3
2	99.7109	2	1
3	99.4219	2	4

Iteration	Linear SVM	False Positives	False Negatives
1	99.5180	0	5
2	99.7109	1	2
3	99.8073	2	0

Iteration	Logistic Regression	False Positives	False Negatives
1	99.4219	1	5
2	99.5183	3	2
3	99.3256	0	7

3.With Feature Selection :- To predict the correct label of an example, it is not necessary that we use whole feature-set. We can choose a set of features depending on the classification model which would classify examples easily with high accuracy. A drawback with this approach is while using few selected features the examples which have similar value in selected features are considered as a single example so for all those examples only one such example would be kept and rest all would be discarded. Though with HW_TESLA it doesn't affect very much.

- **Decision Tree:-** In this case, we used *Gini-Impurity* to get the most important features from the trained model classifier. By default, we are taking 6 most informative features. With such small feature-set it's accuracy is ranging between 98.5% and 99.70%. However, we can reduce the feature set to 2 and still get good accuracy beyond 90% and number of false positives and false negatives increase.

- **Linear SVM and Logistic Regression :-** Here, we select best features based on largest absolute value of coefficients learned by SVM/Logistic model classifier. Top coefficients are the largest entries of weight vector. We use this because if coefficient is larger then that feature is playing more important role in learning classifier.

Observations and Plots :-

We observed that doing PCA on the dataset actually increases the number of false positives and false negatives as compared to the Original full featured dataset. This may be due to the fact that PCA is an unsupervised learning method which doesn't account for the labels of the data while calculating principal components. We can choose to reduce the dimensions using Fischer's Discriminant Analysis which accounts for the labels during feature reduction.

Below are the plots on the features we selected using different classifiers.

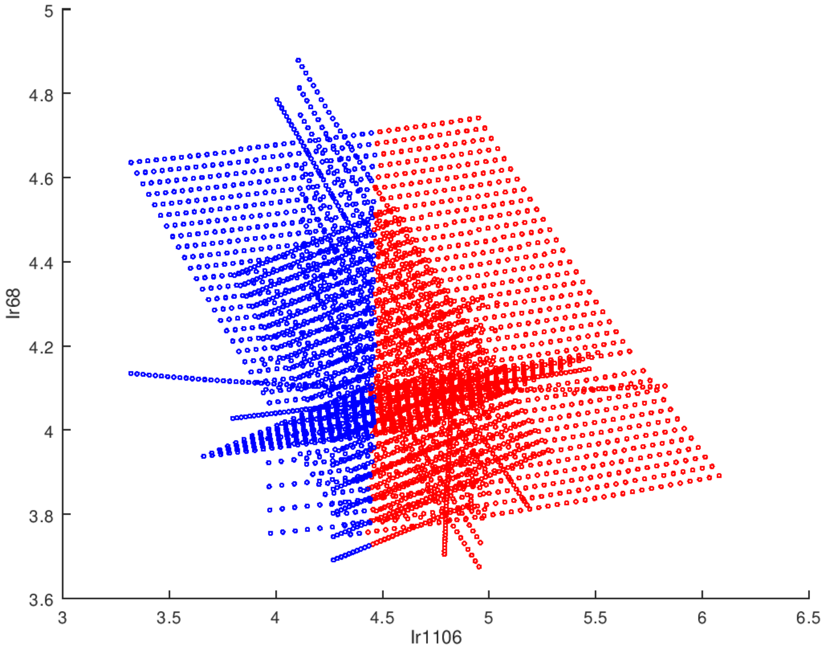

Figure 3: Most Important 2 Features using Decision Tree Model (for random shuffle)

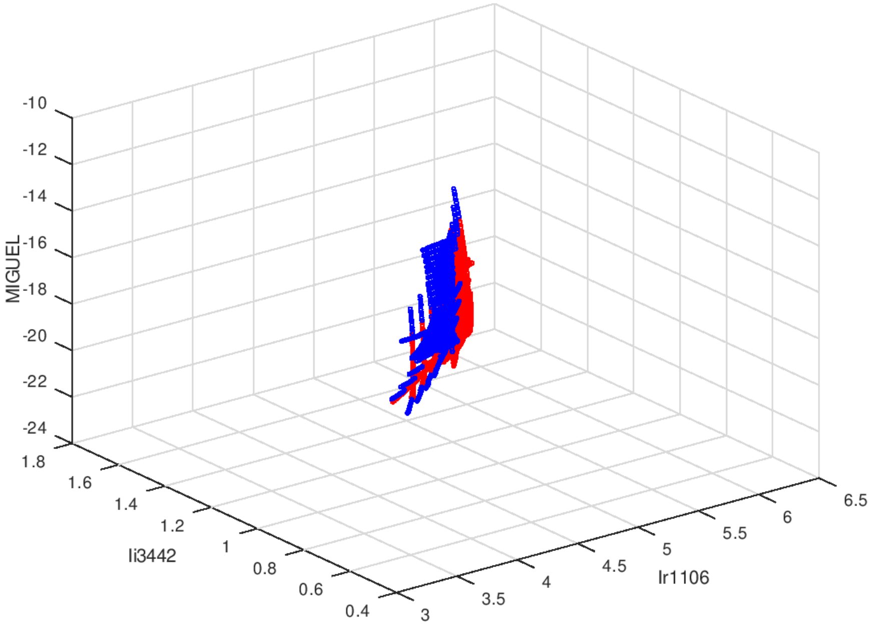

Figure 4: Most Important 3 Features using Decision Tree Model (for random shuffle)

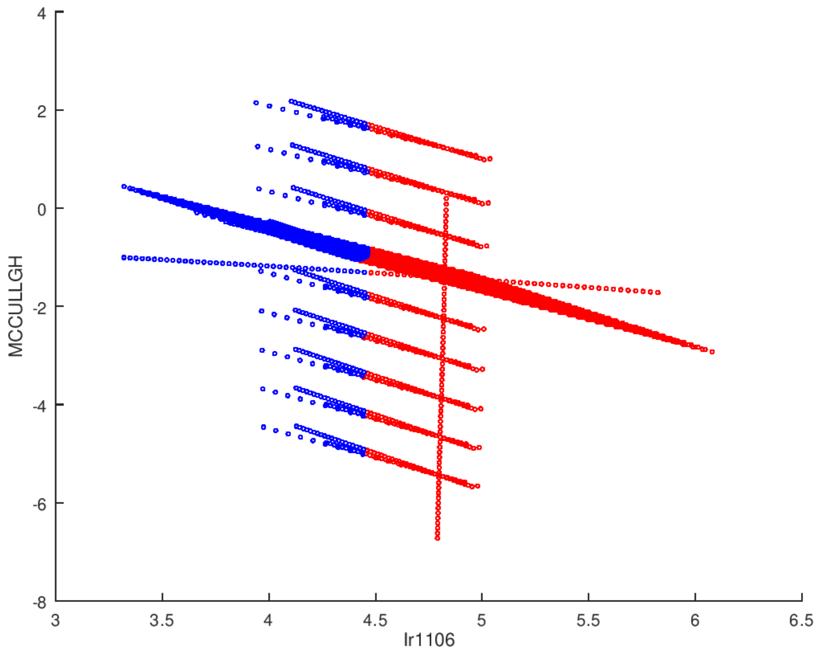

Figure 5: Most Important 2 Features using Linear SVM Model (for random shuffle)

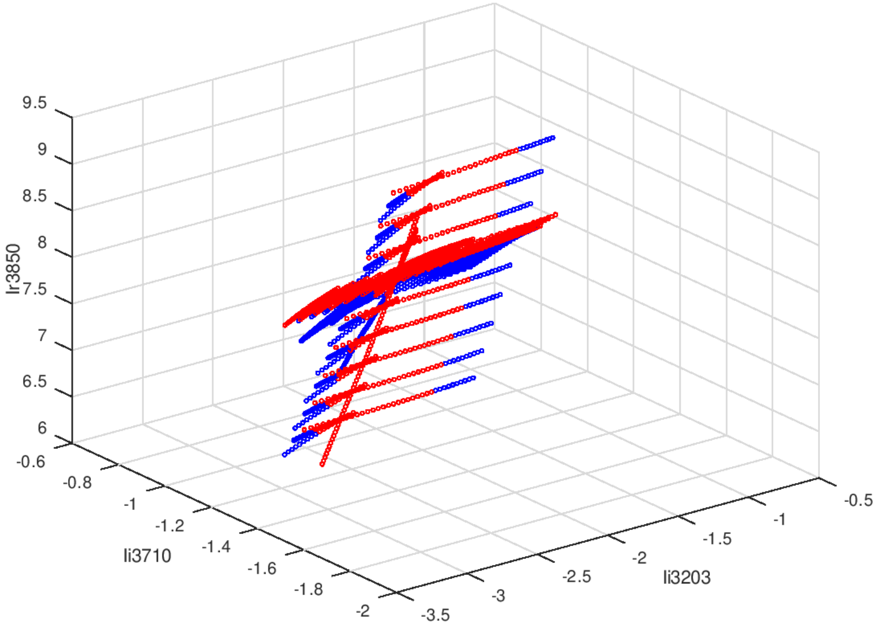

Figure 6: Most Important 3 Features using Linear SVM Model (for random shuffle)

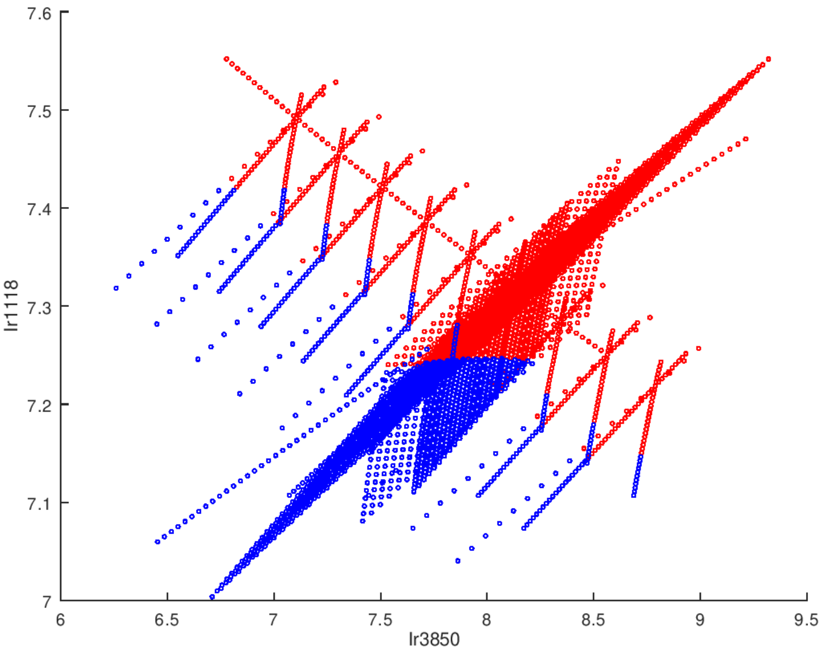

Figure 7: Most Important 2 Features using Logistic Model (for random shuffle)

We observed almost 90% accuracy using linear SVM on many pairwise selected features. We can't be sure of the importance of a subset of features selected as many data points become redundant when we consider only 2 or 3 dimensions.