EEE 591 : Project 1
Name: Siddharth Nilesh Joshi
ASU Id: 1217923356

**Project Report**

**Problem 1**

**Objective**: The problem 1 of the project deals with a statistical analysis of the each variable in the provided csv file. A conclusive report was generated based on the correlations of each of the variables: variance, skewness, curtosis, entropy and class.

This analysis is presented in the correlation plot (figure 2) and variable pair-plot (figure 3).Substantial conclusions are derived from the Correlation Plot as discussed further.

```
C:\Users\sidjo\AppData\Local\Programs\Python\Python37\python.exe "C:/Users/sidjo/OneDrive/Projects/Python_EEE591/Project 1/project1_A.py"
Here are the first 5 observations:
    variance  skewness  curtosis  entropy  class
0   3.62160    8.6661   -2.8073 -0.44699      0
1   4.54590    8.1674   -2.4586 -1.46210      0
2   3.86600   -2.6383    1.9242  0.10645      0
3   3.45660    9.5228   -4.0112 -3.59440      0
4   0.32924   -4.4552    4.5718 -0.98880      0

 Correlations:

      Var 1     Var 2  correlation
0  skewness  curtosis    -0.786895
1  variance     class    -0.724843
2  skewness   entropy    -0.526321
3  skewness     class    -0.444688
4  variance  curtosis    -0.380850
          variance     skewness     curtosis      entropy        class
count  1372.000000  1372.000000  1372.000000  1372.000000  1372.000000
mean      0.433735     1.922353     1.397627    -1.191657     0.444606
std       2.842763     5.869047     4.310030     2.101013     0.497103
min      -7.042100   -13.773100    -5.286100    -8.548200     0.000000
25%      -1.773000    -1.708200    -1.574975    -2.413450     0.000000
50%       0.496180     2.319650     0.616630    -0.586650     0.000000
75%       2.821475     6.814625     3.179250     0.394810     1.000000
max       6.824800    12.951600    17.927400     2.449500     1.000000

Process finished with exit code 0
```

**Fig 1: Output of Problem 1**

The figure above displays the first 5 observations, correlations and statistical analysis of the data in three distinct tables.

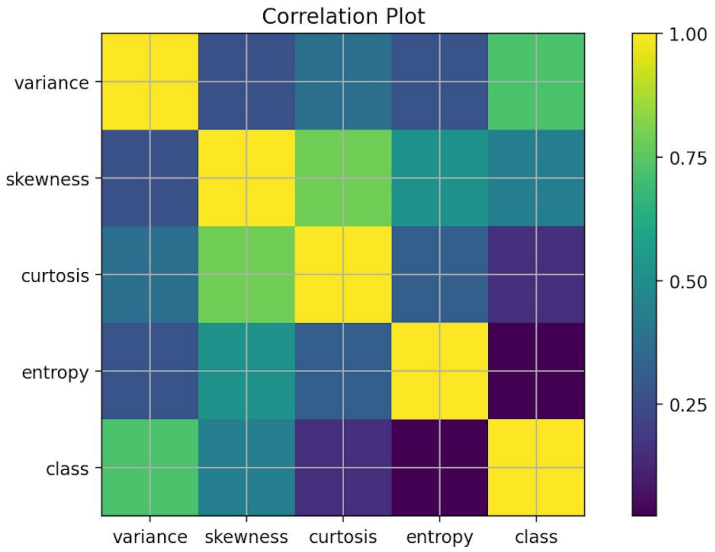


**Fig 2: Correlation Plot between variables**

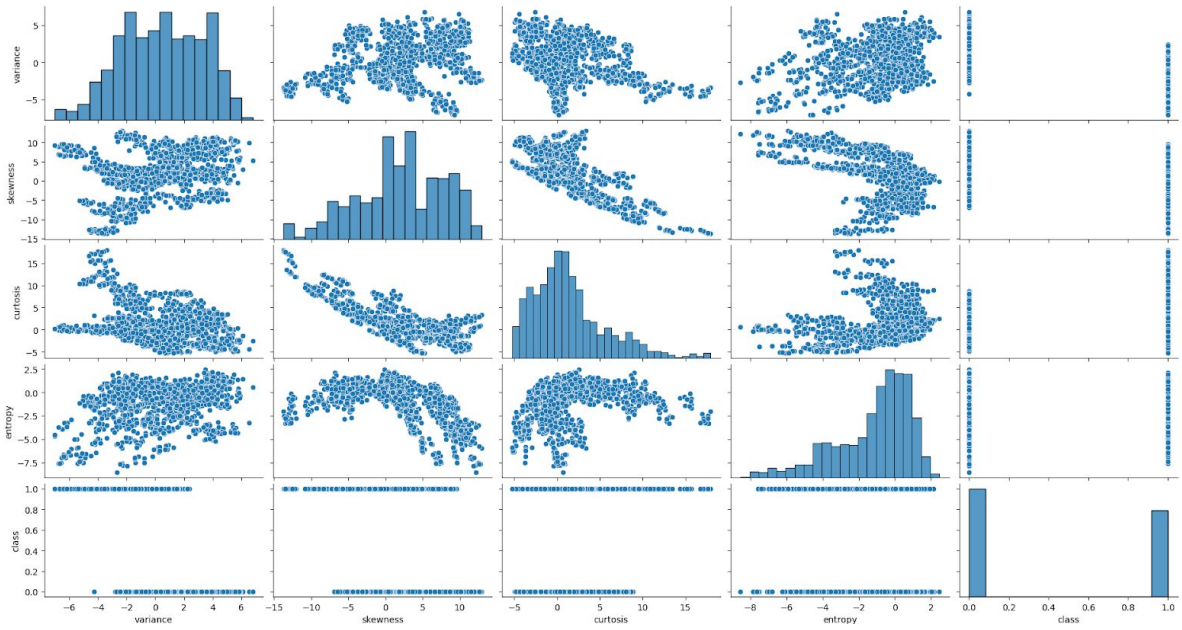


**Fig 3: Variable Pair-Plot**

As observed in the correlation plot, the colorbar on the right is the correlation measurement between variables, the highest being 1.00 along the diagonal (which is obvious). Observing in decreasing order, the optimal correlation-pairs are as follows: curtosis-skewness > variance-class > entropy-skewness > class-skewness > curtosis-variance > entropy-curtosis >= entropy-variance > skewness-variance > class-curtosis > class-entropy. Hence the highest

correlation is observed in case of curtosis-skewness, and the lowest in class-entropy. This proves that using variable curtosis, skewness (with high correlation) would result in an accurately trained model which can predict accurate test results. The class-entropy variables (with low correlation) would lead to a decrease in the model accuracy, with far inaccurate results.

**Problem 2:**

**Objective:** The following problem deals with the training and testing of models based on numerous machine learning techniques namely: perceptron, logistic regression, support vector machines, decision tree learning, random forests and K-nearest neighbors.

| **Algorithm** | **Incorrectly Classified Samples** | **Accuracy (*100 %)** | **Combined incorrectly classified samples** | **Combined (train & test) accuracy (*100 %)** |
|---|---|---|---|---|
| **Perceptron** | 8 | 0.9806 | 28 | 0.9796 |
| **Logistic Regression** | 5 | 0.9879 | 12 | 0.9913 |
| **Support Vector machine** | 6 | 0.9854 | 17 | 0.9876 |
| **Decision Tree Learning** | 14 | 0.9660 | 20 | 0.9854 |
| **Random Forests** | 4 | 0.9903 | 20 | 0.9956 |
| **K-nearest neighbors** | 8 | 0.9806 | 23 | 0.9832 |

The table above elicits the incorrect classification during training of the model as well as along with test data, accuracy in testing and combined accuracy in testing and training. As is it clearly observed, a random forests algorithm proves to be most accurate in predicting if the bill is counterfeit (with 4 mis-classifications). For each of the algorithms, the data provided was standardized through the “standardize” function for testing and training. Random forest hereby yields great results while avoiding overfitting of data. It is hence, highly suitable for classification and regression problems.