EEE 591 Project 1 Report

Problem 1

The most highly correlated columns are below:

Most Highly Correlated columns are:

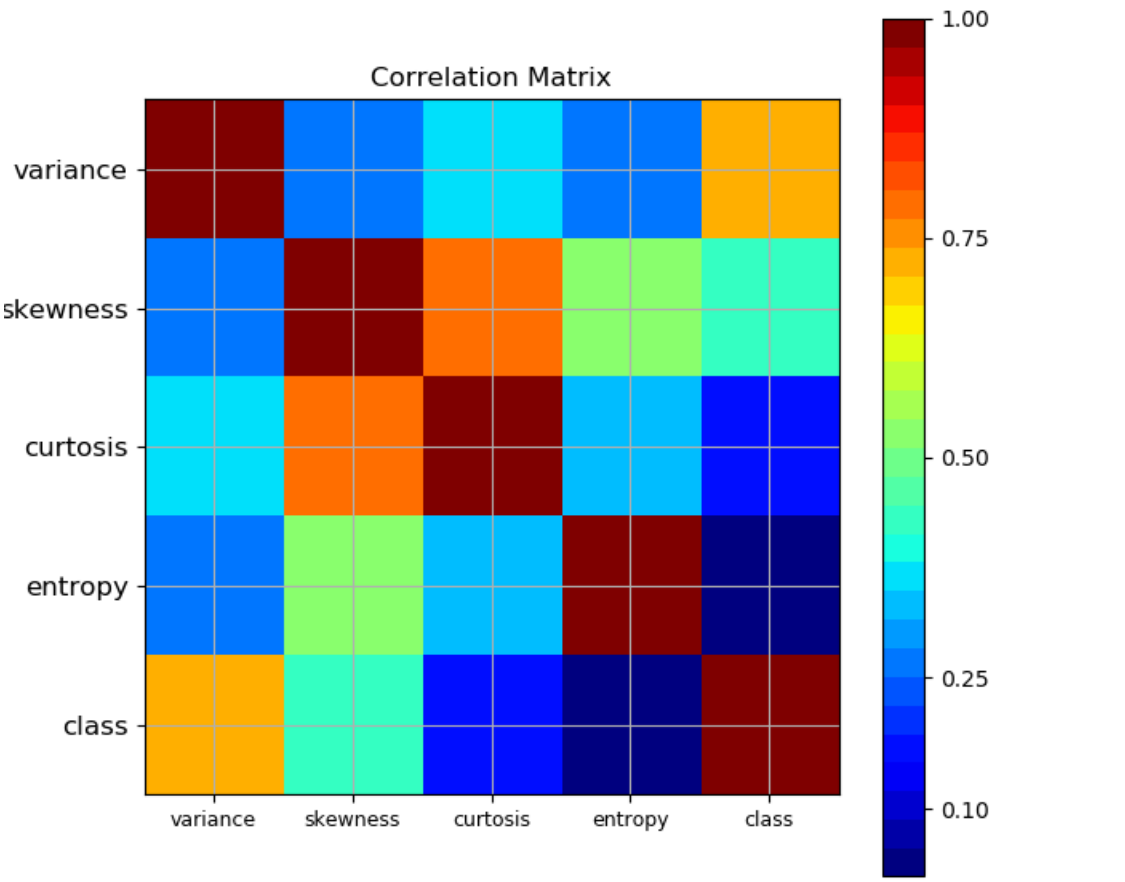
Most Highly Correlated

	FirstVariable	SecondVariable	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
5	curtosis	entropy	0.318841
6	variance	entropy	0.276817
7	variance	skewness	0.264026
8	curtosis	class	0.155883
9	entropy	class	-0.023424

The minus sign indicates that the 2 columns are highly correlated, but in opposite directions. When one decreases, the value of the other column increases. A positive correlation indicates that all of them are directly proportional.

The table above clearly shows that categories (columns that must be predicted) are highly correlated with variance and skewness.

Below is the plot of correlation matrix:



The columns with the highest correlation are skewness and curtosis (the correlation coefficient is -0.786895), followed by the variance and class columns. Skewness and entropy are also related to -0.526321.

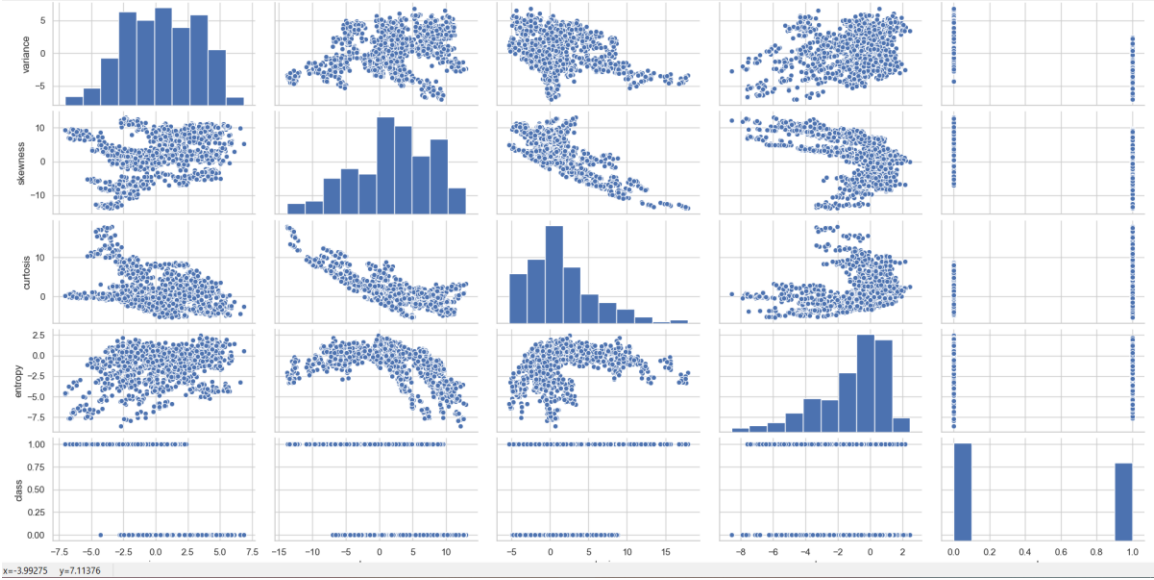
According to the analysis of the data set, the best predictors for real currencies will be the variance and skewness columns because they are highly correlated.

Below is the covariance matrix:

Covariance Matrix

	variance	skewness	curtosis	entropy	class
variance	8.081299	4.405083	-4.666323	1.653338	-1.024310
skewness	4.405083	34.445710	-19.905119	-6.490033	-1.297386
curtosis	-4.666323	-19.905119	18.576359	2.887241	0.333985
entropy	1.653338	-6.490033	2.887241	4.414256	-0.024464
class	-1.024310	-1.297386	0.333985	-0.024464	0.247112

Because the size of the covariance is not easy to normalize, it is not easy to explain, so it leads to the size of the variable, so it is recommended to refer to the correlation chart to understand the degree of correlation between the two columns. According to the covariance matrix, we can only get the positive covariance values of all the columns to indicate the minimum value of one column and the value of the other column.



Based on my analysis, variance and skewness are the columns which are going to impact the class column value (genuine bill) a lot as their correlation coefficient is high.

Problem 2

Following table shows different algorithms used for making the classifier along with their accuracy percentage:

ML Algorithms	Perceptron	Logistic Regression	SVM(kernel=rbf)	Decision Tree	Random Forest	K-NN (k=5)
Misclassified samples	7	5	0	8	4	1
Accuracy	0.98	0.99	1.00	0.98	0.99	1.00
Misclassified combined samples	28	12	0	10	4	2

Combined Accuracy	0.98	0.99	1.00	0.99	1.00	1.00

Simplified SVM (kernel = rbf) is the best classifier because it can correctly classify all samples. For the SVM classifier, the misclassified samples are 0, and for the K-NN, although the accuracy is 1, but 1 sample is misclassified. When the data points are completely different and can be linearly separated, SVM is the preferred classifier.