EEE 591 Project 1 Report

Introduction

The database for the project is in the CSV file format i.e. data_banknote_authentication.txt . This data set contains observations based on measurements made on a number of bills. The last column is whether the bill is genuine (1) or counterfeit (0). The dataset contains the following columns: 1. variance of Wavelet Transformed image (continuous) 2. skewness of Wavelet Transformed image (continuous) 3. curtosis of Wavelet Transformed image (continuous) 4. entropy of image (continuous) 5. class (integer). Based on the measurements, we need to build a predictor to determine whether a bill is genuine or counterfeit.

Problem 1

The most highly correlated columns are below:

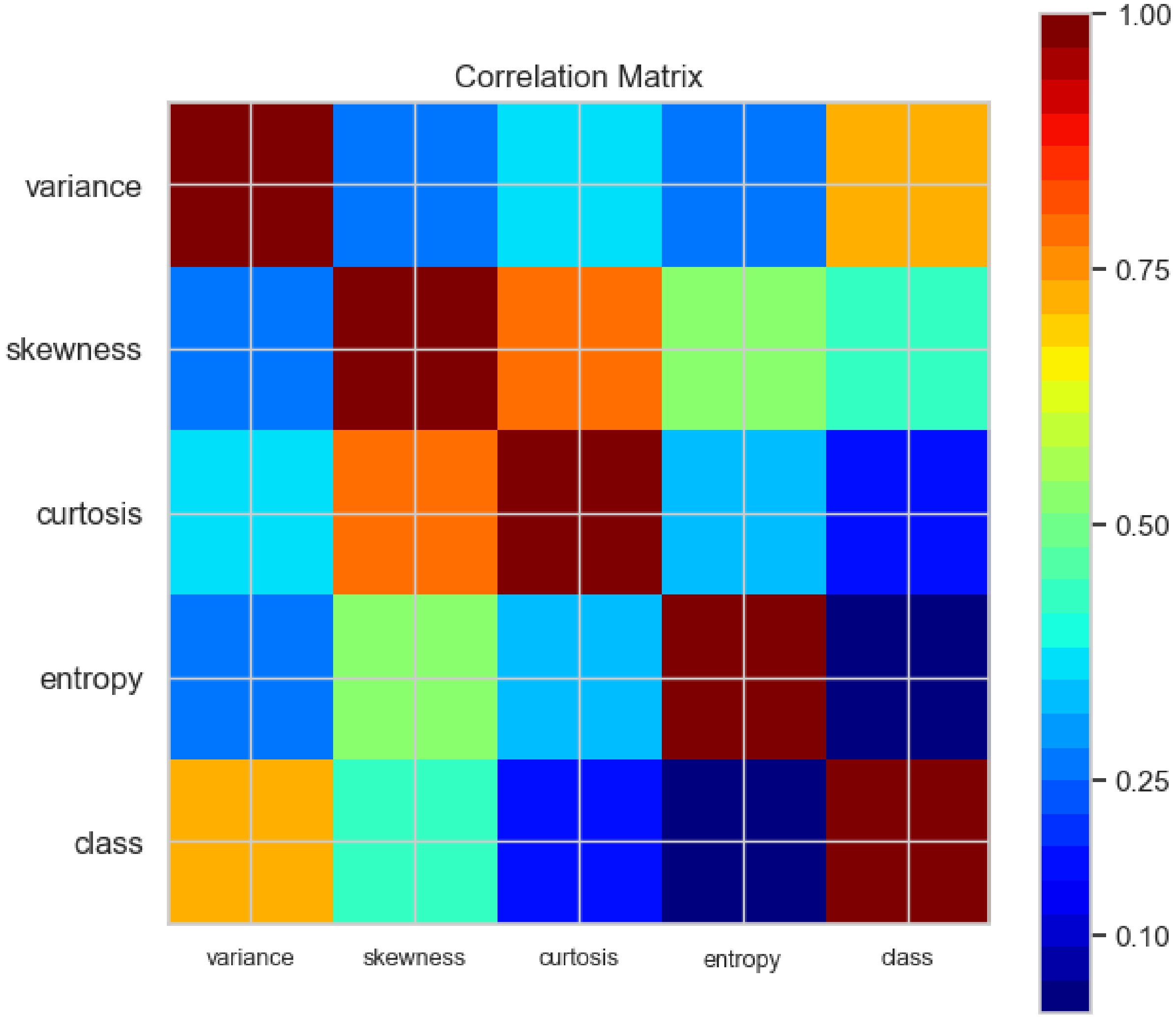
Most Highly Correlated columns are:

	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

Here, negative sign indicates that 2 columns are highly correlated but in the opposite way. When one decreases, the value of other column increases. Whereas the positive correlation indicates that both are directly proportional.

The above table clearly indicates that the class(column that has to be predicted) is highly correlated with variance and skewness.

Below is the plot of correlation matrix:



The most highly correlated columns are : skewness and curtosis with correlation coefficient of -0.786895, followed by variance and class columns. Skewness and entropy are also correlated with a factor of -0.526321.

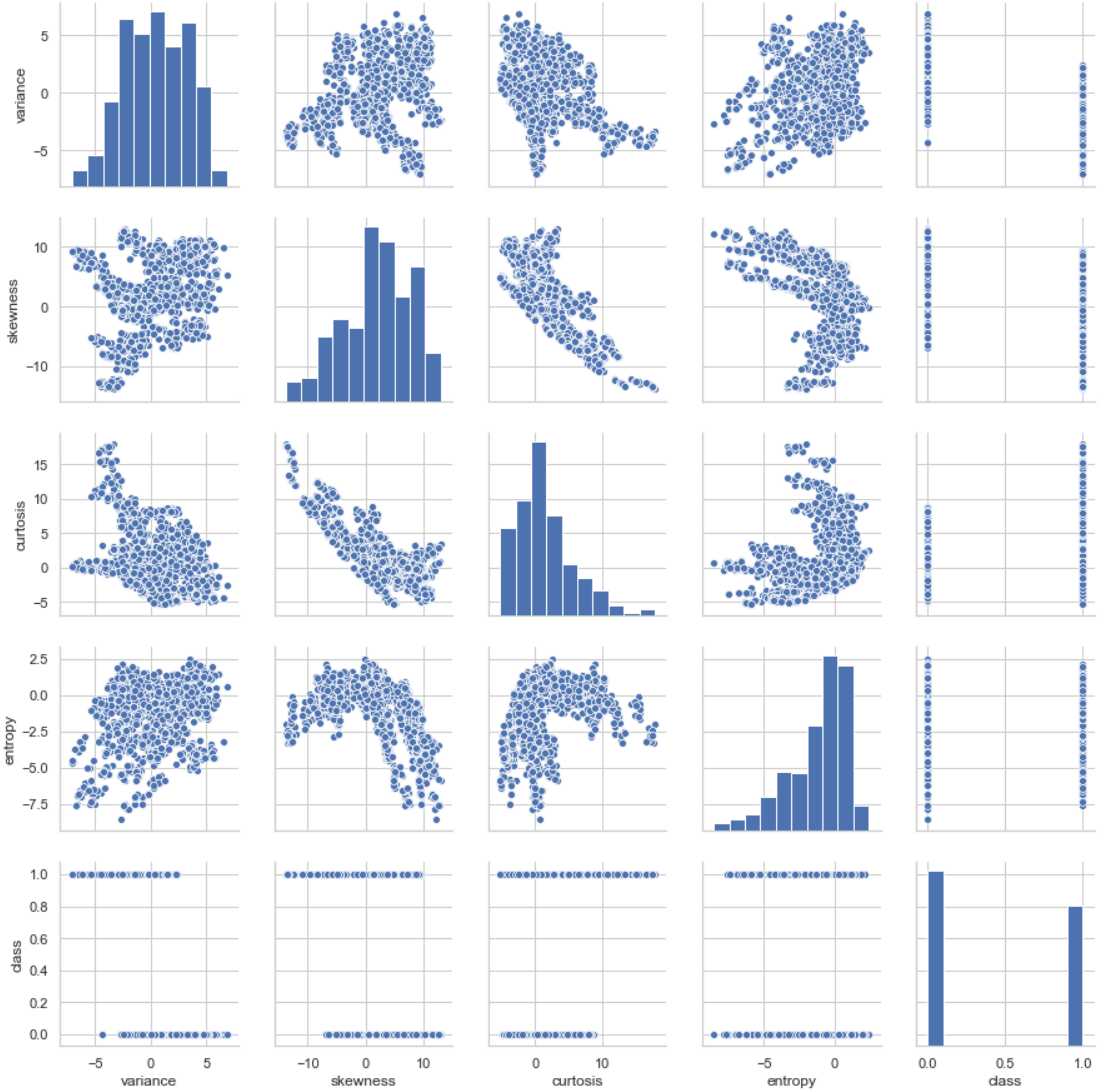
As per the analysis of the dataset, the best predictors of genuine money will be variance and skewness columns as they are highly correlated.

Below is the 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

Since the magnitude of the covariance is not easy to interpret because it is not normalized and hence depends on the magnitudes of the variables, it is more advisable to refer to correlation plot to understand how much two columns are related to each other. From covariance matrix, we can only conclude that all columns depend on each other. Positive covariance value means greater the value of one column, greater the value of the other column.

The pair plot is below:



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)
Combined Accuracy	0.98	0.99	1	0.99	1	1

In my opinion, SVM (kernel=rbf) is the best classifier as it classifies all the samples correctly. Misclassified samples is 0 in case of SVM classifier whereas for K-NN, though the accuracy is 1 but it has misclassified 1 sample. SVM is the preferred classifier when the data points are quite distinct and are linearly separable.