Banknote Authentication Using Unsupervised Machine Learning

Objective

One of the key responsibilities of central bank is to maintain confidence of country's currency. Hence, verifying whether a banknote is genuine or forged is one of important tasks in the finance industry, especially banks. For that reason, identification systems are developed, one is using machine learning algorithms such as unsupervised machine learning algorithms, especially clustering models.

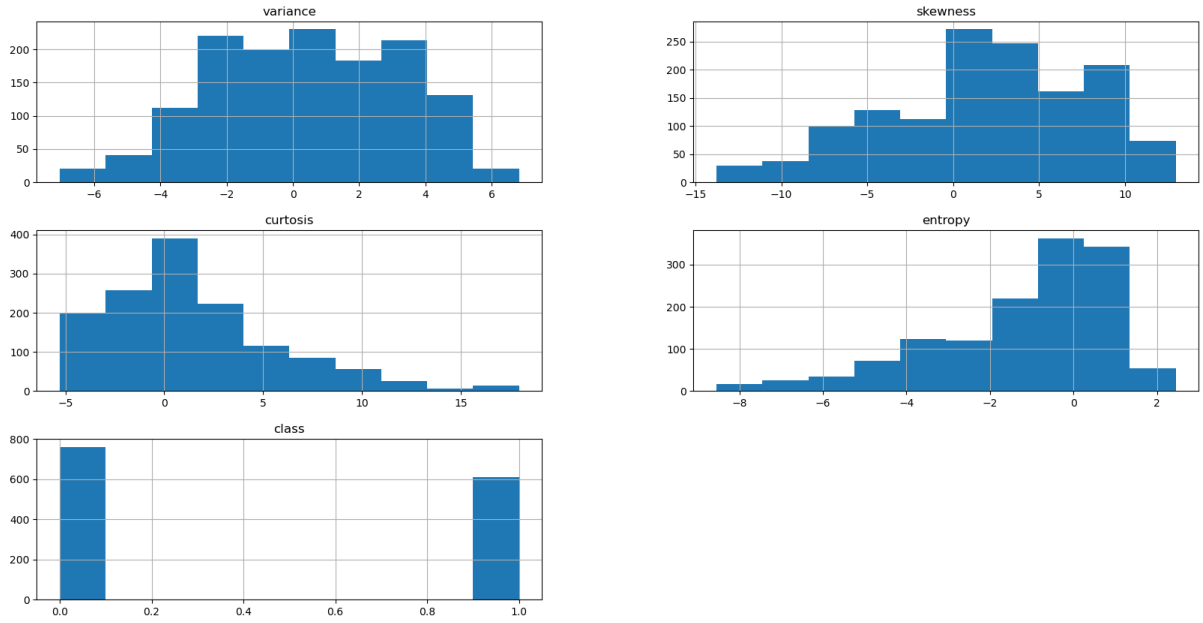
Dataset

This dataset originates from UCI ML repository and contains extracted data from images that were taken from genuine and forge banknote-like specimens. The digital images have 400 × 400 pixels and were taken using an industrial camera that is usually used for print inspection. Due to the object lens and distance to the investigated, object grayscale pictures with a resolution of about 660 dpi were gained. To extract features from those images, a Wavelet Transform tool was also used.

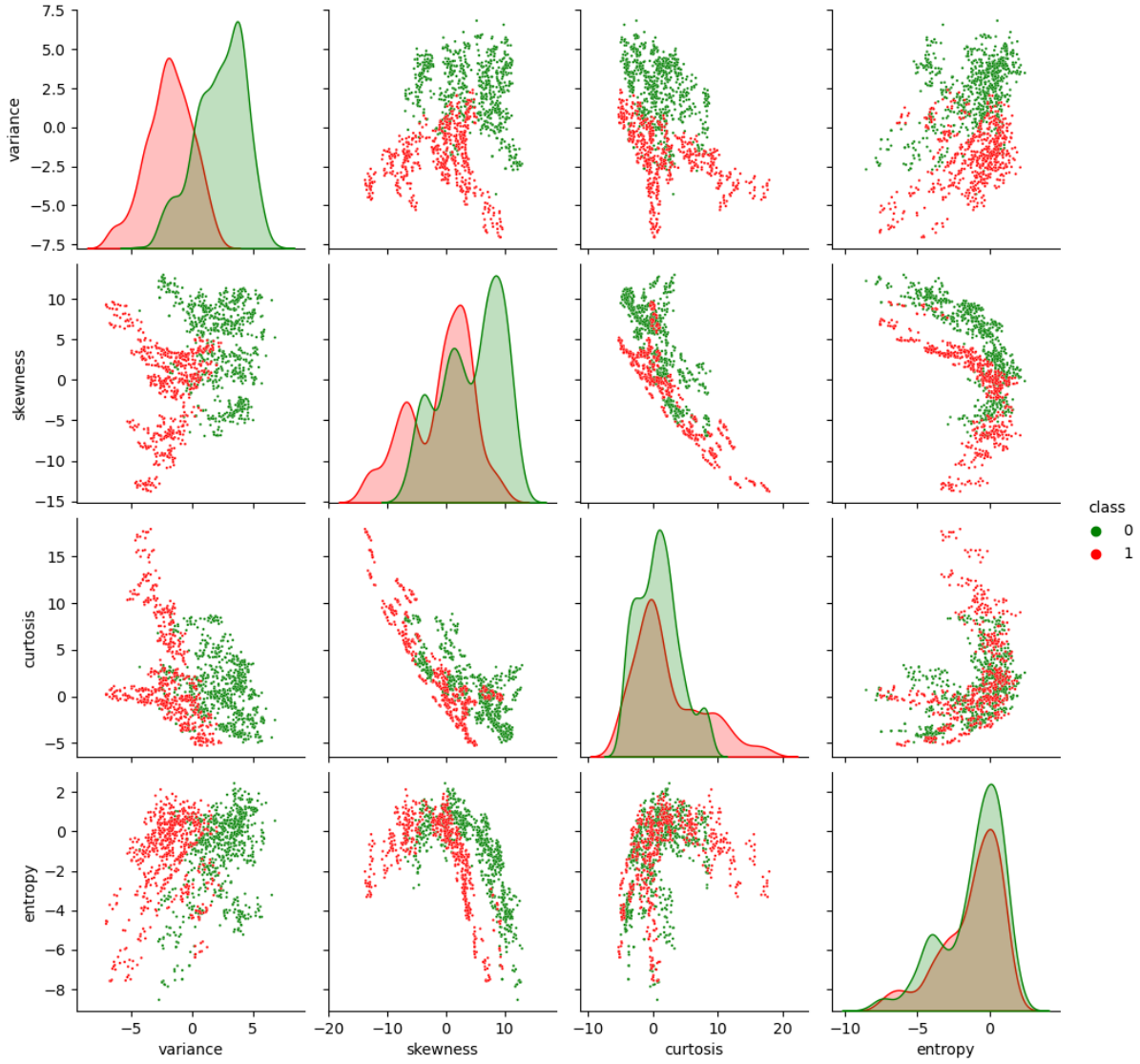
The data table contains four columns with parameters from wavelet analysis; variance, skewness, curtosis, and entropy. Following is the explanation for each parameter.

1. variance is the amount by which something changes or is different from something else.
2. skewness is the amount by which something changes or is different from something else.
3. curtosis refers to the pointedness of a peak in the distribution curve.
4. entropy is the measure of disorder or uncertainty.

The target column is class that is identifying whether the banknote is genuine (0) or forged (1). There are 1372 rows which are divided into 762 rows for genuine banknote and 610 rows for forged ones. For training models purpose, class was not included, but used as the ground truth in the validation process.



Variance seems to follow normal distribution. However, skewness and entropy are more distributed to the positive side, while curtosis is more skewed to the negative side, with maximum data points being near zero. We can also confirm that both classes are near balanced.


variance
skewness
curtosis
entropy
class
0
1


variance
skewness
curtosis
entropy
class
1
0.26
-0.38
0.28
-0.72
0.26
1
-0.79
-0.53
-0.44
-0.38
-0.79
1
0.32
0.16
0.28
-0.53
0.32
1
-0.023
-0.72
-0.44
0.16
-0.023
1
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

Looking at the scatter plot, we can see that both classes are distinct and separate. However, both curtosis and entropy seem to be non-linear correlated with each other. Also, the datapoints are least separable in both plots, so there is no reason to use these two variables together. At the heat map, we can see that variance, skewness, and entropy have negative correlation to target column, where variance is the highest, followed by skewness, and the last is entropy. Meanwhile, only curtosis has positive correlation to class.

Machine Learning Models

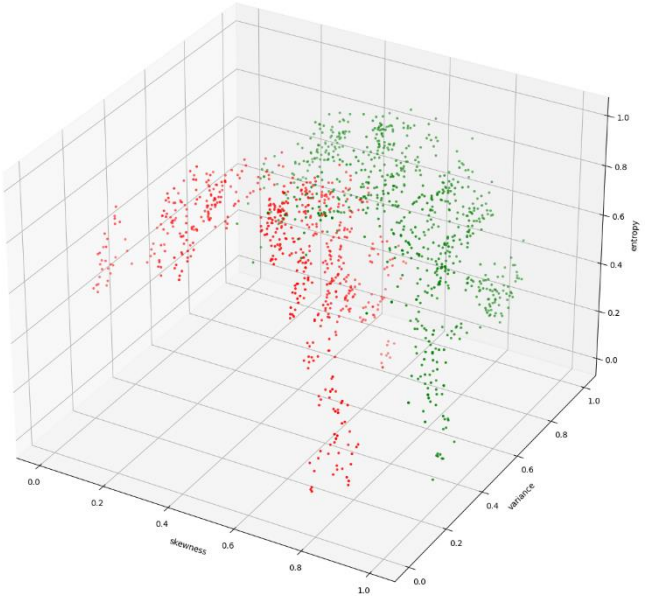
Before training the models, the dataset is scaled to the value between 0 and 1. Then, the analysis is divided into three different datasets. The first dataset includes variance, skewness, and curtosis. The second dataset includes variance, skewness, and entropy. The last dataset only includes variance and skewness. All these dataset are used in three different clustering models; KMeans, DBSCAN, and agglomerative clustering models. KMeans1, DBSCAN1, and Agglomerative1 are trained using first dataset. KMean2, DBSCAN2, and Agglomerative2 are trained using the second dataset. Last, KMeans3, DBSCAN3, and Agglomerative3 are trained using the third dataset. The following table shows the results of analysis.

	variance	skewness	curtosis	entropy	class	KMeans1	KMeans2	KMeans3	DBSCAN1	DBSCAN2	DBSCAN3	Agglomerative1	Agglomerative2	Agglomerative3
0	0.769004	0.839643	0.106783	0.736628	0.0	0	1	1	0	0	0	0	0	0
1	0.835659	0.820982	0.121804	0.644326	0.0	0	1	1	0	0	0	0	0	0
2	0.786629	0.416648	0.310608	0.786951	0.0	1	0	0	0	-1	1	0	1	1
3	0.757105	0.871699	0.054921	0.450440	0.0	0	1	1	0	-1	0	0	0	0
4	0.531578	0.348662	0.424662	0.687362	0.0	1	0	0	-1	-1	2	1	1	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1367	0.537124	0.565855	0.165249	0.726398	1.0	0	1	1	1	0	0	0	0	0
1368	0.407690	0.332868	0.506753	0.808350	1.0	1	0	0	-1	1	2	1	1	1
1369	0.237385	0.011768	0.985603	0.524755	1.0	1	0	0	-1	-1	-1	1	1	1
1370	0.250842	0.201701	0.761587	0.660675	1.0	1	0	0	-1	-1	2	1	1	1
1371	0.324528	0.490747	0.343348	0.885949	1.0	1	0	0	1	1	2	0	1	1

The findings are

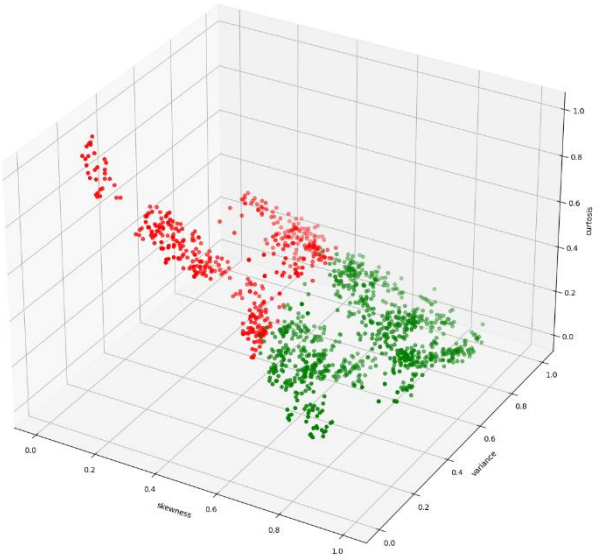
1. DBSCAN models are fail to cluster dataset into two.
2. Among the three datasets, the models that are trained using the first dataset are almost successfully having results that are almost identical to the real value in the target column.
3. By seeing the results above, it seems that Agglomerative1 result is better than Kmeans1 one.

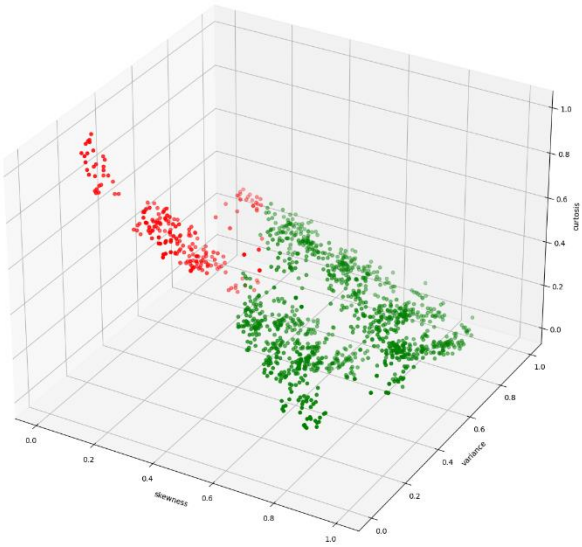
In order to see clearly, which one of this model is better in illustrating the real data, three plots are provided below, where the first one is the ground truth.



K-means

Agglomerative Cluster





By seeing above plots, it is clear that two models are not really perfect in matching the real data. It also seems that both models generate almost identical results.

For future analysis, since the unsupervised machine learning models are not able to catch the real data, it's better to using supervised model to predict whether a banknote is genuine or forged.