1. The task

Classify (cluster) the banknotes into one of the 2 classes - fake or genuine banknote - with the help of data science algorithms. The clustering has to be done upon the given features. Data set description follows below.

2. Data description

The given data set consists of 2 features: variance (V1) and skewness (V2) of banknote's images. These features are 2 of many that were mined with the help of Wavelet Transform – special tool which is used in machine learning for pictures processing. These two features are called to assist with the classifying banknotes into fake or genuine.

General description:

- Features: 2
- Quantity of examples: 1372

	Feature	
	Variance (V1)	**Skewness (V2)**
min value	-7.0421	-13.7731
max value	6.8248	12.9516

Table 1.1

With the purpose to reach the best performance and avoid mistakes, data was *normalized* (values were scaled to range from 0 to 1). And for the sake of initial data analysis, s.a. statistical data distribution, the following measurements on an already *normalized data set* were calculated:

	Feature	
	Variance (V1) normalized	**Skewness (V2) normalized**
standard deviation (std)	0.205	0.587
mean	0.539	0.220

Table 1.2

After all, the data san be presented as following:

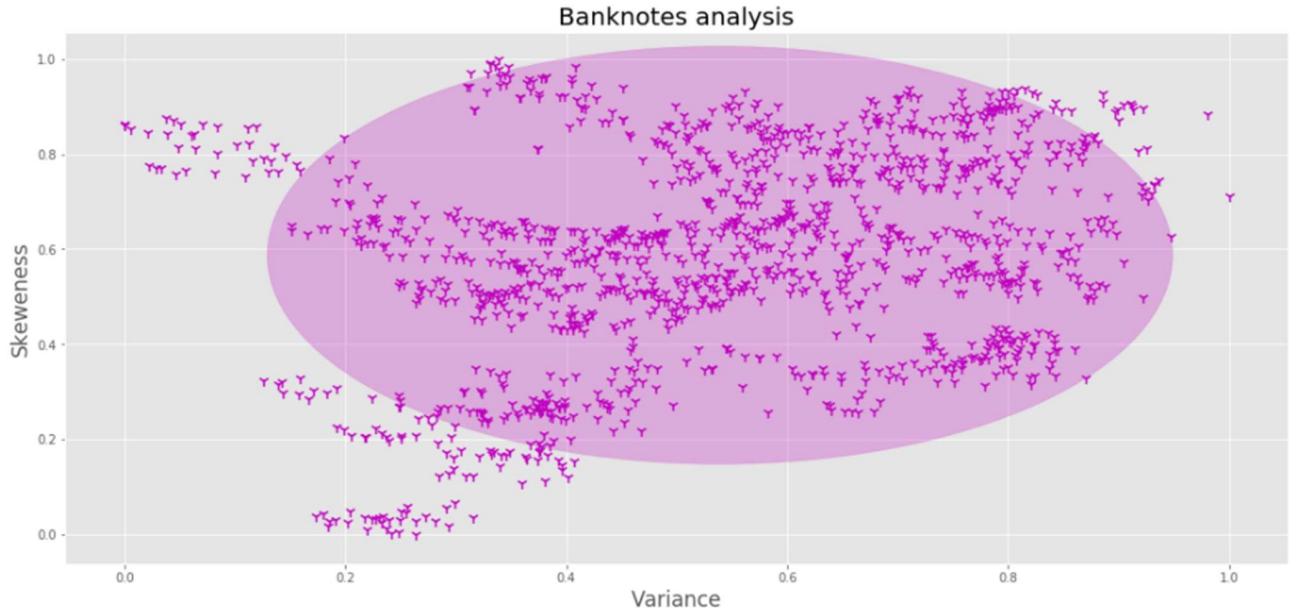

Chart 1.1

This chart undercovers the existence of outliers (data points out of the oval) and illustrates the statistical distribution. The oval figure is build in the following way:

- center coordinates –> *mean*
- radiuses r1 and r2 –> *std * 4*

3. Used method: K-means

As already mentioned above, our final task is to label the data set - assign classes to the given examples with the help of a clustering algorithm. Final aim – define whether the given banknote is fake or genuine, using only 2 features: skewness and variance of an banknotes' watermark image. For the mentioned task K-means algorithm was chosen. At the beginning stage it is hard to estimate whether the clustering algorithm would perform good on the given data set. At first glance there are several decision boundaries that may take place. Only running the algorithm certain number of times and further assessment of its performance will give the proper answer.

4. Results summary

As task was to label 2 classes (fake and genuine), number of principal components (parameter K or quantity of clusters) was set up as equal to 2 as well. K-means algorithm is so, that it cannot surely always converge to a global minima – sometimes only to a local one. Normally, each time when the algorithm runs different coordinates are chosen as cluster initial centroids. These initialization of these centroids plays the crucial role in finding for the K-means clustering algorithm in finding the best solution. Each time algorithm might converge at completely different points. However, in this particular case and the given data set, each run gives the same result - decision boundary appears to be nearly the same, as well as final cluster centroids (magenta crosses on the chart). Fake and genuine banknotes are highlighted with blue and cyan respectively (legend is added to the chart). Depending to what cluster centroid each data point (example of a training set) is closer, to hat cluster it will be assigned.

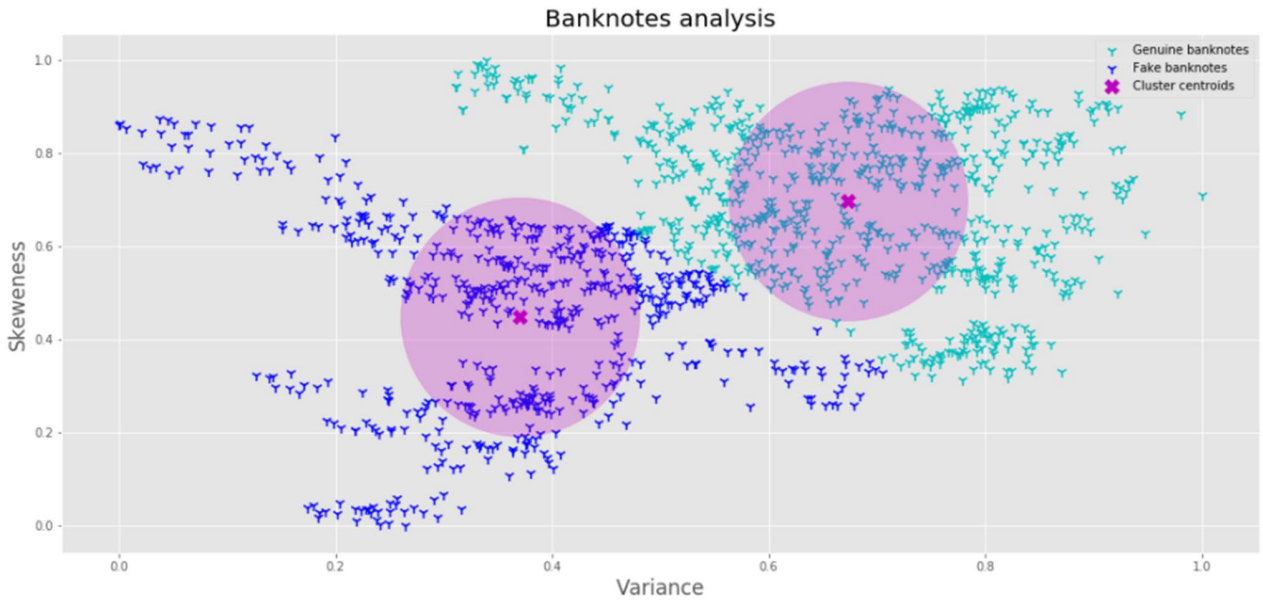

Chart 1.2

Result interpretation:

Banknotes with Variance less than 0.3 (no regard to the skewness value) are predicted to be fake. Whilst all the banknotes with image variance higher than 0.7 (again no regard to the skewness value) are highly likely to be genuine. In terms of the skewness feature: we can surely predict that banknote is genuine (if we would like to disregard the variance) only if it's value close enough to 1. The most problematic group of banknotes is that which has skewness around 0.5-0.7 meanwhile variance value is between 0.4 and 0.6.

5. Recommendations

In a conclusion to the result interpretation, I would recommend collecting more data examples of banknotes with standardized variance value between 0.4 and 0.6, and skewness value of 0.4 - 0.8. It means that generally more data may help to work on this problem closer and may help algorithm to perform better.