BANKNOTE AUTENTIFICATION

Work by: Rodolfo J.

Foundations of Data Science: K-Means Clustering in Python

Motivation.

The following work wants to predict whether if wavelets are genuine or forged, using just two variables (Variance and Skewness), and, in order to classify the data, we aim to employ K-means algorithm.

Data Description.

Original dataset consists on 5 variables:

1. V1. variance of Wavelet Transformed image (continuous)
2. V2. skewness of Wavelet Transformed image (continuous)
3. V3. curtosis of Wavelet Transformed image (continuous)
4. V4. entropy of image (continuous)
5. Class (target). Presumably 0 for genuine and 1 for forged

For purpose of the following project, a shorter dataset was provided; on it, only 2 variables were displayed: V1 and V2.

For both datasets, on each column, 1371 rows are displayed. No missing values.

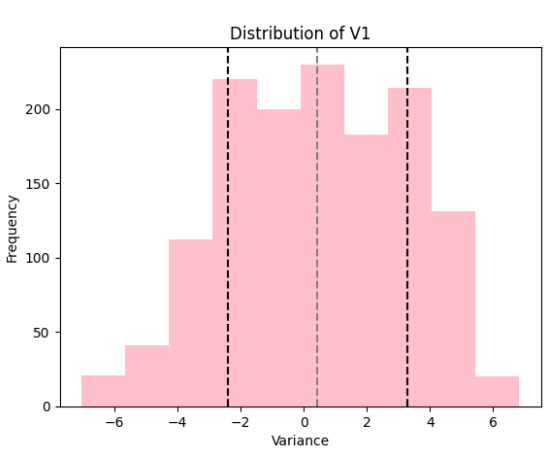
Analysis of the shortened dataset (V1 and V2)

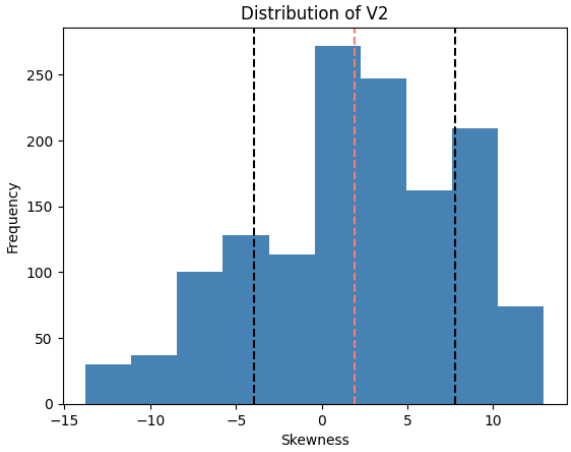
The shortened dataset only consists of two variables: **V1**(Variance of wavelet transformed image) and **V2**(Skewness of Wavelet Transformed image).

The following table will be displayed to describe the dataset, it includes (in order of appearance), the name of the variable, the range, and three statistical measures; mean, median and standard deviation.

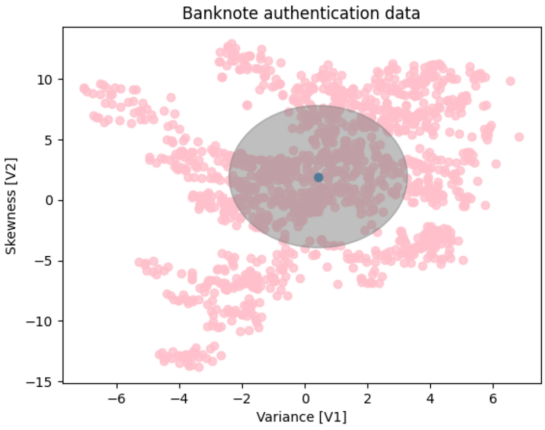
Variable	Description	Range (Inferior limit)	Range (Superior limit)	Mean	Median	Std
V1	Variance of wavelet transformed image	-7.0421	6.8248	0.433735	0.49618	2.841726
V2	Skewness of Wavelet Transformed image	-13.7731	12.9516	1.922353	2.31965	5.866907

To display the variables, two histograms will be displayed, alongside with the mean (in the center), and the standard deviation (dark dotted lines).





It is possible to proceed then to plot the two variables with their means and standard deviations.



In the last graph, we proceed to find out that only a few data fits standard parameters. However, and after having tried different ways to normalize data (I,e; logs and normalization), no significant effect in the distribution of data occurred. Only after running the "OLS" ordinary least square regression, a correlation error was found, explaining why it is not necessary to change the data distribution.

Methods:

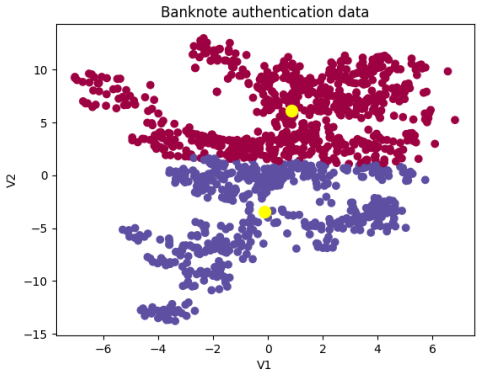
After having understood the composition of the data, there are reasons to believe K-means clustering will be able to predict the veracity of wavelets.

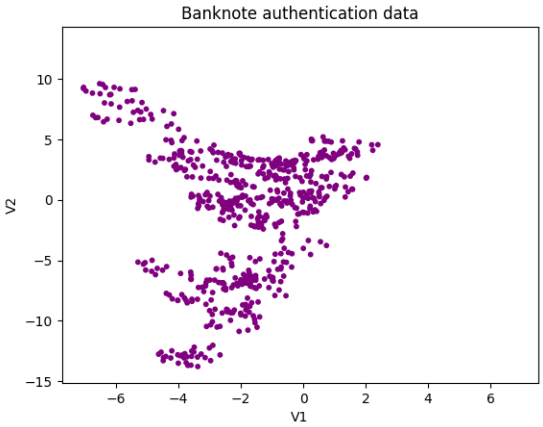
Here are the main reasons:

K-means is suitable for **large datasets**, since this clustering algorithm is relatively fast vis-à-vis other algorithms.
Second, as all our variables are **numeric**, K-means is perfect to cluster all our data.
Third, data can be separated by variance and skewness.
Although the algorithm assumes that clusters have a "spherical shape and similar variances", it will fit data beyond those borders.

K-means algorithm
This algorithm has the boundary of assigning each of the N observations into one of the K clusters, where each observation belongs to a cluster with the closest mean observation.
In this case, we chose to use two clusters: **Real (0)** and **False(1).**

In the graph of the left, it is plotted the prediction with the higher accuracy, while in the right, the results based on the original dataset.





Also, this is how several trials resulted **Real (0)** and **Falsified (1):**

Trial	1	2	3	4	5	6	7	8	9
0	773	780	775	589	773	798	589	599	783
1	599	592	597	783	599	574	783	773	589
Total	1372	1372	1372	1372	1372	1372	1372	1372	1372
Accuracy (%)	98.1967213	97.0491803	97.8688525	28.3606557	98.1967213	94.0983607	28.3606557	26.7213115	96.557377

Recommendations for the client:

The model is biased by the proximity of the variables vis-à-vis the mean. So, more variables are needed to prevent this bias.

However, at the moment of counting the number of predictions (i,e; how many wavelets have not been falsified), the model proves itself with a maximum value of 98% accuracy.

Finally, on practice, the client should be aware that the model, associates two characteristics to the **Falsified (1) cluster: a)** skewness below the mean (1.922353).