Identification of Forged Banknotes using K-Means Clustering Report

Purpose of the project:

Distinguishing the genuine banknotes from the forged ones using the features resulting from the wavelet analysis of the images of both types of banknotes: genuine and forged.

Description of the Data:

The dataset has 1372 data points with 2 features: the variance and skewness, resulting from the wavelet analysis of the images of the genuine and forged banknotes. The data is contained in the form of a CSV file with two columns of these features.

The main ***assumption*** that we make is that this sample of data points is large enough to capture the features of all genuine and forged banknotes.

Methods and limitations:

1. We use the pandas package to read the data of the two features variance and skewness from the CSV file.
2. We use numpy to calculate the statistical properties like the mean and the standard deviation of each of these features.
3. We use matplotlib.pyplot package to plot the two features against each other and visualize them and check visually if there are any sort of clusters visible in the figure. The visualization is shown in ***Figure 1***. We plot the means and standard deviations of the two clusters of the data. These ellipses are centered on the means of skewness and variances for real and counterfeit, with width and heights as 2 standard deviations. The significance of the 2 standard deviations is that they contain 95% of the data.
4. We use the K-Means clustering algorithm to classify the data points of the variance of the wavelets and skewness of the wavelets from the genuine and forged banknotes into two clusters one for the genuine banknotes and one for the forged ones.

The ***limitation*** is that for the 2 resulting clusters from the K-means clustering algorithm, there may be some overlap between the two clusters. However, overall the algorithm should give us a decent estimate of the true and forged banknotes.

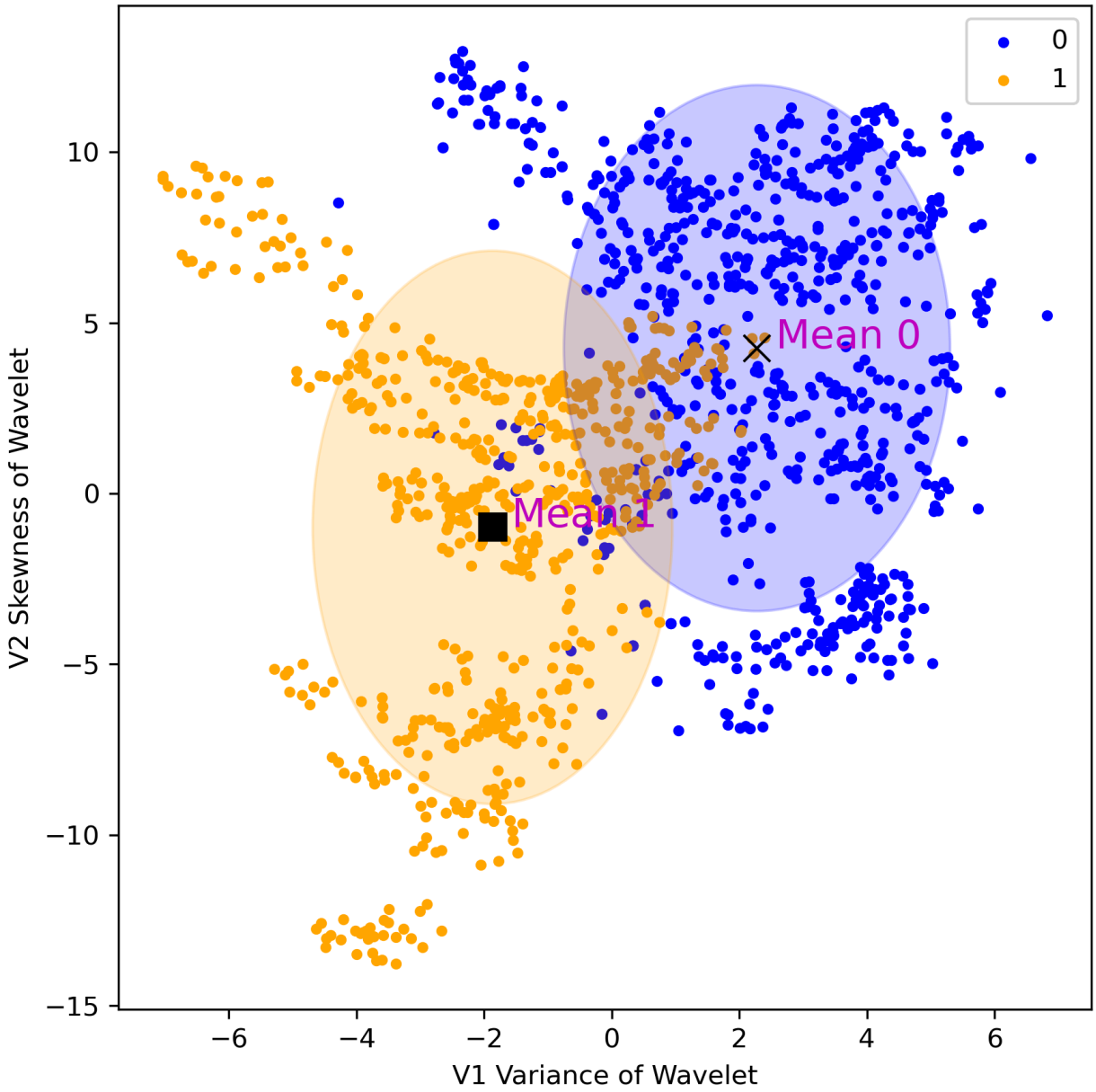

Figure 1: Skewness of the wavelet on the y-axis versus the variance of the wavelet on the x-axis. Blue data points represent the counterfeit banknotes. Orange data points represent the real banknotes. The blue ellipse is centered on the mean of the counterfeit with width and height equal to 2 standard deviations of the variance and skewness of the counterfeit. The orange ellipse is centered on the mean of the real banknotes with width and height equal to 2 standard deviations of the variance and skewness of the real banknotes.

Summary of the results:

After running the K-means clustering algorithm multiple times, we notice that the locations of the centers of the two clusters does not vary, which means that the algorithm is stable. These results are shown in figure 2.

We can classify the data points to genuine and forged banknotes based on the distance of each data point from each cluster center.

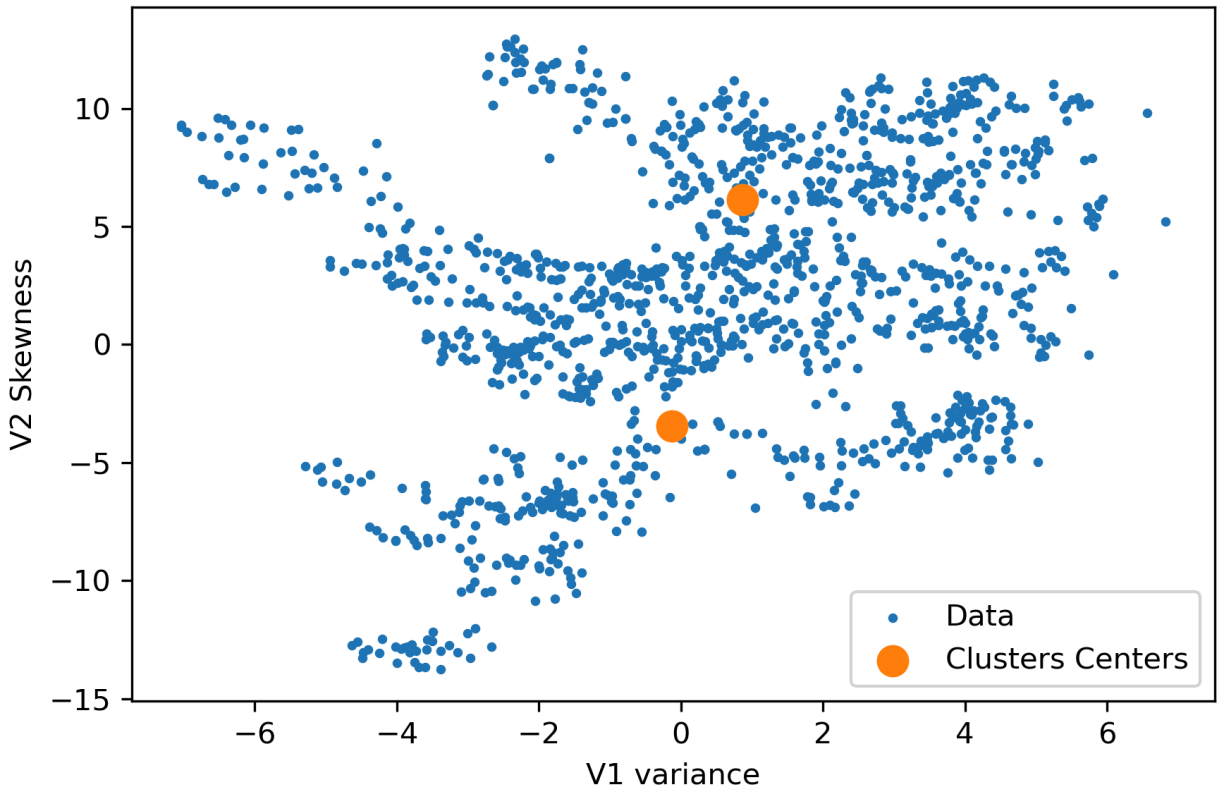

Figure 2: Skewness versus variance. After running the K-means clustering algorithm, the result is two centers represented as orange circles in the Figures. These two points are the centers of the two clusters. One for the genuine banknotes and the other for counterfeit banknotes. Based on that, we can classify the data points to counterfeit banknotes and genuine banknotes.

Suggestions to my client:

The client should implement the K-means clustering algorithm to detect the forged banknotes using the features from the wavelet analysis. It will give a decent estimate if the banknote is forged or not.