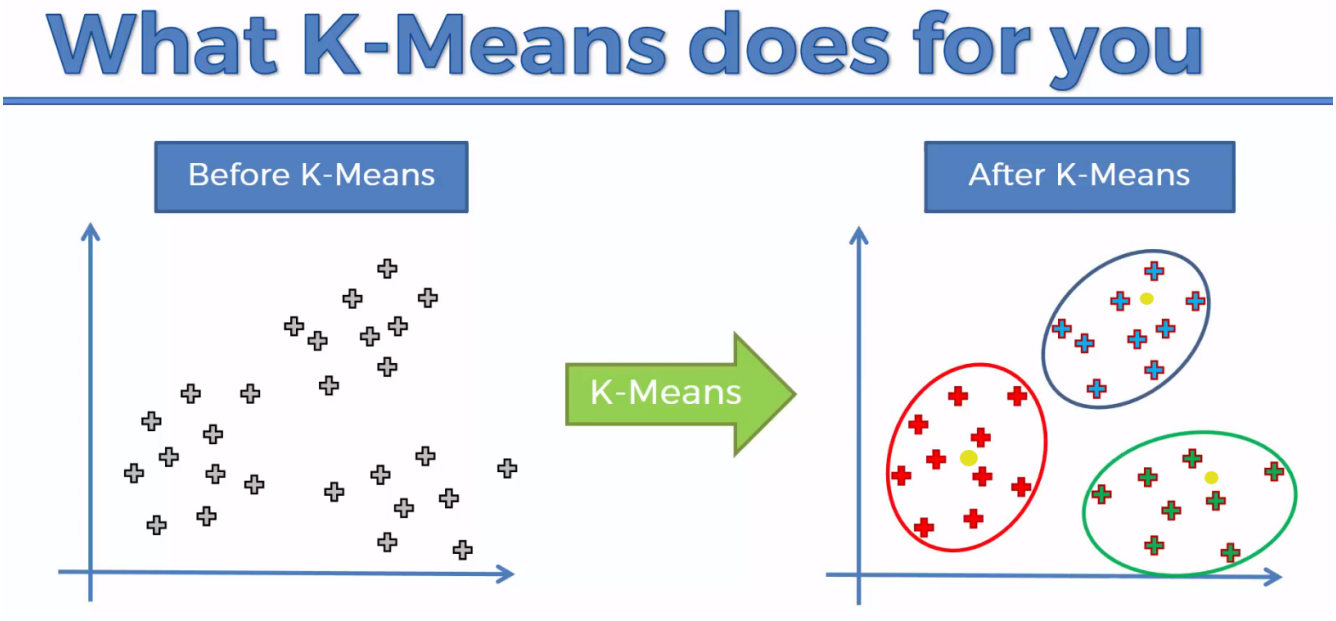
Forged banknote identification with Data Science.

Nowadays, Data Science is one of the most indispensable tools on the toolbox of any modern enterprise, that is due to his ability to bring fast and accurate solutions to complex problems; the problem of identifying forged banknotes is no exception. K-means offer us a good aproximation,
K-means is an algorithm that uses big amounts of data to identify features.

The K-means algorithm

With this algorithm every data element is represented with a point.Kmeans is pretty good at finding patterns. We only need to tell it how many features are we looking for. At first, for each feature, the algorithm generates points(as many as features we asked) placed randomly, these points are called our K-points or centroids. Then it calculates the mean(that's why is called K-means) distance from each data point to his centroid more closest, thus forming a cluster.

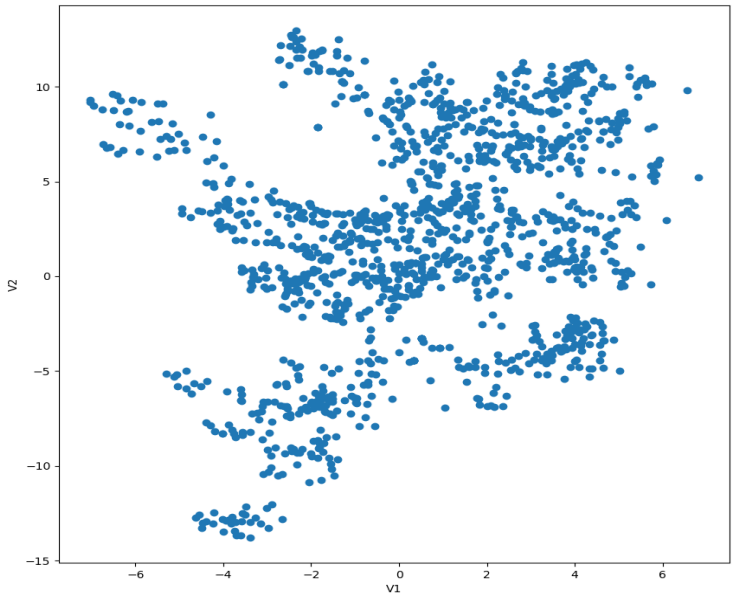


Source:https://github.com/MihailaDumitru/K-Means_Clustering/blob/master/README.md

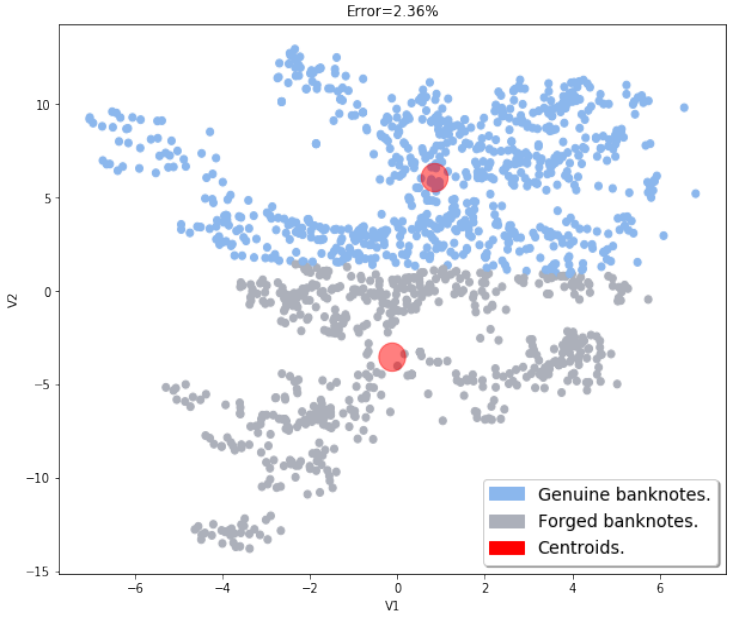
Finally Kmeans repeats this process with the points of each cluster, this to adjust the centroids(one for feature) in stable positions, making the calculation as accurate as possible. For instance, It's like we make a roadtrip in a distant place and we try to get a good phone signal. At first you can move randomly the phone to see when the bars goes up, then based on this new information we move again in that direction until we get the the connection working.

What does it mean for the proposite of our project?

K-means can took the data given by common security elements of banknotes, and with them determine wich ones are in fact genuine or forged. To display the aforementioned, we used a banknote authentication database: https://www.openml.org/d/1462 wich consists of data extracted from images that were taken from genuine and forged banknote-like specimens. Taking for example two columns of features, V1 and V2 ,we can get the following figure:



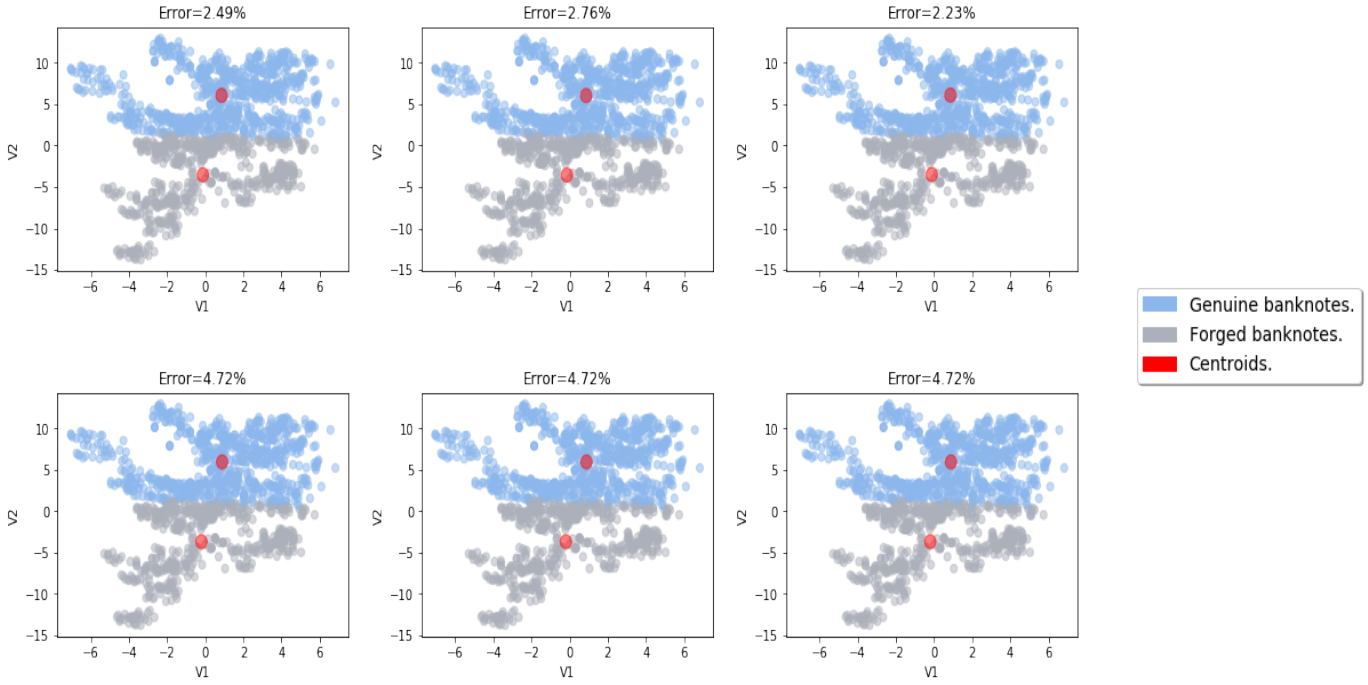
There's no distintion of genuine or forged banknotes, just the scores are plotted on this canvas. Running K-means with the instruction to find two features (genuine and forged), resulting in the following:



Two groups clearly emerged given by the distances of the two centroids we established. So we can label those groups by the amount of genuine and forged banknotes given by our dataset.

How about the reliability of the results?

For this database, we know the quantity of genuine and forged notes, those can be compared to the quantities given by the K-means algorithm to calculate the error percentage.
In the following plots, the K-means algorithm was ran six times, and then compared to the real quantities of genuine and forged notes.



Summary

As we can see in the last figure, the error doesn't surpases the 5%. Hence the method has an accuracy of near 95% to distinguish between forged and genuine banknotes.

Recommendations

The K-means method is widely recomended to clasify big amounts of data, to obtain useful information for us and give the possibility to enhance our security systems.