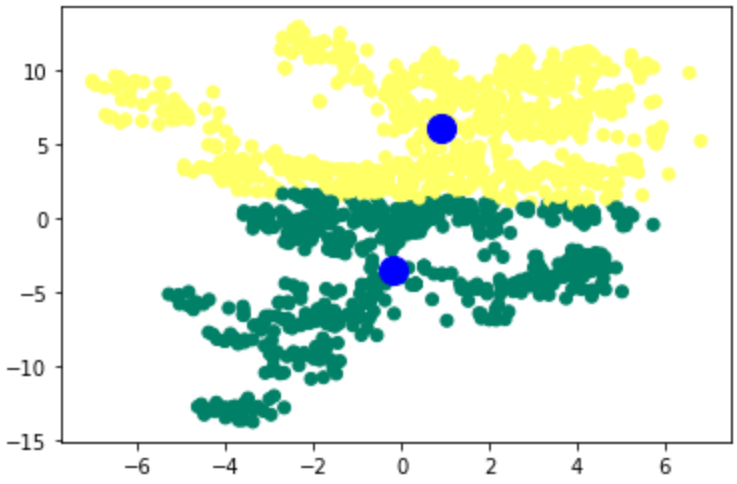
Foundations of Data Science: K-Means Clustering in Python - Week 5 - Coursera

This report is in regards to a data analysis conducted to determine the accuracy and subsequent recommendation or condemnation of a particular automation to differentiate between genuine and forged banknotes.

The dataset used was a simplified version of the one available here: OpenML banknote-authentication. The original dataset was created based on results of variance, skewness, kurtosis and entropy (4 variables). These were condensed into 2 variables for a simplified dataset, and this was the dataset used for the current analysis. The simplified dataset only made use of variance and skewness scores. It is recommended to conduct future data analysis on the automation's ability to differentiate on the other values.

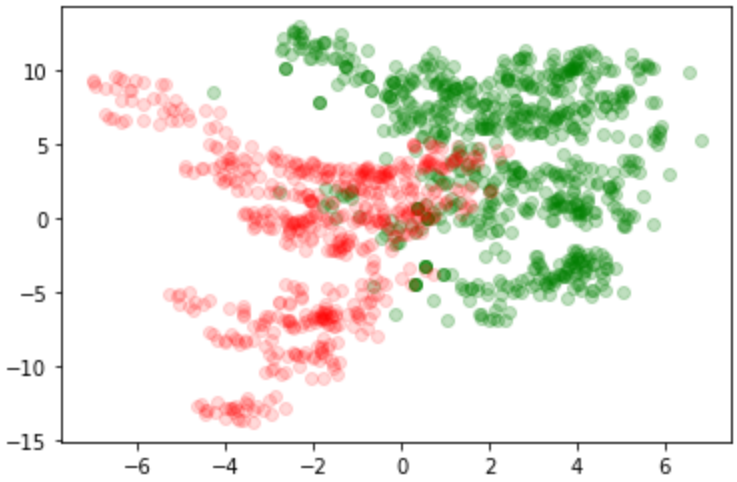
Data were analysed using K-means clustering. ***K*-means clustering** is a method of vector quantization, which partitions a dense amount of observations into "clusters". Repeated K-means are needed for most accurate results. The number of clusters chosen to be discovered is up to the analyst, and for this report, two clusters were chosen. This was so that it was possible to compare the two graphs, one which showed genuine and forged banknotes, and the other which showed the clusters identified by the program.

In the present data analysis, K-means was conducted 10 times, with results always falling in similar ranges. Thus, finally, the results of one K-means was plotted on a graph:

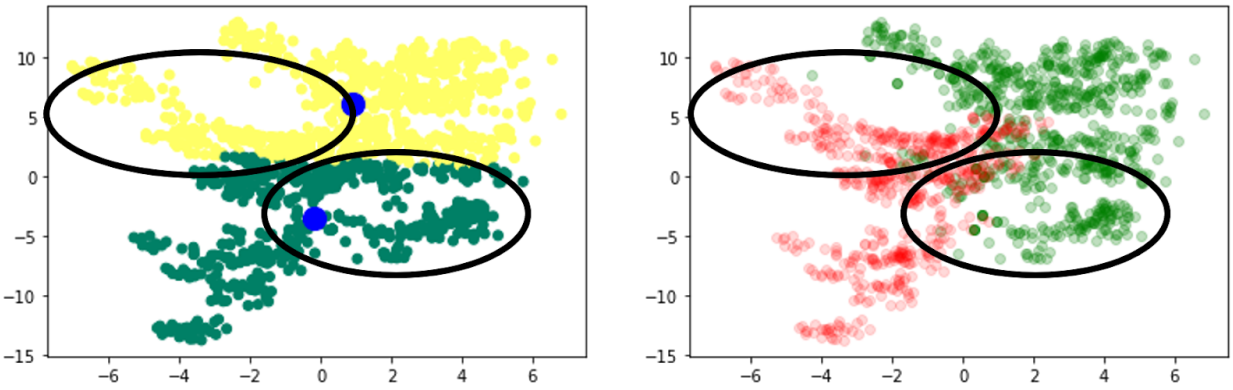


The blue represents cluster points identified in the latest K-means analysis, and the green and yellow represents the different clusters.

The following graph shows which banknotes were genuine in green, and which were forged in red.



Below is a comparison of the graphs. Clearly, the clusters need some work. Continuing with the current automation technique and basing results on only the two variables of variance and skewness scores, there are likely to be some false positives and false negatives, as the circled parts of the graphs show.



Results may differ if the other variables of kurtosis and entropy are entered into the mix, and if the KMeans algorithm is repeated multiple times until its clusters are closer to the second graph.

In conclusion, automation is definitely recommended, but the automation that was used for this analysis is not there yet. But this may be because of the use of only two of four variables. Further data analysis will be conducted on the raining variables. Recommendations are suggested below.

Recommendations:

- More data, i.e more input of known genuine and forged banknotes.
- Use of other variables not included in the simplified dataset

Hawra Ali - https://github.com/Feline17

- Repeating the KMeans clustering algorithm multiple times until cluster points closer to the genuine and forged banknotes graphs are found