**Automating the detection of forged bank notes.**

**Purpose of the Project:**

Due to feasibility, flexibility and currency notes being the source of exchange, banks have become a center of handling the money. The number of transactions and work has been dramatically increasing every year. But the problem of forging is still in practice and as the transactions are been increased, we cannot rely on manual inspection and other inspection techniques as they do take a lot amount of time and energy. In these recent years there has been a lot of development in the field of computer vision and faster computing and machine learning, we do now have access to programs that can classify the forged and legal bank notes in a short duration of time and high accuracy. We have used a Machine Learning algorithm that can classify between real and forged bank notes by the technique of clustering. To show how this process works a real dataset with 1000+ observations of bank notes has been used below.

**Description of the data:**

The banknote authentication dataset is about distinguishing genuine and forged banknotes.

To produce this dataset images were taken from both genuine and forged banknotes. Then a special tool, Wavelet Transform, was used to extract features from these images. These features are: variance, skewness, kurtosis, and entropy of the images. The entire dataset is completely numeric i.e., the data has no text in the columns the features extracted are represented by the magnitude in number notation. The important thing to know is that these features capture the necessary information about the images. The banknote authentication dataset has 1,372 instances and no missing values. We will use a simplified version of this dataset that has only two \*features.

\*Feature: a column which is a variable that is input to a system. for example, in the bank authentication problem above we had four features variance, skewness, kurtosis and entropy each column having their own values related to its property.



Figure:1 A sample of dataset with two features named V1 and V2 and the extracted feature details in the form of number.

**Methods: how the data was analyzed;**

When we look at some data specially the numeric, a statistician or an analyst would like to see a plot of how the data is i.e., where the center of data is and how close or in simple terms how the numbers are with respect to one other as this helps to choose or guess which technique would work to solve our issue. A picture of the center of data is being shown below from our dataset and we do see that the samples are relatively close to one another. To obtain the accuracy and improve the efficiency we reduce the size of numbers by a technique called normalization. The reason for doing this is we want two clusters one representing the forged notes and one representing the real notes and there is strong evidence that reducing the limits makes it easy for the CPU to work well and find the clusters in no time with accuracy.

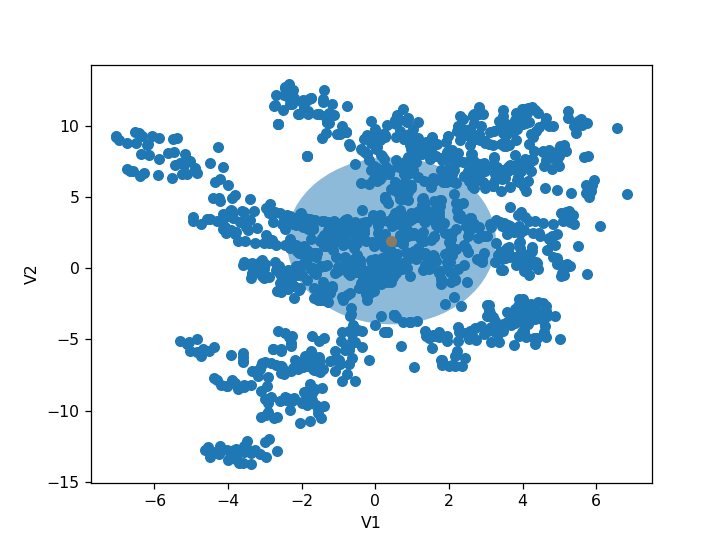


Figure:2 Plotting V1 V/S V2 we see the data isn’t spread out and orange color showing the center of data.

The technique we are going to use is K-Means Clustering, “the objective of K-means is simple: group similar data points together and discover underlying patterns. To achieve this objective, K-means looks for a fixed number (k) of clusters in a dataset.” As we need just two clusters one for the forged bank notes and other for the real ones the number of clusters will be two. A cluster refers to a collection of data points aggregated together because of certain similarities.

You’ll define a target number k, which refers to the number of centroids you need in the dataset. A centroid is the imaginary or real location representing the center of the cluster.

Every data point is allocated to each of the clusters through reducing the in-cluster sum of squares.

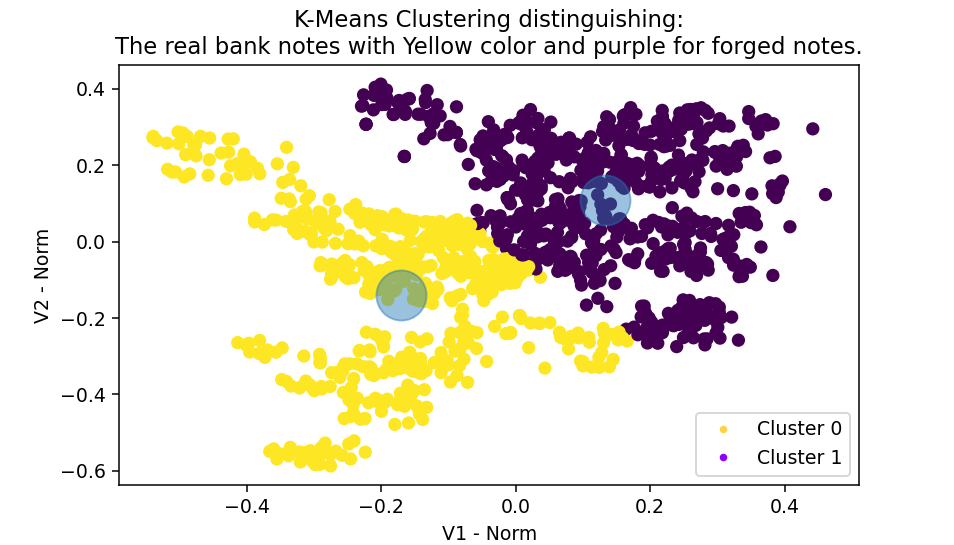
In other words, the K-means algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible.

The ‘means’ in the K-means refers to averaging of the data; that is, finding the centroid.

**Summary of results:**

After running the K-Means clustering technique we have got two clusters with there centers as

(cluster1(x, y), cluster 2(x, y)) = ([[0.13449765,0.11031867], [-0.17000858, -0.13944571]])



**Recommendations for the client:**

We recommend using high quality data first.

Though K-Means is very good clustering algorithm, the working principle of it is more biased towards variance so a small change in variance can affect the cluster patterns.

We highly recommend using large data samples for training the machine and improving some features will do a better help in increasing the accuracy of the model so it can be deployed into the real world for use.