Fake Banknotes Clustering

# Introduction

In this report, we aim to solve the important issue of identifying counterfeit banknotes. Detecting fake currency is crucial for maintaining financial security and preventing fraud.

Our objective is to develop an accurate clustering model that distinguishes between real and fake banknotes. Using the provided dataset, we have identified key features, particularly **variance** and **skewness**, that are most effective for this classification task.

We will preprocess the data to ensure accuracy and apply the **K-Means clustering algorithm**, leveraging the **Euclidean distance function** to group the banknotes. This report will focus on presenting the outcomes of our analysis along with key instructions and insights for clients to easily interpret the results.

## Dataset

The dataset is taken from [OpenLM **banknote-authentication**](https://www.openml.org/search?type=data&sort=runs&id=1462&status=active) dataset, it is about distinguishing genuine and forged banknotes. Data was extracted from images that were captured by an industrial camera usually used for print inspection. All of the images have been converted into 4 features: V1(Variance), V2(Skewness), V3(Kurtosis), and V4(Entropy). Additionally, the uploaded dataset included the expected results (which are the real results that already exist). Hence, in this area, we will compare with the expected results to get the accuracy of our machine.

In this section, we will discuss the significance of each feature in the dataset and determine which features best suit our needs. We aim to avoid including too many features, which could introduce “noise” and lead to incorrect or inaccurate results.

* **Variance**: Measures the spread of pixel values, highlighting differences in the texture between real and fake notes.
* **Skewness**: Captures the uneven distribution of pixels, potentially revealing unique printing defects or patterns in fake notes.
* **Kurtosis**: Reflects how sharp or smooth the edges in the image are, which may help distinguish distinct features in real or fake notes.
* **Entropy**: Assesses the randomness in the image. Fake notes may exhibit higher randomness due to inconsistent printing.

Although **kurtosis** and **entropy** can contribute to identifying fake notes, modern advancements in printing may result in fake and real notes having similar edge sharpness and randomness. Therefore, we focus on the more reliable factors, **variance,** and **skewness**, which remain crucial for distinguishing between real and fake notes.

For more relevant information:

Here is the **Importance** of each feature, which has been measured on the original source.

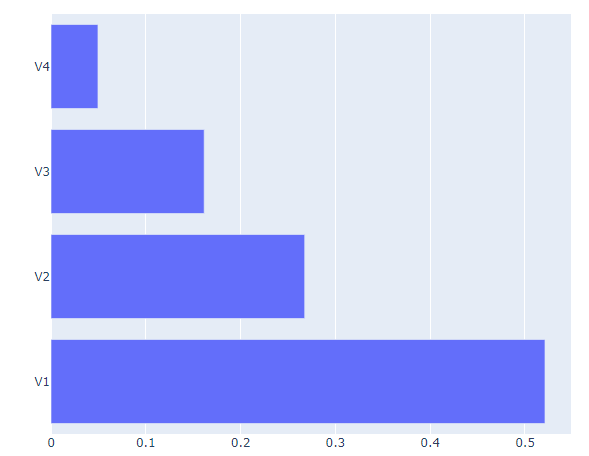


Figure 1-RandomForest Feature Importance taken from [**banknote-authentication**](https://www.openml.org/search?type=data&sort=runs&id=1462&status=active)analysis section

### Methods

To accurately cluster the labels (fake or real notes), we will apply algorithms for data preprocessing, which will "clean" the data and enhance the accuracy of the results. For the core objective, we will use K-Means clustering with the Euclidean distance function.

*\*Note: The detailed working process will not be shown. Instead, the outcomes will be presented along with instructions and notes.*

1. **Pre-Processing data:**

In here, I will display our data in 2 features: V1, and V2.

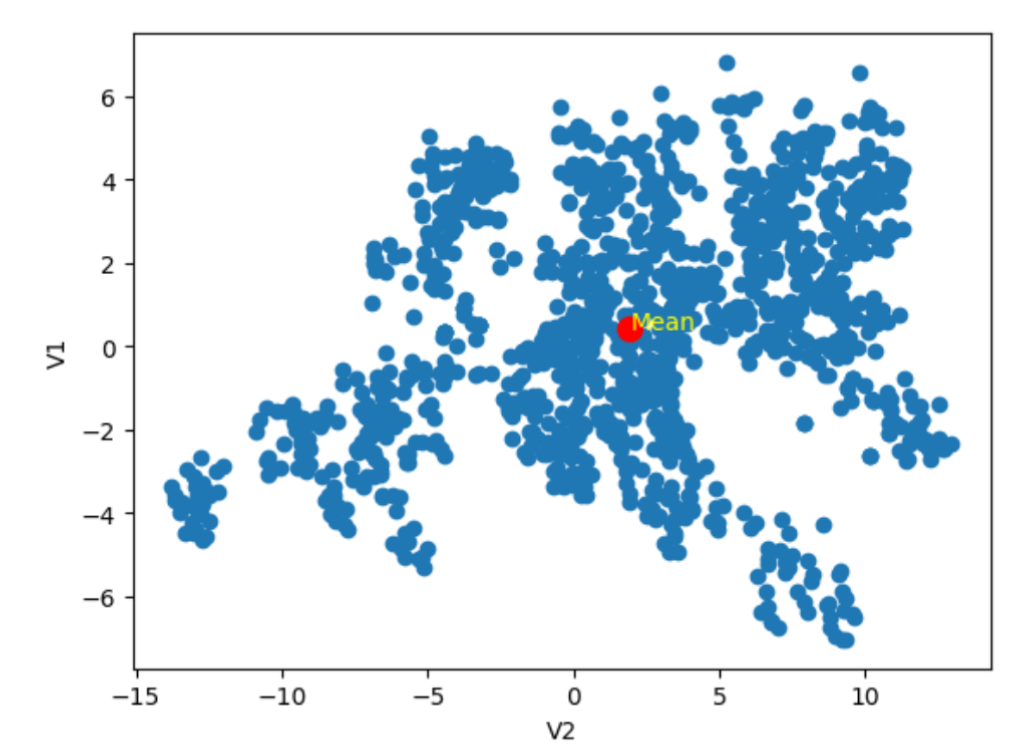


Figure 2 – Taken from previous research file. Illustrate the spread of V1 and V2 data points.

Next, we apply the Min-max normalization to balance the range. Thus, it can improve the performance of clustering algorithms and the consistency of the overall data.

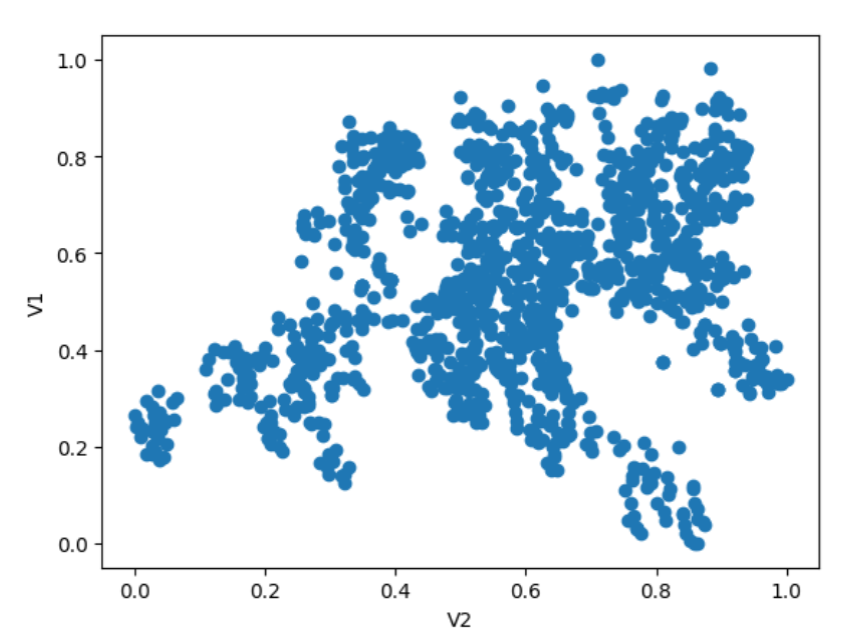
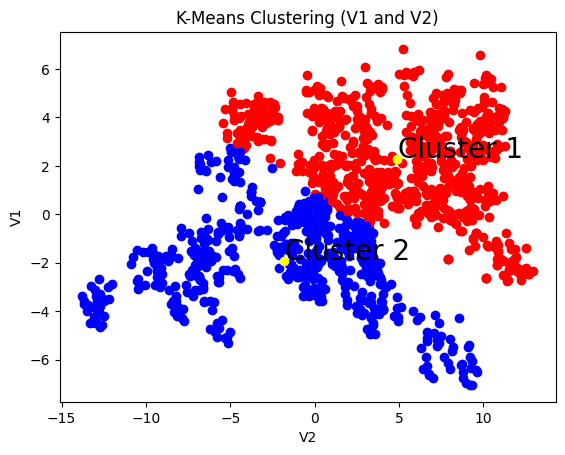


Figure 3 – Illustrate the data points after pre-processing data.

1. **K-Means Clustering:**

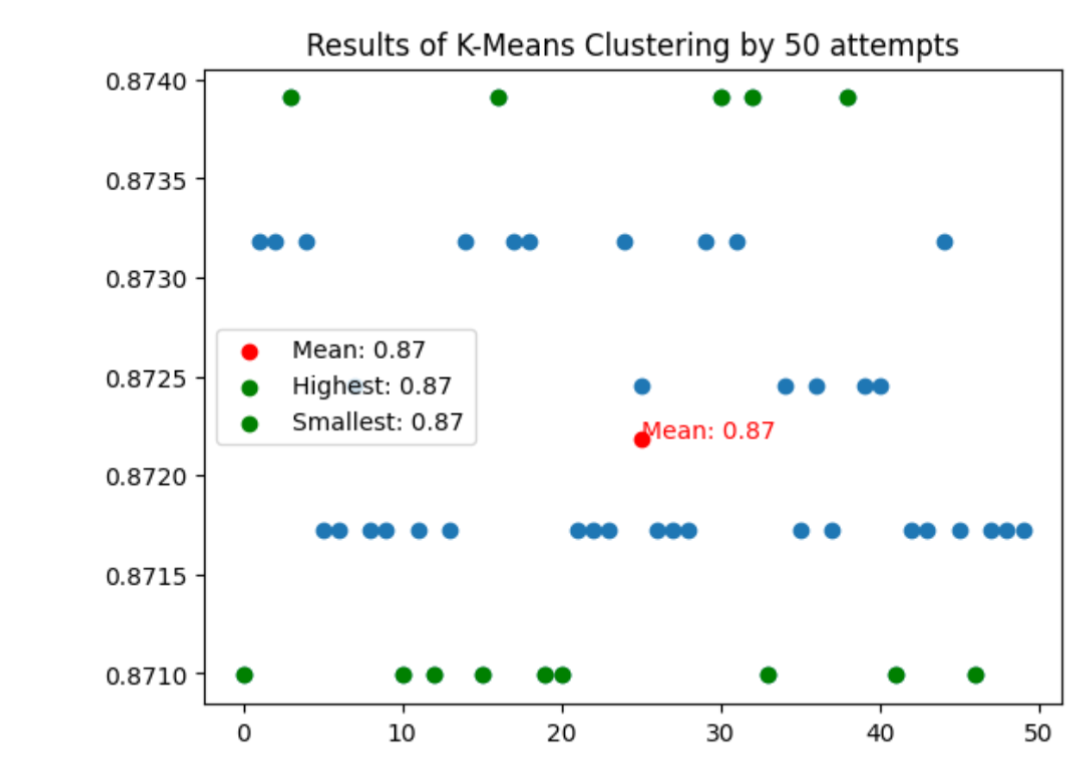
For this task, we used K-Means with two clusters, corresponding to the two possible classifications of the banknotes. The algorithm iteratively refines the cluster centroids, ensuring that the points in each cluster are as similar as possible. By using this method, we can effectively separate real and fake banknotes based on key features such as **variance** and **skewness**.

The result after being processed by K-Means :



With the K-Means cluster machine, we can assume the 2 separated areas which are red and blue. These areas signify *“fake” or “real”* banknotes. As we are getting the final outcome provided by the machine, we can also take a deeper look into how the algorithm works. K-Means is an unsupervised learning algorithm that groups data points into clusters based on their similarities. It works by assigning each data point to the nearest **centroid**—the center of a cluster—using the [Euclidean distance](https://en.wikipedia.org/wiki/Euclidean_distance) function.

However, we can't fully rely on the machine to do all the work, as the results can sometimes be unstable or fluctuate. To address this, the figure below shows the outcome of 50 attempts to test the accuracy of the model. These attempts are compared with the original results provided by the source to ensure consistency.



The K-Means clustering results show the highest accuracy of **87.40%**, and we can observe that the model’s performance is very stable, consistently around **87%**. This high accuracy demonstrates that the model is effectively performing its task. Furthermore, with access to higher-quality data and larger datasets, we can expect even higher accuracy scores in the future.

#### Summary

In conclusion, the K-Means clustering model has proven to be effective in distinguishing between real and fake banknotes, achieving a stable accuracy of approximately **87%**. By focusing on key features like **variance** and **skewness**, and ensuring consistent results through label matching, we successfully addressed the task of banknote classification.

##### Limitation

Although the model performed well, there is room for improvement with larger, higher-quality datasets. Additionally, incorporating more features could further enhance the model's performance and push it to its full potential.

###### Conclude

The machine delivers high and consistent results, along with superior speed and accuracy. This makes it one of the most cost-effective and reliable solutions for handling these tasks. Unlike human inspection, which can be time-consuming and error-prone, machine-based systems can process large volumes of data with minimal errors. Furthermore, the scalability of machine learning allows businesses to handle increasing workloads without additional labor costs. By automating these processes, companies can significantly reduce operational costs while improving overall efficiency and accuracy.