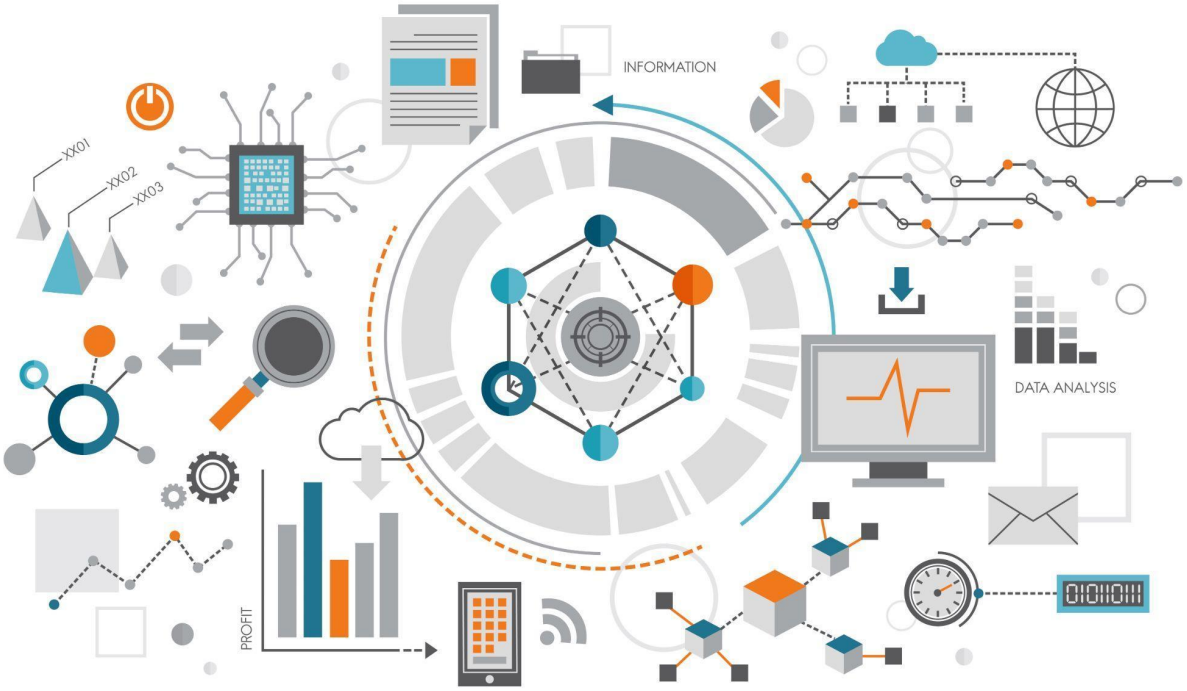
K-Means Clustering in Python

Automated Detection of Forged Banknotes



Introduction

In an era where technological advancements continue to shape the financial industry's landscape, ensuring the security of financial transactions is paramount for banks. Counterfeit banknotes pose a significant threat to financial institutions and their customers, leading to potential financial losses and eroding trust in the banking system. To combat this menace, our Data Science project aims to develop an automated system for detecting forged banknotes.

Purpose of the Data Science Project

The purpose of this Data Science project is to develop an automated system that can effectively detect forged banknotes to enhance security and prevent financial fraud.

By utilizing K-Means clustering, we aim to group banknote samples into distinct clusters based on their features, making it easier to differentiate genuine notes from counterfeit ones. This will provide the bank with a reliable tool to identify potential fraudulent activities and ensure the safety of its customers and financial assets.

Description of the data

The dataset used in this project contains samples of banknotes, each represented by a set of numerical features extracted from images. The features are engineered to describe various attributes of the banknotes, such as variance, skewness, kurtosis, and entropy.

Each entry includes the following attributes:

- Variance: Variance of the wavelet-transformed image of the banknote.
- Skewness: Skewness of the image.
- Curtosis: Curtosis (excess kurtosis) of the image.
- Entropy: Entropy of the image.
- Class: Class label (1 for genuine banknotes, 0 for forged banknotes).

The data has been preprocessed and labeled

Author: Volker Lohweg (University of Applied Sciences, Ostwestfalen-Lippe)

Dataset = click here

A sample from our dataset

	id	V1	V2	V3	V4	Class
0	1	3.62160	8.6661	-2.80730	-0.44699	1
1	2	4.54590	8.1674	-2.45860	-1.46210	1
2	3	3.86600	-2.6383	1.92420	0.10645	1
3	4	3.45660	9.5228	-4.01120	-3.59440	1
4	5	0.32924	-4.4552	4.57180	-0.98880	1
5	6	4.36840	9.6718	-3.96060	-3.16250	1
6	7	3.59120	3.0129	0.72888	0.56421	1
7	8	2.09220	-6.8100	8.46360	-0.60216	1
8	9	3.20320	5.7588	-0.75345	-0.61251	1
9	10	1.53560	9.1772	-2.27180	-0.73535	1

How the Data is Analyzed using K-Means Clustering

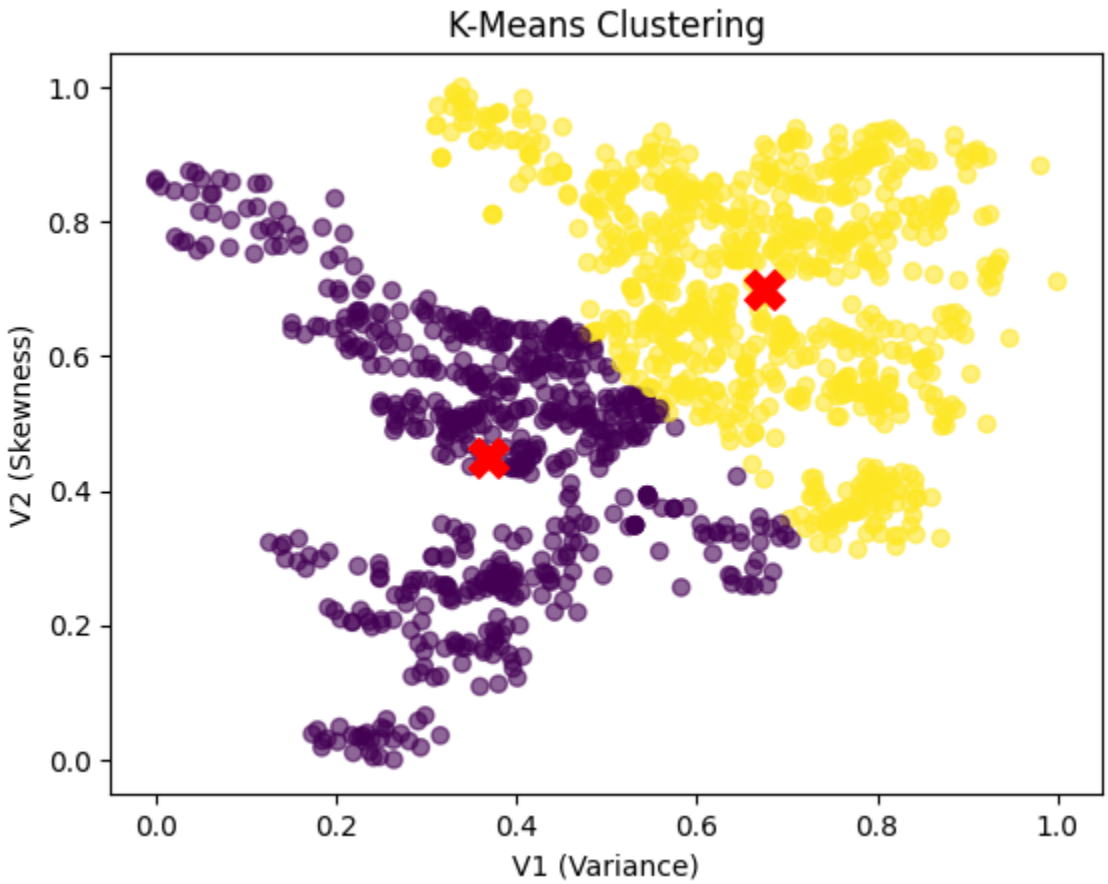
K-Means clustering is a popular unsupervised learning algorithm that partitions data into 'K' clusters, where 'K' is pre-defined.

In our case, we used K=2 as we have both genuine and forged banknotes (binary classification).

The steps followed to analyze the data using K-Means clustering are as follows:

- Data Preprocessing: We standardized the feature values to ensure that all features have a similar scale and contribute equally to the clustering process.
- K-Means Clustering: We applied the K-Means algorithm to the preprocessed data and ran it multiple times to account for different random initializations and increase the stability of the clustering results.
- Cluster Assignment: Each banknote was assigned to one of the two clusters based on its proximity to the cluster centroids.

After performing K-Means clustering on the banknote dataset, we observed two distinct clusters. The clustering results displayed promising performance, the two cluster centroids are represented using the **X** mark



Summary

- Based on a thorough analysis, it is evident that the K-means clustering algorithm has demonstrated its effectiveness in distinguishing between forged and genuine banknotes.
- However, to further enhance the accuracy of the process and improve differentiation between genuine and forged banknotes, it is advisable for the bank to gather additional information on the banknotes. This additional data will provide valuable insights and contribute to a higher success rate in the classification process.
- Continuously update the automated detection system with new data to adapt to evolving fraud techniques and maintain high accuracy in detecting forged banknotes.