Data clustering project report

WRITTEN BY:

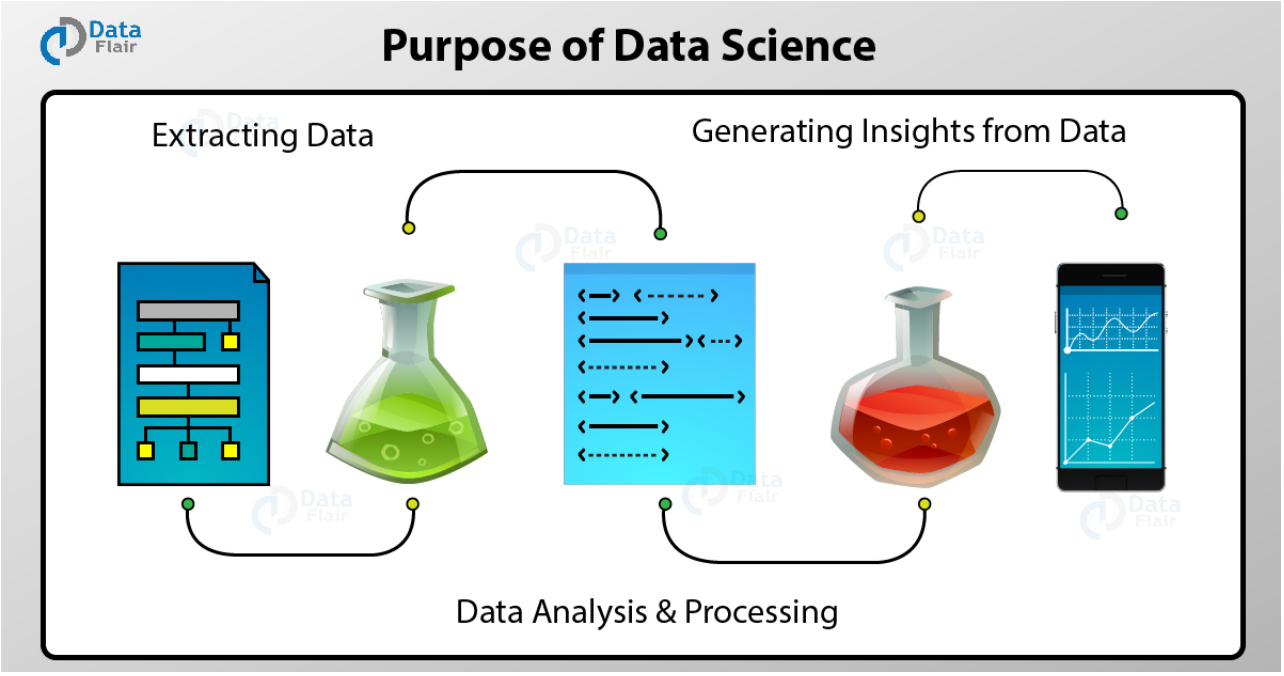
DIAO

Foundations of Data Science: K-Means Clustering in Python

Contents

1 The purpose of the Data Science project

The principal purpose of Data Science project is to find patterns within data. It uses various statistical techniques to analyze and draw insights from the data. From data extraction, wrangling and pre-processing, a Data Scientist must scrutinize the data thoroughly. Then, he has the responsibility of making predictions from the data. The goal of a Data Science project is to derive conclusions from the data. Through these conclusions, the bank will be able to take decisions very quickly concerning the detection of counterfeit banknotes



2 Description of the data

The dataset is avalaible on **URL :** https://www.openml.org/d/1462/.

Dataset about distinguishing genuine and forged banknotes. Data were extracted from images that were taken from genuine and forged banknote-like specimens. For digitization, an industrial camera usually used for print inspection was used. The final images have 400x 400 pixels. Due to the object lens and distance to the investigated object gray-scale pictures with a resolution of about 660 dpi were gained. A Wavelet Transform tool was used to extract features from these images.

For this project we have selected two features:

- **V1** variance of Wavelet Transformed image (continuous)
- **V2** skewness of Wavelet Transformed image (continuous)

We have in total 1372 observations and 2 features. But only two are used in this analysis. There were 1372 rows in the dataset. These images were classified as either genuine or fake. The statistical characteristics of the attributes V1 and V2 are described in the table 1 and 2.

The distribution is shown in the graph 3 and 4.

The boxplot is shown in the graph 5 and 6.

	V1	V2
count	1372.000000	1372.000000
mean	0.433735	1.922353
std	2.842763	5.869047
min	-7.042100	-13.773100
25%	-1.773000	-1.708200
50%	0.496180	2.319650
75%	2.821475	6.814625
max	6.824800	12.951600

Figure 1: Raw data

	V1	V2
count	1372.000000	1372.000000
mean	0.539114	0.587301
std	0.205003	0.219611
min	0.000000	0.000000
25%	0.379977	0.451451
50%	0.543617	0.602168
75%	0.711304	0.770363
max	1.000000	1.000000

Figure 2: Standardized data

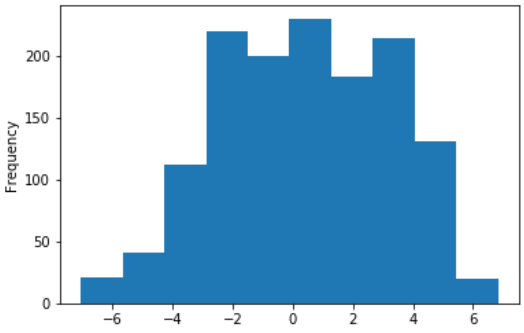


Figure 3: V1 distribution

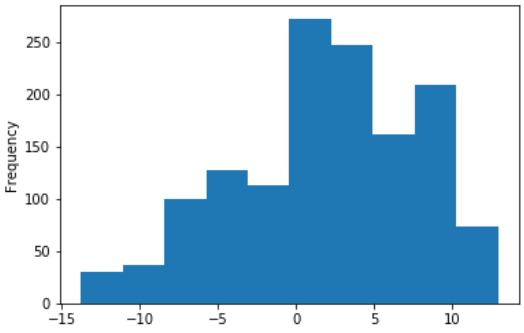


Figure 4: V2 distribution

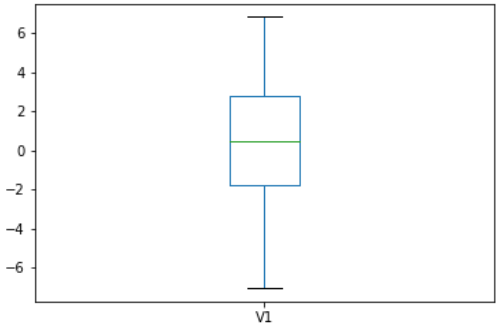


Figure 5: V1 distribution

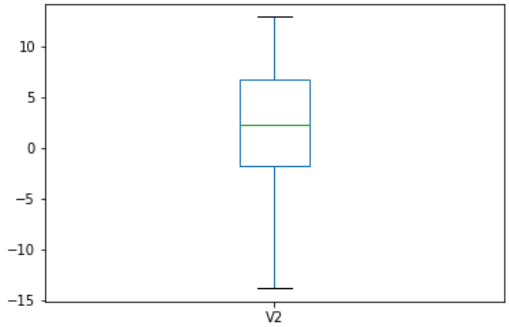
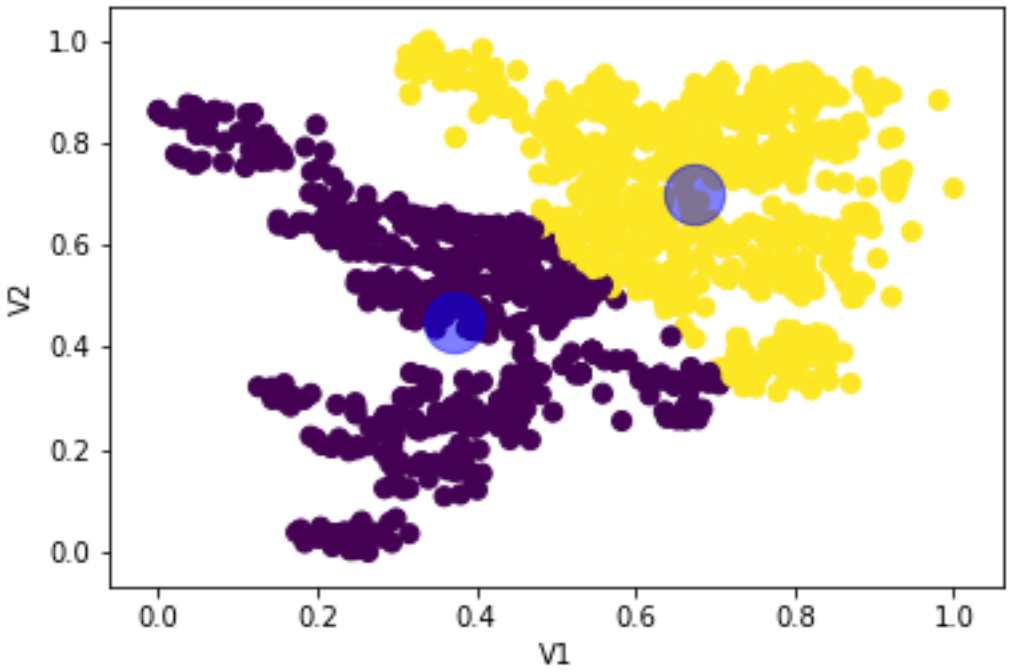


Figure 6: V2 distribution

3 Methods to analyze data

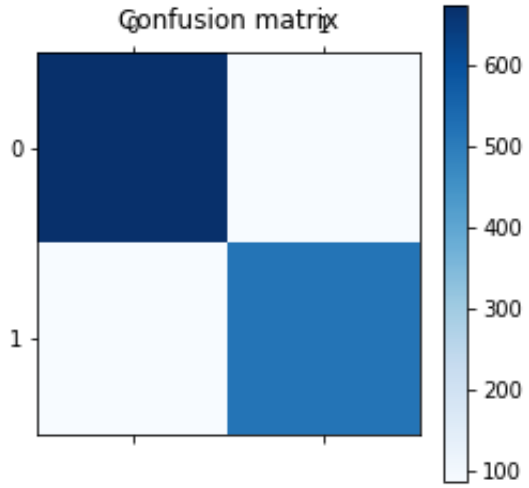
The data has been preprocessed and did not need to be "cleaned" but we are Standardized data. There were no missing values and no points that might be deemed outliers. We used the KMeans algorithm to form two cluster. The initial cluster centers were chosen randomly. The results were stable and converged to the same center for several runs. The below two figures illustrate the result of the clustering algorithm. The figure below illustrates the two clusters clearly.



We obtain the two classes, namely if a banknote is genuine or not. Kmeans divides into 2 depending on whether the values of V1 and V2 are all high or not.

4 Summary of the results

The results of the clustering using just two features are far from encouraging. They are summarized in the Confusion matrix below:



Of the 762 'Authentic" notes the clustering classified 675 as authentic i.e. 88.5 % of the notes were correctly identified as authentic, but 11.5% were classified as "fake". Of the 610 fake notes 521 or 85.4 % were correctly identified as fake, but a staggering 11.6 % were identified as Authentic.

5 Recommendations for your client

Here is a model capable of detecting a counterfeit to a real note with a good prediction rate of 88 %. All the same, one must be careful with new data which will not have been obtained in the same condition as those whose model has been trained.

6 Good elements to include in project report

Have you included to your report the following?:

- Description of your dataset(It is important to include description of the used data to a data science report);
- Statistical measures of the dataset(It is important to include statistical measures of the dataset to the description of the used data).
- Information about the size of the dataset and the number of its features(It is important to include information about the size and number of features to the description of the used data).
- Recommendations(A good data science report includes recommendations made on the basis of the carried out data analysis).
- Limitations of the data analysis(A good data science report clearly states all the limitations of the carried out data analysis)