Course: Foundations of Data Science: kmeans

Task: Final report

Yuri Santa Rosa Nassar dos Santos

Project purpose

We will conduct a proof of concept to evaluate the KMeans clustering method for grouping the banknote authentication dataset. We aim to check if the automation for detecting forged banknotes using KMeans algorithm is suitable for this purpose.

Dataset description

This dataset has four continuous attributes and 1 binary class. The attributes information are the following:

- V1. variance of Wavelet Transformed image
- V2. skewness of Wavelet Transformed image
- V3. curtosis of Wavelet Transformed image
- V4. entropy of image
- Class (target). Presumably 1 for genuine and 2 for forged

This dataset contains 1372 instances which 55.54% belongs to class 1 and 44.46% to class 2. Figure 1 shows the dataset without and with classes information considering the V1 in x-axis and V2 in y-axis.

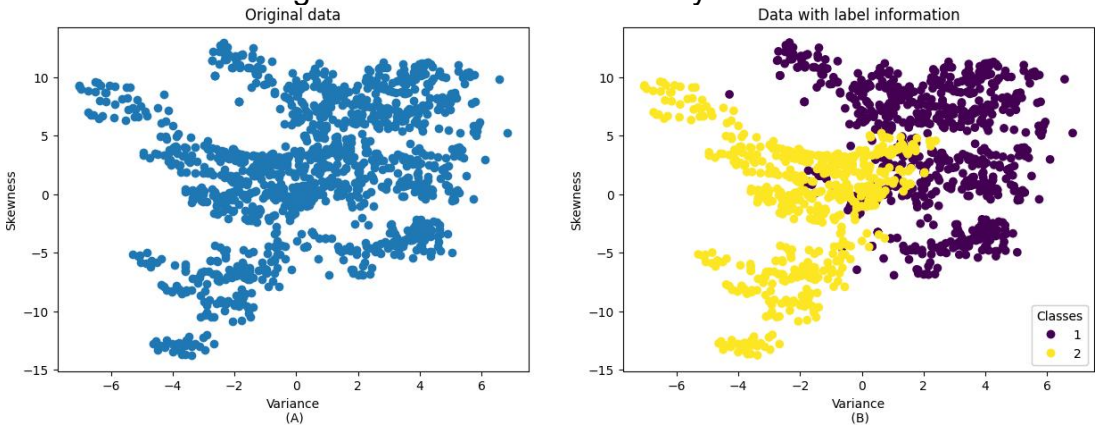


Figure 1. Dataset visualization without and with class information.

Methodology

We executed three steps for this *proof of concept* as follows:

1. Perform a feature selection method to evaluate which attributes are more relevant;
2. Perform the clustering evaluation on train dataset using the kmeans method;
3. Perform the clustering evaluation on test dataset using the kmeans method.

Results

In this section we show the feature selection, and clustering results on train and test datasets with its clustering evaluation.

Figure 1 shows the scores of each attribute generated by a feature selection method (regression using the F score). It can be seen that the first two attributes, V1 and V2, are the most relevant to describe the information in this dataset.

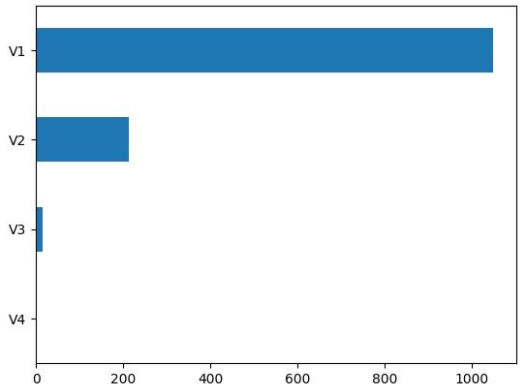
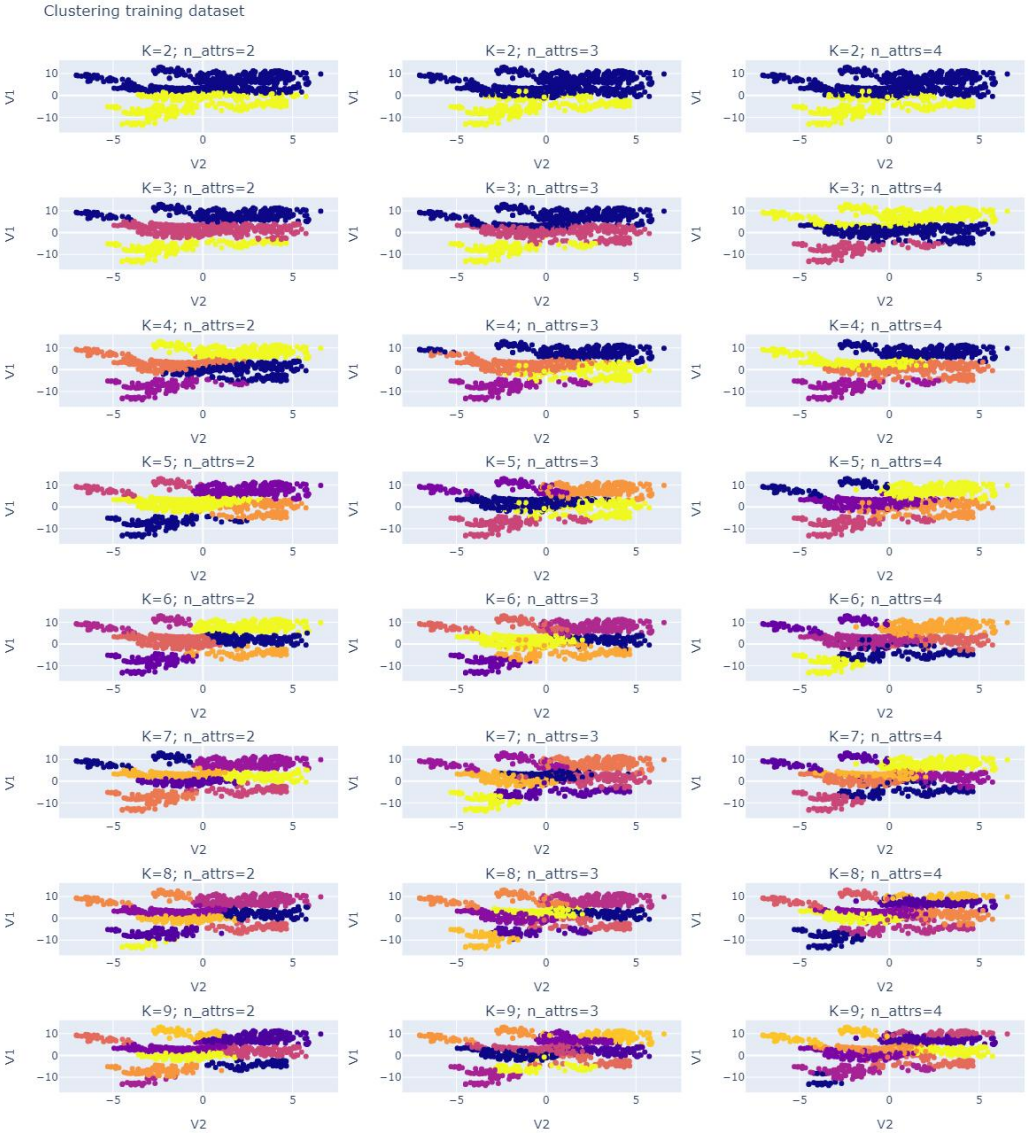


Figure 1. Feature selection evaluation.

We could use these two attributes on clustering, but we will use between 2 and 4 attributes to show the clustering performance when varying some hyperparameters. Figure 2 shows the clustering results when varying the number of clusters *k* and the number of attributes *n_attrs*.



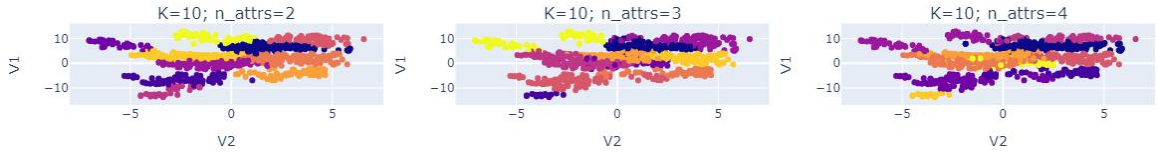


Figure 2. Clustering results on training dataset

Figure 3 shows the clustering evaluation scores (internal and external metrics) to assess the clustering quality after identifying the groups. It is important to measure these evaluation scores to help us to identify if a clustering result is good since the visualization process sometimes is not enough. We can see that clustering with k=2 produced better results than the other clustering setups.

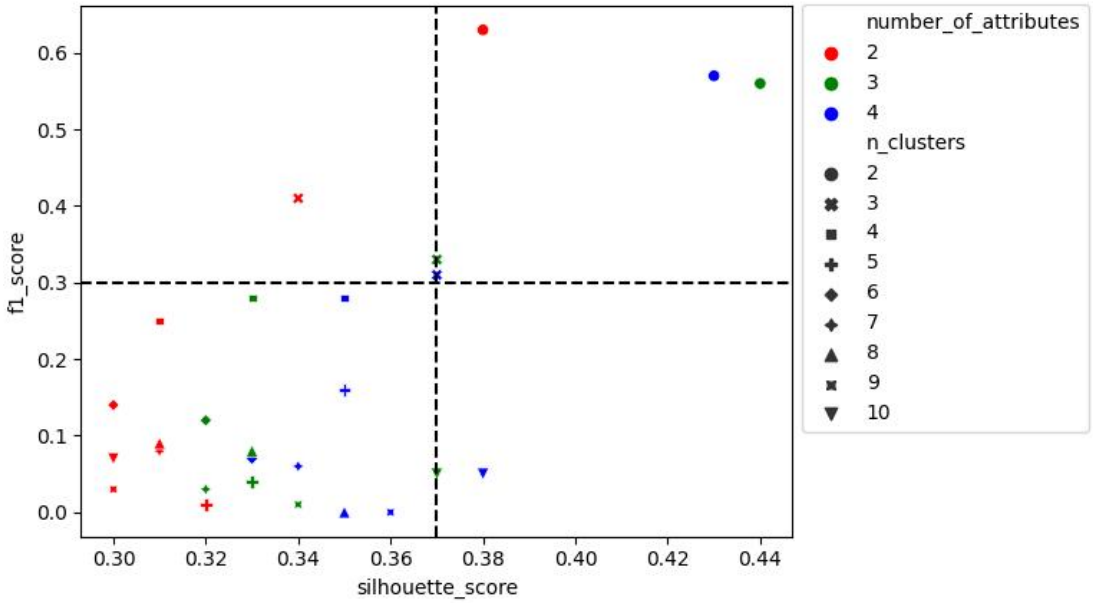
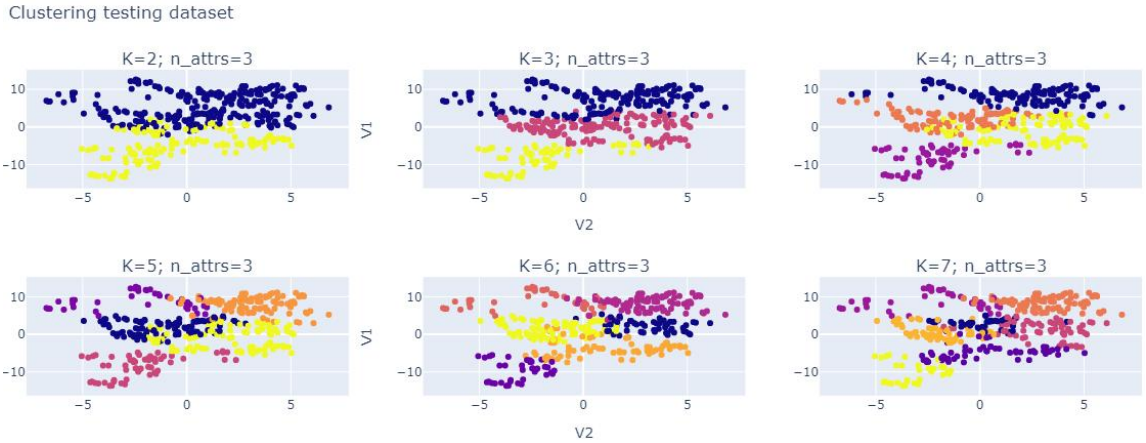


Figure 3. Clustering evaluation scores on training dataset

After training the kmeans method to identify the clustering centroids, we can proceed to predict the labels for new data points. Figure 4 shows the clustering results on the test dataset varying the number of clusters but with a fixed number of attributes (the three most relevant).



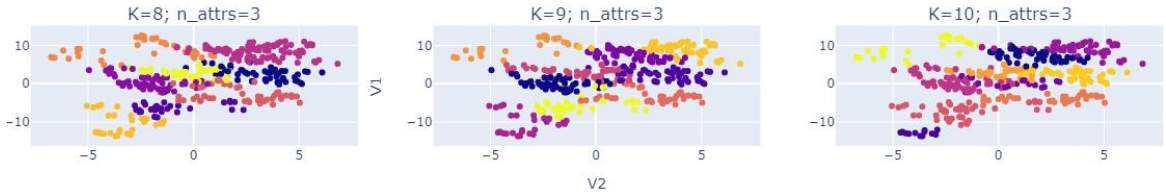


Figure 4. Clustering result on train dataset

Figure 5 shows the same patterns as in Figure 3, therefore, clustering with k=2 keep generating better evaluation scores.

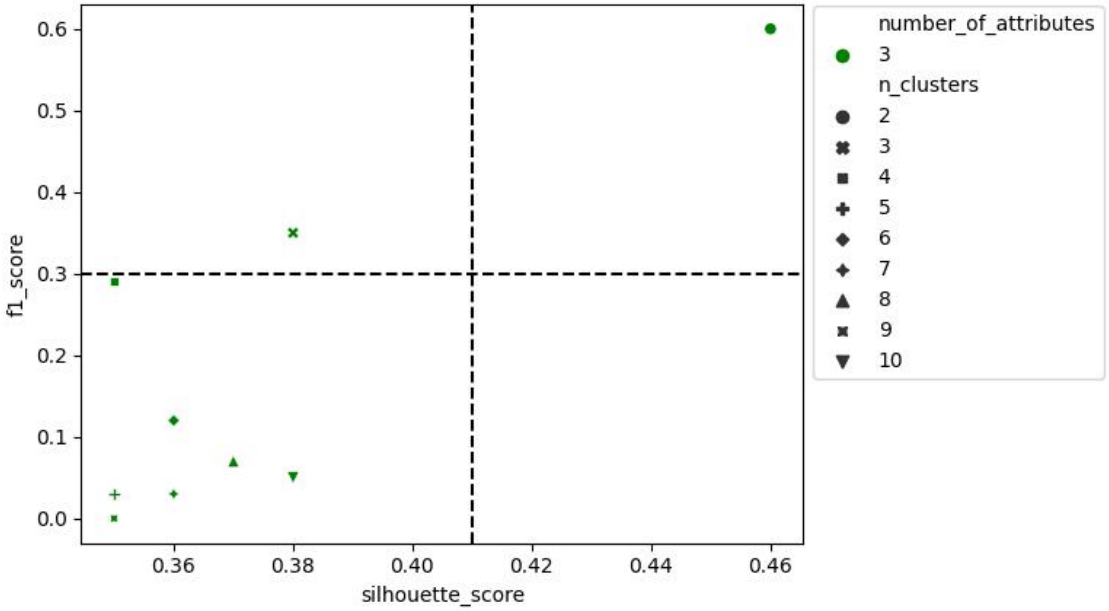


Figure 5. Clustering evaluation scores on test dataset

Recommendations

This project shows that the kmeans method can identify the labels with close to 60% of correctness. The kmeans method is a well-known algorithm, easy to test and implement, however, it has some limitation which mainly depends on the dataset characteristics. It is known that kmeans perform better in spherical shape datasets and are well separated. Originally, the banknotes dataset is for classification purposes and it has overlapping among the classes, consequently, the kmeans method can have problems separating the groups. In general, bank automation systems require a high precision rate to identify the desired target. Thus, in our view, 60% of correctness is not viable for a system to identify fraud and build automation using this method. We recommend exploring other clustering methods aim to identify an algorithm that can model the dataset in a better way than the kmeans method.