PROJECT REPORT

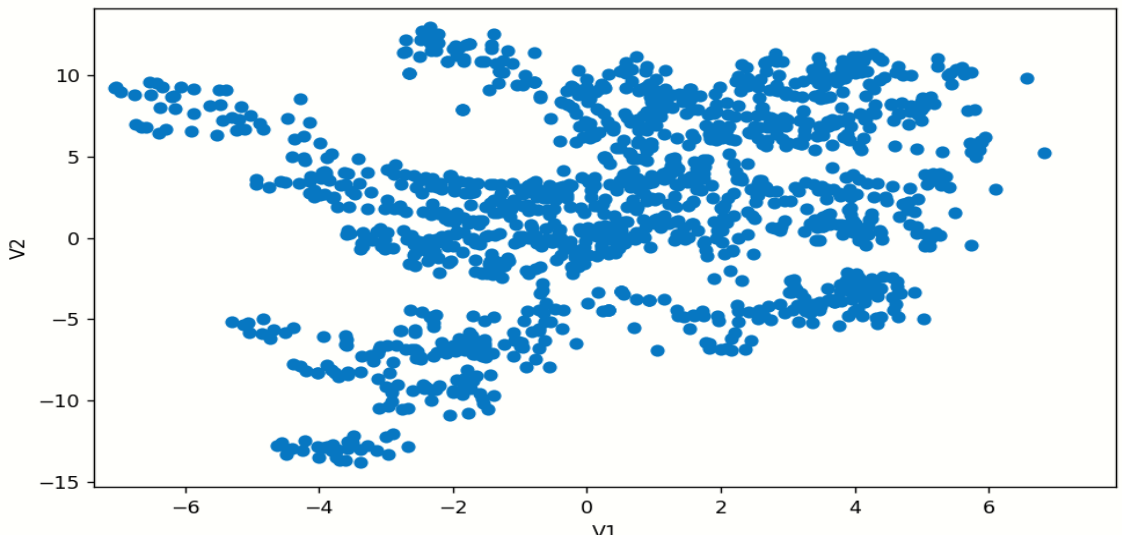
Mausam Singh 11/10/2020

Aim: Examining the authenticity of banknotes is a very baffling task that one had to perform manually around 10-15 years ago. It can get tedious when there is a huge number of notes to examine. The purpose of this project is to apply the K-means clustering algorithm on the data obtained from the banknotes and to determine whether the algorithm manages to differentiate the forged notes from genuine notes. K-means clustering is one of the unsupervised machine learning algorithm. It is basically used for classification.

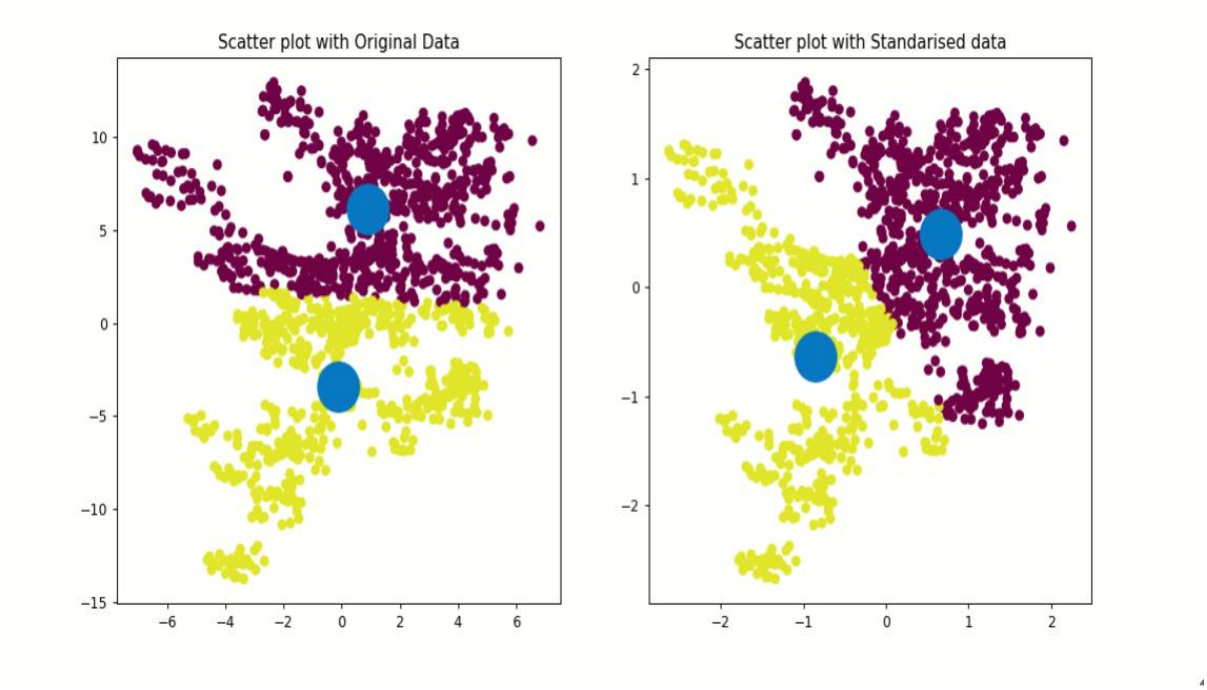
Summary of Dataset: The dataset is obtained from the website https://www.openml.org/. Data were extracted from images that were taken from genuine and forged banknotes like specimens. It consists of 1373 observations of two variables "V1" and "V2". "V1" is the variance of Wavelet Transformed image (continuous) and "V2" is the skewness of Wavelet Transformed image (continuous). A wavelet transformation tool was used to extract all these features from images.

	V1	V2
count	1372.000000	1372.000000
mean	0.433735	1.922353
std	2.842763	5.869047
min	-7.042100	-13.773100

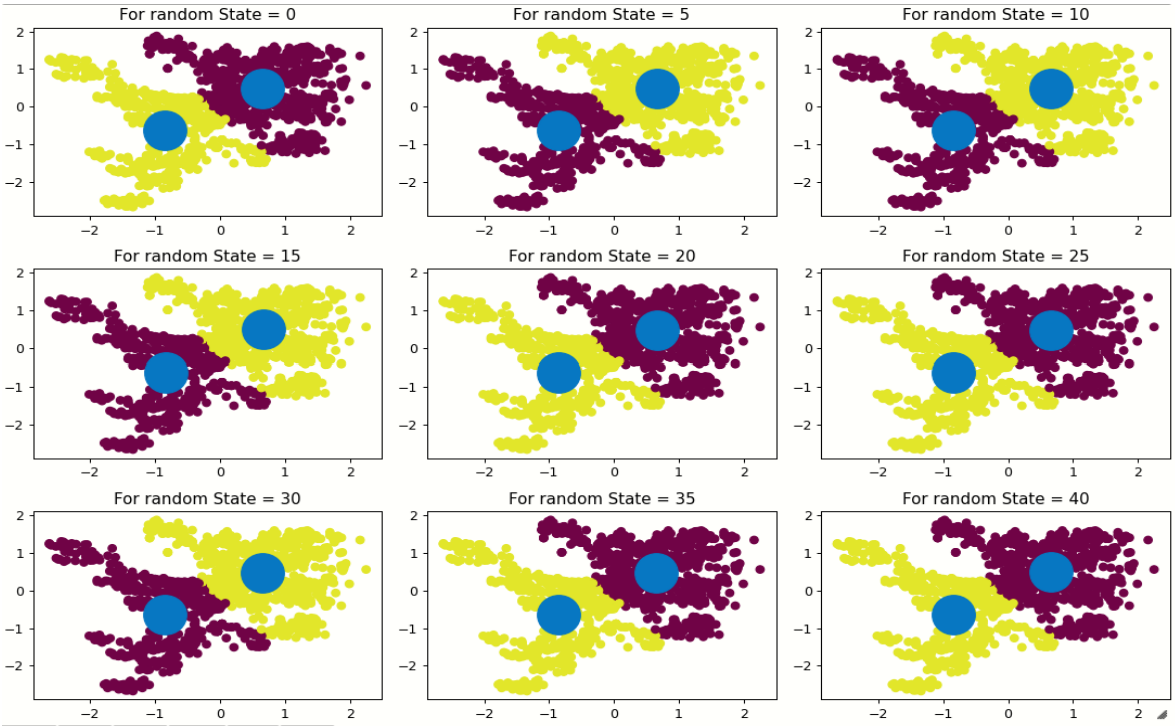
Scatter Plot of "V2" vs "V1"



Analysis: Data obtained from website were in tidy form so there wasn't any need to clean the data. I had to standardize the data so that a proper cluster and a constant cluster center could be obtained. Here is the scatter plot in both of cases. "V1" is plotted on X-axis and "V2" is plotted on Y-axis.



I applied the K-means clustering algorithm on standardized for 9 different random_state. Here is the plot for 9 different random_state.



As we can see from the figure K-means cluster centers is almost same in all of the cases. We can say that K-means algorithm is stable.

Result:

Accuracy of algorithm = 0.8782

F1 Score = 0.8913

Recommendations:

As we can see that the accuracy is above 0.98 which is pretty good for unsupervised machine learning. If we can ignore the rest of 0.02, algorithm is working pretty well. If we have a huge chunk of banknotes to examine then it wouldn't be a bad idea to use this algorithm to examine them.

*Note: All the observations were recorded for random_state = 10. If we change the randome_state, observations can differ by 2,3 counts.