|  |  |  |
| --- | --- | --- |
| Students | | |
| Groupe | Prénom | Nom |
| LSI 1 | Mahamat Moussa | **ADJI** |
| LSI 2 | Bill Williams | **TSOFACK FOZEING** |
| LSI 1 | Phuoc-Duy | **NGUYEN** |

DATA VISUALIZATION PROJECT

**Subject:** Use the tools and methodology seen in class to reduce the dimensionality of your data to be able to focus on the most important features. Choose the best approach to reduce the dimensionality and to project the data. Explain the model and why it works best for this problem/dataset. Use Python with matplotlib and/or KNIME to plot the dataset and the knowledge extracted from it. Explain the knowledge you extracted. The same dataset can be used for both projects.

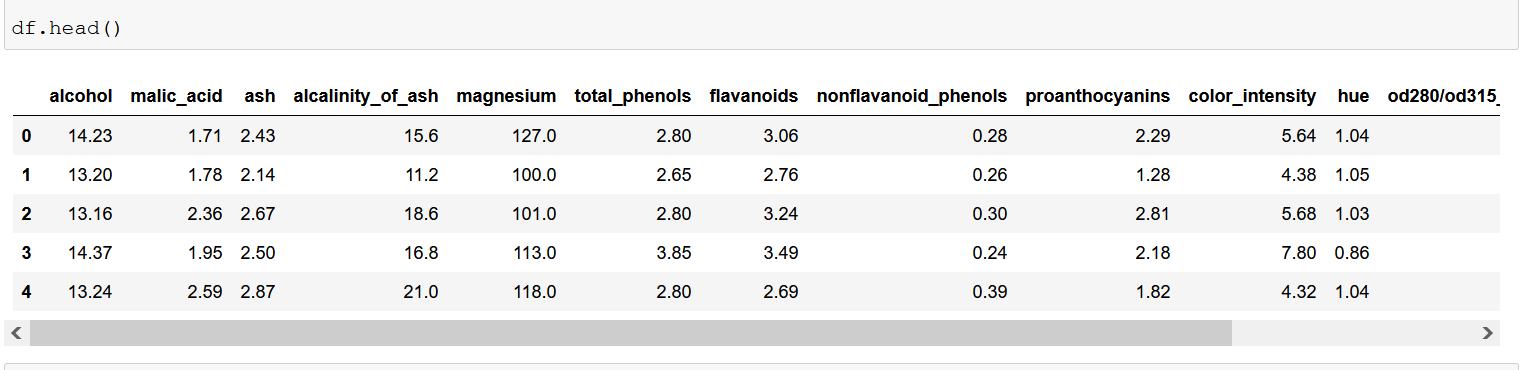
**Process**:

1. Import the all the libraries and import the dataset

Une image contenant texte

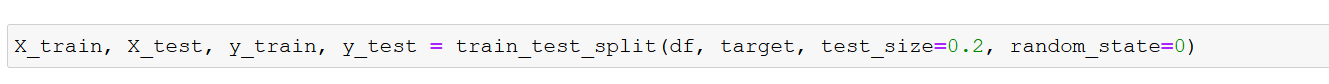
Description générée automatiquement

1. Take look of what the dataset looks like



1. Preprocessing part

Here we are going to divide our data into training and test sets



1. Normalization (PCA case)

Since PCA model performs best with normalized feature set, we will use standard normalization to normalize our feature set.

Une image contenant texte

Description générée automatiquement

1. Applying the model

Une image contenant texte, orange

Description générée automatiquement

As we can see above, applying PCA using python’s Scikit-Learn library is a two steps process:

* Initializing PCA by passing the number of components to the constructor
* Call the fit and the transform methods and pass the features to set these methods.

1. Training and making prediction

Here, we are going to use Random Forest classification to make predictions

Une image contenant texte

Description générée automatiquement

1. Performance evaluation

This part consists of determining the performance of the algorithm

Une image contenant texte

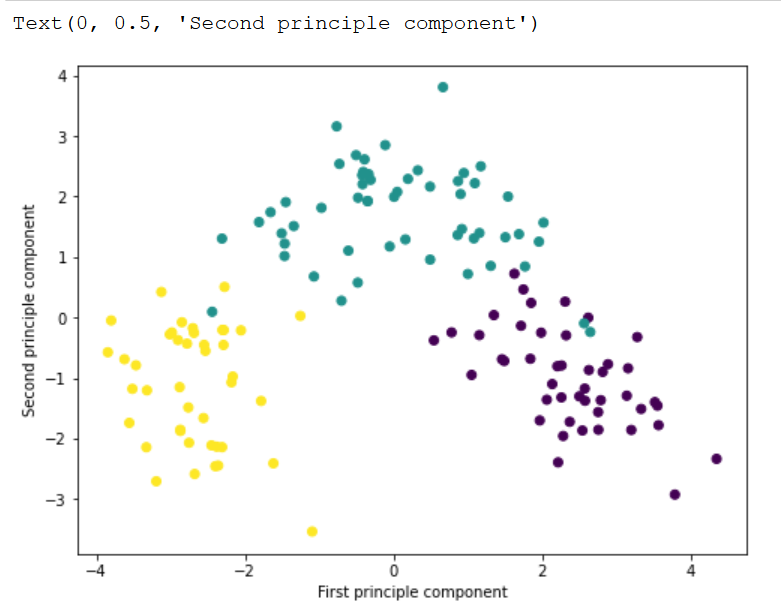
Description générée automatiquement

As we can see with only 02 features, the Random Forest classification is able to correctly predict 34 out of 36 instances, resulting in 97, 22% accuracy.

1. Plot the dataset

Une image contenant texte

Description générée automatiquement



The process is the with LDA model, the differences are only in :

* Part 5:

Une image contenant texte

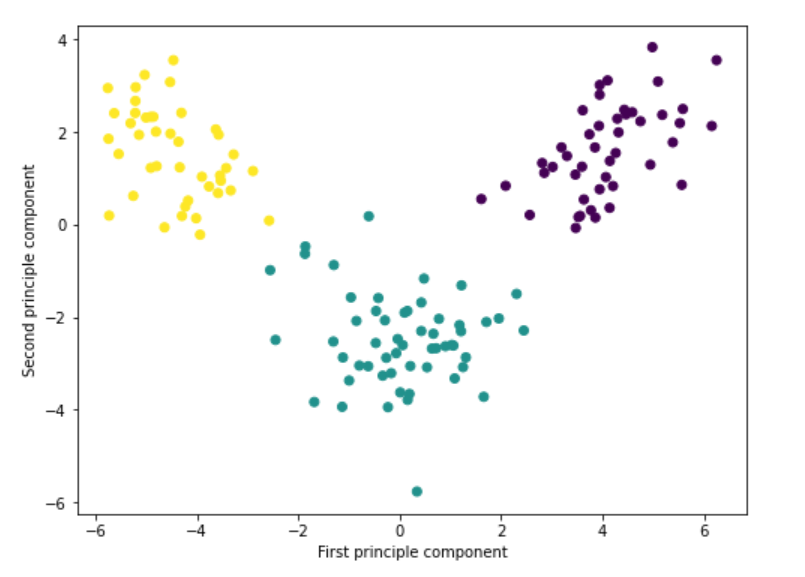
Description générée automatiquement

* Part 7:

Une image contenant texte

Description générée automatiquement

* Part 8:



Same with MDS:

* Part 5:

Une image contenant texte

Description générée automatiquement

* Part 8

