# **Université des Sciences et de la Technologie Houari Boumediene**

#### 

**Faculté d'Electronique** **et d'Informatique**

**Département d'Informatique**

**Master 2**

**MIND**

***Projet Data Mining***

***ACP & Clustering***

|  |  |  |
| --- | --- | --- |
| **Binome** : ZOUILEKH Belkacem  HARHAD Selmene |  |  |
|  |  |  |

**08 janvier 2020**

**Sommaire**

1. Résumé
2. Partie theorique :

- Analyse exploratoire multidimensionnelle

- Définition et L’intérêt de l'ACP et le Clustering

1. ACP et Clustering avec code python

- pour quoi python ?

- les Library utilisé dans le script

1. Réaliser une ACP
2. Réaliser un Clustering
3. Application web

- Framework Django

- Démonstration

1. Bibliographie

**Résumé** :

Dans ce projet, nous apprendrons à effectuer une **analyse exploratoire multidimensionnelle.**Nous utiliserons des méthodes populaires pour**analyser rapidement notre échantillon** en réduisant la dimension du nombre d'individus ou de variables.

Nous aborderons des méthodes emblématiques comme l'**Analyse en Composantes Principales**ou encore le fameux ***clustering :***

L'**Analyse en Composantes Principales** (ACP ou **PCA** en anglais) permet de dégager rapidement les principales tendances de votre échantillon, en diminuant le nombre de variables nécessaires à la représentation de vos données tout en perdant le moins d'informations possible.

Nous aborderons également une méthode de clustering la plus populaires : l'algorithme du **k-means**. Celle-ci permet de regrouper nos individus selon leurs similarités.

Nous allons utiliser le langage de programmation python afin de programmer les deux méthodes.

Nous terminerons avec une application qui est implémenté avec le framework python ‘django’.

**Partie théorique :**

**Analyse exploratoire multidimensionnelle :**

Exploratoire : synonyme de « descriptif ».

**Multidimensionnel** : étudiera les relations entre plus de 2 variables à la fois

L’analyse multivariée est utile quand on souhaite étudier des profils, c'est-à-dire un ensemble de caractéristiques d'un individu.

Nous allons évoquer 2 familles de méthodes :

* les méthodes **factorielles ;**
* les méthodes de **classification non supervisée**, aussi appelées de partitionnement de données (plus connues sous le terme anglophone de **clustering**).

Chacune de ces 2 familles possède une méthode emblématique :

* l'**analyse en composantes principales** (ACP) ou *Principal component analysis* (PCA) en anglais, qui est la plus connue des méthodes factorielles ;
* l'algorithme **k-means** (en français "K-moyennes"), qui est le plus connu des algorithmes de clustering.

**Définition et l’intérêt de l'ACP et le Clustering**

l'ACP est une methode qui permet de réduire le nombre de variables en trouvant de nouvelles variables qui en **synthétisent** plusieurs. Trouver une variable synthétique permet de remplacer plusieurs colonnes du tableau par une seule. Malheureusement, cette transformation nous fera perdre un peu d'information.

Le clustering quant à lui se chargera de **regrouper** des individus similaires, c'est-à-dire qu'il va **partitionner** l'ensemble des individus. Regrouper des individus est ici synonyme de regrouper des lignes. Parfois, il est possible de regrouper 100 000 lignes en 3 groupes assez homogènes pour n'étudier finalement que le profil général de chacun de ces 3 groupes, c'est-à-dire 3 lignes !

Mais ce n'est pas tout ! Au-delà de la réduction des dimensions du tableau de données, ces méthodes ont d'autres intérêts.

L'ACP, tout d'abord, permet d'étudier :

* la variabilité entre les individus, c'est-à-dire quelles sont les différences et les ressemblances entre les individus ;
* les liaisons entre les variables : y a-t-il des groupes de variables très corrélées entre elles qui peuvent être regroupées en de nouvelles variables synthétiques ?

Le clustering a de multiples applications :

En statistiques, on distingue les traitements **supervisés** des traitements **non supervisés**.

La **classification non supervisée** consiste en l'organisation d'individus en groupes homogènes. En gros, on définit des classes que l'on ne connaît pas à l'avance.

La **classification supervisée** consiste à "ranger" les individus dans des classes connues. Ici, il y a une question préalable. Vous rencontrerez peut-être le terme de classement, qui est synonyme de classification supervisée.

la classification non supervisée est appelée en anglais clustering.

**ACP et Clustering avec un scripte python:**

**Pour quoi python :**

Python est un langage de programmation orienté objet interprété.

Un programme écrit en Python n'est opérationnel que si l'interpréteur est disponible sur la machine (bien que des solutions de compilation existent). En contrepartie, il peut fonctionner dès lors que l'interpréteur est présent, quel que soit le système d'exploitation de la machine. Sous cet angle, on peut le considérer comme un langage multiplateformes.

La syntaxe de Python est simple et claire, elle respecte les standards du domaine. Python propose les principales fonctionnalités de la programmation (actions conditionnelles, boucles, programmation modulaire), y compris les mécanismes de classes (héritage, surcharge des méthodes, polymorphisme). Python se marie très bien avec un cours d'algorithmie.

La distribution Python intègre un grand nombre de librairies. Elles couvrent un large choix de domaines (bases de données, accès réseaux, multimédia, traitements systèmes, compression, multithreading, ...).

Outre les librairies standards, un grand nombre de paquetages (packages) développés par des contributeurs indépendants donne accès à des fonctionnalités spécialisées performantes. Ici également, la liste est particulièrement longue. Ils nous donnent la possibilité de programmer des applications dans quasiment tous les secteurs de l'informatique.

Nous nous intéresserons en particulier à la programmation statistique, au machine learning, au big data et au data science. Python y occupe de plus en plus une place de choix et vient même taquiner R (News KDnuggets, Mai 2015, "R vs Python for Data Science: The Winner is..." ; ou encore DataCamp, Mai 2015, "Choosing R or Python for data analysis? An infographic.")

**Les librery python utiliser dans le projet :**

**Pandas** est une bibliothèque écrite pour le langage de programmation Python permettant la manipulation et l'analyse des données. Elle propose en particulier des structures de données et des opérations de manipulation de tableaux numériques et de séries temporelles.

**NumPy** est une extension du langage de programmation Python, destinée à manipuler des matrices ou tableaux multidimensionnels ainsi que des fonctions mathématiques opérant sur ces tableaux.

**Scikit-learn** est une bibliothèque libre Python destinée à l'apprentissage automatique. Elle est développée par de nombreux contributeurs2 notamment dans le monde académique par des instituts français d'enseignement supérieur et de recherche comme Inria3 et Télécom Paris. Elle comprend notamment des fonctions pour estimer des forêts aléatoires, des régressions logistiques, des algorithmes de classification, et les machines à vecteurs de support. Elle est conçue pour s'harmoniser avec d'autres bibliothèques libres Python, notamment NumPy et SciPy.

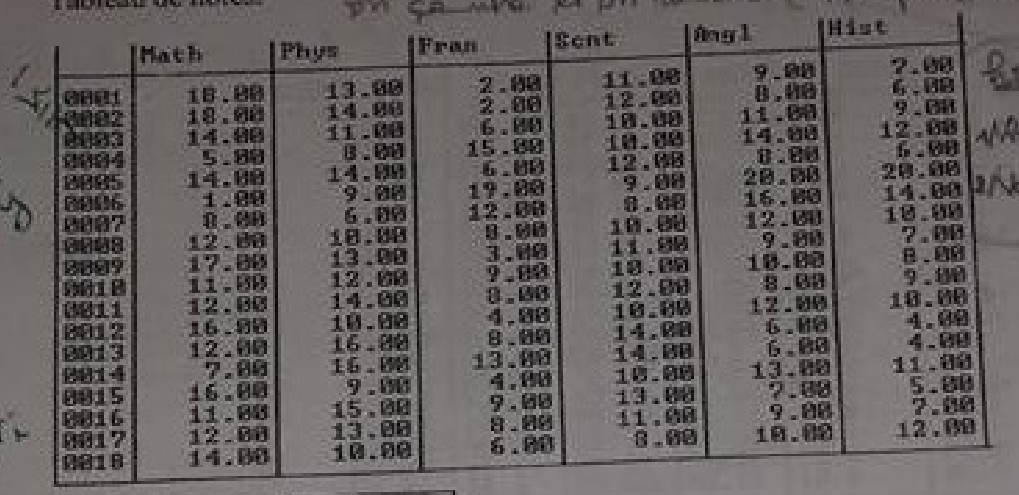
**Matplotlib** est une bibliothèque du langage de programmation Python destinée à tracer et visualiser des données sous formes de graphiques5. Elle peut être combinée avec les bibliothèques python de calcul scientifique NumPy et SciPy6.

**Réaliser une ACP :**

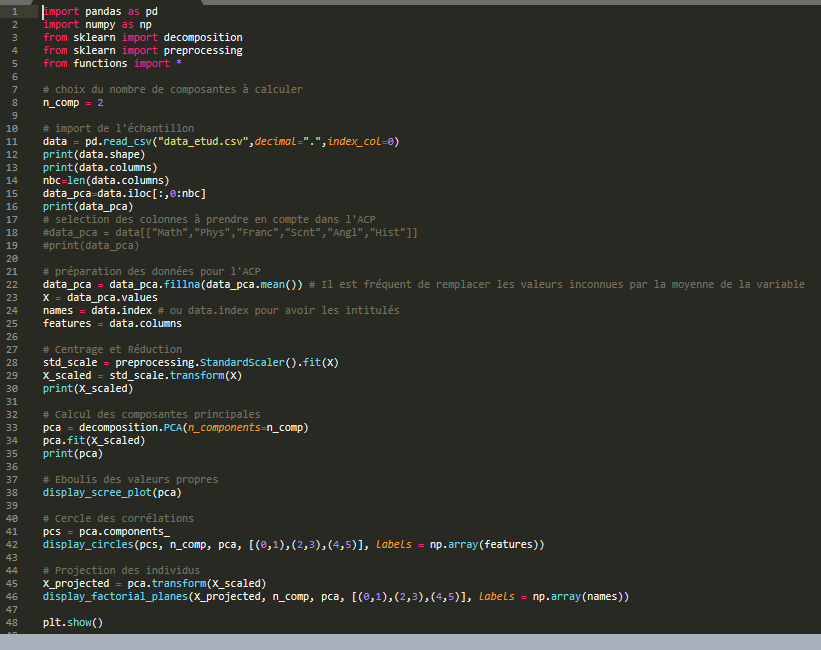
Nous souhaitons étudier une population composée d’individus. Ces individus ne sont pas forcément des personnes, mais peuvent être des objets, des animaux, des relevés effectués par des capteurs (relevés de température, par exemple), ou beaucoup d’autres choses !

Souvent, une population est difficile à étudier dans sa globalité (surtout quand elle contient un grand nombre d’individus et qu’il est impossible de tous les observer). Pour cela, on extrait un échantillon de la population, c’est-à-dire que l’on sélectionne certains individus pour les étudier précisément. Étudier un individu, c’est observer ses caractéristiques : chaque caractéristique est décrite par une variable.

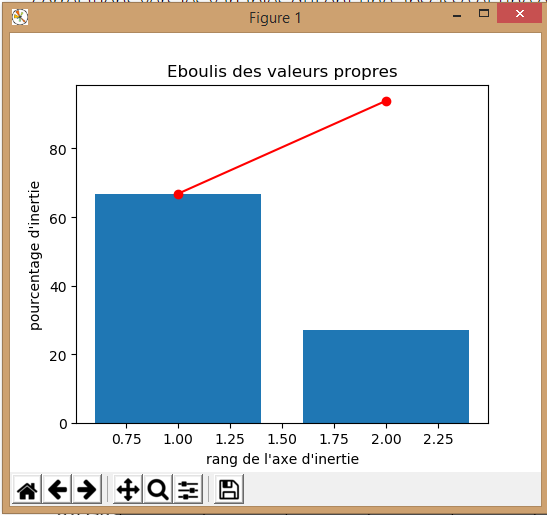
**Table de données (variable, individu) :**

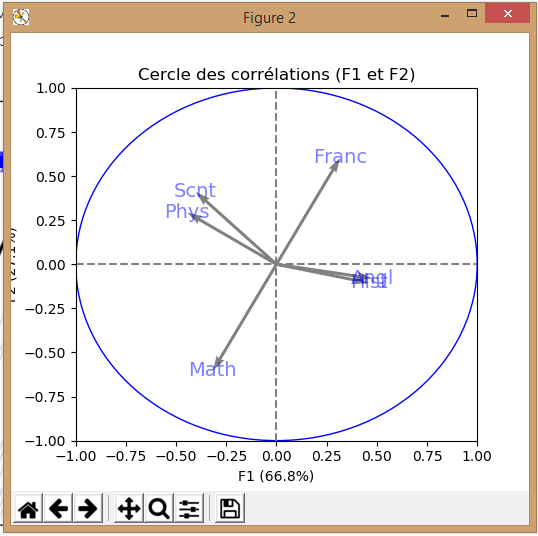


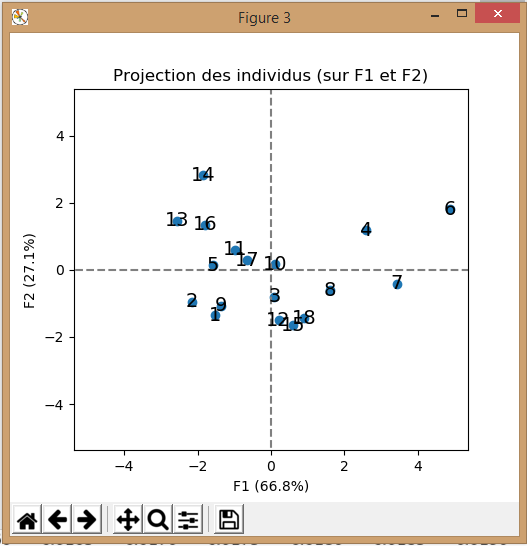
**Le scripte de python pour réaliser une ACP :**



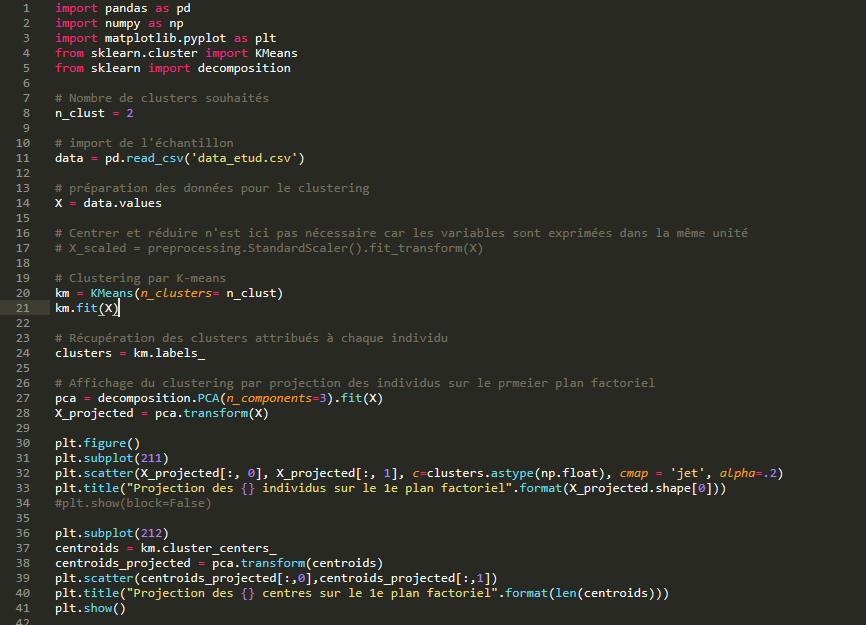
**Exécution de script sur le tableau des données pca\_data\_etud.csv :**

Affichons les valeurs propres, le cercle de corrélation  et la projection des individus :

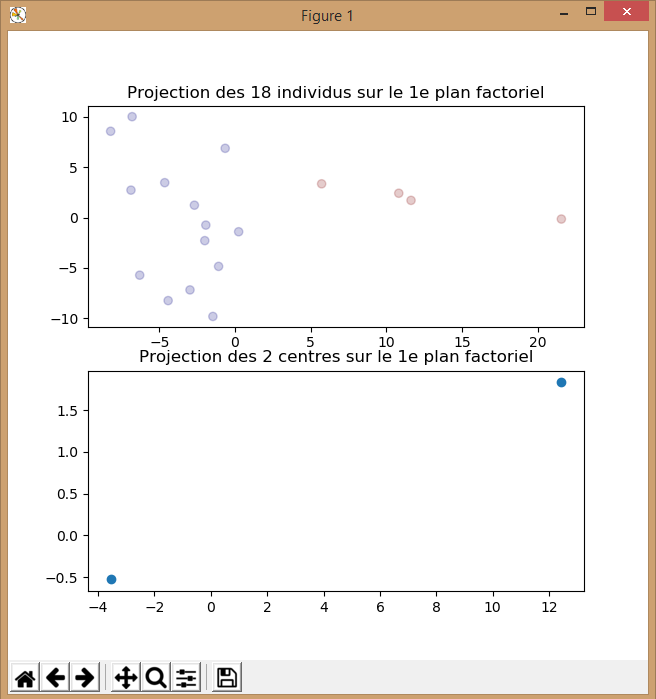




**Le scripte de python pour réaliser un Clustering:**



**Exécution de script sur le tableau des données clustering\_data\_etud.csv :**



**Application web:**

**Le framwork Django**

Django est un cadre de développement web open source en Python. Il a pour but de rendre le développement web 2.0 simple et rapide. Pour cette raison, le projet a pour slogan « Le framework pour les perfectionnistes avec des deadlines. ». Développé en 2003 pour le journal local de Lawrence (Kansas), Django a été publié sous licence BSD à partir de juillet 2005.

Depuis juin 2008, la Django Software Foundation s'occupe du développement et de la promotion du cadre. En plus de cette promotion régulière, des conférences entre développeurs et utilisateurs de Django sont organisées deux fois par an depuis 2008. Nommées DjangoCon, une se déroule en Europe et l'autre aux États-Unis.

Plusieurs sites grand public sont désormais fondés sur Django, dont Pinterest3 et Instagram4 au moins en 2011 ou encore Mozilla5.

Django est un cadre de développement qui s'inspire du principe MVC ou MTV (la vue est gérée par un gabarit) composé de trois parties distinctes :

* 1. Un langage de gabarits flexible qui permet de générer du HTML, XML ou tout autre format texte ;
  2. Un contrôleur fourni sous la forme d'un « remapping » d'URL à base d'expressions rationnelles ;
  3. Une API d'accès aux données est automatiquement générée par le cadre compatible CRUD. Inutile d'écrire des requêtes SQL associées à des formulaires, elles sont générées automatiquement par l'ORM.

**Démonstration de l’application :**

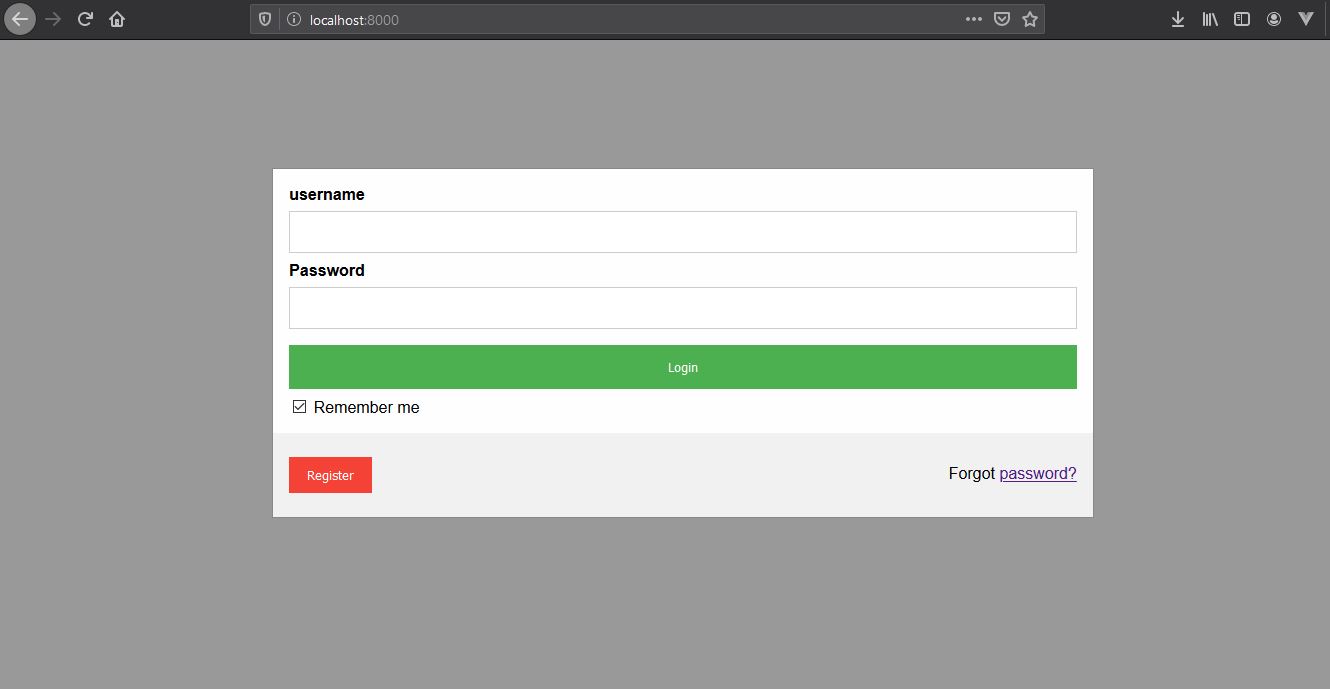
Le but de cette application est d’automatiser d’éxécution l’annalyse des données avec l’ACP, Clustering et vésualiser les résultats.

**Spécification fonctionnelle :**

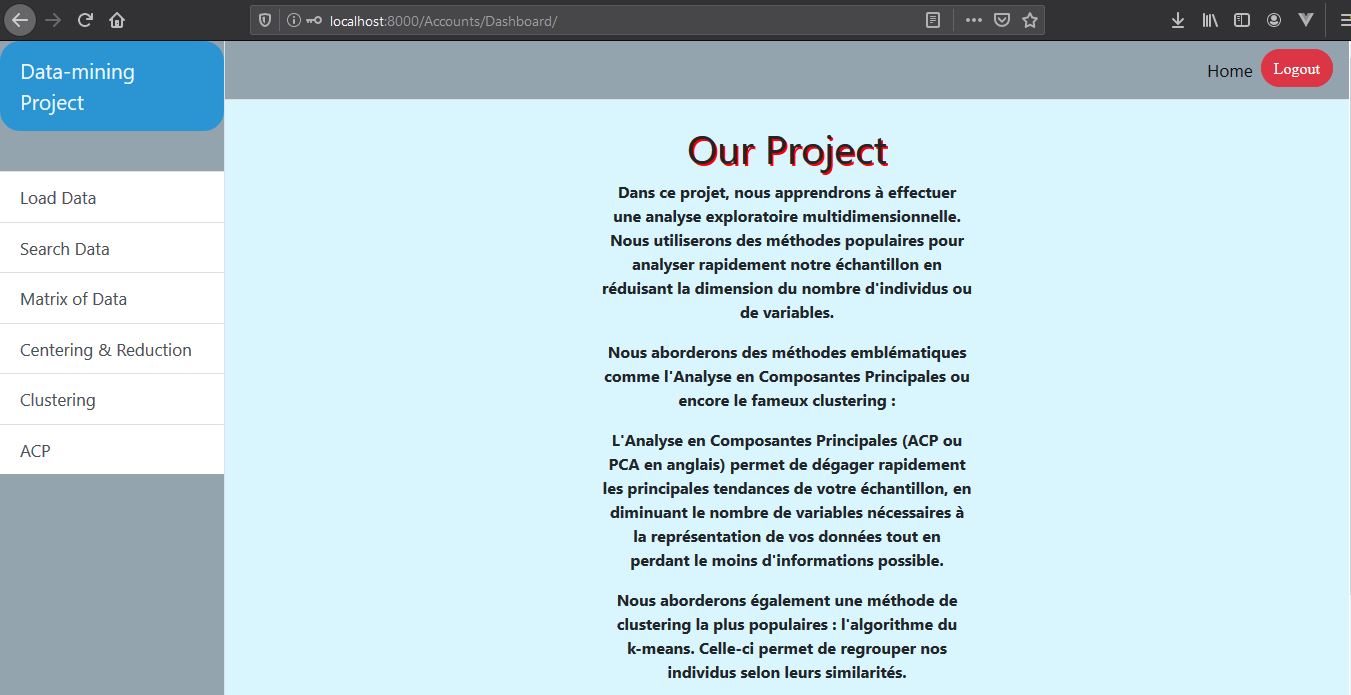
1. Stocker les fichier des données dans une base de données.
2. Lires les ficher stockés et affichage des données
3. Faire un Clustering avec nombre de cluster au chois
4. Réaliser une ACP avec un chois de nombre d’axes
5. Affichage des résultats.
6. System d’autentification

**Démonstration des fonctionnalité :**

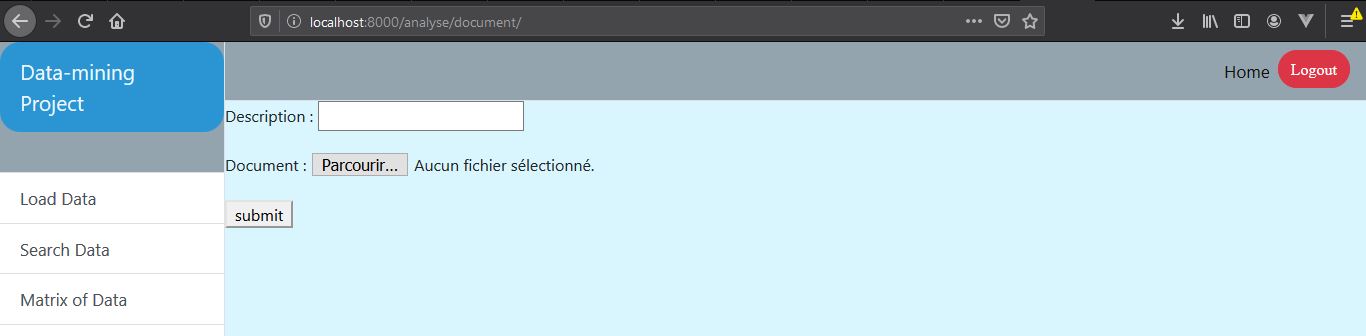
1. System d’autentification

****

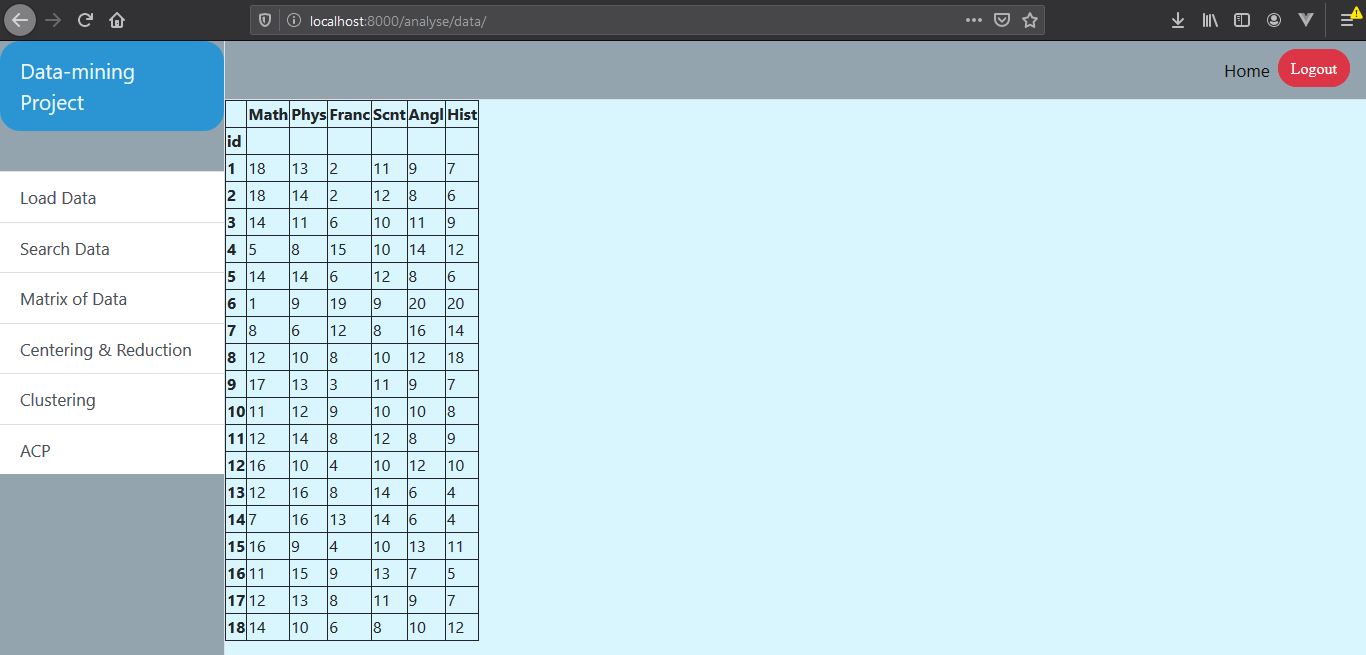
1. Description

****

1. Stocker les fichier des données dans une base de données.



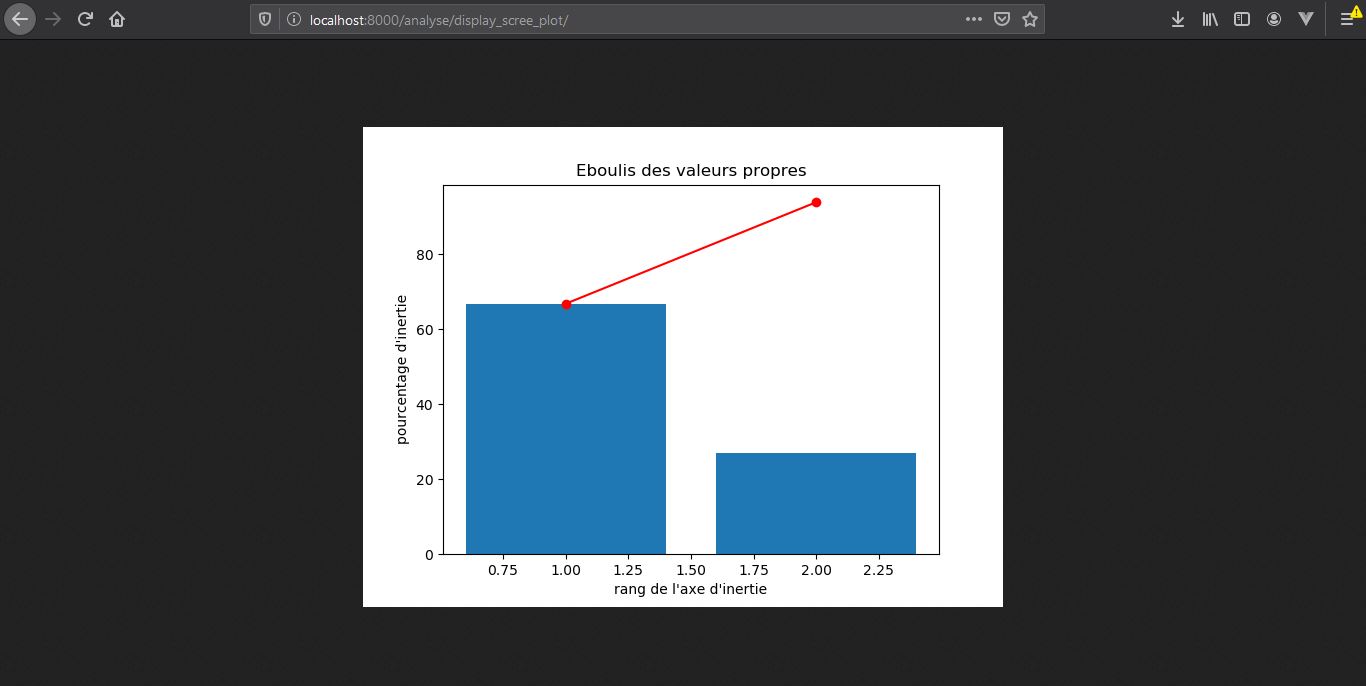
1. Lires les ficher stockés et affichage des données

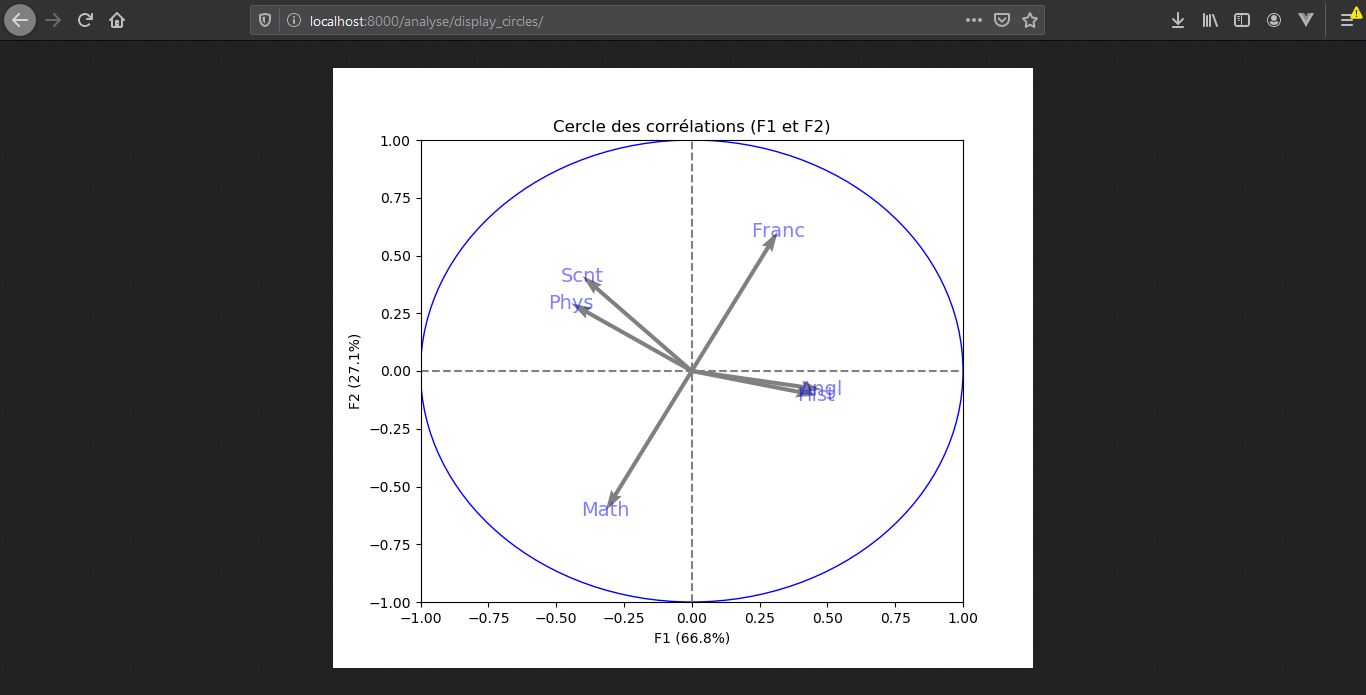


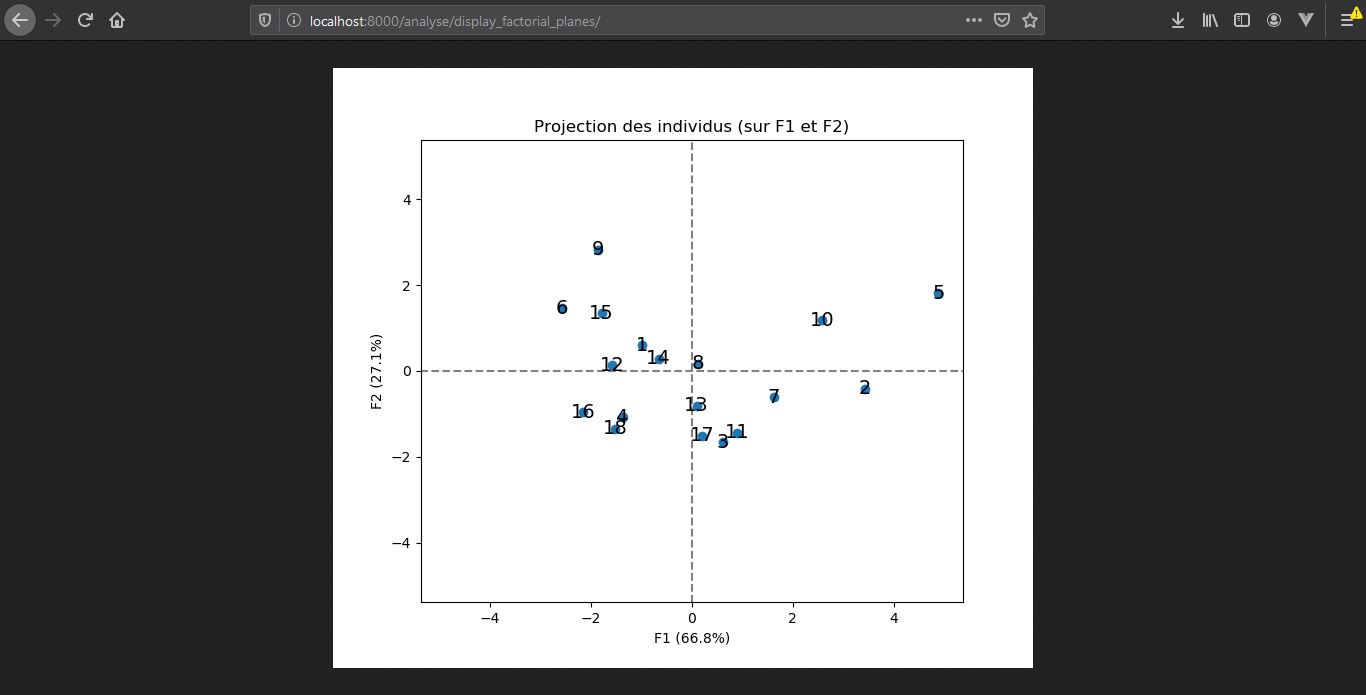
1. Faire un Clustering avec nombre de cluster au chois

Réaliser une ACP avec un chois de nombre d’axes

Affichage des résultats.

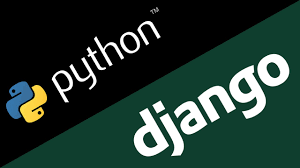




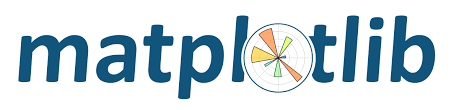


**Spécification téchnique :**









**Bibliographie**

1. <https://fr.wikipedia.org/wiki/>
2. <https://openclassrooms.com/fr/courses/4525281-realisez-une-analyse-exploratoire-de-donnees>
3. <https://docs.djangoproject.com/fr/3.0/>
4. (SKLEARN) ‘’scikit-learn – Machine Learning in Python’’, <http://scikit-learn.org/stable/index.html>
5. <http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/fr_Tanagra_ACP_Python.pdf>