**Table des matières**

[**Introduction**](#_24zjxbkhcsuc) **2**

[**1. Choix de de la base de données**](#_fpv3v19v8n58) **3**

[**2. Description détaillée des données**](#_xa3hk821vdug) **3**

[2.1. Attributs du jeu de donnée](#_2ne6kr7am6kr) 3

[**3. Analyse Exploratoire de chaque Attribut**](#_5t6mfdfamdfi) **3**

[**4.Analyse de lien entre chaque paire d’attributs.**](#_qz272tm06knp) **5**

[**Conclusion**](#_nudi8hohtd0s) **8**

[**Référence**](#_vvpmh9xch8n6) **9**

# Introduction

La fouille de données consiste à rechercher et extraire de l’information utile et inconnue sur des gros volumes de données stockées dans les bases ou des entrepôts de données. comprendre, analyser et prendre des décisions face à une telle masse de données devient quelque chose de plus complexe.

Alors le développement récent de la fouille de données (depuis le début des années 1990) est lié à ces contraintes à savoir: puissance de calcul importante est disponible sur des ordinateurs de bureau ou même à domicile; le volume des bases de données augmente énormément (la quantité de données dans le monde est estimée à doubler tous les 20 mois); l'accès aux réseaux de taille mondiale, ces réseaux ayant un débit sans cesse croissant qui rendent le calcul distribué et la distribution d’information sur un réseau d’échelle mondiale viable

Au fait, la finalité n’est juste pas d’avoir les données, mais d’arriver par des méthodes (Analyse en composantes principales, Analyse discriminante, Analyse factorielle, Classification automatique, Analyse en composantes indépendantes ) à donner les liens pouvant exister entre les données ainsi qu’en tirer une information statistique qui sert à décrire de façon plus succincte les principales informations contenues en elles. On peut aussi chercher à les classer en différents sous groupes plus homogènes.

Ces méthodes permettent surtout de manipuler et de synthétiser de l’information provenant des tableaux de données de grande taille. Pour cela, il est important de bien estimer les corrélations entre les variables qu’on étudie.

Dans ce rapport, il nous est demandé de faire :

* Choisir un jeu de données qui constitue l’objet d’études pour tous le temps du cours de fouilles de données;
* Faire une description détaillée de ce jeu de données;
* Faire une analyse exploratoire de chaque attribut;
* Faire une analyse de lien entre chaque paire d’attributs.

# **1.** **Choix de de la base de données**

Notre choix s’est porté sur le jeu de donnée **Iris Data Set** vu sa popularité et sa facilité dans le processus de fouille de données dont voici le lien de téléchargement :<http://archive.ics.uci.edu/ml/datasets/Iris>

# **2.** **Description détaillée des données**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | Ci-dessous est présenté un tableau qui relate les nombres d’attributs, leurs types, le nombre de valeurs manquantes, qui est d’ailleurs nulle, la tâche associée et les nombres d’instances dans la base de données |  |      |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | **Data Set Characteristics:** | Multivariate | **Number of Instances:** | 150 | **Area:** | Life | | **Attribute Characteristics:** | Real | **Number of Attributes:** | 4 | **Date Donated** | 1988-07-01 | | **Associated Tasks:** | Classification | **Missing Values?** | No | **Number of Web Hits:** | 1722921 | |

## **2.1. Attributs du jeu de donnée**

Dans notre jeu de donnée nous avons les attributs suivant :

* **sepal length in cm** : qui est de type float ou double, c’est la longueur du sépale.
* **sepal width in cm** : idem pour le type et est la largeur du sépale.
* **petal length** in cm : idem pour le type de donnée, c’est la longueur du pétale.
* **petal width** in cm : idem pour le type de donnée, c’est la largeur du pétale.
* **class** : qui est de type nominal ou catégoriel, qui est le type de fleur.

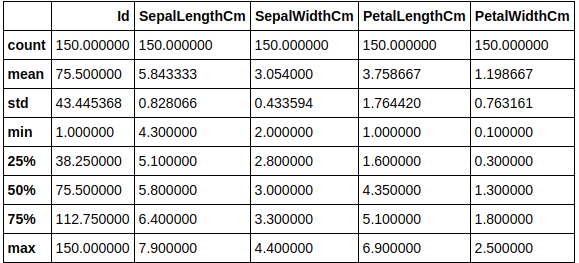
# **3. Analyse Exploratoire de chaque Attribut**

L’analyse exploration sur le jeu de donnée choisie va consister à trouver la médiane, la moyenne, l’écart-type, les quartiles, la valeur maximum, la valeur minimum.

Pour se faire nous avons utilisé un outil puissant afin de nous aider dans le processus de fouille de données. Notre choix s’est porter sur python, un des langages de programmation orienté objet le plus populaire dans le monde et qui a une forte communauté de développeur.

Nous l’avons associé avec une plateforme de développement python ANACONDA, qui est interactif et très approprié pour l’analyse exploratoire.

Ci-dessous, il nous est présenté le tableau qui regroupe les éléments qui ont été cité ci-haut:



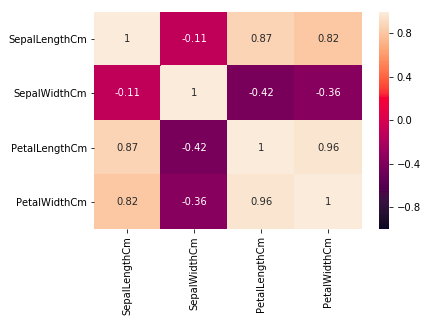
Le tableau ci-haut est le résultat de la méthode describe sur le jeu de donnée.

Etant donné que nous avons finis avec l’analyse exploratoire, nous allons poursuivre la dernière étape de ce rapport en analysant les liens qui existent entre les attributs.

# 4.Analyse de lien entre chaque paire d’attributs.

Dans cette section du rapport nous allons montrer les liens qui existent entre les attributs, d’un point de vu statistique cela revient à trouver la relation entre chaque paire des attributs et pour ce faire nous allons calculer la corrélation entre ces attributs.

* **Matrice de corrélation:**

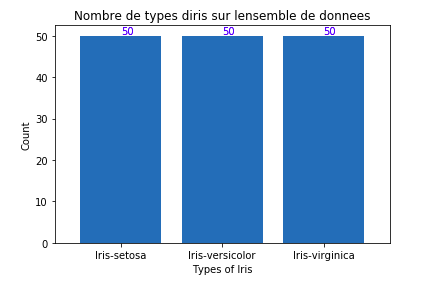


Cette matrice nous montre en effet la dépendance entre différents attributs, nous pouvons déduire que:

* La longueur du Sepal est fortement corrélée avec la longueur et la largeur de Petal avec des taux de corrélations respectifs de 0.87 et 0.82.
* Longueur du Pétale est fortement corrélée avec sa largeur avec un taux de corrélation de 0.96

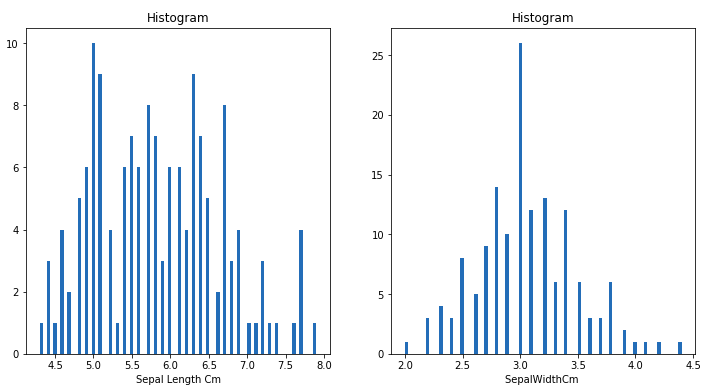
Les autres relations entre attribut sont faiblement corrélées et indépendantes entre elles.

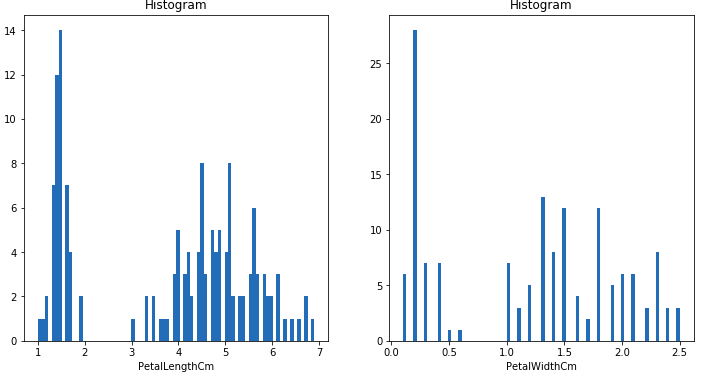
* **Histogrammes**
* **Histogramme de classe d’Iris**

****

cet histogramme nous montre la fréquence de chaque classe d’Iris; en fait il montre le nombre de fleur Iris dans chaque classe. Alors il y’a 50 fleurs dans chaque classe (Setosa, Versicolor et Virginica) d’Iris.

* **Histogramme des attributs**

****

****

cette histogramme montre la fréquence de données

# Conclusion

Bien que cela ne fut pas facile , nous avons en définitive répondu à toutes les questions posé dans ce rapport.En jetant un oeil de près au tableau de corrélation ,nous pouvons voir que le type ou la classe d’une fleur est plus dépendante de la longueur du sépale,la longueur de pétale et à la largeur du pétale

# Référence

* <http://archive.ics.uci.edu/ml/datasets/Iris>
* <http://www.suristat.org/article185.html>
* <https://alstatr.blogspot.com/2014/03/python-numerical-description-of-data.html>
* <https://machinelearningmastery.com/machine-learning-datasets-in-r/>
* <https://www.statcan.gc.ca/edu/power-pouvoir/glossary-glossaire/5214842-fra.htm>