Introduction

La Data science est en fait un terme générique englobant toute une famille d'outils facilitant l'exploration et l'analyse des données pour des fins décisionnelles.

Les techniques mises en action lors de l'utilisation de cet instrument d'analyse et de prospection sont particulièrement efficaces pour extraire des informations significatives depuis de grandes quantités de données.

En dépit des méthodes classiques d'analyses statistiques, Cet instrument d'analyse est particulièrement adapté au traitement de grands volumes de données et avec l'augmentation de la capacité de stockage des supports informatiques, un maximum de renseignements sera capté, ordonnés et rangés « Comportement des acheteurs, caractéristiques des produits, historisation de la production », désormais plus rien n'échappe à la collecte.

Le travail présenté dans ce rapport rentre dans ce cadre et consiste en la réalisation d’une étude sur plusieurs sources d’informations afin de faire une comparaison entre ces différentes techniques d’analyse.

Le présent rapport est organisé de la manière suivante :

Chapitre 1 : Etat de l’art

C’est la partie théorique où nous présentons les concepts de bases relatifs aux Machine Learning en abordant ses différentes approches.

Chapitre 2 : Réalisation de la solution

Ce chapitre est composé de deux parties à savoir :

* Partie 1 : Etude des données synthétiques
* Exploratoire préliminaire : qui consiste en l’utilisation des méthodes classiques d’analyses statistiques.
* Classification supervisée : Dans cette étape nous appliquerons et comparerons les différentes approches de classifications supervisées .
* Partie 2 : Etude des données réels :

Dans cette partie nous allons faire une comparaison entre les approches de classifications supervisées sur les données «visa premiere, credit card fraude »

Enfin nous terminons par une conclusion générale.

Chapitre 1 : Etat de l’art

Introduction :

L’exploitation de l’information est l’un des problèmes majeurs que rencontre les entreprises ce qui a fait naitre la data science.

Cette technologie fait appel à différentes approches à des fin d’exploitation et d’analyse de données, ces approches ce divisent en deux grande familles de machine learning «  suppervised learning » et « unsupervised learning ».

Dans notre cas nous allons aborder celle du suppervised learning.

Machine learning :

L’apprentissage automatique « machine learning » est une discipline qui se base sur les statistiques, probabilités, intelligence artificielle et l’optimisation tout en entrainant des algorithmes sur des données connues pour des fin décisionnelles et prédictives.

Suppervised learning :

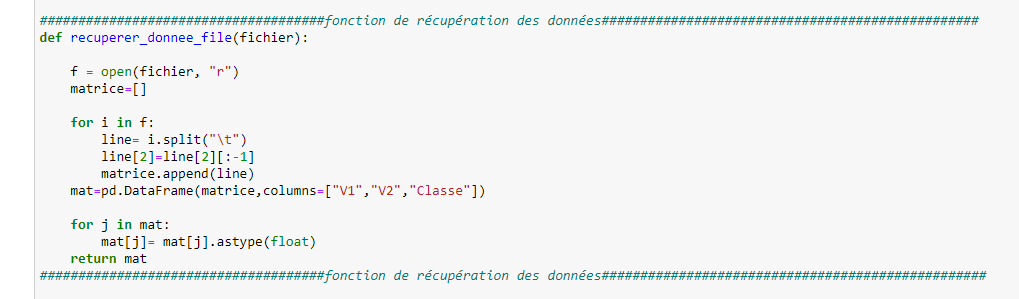
L’apprentissage supervisé «  suppervised learning » est le fait d’entrainer notre modèle afin qu’il puisse faire une liaison entre les inputs donnés et les output voulus.

L’apprentissage supervisé est constitué d’un ensemble d’approches de prédictions, dans notre travail nous avons fait appel à certaines d’elles à savoir : « KNN, Logistic regression, SVC, LDA, QDA, Naive Bayes, Random Forest, Decision Tree …etc. ».

KNN :

Partie 1 : données synthétiques

Tout d’abord nous allons commencer par la récupération des données via une fonction.



Après avoir récupérer les données nous allons les normaliser

