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 «  supervised learning » et « unsupervised learning ».

Dans notre cas nous allons aborder celle du supervised learning.

L’apprentissage automatique :

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.

L’apprentissage supervisé :

L’apprentissage supervisé «  supervised 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** (K-nearest neighbors) : l’algorithme du K plus proche voisin est une méthode pouvant être utilisée pour les cas de régression et de classification. L’idée est de pouvoir classer une donnée à partir d’un ensemble de données labélisés (classés). Les étapes de cet algorithme sont les suivantes :

* Choisir un nombre K de voisins.
* Calculer la distance (euclidienne, Manhattan…) de la donnée à classer aux autres données (labélisés).
* Prenez les K voisins ayant la plus petite distance par rapport à la donnée à classer.
* Déterminez à quelles catégories appartiennent les K données.
* Attribuez à la donnée à classer, la classe majoritaire parmi les K données.

**Bayésien-Naïf** : cette méthode de classification probabiliste repose, comme son nom l’indique, sur le théorème de Bayes (fondé sur les probabilités conditionnelles). Cette méthode devra attribuer une classe à une donnée en calculant la probabilité que cette donnée appartienne à telle ou telle classe sachant un certain nombre de caractéristiques. La classe correspondant à la probabilité la plus grande sera ensuite sélectionnée. Cette méthode est dite « naïve » puisqu’on suppose que les variables explicatives sont indépendantes, ce qui est en réalité faux.

**Régression Logistique :** cet algorithme d’apprentissage automatique est le plus couramment utilisé pour les problèmes de classification binaire, c’est-à-dire pour les classes pouvant prendre les valeurs « OUI/NON », « VRAI/FAUX », « A/B » … Cette méthode consiste à prédire la probabilité qu’un évènement arrive ou non et à déterminer une relation entre les variables explicatives (caractéristiques) et la variable à expliquer.

**SVM linéaire** (séparateur à vaste marge linéaire) : pour déterminer la classe d’un objet, il faut connaître la frontière séparant les classes (sur un plan) afin de déterminer à quelle catégorie appartient cet objet. Cette frontière est justement déterminée par le SVM, qui va faire en sorte de la placer le plus loin possible des points d’entrainement.

**SVM non linéaire** : ici, il est impossible de trouver une ligne droite qui permet de séparer les données. Il faut donc trouver une transformation qui va permettre de classer les données (on appelle cette méthode l’astuce du noyau).

**CART (Classification and Regression Tree) :** Les arbres de régression (Regression tree) permettent de prédire une valeur réelle (donnée quantitative) et les arbres de classification permettent de déterminer à quelle classe appartient une donnée. Ces arbres sont appelés arbres de décision et sont construits de manière itérative où sont appliquées des règles, test à chaque nœud. Chaque branche représente le résultat du test et les feuilles représentent les différentes valeurs ou classes possibles pour la variable à prédire.

**Random Forest** : cet algorithme utilise un ensemble d’arbres de décision indépendants. Il fonctionne selon le principe du bagging, c’est-à-dire qu’on divise les données en plusieurs sous-ensembles aléatoirement constitués, on entraîne un modèle sur chaque sous-ensemble puis avec les prédictions obtenues sur les différents arbres, on détermine le résultat en choisissant celui qui a la catégorie la plus fréquente ou bien en faisant la moyenne des valeurs prédites.

**LDA (Linear Discriminant Analysis)** : On modélise la distribution des variables explicatives par une loi de probabilité gaussienne puis on détermine les paramètres de la loi. Puis on applique la loi de Bayes pour déterminer la probabilité d’une classe sachant les variables explicatives.

**QDA (Quadratic Discriminant Analysis)** : elle constitue une généralisation de la LDA, sauf qu’ici on ne considère pas la matrice de covariance comme indépendante de la classe.

Chapitre 2 : Réalisation de la solution

Introduction

Après avoir terminé l’état de l’art, nous entamons la partie réalisation et la mise en œuvre de la solution. Ce chapitre présente les étapes de réalisation de notre solution et qui sera diviser en deux parties de traitement de données à savoir : »Données synthétiques et réelles ».

Partie 1 : Données synthétiques

Descriptif des données :

Dans cette partie nous avons utilisé trois datasets « spiral, flame et aggregation » décris comme suit :

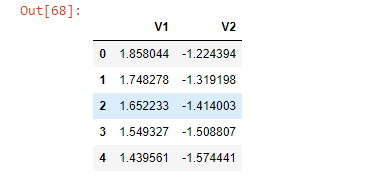
* Spiral : Ce dataset contient 312 observations et 3 variables d’études sachant que l’une d’elles et la variable à expliquer et qui contient 3 classes.
* Flame : Ce dataset est constitué de 240 observations, 2 variables et une seule caractéristique qui est composée de 2 classes.
* Aggregation : Ce dataset contient 788 observations et 3 variables tel que l’une d’elles et caractérisée par 7 classes.

Etude exploratoire préliminaire :

Dans cette partie nous allons effectuer une étude statistique sur les données de chaque dataset afin de savoir leurs comportement.

Données spiral :

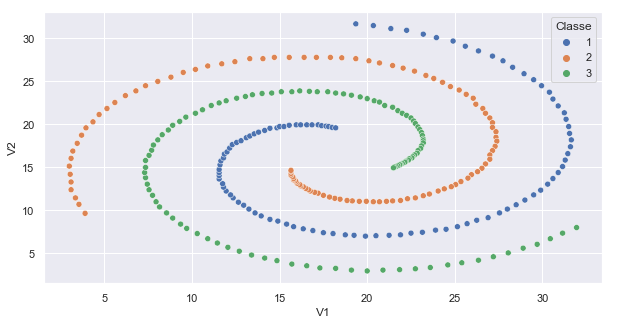
1. Normalisation des données :



1. Visualisation du nuage de points

Qu’est-ce qu’un nuage de points :

L’interprétation d’un nuage de point sert à vérifier certaine critères entre deux variables à savoir « intensité, affinité, corrélation, les points aberrants ».

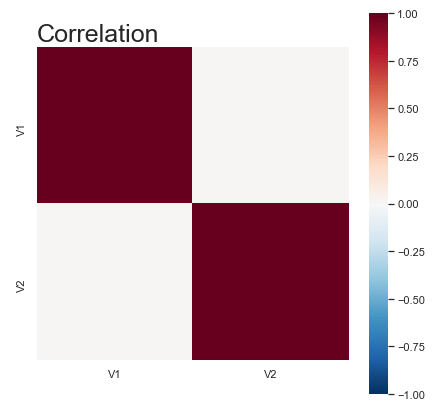


D’après le graphe ci-dessus nous trouvons aucune corrélation, affinité ou points aberrants entre les deux variables.

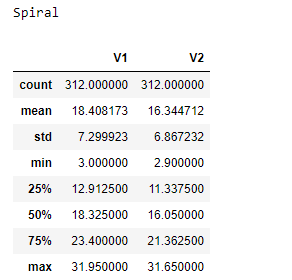
1. Matrice de corrélation

Après avoir fait la description du nuage de points nous pouvons vérifier la corrélation avec la matrice de corrélations.

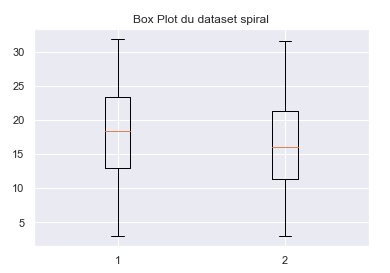
D’après la figure 2 nous trouvons que les 2 variables ne sont pas corrélées entre elles.



1. La description statistique des variables et le boxplot

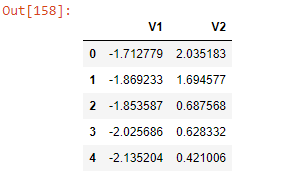


Dans le tableau ci-dessus nous trouvons la moyenne, l’écart-type, min, max, le premier et troisième quartile et la médiane des deux variables qui sont illustrés dans la figure 4.

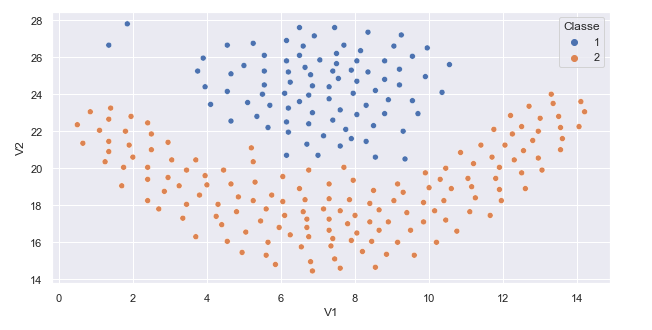


Données flame :

1. Normalisation des données :



1. Visualisation du nuage de points

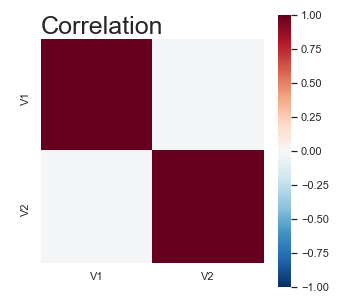


D’après le graphe ci-dessus nous trouvons aucune corrélation, affinité ou points aberrants entre les deux variables.

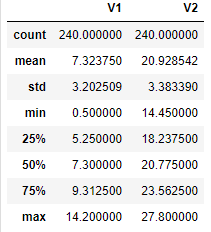
1. Matrice de corrélation

Après avoir fait la description du nuage de points nous pouvons vérifier la corrélation avec la matrice de corrélations.

D’après la figure 2 nous trouvons que les 2 variables ne sont pas corrélées entre elles.

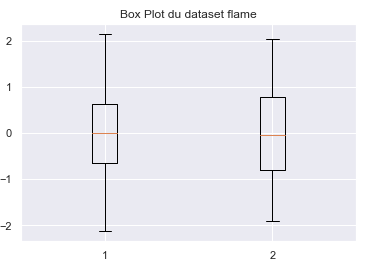


1. La description statistique des variables et le boxplot



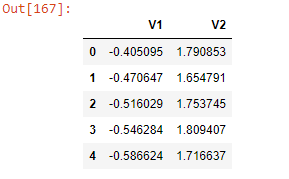
Le tableau de la figure ci-dessus résume les différents comportements de chaque variable à savoir :  « la moyenne, l’écart-type, min, max, le premier et troisième quartile et la médiane ».

Ces différentes valeurs sont illustrées ci-dessous.



Données Aggregation :

1. Normalisation des données :



1. Visualisation du nuage de points

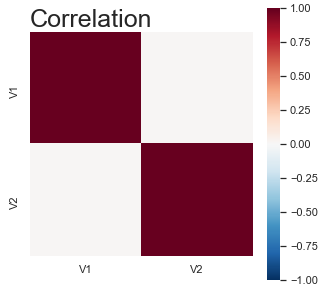


D’après le graphe ci-dessus nous trouvons aucune corrélation, affinité ou points aberrants entre les deux variables.

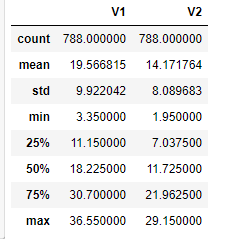
1. Matrice de corrélation

Après avoir fait la description du nuage de points nous pouvons vérifier la corrélation avec la matrice de corrélations.

D’après la figure 2 nous trouvons que les 2 variables ne sont pas corrélées entre elles.

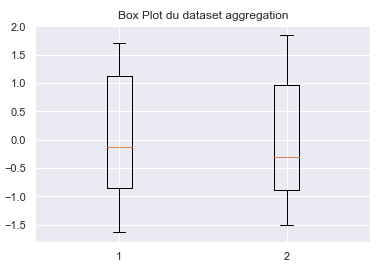


1. La description statistique des variables et le boxplot



Le tableau de la figure ci-dessus résume les différents comportements de chaque variable à savoir :  « la moyenne, l’écart-type, min, max, le premier et troisième quartile et la médiane ».

Ces différentes valeurs sont illustrées ci-dessous.



Comparaison des méthodes :

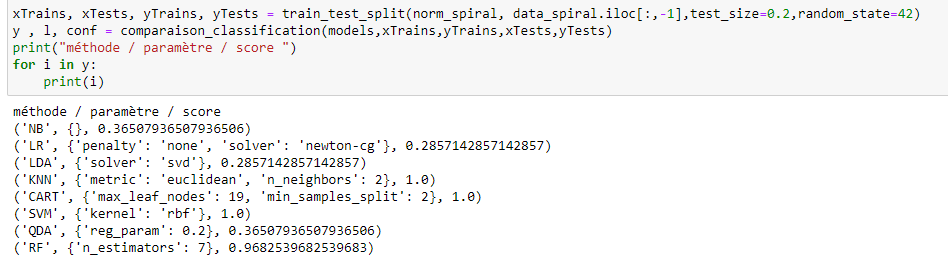
Dans cette partie nous avons comparé les différents modèles de classification cités dans le chapitre précédents afin de sortir avec le meilleur d’entre eux pour chaque dataset.

Pur faire la comparaison nous avons créé trois fonctions à savoir :

* Fonction retournant les meilleurs paramètres de chaque modèle.
* Fonction retournant Les résultats de chaque modèle
* Fonction retournant les meilleurs modèles avec un rapport de classification.

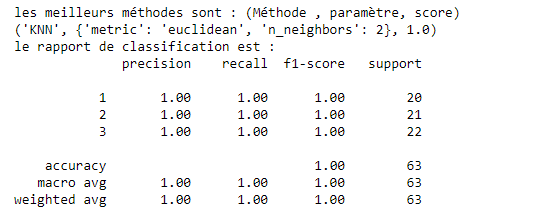
Dataset spiral :

Le résultat des différentes méthodes :



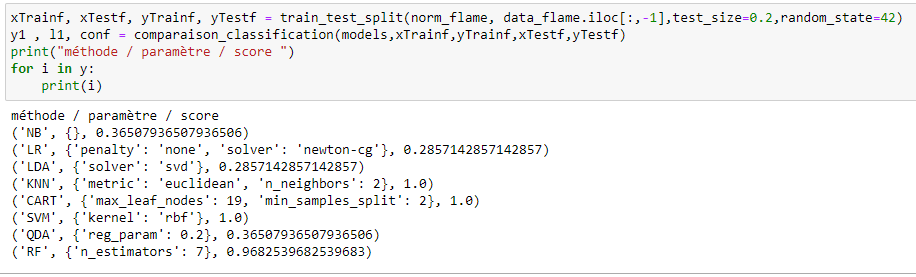
La prise des meilleures méthodes :

Les meilleures méthodes ont été prises en faisant une comparaison entre les valeurs obtenues sur le test «  nous avons pris le max »



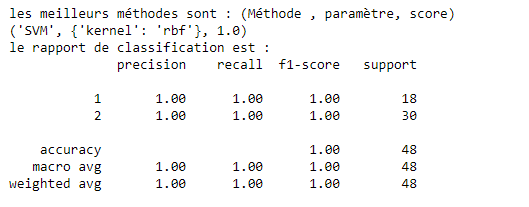
Data flame :

Le résultat des différentes méthodes :



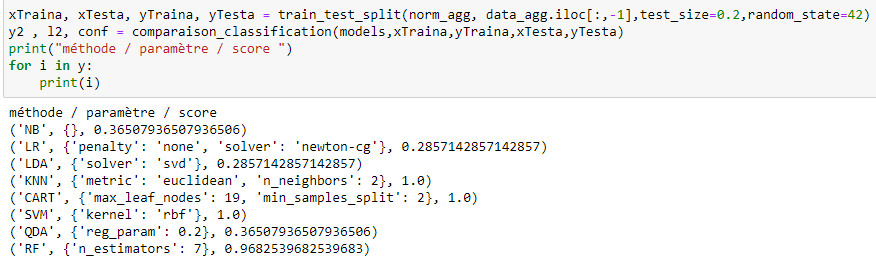
La prise des meilleures méthodes :

Les meilleures méthodes ont été prises en faisant une comparaison entre les valeurs obtenues sur le test «  nous avons pris le max »



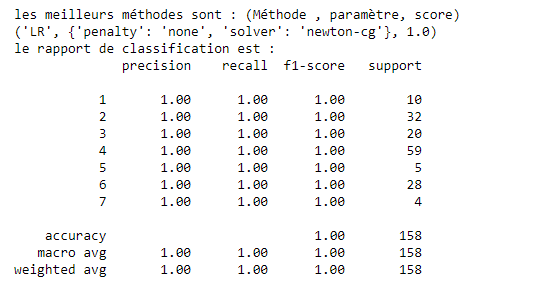
Data Aggregation

Le résultat des différentes méthodes :



La prise des meilleures méthodes :

Les meilleures méthodes ont été prises en faisant une comparaison entre les valeurs obtenues sur le test «  nous avons pris le max »



Partie 2 : Données réelles

Dans cette partie nous allons faire une étude comparative sur les données réelles «  VisaPremier, Creditcard ».

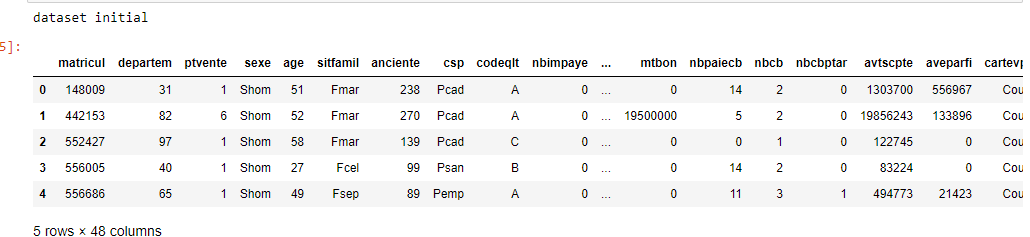
VisaPremier :

Ce dataset est composé de 1073 individus, 47 variables tel qu’une à expliquer qui est binaire.

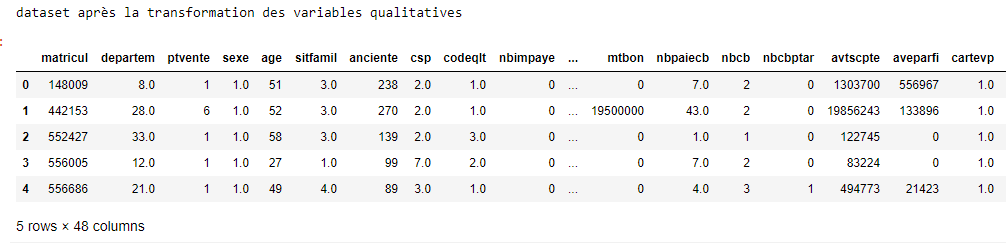
Traitement du dataset :

1. Encodage des variables qualitatives :

Données initiales



Données après la transformation des variables qualitatives



1. La suppression des variables insignifiantes :

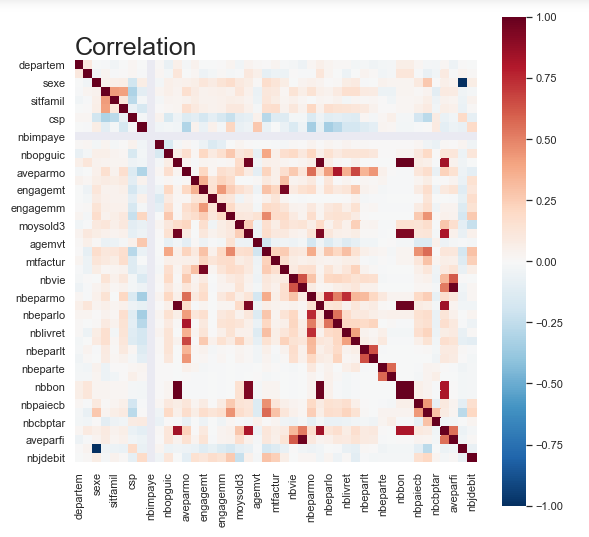
En premier nous avons supprimé les variables :

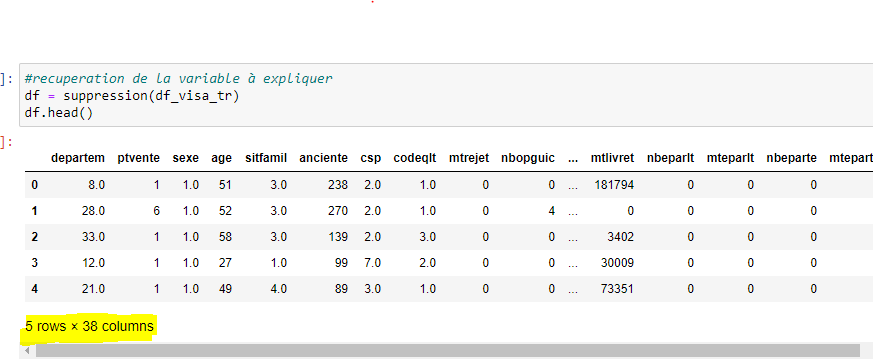
* Variable à expliquer : Après avoir récupéré la variable « cartevpr » nous l’avons supprimé.
* Variable matricul : La variable « matricul » n’apporte aucune signification aux analyses car ce n’est qu’un identifiant des clients.

Puis, nous avons entamé l’étape de suppression des variables corrélées comme suit :

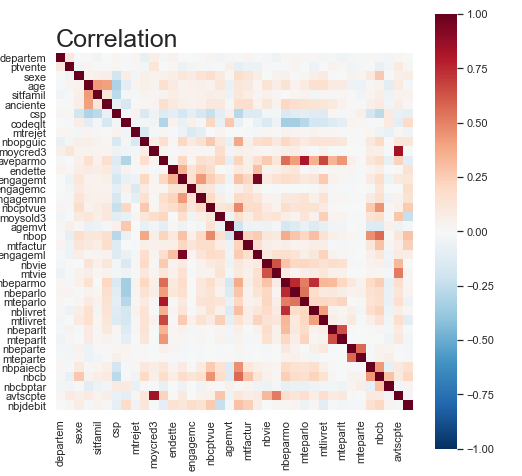
1. L’application de la matrice de corrélation :

Vu que les variables sont qualitatives donc nous avons appliqué la corrélation afin de supprimer les doubles, tel que le seuil utilisé est «  0.95, -0.95 »

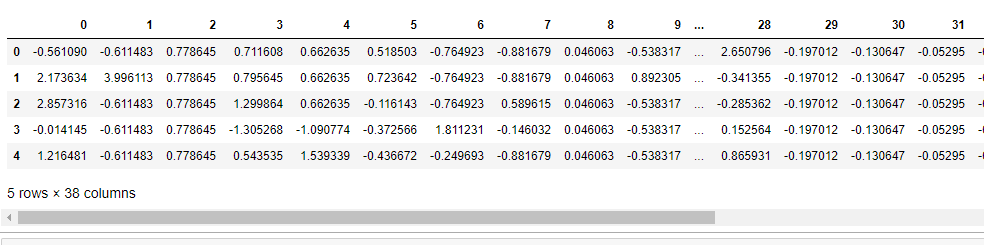




Après avoir supprimé les variables, nous allons appliquer la matrice de corrélation afin de montrer que toutes les autres variables sont décorrélées.

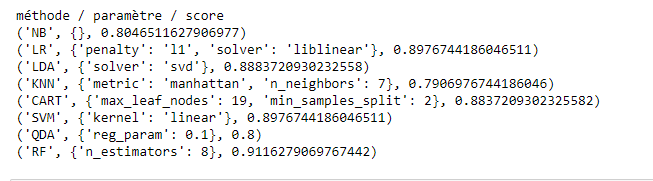


1. La normalisation des données :



1. Comparaison des modèles

Le résultat des différentes méthodes :



La prise des meilleures méthodes :

Les meilleures méthodes ont été prises en faisant une comparaison entre les valeurs obtenues sur le test «  nous avons pris le max »

