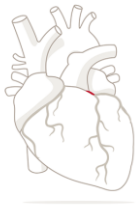
« Bio-Informatique, Modélisation des Systèmes Complexes Appliquée à la Santé »

Module: Data && Web Mining

Sujet:

Data Mining and Machine Learning techniques in Heart Disease Prediction



Réalisé par :

- BENLAMLIH Othmane
- EL ABBADI Youness
- MAISSOU Rahma
- OUHMOUK Maryem

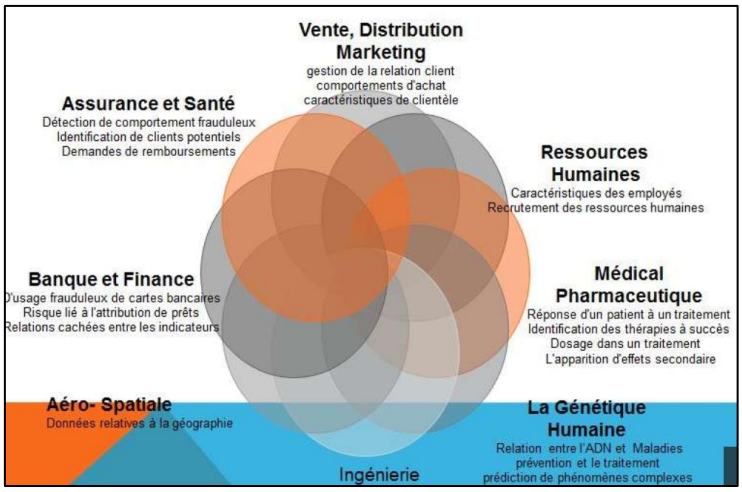
Supervisé par :

Pr H. BENBRAHIM

Année universitaire : 2020-2021

INTRODUCTION

Le terme de *Data Mining* est un terme anglo-saxon qui peut être traduit par « exploration de données » ou « extraction de connaissances à partir de données ». Ainsi le *Data Mining* consiste en une famille d'outils -- qu'ils soient automatiques ou semi-automatiques -- permettant l'analyse d'une grande quantité de données contenues dans une base. Objectif : faire apparaître des corrélations entre des phénomènes en apparence distincts afin d'anticiper des tendances.



Comme vous voyez parmi les domaines de Data Mining on a Médical Pharmaceutique, génétique Humain, Assurance et santé alors il participe d'une manière très importante dans le volet Biologie, santé et biomédical.

En règle générale, le terme **Data Mining** désigne l'analyse de données depuis différentes perspectives et le fait de transformer ces données en informations utiles, en établissant des relations entre les données ou en repérant des patterns.

Pour mener à bien un projet de *Data Mining*, il faut évidemment d'abord définir clairement la problématique à étudier. Ensuite, il est crucial de sélectionner parmi l'ensemble des données disponibles, celles qui pourront être utilisées. C'est-à-dire celle dont la qualité ne laisse aucune place au doute. Vient alors l'étape de paramétrage du modèle construit à partir de techniques issues des méthodes statistiques, des analyses de données et de l'informatique. Enfin, il faut procéder à l'étude des résultats. Les logiciels ne sont en effet pas autosuffisants et l'intervention d'un analyste spécialisé reste indispensable.

TABLE DES MATIERES

INTRODUCTION 2
Table des matières 3
Liste des figures et tableaux 5
I. Problématique 6
1. Les maladies cardiovasculaires 6
a. Définition 6
b. Les causes 6
2. Objectif 7
3. Data 7
a. Data description 7
b. Data exploration 9
II. Méthodes et techniques 11
1. Méthodes de prétraitement 11
a. Concaténation de data (Talend) 11
b. Les données manquantes 14
c. Analyse en composantes principales (ACP) 16
d. Les points aberrants 18
e. Standardisation des données 19
2. algorithmes du Machine Learning 21
a. Régression logistique 21
b. Random Forest 21
c. Arbre de décision 22
d. Naive Baiyes 22
e. SVM (Support vector machine) 23
f. Algorithme KNN (k-nearest neighbors) 23
III. MODELISATION ET EVALUATION 24
1. Importation des bibliothèques 25
2. Outils d'evaluation 25
a. Cross Validation (**K-fold**) 25
b. Matrice de confusion 26
3. Résultats 27
1. Régression logistique 27
2. Random Forest 28

3. Arbre de decision 29
4. Naive Baiyes 30
5. SVM (Support vector machine) 31
6. Algorithme KNN (k-nearest neighbors) 32
7. BENCHMARK algorithmes 33
CONCLUSION 34
REFERENCES 35

LISTE DES FIGURES ET TABLEAUX

Tableau 1: Représentation des champs avec la description 8

Figure 1:Une représentation des variables avec les données. 7
Figure 2:Distribution de la maladie 9
Figure 3:Distrubution au niveau de sexe 9
Figure 4:Relation entre le sexe et la maladie 10
Figure 5:Distribution de l'âge 10
Figure 6: Relation entre l'activité du cœur et la maladie 11
Figure 7:Importation_data 12
Figure 8 : importation de composant tUnite 12
Figure 9 la liaison des datas avec tUnite. 13
Figure 10 : choix des colonnes des datas qu'on veut fusionner 13
Figure 11 : compilation de modèle 13
Figure 12 : résultat de model. 14
Figure 13 : Sauvegarde de résultat dans un fichier csv. 14
Figure 14:Exemple de manque de data 15
Figure 15: Résultats de calcul de la moyenne. 15
Figure 16:Résultat après la modification. 16
Figure 17:Informations significatives 17
Figure 18:Cercle des corrélations 17
Figure 19:Représentation de l'âge des individus. 18
Figure 20:Le calcul des points aberrants 19
Figure 21:Résultats de standardisation. 20
Figure 22: Schéma descriptive de Random Forest 21
Figure 23: Schéma descriptive d'arbre de décision 22
Figure 24: Schéma descriptive de SVM 23
Figure 25: Schéma descriptive de K-NN 24
Figure 26: les étapes de CRISP-DM. 24
Figure 27:Matrice de confusion 26

I. PROBLEMATIQUE

1. LES MALADIES CARDIOVASCULAIRES

a. Définition

Les maladies cardiovasculaires regroupent les pathologies qui touchent le cœur et l'ensemble des vaisseaux sanguins, comme l'athérosclérose, les troubles du rythme cardiaque, l'hypertension artérielle, l'infarctus du myocarde, l'insuffisance cardiaque ou encore les accidents vasculaires cérébraux.

Les maladies cardiovasculaires (MCV) sont la première cause de décès dans le monde, faisant environ 17,9 millions de morts chaque année. Quatre décès dus à la MCV sur 5 sont dus à des crises cardiaques et des accidents vasculaires cérébraux, et un tiers de ces décès surviennent prématurément chez des personnes de moins de 70 ans.

Les personnes à risque de maladie cardiovasculaire peuvent présenter une pression artérielle, une glycémie et des lipides élevées ainsi qu'un surpoids et une obésité. Ceux-ci peuvent tous être facilement mesurés dans les établissements de soins primaires. Identifier les personnes les plus à risque de maladies cardiovasculaires et s'assurer qu'elles reçoivent un traitement approprié peut prévenir les décès prématurés. L'accès aux médicaments essentiels contre les maladies non transmissibles et aux technologies de santé de base dans tous les établissements de soins de santé primaires est essentiel pour garantir que ceux qui en ont besoin reçoivent un traitement et des conseils.

b. Les causes

Les maladies cardiovasculaires sont favorisées par :

- Le tabagisme ;
- l'obésité ;
- Une mauvaise alimentation ;
- l'excès d'alcool ;
- le manque d'activité physique ;
- L'hypertension artérielle et le diabète sont aussi des facteurs de risque.

2. OBJECTIF

L'apprentissage des composantes de risque liées aux maladies cardiaques aide les experts des services médicaux à reconnaître les patients à haut risque de maladie cardiaque. L'analyse statistique a identifié les facteurs de risque associés au cœur maladie à l'âge, tension artérielle, cholestérol total, diabète, hypertension, antécédents familiaux de maladie cardiaque, l'obésité et le manque d'exercice physique, la glycémie à jeun, etc.

Le diagnostic médical joue un rôle vital et pourtant une tâche compliquée qui doit être exécutée de manière efficace et précise.

Pour réduire le coût de réalisation des tests cliniques, une application des méthodes informatique appropriée ainsi qu'une aide à la décision devrait être intégré.

Les chercheurs ont appliqué différentes techniques de data mining et de machine learning pour aider les experts en services médicaux à progresser dans le jugement des maladies cardiaques. Le réseau neuronal, l'arbre de décision, etc. sont quelques techniques utilisées dans le diagnostic des maladies cardiaques.

3. DATA

a. Data description

La figure ci-dessous illustre une représentation des variables avec les données.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	target	ca	thal
0	63	1	3	145	233	1	0	150	0	2.3	0	1	0.0	1.0
1	37	1	2	130	250	0	1	187	0	3.5	0	1	0.0	2.0
2	41	0	1	130	204	0	0	172	0	1.4	2	1	0.0	2.0
3	56	1	1	120	236	0	1	178	0	0.8	2	1	0.0	2.0
4	57	0	0	120	354	0	1	163	1	0.6	2	1	0.0	2.0

Figure 1:Une représentation des variables avec les données.

Le tableau ci-dessous représente chaque champ avec sa description.

Tableau 1: Représentation des champs avec la description

Champ	Description
age	âge en années
sex	sexe (1 = homme; 0 = femme).
cp	Type de douleur thoracique : Valeur 1 : angine typique Valeur 2 : angine de poitrine atypique Valeur 3 : douleur non angineuse Valeur 4 : asymptomatique
trestbps	tension artérielle au repos (en mm Hg lors de l'admission à l'hôpital).
chol	cholestérol en mg/dl
fbs	(glycémie à jeun > 120 mg/dl) (1 = vrai ; 0 = faux).
restecg	les résultats de l'électrocardiographie au repos.
thalach	la fréquence cardiaque maximale atteinte
exang	l'angine induite par l'exercice (1 = oui ; 0 = non).
oldpeak	Dépression ST induite par l'exercice par rapport au repos.
slope	La pente du segment ST de l'exercice de pointe : Valeur 1 : en hausse Valeur 2 : plat Valeur 3 : en baisse
Target	si la personne et malade ou non
ca	nombre de grands navires (0-3) colorés par flourosopie.
thal	0-3 = normal; 3-6 = fixed defect; 6-7 = reversable defect.

b. Data exploration

Dans cette partie on va explorer notre data, on se focalise sur la distribution et la relation entre leur attribut.

- Distribution de la maladie

En va commencez par l'étude de la distribution de data de sortie (output) :

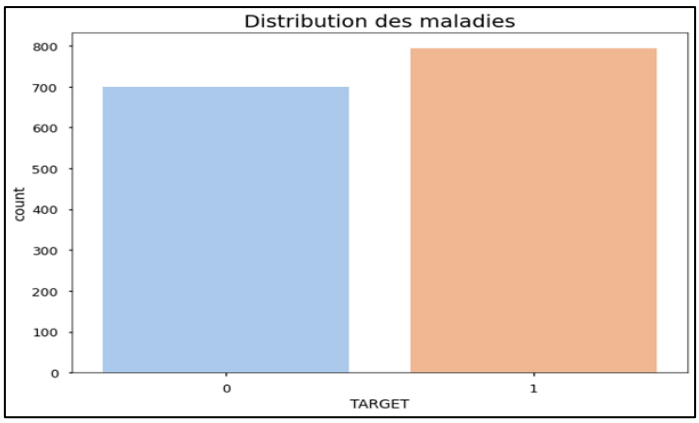


Figure 2:Distribution de la maladie

Nous constatons que notre base de données est assez équilibrée, ou les cas de maladie cardiovasculaires représenté par presque de 800 cas par rapport à 690 du cas normal, avec une infériorité de presque de 100 cas.

- Distribution de sexe

Dans cette partie en illustre un diagramme circulaire qui montre le pourcentage des participants au test de maladies cardiovasculaires.

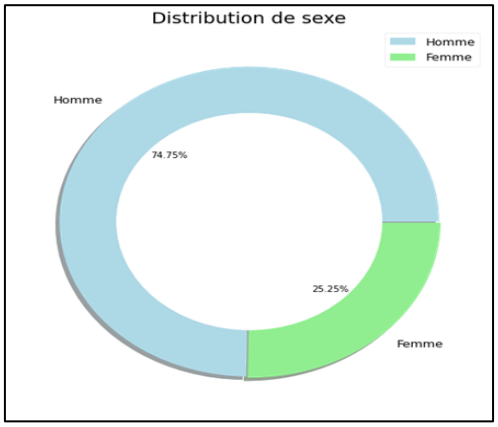


Figure 3:Distrubution au niveau de sexe

Selon notre diagramme, en remarque que 74,75% des patients sont des hommes et 25,25% des femmes, ce qui montre que plus d'hommes ont participé aux tests de maladies cardiovasculaires plus que les femmes.

- Relation entre le sexe et la maladie

Ce diagramme circulaire montre le pourcentage des hommes et femmes qui ont la maladie

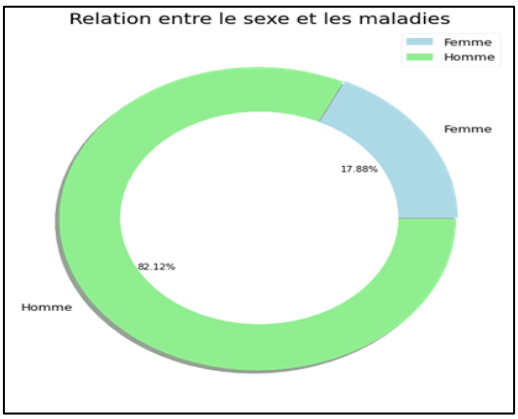


Figure 4:Relation entre le sexe et la maladie

- Distribution de l'âge

Ce diagramme montre les personnes qui ont la maladie cardiovasculaire en fonction de l'âge.

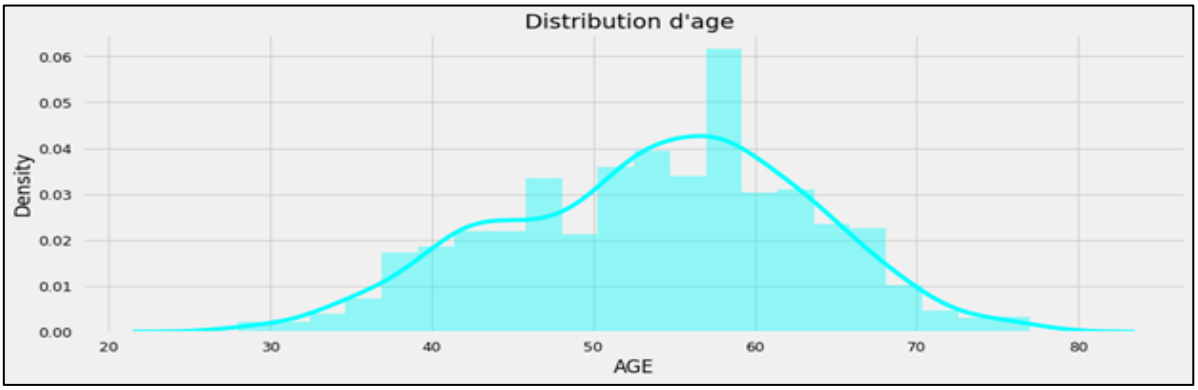


Figure 5:Distribution de l'âge

Le diagramme montre que le nombre le plus élevé de personnes atteintes de maladies cardiovasculaires se situe dans la tranche d'âge des 50-65 ans. Les patients dans la tranche d'âge 20-30 ans sont beaucoup moins susceptibles de souffrir. Le nombre de patients dans la tranche d'âge 65-80 ans étant très faible, la répartition est également plus restreinte.

- Relation entre l'activité électrique du cœur et la maladie

Ce diagramme montre que les patients présentant une activité électrique du cœur représentant une hypertrophie du ventricule gauche sont plus susceptibles d'avoir une maladie cardiovasculaire, tandis que les patients ayant une activité normale sont peu susceptibles d'avoir des problèmes cardiovasculaires.

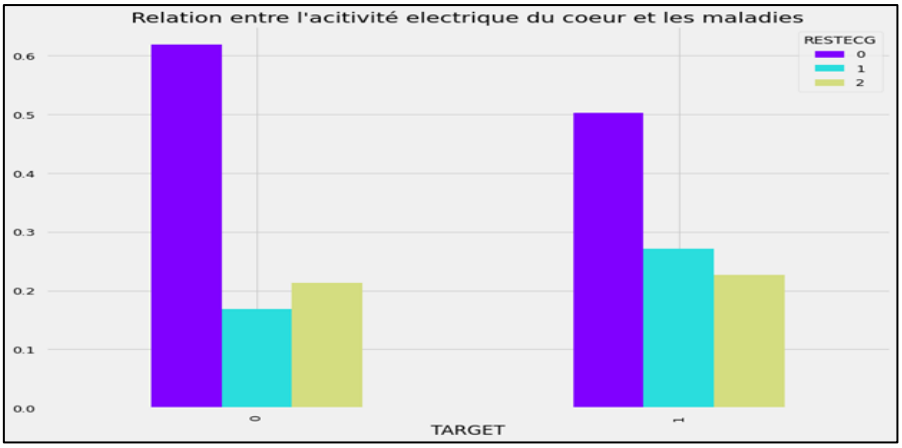


Figure 6: Relation entre l'activité du cœur et la maladie

II. METHODES ET TECHNIQUES

1. METHODES DE PRETRAITEMENT

a. Concaténation de data (Talend)

La concaténation et la séparation d'observations des variables sont deux manipulations courantes dans la réorganisation des données, servant à refaçonner les jeux de données afin de mieux les exploiter.

Alors que **Talend** est un éditeur de logiciel spécialisé dans l'intégration de données.

La société fournit des logiciels et services dans les domaines de l'intégration de données, de la gestion des données, du Master Data Management (gestion des données maître), de la qualité de données, de la préparation des données et de l'intégration d'applications, du Big Data, du Cloud [4].

- Les étapes suivies pour la concaténation des donnés

Premièrement on importe notre data :

Figure 7:Importation_data

Deuxièmement, on importe tUnite qui va nous aider à fusionner les données :

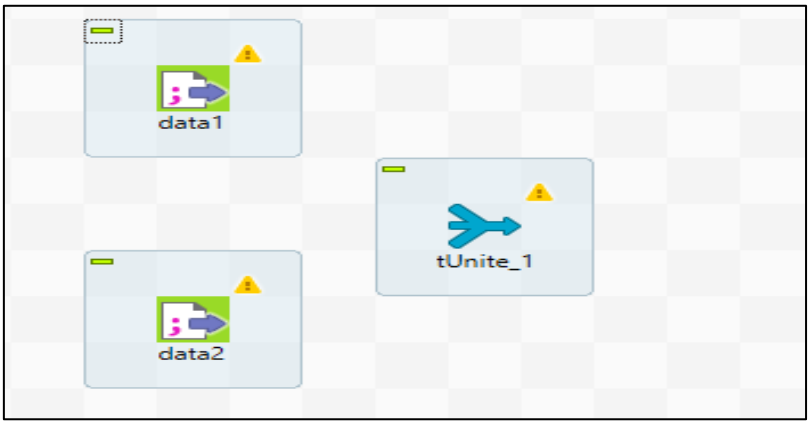


Figure 8 : importation de composant tUnite

Troisièmement, on fait une liaison entre les données qu'on veut fusionner avec **tUnite** et on ajoute **tLogRow** pour afficher notre nouvelle data :

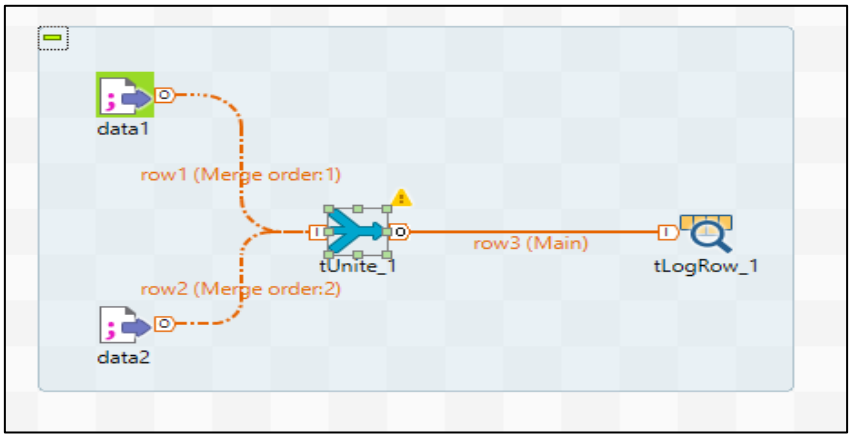


Figure 9 la liaison des datas avec tUnite.

Et puis, on choisit les colonnes de data qu'on veut fusionner :

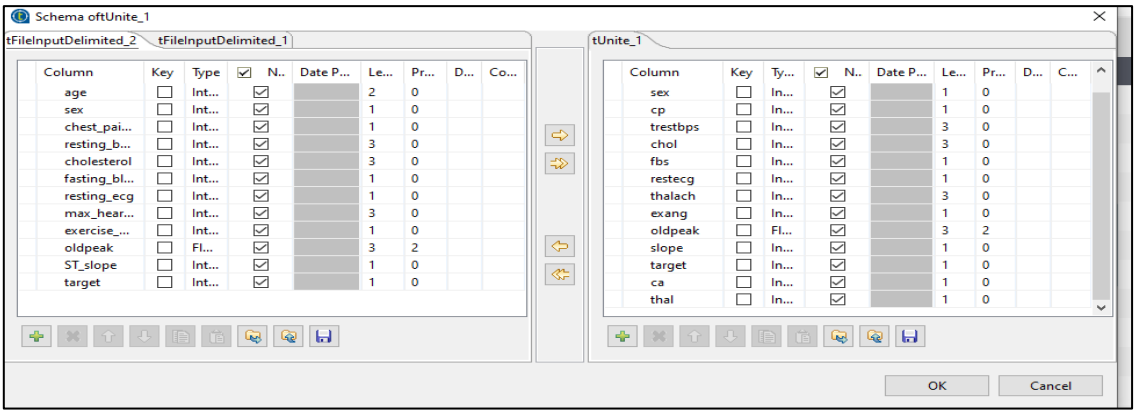


Figure 10 : choix des colonnes des datas qu'on veut fusionner

Ensuite, on exécute notre modèle et on visualise la nouvelle data :

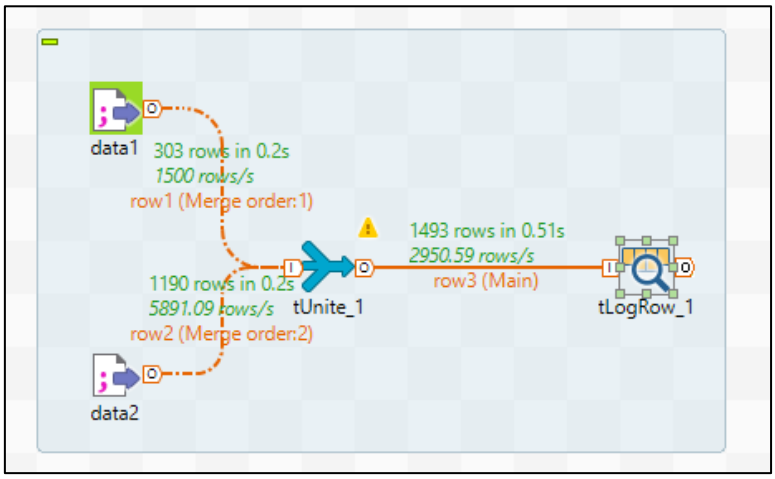


Figure 11 : compilation de modèle

On observe que notre nouvelle data contient 1493 lignes qui sont fusionnés dans un temps de 0.51s.

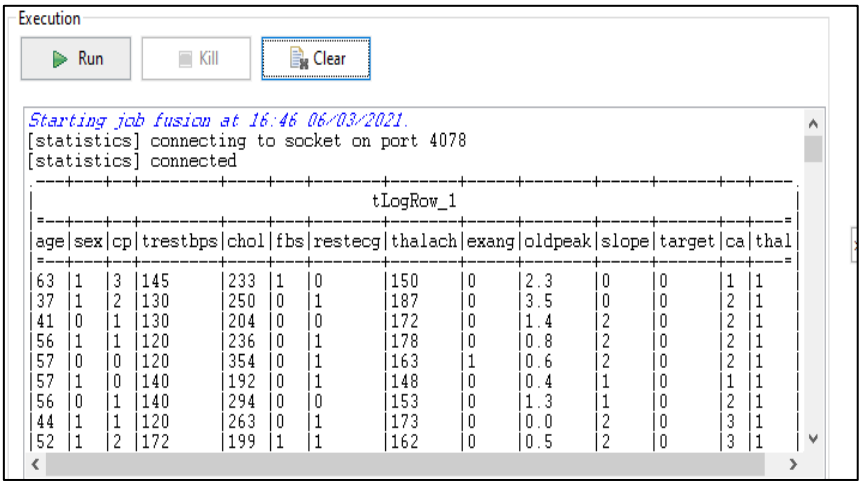


Figure 12 : résultat de model.

Finalement, on ajoute le fichier **tOutputDelimited** dans lequel on va stocker les résultats, et on exécute le modèle.

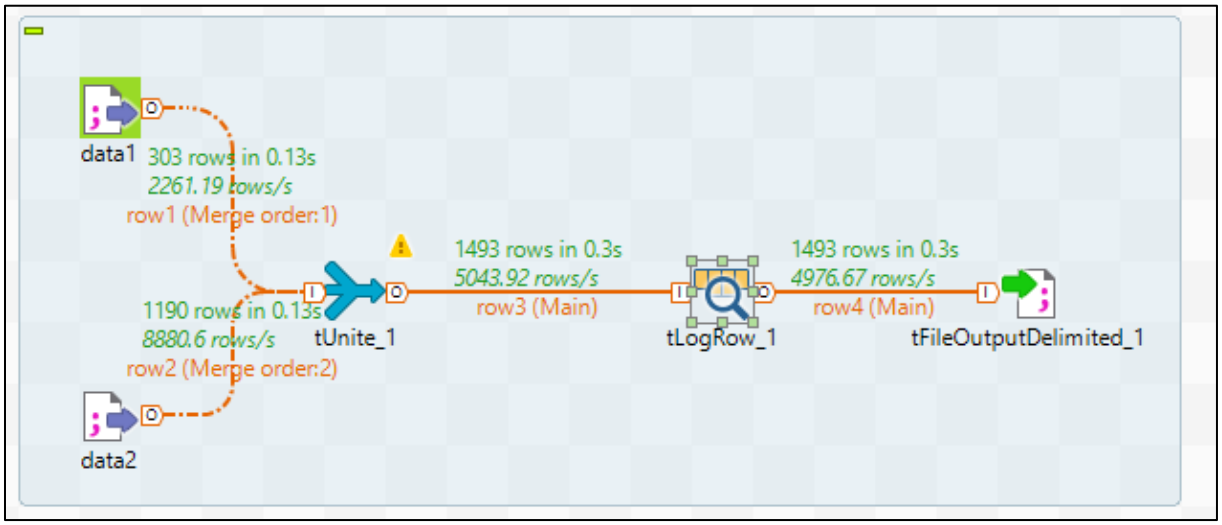


Figure 13 : Sauvegarde de résultat dans un fichier csv.

b. Les données manquantes (missing data)

Dans de nombreuses applications, les données sont affectées par des valeurs manquantes, qui perturbent l'analyse statistique.

L'apprentissage supervisé consiste à approcher une fonction de lien entre des Co-variables, que l'on représente par un vecteur X, et une réponse Y, avec l'objectif de minimiser une fonction de perte entre la valeur prédite et la vraie réponse (Vapnik 1999). Afin de mesurer la capacité de généralisation d'un prédicteur, on distingue généralement un jeu d'apprentissage, et un jeu de validation sur lequel on mesure l'erreur. En apprentissage supervisé, les données manquantes

apparaissent naturellement dans X, autant dans le jeu d'apprentissage que dans le jeu de validation, d'où la nécessité d'une méthode de traitement adaptée à cet objectif particulier.

En pratique, les méthodes les plus souvent utilisées dans la communauté de la machine Learning sont l'imputation de chaque colonne par sa moyenne, et l'ajout d'une variable supplémentaire indiquant la présence ou non de données manquantes dans la colonne en question.

Dans le cas de notre data on se trouve avec données manquantes alors, on a remplacé les données par **la moyenne** de la colonne. [1]

Avant :

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	target	ca	thal
0	63	1	3	145	233	1	0	150	0	2.3	0	1	0.0	1.0
1	37	1	2	130	250	0	1	187	0	3.5	0	1	0.0	2.0
2	41	0	1	130	204	0	0	172	0	1.4	2	1	0.0	2.0
3	56	1	1	120	236	0	1	178	0	0.8	2	1	0.0	2.0
4	57	0	0	120	354	0	1	163	1	0.6	2	1	0.0	2.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1488	45	1	1	110	264	0	0	132	0	1.2	2	1	NaN	NaN
1489	68	1	4	144	193	1	0	141	0	3.4	2	1	NaN	NaN
1490	57	1	4	130	131	0	0	115	1	1.2	2	1	NaN	NaN
1491	57	0	2	130	236	0	2	174	0	0.0	2	1	NaN	NaN
1492	38	1	3	138	175	0	0	173	0	0.0	1	0	NaN	NaN

Figure 14:Exemple de manque de data

Après :

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	target	ca	thal
0	63	1	3	145	233	1	0	150	0	2.3	0	1	0.000000	1.000000
1	37	1	2	130	250	0	1	187	0	3.5	0	1	0.000000	2.000000
2	41	0	1	130	204	0	0	172	0	1.4	2	1	0.000000	2.000000
3	56	1	1	120	236	0	1	178	0	0.8	2	1	0.000000	2.000000
4	57	0	0	120	354	0	1	163	1	0.6	2	1	0.000000	2.000000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
795	75	1	4	170	203	1	1	108	0	0.0	2	1	0.729373	2.313531
796	49	1	1	130	0	0	1	145	0	3.0	2	1	0.729373	2.313531
797	51	1	3	137	339	0	0	127	1	1.7	2	1	0.729373	2.313531
798	60	1	4	142	216	0	0	110	1	2.5	2	1	0.729373	2.313531
799	64	0	4	142	276	0	0	140	1	1.0	2	1	0.729373	2.313531

Figure 15: Résultats de calcul de la moyenne.

-La moyenne des valeurs de la colonne "ca" donne la valeur 0.729 et la colonne "thal" donne la valeur 2.31.

- Les valeurs de "ca" et "thal" sont des nombres décimaux. Pour cela on a remplacé :

→ 0.729 par 1.

→ 3.31 par 2

	AGE	SEX	CP	TRESTBPS	CHOL	FBS	RESTECG	THALACH	EXANG	OLDPEAK	SLOPE	TARGET	CA	THAL
0	63	1	3	145	233	1	0	150	0	2.3	0	1	0.0	1.0
1	37	1	2	130	250	0	1	187	0	3.5	0	1	0.0	2.0
2	41	0	1	130	204	0	0	172	0	1.4	2	1	0.0	2.0
3	56	1	1	120	236	0	1	178	0	0.8	2	1	0.0	2.0
4	57	0	0	120	354	0	1	163	1	0.6	2	1	0.0	2.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
795	75	1	4	170	203	1	1	108	0	0.0	2	1	1.0	2.0
796	49	1	1	130	0	0	1	145	0	3.0	2	1	1.0	2.0
797	51	1	3	137	339	0	0	127	1	1.7	2	1	1.0	2.0
798	60	1	4	142	216	0	0	110	1	2.5	2	1	1.0	2.0
799	64	0	4	142	276	0	0	140	1	1.0	2	1	1.0	2.0

Figure 16:Résultat après la modification.

c. Analyse en composantes principales (ACP)

L'**analyse en composantes principales** (**ACP** ou **PCA** en anglais pour *principal component analysis*), est une méthode de la famille de l'analyse des données et plus généralement de la statistique multivariée, qui consiste à transformer des variables liées entre elles (dites « corrélées » en statistique) en nouvelles variables dé corrélées les unes des autres.

Ces nouvelles variables sont nommées « composantes principales », ou axes principaux. Elle permet au praticien de réduire le nombre de variables et de rendre l'information moins redondante [3].

Alors après une normalisation des données par le calcul des moyennes et des écart-type, aussi la définition des valeurs propre et de matrice de corrélation, on a trouvé 5 composantes principales qu'on peut les gardent pour réduire le nombre des variables.

```
Nombre de composantes principales :  5

Valeurs propres : [1.69195277 1.19519351 0.82556454 0.72920901 0.56143137]

Proportion des variances (pourcentage) :  [33.8163903  23.88785951 16.5002317  14.57441185 11.22110662]

Matrice des covariances :
 [[ 1.00067024  0.26154062 -0.01254679 -0.36494283  0.23863812]
 [ 0.26154062  1.00067024  0.09684336 -0.09231827  0.17888974]
 [-0.01254679  0.09684336  1.00067024  0.22816965  0.06121456]
 [-0.36494283 -0.09231827  0.22816965  1.00067024 -0.20516731]
 [ 0.23863812  0.17888974  0.06121456 -0.20516731  1.00067024]]
```

Figure 17:Informations significatives

- Cercle des corrélations

La figure ci-dessous illustre le cercle des corrélations pour les cinq composantes principales.

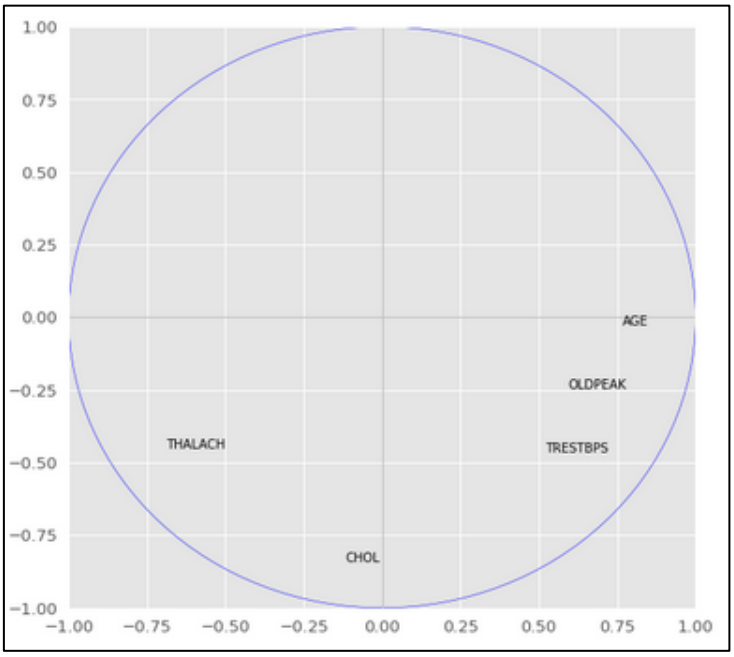


Figure 18:Cercle des corrélations

- Représentation de l'âge des individus

Pour visualiser et vérifier les résultats, on affiche l'âge des individus dans un dataframe et dans un plot à l'emplacement correspondant à l'ACP. L'âge des individus proches auront à chaque fois leurs composantes proches.

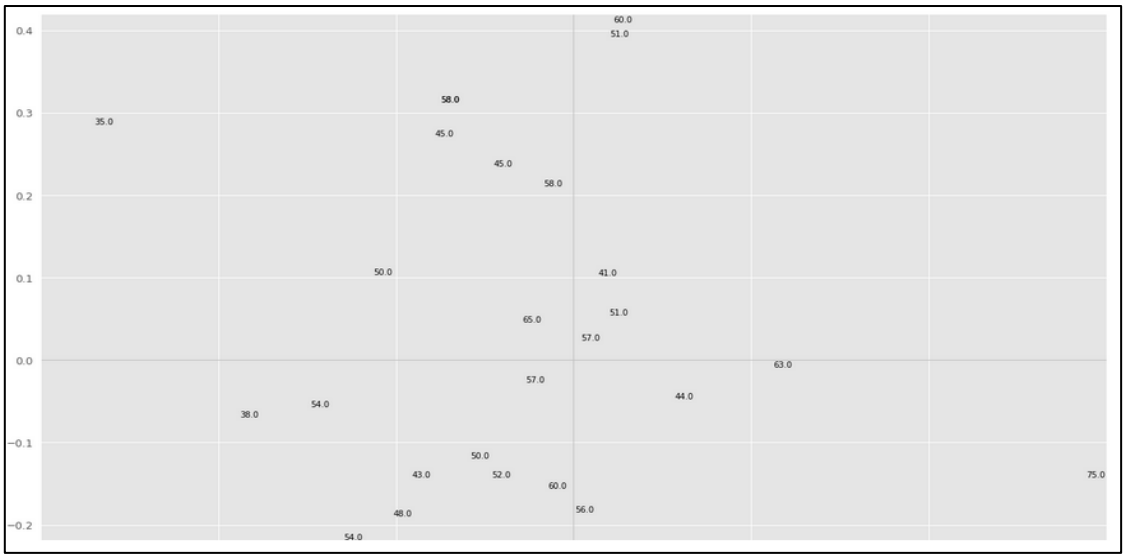


Figure 19:Représentation de l'âge des individus.

Nous avons pu voir à l'aide de cette application comment effectuer une Analyse à Composantes Principales et exploiter ses données. En effet, elle nous a donné assez peu d'informations significatives.

Remarque : on a décidé de ne pas travailler avec les résultats donnés par l'ACP, car il n'a pas amélioré les résultats de training comme prévus.

d. Les points aberrants

En statistique, une donnée aberrante (outlier) est une valeur ou une observation qui est « distante » des autres observations effectuées sur le même phénomène, c'est-à-dire qu'elle contraste grandement avec les valeurs « normalement » mesurées.

Une donnée aberrante peut être due à la variabilité inhérente au phénomène observé ou bien elle peut aussi indiquer une erreur expérimentale. Les dernières sont parfois exclues de la série de données.

Heureusement, on n'a pas des outliers dans notre data.

→ Le calcul des points aberrants :

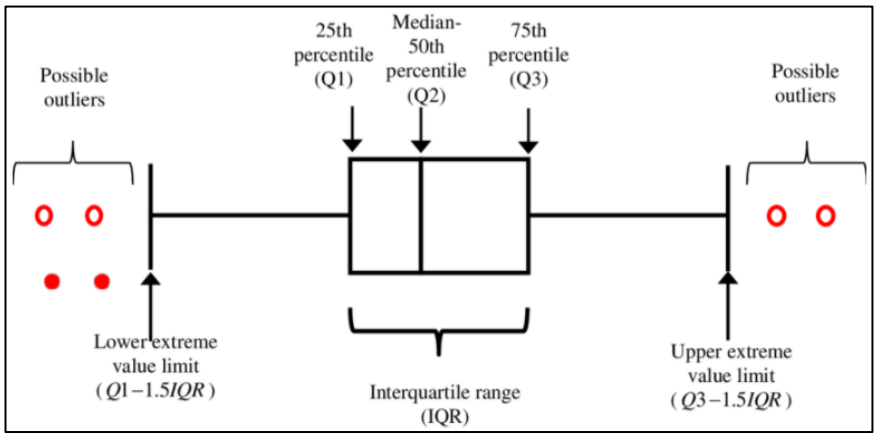


Figure 20:Le calcul des points aberrants

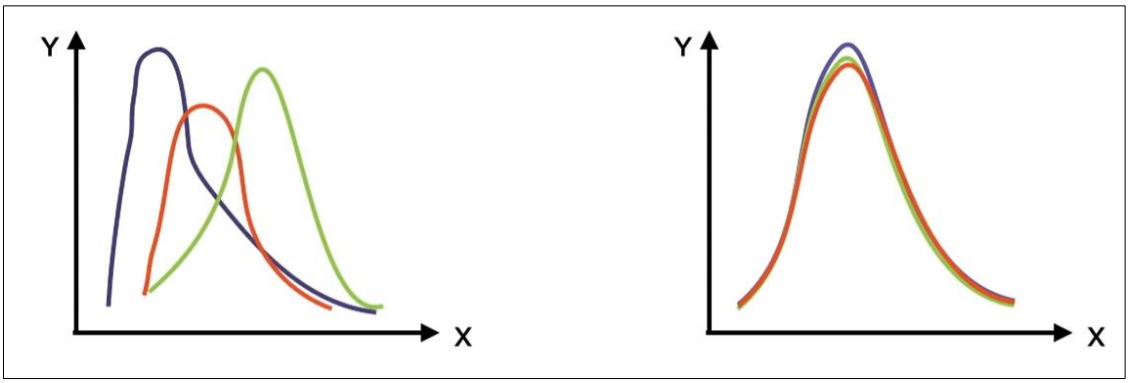
- Q1 = 1. Quartile 25%
- Q2 = 2. Quartile 50% (median)
- Q3 = 3. Quartile 75%
- IQR = Q3 - Q1
- Outliers = (Q1 - 1.5 IQR) U (Q3 + 1.5 IQR)

e. Standardisation des données

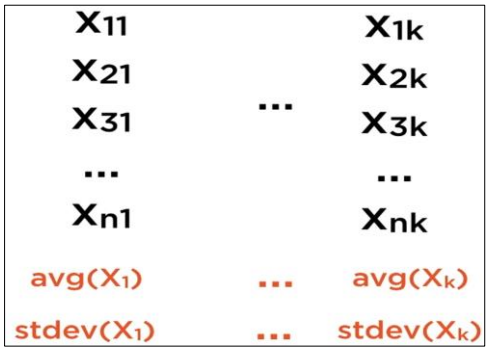
Dans une base de données ou on a beaucoup de valeurs numériques, l'intervalle de chaque colonne sera différent de l'autre, la moyenne et la dispersion seront différentes aussi ce qui cause que la comparaison de colonnes sera difficile.

Les algorithmes du Machine Learning ne fonctionnent pas bien quand les données numériques ont une échelle différente, pour résoudre ce problème il faut standardiser les données.

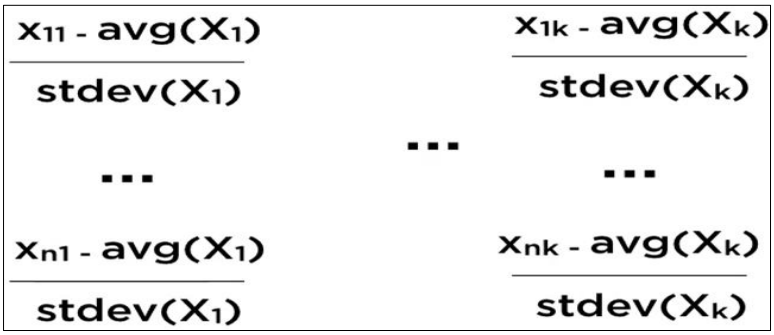
Standardisation c'est le fait de centrer les données d'avoir une moyenne de 0 et une variance de 1, toutes les colonnes seront réparties sur cette valeur centrale de 0.



Pour mieux évaluer le rôle de cette phase on va schématiser un exemple :



On a k nombre de colonnes, pour chaque colonne on a calculé la moyenne et l'écart-type, pour standardiser ces données il faut soustraire la moyenne de la colonne de chaque valeur de cette colonne et diviser tout par l'écart-type de la colonne.



Ce qui fait que chaque colonne a une moyenne de 0 et une variance ou un écart type de 1.

- Résultats de standardisation

	AGE	SEX	CP	TRESTBPS	CHOL	FBS	RESTECG	THALACH	EXANG	OLDPEAK	SLOPE	TARGET	CA	THAL
0	0.983695	1	3	0.712033	0.162316	1	0	0.326165	0	1.228071	0	1	0.0	1.0
1	-1.811902	1	2	-0.112475	0.342076	0	1	1.788047	0	2.316853	0	1	0.0	2.0
2	-1.381810	0	1	-0.112475	-0.144334	0	0	1.195392	0	0.411484	2	1	0.0	2.0
3	0.231034	1	1	-0.662146	0.194038	0	1	1.432454	0	-0.132907	2	1	0.0	2.0
4	0.338557	0	0	-0.662146	1.441785	0	1	0.839800	1	-0.314371	2	1	0.0	2.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1488	-0.951719	1	1	-1.211818	0.490114	0	0	-0.385021	0	0.230020	2	1	1.0	2.0
1489	1.521309	1	4	0.657065	-0.260649	1	0	-0.029428	0	2.226121	2	1	1.0	2.0
1490	0.338557	1	4	-0.112475	-0.916245	0	0	-1.056696	1	0.230020	2	1	1.0	2.0
1491	0.338557	0	2	-0.112475	0.194038	0	2	1.274413	0	-0.858762	2	1	1.0	2.0
1492	-1.704379	1	3	0.327263	-0.450983	0	0	1.234903	0	-0.858762	1	0	1.0	2.0

1493 rows × 14 columns

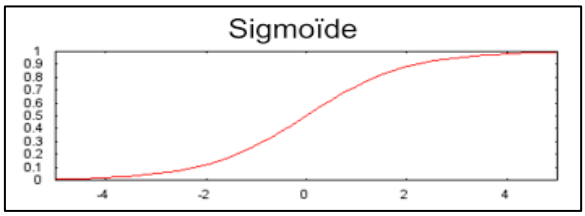
Figure 21:Résultats de standardisation.

2. ALGORITHMES DU MACHINE LEARNING

a. Régression logistique

La régression logistique est un algorithme de classification utilisé pour attribuer des observations à un ensemble discret de classes.
Certains des exemples de problèmes de classification sont les spams par courrier électronique ou non, les transactions en ligne frauduleuses ou non, les tumeurs malignes ou bénignes.
La régression logistique transforme sa sortie en utilisant la fonction logistique sigmoïde pour renvoyer une valeur de probabilité.

- La fonction sigmoïde



Afin de mapper les valeurs prédites aux probabilités, nous utilisons la fonction Sigmoïde. La fonction mappe toute valeur réelle en une autre valeur comprise entre 0 et 1.

En apprentissage automatique, nous utilisons le sigmoïde pour mapper les prédictions aux probabilités.

b. Random Forest

Random Forest Classifier est un algorithme de classification qui réduit la variance des prévisions d'un arbre de décision seul, améliorant ainsi leurs performances. Pour cela, il combine de nombreux arbres de décisions dans une approche de type bagging.

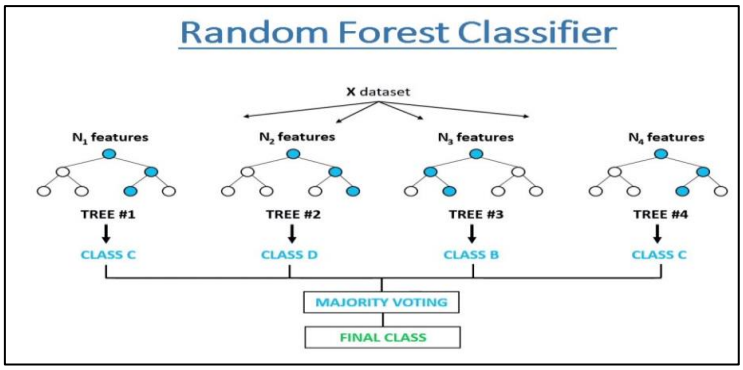


Figure 22: Schéma descriptive de Random Forest

c. Arbre de décision

Un arbre de décision peut être décrit comme un diagramme de flux de données où chaque nœud interne décrit un test sur une variable d'apprentissage, chaque branche représente un résultat du test, et chaque feuille contient la valeur de la variable cible (une étiquette de classe pour les arbres de classification, une valeur numérique pour les arbres de régression).

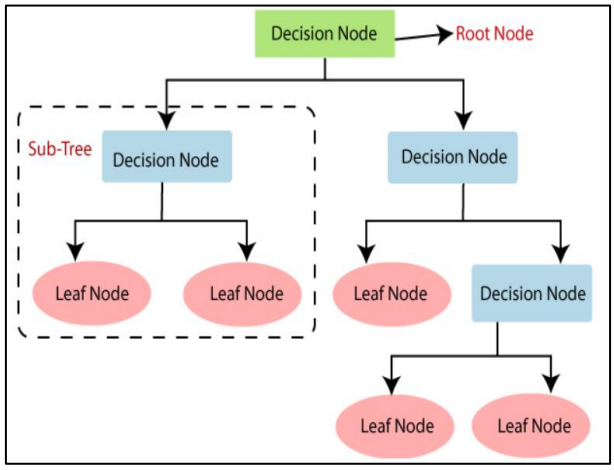
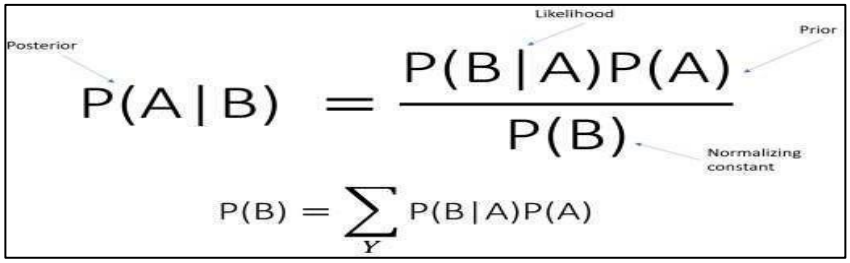


Figure 23: Schéma descriptive d'arbre de décision

d. Naive Baiyes

Les modèles naïfs de Bayes sont appelés algorithmes « naïfs » car ils supposent que les variables prédictives sont indépendantes les unes des autres. En d'autres termes, la présence d'une certaine fonctionnalité dans un jeu de données n'a aucun lien avec la présence d'une autre fonctionnalité. Ils fournissent un moyen facile de construire des modèles précis avec de très bonnes performances compte tenu de leur simplicité. Ils le font en fournissant un moyen de calculer la probabilité « postérieure » qu'un certain événement A se produise, compte tenu de certaines probabilités d'événements « antérieurs ».



e. SVM (Support vector machine)

L'objectif de l'algorithme de la machine à vecteurs de support est de trouver un hyperplan dans un espace à N dimensions (N - le nombre de caractéristiques) qui classe distinctement les points de données. Pour séparer les deux classes de points de données, de nombreux hyperplans possibles peuvent être choisis.

Notre objectif est de trouver un plan qui présente la marge maximale, c'est-à- dire la distance maximale entre les points de données des deux classes. La maximisation de la distance de marge fournit un renforcement permettant de classer les futurs points de données avec plus de confiance.

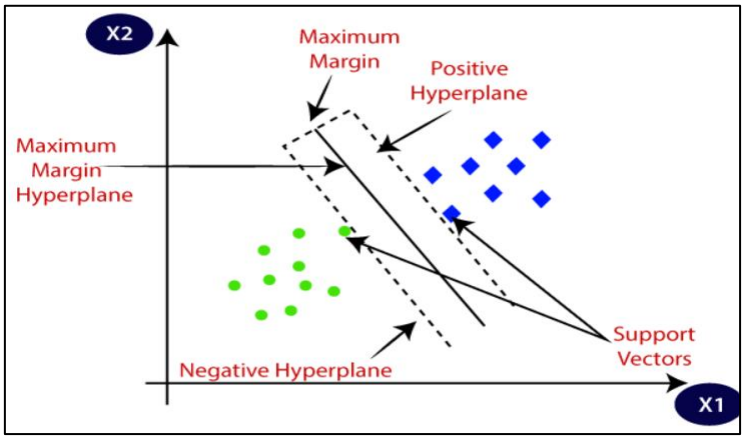


Figure 24: Schéma descriptive de SVM

f. Algorithme KNN (k-nearest neighbors)

L'algorithme des K voisins les plus proches classifie un objet à la majorité des voix des voisins de l'objet, dans l'espace du paramètre d'entrée. L'objet est assigné à la classe qui est la plus commune parmi ses k voisins les plus proches voisins.

C'est un algorithme non paramétrique, paresseux. C'est non paramétrique car il ne fait aucune hypothèse sur la distribution des données (les données ne doivent pas être normalement distribuées). Il est paresseux car il n'apprend vraiment aucun modèle et ne fait pas de généralisation des données (il n'entraîne pas certains paramètres d'une fonction dans laquelle l'entrée X donne la sortie y).

Cet algorithme est simple à mettre en œuvre, robuste aux données d'entraînement bruyantes et efficace si les données d'entraînement sont volumineuses.

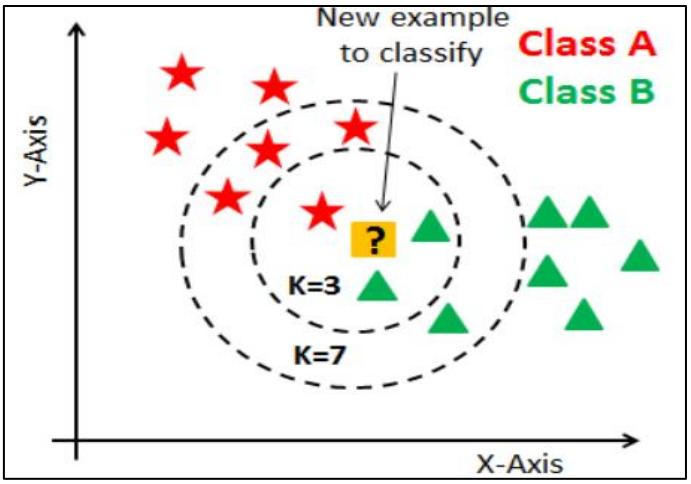


Figure 25: Schéma descriptive de K-NN

III. MODELISATION ET EVALUATION

Dans cette partie on va parcourir les différentes méthodes d'apprentissage automatique, qu'on va l'appliquer à notre data.

Avant de commencer l'implémentation direct des méthodes de machine learning, on a suivi La méthode **CRISP-DM** qui se décompose en 6 étapes commençant par la compréhension du problème métier, la compréhension des données, la préparation des données, la modélisation, et enfin l'évaluation.

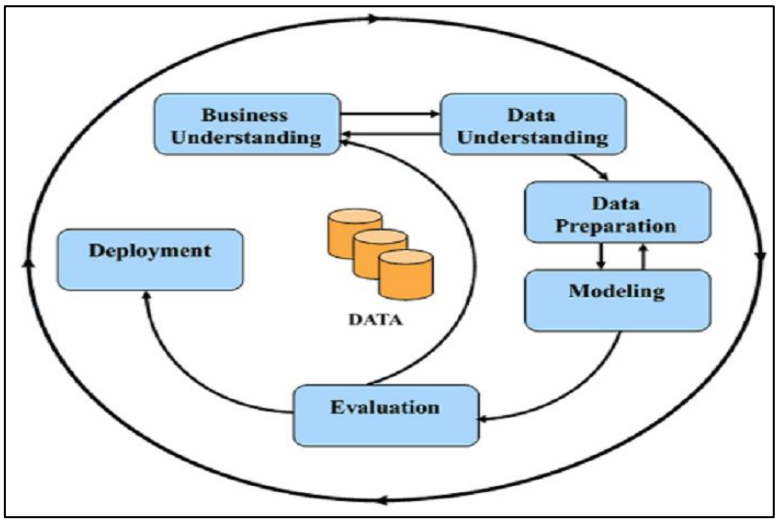


Figure 26: les étapes de CRISP-DM.

1. IMPORTATION DES BIBLIOTHEQUES

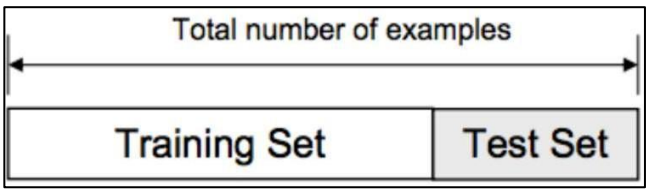
```
from sklearn.model_selection import train_test_split
from sklearn import preprocessing
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.neighbors import KNeighborsClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.svm import SVC
from sklearn import metrics
from sklearn.metrics import confusion_matrix, roc_curve
from sklearn.model_selection import KFold
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import cross_val_predict
import warnings
warnings.filterwarnings('ignore')
from IPython.display import Image
```

2. OUTILS D'EVALUATION

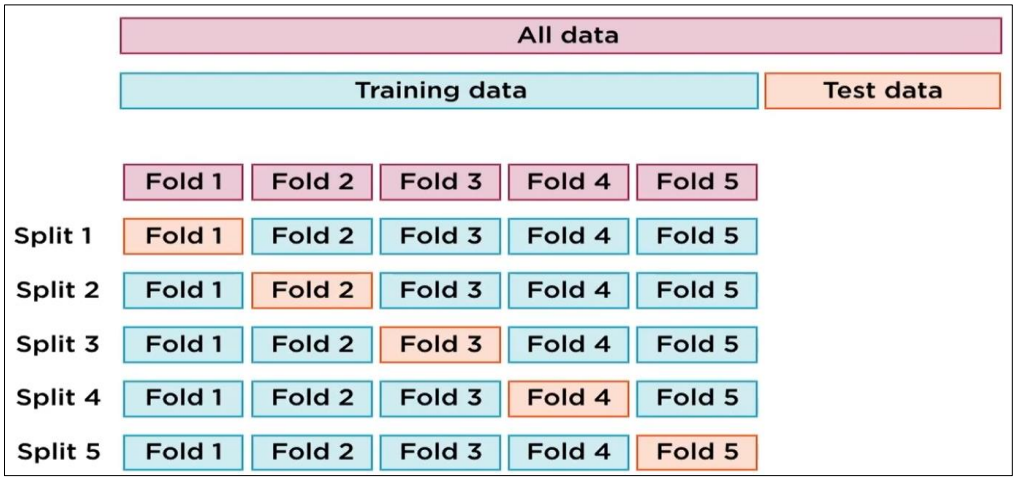
a. Cross Validation (**K-fold**)

```
import pandas as pd
data_train, data_test = train_test_split(df,random_state= 42, stratify=df['TARGET'])
#target
y_train = data_train['TARGET']
y_test = data_test['TARGET']
#features
X_train = data_train.drop('TARGET',axis=1)
X_test = data_test.drop('TARGET',axis=1)

k_fold = KFold(n_splits=5, shuffle=True, random_state=0)
```

Pour évaluer nos modèles supervisés, jusqu'à présent, nous avons divisé notre ensemble de données en un ensemble d'apprentissage et un ensemble de tests à l'aide de la fonction **train_test_split.**



Cette méthode construit un modèle sur l'ensemble d'apprentissage en appelant la méthode d'ajustement, et évalue sur l'ensemble de tests à l'aide de la méthode du score, qui pour la classification calcule la fraction d'échantillons correctement classés, mais il y a une méthode plus facile et mieux c'est **Validation croisée**.



La validation croisée est une méthode statistique d'évaluation des performances de généralisation qui est plus stable et approfondie que l'utilisation d'une répartition en une formation et un ensemble de tests.

Lors de la validation croisée, les données sont plutôt fractionnées à plusieurs reprises et plusieurs modèles sont formés. La version la plus couramment utilisée de la validation croisée est généralement 5 ou 10.

b. Matrice de confusion

En apprentissage automatique supervisé, la matrice de confusion est une matrice qui mesure la qualité d'un système de classification. Chaque ligne correspond à une classe réelle, chaque colonne correspond à une classe estimée.

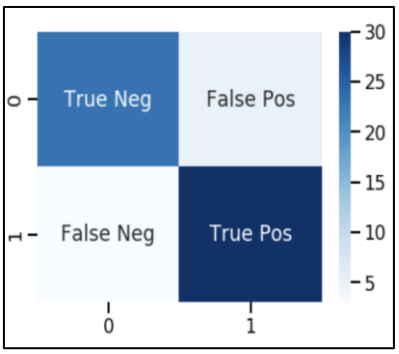
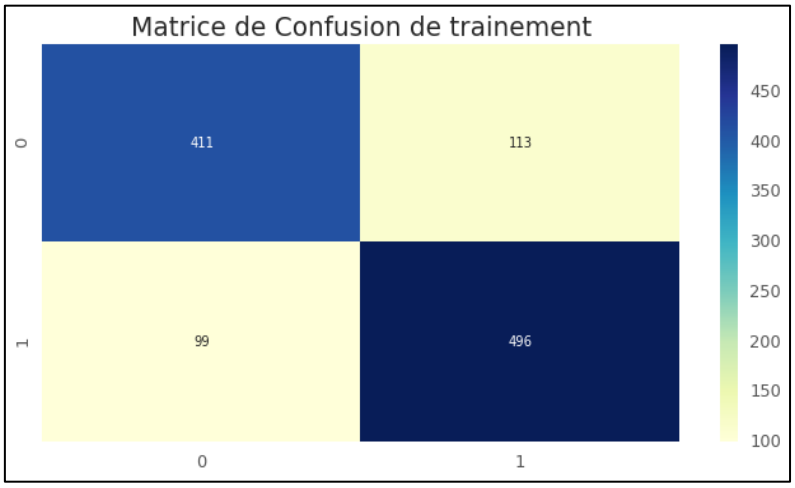


Figure 27:Matrice de confusion

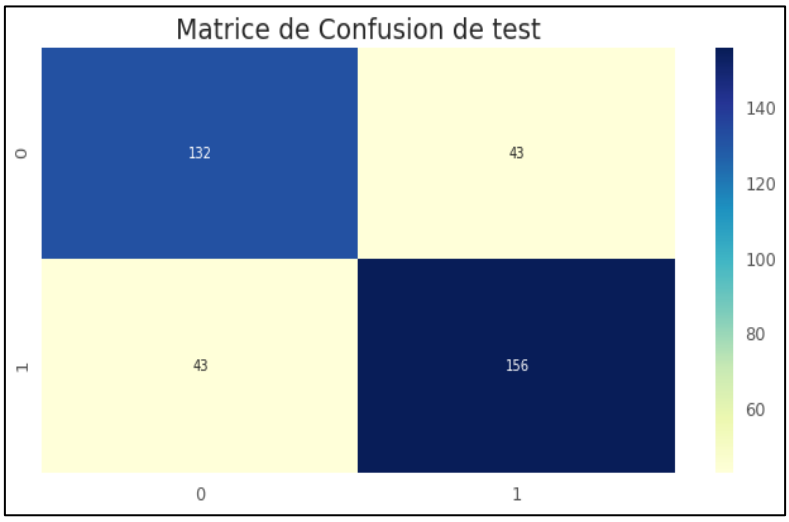
3. RESULTATS

1. Régression logistique

- Résultats de traitement



- Résultats de test

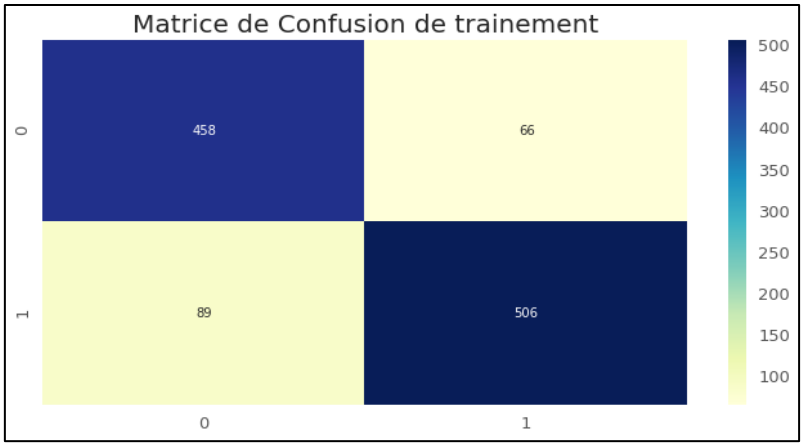


- Score de précision

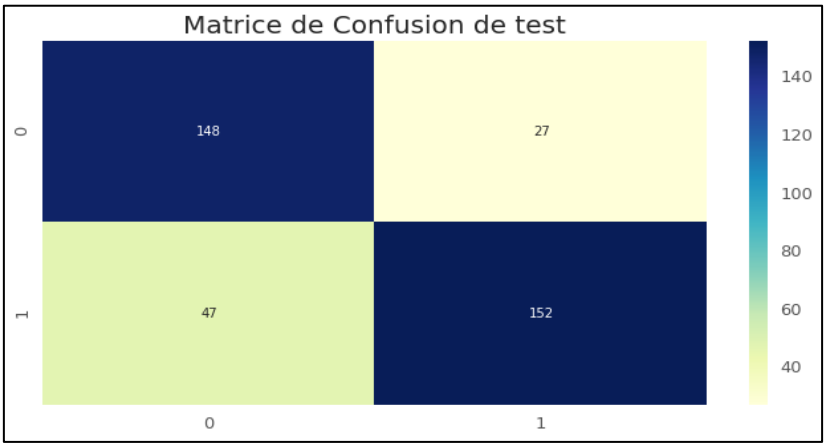
```
round(scoreLR.mean(),5)

0.80976
```

2. Random Forest

- a. Résultat de traitement



- b. Résultat de test



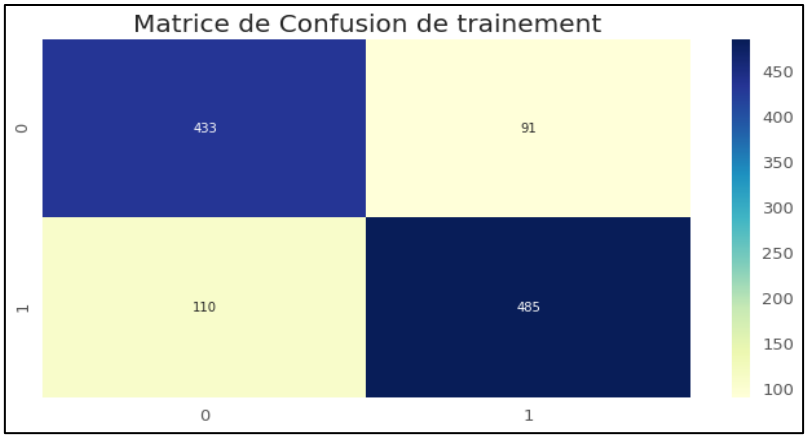
- Score de précision

```
round(scoreRF.mean(),5)
```

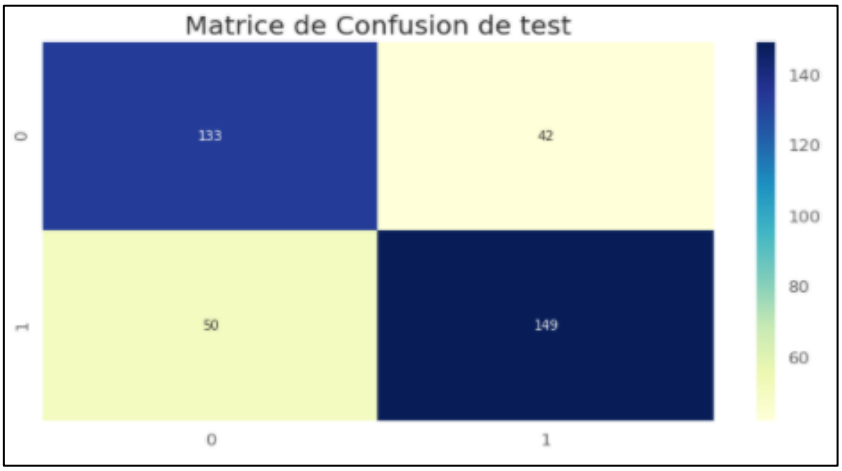
```
0.88077
```

3. Arbre de décision

- Résultats de traitement



- Résultats de test



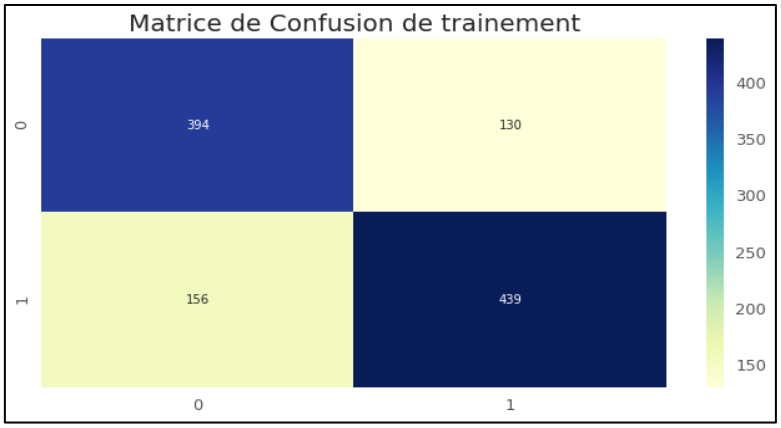
- Score de précision

```
round(scoreDT.mean(),5)
```

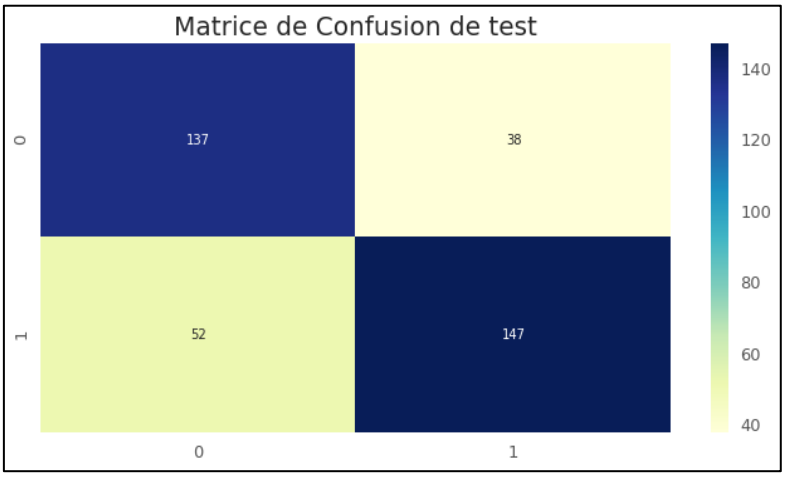
```
0.85465
```

4. Naive Baiyes

- Résultats de traitement



- Résultats de test

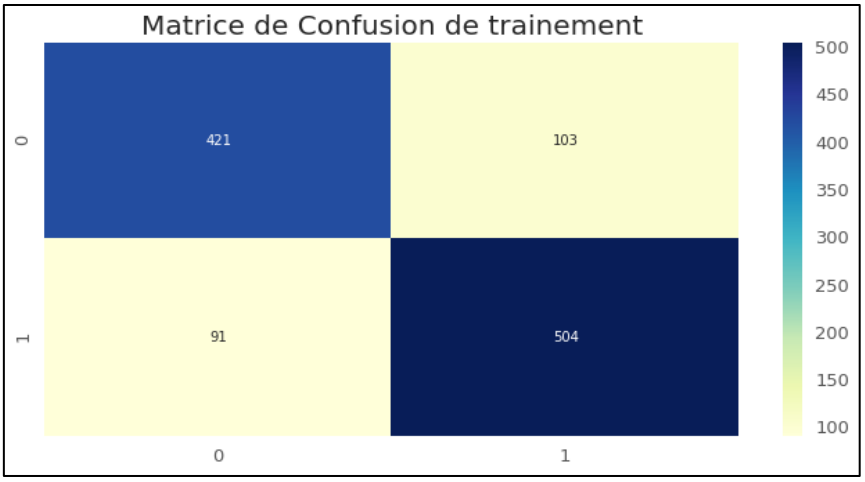


- Score de précision

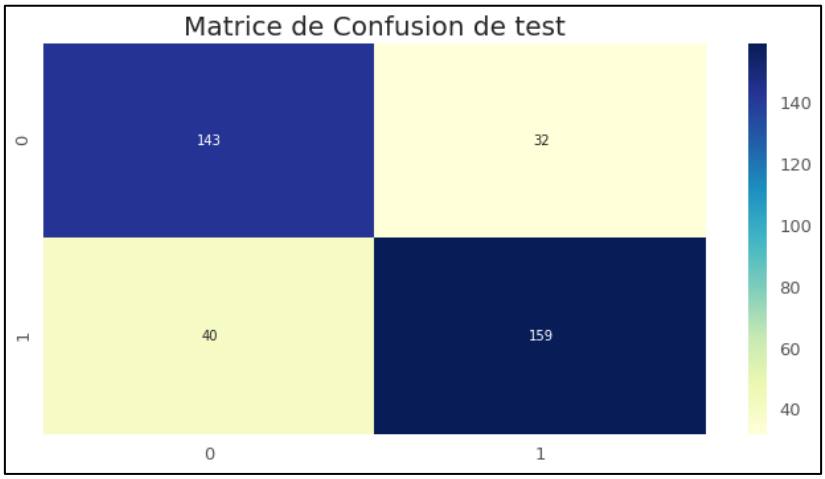
```
round(scoreNB.mean(),5)

0.75016
```

5. SVM (Support vector machine)

- Résultats de trainement



- Résultats de test



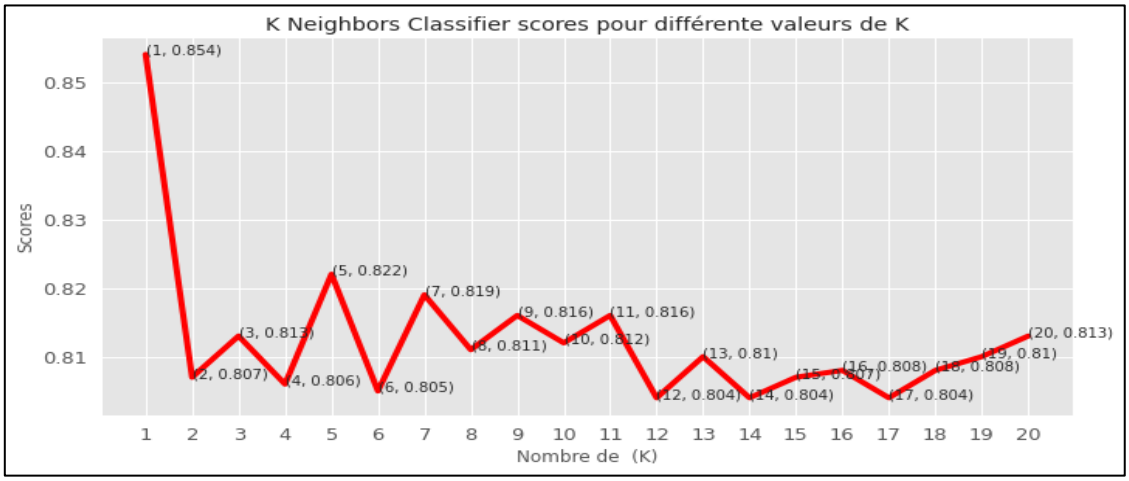
- Score de précision

```
round(scoreSVM.mean(),5)
```

```
0.82718
```

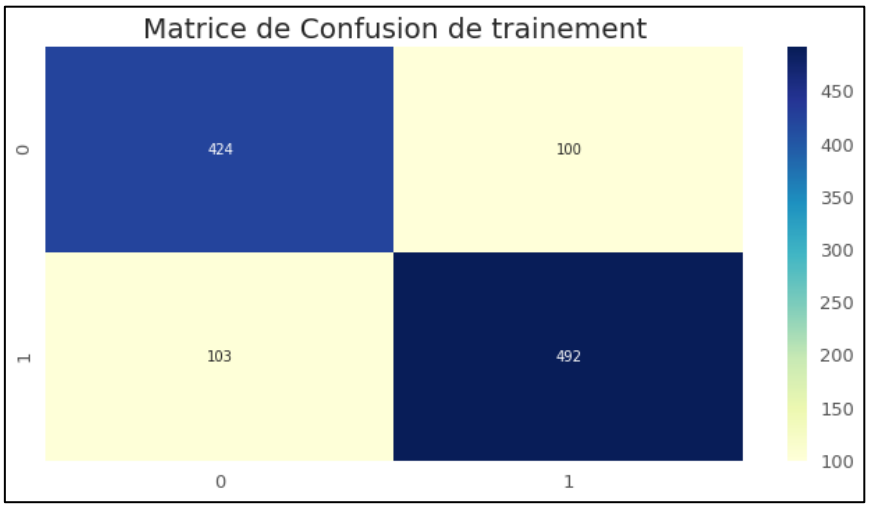
6. Algorithme KNN (k-nearest neighbors)

Pour choisir le meilleur K, on a lancé une boucle pour essayer tous les K possible de 1 à 21 :

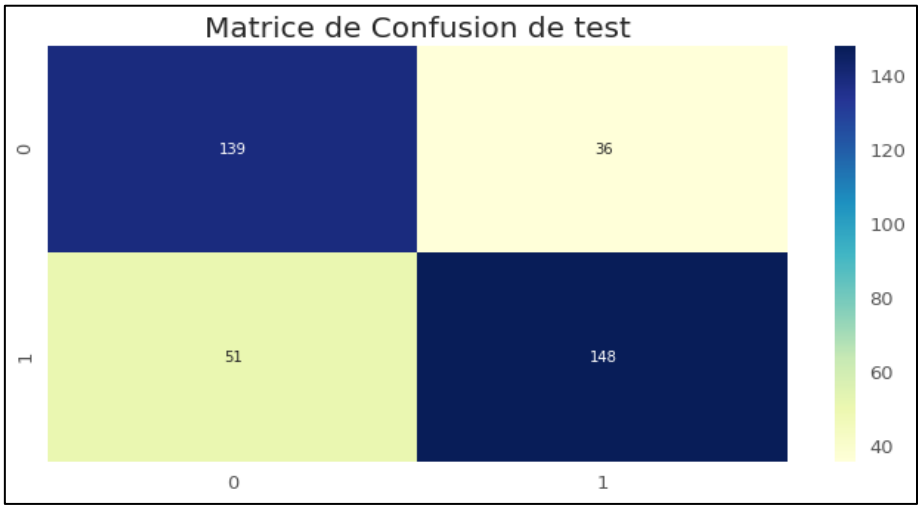


Le meilleur K alors est 1 car il donne le meilleur résultat.

- Résultats de trainement



- Résultats de test

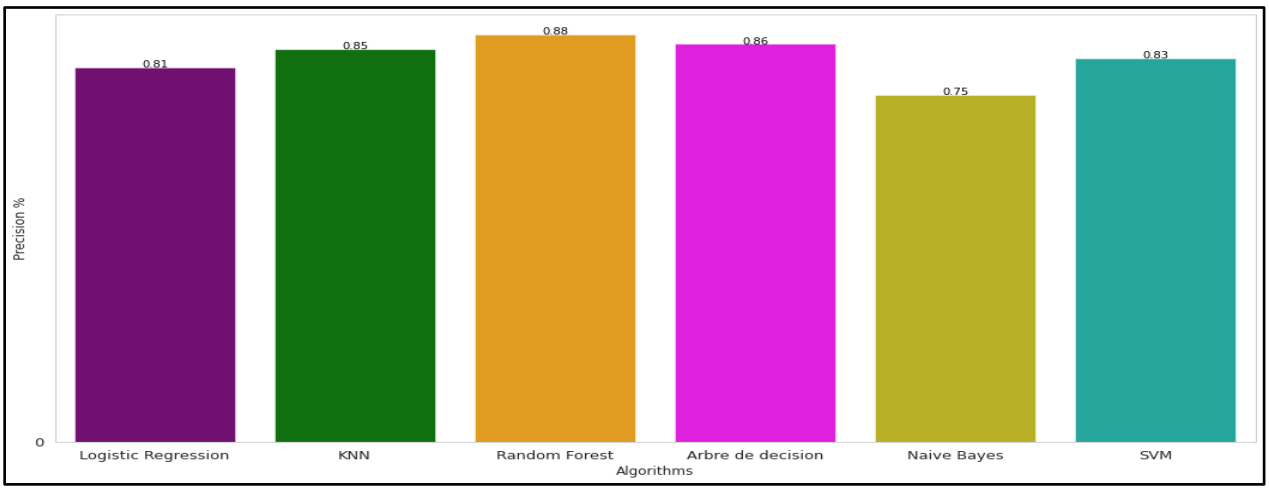


- Score de précision

```
round(scoreKNN.mean(),5)

0.85397
```

7. BENCHMARK algorithmes

Pour conclure, on va benchmarker les résultats de notre data **mining process** sur notre data. Alors comme il est affiché sur notre graphe l'algorithme de Random Forest a donné la meilleure valeur de score de précision avec 88% ;



Ce score montre que le modèle est robuste ; par rapport aux algorithmes KNN régression logistique, arbre de décision et SVM, même si ces derniers ils ont donné des bons résultats et ça reflète le travail de prétraitement et la procédure qu'on l'a effectué à notre data.

CONCLUSION

Le processus de la gestion et le prétraitement du data joue un rôle primordial dans les résultats attendus d'un modèle de prédiction ou de classification de machine Learning. Pour cette raison le data Mining joue un rôle.

Pour conclure, le data Mining consiste à extraire d'une base de données les informations les plus utiles pour détecter les nouvelles tendances adoptées par les consommateurs. Une fois triées, celles-ci peuvent aider le Machine Learning dans sa tâche de formation des systèmes informatisés pour la réalisation de tâches complexes, sans avoir besoin d'intervention humaine. Ces deux technologies ne se chevauchent donc pas, mais sont plutôt complémentaires. En termes simples, le data Mining est une ressource sur laquelle le Machine Learning peut compter pour accomplir ses fonctions. À ce jour, le data Mining n'est toujours pas une technologie évolutive. De nature statique, elle nécessite une intervention humaine pour tirer parti des données qu'elle trie. D'autre part, l'un des points forts du Machine Learning est qu'il permet aux machines d'ajuster leurs algorithmes de manière autonome afin d'atteindre un plus grand degré d'intelligence au fil du temps.

REFERENCES

1- « Apprentissage supervisé avec données manquantes » Nicolas Prost, Julie Josse, Erwan Scornet & Gael Varoquaux 2019.

2- https://www.researchgate.net/publication/339507978_L%27intelligence_artificielle_au_service_des_cardiopathies_congenitales_de_l%27adulte_un_cas_de_figure_exemplaire

3- https://fr.wikipedia.org/wiki/Analyse_en_composantes_principales

4- https://fr.wikipedia.org/wiki/Talend

5- https://www.kaggle.com/ronitf/heart-disease-uci?fbclid=IwAR3_-p1FsCgmwOM5uetYX2bKODesW2l4leKDa4d9c7SpFLkWL2kKxxutsWs

6- https://archive.ics.uci.edu/ml/datasets/Heart+Disease?fbclid=IwAR02TtxM7uVVeTltmW6FiEZYFoO4N3v1H68xx7WygUH7HllVeV91_xjH__w

7- https://www.kaggle.com/maxcobra/projet-acp-pubg

8- https://help.talend.com/r/MW~bv2NtBd_sCQPsRNN2pA/lCqoKBO1L5I3HJrhQ49C8g