Université Mohammed V - Rabat
Master Bio-informatique
et
Modélisation des Systèmes Complexes Appliquée à la Santé

Application des modèles du Machine Learning pour la prédiction et visualisation de Diabetis Melitus (DM) aux pays de GOLF

Réalisé par :
REDOUANE FARDOUZ
SADIKI NOUR-EDDINE

Encadré par :
Dr. Abdellah Mahmoudi

Année universitaire 2019-2020

Table des matières

1 PRESENTATION DU SUJET **2**

2 Définition de la population cible : **3**
2.1 Critères d'inclusion : . . . 3
2.2 Critères d'exclusion : . . . 3

3 Définition des variables : **3**

4 Caractéristiques de la population étudiée : **5**
4.1 Profil socio-démographiques . . . 5
4.2 Profil glycémique . . . 8
4.3 Profil lipidique . . . 8
4.4 Profil clinique . . . 8

5 Nettoyage de données **9**
5.1 Filtrage des données éparses : . . . 9
5.2 Vérification du contenu (valeurs contradictoires et hors normes) : . . . 11
5.2.1 valeurs contradictoires : . . . 11
5.2.2 valeurs hors normes . . . 11
5.3 gestion des valeurs manquants . . . 13
5.4 Regroupements des modalités (modalités multi-valeurs) : . . . 14

6 APPLICATION D'ALGORITHMES DU MACHINE LEARNING : **15**
6.1 régression logistique : . . . 15
6.2 Random Forest : . . . 17
6.3 Arbre de décision : : . . . 18
6.4 Visualisation graphique des algorithmes : . . . 19

1 PRESENTATION DU SUJET

Le diabète est une véritable épidémie du nouveau millénaire. L'OMS estime que le nombre des diabétiques est de 422 millions (Avril 2014) et il devrait devenir l'une des principales causes d'incapacité et de décès dans le monde d'ici les vingt-cinq prochaines années.

Le problème de diabète a toujours été source de recherches aussi bien médicale que biologiste. Cependant, les maths et l'outil informatique apporte une touche très intéressante pour pouvoir mieux comprendre, décrire et analyser cette maladie qui touche de plus en plus de peuple dans le monde et en particulier dans les pays de Golf.

Le Machine Learning et l'intelligence artificielle vont avoir un impact dramatique sur le domaine de la santé ; par conséquent, se familiariser avec les techniques de traitement des données appropriées pour les données de santé numériques et les algorithmes les plus utilisés pour les tâches de classification est une utilisation extrêmement précieuse.

Notre étude a pour objectif, d'utilisation des plusieurs algorithmes de classification qui nous permet de créer des modèles qui tente à comprendre les facteurs de risque de cette maladie et prédire si un patient est diabétique or non .

Ces pratiques se basent sur deux parties : La première consiste à visualiser et nettoyer les données afin d'être consolidé et accessibles et structurées. Après les données sont nettoyées et pris nous avons passé au deuxième travail qui sert à implémenter des algorithme de classification et interpréter les résultats.

2 Définition de la population cible :

La population cible de l'étude est représentée par les patients diabétiques type 2 habitants aux quatre pays de Golf (UAE, Oman, Kuwait, Bahreïn) suivis aux niveaux des centres hospitaliers de ces dernièrs.

2.1 Critères d'inclusion :

- Tout patient non diabétique ou diabétique de type 2
- Tout patient ayant 90% des informations saisies serait inclus dans cette étude pour obtenir des résultats fiables.

2.2 Critères d'exclusion :

- Les patients diabétiques type 1 et aussi les patients qui nous ne savons pas leurs statuts de DM.
- Les patients qui ont un manque de 10% aux niveaux de leurs profils seraient exclu.

3 Définition des variables :

Chez tous les patients, nous avons étudié les variables suivantes :

— Sociodémographiques : Age, sexe, niveau socio-économique, niveau d'instruction et le statut marital.

— Cliniques : à l'anamnèse nous recherchons les facteurs de risque, le type du diabète, l'ancienneté, à l'examen clinique nous avons procédé à la prise du poids, de la taille et le tour de taille ainsi que le calcule de masse corporelle.

— Biologiques : le taux d'hémoglobine glyquée (HbA1C), la glycémie à jeun et les paramètres lipidiques.

— Les mesures hygiéno-diététiques (MHD) et l'activité physiques.

— Hypertension : tension diastolique, tension systolique

	Variable	Explication
2	Country	pays de patient
3	Gender	Sexe de patient
4	Marital_Status	État civil de patient
5	Hypertension	variable binaire décrit le statut de hypertention chez le patient
6	Dyslipidemia	variable binaire décrit le statut de Dyslipidémie chez le patient
7	DM	variable binaire décrit le statut de diabéte chez le patient
8	Year_DM_Diagnosed	Année de diagnostic
9	DM_Duration	la durrée du malade de diabéte chez le patient (diabétique or non diabétique)
10	DM_Type	le type de diabéte
11	DM_Treatment_diet	variable binaire décrit le statut de Régime de traitement DM
12	DM_Treatment_insulin	variable binaire décrit est-ce que le patient prendre l'Insulin
13	DM_Treatment_oralhypoglycemicdrugs	variable binaire décrit est que le patient prendre des médicaments hypoglycémiants oraux
14	Family_History_DM_father	variable binaire décrit est-ce que le pére de patient est diabétique ou non
15	Family_History_DM_mother	variable binaire décrit est-ce que la mére de patient est diabétique ou non
16	Family_History_DM_siblings	variable binaire décrit est-ce que l'un de ses Fratries est diabétique
17	BMI	Indice de masse corporelle (IMC)

Data | Data_codifier | EXPLICATION DES VARIABLES

18	Admission_Blood_Glucose_Value_SI_Units	le taux de glycémie
19	Fasting_Blood_Glucose_Value_SI_Units	le taux de glycémie à jeun
20	HbA1C_Admission_Value	L'hémoglobine glyquée
21	Cholesterol_Value_SI_Units	le taux de cholestérol total
22	Triglycerides_Value_SI_Units	taux de triglycérides
23	LDL_Value_SI_Units	un taux de LDL-cholestérol "mauvais cholestérol"
24	HDL_Value_SI_Units	un taux de HDL-cholestérol "bon cholestérol"
25	Creatinine_Clearance	le taux de clairance urinaire de créatinine
26	Heart_Rate	rythme cardiaque
27	Systolic_BP	la valeur du tention systolique
28	Diastolic_BP	la valeur du tention diastolique
29		
30		

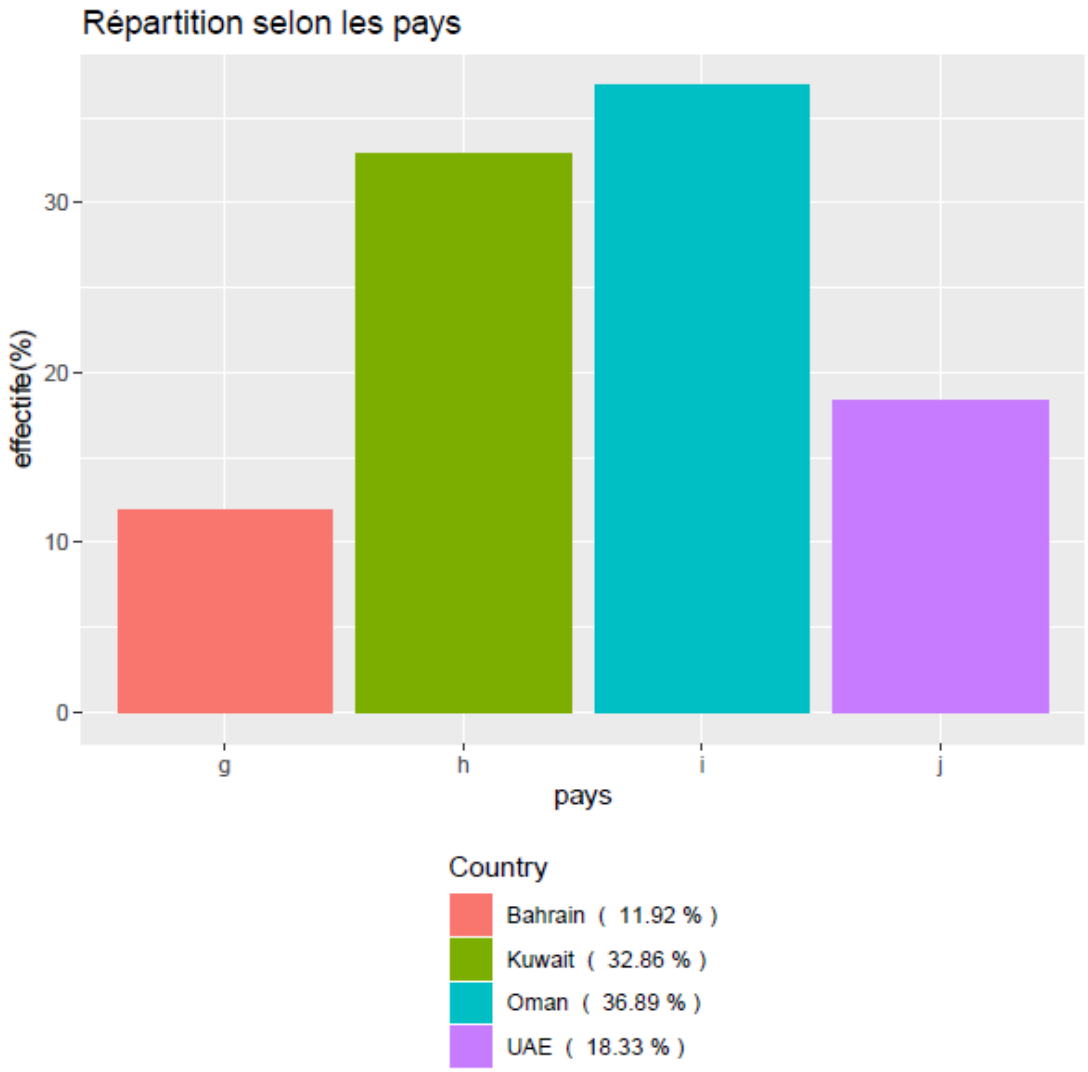
Data | Data_codifier | EXPLICATION DES VARIABLES

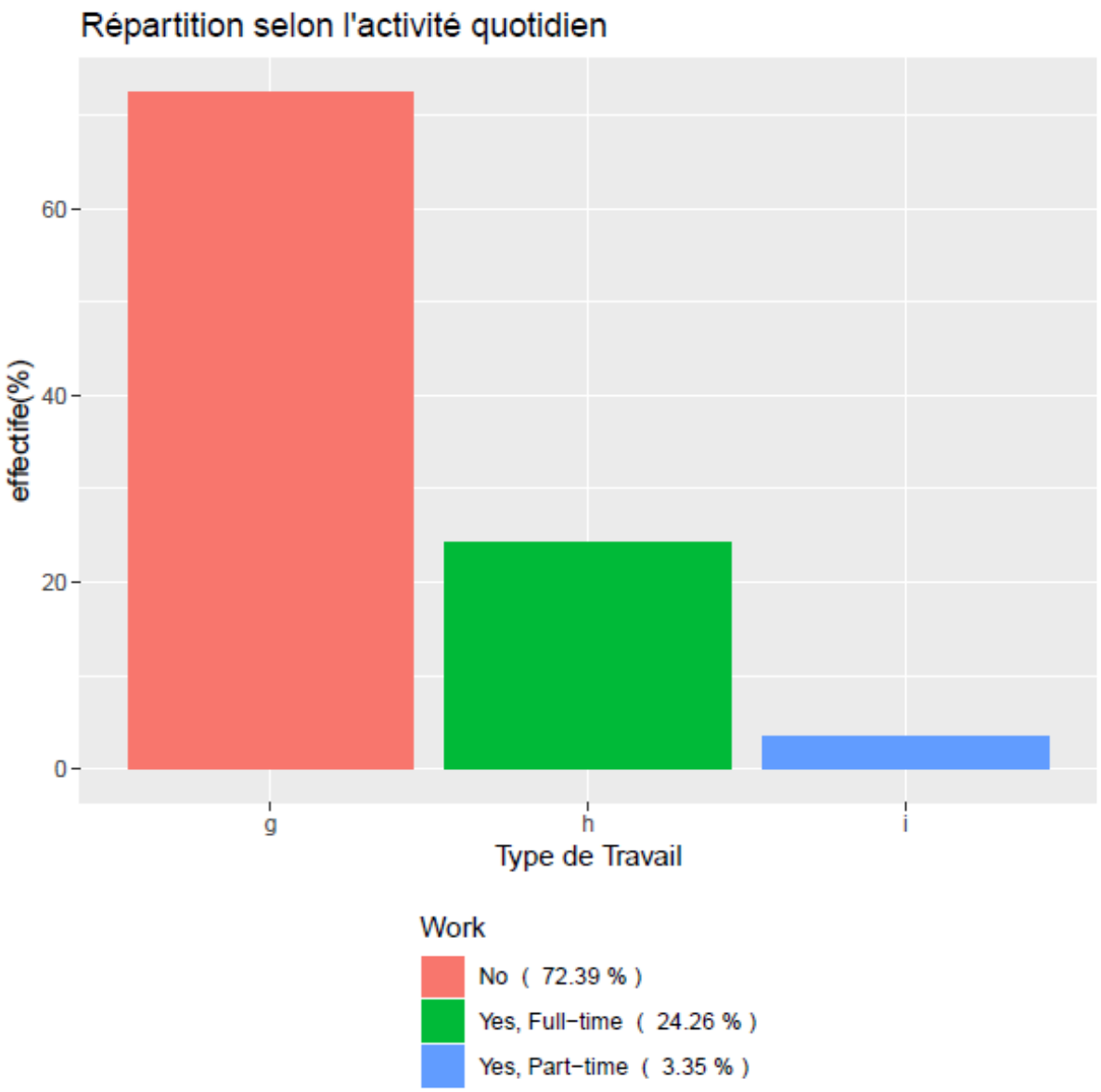
4 Caractéristiques de la population étudiée :

4.1 Profil socio-démographiques

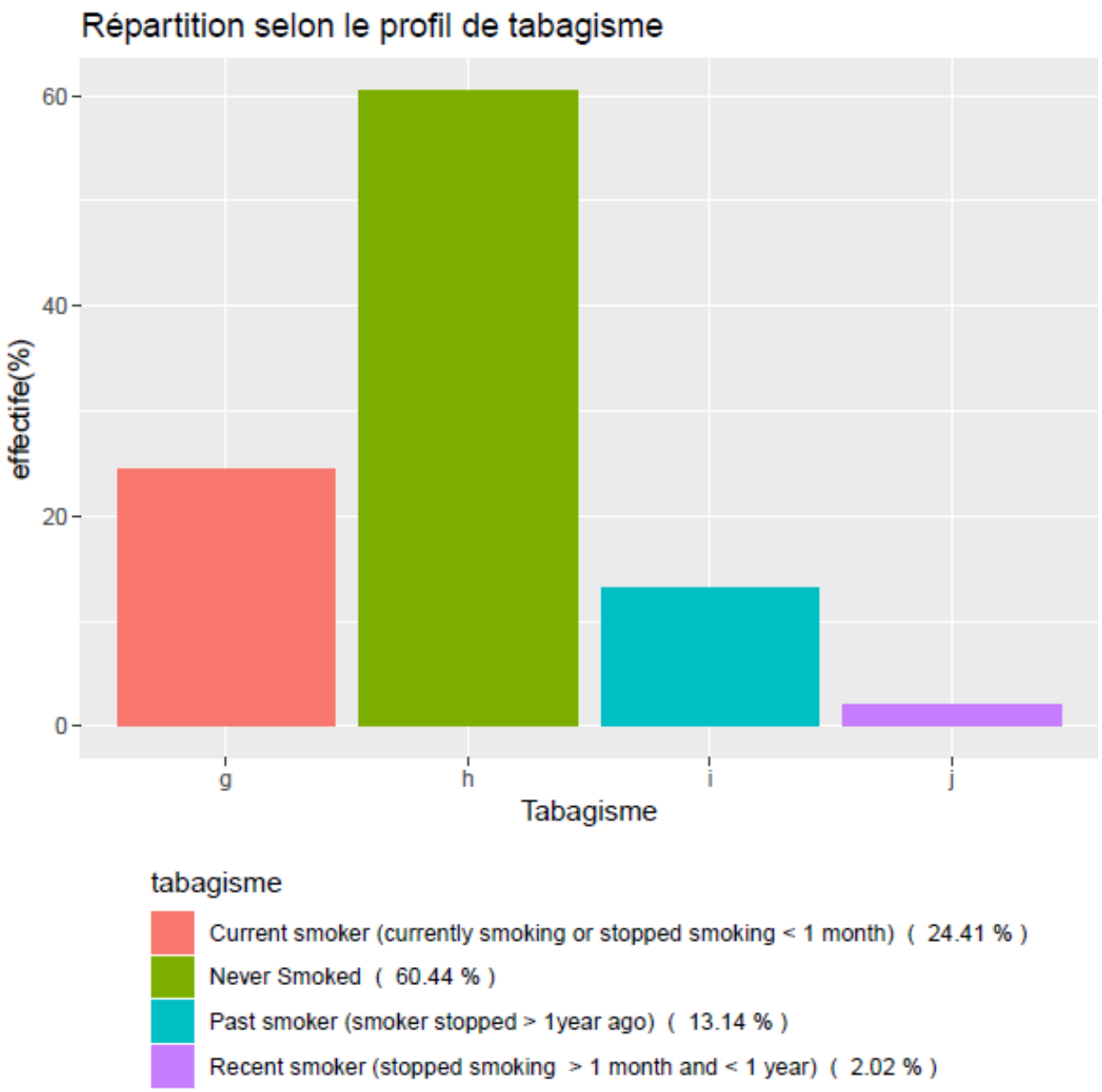
La population étudiée est de 3372 patients distribué géographiquement par (UAE, Oman, Kuwait, Bahreïn). La population est constitué de plus de homme (66,4%) que de femme (33,6%) avec un sexe ratio $H/F = 1,98$. Les extrêmes d'âge des patients oscillent de 18 à 90 ans avec une moyenne d'âge de 60,28 ans ±12,41.

	Observation	Moyenne	Ecart type	Médiane	Min-Max
Age	3372	60.28	12.41	60	18-99





Concernant le profil de travail des individus qui nous donne une information d'activité, 72,39% des patient ne travaillent pas, en revanche plus de 27,61% exercent une activité soit travailler à plein temps (24,26%) ou mi-temps (3,35%).
Cette description nous informe que la majorité des individus probablement n'exercent aucune activité.



le profil de tabac nous avons remarqué que 60,44% ne consomment pas le tabac, d'autre part 39,57% sont tabagiques.

4.2 Profil glycémique

	Observation	Moyenne	Ecart type	Médiane	Min-Max
Hémoglobine glyqué	3372	7.52	2.05	7.01	3.2-20
Glycémie à jeun	3372	7.97	3.26	7	0.81-31.8
Glycémie	3372	10.76	5.48	9	0.92-33

On constate que 51,39% des patients ont une hémoglobine glyquée insuffisante, alors que seulement 48,61% de ces patients ont un diabète équilibré. Concernant les valeurs de la glycémie à jeun nous retrouvons 50,65% des patients ont des valeurs glycémiques insuffisantes et seulement 49,35% ont des valeurs glycémiques suffisantes

4.3 Profil lipidique

	Observation	Moyenne	Ecart type	Médiane	Min-Max
Triglyceride	3372	1.67	1.16	1.4	0.01-17.6
LDL	3372	3	1.1	2.9	0.4-9.94
HDL	3372	1.01	0.31	0.98	0.3-4.29
cholestérol total	3372	4.74	1.29	4.6	1.66-11.8

Concernant le profil lipidique, 39,5% des patients avaient un taux du cholestérol total dépassant 2 g/l ($\simeq$ 5 mmol/l) qui signifie que ces patients ont plus d'éventualité d'être diabétique .

4.4 Profil clinique

	Observation	Moyenne	Ecart type	Médiane	Min-Max
IMC	3372	28.89	5.71	27.99	15.78-61.33
La tension systolique	3372	141.32	27.4	140	42-253
La tension diastolique	3372	80.53	16.16	80	29-168

Pour l'indice de masse corporelle (IMC) , plus de 74,56 % sont en surpoids ou obèses, 57,33 % des patients avaient une tension systolique qui correspond à un contrôle insuffisant et 74,56 % avaient une tension diastolique dépassant les valeurs limites.

5 Nettoyage de données

Les données présentes dans les bases de données peuvent avoir plusieurs types d'erreurs comme des erreurs de frappe, des informations manquantes, des imprécisions etc. La partie impropre de la donnée traitée peut être remplacée, modifiée ou supprimée. Le processus de nettoyage identifie les données erronées et les corrige automatiquement avec un programme informatique ou les propose à un humain pour qu'il effectue les modifications.
Le processus de nettoyage est basé sur :

5.1 Filtrage des données éparses :

Les données éparses caractérisent des observations qui possèdent de nombreuses valeurs manquantes et qui n'apportent que peu d'information dans l'analyse et la modélisation. Après la première visualisation des données nous avons remarqué des valeurs manquantes aux profils des patients. Chose qu'il faut traiter, voici un aperçu ci-dessous qui résume l'état initial des données :

col	index	mod count	NAs	NAp
Country	1	4	0	0
Gender	2	2	0	0
Age	3	-1	0	0
Marital_Status	4	5	0	0
Work	5	3	0	0
Cardiac_Arrest_Admission	6	3	0	0
Non_Cardiac_Condition	7	2	0	0
Hypertension	8	2	0	0
Dyslipidemia	9	2	0	0
DM	10	2	0	0
Year_DM_Diagnosed	11	-1	2051	50.89
DM_Duration	12	-1	1970	48.88
DM_Type	13	1	1867	46.33
DM_Treatment	14	9	1868	46.35
Family_History_DM	15	8	1872	46.45
Smoking_History	16	4	1	0.02
Waist	17	-1	87	2.16
BMI	18	-1	55	1.36
Admission_Blood_Glucose_Value_SI_Units	19	-1	242	6
Fasting_Blood_Glucose_Value_SI_Units	20	-1	1276	31.66
HbA1C_Admission_Value	21	-1	1484	36.82
Lipid_24_Collected	22	2	2	0.05
Cholesterol_Value_SI_Units	23	-1	329	8.16
Triglycerides_Value_SI_Units	24	-1	369	9.16
LDL_Value_SI_Units	25	-1	830	20.6
HDL_Value_SI_Units	26	-1	802	19.9
Stress	27	2	2	0.05
Creatinine_Clearance	28	-1	64	1.59
Education	29	6	0	0
Sleep_Apnea	30	2	0	0
Heart_Rate	31	-1	1	0.02
Systolic_BP	32	-1	1	0.02
Diastolic_BP	33	-1	1	0.02

Pour résoudre ce problème on opte une condition d'inclusion que tout patient ayant 90% des données aux leurs profils serait choisi aux effectifs finals de cette étude afin d'avoir des résultats fiables. La figure suivante montre l'état des données après l'application de cette démarche :

col	index	mod count	NAs	NAp
Country	1	4	0	0
Gender	2	2	0	0
Age	3	-1	0	0
Marital_Status	4	5	0	0
Work	5	3	0	0
Cardiac_Arrest_Admission	6	3	0	0
Non_Cardiac_Condition	7	2	0	0
Hypertension	8	2	0	0
Dyslipidemia	9	2	0	0
DM	10	2	0	0
Year_DM_Diagnosed	11	44	154	4.57
DM_Duration	12	42	85	2.52
DM_Type	13	2	0	0
DM_Treatment	14	10	0	0
Family_History_DM	15	8	1464	43.42
Smoking_History	16	4	0	0
Waist	17	-1	34	1.01
BMI	18	-1	12	0.36
Admission_Blood_Glucose_Value_SI_Units	19	-1	116	3.44
Fasting_Blood_Glucose_Value_SI_Units	20	-1	862	25.56
HbA1C_Admission_Value	21	-1	973	28.86
Lipid_24_Collected	22	2	0	0
Cholesterol_Value_SI_Units	23	-1	5	0.15
Triglycerides_Value_SI_Units	24	-1	13	0.39
LDL_Value_SI_Units	25	-1	276	8.19
HDL_Value_SI_Units	26	-1	246	7.3
Stress	27	2	0	0
Creatinine_Clearance	28	-1	21	0.62
Education	29	6	0	0
Sleep_Apnea	30	2	0	0
Heart_Rate	31	-1	0	0
Systolic_BP	32	-1	0	0
Diastolic_BP	33	-1	0	0

5.2 Vérification du contenu (valeurs contradictoires et hors normes) :

5.2.1 valeurs contradictoires :

Il existe ainsi des valeurs contradictoires prises par les observations, pour cela nous avons retiré ces observations de l'analyse, afin d'avoir des résultats fiables. La figure suivante montre les observations dont ces valeurs sont contradictoires :

DM	Year_DM_Diagnosed	DM_Type	DM_Treatment
No	2008.00	Type 2	Diet
No	1992.00	Type 2	
No	2009.00	Type 2	Diet, Oral Hypoglycemic drugs
No	1997.00	Type 2	Insulin
No	2007.00	Type 2	Diet, Insulin
No	2002.00	Type 2	Diet, Oral Hypoglycemic drugs
No	1992.00	Type 2	Diet, Oral Hypoglycemic drugs, Insulin
No	2009.00	Type 2	Diet
No	2006.00	Type 2	Diet, Oral Hypoglycemic drugs
No	2012.00	Type 2	Oral Hypoglycemic drugs
No	2007.00	Type 2	Oral Hypoglycemic drugs, Insulin
No	2002.00	Type 2	Oral Hypoglycemic drugs
No	2002.00	Type 2	Oral Hypoglycemic drugs
No	2007.00	Type 2	Oral Hypoglycemic drugs
No	2000.00	Type 2	Oral Hypoglycemic drugs
No	1992.00	Type 2	Diet, Oral Hypoglycemic drugs

5.2.2 valeurs hors normes

Après d'extraction de l'échantillon final des patients inclut dans cette étude, on passe à la vérification de ces valeurs prises par chaque variable. On observe qu'il existe des valeurs ne respecte pas les normes adaptées par chaque variable. La figure ci-dessous montre quelque valeurs irrespectables prise par les variables :

```
      Age             Waist              BMI
 Min.   :18.00   Min.   : 25.00   Min.   : 13.34
 1st Qu.:52.00   1st Qu.: 89.00   1st Qu.: 24.97
 Median :60.00   Median : 98.00   Median : 27.97
 Mean   :60.28   Mean   : 98.54   Mean   : 29.16
 3rd Qu.:69.00   3rd Qu.:110.00   3rd Qu.: 31.63
 Max.   :99.00   Max.   :194.00   Max.   :342.78
                 NA's   :34       NA's   :12
 Admission_Blood_Glucose_Value_SI_Units Fasting_Blood_Glucose_Value_SI_Units
 Min.   :  0.29                         Min.   :  0.330
 1st Qu.:  6.40                         1st Qu.:  5.500
 Median :  8.90                         Median :  6.900
 Mean   : 11.29                         Mean   :  8.279
 3rd Qu.: 13.80                         3rd Qu.:  9.500
 Max.   :544.00                         Max.   :235.000
 NA's   :116                            NA's   :862
 HbA1C_Admission_Value Cholesterol_Value_SI_Units Triglycerides_Value_SI_Units
 Min.   : 0.000        Min.   :  0.070            Min.   : 0.010
 1st Qu.: 5.900        1st Qu.:  3.800            1st Qu.: 0.980
 Median : 7.000        Median :  4.600            Median : 1.400
 Mean   : 7.595        Mean   :  4.817            Mean   : 1.669
 3rd Qu.: 9.000        3rd Qu.:  5.515            3rd Qu.: 2.030
 Max.   :88.400        Max.   :308.000            Max.   :17.600
 NA's   :973           NA's   :5                  NA's   :13
 LDL_Value_SI_Units HDL_Value_SI_Units Creatinine_Clearance   Heart_Rate
 Min.   : 0.015     Min.   : 0.000     Min.   :   0.29      Min.   :  0.00
 1st Qu.: 2.180     1st Qu.: 0.800     1st Qu.:  57.39      1st Qu.: 71.00
 Median : 2.900     Median : 0.960     Median :  82.84      Median : 82.00
 Mean   : 3.009     Mean   : 1.103     Mean   :  88.57      Mean   : 84.83
 3rd Qu.: 3.700     3rd Qu.: 1.160     3rd Qu.: 111.34      3rd Qu.: 95.00
 Max.   :18.000     Max.   :91.000     Max.   :1164.70      Max.   :200.00
 NA's   :276        NA's   :246        NA's   :21
  Systolic_BP     Diastolic_BP
 Min.   :  0.0   Min.   :  0.00
 1st Qu.:122.8   1st Qu.: 70.00
 Median :140.0   Median : 80.00
 Mean   :141.2   Mean   : 80.46
 3rd Qu.:158.0   3rd Qu.: 90.00
 Max.   :253.0   Max.   :168.00
```

Pour cela nous avons cité les normes que chaque variable doit être varié entre le min et le max d'aprés notre recherche

Variables	MIN	MAX
Waist	24.00	160.00
BMI	15.00	65.00
Admission_Blood_Glucose_Value_SI_Units	0.71	33.00
Fasting_Blood_Glucose_Value_SI_Units	0.71	33.00
HbA1C_Admission_Value	1.50	20.00
Cholesterol_Value_SI_Units	1.29	12.93
Triglycerides_Value_SI_Units	0.00	56.45
LDL_Value_SI_Units	0.30	10.00
HDL_Value_SI_Units	0.30	5.00
Creatinine_Clearance	5.00	150.00
Systolic_BP	40.00	255.00
Diastolic_BP	20.00	170.00
Heart_Rate	30.00	200.00

5.3 gestion des valeurs manquants

Les valeurs manquantes posent problème pour de nombreux algorithmes de data mining. il est donc nécessaire de traiter ces valeurs manquantes afin d'éviter de produire des modèles peu performants ou incomplets, tout en restant vigilant quant à la manière retenue afin d'éviter d'introduire un biais systématique. Pour cela nous avons interpolé les valeurs manquantes par l'algorithme **K-NN (nearest-neighbor)** qui est très adaptable avec nos données afin définir l'origine des données manquants. L'algorithme K-NN (nearest-neighbor) : K-NN est une méthode d'apprentissage supervisé. Il peut être utilisé aussi bien pour la régression que pour la classification. Son fonctionnement peut être assimilé à l'analogie suivante **"dis-moi qui sont tes voisins, je te dirais qui tu es"**

Principe de fonctionnement de l'algorithme de K-NN :

Pour effectuer une prédiction, l'algorithme K-NN va se baser sur le jeu de données en entier. En effet, pour une observation, qui ne fait pas parti du jeu de données, qu'on souhaite prédire, l'algorithme va chercher les K instances du jeu de données les plus proches de notre observation. Ensuite pour ces voisins, l'algorithme se basera sur leurs variables de sortie (output variable) pour calculer la valeur de la variable de l'observation qu'on souhaite prédire. Par ailleurs :

- Si K-NN est utilisé pour la régression, c'est la moyenne (ou la médiane) des variables des plus proches observations qui servira pour la prédiction.
- Si K-NN est utilisé pour la classification, c'est le mode des variables des plus proches observations qui servira pour la prédiction.

5.4 Regroupements des modalités (modalités multi-valeurs) :

Nous avons regroupé les variables qui ont pris des modalités multi-valeurs en variables indicatrices dans le but d'améliorer la qualité de représentation et les coder après pour être adaptable pour apliquer les modeles de machine learning .

les modalité avant répartition

DM_Treatment	Family_History_DM
Oral Hypoglycemic drugs, Insulin	None
Diet, Oral Hypoglycemic drugs, Insulin	None
Diet, Oral Hypoglycemic drugs	Father, Siblings
Insulin	Mother, Siblings

les modalité aprés répartition

DM_Treatment	Family_History_DM
Oral Hypoglycemic drugs, Insulin	None
Diet, Oral Hypoglycemic drugs, Insulin	None
Diet, Oral Hypoglycemic drugs	Father, Siblings
Insulin	Mother, Siblings

6 APPLICATION D'ALGORITHMES DU MACHINE LEARNING :

Après le diagnostic de la qualité des données, nous avons codifié ces derniers pour les préparer afin d'appliques les modeles du machine learning , chaque modalité de tel variable prend une valeur numérique unique et bien spécifique.

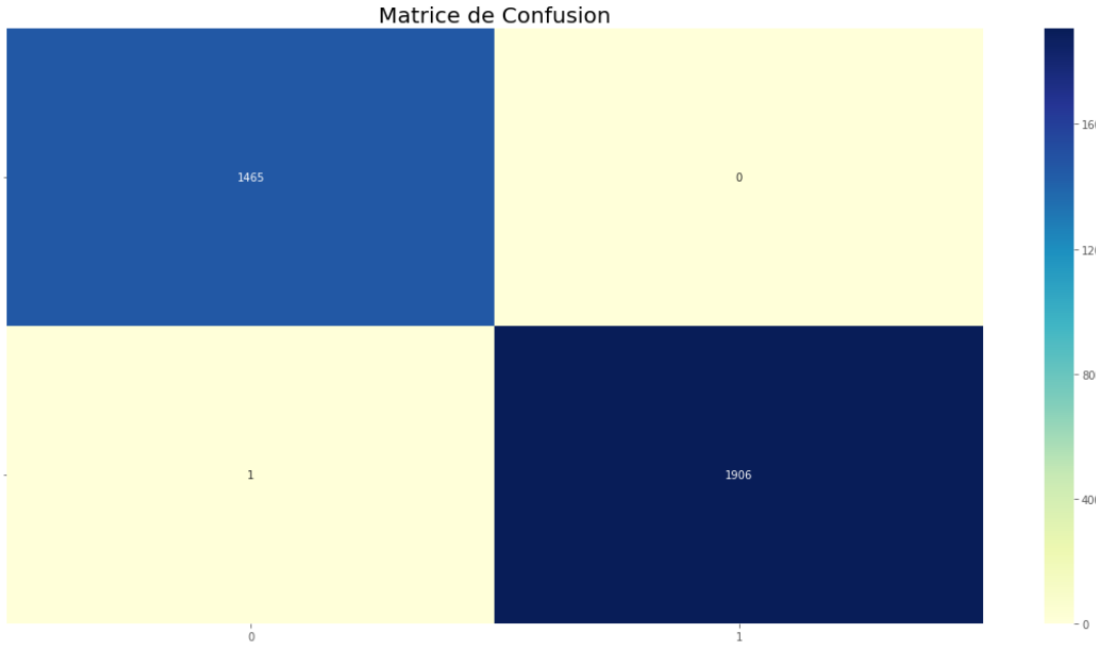
	Country	Gender	Age	Marital_Status	Work	Hypertension	Dyslipidemia	DM	Year_DM_Diagnosed	DM_Duration	DM_Type	Family_History_DM_father	Fam
1	0	0	73	1	0	1	1	0	0	0	0	0	
2	2	0	64	4	0	1	1	1	32	1	1	0	
3	0	0	56	1	0	0	1	1	18	18	1	0	
4	1	0	78	4	0	1	0	0	0	0	0	0	
5	2	1	54	1	0	1	0	0	0	0	0	0	
6	2	1	62	1	0	1	1	1	30	5	1	0	
7	2	1	59	1	0	0	1	1	32	1	1	0	
8	2	0	75	4	0	1	0	1	32	1	1	0	
9	2	0	62	1	0	1	0	1	29	6	1	0	
10	2	1	61	1	2	0	0	0	0	0	0	0	
11	2	1	78	1	0	1	1	0	0	0	0	0	
12	2	0	70	1	0	1	1	0	0	0	0	0	
13	3	1	79	1	0	1	1	1	19	18	1	0	
14	2	0	66	4	0	1	0	1	38	34	1	0	
15	2	0	62	1	0	1	1	1	30	5	1	0	
16	2	1	57	1	2	0	0	1	32	1	1	0	
17	2	1	73	1	0	1	1	1	30	5	1	0	
18	0	0	79	0	0	1	1	0	0	0	0	0	
19	0	1	68	1	0	1	1	0	0	0	0	0	

Maintenant les données sont prêtes il ne nous reste que la construction de notre modèle du Machine Learning, nous allons en utiliser 3 modèles et voir lequel d'entre eux donne le meilleur score de précision.

6.1 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.

Matrice de confusion :




```
k = KFold(n_splits=5, shuffle=True, random_state=0)
LR = LogisticRegression()
scoring = 'accuracy'
scoreLR = (cross_val_score(LR,x,y,cv=k, n_jobs=1, scoring=scoring))
scoreLR.round(1)
```

```
array([1., 1., 1., 1., 1.])
```

```
round(scoreLR.mean(),9)
```

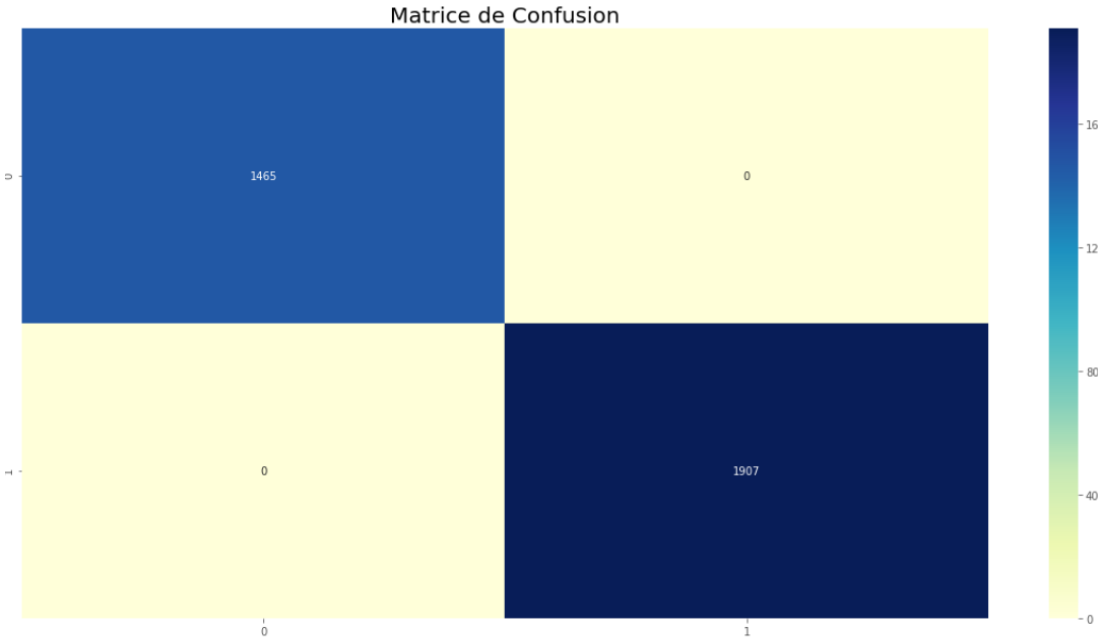
```
0.999703264
```

le score de précision est très robuste et converge vers 1

6.2 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.

Matrice de confusion :




```
randomforest_classifier= RandomForestClassifier(n_estimators=10)
scoring = 'accuracy'
scoreRF = (cross_val_score(randomforest_classifier, x, y, cv=k, n_jobs=1, scoring=scoring))
scoreRF.round(1)
```

```
array([1., 1., 1., 1., 1.])
```

```
round(scoreRF.mean(),200)
```

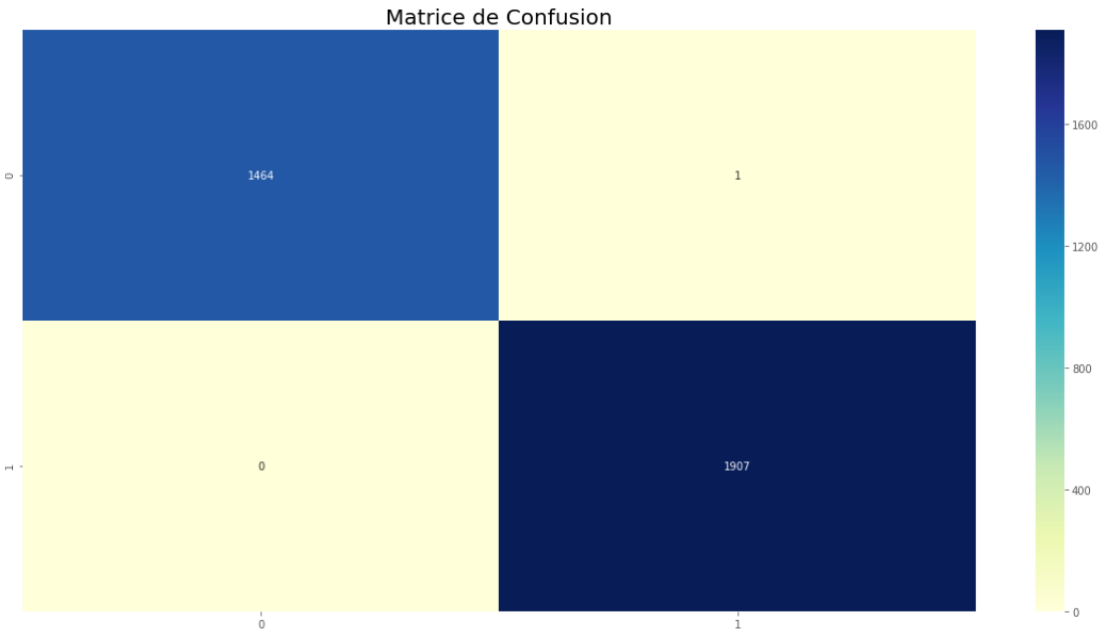
```
1.0
```

le score de précision de ce modèle est très robuste qui vaut 1

6.3 Arbre de décision : :

Un arbre de décision peut être décrit comme un diagramme de flux de données où chaque noeud 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).

Matrice de confusion :




```
1 DecisionTree = DecisionTreeClassifier()
2 scoring = 'accuracy'
3 scoreDT = (cross_val_score(DecisionTree, x, y, cv=k, n_jobs=1, scoring=scoring))
4 scoreDT.round(1)
```

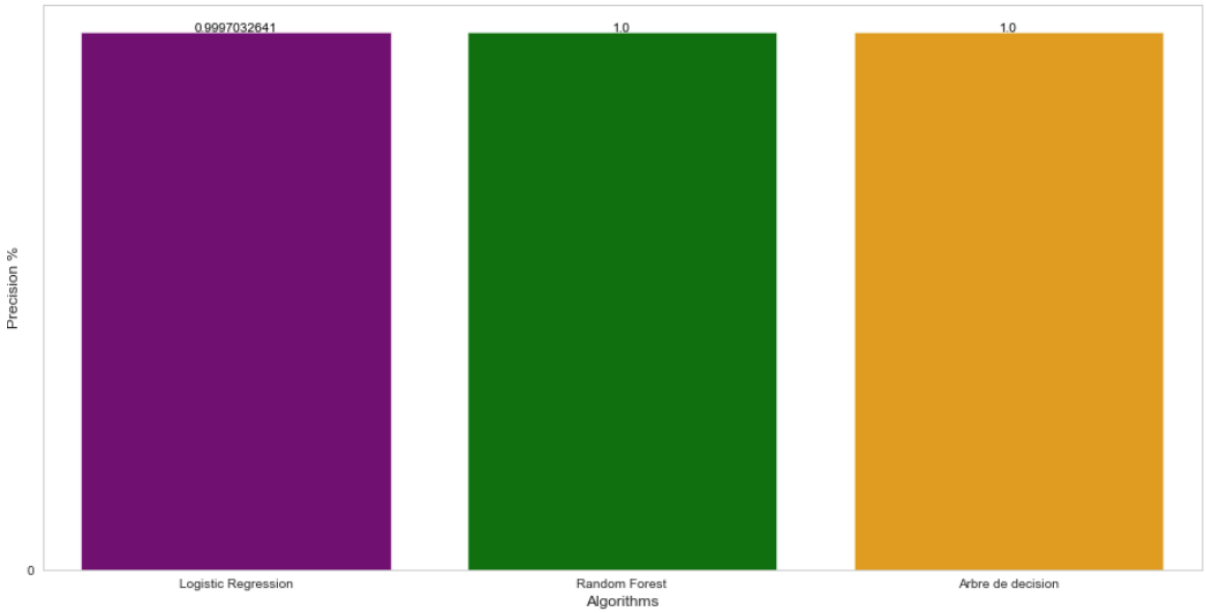
```
array([1., 1., 1., 1., 1.])
```

```
1 round(scoreDT.mean(),6)
```

```
1.0
```

le score de précision de ce modèle est très robuste qui vaut 1

6.4 Visualisation graphique des algorithmes :



Alors comme il est affiché sur notre graphe les trois algorithmes tel que **arbre de décision et régression logistique et Random Forest** ont donné la meilleure valeur de score de précision 100 % .Ce score montre que les trois modèles sont robuste grâce au processus de nettoyage adapté dés le début qui nous ramener à bien préparer notre données et avoir des résultats très fiables.