Master :

Bioinformatique et Modélisation des systèmes complexes appliquée à la santé

Projet Data Mining :

Prédiction de la maladie rénale chronique

Réalisé par :

AMERAN ISMAIL

EDDARIF HASNAA

FERDOUZ REDOUANE

SADIKI NOUR-EDDINE

Encadré par :

Dr. BENBRAHIM HOUDA

Année Universitaire : 2020 - 2021

SOMMAIRE

Liste des tableaux

Liste des Figures

Abréviations

ANN : Artificiel Neural Network (Réseau de neurones artificiel)

CKD : Chronical Kidney Diseas (Maladie Rénale Chronique)

CRISP-DM : Cross-Industry Process For Data Mining

DM : Data Mining

HIS : Hospital information system (systéme d'information hospitalier)

IRC : Infection Rénale Chronique

OMS : Organisation Mondiale de la Santé

1. Résumé

La maladie rénale chronique (MRC) est une affection caractérisée par une perte graduelle de la fonction rénale au fil du temps.

Dans les normes internationales, la MRC est organisée en différents degrés de stratification des risques à l'aide des signes prédéfinies.

Elle est généralement asymptomatique à ses débuts et une détection précoce est importante pour réduire les risques futurs. Cette étude a utilisé la méthodologie CRISP-DM (Cross Industry Standard Process for Data Mining) et les algorithmes de machine learning sous python pour construire un système capable de classer l'état chronique de l'insuffisance rénale en fonction de l'exactitude, de la sensibilité, de la spécificité et de la précision.

Les résultats obtenus ont été jugés satisfaisants, obtenant le résultat le plus approprié de 100% de précision.

2. INTRODUCTION

La maladie rénale chronique (MRC) est une affection courante, affectant des millions de personnes dans le monde, qui fait référence à une perte à long terme de la fonction rénale. Plus de deux millions de personnes dans le monde reçoivent une dialyse, alors que 10 % des cas atteignant la maladie subit à une greffe de rein ou un autre type de traitement pour rester en vie. Un autre aspect est que seulement 20% de la population est traitée dans les pays en développement, soit seulement la moitié de la population mondiale[1].

La MRC est une affection qui ne se manifeste pas immédiatement, entraînant l'absence de symptômes dans les premiers stades, étant dévaluée et ignorée par les personnes atteintes de cette maladie. La perte progressive de fonction, conduit à l'apparition de la pathologie à un stade très avancé. En tant qu'une maladie asymptomatique, ses complications ne sont pas facilement détectées et peuvent être confondues avec d'autres types de maladies. La plupart du temps, la MRC est détectée lors d'une insuffisance rénale, ce qui entraîne des mesures extrêmes telles que la greffe rénale et, si elles n'en résultent pas, la mort.

La MRC comporte cinq stades de lésions rénales, allant de très légères lésions au stade 1 à une insuffisance rénale complète au stade 5 (figure1). En dessous d'un certain seuil de filtration des reins, on parle d'insuffisance rénale chronique (IRC) dont l'évolution naturelle est plus ou moins lente mais peut aller jusqu'à la perte totale de la fonction rénale. C'est l'insuffisance rénale terminale, nécessitant un traitement de suppléance par dialyse et/ou greffe de rein, et ne peut être diagnostiquée à un stade précoce que par des néphrologues et des urologues expérimentés en utilisant les antécédents de la maladie, les symptômes et les tests de laboratoire.

[1] L'Organisation Mondiale de la Santé (OMS)

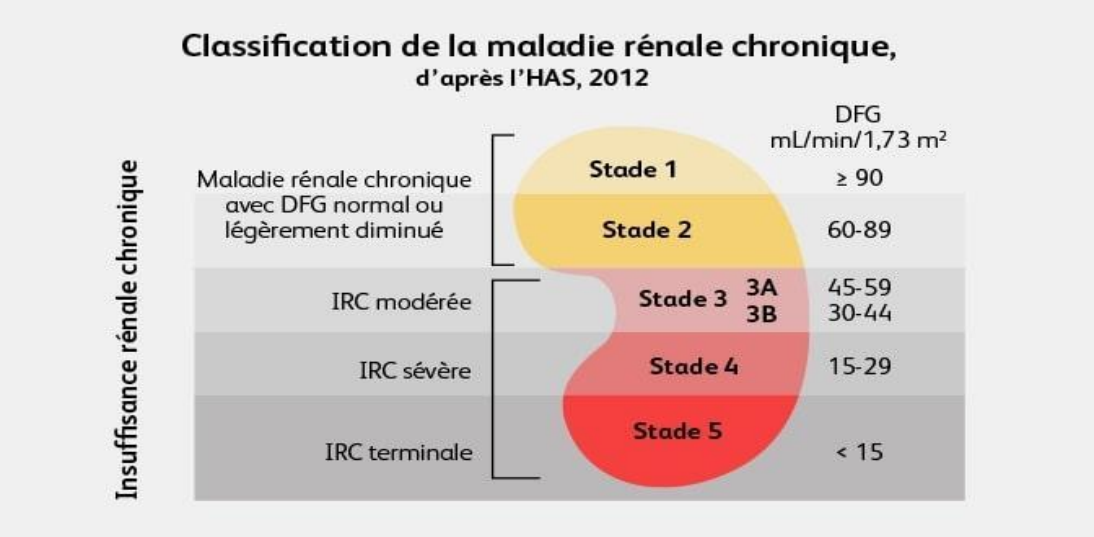


Figure 1 : les stades de la maladie rénale chronique

L'état des personnes qui nécessitent l'un des deux types de thérapie de remplacement rénal, la dialyse ou la greffe, est appelée maladie rénale en phase terminale.

Afin d'empêcher les personnes atteintes de cette maladie d'atteindre la phase terminale, les techniques de prédiction de la maladie à un stade précoce deviennent un avantage et peuvent sauver des vies.

Dans le domaine de la santé, la "Data Mining" DM a été largement utilisé pour prédire et classer les variables cliniques à l'aide des données stockées dans les systèmes d'information hospitaliers (HIS), qui peuvent être nettoyées et analysées à l'aide des techniques de DM.

DM est généralement utilisé pour extraire des informations utiles à partir de données brutes. Les techniques de DM peuvent aider à découvrir des informations cachées pour une meilleure prise de décision, découvrir des modèles cachés et des relations inexploitées.

Ce projet entre dans le cadre de l'application des techniques et outils appris dans le cours de module de Data & Web Mining.

Pour ce, notre travail se concentre sur les techniques de classification, en particulier dans le développement d'un classificateur capable de prédire si un individu aura une maladie rénale chronique noté « ckd » (chronical kidney diseas) ou « notckd » (not chronical kidney diseas), en utilisant la méthodologie CRISP-DM et les algorithmes du "Machine Learning".

L’exploration de données, le nettoyage et la construction de modèle sont programmés en langage python sur Jupyter Notebook, alors que le déploiement s’est posé sur une application web développée sous Flask et Javascript.

3. Littérature

Il existe des études liées au diagnostic automatique de la MRC dans la littérature, qui visent également, comme objectif, le mise en place d’un modèle d’apprentissage automatique pour aider les experts médicaux.

Pour la MRC, les auteurs utilisent généralement des modèles d'analyse prédictive pour prédire sa progression et, dans le meilleur des scénarios, tentent d'arrêter la maladie.

En 2010, une étude de cas pour la prédiction de la MRC dans un hôpital local en Angleterre a été présentée dans laquelle deux conditions ont été prises en compte : la MRC modérée à sévère et l'insuffisance rénale terminale, les algorithmes développés fournissent une base pour identifier les patients à haut risque qui pourraient bénéficier d'une évaluation plus détaillée, d'une surveillance plus étroite ou d'interventions pour réduire leur risque.

Plus tard, en 2011, Tangri et al développent des modèles de prédiction en utilisant des données démographiques, cliniques et des analyses biologiques de laboratoire. Selon cette étude, le modèle le plus précis comprenait l'âge, le sexe, le taux de filtration glomérulaire estimé, l'albuminurie, la calcémie, le phosphate sérique, le bicarbonate sérique et l'albumine sérique.

En 2014, l'utilisation des techniques de data mining pour prédire la survie des patients sous la dialyse rénale a été présentée, où trois techniques de data mining et de machine learning ont été utilisées, à savoir l'ANN, l'arbre de décision et la régression logistique, étant la première parmi les trois à atteindre une valeur de précision de 93,852%, une sensibilité de 93,87% et une spécificité de 93,87%.

4. Méthodologie : CRISP-DM

CRISP-DM (cross-industry process for data mining) est synonyme de processus intersectoriel pour l'exploration de données. La méthodologie CRISP-DM fournit une approche structurée pour planifier un projet d'exploration de données. C'est une méthodologie robuste et que sa flexibilité et son utilité sont éprouvée lors de l'utilisation de l'analytique pour résoudre des problèmes épineux.

Le modèle CRISP-DM schématisé sur la figure 2 ci-dessous est une séquence idéalisée d'événements. Dans la pratique, de nombreuses tâches peuvent être effectuées dans un ordre différent et il sera souvent nécessaire de revenir aux tâches précédentes et de répéter certaines actions. Le modèle n'essaie pas de capturer toutes les routes possibles à travers le processus d'exploration de données.

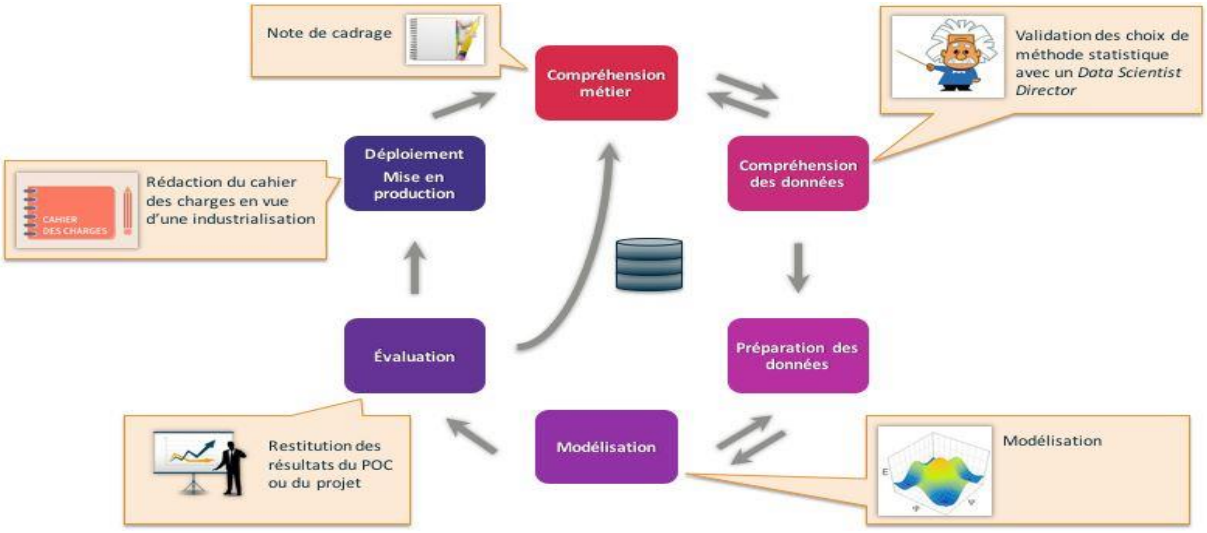


Figure 2 : Modèle CRISP-DM

4.1 Compréhension du métier (Business Understanding)

La prédiction précoce et les traitements appropriés peuvent éventuellement arrêter ou ralentir la progression de la MRC au stade terminal, où la dialyse ou la transplantation rénale sont le seul moyen de sauver la vie du patient. Il devient de plus en plus important d'utiliser ce type de prédictions car il est basé sur des informations de patients réels, permettant la création de modèles de prédiction qui aident les professionnels de santé dans leurs décisions.

Dans ce contexte, ce projet vise à prédire les cas de la MRC à un stade précoce en se basant sur des données cliniques grâce à l'utilisation de techniques de DM.

4.2 Compréhension de données (Data Understanding)

Dans ce travail, on a utilisé une base de données liées à la détection de la MRC, collectées dans un hôpital en Inde sur une période de 2 mois.

Cette dataset contient des informations sur 400 patients et 24 caractéristiques importantes pour identifier les facteurs de risque de la MRC et améliorer le diagnostic des cas de patients atteints de la MRC. L'ensemble de données contient des données personnelles, des données de tests de laboratoire sur le sang et l'urine et des données sur les antécédents cliniques du patient.

Les attributs de l'ensemble de données sont présentés ci-après.

Tableau 1 : Listes des attributs avec description et type

Attribut	Description de la variable	Type
id	Numéro d'identification	Numérique
Age	Age en année	Numérique
blood_pressure	Tension artérielle en mm/Hg	Numérique
Specific_gravity	Score de degré de la densité urinaire	Nominal
albumin	Score de degré d'albumine dans le sang	Nominal
sugar	Score de degré de glycémie dans le sang	Nominal
red_blood_cells	La forme des globules rouge	Binaire
pus_cell	La forme des cellule de pus	Binaire
pus_cell_clumps	Présence d'amas dans les cellule de pus	Binaire
bacteria	Présence de la bactérie	Binaire
blood_glucose_random	Mesure de glucose dans le sang en mgs/dl	Numérique
blood_urea	Mesure d'urée sanguine en en mgs/dl	Numérique
serum_creatinine	Mesure de la créatinine sérique en mgs/dl	Numérique
sodium	Mesure du sodium en mEq/L	Numérique
potassium	Mesure du potassium en mEq/L	Numérique
hemoglobine	Mesure d'hémoglobine en gms	Numérique
packed_cell_volume	Mesure d'hématocrite	Numérique
white_blood_cell_count	Nombre de globule blanc dans le sang en cells/cumm	Numérique
red_blood_cell_count	Nombre de globule rouge dans le sang en millions/cmm	Numérique
hypertension	Présence de la maladie d'hypertension artérielle	Binaire
diabetes_mellitus	Présence de la maladie de diabète	Binaire
coronary_artery_disease	Présence de la maladie de l'artère coronaire	Binaire
appetite	Etat de l'appétit	Binaire
pedal_edema	Présente de la maladie de œdème de la pédale	Binaire
anemia	Présence de la maldie d'anémie	Binaire
Classification	Présence de la maladie rénale chronique	Binaire

La variable cible "classification " indique "ckd" pour les cas qu'ont un IRC et "notckd" pour les personnes qui n'ont pas d'IRC.

L'attribut cible a une distribution déséquilibrée avec 250 cas correspondant à ckd et seulement 150 à notckd

Les principales mesures descriptives de données sur le tableau 2 ci-dessous.

Tableau 2 : Mesures descriptives des attributs

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Age	391	NaN	NaN	NaN	51.48	17.17	2	42	55	64.5	90
blood_pressure	388	NaN	NaN	NaN	76.46	13.68	50	70	80	80	180
specific_gravity	353	NaN	NaN	NaN	1.017	0.006	1.00	1.01	1.02	1.02	1.025
albumin	354	NaN	NaN	NaN	1.02	1.35	0	0	0	2	5
Sugar	351	NaN	NaN	NaN	0.45	1.09	0	0	0	0	5
red_blood_cells	248	2	normal	201	NaN	NaN	NaN	NaN	NaN	NaN	NaN
pus_cell	335	2	normal	259	NaN	NaN	NaN	NaN	NaN	NaN	NaN
pus_cell_clumps	396	2	notpresent	354	NaN	NaN	NaN	NaN	NaN	NaN	NaN
bacteria	396	2	notpresent	374	NaN	NaN	NaN	NaN	NaN	NaN	NaN
blood_glucose_random	356	NaN	NaN	NaN	148.04	79.28	22	99	121	163	490
blood_urea	381	NaN	NaN	NaN	57.43	50.50	1.5	27	42	66	391
serum_creatinine	383	NaN	NaN	NaN	3.07	5.74	0.4	0.9	1.3	2.8	76
Sodium	313	NaN	NaN	NaN	137.53	10.41	4.5	135	138	142	163
potassium	312	NaN	NaN	NaN	4.63	3.19	2.5	3.8	4.4	4.9	47
haemoglobin	348	NaN	NaN	NaN	12.53	2.91	3.1	10.3	12.65	15	17.8
packed_cell_volume	330	44	41	21	NaN	NaN	NaN	NaN	NaN	NaN	NaN
white_blood_cell_count	295	92	9800	11	NaN	NaN	NaN	NaN	NaN	NaN	NaN
red_blood_cell_count	270	49	5.2	18	NaN	NaN	NaN	NaN	NaN	NaN	NaN
hypertension	398	2	no	251	NaN	NaN	NaN	NaN	NaN	NaN	NaN
diabetes_mellitus	398	5	no	258	NaN	NaN	NaN	NaN	NaN	NaN	NaN
coronary_artery_disease	398	3	no	362	NaN	NaN	NaN	NaN	NaN	NaN	NaN
appetite	399	2	good	317	NaN	NaN	NaN	NaN	NaN	NaN	NaN
pedal_edema	399	2	no	323	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Anemia	399	2	no	339	NaN	NaN	NaN	NaN	NaN	NaN	NaN
classification	400	3	ckd	248	NaN	NaN	NaN	NaN	NaN	NaN	NaN

4.3 Préparation des données (Data Preparation)

Une préparation et un nettoyage des données est nécessaire avant de passer aux phases suivantes. La première étape consiste à vérifier les incohérences dans les données, les valeurs en double, les valeurs aberrantes et les valeurs manquantes.

Il a été constaté qu'il n'y avait pas de valeurs en double dans les données. Toutes les valeurs manquantes détectées ont été remplacées par le mode pour les attributs nominaux et par la moyenne pour les attributs numériques. Puis, la suppression des valeurs aberrante détectées sur les données.

Après vérification de la corrélation, il a été constaté que notre variable cible 'classification' est corrélée aux variables « hemoglobine » et « packed cell volume » de 69% et 65% respectivement, pour ce, on a enlevé ces attributs.

Finalement les données ont été normalisées et mises en même échelle avec la méthode de normalisation minimax.

4.3.1 Ajustement des incohérences des données

Les incohérences détectées sur les variables de type binaire, consistent à la mal écriture des modalités, ces erreurs ont été corrigées sur les variables « diabete_mellitus », « coronary_artery_disease » et « classification ».

Pour les variables numériques, certaines données ont figuré de type objets ou bien chaine de caractères, ce qui a nécessité des redressions de ces données notamment sur les variables ; « red_blood_cell_count », « packed_cell_volume » et « white_blood_cell_count ».

4.3.2 Détection et imputation des valeurs manquantes

La présentation des valeurs manquantes est sur la figure 3 ci-dessous :

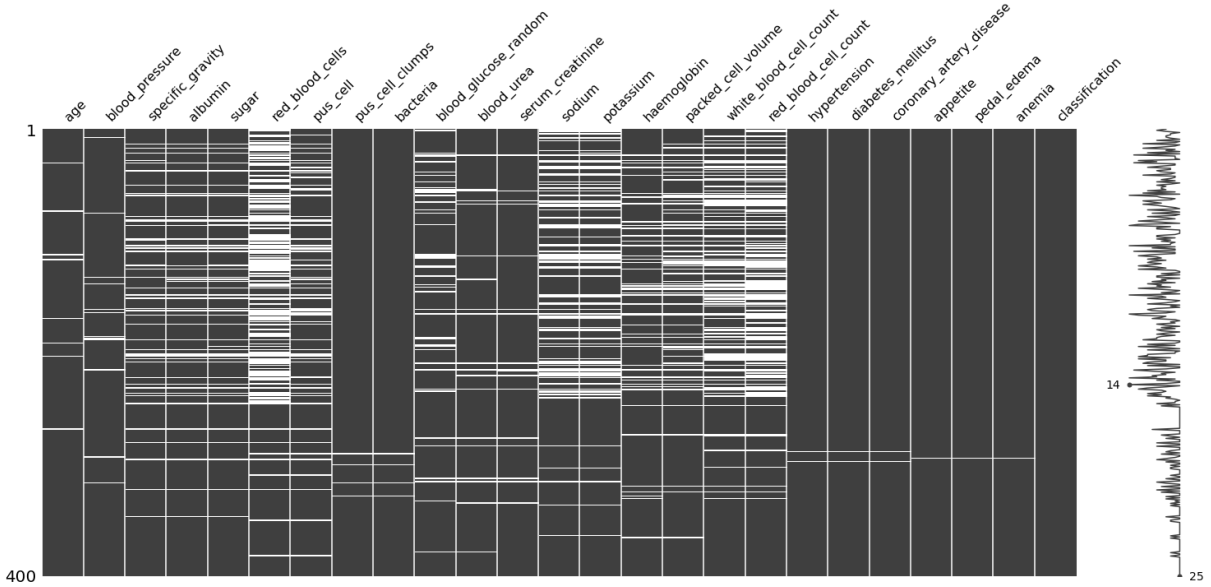


Figure 3 : les valeurs manquantes par attribut

Ci-après, sur le tableau 3 le nombre des valeurs manquantes présentées d'une façon descendante par attribut.

Tableau 3 : Nombre des valeurs manquantes par attributs

red_blood_cells	152
red_blood_cell_count	130
white_blood_cell_count	105
potassium	88
sodium	87
packed_cell_volume	70
pus_cell	65
haemoglobin	52
sugar	49
specific_gravity	47
albumin	46
blood_glucose_random	44
blood_urea	19
serum_creatinine	17
blood_pressure	12
age	9
bacteria	4
pus_cell_clumps	4
hypertension	2
diabetes_mellitus	2
coronary_artery_disease	2
anemia	1
appetite	1
pedal_edema	1
classification	0

- L'imputation des valeurs manquantes sur les variables numériques s'est faite par la moyenne

- l'imputation des valeurs manquantes sur les variables catégorielles binaires, s'est faite par le mode après transformation des modalités binaires en format numérique {"0", "1"}

Tableau 4 : Présentation des attributs avant et après imputation et corrections de type

Avant imputation

age	391 non-null	float64
blood_pressure	388 non-null	float64
specific_gravity	353 non-null	float64
albumin	354 non-null	float64
sugar	351 non-null	float64
red_blood_cells	248 non-null	object
pus_cell	335 non-null	object
pus_cell_clumps	396 non-null	object
bacteria	396 non-null	object
blood_glucose_random	356 non-null	float64
blood_urea	381 non-null	float64
serum_creatinine	383 non-null	float64
sodium	313 non-null	float64
potassium	312 non-null	float64
haemoglobin	348 non-null	float64
packed_cell_volume	329 non-null	float64
white_blood_cell_count	294 non-null	float64
red_blood_cell_count	269 non-null	float64
hypertension	398 non-null	object
diabetes_mellitus	398 non-null	object
coronary_artery_disease	398 non-null	object
appetite	399 non-null	object
pedal_edema	399 non-null	object
anemia	399 non-null	object
classification	400 non-null	object

Après imputation

age	400 non-null	float64
blood_pressure	400 non-null	float64
specific_gravity	400 non-null	float64
albumin	400 non-null	float64
sugar	400 non-null	float64
red_blood_cells	400 non-null	float64
pus_cell	400 non-null	float64
pus_cell_clumps	400 non-null	float64
bacteria	400 non-null	float64
blood_glucose_random	400 non-null	float64
blood_urea	400 non-null	float64
serum_creatinine	400 non-null	float64
sodium	400 non-null	float64
potassium	400 non-null	float64
haemoglobin	400 non-null	float64
packed_cell_volume	400 non-null	float64
white_blood_cell_count	400 non-null	float64
red_blood_cell_count	400 non-null	float64
hypertension	400 non-null	float64
diabetes_mellitus	400 non-null	float64
coronary_artery_disease	400 non-null	float64
appetite	400 non-null	float64
pedal_edema	400 non-null	float64
anemia	400 non-null	float64
classification	400 non-null	int64

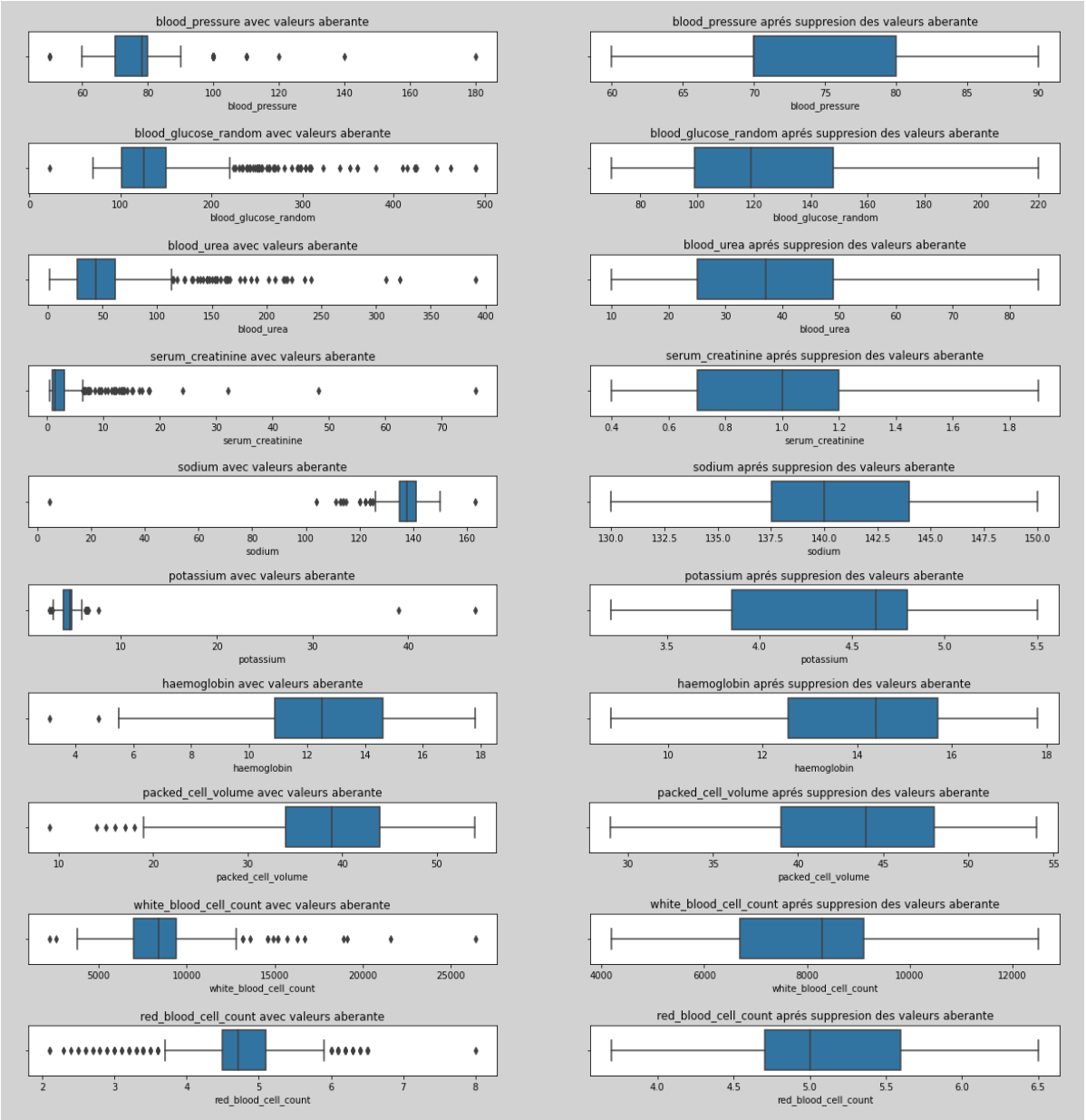
4.3.3 Détection et suppression des valeurs aberrantes

Dans cette opération, on a utilisé la méthode de l'écart interquartiles pour l'identification des valeurs aberrantes avec l'aide de boite à moustache (boxplot).

Une valeur est dite aberrante si la valeur absolue de l'écart avec Q1 ou Q3 est supérieur à plus de 1,5x l'écart interquartile (EQ). Plus précisément, une valeur aberrante est faible s'elle est inférieure à Q1-1.5 EQ et élevée s'elle est supérieure à Q3+1,5EQ.

La figure 4 ci-après illustre les "boxplot" des variables numériques avant et après suppression des valeurs aberrantes.

Figure 4 : les variables avant et après suppression des valeurs aberrantes



4.3.4 Etude de corrélation

La corrélation entre deux variables aléatoires, est l'intensité de la liaison qu'il existe entre ces deux variables.

Afin de déterminer cette liaison il suffit de calculer le coefficient de corrélation par la formule suivante :

$$\rho(X;Y) = \frac{\sum(X-\bar{X}).(Y-\bar{Y})}{\sqrt{\sum(X-\bar{X})^2}.\sqrt{\sum(Y-\bar{Y})^2}} = \frac{Cov(X;Y)}{\sqrt{V(X)}.\sqrt{V(Y)}}$$

X et Y : *deux variables aléatoires*
$\bar{X}$ et $\bar{Y}$: *moyenne respective des variables X et Y*
$Cov(X;Y)$: *Covariance entre X et Y*
$V(X)$ et $V(Y)$: *Variance respective de X et Y*

Le coefficient de corrélation est toujours compris entre -1 et +1, deux variables sont corrélées si le coefficient est proche des extrémités (-1 , +1), c'est à dire sont dépendantes l'une à l'autre.

Une corrélation égale à +1 peut être interprété qu'il existe une relation linéaire positive entre les variables, de même, quand le coefficient égal -1 c'est une liaison négative entre les variables.

Si la corrélation est nulle, on constate que les deux variables sont décorrélées. Autrement dit, qu'il n'existe pas de relation entre elles.

Dans notre cas, la figure 5 présente la matrice de corrélation des attributs de type numérique.

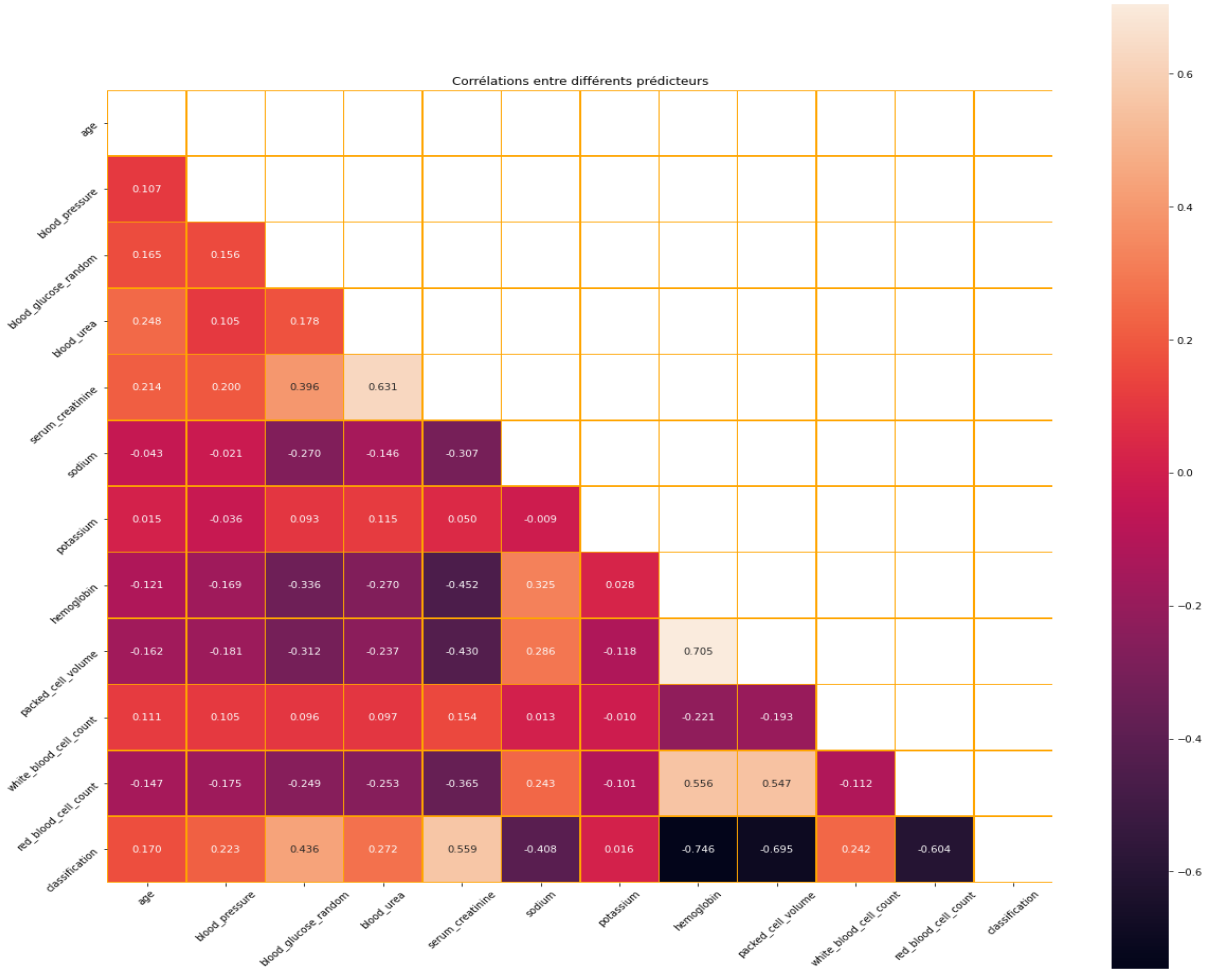


Figure 5 : La matrice de corrélation entre les attributs de type numérique

Comme c'est illustré sur la matrice de corrélation ci-dessus ; les deux variables "hemoglobine" et "packed cell volume » ont un coefficient de corrélation respectivement de 74,6% et 69,5% qu'on peut les considérer fortement corrélées et par conséquence enlever ces deux attributs.

4.3.5 Normalisation des données

Les méthodes de normalisation des données sont utilisées pour que les variables, mesurées à différentes échelles, aient des valeurs comparables. Cette étape de prétraitement est importante pour le regroupement et la visualisation des heatmap, l'analyse en composantes principales et d'autres algorithmes d'apprentissage automatique basés sur des mesures de distance.

La normalisation standardise la moyenne et l'écart-type de tout type de distribution de données, ce qui permet de simplifier le problème d'apprentissage en s'affranchissant de ces deux paramètres.

Lorsque les variables des données proviennent de distributions éventuellement différentes (et non normales), d'autres transformations peuvent être nécessaires. Une autre possibilité consiste à normaliser les variables pour amener les données sur l'échelle de 0 à 1 en soustrayant le minimum et en divisant par le maximum de toutes les observations.

Formulation pour normaliser les données entre 0 et 1 :

$$Transformed.Values = \frac{Values - Minimum}{Maximum - Minimum}$$

4.4 Modélisation (Modeling)

Après la préparation des données, il était nécessaire de construire différents modèles de Machine Learning.

Tous les modèles ont été construits selon une approche supervisée de classification. De plus, 9 différents modèles ont été utilisés : Régression logistique, Arbre de Décision, XGB classifier, Gradient Boosting, Réseaux de Neurones, Gaussian Naive Bayes, k-Nearest Neighbours, Suport Vector Machine et Random Forest.

La répartition des données a été proportionnellement de 66% pour l'entrainement (training data) et le nombre de données restantes ont été réservées pour le test (testing data).

Les résultats d'entrainement par modèle sont affichés sur le tableau 5 ci-dessous

Tableau 5 : Accuracy et marge d'erreur par modèle de classification

Modèle	Accuracy	Erreur
Logistic Regression	1.000000	(0.000000)
Decision Tree	0.994118	(0.017647)
XGB	0.994118	(0.017647)
Gradient Boosting	0.994118	(0.017647)
K-Neighbors	1.000000	(0.000000)
Gaussian NB	0.947712	(0.060753)
Neural Network	0.930719	(0.106936)
Random Forest	1.000000	(0.000000)

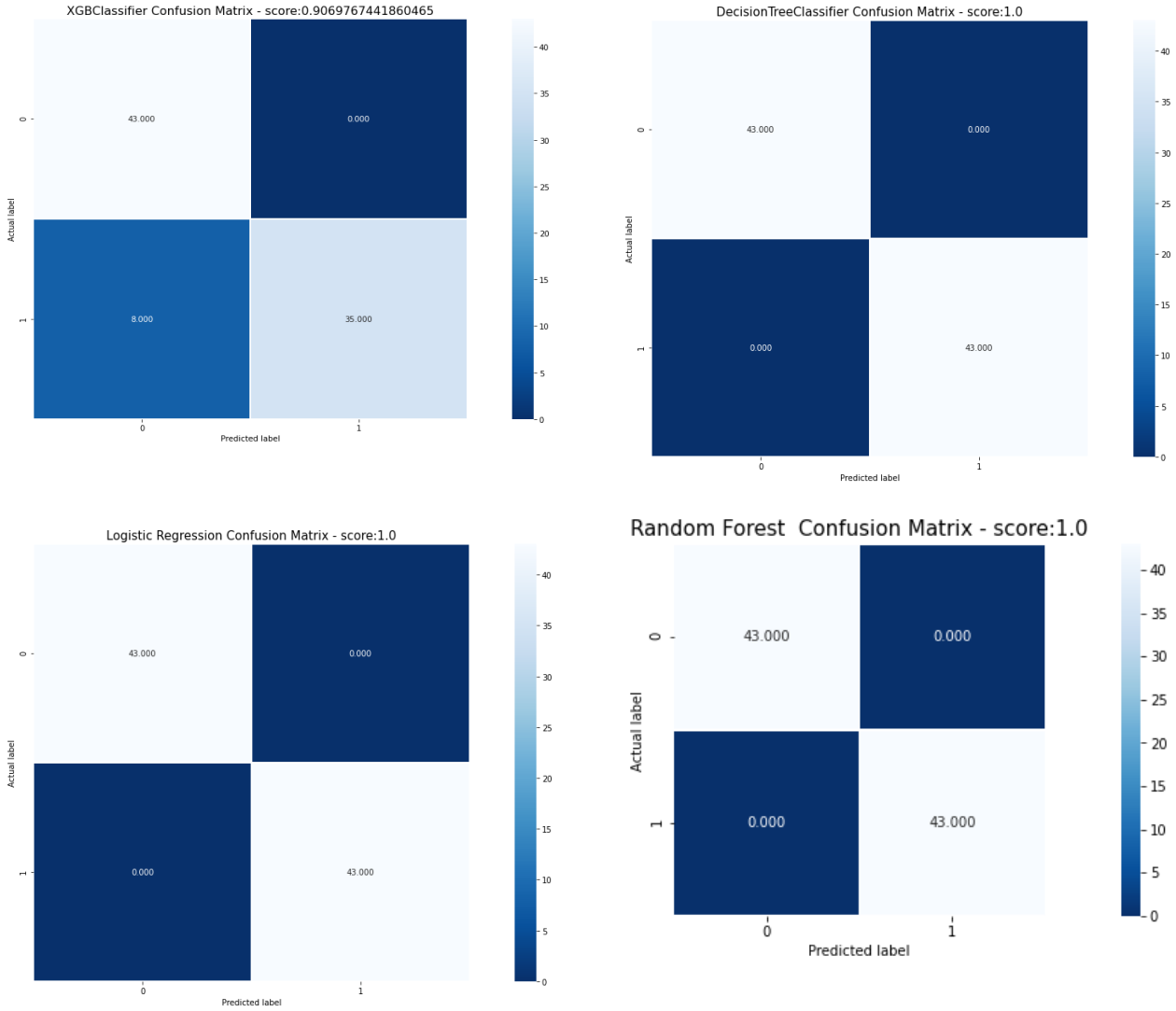
4.5 Evaluation

Les performances de chaque modèle de machine learning ont été évaluées au moyen de certaines métriques pour évaluer la qualité des modèles en garantissant la fiabilité des résultats. Toutes les métriques ont été obtenues à partir de la matrice de confusion.

La matrice de confusion pour une classification binaire présente une matrice 2x2 avec les classes prédites, montrant le nombre de vrais négatifs, faux négatifs, faux positifs et vrais positifs, qui peuvent être combinés pour mesurer comment le modèle a fonctionné.

Quatre paramètres ont été calculés, utilisés pour l'évaluation de la présente étude, à savoir l'exactitude, la sensibilité, la spécificité et la précision.

Figure 6 : Matrice de confusion des modèles de machine learning

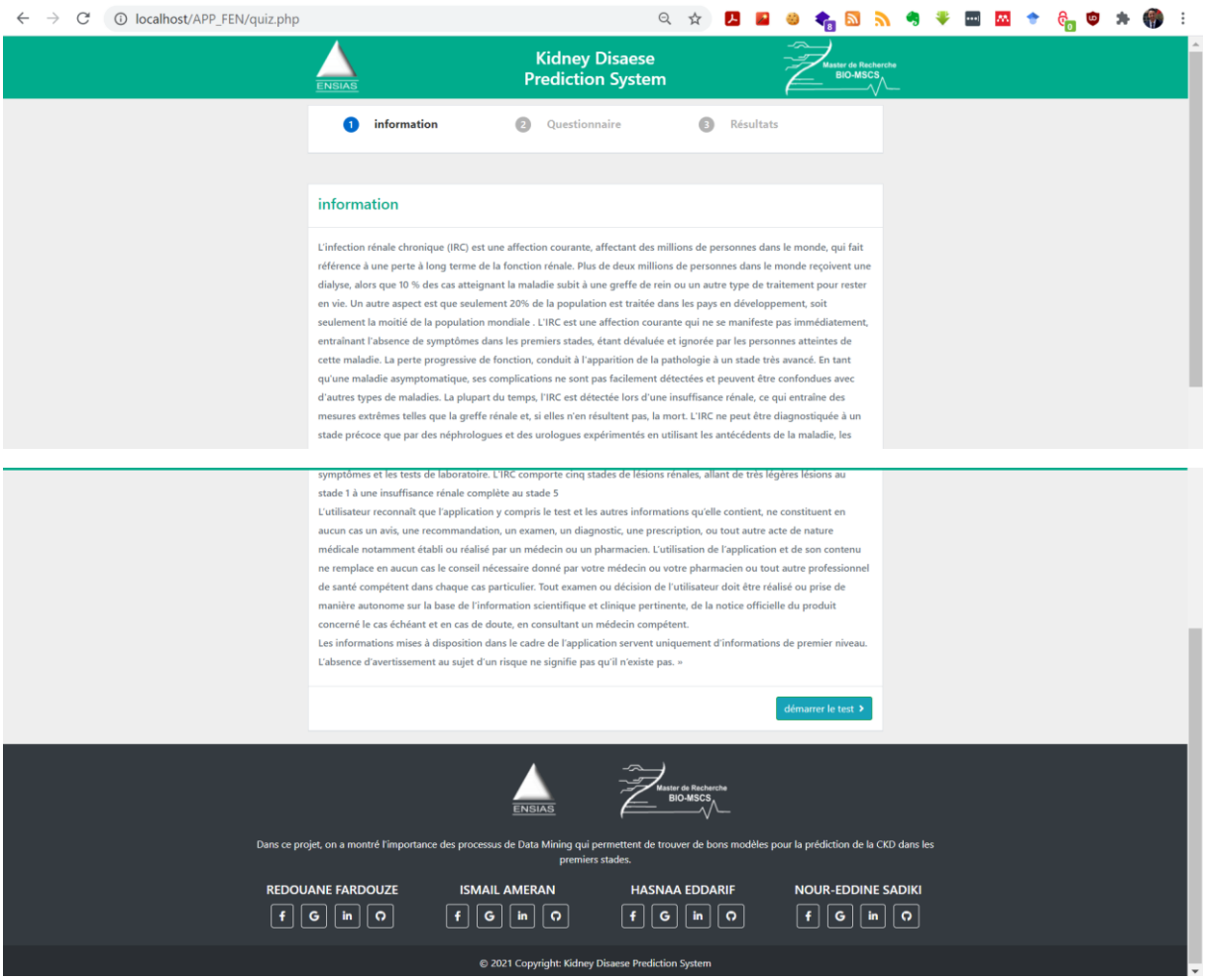


4.6 Déploiement

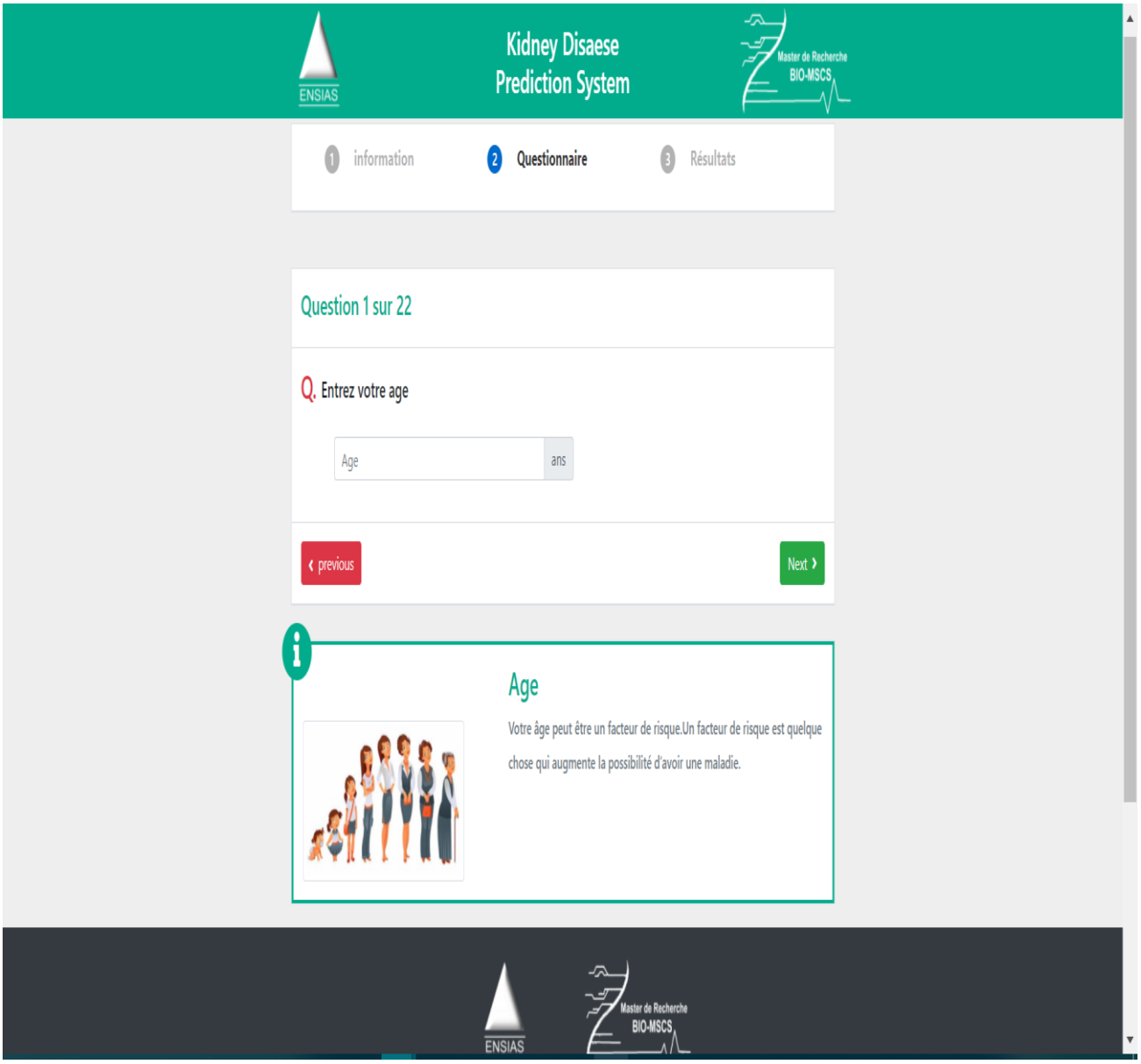
Le déploiement du modèle de prédiction (**Random Forest**) qu'est déjà entrainé et évalué nous l'avons appelé à notre application web développée sous **le Framework Flask** de Python. Pour réutilise ce modèle en mode extérieure, nous l'avons sauvegardé sous forme **« `'model_random.pkl'`»**

Présentation de l'application :

L'interface graphique de notre application (Front End) est développée à l'aide de Framework **Bootstrap 4** qui utilise les **langages HTML, CSS et JavaScript** nous facilite à créer ce site.
Ce Framework est pensé pour développer des sites avec un design responsif, qui s'adapte à tout type d'écran, et en priorité pour les smartphones.
L'interface ci-dessous nous permettra de faciliter l'utilisation de l'application de la part des personnes non-informaticien tel que les docteurs, les infirmières et d'autres.



L'application stimule les attributs de prédiction sous forme d'un quiz, des questions posées à une personne susceptible d'avoir une maladie rénale, pour en déduire et prédire à la fin de quiz si la personne pourra avoir la maladie ou non en se basant sur l'algorithme d'apprentissage (**Random Forest**) d'on a déjà entrainé et évaluée par nous.



5. Résultat et discussion

Les matrices obtenues ont été analysées afin de choisir le modèle qui est capable de réaliser la meilleure prédiction de la maladie. Cette étude s'inscrivant dans le cadre du diagnostic médical, le principal critère de sélection du meilleur modèle doit être la précision, car il s'agit d'un domaine où le taux associé au faux négatif est élevé.

comme l'exactitude était la mesure qui obtenait les valeurs les plus élevées, une meilleure précision étaient le critères de sélection pour définir le meilleur modèle, afin d'éviter un nombre élevé de faux positifs.

Comme on peut le voir dans les résultats, tous les modèles ont atteint, la plupart du temps, des valeurs de haute précision, mais il a été décidé de ne pas considérer ces résultats comme les meilleurs car le jeu de données utilisé dans ce cas a été mis en équilibre, ce qui peut conduire à une mesure non fiable des performances du modèle.

En ce sens, à première vue, on pourrait considérer que le meilleur modèle était celui réalisé sous l'algorithme de Random Forest, qui a obtenu une valeur de précision de 100%.

6. CONCLUSION

Au fur et à mesure que la MRC progresse lentement, une détection précoce et un traitement efficace sont les piliers pour réduire les impacts que cette maladie peut causer.

Dans cette étude, une approche de data mining a été élaborée pour prédire le stade précoce des cas de CKD, étant le meilleur résultat obtenu avec l'algorithme Random Forest et un balançage des données.

Le résultat a été obtenu avec le modèle de machine learning, et a abouti aux mesures parfaites avec des 100% de précision, de sensibilité et de spécificité.

Dans ce projet, on a montré l'importance des processus de Data Mining qui permettent de trouver de bons modèles pour la prédiction de la CKD dans les premiers stades.

Les connaissances acquises avec l'utilisation des techniques de DM peuvent être utilisées pour prendre des décisions cliniques réussies en ce qui concerne le diagnostic de la MRC, en aidant à soutenir les professionnels de la santé dans la prise de décision concernant le diagnostic du patient et en améliorant la qualité des services fournis aux patients ainsi que les chances de ne pas avoir besoin d'un traitement de remplacement rénal, comme la dialyse ou la transplantation rénale.

En outre, il serait extrêmement important d'utiliser une autre base de données avec plus d'instances afin d'avoir un jeu de données plus grand et équilibré, permettant d'améliorer la fiabilité de cette étude. En outre, d'autres métriques peuvent être utilisées afin d'explorer plus de paramètres d'évaluation, correspondant ainsi à des conclusions plus solides.

7. Références

Aljaaf, A.J., Al-Jumeily, D., Haglan, H.M., Alloghani, M., Baker, T., Hussain, A.J., Mustafina, J., 2018. Early prediction of chronic kidney disease using machine learning supported by predictive analytics, in : 2018 IEEE Congress on Evolutionary Computation (CEC), IEEE.

Ferreira, D., Silva, S., Abelha, A., Machado, J., 2020. Recommendation system using autoencoders. Applied Sciences 10, 5510.

Fraser, S.D., Blakeman, T., 2016. Chronic kidney disease: identification and management in primary care. Pragmatic and observational research 7, 21.

Hippisley-Cox, J., Coupland, C., 2010. Predicting the risk of chronic kidney disease in men and women in england and wales: prospective derivation and external validation of the qkidneyR scores. BMC family practice 11, 49.

Lakshmi, K., Nagesh, Y., Krishna, M.V., 2014. Performance comparison of three data mining techniques for predicting kidney dialysis survivability. International Journal of Advances in Engineering & Technology 7, 242. Levey, A.S., Coresh, J., 2012. Chronic kidney disease. The lancet 379, 165–180.

Neto, C., Brito, M., Lopes, V., Peixoto, H., Abelha, A., Machado, J., 2019. Application of data mining for the prediction of mortality and occurrence of complications for gastric cancer patients. Entropy 21, 1163.

Palaniappan, S., Awang, R., 2008. Intelligent heart disease prediction system using data mining techniques, in: 2008 IEEE/ACS International Conference on Computer Systems and Applications, pp. 108–115.

Tangri, N., Stevens, L.A., Griffith, J., Tighiouart, H., Djurdjev, O., Naimark, D., Levin, A., Levey, A.S., 2011. A predictive model for progression of chronic kidney disease to kidney failure. Jama 305, 1553–1559.

UCI, 2015. Uci machine learning repository: chronic kidney disease data set. URL : https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease.

OMS : The global burden of kidney disease and the sustainable development goals.