Rapport du projet Machine Learning

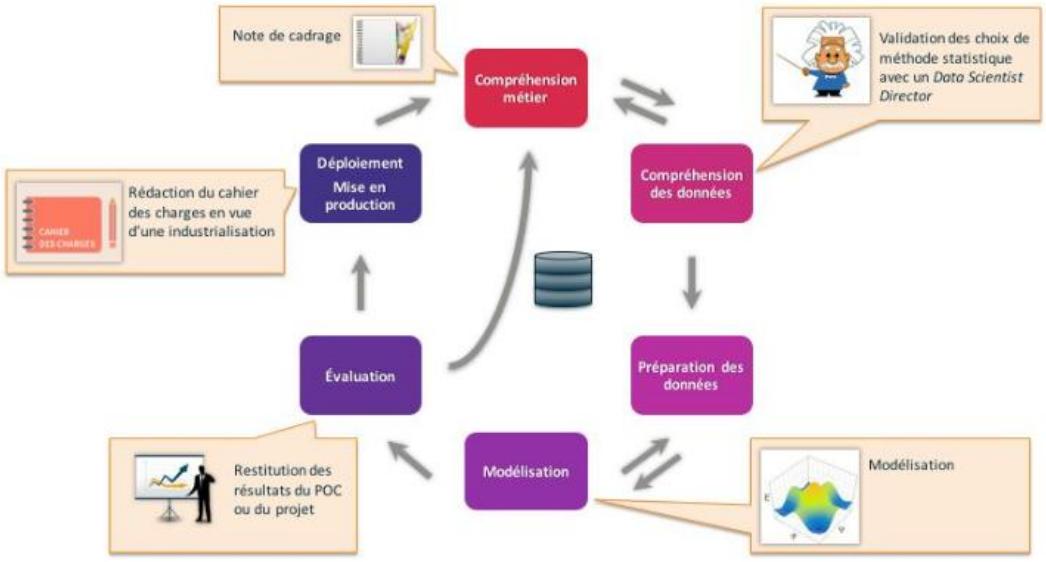
Réalisé par :

Sana Bouhaouala , Melek Abid , Malek Zommit Chatti , Yoser Walha , Bilel Kammoun , Omar Nouri

4DS5

La Méthodologie :

Le **Cross Industry Standard Process for Data Mining** (**CRISP-DM**) est un modèle de processus d'exploration de données qui décrit une approche communément utilisée pour résoudre les problèmes du domaine de l'analyse, de l'extraction et des sciences des données. Cette méthodologie a la particularité d'adopter une démarche cyclique et itérative, semblable à celle du modèle Agile, permettant une meilleure appréhension des spécificités de chaque projet. Nous avons choisi de réaliser notre travail en utilisant la méthode CRISP-DM.



La méthode CRISP illustrée

I. Compréhension du métier :

La première étape de la méthode CRISP-DM consiste à bien comprendre le métier, c'est à dire le sujet de notre base de données. Dans notre cas, il s'agit de la maladie : insuffisance rénale chronique.

L'insuffisance rénale chronique est une maladie irréversible d'apparition lente, le plus souvent liée au diabète et à l'hypertension artérielle. Les reins cessent progressivement de fonctionner et les déchets du métabolisme s'accumulent dans le corps. La dialyse et la greffe

de rein sont les traitements les plus fréquents de cette maladie grave dont les symptômes n'apparaissent que tardivement

L'insuffisance rénale chronique (IRC) est une maladie causée par la fonction de dégénérescence des reins. IRC est le top 10 en tête des causes de décès dans le monde. Il y a deux causes principales d'IRC : le diabète et l'hypertension.

Cependant, les signes et les symptômes ne sont pas souvent spécifiques. C'est pour cette raison que, dans certains cas, le diagnostic du personnel médical peut être subjectif et varier.

II. Compréhension des données :

La deuxième étape de la méthode CRISP-DM consiste à bien comprendre les données sur lesquels nous allons travailler.

- Notre base de données est sous forme d'un tableau
- Notre base de données contient :
  - ❖ 400 observations : `250 CKD`
    `150 notckd`
  - ❖ 24 Features + class = `25 ( 11 numeric ,14 nominal )`
  - ❖ 1 variable dite "Target" : classification ("ckd", "notckd")

Les Features	Type	numérique/ nominal	nombre de valeurs manquantes
Âge (âge)	float64	numérique	9
Blood Pressure (bp)	float64	numérique (mm/Hg)	12
Specific Gravity (sg)	float64	nominal	47
Albumin (al)	float64	nominal	46
Sugar (su)	float64	nominal	49
Red Blood Cells (rbc)	object	nominal	152
Pus Cell (pc)	object	nominal	65
Pus Cell clumps (pcc)	object	nominal	4
Bacteria (ba)	object	nominal	4
Blood Glucose Random (bgr)	float64	numérique (mgs/dl)	44
Blood Urea (bu)	float64	numérique (mgs/dl)	19
Serum Creatinine (sc)	float64	numérique (mgs/dl)	17

Sodium (sod)	float64	numérique (mEq/L)	87
Potassium (pot)	float64	numérique (mEq/L)	88
Hemoglobin (hemo)	float64	numérique (gms)	52
Packed Cell Volume (pcv)	object	numérique	70
White Blood Cell (wc)	object	numérique (cells/cumm)	105
Red Blood Cell Count (rc)	object	numérique (milloins/cmm)	130
Hypertension (htn)	object	nominal	2
Diabetes Mellitus (dm)	object	nominal	2
Coronary Artery Disease (cad)	object	nominal	2
Appetite (appet)	object	nominal	1
Pedal Edema (pe)	object	nominal	1
Anemia (ane)	object	nominal	1
Class	object	nominal	0

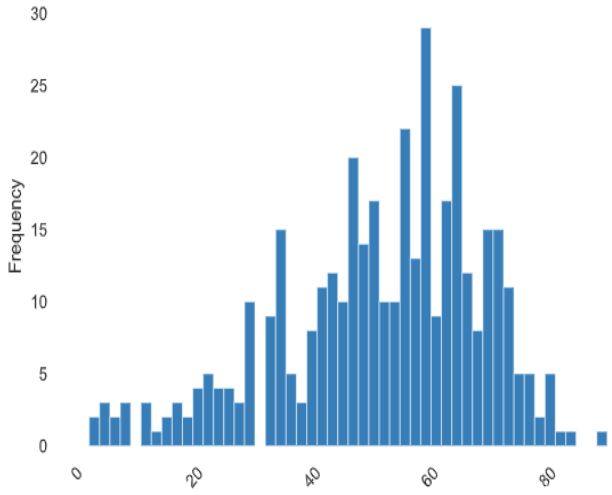
Nous avons donc décelé plus de 1000 valeurs manquantes dans toute la base de données. Chaque feature comportait au moins une donnée manquante.

Nous avons également remarqué que certaines features étaient de type object mais réellement, elles ne contenaient que des valeurs numériques.

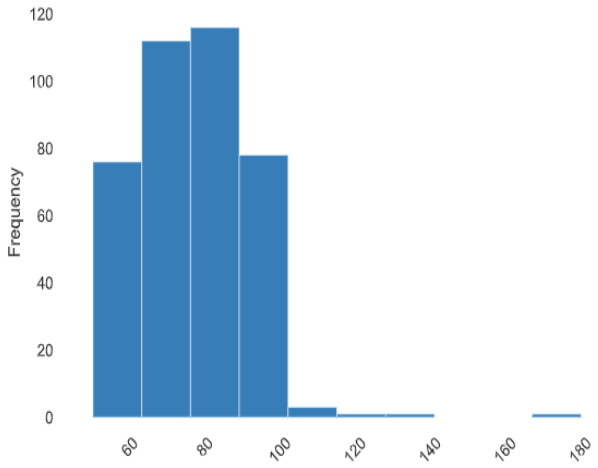
Enfin, les noms des colonnes représentent des abréviations et ne sont visiblement pas pratiques pour une personne qui d'emblée n'était pas du domaine médical.

Par ailleurs, afin de mieux comprendre les données, nous avons décidé de réaliser des graphiques permettant de mieux visualiser les données. Ces graphiques nous ont permis de voir d'une part la distribution de certaines données (majoritairement une distribution gaussienne), et d'autre part de visualiser les différentes valeurs de toutes les variables. Nous avons également remarqué que notre base de données était déséquilibrée étant donné que 100 observations séparent les personnes malades et les personnes non malades.

Age



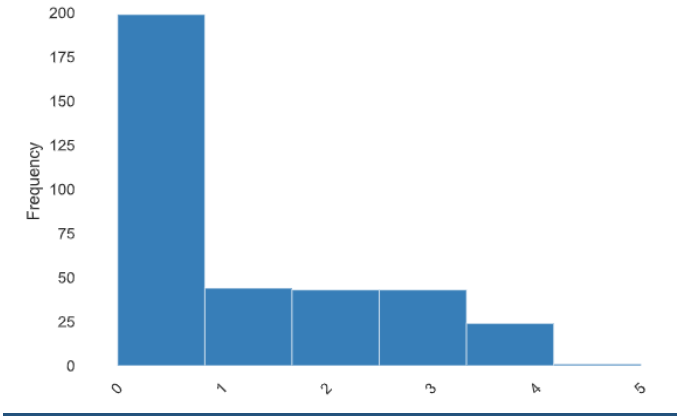
blood_pressure



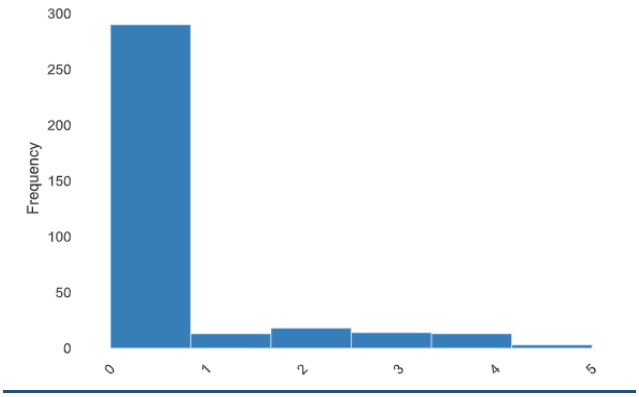
specific_gravity

Value	Count	Frequency (%)
1.02	106	30.0%
1.01	84	23.8%
1.025	81	22.9%
1.015	75	21.2%
1.005	7	2.0%

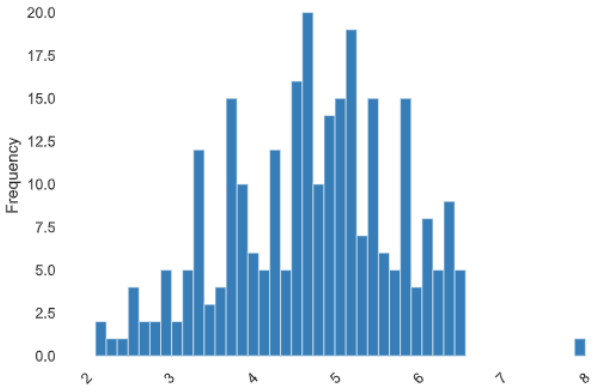
albumin



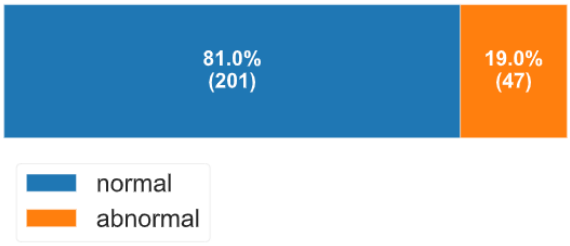
Sugar



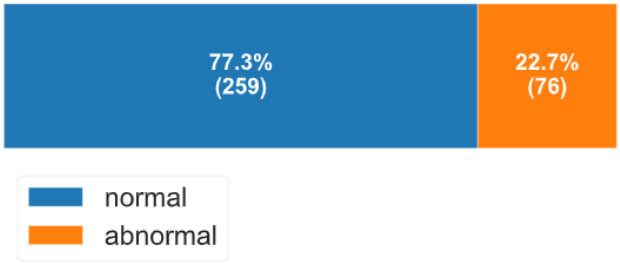
red_blood_cell_count



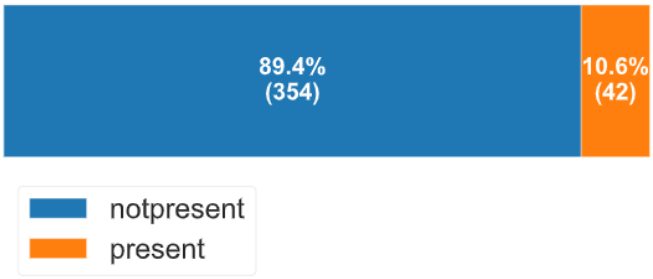
red_blood_cells



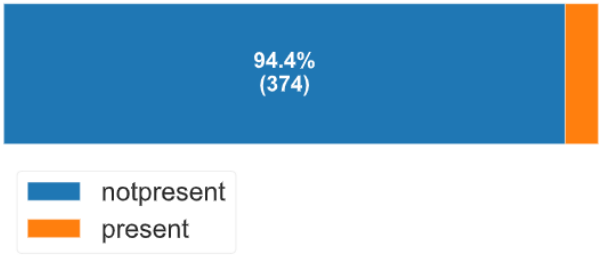
pus_cell



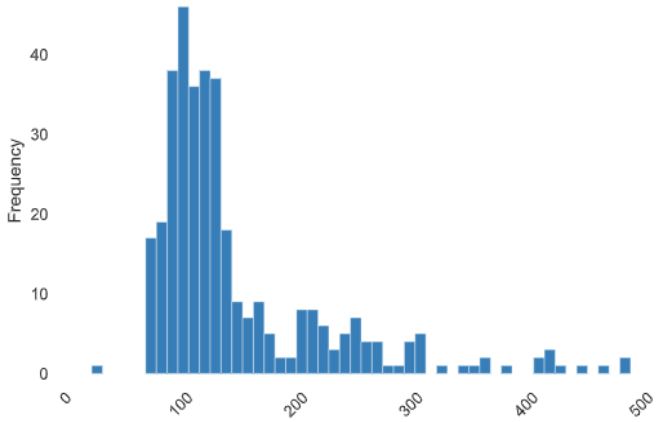
pus_cell_clumps



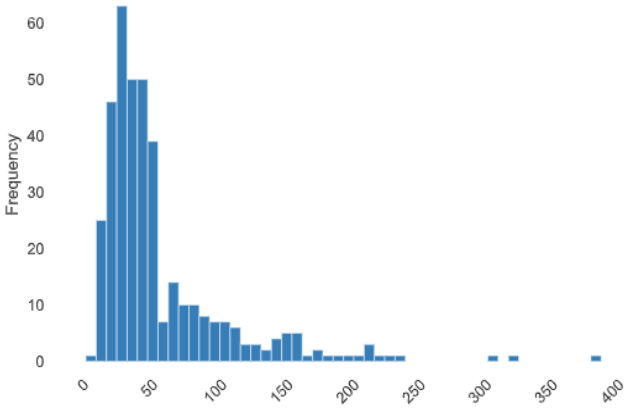
bacteria



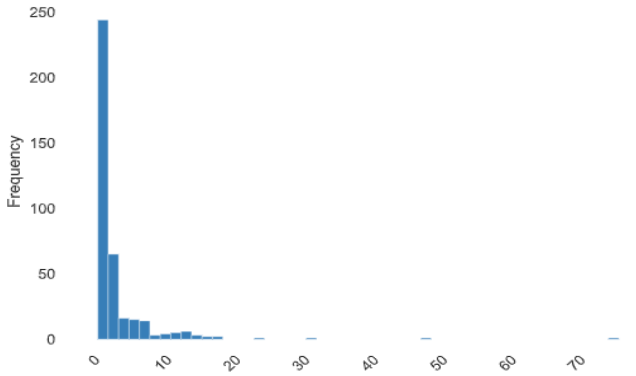
blood_glucose_random



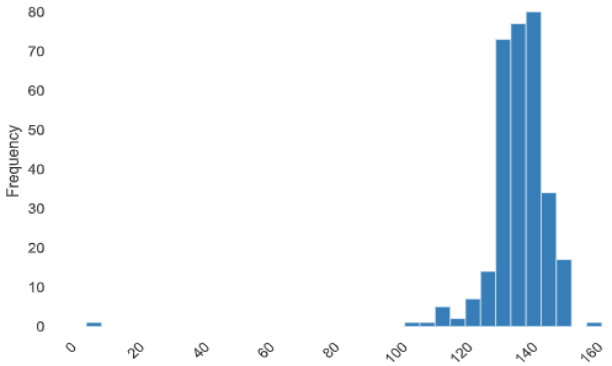
blood_urea



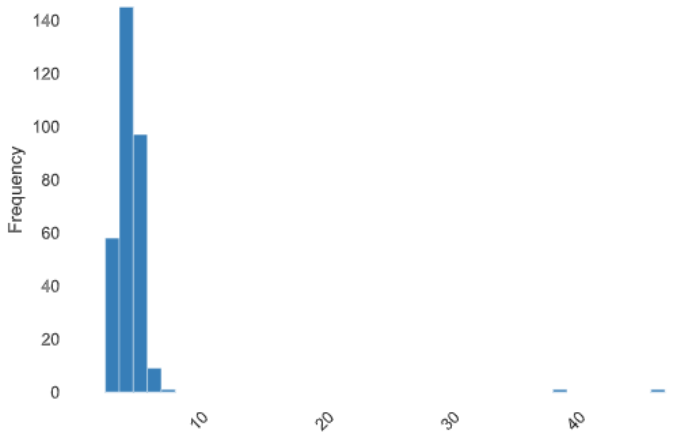
serum_creatinine



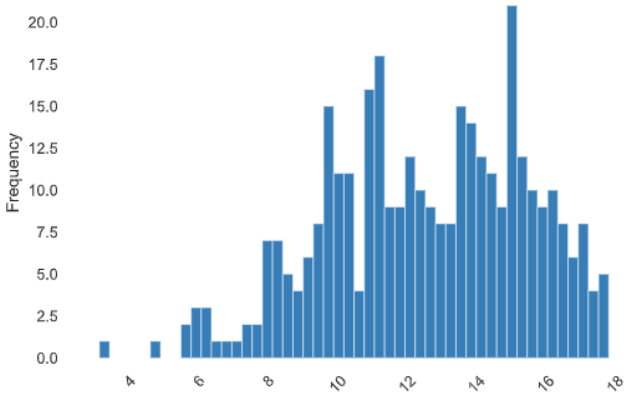
sodium



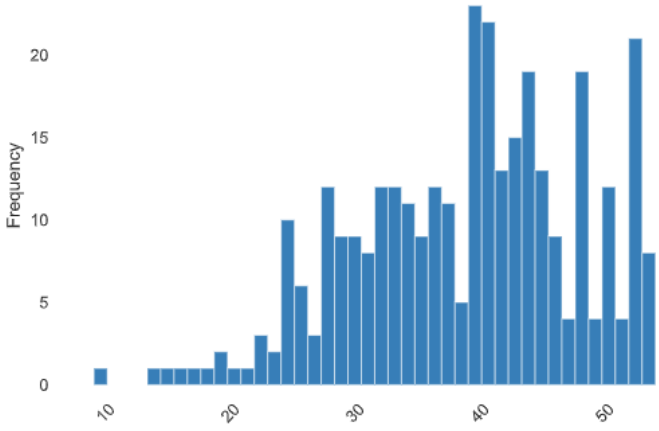
potassium



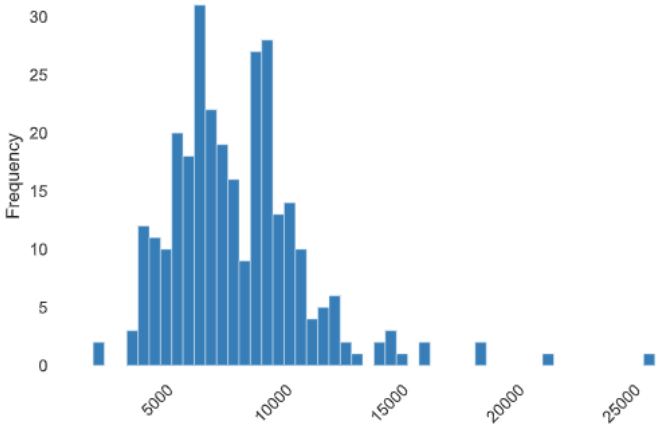
haemoglobin



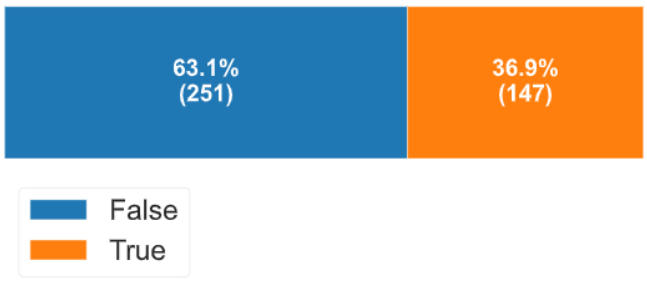
packed_cell_volume



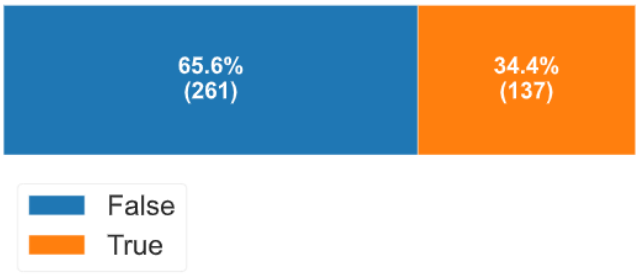
white_blood_cell_count



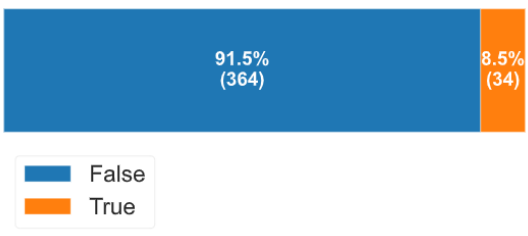
hypertension



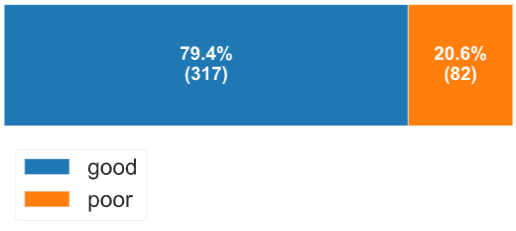
diabetes_mellitus

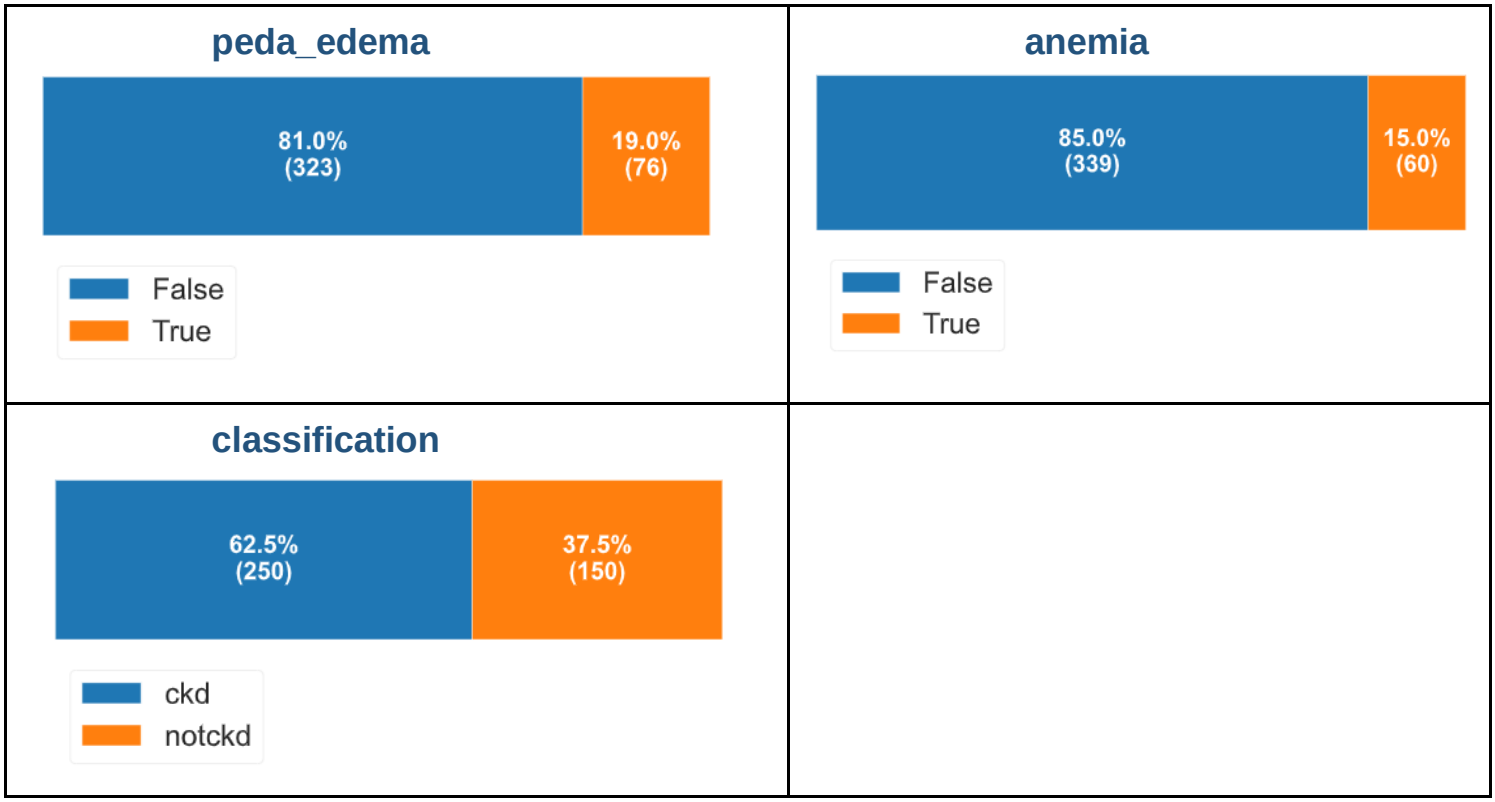


coronary_artery_disease

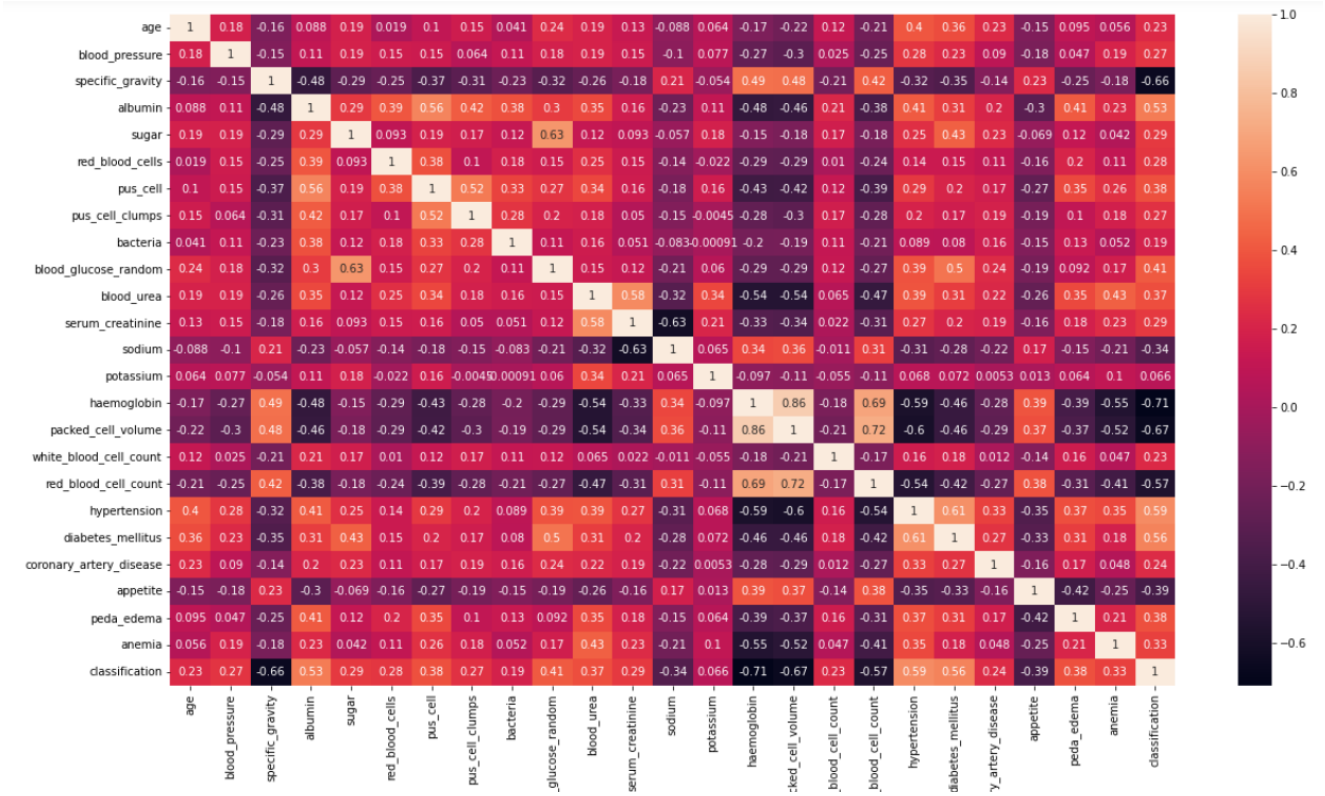


appetite

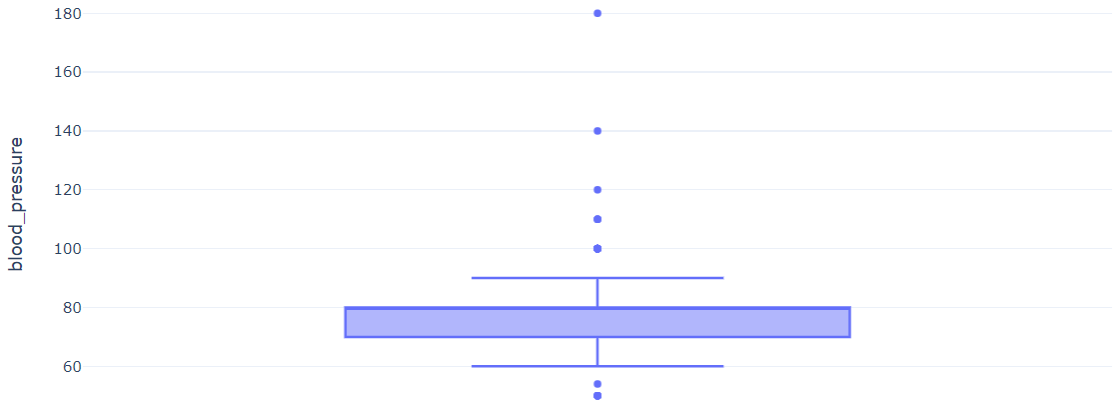




Nous avons également réalisé une matrice de corrélation qui nous a permi de dénicher des corrélations entre certaines features. Par exemple, il existe une forte corrélation entre Haemoglobin et packed_cell_volume. De même, nous avons pu voir que Haemoglobin était la variable la plus fortement corrélée avec notre target “classification”.

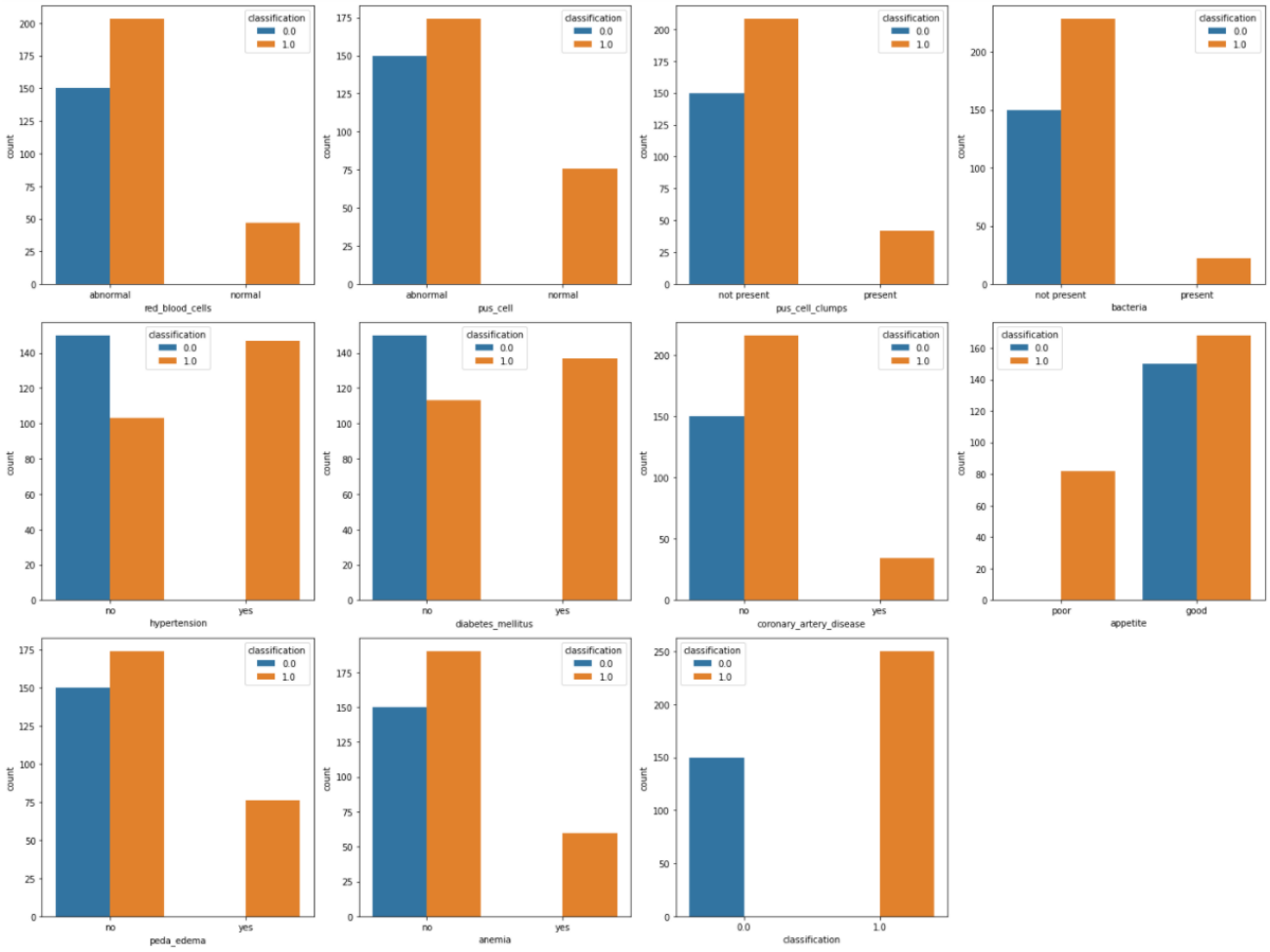


D’autre part, grace aux différents boxplots réalisés, nous avons remarqué l’existence de valeurs aberrantes. Celle-ci pourraient empiéter sur notre travail et biaiser les résultats des différents modèles utilisés. Voici un exemple de boxplot :



Nous pouvons voir que certains points sont extremes. Ceux-ci représentent les valeurs aberrantes.

Enfin, nous avons visualisé la relation entre les différentes variables catégoriques et notre target.



Nous avons remarqué que la plupart des personnes atteintes d'insuffisance rénales souffrent, dans bien des cas, d'hypertension, d'une perte d'appétit, de diabète, d'oedèmes pédaux ou encore d'anémie.

III. La préparation des données

Une fois la phase de compréhension et visualisation des données terminées, notre base de données doit endurer plusieurs transformations dans le but d'enlever toute sorte de bruit. L'objectif est de résoudre les différents problèmes de : types des variables, variables manquantes, variables dupliquées et les valeurs aberrantes.

IV. Modélisation

La quatrième phase est la phase de modélisation. Notre base de données est prete à l'emploi.

Dans notre travail nous avons utilisé plusieurs modèles, tels que KNN, SVM, Random Forest afin de comparer leur efficacité.

V. Evaluation

La cinquième phase est l'évaluation. Il s'agit de l'étape dans laquelle chaque algorithme est testé et comparé aux autres pour trouver la meilleure prédiction possible.
NB: Les phases 3 à 5 sont détaillées dans nos Notebooks.

VI. Déploiement

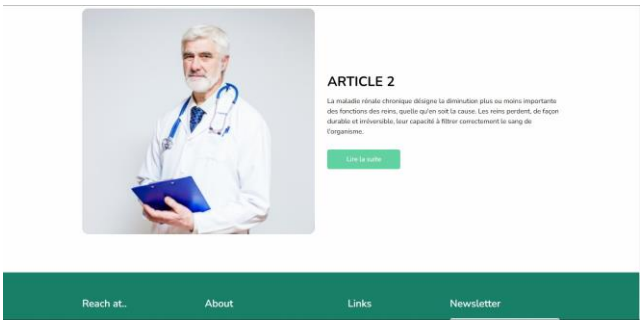
La dernière étape est le déploiement. Le déploiement est défini comme un processus par lequel un modèle d'apprentissage automatique est intégré dans un environnement de production existant pour obtenir des décisions commerciales efficaces basées sur des données.

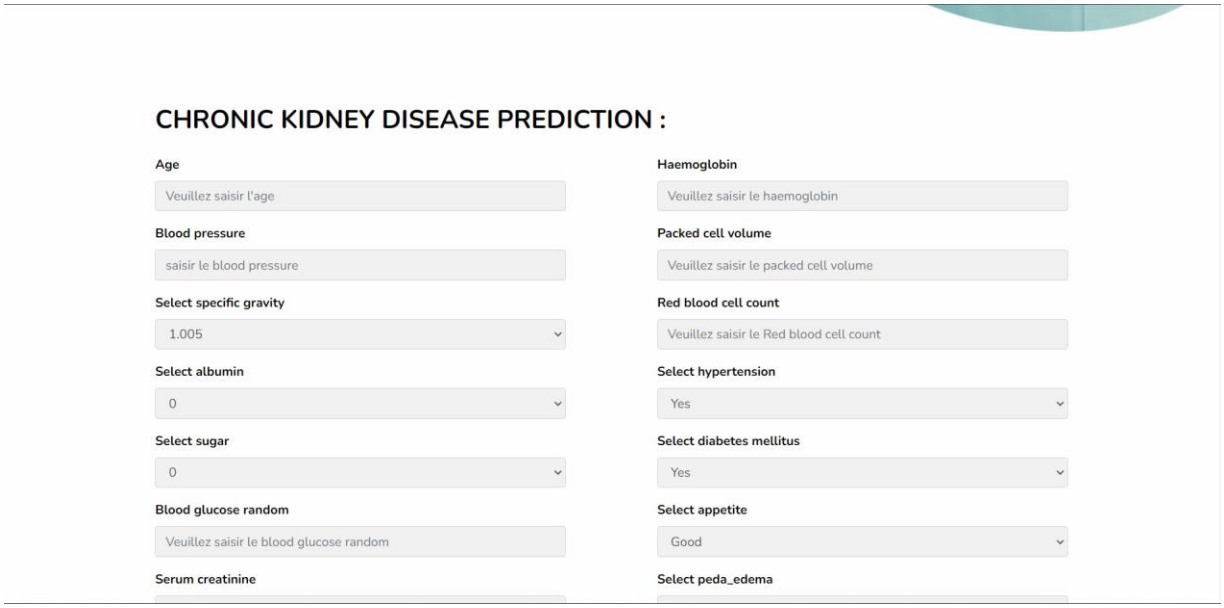
Nous avons donc réalisé une plateforme Web dans lequel une personne peut renseigner les différents champs pour savoir si son bilan d'analyse est celui d'une personne atteinte d'insuffisance rénale chronique ou non.

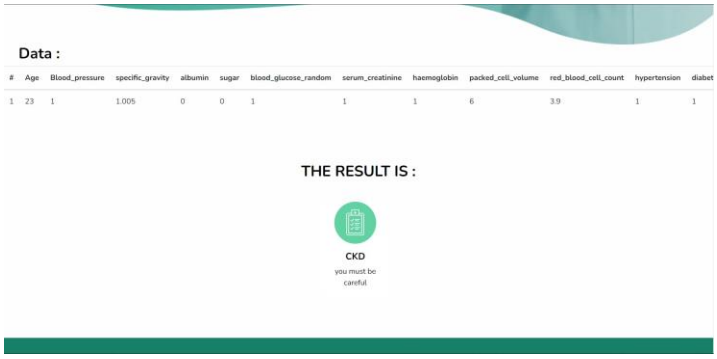
Le code source se trouve sous ce lien github :
https://github.com/yoserwalha/ML_Project_CKD/tree/yoser?fbclid=IwAR3_d-ODgCUtrQBiFGVlvmSiyiByChUW5OG6MVn0-IQmNrKJn6DZ69g4k7Y

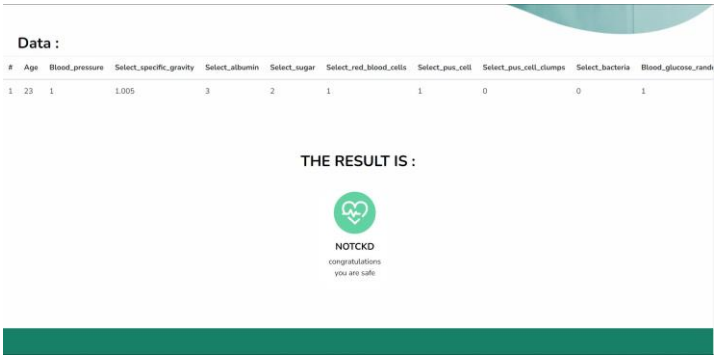
Voici notre plateforme en images :


ARTICLE 1
La maladie rénale chronique est une diminution du fonctionnement des reins qui ne filtrent plus correctement le sang de l'organisme. Cette insuffisance rénale chronique a deux causes principales : le diabète et l'hypertension artérielle.
Lire la suite


ARTICLE 2
La maladie rénale chronique désigne la diminution plus ou moins importante des fonctions des reins, quelle qu'en soit la cause. Les reins perdent, de façon durable et irréversible, leur capacité à filtrer correctement le sang de l'organisme.
Lire la suite
Reach at..
About
Links
Newsletter


CHRONIC KIDNEY DISEASE PREDICTION :
Age
Veuillez saisir l'age
Blood pressure
saisir le blood pressure
Select specific gravity
1.005
Select albumin
0
Select sugar
0
Blood glucose random
Veuillez saisir le blood glucose random
Serum creatinine
Haemoglobin
Veuillez saisir le haemoglobin
Packed cell volume
Veuillez saisir le packed cell volume
Red blood cell count
Veuillez saisir le Red blood cell count
Select hypertension
Yes
Select diabetes mellitus
Yes
Select appetite
Good
Select peda_edema


Data :
Age Blood_pressure specific_gravity albumin sugar blood_glucose_random serum_creatinine haemoglobin packed_cell_volume red_blood_cell_count hypertension diabet
1 23 1 1.005 0 0 1 1 1 6 3.9 1 1
THE RESULT IS :
CKD
you must be careful


Data :
Age Blood_pressure Select_specific_gravity Select_albumin Select_sugar Select_red_blood_cells Select_pus_cell Select_pus_cell_clumps Select_bacteria Blood_glucose_rand
1 23 1 1.005 3 2 1 1 0 0 1
THE RESULT IS :
NOTCKD
congratulations you are safe