REPUBLIQUE DU BENIN

***

MINISTERE DE L'ENSEIGNEMENT SUPERIEURE

ET DE LA RECHERCHE SCIENTIFIQUE

***

UNIVERSITE D'ABOMEY-CALAVI

******

INSTITUT DE FORMATION ET DE RECHERCHE EN INFORMATIQUE

****

PROJET AI4CKD HACKATHON – LICENCE 3 - IA

****

Thème **: Projet AI4CKD - Prédiction des Stades de l'Insuffisance Rénale Chronique (IRC)**

Membres du groupe 10

1. HOUNDAYI Frédy
2. KPELLY Brejnev
3. RADJI Anlim
4. TONON David
5. YEHOUETA Jorias

Sous la supervision de :

Dr (MA) Ratheil HOUNDJI

M. Emeric GBODO

Mme Mélène TONOU

Année Académique **: 2024-2025**

1. Introduction

1.1. Contexte général et motivation du projet

La maladie rénale chronique (IRC) touche environ 10 % de la population mondiale et représente un défi majeur pour la santé publique, notamment en raison des complications graves comme l'insuffisance rénale terminale. Un diagnostic précoce est essentiel pour améliorer les pronostics et réduire les complications. L'Intelligence Artificielle (IA) offre des solutions prometteuses pour analyser des données complexes et prédire les stades de l'IRC, facilitant ainsi une prise en charge plus rapide et ciblée des patients. Ce projet, dans le cadre de l'hackathon AI4CKD de l'IFRI-UAC, vise à développer un modèle prédictif pour identifier les stades de l'IRC à partir de données cliniques et biologiques, servant d'outil d'aide à la décision pour les professionnels de santé.

1.2. Problématique et objectifs spécifiques

La prédiction des stades de l'IRC est complexe en raison de la diversité des facteurs et de la qualité variable des données. L'objectif principal de ce projet est de développer un modèle IA robuste et interprétable pour classifier les stades de l'IRC. Ce modèle doit être fiable et compréhensible pour les cliniciens, tout en contribuant à un système d'aide à la décision pour une meilleure gestion des patients.

2. Revue de littérature

2.1 État de l'art sur la maladie rénale chronique (MRC)

La **maladie rénale chronique (MRC)** est une affection progressive et irréversible des reins, caractérisée par la perte graduelle de leur fonction sur plusieurs mois ou années. Elle représente un problème de santé mondial majeur, affectant des millions de personnes. Les stades de la MRC sont classés selon le taux de filtration glomérulaire (TFG) et sont divisés en cinq étapes, allant du stade 1 (fonction rénale presque normale) au stade 5 (insuffisance rénale terminale).

Les principales causes de la MRC incluent des pathologies comme le **diabète de type 2**, l'**hypertension artérielle (HTA)** et les **maladies cardiovasculaires**. Les facteurs de risque, tels que l'âge avancé, les antécédents familiaux de MRC et certains comportements (fumeur, alimentation déséquilibrée), contribuent également à la progression de la maladie.

Méthodes de diagnostic traditionnelles :

Le diagnostic de la MRC repose principalement sur :

- **La mesure du taux de filtration glomérulaire (TFG)**, calculé à partir des niveaux de créatinine sanguine et d'autres paramètres comme l'âge, le sexe et la race.

- **L'analyse de la créatinine sérique**, un biomarqueur clé pour évaluer la fonction rénale.

- **Les tests de la protéinurie**, qui détectent des niveaux anormaux de protéines dans l'urine, indicateurs de lésions rénales.

- **Les échographies rénales** et autres imageries médicales permettent également de visualiser l'état physique des reins.

Le suivi des patients est souvent effectué par des tests réguliers de ces paramètres, mais le diagnostic précoce des stades avancés de la MRC demeure difficile, d'où l'importance d'améliorer les méthodes de prédiction et de diagnostic.

2.2 Études antérieures sur l'utilisation de l'IA pour la prédiction des stades de l'IRC

L'intelligence artificielle (IA), et plus précisément les **modèles d'apprentissage automatique**, ont récemment fait leurs preuves dans la prédiction des maladies rénales chroniques. Plusieurs études ont démontré l'efficacité des modèles d'IA pour classer les patients en fonction de l'étendue de leur MRC, permettant une détection précoce et une prise en charge personnalisée.

Exemples d'études :

1. **Hassan et al. (2019)** ont utilisé un modèle **Random Forest** pour prédire la progression de la MRC à partir de données cliniques et biologiques. L'étude a montré que l'utilisation des variables cliniques (comme la créatinine sérique et la pression artérielle) améliore significativement la précision du modèle de prédiction.

2. **Ravi et al. (2020)** ont développé un modèle d'apprentissage profond basé sur des réseaux neuronaux pour classifier les stades de l'IRC. Ils ont montré que les modèles basés sur l'IA peuvent surpasser les approches traditionnelles en termes de précision et de capacité de généralisation, surtout lorsqu'ils sont alimentés par de grands ensembles de données patient.

3. **Cheng et al. (2021)** ont exploré l'utilisation des **modèles de machine learning supervisés** pour prédire le stade de la MRC à partir de données démographiques, biométriques et biochimiques. Leur modèle a atteint une précision de 85 % et a

permis de prédire les stades précoces de la MRC avec une grande fiabilité.

Ces études montrent que l'IA a un potentiel considérable pour transformer le diagnostic et la gestion des maladies rénales chroniques, notamment en améliorant l'accès à des diagnostics rapides et précis.

3. Méthodologie

3.1 Description des données

Source des données :

Les données sont issues d'un **dataset médical**, avec des informations cliniques et biologiques relatives à des patients souffrant de maladies rénales chroniques. Ce dataset a été construit en collaboration avec des experts médicaux afin de refléter la diversité des cas d'IRC dans une population donnée. La collecte a eu lieu dans un environnement hospitalier, sous l'approbation des comités éthiques appropriés.

Variables du dataset :

Le dataset comporte un total de **201 colonnes** et **300 lignes**, comprenant des informations cliniques détaillées. Parmi les variables clés présentes, on retrouve des informations démographiques, des résultats biologiques et des symptômes spécifiques aux patients atteints de MRC. Pour faciliter l'analyse, une sélection de variables a été effectuée en fonction de leur pertinence par rapport au diagnostic et à la gestion de la maladie rénale chronique.

Voici une vue d'ensemble des types de variables sélectionnées pour l'analyse :

1. **Identifiants et démographie :** ID, Sexe, Age.
2. **Variables cliniques critiques :** Créatinine, Urée, Protéinurie, Tension artérielle (TA), etc.
3. **Comorbidités :** Hypertension, Diabète, Maladies cardiovasculaires.
4. **Symptômes clés :** Œdème, Oligurie, Asthénie, etc.
5. **Anthropométrie :** Poids, Taille, IMC.

6. **Biologie :** Hb, Na+, K+, Ca2+.

7. **Données échographiques :** Taille des reins, Echogénicité, Contours rénaux.

8. **Stade IRC (si disponible).**

Taille du dataset :

Le dataset contient **300 lignes** représentant 300 patients différents et **201 colonnes**. Les données sont variées, avec des informations numériques (mesures biologiques, tension artérielle, poids, etc.), des informations catégorielles (sexe, antécédents médicaux) et des informations ordinales (stades de la MRC).

3.2 Explication de la sélection des variables pertinentes

La sélection des variables a été réalisée en se basant sur les **recommandations de la National Kidney Foundation (NKF) https://www.kidney.org**, ainsi que sur des connaissances médicales actuelles concernant la gestion et le suivi de la maladie rénale chronique. La **NKF** souligne l'importance de certaines variables cliniques qui entre dans le diagnostic de la maladie par les medecins (notamment:un **test sanguin** connu sous le nom de débit de filtration glomérulaire estimé (DFGe)

- un **test d'urine** connu sous le nom de rapport albumine-créatinine urinaire (uACR))

, biologiques et démographiques dans le diagnostic et le suivi des stades de l'IRC.

Les variables sélectionnées comprennent :

1. **Les facteurs démographiques :** Le sexe et l'âge sont des facteurs importants dans la gestion de l'IRC. Par exemple, l'âge avancé est souvent un facteur de risque majeur dans la progression de la MRC.

2. **Les mesures biologiques critiques :** Des variables telles que la **créatinine**, l'**urée**, la **protéinurie**, et la **tension artérielle** sont cruciales pour évaluer la fonction rénale et détecter d'éventuelles anomalies dans le fonctionnement des reins.

3. **Les comorbidités :** Le diabète, l'hypertension artérielle et les maladies cardiovasculaires sont fréquemment associées à la MRC et influencent son évolution. Ces informations sont essentielles pour mieux comprendre les causes sous-jacentes

de la maladie et ses stades.

4. **Les symptômes cliniques :** Les symptômes tels que l'**œdème**, l'**oligurie** et l'**asthénie** sont des indicateurs importants de la progression de la MRC. Leur prise en compte permet de mieux suivre l'évolution clinique des patients.

5. **Les mesures anthropométriques :** L'**IMC** (indice de masse corporelle) est utilisé pour évaluer le poids corporel du patient par rapport à sa taille et peut influencer la progression de l'IRC.

6. **Les données échographiques :** Les dimensions des reins et leur échogénicité sont des indicateurs importants de l'état rénal.

A l'issu de cette phase nous obtenons un dataset comprenant les colonnes suivantes: à utiliser pour l'analyse

```
colonnes_selectionnees = [
  # Identifiants et démographie
  'ID',
  'Sexe',
  'Age',

  # Variables cliniques critiques
  'Créatinine (mg/L)',
  'Urée (g/L)',
  'Protéinurie',
  'Protéinurie à la bandellette urinaire (g/24h)',
  'TA (mmHg)/Systole',
  'TA (mmHg)/Diastole',

  # Comorbidités
  'Personnels Médicaux/HTA',
```

'Personnels Médicaux/Diabète 1',

'Personnels Médicaux/Diabète 2',

'Personnels Médicaux/IRC',

'Personnels Médicaux/Maladies Cardiovasculaire(Cardiopathie, AVC, preeclampsie)',

Symptômes clés

"Motif(s) d'Admission/Œdème",

'Symptômes/Oligurie',

'Symptômes/Asthénie',

'Symptômes/Nausées',

'Symptômes/Vomissements',

'Symptômes/Perte de poids',

'Etat Général (EG)/OMI',

Anthropométrie

'Poids (Kg)',

'Taille (m)',

'IMC',

Biologie

'Hb (g/dL)',

'Hte (%)',

'Na^+ (meq/L)',

'K^+ (meq/L)',

'Ca^2+ (meq/L)',

```
  # Données échographiques
  'Grosseur Rein Gauche ',
  'Grosseur Rein Droit ',
  'Echogénicité',
  'Contour régulier/Rein droit',
  'Contour régulier/Rein gauche',

  # Stade IRC
  "Stage de l'IRC"
]
```

3.3 Prétraitement des données

Gestion des valeurs manquantes :

- **Variables catégorielles** : Pour les variables catégorielles (telles que 'Sexe', 'Protéinurie'), les valeurs manquantes ont été imputées par la **mode** (valeur la plus fréquente) de chaque colonne. Cela permet d'assurer que les données manquantes dans ces colonnes sont remplacées par une valeur représentative.

- **Variables ordinales** : Pour les variables ordinales comme 'Âge' et 'Stage de l'IRC', les valeurs manquantes ont été imputées par la **médiane** de la colonne, car cette méthode est plus robuste aux valeurs extrêmes que la moyenne.

- **Variables numériques continues** : Pour les variables numériques continues, telles que 'Créatinine (mg/L)', 'TA systolique', 'TA diastolique', les valeurs manquantes ont été imputées par la **moyenne** des valeurs existantes dans chaque colonne.

Suppression des colonnes non représentatives :

- Certaines colonnes ont été supprimées en raison de la présence de trop nombreuses valeurs manquantes ou d'un manque de données représentatives. Cela comprend des variables comme 'Poids (Kg)', 'Taille (m)', 'Protéinurie à la bandellette urinaire (g/24h)', 'Hte (%)', et d'autres qui ont montré un nombre trop faible de valeurs non-nulles pour être imputées de manière fiable.

Traitement des valeurs aberrantes (Outliers)

Les valeurs aberrantes peuvent fausser les résultats des modèles d'apprentissage automatique. Pour cela, une méthode de **cap** (limitation des valeurs extrêmes) a été utilisée pour certaines variables clés :

- **Créatinine (mg/L)** : Des valeurs extrêmes ont été limitées en fonction des normes médicales.

- **TA systolique et diastolique** : Les valeurs de la tension artérielle ont été régularisées pour correspondre aux seuils médicaux attendus.

Cette approche garantit que les valeurs extrêmes ne perturbent pas l'apprentissage du modèle tout en préservant les informations utiles des autres observations.

4.2. Corrélations

Corrélations avec la variable cible : stade de l'IRC

L'analyse des corrélations entre les variables explicatives et le **stade de l'IRC** a révélé plusieurs relations significatives :

- **Créatinine (mg/l)** : présente une forte corrélation positive ($r \approx +0.7$) avec le stade de la maladie. Cela confirme son rôle de biomarqueur clinique central dans l'évaluation de la fonction rénale.

- **DFGe** : montre une corrélation négative forte ($r \approx -0.7$). Ce lien inverse est attendu, car le DFGe diminue lorsque la maladie progresse.

- **Électrolytes (Na^+ et Ca^{2+})** : affichent des corrélations modérées et négatives ($r \approx -0.3$), probablement dues aux déséquilibres métaboliques observés dans les stades avancés.

- **Pression artérielle** : une corrélation faible mais positive ($r \approx +0.2$) est observée, ce qui soutient l'hypothèse que l'hypertension peut être à la fois une cause et une conséquence de l'IRC.

- **Âge** : la corrélation est faible et négative ($r \approx -0.1$), suggérant que, bien qu'il soit un facteur de risque, l'âge seul ne permet pas de prédire efficacement le stade de la maladie dans ce dataset.

Corrélations entre variables explicatives (multicolinéarité)

L'étude des corrélations entre les variables indépendantes a mis en évidence quelques cas de **multicolinéarité** :

- Une corrélation négative forte entre la créatinine et le DFGe ($r \approx -0.7$), ce qui est logique puisque le DFGe est dérivé à partir de la créatinine.

- Une corrélation positive importante ($r \approx +0.7$) entre la **pression systolique** et la **pression diastolique**, deux mesures naturellement liées.

- Les autres variables (âge, électrolytes, etc.) présentent des corrélations faibles ($< \pm 0.4$), suggérant une bonne indépendance statistique.

Dans le cadre de modèles sensibles à la redondance, comme la régression linéaire, il faudra veiller à **ne pas inclure simultanément** des variables trop corrélées (ex. : créatinine et DFGe), sous peine de fausser l'interprétation des coefficients.

4.4. Interprétations

Cette phase exploratoire a permis d'identifier **les variables les plus pertinentes** pour la prédiction du stade de l'IRC. La **créatinine**, le **DFGe**, ainsi que certains **biomarqueurs électrolytiques** et la **pression artérielle** semblent être les meilleurs indicateurs du stade de la maladie. De plus, l'absence de multicolinéarité excessive (hors quelques exceptions attendues) constitue un bon signe pour la suite du processus de modélisation. Enfin, le **déséquilibre** observé dans certaines variables catégorielles notamment dans la variable cible appelle à la prudence dans le choix des algorithmes et à l'éventuelle utilisation de techniques de rééquilibrage.

- **Choix des modèles d'apprentissage automatique**

5. Choix des modèles d'apprentissage automatique

Pour prédire les **stades de l'insuffisance rénale chronique (IRC)** à partir des données cliniques, plusieurs modèles d'apprentissage supervisé ont été envisagés, chacun avec ses avantages en termes de performance, d'interprétabilité et de robustesse face aux spécificités du dataset.

5.1. Justification des modèles sélectionnés

Trois familles de modèles ont été sélectionnées :

Régression logistique multinomiale

Ce modèle est un **classique des problèmes de classification multiclasse**. Il a été choisi pour son **interprétabilité**, notamment en lien avec l'étude des **coefficients** qui permet de quantifier l'impact de chaque variable sur les prédictions. Bien que moins performant sur des relations non linéaires, il sert de **baseline solide**.

Random Forest

En tant que **modèle d'ensemble basé sur des arbres de décision**, Random Forest est capable de gérer la non-linéarité, les interactions complexes entre les variables, et il est robuste au bruit. Il offre également une mesure utile de **l'importance des variables**. C'est un excellent choix pour ce type de problème médical, car il tolère bien les données déséquilibrées et hétérogènes.

🔹 XGBoost

Ce modèle de **gradient boosting** est souvent supérieur en termes de **précision**, notamment sur des jeux de données de taille modérée comme le nôtre (~300 lignes). Il permet un **contrôle fin des erreurs**, une **régularisation intégrée** (via les paramètres lambda et alpha) et une **grande capacité de personnalisation**. XGBoost est particulièrement bien adapté à la compétition et à l'optimisation des performances, tout en permettant une gestion efficace des données déséquilibrées grâce à des pondérations personnalisées.

5.2. Critères de sélection des hyperparamètres

Pour chacun des modèles, les **hyperparamètres ont été sélectionnés** en combinant :

- Une **recherche par grille (GridSearchCV)** ou par validation aléatoire (RandomizedSearchCV) pour explorer les combinaisons possibles.

- Une **validation croisée stratifiée à k-folds (k=5 ou k=10)** pour évaluer la stabilité des performances sur des sous-échantillons équilibrés selon les classes.

- L'**évaluation basée sur plusieurs métriques** (accuracy, recall, F1-score macro) pour s'assurer que le modèle ne favorise pas uniquement les classes majoritaires.

- Pour XGBoost, des **paramètres de régularisation** (max_depth, learning_rate, n_estimators) ont été ajustés pour éviter le sur-

apprentissage, avec scale_pos_weight utilisé dans certains cas pour compenser le déséquilibre des classes.

-

4. Développement du modèle

Entraînement du modèle

L'entraînement a été réalisé à l'aide d'un pipeline combinant plusieurs étapes :

- **Normalisation** des variables numériques à l'aide de StandardScaler.
- **Rééquilibrage des classes** via la technique SMOTE (Synthetic Minority Over-sampling Technique), afin de pallier la distribution inégale des stades de l'IRC dans le jeu de données.
- **Modélisation avec XGBoost**, un classificateur de gradient boosting puissant, adapté aux tâches multi-classes, avec l'objectif "multi:softmax".

Pour optimiser la performance du modèle, une recherche d'hyperparamètres a été menée à l'aide de **GridSearchCV** avec validation croisée à 5 plis (cv=5). Les hyperparamètres explorés incluaient :

- n_estimators : [100, 150]
- max_depth : [3, 5, 7]
- learning_rate : [0.05, 0.1]

✓ **Meilleurs paramètres obtenus** : n_estimators=150, max_depth=5, learning_rate=0.1.

Méthodes utilisées pour éviter le sur-apprentissage

Plusieurs approches ont été utilisées pour éviter l'overfitting :

- **Validation croisée à 5 plis** : permet de valider la robustesse du modèle.
- **Régularisation naturelle de XGBoost** via les hyperparamètres comme *max_depth* ou *learning_rate*.
- **Rééquilibrage SMOTE** : limite l'effet des classes majoritaires et favorise une généralisation plus stable.

Évaluation des performances

Présentation des métriques

Le modèle a été évalué à l'aide de plusieurs métriques :

- **Précision**
- **Rappel**
- **F1-score**
- **Accuracy globale**
- **Matrice de confusion**

Résultats du modèle final

➢ *Sur le jeu d'entraînement :*

- Accuracy : **100 %** (indiquant un risque d'overfitting sur ce set)
- Toutes les classes ont obtenu un **F1-score de 1.00**

➢ *Sur le jeu de test (20 % des données) :*

- Accuracy globale : **81 %**

- Scores par classe :
    - Stade 0 : F1-score = **0.84**
    - Stade 1 : F1-score = **0.67**
    - Stade 2 : F1-score = **0.82**
    - Stade 3 : F1-score = **0.76**
    - Stade 4 : F1-score = **0.91**

Matrice de confusion :

css

CopierModifier

```
[[ 8  0  1  0  0]
 [ 2  6  2  1  0]
 [ 0  1 18  1  0]
 [ 0  0  2  8  0]
 [ 0  0  1  1 10]]
```

Ces résultats montrent que le modèle est particulièrement performant pour les stades 2, 3 et 4, mais qu'il reste encore des confusions entre les stades 0 et 1.

Comparaison des modèles

Un **RandomForestClassifier** a également été testé. Bien qu'il ait fourni des informations intéressantes sur l'importance des variables, ses performances étaient légèrement inférieures à celles de XGBoost. De plus, XGBoost a mieux tiré parti de la structure multi-classes et du rééquilibrage SMOTE.

✅ Le modèle XGBoost optimisé a donc été sélectionné comme **modèle final**.

Interprétabilité du modèle

Importance des variables

Plusieurs méthodes ont été utilisées pour évaluer l'interprétabilité :

- **Feature importance de XGBoost** : les variables les plus importantes comprenaient *dfge*, *creatinine*, *echogenicite*, *na*, *ca* et *grosseur_rein_droit*.
- **Sélection par ANOVA (SelectKBest)** : a mis en évidence des variables cliniques critiques.
- **LassoCV** : a aidé à identifier les variables ayant un poids significatif tout en tenant compte de la régularisation.

Des visualisations sous forme de **barres horizontales** ont été générées pour faciliter l'analyse.

---

5. Résultats et discussion

Analyse des résultats

Le modèle XGBoost final atteint **81 % de précision globale** sur l'ensemble de test, avec des performances équilibrées sur les différents stades de l'IRC, ce qui est satisfaisant dans un contexte médical sensible où la détection précoce est critique.

Le **stage 4** est prédit avec une grande fiabilité (F1 = 0.91), ce qui est crucial car les décisions thérapeutiques deviennent plus urgentes à ce stade. Le **stage 1** reste le plus difficile à prédire, probablement en raison de la **ressemblance des symptômes** avec le stage 0 et du **petit nombre d'observations**.

Limites

- Le **petit effectif** pour certaines classes peut affecter la robustesse du modèle.

- Le **sur-apprentissage** constaté sur le jeu d'entraînement (accuracy = 100 %) suggère que des stratégies supplémentaires comme la validation croisée répétée ou une pénalisation plus forte pourraient être envisagées.

---

6. Annexe.

https://ca-ra.org/fr/tableau-du-dfg-des-reins-par-%c3%a2ge-et-stades-de-lirc/

https://www.omnicalculator.com/fr/sante/calculateur-dfg-debit-filtration-glomerulaire-estime#quest-ce-que-le-dfge

https://pmc.ncbi.nlm.nih.gov/articles/PMC9874070/#s0165

https://www.kidney.org/kidney-topics/stages-chronic-kidney-disease-ckd

https://www.chatgpt.com

https://chat.deepseek.com

https://claude.ai/