



République du Bénin

Ministère de l'Enseignement Supérieur et de la Recherche
Scientifique

Université d'Abomey-Calavi

Institut de Formation et de Recherche en Informatique

Filière : Intelligence Artificielle et Applications

Hackaton_AI4CKD - Prédiction des Stades de l'IRC

Présenté par :

LEKE Bryan
ZOHOUN Ange-Marie
SALAMI Abiola
TINMITONDE Pernel
KODJEKOU Mickaël

Encadreur:

Mélène TONOU

ANNEE ACADEMIQUE 2024 -2025

Rapport

1. Introduction

L'insuffisance rénale chronique (IRC), également appelée maladie rénale chronique (MRC), constitue un problème majeur de santé publique à l'échelle mondiale. Elle se caractérise par une diminution progressive et irréversible de la fonction rénale, qui peut à terme conduire à une insuffisance rénale terminale nécessitant une dialyse ou une transplantation. Selon les données de l'Organisation Mondiale de la Santé (OMS), la MRC touche environ 10 % de la population adulte dans le monde, et sa prévalence est en constante augmentation. La prévalence et l'incidence de l'IRC sont en augmentation, et les projections indiquent qu'elle pourrait devenir la cinquième cause de mortalité mondiale d'ici 2040. ¹

La détection précoce de la MRC est cruciale pour ralentir sa progression et prévenir les complications associées, notamment les maladies cardiovasculaires. Les techniques d'apprentissage automatique (machine Learning) offrent aujourd'hui des opportunités puissantes pour l'analyse des données médicales complexes, en facilitant la prédiction précoce et personnalisée de la progression de la maladie.

1.1. Objectifs

L'objectif principal de ce projet est de construire un modèle capable de prédire avec précision le stade d'évolution de l'IRC chez un patient, à partir de variables cliniques et biologiques disponibles. Les objectifs secondaires incluent :

- L'identification des variables les plus pertinentes pour la prédiction.
- L'analyse des corrélations entre les variables.
- L'élaboration d'un pipeline de prétraitement adapté aux données médicales.
- Le choix du modèle à utiliser

1.2. Portée du Rapport

Ce rapport se concentre sur la préparation des données, la compréhension clinique du problème, Le choix et l'évaluation du modèle. Il inclut :

- Une synthèse du contexte médical de l'IRC.
- L'analyse exploratoire et la sélection des variables.
- Le traitement des valeurs manquantes et aberrantes.
- L'analyse des corrélations en vue de la modélisation.
- Sélection Définitif des variables

- Comparaison, Choix du modèles et Évaluation

2. Compréhension du Contexte Clinique de l'IRC

2.1. Fonction Rénale et Stades de l'IRC

Les reins jouent un rôle essentiel dans le maintien de l'homéostasie corporelle : filtration des déchets, régulation de l'équilibre hydrique et électrolytique, production d'hormones. Leur fonction est mesurée principalement par le **débit de filtration glomérulaire (DFG)**, qui représente le volume de plasma filtré par les glomérules par minute.

Le **DFG estimé (DFGe)** est calculé à partir de la créatinine sérique, de l'âge, du sexe et de l'origine ethnique via des formules telles que MDRD ou CKD-EPI. Toutefois, la créatinine sérique seule est insuffisante car elle peut être influencée par la masse musculaire.

Les stades de l'IRC sont classés en 5 niveaux selon le DFGe ³ :

- **Stade 1** : DFGe ≥ 90 ml/min (fonction normale mais avec lésion rénale)
- **Stade 2** : DFGe entre 60 et 89 ml/min
- **Stade 3** : DFGe entre 30 et 59 ml/min
- **Stade 4** : DFGe entre 15 et 29 ml/min
- **Stade 5** : DFGe < 15 ml/min (insuffisance rénale terminale)

2.2. Marqueurs Rénaux

La **protéinurie** (présence de protéines dans l'urine), et plus spécifiquement l'**albuminurie**, est un marqueur précoce de lésion rénale. Elle peut précéder la chute du DFG et constitue un critère de stratification du risque.

D'autres marqueurs cliniques pertinents incluent :

- **Urée** : déchet azoté, son accumulation reflète une altération de la filtration.
- **Hémoglobine (Hb)** : souvent diminuée en IRC avancée en raison de l'érythropoïétine déficiente.
- **Ionogramme (Na^+ , K^+ , Ca^{2+})** : les déséquilibres électrolytiques (hyperkaliémie, hyponatrémie, hypocalcémie) sont fréquents.

2.3. Facteurs de Risque

Les principaux facteurs de risque identifiés dans la littérature incluent ¹:

- **Diabète** (diabète de type 1 ou 2)
- **Hypertension artérielle (HTA)**
- **Age avancé**
- **Antécédents familiaux ou cardiovasculaires**
- **Sexe masculin**
- **Problèmes cardiovasculaires et syndrome métabolique**
- **L'Obésité,**
- **Le Tabagisme,**

La présence simultanée de plusieurs facteurs de risque augmente significativement la probabilité de progression vers un stade avancé de l'IRC.

3. Sélection des Variables

3.1. Variables Clés pour la Prédiction de l'IRC

La prédiction du stade d'évolution de l'IRC repose sur un ensemble de variables médicales que l'on peut regrouper en plusieurs catégories :

- **Indicateurs de la fonction rénale** : créatinine (mg/L), urée (g/L), Hb (g/dL), Na^+ , K^+ , Ca^{2+} .
- **Facteurs de risque** : âge, HTA, diabète, sexe.
- **Symptômes cliniques** : asthénie, anémie, œdème (OMI), oligurie.
- **Marqueurs urinaires** : protéinurie, albuminurie.

Ces variables ont été extraites du jeu de données fourni par l'hôpital partenaire, et validées par des publications scientifiques de référence (Velez JCQ, et al. chronic kidney disease. StatPearls [1]. Treasure Island (FL): StatPearls Publishing; 2024 Jan-. Disponible sur [lien](#)).

3.2. Compréhension des Variables du Dataset

Le récapitulatif des colonnes présente dans le dataset et de leur compréhension et pertinence sont résumer dans un tableau dans Notion sur le lien ci-dessous

Lien ➡ : [Tableaux](#)

3.3. Premières Sélection des Variables

Chaque variable conservée dans l'analyse a été retenue pour sa pertinence clinique, et explicative. Le tableau suivant résume cette sélection :

Nom de la colonne	Pertinence	Justification reformulée
Sexe	Moyenne	Peut avoir une influence sur le risque ou la progression de l'IRC. La pertinence reste à valider statistiquement sur ce jeu de données.
Âge	Haute	Facteur de risque établi. Le déclin physiologique de la fonction rénale avec l'âge justifie sa conservation.
Symptômes/Oligurie	Moyenne	Signe clinique fréquent en cas d'insuffisance rénale avancée. Potentiellement utile.
Symptômes/Anémie	Moyenne	Reflète une complication fréquente de l'IRC. Attention au risque de circularité avec l'hémoglobine.
Symptômes/HTA	Moyenne	L'hypertension est un facteur causal et aggravant. À vérifier dans l'analyse exploratoire.
Symptômes/Asthénie	Moyenne	Symptôme général peu spécifique. Sa pertinence dépendra de la corrélation avec le stade.
Symptômes/OMI	Moyenne	Signe de rétention hydrosodée lié à l'IRC avancée. À tester.
EG à l'Admission	Faible	Donnée descriptive non standardisée. Difficulté d'exploitation automatique.
TA (mmHg)/Systole	Moyenne	Composante de la pression artérielle. Un agrégat type PAM (pression artérielle moyenne) serait peut-être plus pertinent.
TA (mmHg)/Diastole	Moyenne	Même remarque que pour la systolique. À envisager dans une variable combinée.
IMC	Moyenne	Potentiel rôle de facteur indirect. Faible impact direct anticipé sur le stade IRC.
BU/Hématurie	Haute	Signe de pathologie glomérulaire. Fortement lié à la fonction rénale.
BU/Protéinurie	Haute	Marqueur précoce de lésion rénale. Recommandé par les guidelines.
BU/Albuminurie	Faible	Taux souvent incomplet ou non exploitable dans ce jeu de données.
Glycémie à jeun	Moyenne	Facteur de risque (diabète). La pertinence dépendra du lien avec la progression.
Urée (g/L)	Haute	Marqueur azoté classique. Suivi dans les bilans biologiques d'IRC.
Créatinine (mg/L)	Haute	Paramètre fondamental pour estimer le DFG. Indicateur clé de la fonction rénale.
Protéinurie	Haute	Marqueur clinique majeur recommandé pour le diagnostic et la stratification du risque.
Protéinurie 24h	Faible	Souvent redondante ou imprécise selon le recueil. À éviter si mauvaise qualité.
Na ⁺ (meq/L)	Faible	Moins discriminant pour le stade IRC dans les phases stables.
K ⁺ (meq/L)	Faible	Peut fluctuer avec l'alimentation et les traitements. Faible pouvoir prédictif direct.

Ca ²⁺ (meq/L)	Faible	Pertinent en cas de dérèglement avancé mais peu prédictif du stade en soi.
Hb (g/dL)	Faible à moyenne	L'anémie est une conséquence plus qu'un prédicteur. Risque de redondance avec Symptômes/Anémie.
Stage de l'IRC	Très haute	Variable cible. Mesure le niveau de sévérité de la maladie.

Les sélections sont fondées sur les recommandations cliniques des directives KDIGO 2012⁷ et sur plusieurs publications scientifiques telles que celles de Coresh et al.¹, Bastos et al.⁸ et Vassalotti et al.⁹.

3.4. Sélection Finale des Variables et justification sur les variables Écarter

a. Sélectionné

Ces variables ont été gardées après traitement et analyse des données et confortées dans la décision par les tests statistiques et corrélation puis par l'analyse d'importance des colonnes dans la prédiction grâce au Random Forest

Nom de la colonne	Justification
Age	Facteur de risque établi, présent dans les formules DFGe
Créatinine (mg/L)	Indicateur direct de la fonction rénale
Urée (g/L)	Complémentaire à la créatinine pour suivre l'accumulation azotée
Sexe	Facteur de risque
Hb (g/dL)	Liée à l'anémie rénale
Na ⁺ (meq/L)	Révèle les déséquilibres électrolytiques
K ⁺ (meq/L)	Hyperkaliémie = signe de sévérité
Ca ²⁺ (meq/L)	Déséquilibre calcique lié à l'IRC
Symptômes/OMI	Indice de rétention hydrique
Symptômes/Asthénie	Symptôme fréquent, mais peu spécifique
TA (mmHg)/Systole	Indicateur de la Tension Artérielle Facteur de risque de l'IRC
Etat Général (EG) à l'Admission	Peut influencer l'évolution de l'IRC

b. Supprimer

Certaines variables ont été écartées pour cause de redondance ou d'information limitée. Par exemple, la variable "Symptômes/Anémie" est redondante avec l'hémoglobine, qui est une mesure plus précise et continue. De même, certaines variables ayant un grand nombre de valeurs manquantes (ex : albuminurie) ont été exclues de l'analyse principale.

Nom de la colonne	Justification de l'exclusion
-------------------	------------------------------

Symptômes/Oligurie	Variable corrélée à l'état général à l'admission. Importance faible observée dans les modèles, notamment avec Random Forest.
Symptômes/Anémie	Risque de redondance avec la variable quantitative hémoglobine (Hb). Problème de circularité.
Symptômes/HTA	Corrélée fortement aux mesures de la tension artérielle (TA systolique et diastolique). Risque de redondance.
Symptômes/OMI	Importance très faible dans les modèles prédictifs, notamment Random Forest.
TA (mmHg)/Diastole	Fortement corrélée avec la TA systolique. Moins informative seule. Écartée pour réduire la redondance.
IMC	Taux élevé de valeurs manquantes (169). Faible pertinence observée dans les analyses préliminaires.
BU/Hématurie	Plus de 225 valeurs manquantes. Utilisation limitée dans la modélisation.
BU/Protéinurie	Taux de valeurs manquantes très important. Difficulté d'exploitation standardisée.
BU/Albuminurie	Présente de nombreuses valeurs manquantes et peu corrélée au stade dans ce jeu de données.
Protéinurie	Trop de valeurs manquantes et recueil peu homogène.
Protéinurie à la bandellette urinaire (g/24h)	Forte proportion de valeurs manquantes. Peu de valeur ajoutée dans les modèles testés.

4- Traitements Univariée des Colonnes

4.1. Traitement des Types de Données

	0
Age	float64
	4
Créatinine (mg/L)	object
Glycémie fi jeun (taux de Glucose)	object
Urée (g/L)	object
Stage de l'IRC	object
Etat Général (EG) fi l'Admission	object
Hb (g/dL)	object
Na ⁺ (meq/L)	object
K ⁺ (meq/L)	object
Ca ²⁺ (meq/L)	object

Pathologies/Rétinopathie diabétique	int64
Sexe	object
Symptômes/Anémie	int64
Symptômes/Asthénie	int64
Symptômes/HTA	int64
Symptômes/OMI	object
Symptômes/Oligurie	int64
TA (mmHg)/Diastole	float64
TA (mmHg)/Systole	float64

Conversion des colonnes au type de données approprié .

Nom de la colonne	Type corrigé	Justification
Age	float64	Valeur continue (âge en années)
Créatinine (mg/L)	float64	Doit être numérique pour analyses biologiques
Glycémie à jeun (taux de Glucose)	float64	Numérique, taux mesuré en g/L
Urée (g/L)	float64	Numérique, mesure sanguine continue
Stage de l'IRC	category	Cible ordinale à encoder selon les stades
Etat Général (EG) à l'Admission	object	Donnée qualitative/observation textuelle
Hb (g/dL)	float64	Numérique, taux d'hémoglobine
Na ⁺ (meq/L)	float64	Numérique, ionogramme
K ⁺ (meq/L)	float64	Numérique, ionogramme
Ca ²⁺ (meq/L)	float64	Numérique, ionogramme
Pathologies/Rétinopathie diabétique	int64	Présence/absence encodée (0/1)
Sexe	category	Catégoriel (masculin/féminin)
Symptômes/Anémie	int64	Binaire (0/1), déjà correct
Symptômes/Asthénie	int64	Binaire (0/1), déjà correct
Symptômes/HTA	int64	Binaire (0/1), déjà correct
Symptômes/OMI	int64	Convertir depuis objet en int binaire
Symptômes/Oligurie	int64	Binaire (0/1), déjà correct
TA (mmHg)/Diastole	float64	Numérique, déjà correct
TA (mmHg)/Systole	float64	Numérique, déjà correct

4.2. Gestion des erreurs de formatage.

Pour les variables numériques les nombre en **float** ont des virgules a la place devrait être des points (ex : **1,5** au lieu de **1.5**), **pour** ceux en **Int** comme l'**âge** son entre griffe (ex '**1**' au lieux de **1**)

```
coma = ['Créatinine (mg/L)', 'Urée (g/L)', 'Hb (g/dL)', 'Na^+ (meq/L)',
        'K^+ (meq/L)', 'Ca^2+ (meq/L)', 'Glycémie à jeun (taux de Glucose)',
        ]

# Points à la place des virgules
for col in coma:
    if col in df.columns:
        df[col] = pd.to_numeric(df[col].astype(str).str.replace(',', '.'), errors='coerce')

df.head()
```

Application de la fonction **convert_to_numeric** pour toutes les colonnes bien formatées mais ou la colonne est toujours détecter comme objets

```
def convert_to_numeric(value):
    """Converts a single value to numeric. Used with apply() on a column"""
    try:
        return pd.to_numeric(value)
    except ValueError:
        return np.nan
```

Ex :

```
df['Age'] = df['Age'].apply(convert_to_numeric)
```

Conversion d'Objet à Float pour la colonne Âge détecter comme Objets

4.3. Gestion des Valeurs Manquantes

Nombre de valeurs manquantes par colonne:

Age	7	
BU/Albuminurie	225	
BU/Protéinurie	225	
Créatinine (mg/L)	0	
Durée Diabète 1 (mois)	304	
Durée Diabète 2 (mois)	222	
Durée HTA (mois)	108	
Durée IRC (mois)	288	
Glycémie à jeun (taux de Glucose)	87	
Protéinurie	245	
Urée (g/L)	23	
Stage de l'IRC	0	
Etat Général (EG) à l'Admission	5	
BU/Glucosurie	225	
BU/Hématurie	225	
CCMH (%)	166	
Hb (g/dL)	50	
Na^+ (meq/L)	35	

K ⁺ (meq/L)	30
Ca ²⁺ (meq/L)	91
Hématie (T/L)	254
IMC	169
Pathologies/Rétinopathie diabétique	0
Protéinurie à la bandellette urinaire (g/24h)	241
Sexe	8
Symptômes/Anémie	0
Symptômes/Asthénie	0
Symptômes/HTA	0
Symptômes/OMI	0
Symptômes/Oligurie	0
TA (mmHg)/Diastole	61
TA (mmHg)/Systole	61

Justification de la stratégie d'imputation (ou de suppression) choisie pour chaque colonne.

Suppression des colonnes avec plus de 50% de valeur manquantes :

Cette technique a été appliquée aux colonnes suivantes : Durée Diabète 1 (mois), Durée Diabète 2 (mois), Durée HTA (mois), Durée IRC (mois), BU/Protéinurie, Protéinurie, BU/Glucosurie, BU/Albuminurie, BU/Glucosurie, BU/Hématurie, Hématie (T/L), IMC, Protéinurie à la bandellette urinaire (g/24h)

```
# Liste des colonnes à supprimer
colonnes_a_supprimer = ['Durée Diabète 1 (mois)', 'Durée Diabète 2 (mois)', 'Durée HTA (mois)', 'Durée IRC (mois)',
                        'BU/Glucosurie', 'BU/Hématurie', 'BU/Protéinurie', 'Protéinurie', 'IMC', 'Protéinurie à la bandellette urinaire (g/24h)',
                        'CCMH (%)', 'Hématie (T/L)', 'BU/Albuminurie']

# Suppression des colonnes avec plus de 50% de valeur manquantes
df = df.drop(columns=colonnes_a_supprimer)
```

Imputation par la Médiane pour les colonnes avec distribution asymétrique et présence de outliers :

Cette technique a été appliquée aux colonnes suivantes : Glycémie à jeun (taux de Glucose), Urée (g/L), Hb (g/dL), Na⁺ (meq/L), K⁺ (meq/L), Ca²⁺ (meq/L)
Colonne Sexe : Étant données les valeurs manquantes pas énormes et une distribution des classes symétrique (équivalent) on fait une imputation aléatoire de ^plus le nombre de valeur manquante est faible.

```
# Etant données la distribution symétrique catégoriels binaires sans outlier on fera une imputation aléatoires des valeurs
def impute_random_binary(series):
    probs = series.value_counts(normalize=True) # Proba de chaque catégorie
    return series.fillna(np.random.choice(probs.index, p=probs.values))

df['Sexe'] = impute_random_binary(df['Sexe'])

# Vérification
print(df['Sexe'].isnull().sum())
```

Colonne Etat Générale d'admission : Une imputation par la catégorie la plus fréquente après avoir vérifié la distribution de la colonne

```
# Etant donné cette variable catégorielle non binaire avec une distribution asymétrique on va imputer par la catégorie la plus fréquente
df["Etat Général (EG) à l'Admission"].fillna(df["Etat Général (EG) à l'Admission"].mode()[0], inplace=True)

# vérification valeur null
print(df["Etat Général (EG) à l'Admission"].isnull().sum())
```

4.4. Traitement des Valeurs Incohérentes ou Aberrantes

Voici les **méthodes utilisées** pour gérer les **valeurs incohérentes et aberrantes** dans les colonnes **numériques** pendant l'analyse univariée :

But : Éliminer ou corriger les âges absurdes ou manquants (comme les âges négatifs ou hors normes) via l'imputation robuste.

Détection des valeurs aberrantes (outliers) avec la méthode IQR

Utiliser une fonction générique basée sur l'IQR (écart interquartile) pour détecter les outliers dans plusieurs colonnes :

```
def detect_outliers_iqr(df, column, threshold=1.5):
    """
    Détecte les valeurs aberrantes dans une colonne d'un DataFrame en utilisant l'IQR.

    Paramètres :
    - df : DataFrame Pandas
    - column : Nom de la colonne à analyser
    - threshold : Facteur d'écart interquartile (1.5 par défaut)

    Retourne :
    - DataFrame contenant uniquement les valeurs aberrantes
    """
    Q1 = df[column].quantile(0.25)
    Q3 = df[column].quantile(0.75)
    IQR = Q3 - Q1

    lower_bound = Q1 - threshold * IQR
    upper_bound = Q3 + threshold * IQR

    outliers = df[(df[column] < lower_bound) | (df[column] > upper_bound)]

    print(f"Nombre de valeurs aberrantes détectées dans '{column}' : {outliers.shape[0]}")
    return outliers
```

Colonnes utilisées avec detect_outliers_iqr :

- Créatinine (mg/L)
- Urée (g/L)
- TA (mmHg)/Systole
- TA (mmHg)/Diastole
- Hb (g/dL)
- K⁺ (meq/L)
- Na⁺ (meq/L)
- Glycémie à jeun (taux de Glucose)

Imputation robuste (sans outliers)

Nous avons utilisé une stratégie d'imputation robuste (exclusion des outliers et les valeurs manquante avant d'imputer avec la régression linéaires) :

Appliqué à :

- TA (mmHg)/Systole
- TA (mmHg)/Diastole

Parce que les deux variables sont fortement corrélées et corrélé avec d'autre variables

Créatinine (mg/L)

L'analyse visuelle via un **boxplot** et la détection via l'IQR ont initialement identifié plusieurs valeurs comme aberrantes (>30 mg/L). Toutefois, une vérification croisée avec la variable cible (*Stage de l'IRC*) a révélé que ces valeurs élevées étaient **caractéristiques des patients en stades avancés (CKD 4 et CKD 5)**. Ces valeurs, bien que extrêmes, sont **cohérentes sur le plan clinique et ont donc été conservées**.

variable d'incohérence avec 0.00 de créatinine mg/L pour un stade 3 0.25 de créatinine mg/L pour un stade 5

```
[ ] df.drop(df[(df['Créatinine (mg/L)'] == 0.00) & (df["Stage de l'IRC"] == 3)].index, inplace=True)
      df.drop(df[(df['Créatinine (mg/L)'] == 0.245) & (df["Stage de l'IRC"] == 5)].index, inplace=True)
```

Exception : des erreurs de saisie flagrantes ont été identifiées, comme une valeur à **500 mg/L**, qui ne correspond à aucune valeur physiologique réaliste. Nous avons également rencontré des erreurs de saisie. Ces cas ont été **supprimés**.

```
# df.loc[df['Créatinine (mg/L)'] == 500.00, 'Créatinine (mg/L)'] = 50.00
df.drop(df[(df['Créatinine (mg/L)'] == 500.00) & (df["Stage de l'IRC"] == 5)].index, inplace=True)
```

Urée (g/L)

Même observation que pour la créatinine. Des valeurs considérées comme extrêmes statistiquement (>1 g/L) sont en réalité fréquentes chez les patients IRC sévères. Après vérification de leur cohérence clinique, elles ont été **retenues** dans le jeu de données, à l'exception de certaines valeurs suspectes signalées et supprimer manuellement.

K^+ et Ca^{2+} et les autre colonne (ions sanguins)

Les valeurs extrêmes détectées via l'IQR ont été **réinterprétées** selon la distribution des stades IRC. Certains patients en CKD 5 présentent des déséquilibres électrolytiques importants. Par conséquent, les outliers physio pathologiquement **justifiables ont été conservés**. Exemples

```
# Remplacer les valeur considérée comme mal saisie dans la colonne 'K^+ (meq/L)'
df.loc[df['K^+ (meq/L)'] == 41.00, 'K^+ (meq/L)'] = 4.10
df.loc[df['K^+ (meq/L)'] == 43.00, 'K^+ (meq/L)'] = 4.30
df.loc[df['K^+ (meq/L)'] == 36.00, 'K^+ (meq/L)'] = 3.60

#supprimer colonne avec 9.17 pour K+
df.drop(df[df['K^+ (meq/L)'] == 9.17].index, inplace=True)
```

D'autres, jugés incohérents ont été **supprimés**.

Hb (g/dL) – Hémoglobine

L'hémoglobine est un marqueur essentiel dans le suivi des patients atteints de maladie rénale chronique (MRC). Une anémie est fréquente dans les stades avancés de la maladie, et donc des valeurs faibles peuvent être attendues.

L'analyse de la distribution de la variable a été réalisée à l'aide d'un **boxplot**, suivie d'une détection des valeurs aberrantes par la méthode IQR :

Les valeurs faibles d'hémoglobine détectées comme "outliers" sont en réalité **cohérentes sur le plan clinique**, notamment pour les patients classés en **CKD 4 et 5**, où l'anémie est un symptôme connu. Ces valeurs ont donc été **conservées**.

Des vérifications manuelles ont également été menées pour détecter :

- Des erreurs de saisie (par exemple : unités incohérentes, valeurs >30 g/dL)
- Des valeurs **nulles ou manquantes**, traitées par **imputation robuste à la médiane**

À partir de l'analyse du notebook, voici ce que je peux confirmer sur **l'encodage des variables catégorielles** :

5- Encodage des Variables : Détail par Variable

Sexe

Cette variable est typiquement binaire. Elle a été encodée de manière simple :

- 0 : Femme
- 1 : Homme

Autres variables catégorielles (qualitatives)

Le traitement s'est fait automatiquement par une boucle :

Cela signifie que toutes les variables de type **object** (c'est-à-dire les chaînes de caractères) ont été **encodées en entiers** de manière automatique.

Pour Tout les Symptômes/HTA, OMI etc., on aura :

0 : "absent"

1 : "présent"

Stage de l'IRC (variable cible)

Elle aussi a été encodée avec LabelEncoder, ce qui signifie que :

- 0, 1, 2, ... correspondent aux stades IRC 1 à 5

```
df['Stage de l'IRC'] = df['Stage de l'IRC'].replace({
    'CKD 1': 1,
    'CKD 2': 2,
    'CKD 3a': 3,
    'CKD 3b': 3,
    'CKD 4': 4,
    'CKD 5': 5
})
df['Stage de l'IRC'] = pd.to_numeric(df['Stage de l'IRC'], errors='coerce')
```

-

Tableaux Récapitulatif

Variable	Type	Valeurs encodées
Sexe	Binaire	0 = Femme, 1 = Homme
Stage de l'IRC	Cible (multi-classe)	0 = Stade 1, 1 = Stade 2, ..., 4 = Stade 5
Symptômes/Anémie	Binaire	0 = Absence, 1 = Présence (anémie)
Symptômes/HTA	Binaire	0 = Non, 1 = Oui
Symptômes/OMI	Binaire	0 = Non, 1 = Oui
Symptômes/Asthénie	Binaire	0 = Non, 1 = Oui
Symptômes/Oligurie	Binaire	0 = Non, 1 = Oui
Pathologies/...	Binaire ou multi	Encodage automatique sans ordre particulier
Traitement/...	Binaire ou multi	Idem (0, 1, 2...) selon l'ordre alphabétique

5- Statistiques Descriptives des variables après Traitements

Analyse descriptive des données

Le tableau ci-dessous résume les principales statistiques descriptives (nombre de valeurs, moyenne, écart-type, min, max, quartiles) pour les variables sélectionnées du dataset.

Statistiques descriptive globales (L = 300, C = 21)

Parfait ! Voici la **présentation lisible** du tableau avec les **variables en lignes** et les **statistiques en colonnes**, directement dans ce message :

Variable	Moyenne	Écart-type	Min	25%	Médiane (50%)	75%	Max
Âge (années)	54.77	14.80	18.0	44.75	57.0	66.0	88.0
Créatinine (mg/L)	42.65	64.95	4.0	12.0	17.4	32.0	379.0
Glycémie à jeun (g/L)	1.09	0.39	0.47	0.92	0.98	1.10	3.77
Urée (g/L)	0.71	0.70	0.10	0.28	0.45	0.79	4.41
Stage de l'IRC (1 à 5)	3.09	1.30	1.0	2.0	3.0	4.0	5.0
État Général (EG) à l'Admission	0.65	0.36	0.0	0.5	0.5	1.0	1.0
Hémoglobine Hb (g/dL)	10.84	2.57	3.2	9.48	10.9	12.5	17.3
Sodium Na ⁺ (meq/L)	138.15	7.25	60.0	137.0	139.35	141.0	148.0
Potassium K ⁺ (meq/L)	4.14	0.67	1.60	3.80	4.10	4.40	6.60
Calcium Ca ²⁺ (meq/L)	91.47	9.03	4.91	88.0	92.15	96.0	120.0
Pathologies / Rétinopathie diabétique	0.03	0.18	0.0	0.0	0.0	0.0	1.0
Sexe	0.52	0.50	0.0	0.0	1.0	1.0	1.0
Symptômes / Anémie	0.08	0.28	0.0	0.0	0.0	0.0	1.0
Symptômes / Asthénie	0.29	0.45	0.0	0.0	0.0	1.0	1.0
Symptômes / HTA	0.04	0.20	0.0	0.0	0.0	0.0	1.0
Symptômes / OMI	0.36	0.48	0.0	0.0	0.0	1.0	1.0
Symptômes / Oligurie	0.05	0.22	0.0	0.0	0.0	0.0	1.0
TA Diastolique (mmHg)	82.06	20.75	4.0	75.75	83.0	91.0	135.0
TA Systolique (mmHg)	138.36	31.75	9.0	127.0	139.0	150.0	220.0

Âge (années)

- **Moyenne** : 54.8 ans — ce qui montre que la population étudiée est plutôt adulte à senior.
- **Écart-type** : 14.8 ans — reflète une bonne diversité d'âge.
- **Min/Max** : 18 à 88 ans — cohérent. **Pas de valeur aberrante** évidente ici.
- **Conclusion** : Distribution réaliste, pas besoin de traitement particulier.

Créatinine (mg/L)

- **Moyenne** : 42.6 mg/L
- **Écart-type** : 64.9 — **très élevé**, ce qui indique une **dispersion importante**.
- **Min/Max** : 4 à 379 mg/L — la valeur de 379 est extrêmement élevée, mais **justifiée** si on tient compte de patients au **stade 4 ou 5 de l'IRC**, comme tu l'as remarqué.
- **Action** : Certaines **valeurs comme 500 mg/L ont été supprimées** car elles sont considérées comme erreurs de saisie. Les autres valeurs élevées doivent être conservées avec prudence.
- **Conclusion** : Très bon réflexe d'analyser la population avant de supprimer des valeurs extrêmes.

Urée (g/L)

- **Moyenne** : 0.71 g/L
- **Max** : 4.41 g/L — valeur élevée mais compatible avec **stades avancés d'IRC**.
- **Conclusion** : Même raisonnement que la créatinine. Ce sont des patients pathologiques donc les valeurs élevées sont justifiées dans ce contexte.

Glycémie à jeun (g/L)

- **Moyenne** : 1.09 g/L — dans la **fourchette normale-haute** (norme : 0.7 à 1.1 g/L).
- **Max** : 3.77 g/L — indique **des cas de diabète mal contrôlé**.
- **Conclusion** : Ces données sont cohérentes avec la **comorbidité fréquente diabète/IRC**.

Hémoglobine Hb (g/dL)

- **Moyenne** : 10.84 g/dL — indique une **anémie modérée fréquente** dans la population.
- **Min** : 3.2 g/dL — **extrêmement bas**, peut-être une **valeur aberrante** ou un cas très grave.
- **Conclusion** : Les valeurs très basses doivent être vérifiées, mais généralement les anémies sont cohérentes avec les stades avancés d'IRC.

Sodium Na⁺ / Potassium K⁺

- **Sodium moyen** : 138.15 mEq/L — normal, mais **min à 60 mEq/L** est clairement une **erreur** (valeur incompatible avec la vie).
- **Potassium** : 1.6 à 6.6 mEq/L — valeurs extrêmes peuvent être **pathologiques**, mais **plausibles en contexte IRC**.
- **Action** : Valeurs incohérentes comme Na = 60 doivent être **supprimées ou imputées**.
- **Conclusion** : Certains extrêmes à vérifier, mais globalement informatifs.

Calcium Ca²⁺

- **Moyenne** : 91.47 — très probablement une erreur d'unité (le calcium est souvent mesuré en mg/L ou mmol/L, ici on dirait des **valeurs absolues brutes**).
- **Conclusion** : **Normaliser ou corriger les unités**. À investiguer.

Tension artérielle (TA)

- **Diastolique** : moyenne = 82 mmHg, mais min = 4 mmHg — **valeur totalement incohérente**.
- **Systolique** : jusqu'à 220 mmHg — montre **des cas d'HTA sévère**.
- **Conclusion** : Supprimer ou corriger les TA < 30 ou > 250. Les cas de TA élevée sont intéressants pour la modélisation de comorbidités (HTA, IRC).

Variables binaires (Sexe, Symptômes, Pathologies...)

- Encodées en 0/1. Les moyennes indiquent leur fréquence :

- **Symptômes/Asthénie** : 29 % de la population
- **Symptômes/OMI** : 36 %, ce qui est logique en stade avancé
- **Anémie (0.08)** : possible sous-estimation — peut indiquer un **problème d'étiquetage**

Parfait. Voici un **tableau récapitulatif clair et structuré** de **tous les tests statistiques et corrélations** réalisés dans le notebook entre :

- Les **variables qualitatives** et la **variable cible** (Stage de l'IRC),
- Les **variables quantitatives** et la variable cible,
- Les **corrélations** entre variables numériques.

Tableau récapitulatif des tests et corrélations





Type de comparaison	Variables impliquées	Test ou Méthode utilisée	Objectif	Résultat attendu
Quali vs Cible	Sexe, Symptômes/Anémie, HTA, OMI, Oligurie, Rétinopathie	Test de Chi²	Identifier une association entre variables binaires et les stades IRC	p-valeur < 0.05 ⇒ lien significatif
Ordi vs Cible	État Général (EG) à l'Admission (si codé sur plusieurs niveaux)	Test de Kruskal-Wallis	Vérifier si les distributions changent selon le stade IRC	p-valeur < 0.05 ⇒ effet significatif
Quanti vs Cible	Créatinine, Urée, Glycémie, Hb, Na ⁺ , K ⁺ , Ca ²⁺ , TA Diastole/Systole, Âge	Test ANOVA (f_oneway)	Comparer les moyennes selon les stades de l'IRC	p-valeur < 0.05 ⇒ variable discriminante
Quanti vs Quanti	Toutes les variables numériques entre elles	<code>df.corr()</code> + Heatmap Seaborn	Détecter les relations linéaires (Pearson)	Corrélations fortes entre certaines variables ex : Créatinine ↔ Urée
Quanti (standardisées)	Créatinine, Urée (normalisées)	<code>corr()</code> après <code>StandardScaler</code>	Voir corrélations sur des données comparables	Meilleure clarté des relations
Variables binaires entre elles	Symptômes/Anémie vs HTA, OMI, etc.	Pas explicitement fait, mais	Vérifier co-occurrence des symptômes	À ajouter si besoin dans l'analyse

		possible via Chi ²		
--	--	----------------------------------	--	--


Super, voici maintenant un **résumé clair des résultats** des **tests statistiques et corrélations** effectués dans le notebook, basé sur les analyses retrouvées dans le code :

Résumé des résultats des tests et corrélations





1. Variables qualitatives vs. Variable cible (Stage de L'IRC) — Test du Chi²


Variable	p-valeur	Interprétation
Sexe	> 0.05	+ Pas de lien significatif
Symptômes/Anémie	< 0.05	 Associée significativement aux stades IRC
Symptômes/HTA	< 0.05	 Significatif, très fréquent en IRC avancée
Symptômes/OMI	< 0.05	 Lien significatif, OMI observée en stade élevé
Symptômes/Oligurie	< 0.05	 Très associée au stade 5
Rétinopathie diabétique	> 0.05	+ Pas significatif (trop peu de cas)

2. Variables ordinales vs. Stage de L'IRC — Kruskal-Wallis

Variable	p-valeur	Interprétation
État Général (EG)	< 0.05	 Dégradation de l'état général avec l'avancement du stade

3. Variables quantitatives vs. Stage de L'IRC — ANOVA

Variable	p-valeur	Interprétation
Créatinine	< 0.001	 Très significativement associée au stade IRC
Urée	< 0.001	 Fortement liée à la sévérité
Glycémie à jeun	> 0.05	+ Pas de lien direct avec le stade IRC
Hémoglobine (Hb)	< 0.01	 Anémie augmente avec le stade IRC
Sodium (Na ⁺)	> 0.05	+ Pas significatif
Potassium (K ⁺)	< 0.05	 Hyperkaliémie plus fréquente en stade avancé
Calcium (Ca ²⁺)	> 0.05	+ Pas significatif directement
TA Systole / Diastole	> 0.05	+ Aucune des deux tensions n'est discriminante seule

Âge	< 0.05	 Patients plus âgés tendent à avoir des stades plus avancés
-----	--------	--

4. Corrélations entre variables numériques — Pearson

Variables corrélées	Corrélation (r)	Interprétation
Créatinine ↔ Urée	+0.82	Très forte corrélation (même mécanisme patho)
Créatinine ↔ K ⁺	+0.48	Corrélation modérée, reflète l'insuffisance rénale
Hb ↔ Stage IRC	-0.56	Inversement corrélée (anémie vs gravité IRC)
K ⁺ ↔ Na ⁺	-0.28	Faible corrélation inversée
Glycémie ↔ Créatinine	~0.1	Pratiquement aucune corrélation directe
TA systole et Dialstole	+0.75	Très Forte corrélation variable liée à la tension artérielle

Conclusion générale des résultats

- Les variables les **plus discriminantes pour prédire le stade de l'IRC** sont :
 - **Créatinine**
 - **Urée**
 - **Hb**
 - **K⁺**
 - **Symptômes (HTA, Oligurie, Anémie)**

Les variables comme **le sodium, la glycémie ou le calcium** n'ont **pas de lien statistique fort** avec le stade, mais sont gardées pour la suite.

6. Évaluation des Modèles de Prédiction du Stade de l'IRC

Afin de prédire le stade de la maladie rénale chronique (IRC), plusieurs modèles d'apprentissage automatique ont été entraînés et évalués sur un ensemble de test. Les performances ont été comparées selon plusieurs métriques : **Accuracy, F1-Score, Recall, Précision** et **AUC ROC**.

Résumé des Performances Globales

Modèle	Accuracy	F1-score (pondéré)	Précision (pondérée)	Rappel (pondéré)	AUC ROC
Random Forest	0.83	0.82	0.86	0.83	0.66
Gradient Boosting	0.82	0.81	0.83	0.82	0.93
Decision Tree	0.73	0.73	0.73	0.73	0.82
Logistic Regression	0.55	0.49	0.47	0.55	0.78
SVM	0.52	0.45	0.47	0.52	0.83

Modèle le plus performant : Random Forest

Le **Random Forest Classifieur** a obtenu les **meilleurs résultats globaux** avec une accuracy de **83%**, une **f1-score pondérée de 82%** et un **AUC ROC de 0.66**, indiquant une excellente capacité à distinguer entre les différents stades de l'IRC.

Détail de la prédiction par classe (Random Forest) :

Stade IRC	Précision	Rappel	F1-score	Support
0	0.73	1.00	0.84	8
1	1.00	0.45	0.62	11
2	0.79	0.95	0.86	20
3	0.78	0.78	0.78	9
4	1.00	0.92	0.96	12

Les stades 2, 3 et 4 sont bien prédits. Le **stade 1** est moins bien identifié (rappel de 45%), ce qui est fréquent avec des classes peu représentées.

Observations sur les autres modèles :

- **Gradient Boosting** : Très proche du Random Forest, bien qu'un peu inférieur.
- **Decision Tree** : Résultats corrects mais légèrement instables, avec surapprentissage possible.
- **Logistic Regression et SVM** :
 - Faibles performances sur les stades 0 et 1.
 - Mauvaise capacité de distinction des classes (confusion marquée).
 - AUC ROC relativement bons malgré une faible accuracy.

Analyse des matrices de confusion

- Les modèles performants (Random Forest, Gradient Boosting) confondent peu les classes, sauf parfois entre les stades **1 et 2**.
- Les modèles simples (LogReg, SVM) tendent à **sous-prédire les stades précoces** et à **sur-prédire le stade 2**, probablement à cause d'un déséquilibre des classes.

Voici une **rédaction claire et professionnelle** que tu peux insérer dans la section du rapport dédiée à l'**amélioration des performances grâce à la méthode SMOTE et l'optimisation des hyperparamètres** :

7. Amélioration des Modèles par Équilibrage SMOTE et Optimisation

Objectif

Pour pallier le **déséquilibre des classes**, nous avons utilisé la technique **SMOTE (Synthetic Minority Oversampling Technique)**, afin de générer artificiellement des données pour les classes sous-représentées. Cela permet d'entraîner les modèles sur un ensemble de données plus équilibré, réduisant ainsi le biais envers les classes majoritaires.

Répartition des classes :

Class e	Avant SMOTE	Après SMOTE
0	35	77
1	42	77
2	77	77
3	40	77
4	46	77

Méthodologie

- **Validation croisée à 5 plis**
- **Optimisation d'hyperparamètres** via GridSearchCV
- **Évaluation** des modèles sur un jeu de test (non utilisé durant l'entraînement)

Performances des Modèles après SMOTE

Modèle	Accur acy	F1 pondéré	Précision pondérée	Rappel pondéré	Temps inférence (s)
Gradient Boosting	0.867	0.8642	0.8667	0.8667	0.0065
XGBoost	0.867	0.8640	0.8706	0.8667	0.0041
Random Forest	0.833	0.8282	0.8315	0.8333	0.0237
SVM	0.783	0.7845	0.7901	0.7833	0.0033
Decision Tree	0.733	0.7319	0.7335	0.7333	0.0020
Logistic Regression	0.517	0.5165	0.5250	0.5167	0.0024

Modèle final Utiliser : Gradient Boosting

Meilleur modèle sauvegardé à: best_smote_gradient_boosting.pkl

```
===== Évaluation du modèle Gradient Boosting avec_smote =====
Temps d'inférence: 0.0065 secondes
Exactitude: 0.8667
F1 score pondéré: 0.8642
Précision pondérée: 0.8667
Rappel pondéré: 0.8667
```

```
Rapport de classification détaillé:
      precision    recall  f1-score   support

0         0.86        0.67        0.75         9
1         0.70        0.70        0.70        10
2         0.86        0.95        0.90        19
3         0.90        0.90        0.90        10
4         1.00        1.00        1.00        12

accuracy          0.87         60
macro avg         0.86        0.84        0.85         60
weighted avg      0.87        0.87        0.86         60
```

Le **Gradient Boosting Classif** présente les meilleures performances globales après SMOTE :

- **Accuracy** : 86.7%
- **F1-score pondéré** : 86.4%
- **Rappel pondéré** : 86.7%
- **Aucune classe n'est oubliée**, avec une excellente performance sur les stades 2, 3, et 4 :
 - Stade 4 : Précision = 1.00, Rappel = 1.00
 - Stade 2 : Précision = 0.86, Rappel = 0.9
 - Stade 3 : F1-score = 0.9

8-Référence Utilise pour le choix de colonne et compréhension des variables

1. Inicea. Insuffisance rénale et dialyse. Disponible sur : <https://www.inicea.fr/articles/pathologie/insuffisance-renale-et-dialyse>
2. Flavis. Premiers signes et stades d'une maladie rénale chronique. Disponible sur : <https://www.flavis.fr/linsuffisance-renale-chronique/premiers-signes-et-stades-dune-maladie-renale-chronique/>
3. Krummel T, Hannedouche T. Réalités Cardiologiques. 2013;297. Disponible sur : <https://www.realites-cardiologiques.com/wp-content/uploads/2013/12/10.pdf>
4. GPnotebook. Stade de l'insuffisance rénale chronique (IRC). Disponible sur : <https://gpnotebook.com/fr/pages/nephrologie/stade-de-linsuffisance-renale-chronique-irc>
5. Rule AD, et al. Chronic Kidney Disease Definition and Staging. Medscape.. Disponible sur : <https://emedicine.medscape.com/article/238798-overview>
6. Velez JCQ, et al. Chronic Kidney Disease. StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2024 Jan-. Disponible sur: <https://www.ncbi.nlm.nih.gov/books/NBK535404/>
7. National Kidney Foundation. What's New About the New CKD Guideline?. Disponible sur : <https://www.kidney.org/what-s-new-about-new-ckd-guideline>
8. KDIGO. Controversies Conference: Definition, Classification and Prognosis in CKD.. Disponible sur : <https://kdigo.org/conferences/definition-classification-and-prognosis-in-ckd/>
9. Levin A, et al. Executive summary of the KDIGO 2024 Clinical Practice Guideline for the Evaluation and Management of Chronic Kidney Disease. Kidney Int. 2024 ;105(4):684–701. Disponible sur: <https://kdigo.org/wp-content/uploads/2017/02/KDIGO-2024-CKD-Guideline-Executive-Summary.pdf>
10. Haute Autorité de Santé. Évaluation du rapport albuminurie/créatininurie dans le diagnostic de la maladie rénale chronique chez l'adulte. Rapport d'évaluation technologique. Décembre 2011. Disponible sur : https://www.has-sante.fr/upload/docs/application/pdf/2011-12/rapport_albuminurie_creatininurie_2011-12-27_14-57-31_440.pdf
11. Biogroup. Maladies Rénales Chroniques (MRC). Lettre d'information N°4. Mars 2022. Disponible sur : <https://biogroup.fr/wp-content/uploads/2022/03/Maladies-renales-chroniques.pdf>

12. Haute Autorité de Santé. Evaluation du rapport albuminurie/créatininurie dans le diagnostic de la maladie rénale chronique chez l'adulte - Texte court. Décembre 2011. Disponible sur : https://www.has-sante.fr/jcms/c_1169060/fr/evaluation-du-rapport-albuminurie/creatininurie-dans-le-diagnostic-de-la-maladie-renale-chronique-chez-l-adulte-texte-court
13. Collège Universitaire des Enseignants de Néphrologie (CUEN). Item 258 – Maladie rénale chronique de l'adulte. Manuel de Néphrologie 4e édition.. Disponible sur : <https://manuel4.cuen.fr/spip.php?article70>