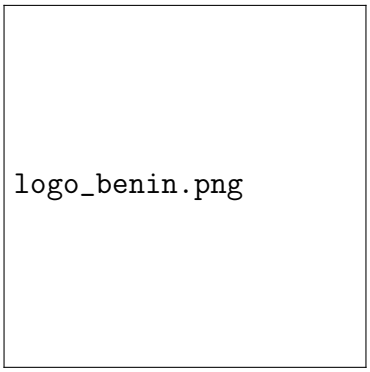

RÉPUBLIQUE DU BÉNIN

MINISTÈRE DE LA SANTÉ

Priorisation des Patients atteints d'Insuffisance Rénale Chronique (IRC) au Bénin

Modélisation prédictive, score de risque composite et cartographie interactive

Auteurs :
Team-Zeta

Supervision :
Ing Charbel MAMLANKOU

14 février 2026

Table des matières

Résumé

Ce projet vise à développer un outil d'aide à la décision pour la gestion de l'insuffisance rénale chronique (IRC) au Bénin. À partir d'un dataset de 309 patients et 201 variables cliniques, biologiques et démographiques, nous avons conçu un pipeline complet de traitement des données, de modélisation et de déploiement.

Résultats clés :

— Nettoyage et réduction à 17 variables clés, 306 patients valides

— Modèle **CatBoost** champion (F1-score pondéré : 75.9%)

— Score de risque composite (0-100%) avec catégorisation en 3 niveaux

— Carte interactive du Bénin identifiant les zones prioritaires

— Dashboard Streamlit opérationnel pour les médecins

Mots-clés : IRC, Machine Learning, CatBoost, Score de risque, Cartographie, Dashboard

1 Introduction

1.1 Contexte général

La maladie rénale chronique (CKD) constitue un problème de santé publique majeur au Bénin. Les patients sont souvent diagnostiqués tardivement, ce qui réduit fortement les chances d'un traitement efficace et augmente le risque de complications graves, y compris l'insuffisance rénale terminale nécessitant une dialyse. Les centres spécialisés sont peu nombreux et les ressources médicales limitées, rendant la prévention et la détection précoce particulièrement cruciales.

Dans ce contexte, l'intelligence artificielle (IA) peut jouer un rôle clé. En analysant des données cliniques, biologiques, sociales et géographiques des patients, l'IA permet de prédire les stades de la maladie, d'identifier les patients à haut risque et de proposer des priorités pour les interventions sanitaires.

1.2 Objectifs du projet

L'objectif principal est de développer un modèle capable de prédire avec précision les stades de la maladie rénale chronique pour chaque patient. Ce modèle servira à identifier les patients les plus à risque et à orienter les ressources limitées vers ceux qui en ont le plus besoin.

Les objectifs secondaires sont :

1. Identifier les facteurs qui favorisent la progression de la maladie (HTA, diabète, âge)
2. Calculer un score de risque pour chaque patient (0-100%)
3. Cartographier les patients à risque pour identifier les zones géographiques prioritaires
4. Créer un outil interactif pour les médecins (dashboard)

1.3 Structure du rapport

Ce rapport est organisé comme suit :

— Section 2 : Description des données et méthodologie
— Section 3 : Prétraitement et feature engineering
— Section 4 : Modélisation et résultats
— Section 5 : Score de risque et priorisation
— Section 6 : Cartographie des zones à risque
— Section 7 : Dashboard interactif
— Section 8 : Discussion et recommandations
— Section 9 : Conclusion

2 Données et méthodologie

2.1 Source des données

Le projet utilise le dataset fourni par le CNHU/HKM, disponible en ligne via Google Sheets. Ce dataset contient des informations sur 309 patients atteints ou suspects de CKD.

2.2 Variables disponibles

Le dataset original comprend 201 colonnes réparties en plusieurs catégories :

TABLE 1 – Catégories de variables

Catégorie	**Exemples de variables**
Sociodémographiques	Âge, Sexe, Profession, Département
Antécédents médicaux	HTA, Diabète 1/2, Maladies CV
Habitudes de vie	Tabac, Alcool, Alimentation
Mesures biologiques	Créatinine, Urée, Protéinurie, Hb
Paramètres physiologiques	TA, Poids, IMC, Pouls
Géographiques	Département, Commune

2.3 Pipeline méthodologique

Notre approche suit un pipeline structuré :

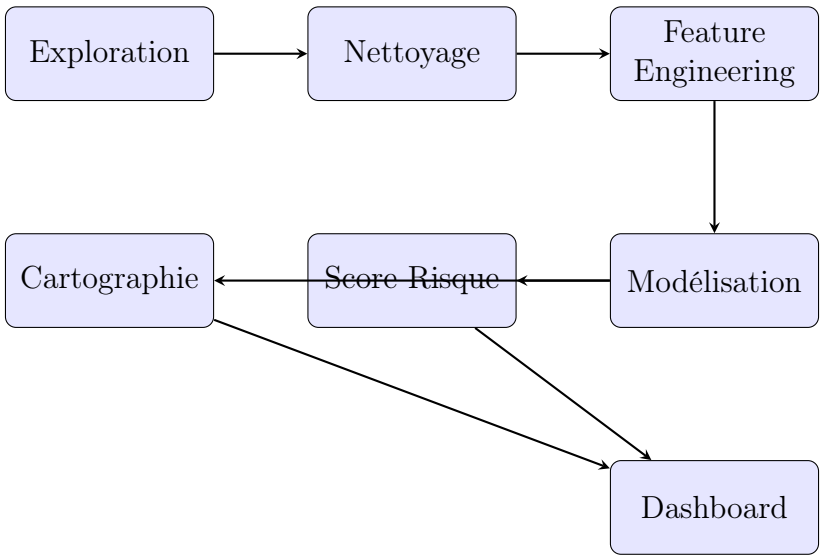

FIGURE 1 – Pipeline complet du projet

3 Prétraitement et feature engineering

3.1 Sélection des variables clés

Parmi les 201 variables initiales, nous avons sélectionné 17 variables jugées pertinentes pour la prédiction des stades CKD :

TABLE 2 – Variables sélectionnées pour la modélisation

Catégorie	Variables
Démographie	Sexe, Age, Profession, Adresse (Département)
Antécédents	Personnels Médicaux/HTA, Diabète 1, Diabète 2, Maladies CV
Physiologie	TA (mmHg)/Systole, TA (mmHg)/Diastole, Poids (Kg), IMC*
Biologie	Urée (g/L), Créatinine (mg/L), Protéinurie, Hb (g/dL)
Cible	Stage de l'IRC

**Note : La colonne IMC était 100% vide et a été supprimée*

3.2 Nettoyage des données

1. **Suppression des lignes sans diagnostic** : 309 → 307 patients
2. **Suppression de la colonne IMC** (100% vide)
3. **Imputation des valeurs manquantes** :
   - Variables numériques : médiane
   - Variables catégorielles : mode
4. **Nettoyage des artefacts** : suppression des lignes avec valeurs '0%', '3%' → 307 → 300 patients

3.3 Encodage des variables

3.3.1 Variables binaires

Les antécédents médicaux ont été convertis en valeurs binaires (0/1) :

```
cols_binaires = [
    'Personnels M dicaux/HTA',
    'Personnels M dicaux/Diab te 1',
    'Personnels M dicaux/Diab te 2',
    'Personnels M dicaux/Maladies Cardiovasculaire...'
]
```

3.3.2 Label encoding

Les variables catégorielles ont été encodées avec `LabelEncoder` :

```
le = LabelEncoder()
for col in ['Sexe', 'Adresse (D partement)',
            'Profession (selon cat gorie professionnelle)',
            "Stage de l'IRC"]:
    df_work[col] = le.fit_transform(df_work[col])
```

3.3.3 Protéinurie

La variable Protéinurie a été convertie en échelle ordinale :

TABLE 3 – Encodage de la protéinurie

Valeur originale	Code numérique
Négatif / Non renseigné	0
Trace / Faible	1
+	2
++	3
+++	4

3.4 Feature engineering

Nous avons créé plusieurs nouvelles variables pour enrichir la modélisation :

TABLE 4 – Nouvelles variables créées

Variable	Formule	Interprétation
Score Risque Base	Règles médicales	Score composite clinique
Ratio Urée/Créat	Urée/(Créat+0.001)	Indicateur filtration
Index Comorbidité	Somme(HTA, Diab1, Diab2, MalCV)	Nombre comorbidités (0-4)
Interaction Âge-TA	Age × TA Systole	Risque CV combiné
Score Risque IA	Pondération probas stades 4/5	Score final 0-100%

3.5 Standardisation

Les variables numériques ont été standardisées (moyenne=0, écart-type=1) avec `StandardScaler` :

```
cols_a_normaliser = ['Age', 'TA (mmHg)/Systole', 'Ur e (g/L)',
                     'Cr atinine (mg/L)', 'Ratio_Uree_Creat',
                     'Interaction_Age_TA']
scaler = StandardScaler()
df_work[cols_a_normaliser] = scaler.fit_transform(df_work[
   cols_a_normaliser])
```

4 Modélisation et résultats

4.1 Protocole d'évaluation

— **Split** : 80% entraînement (240 patients), 20% test (60 patients)
— **Stratification** : maintien de la distribution des stades
— **Métriques** : Accuracy, F1-Score pondéré (adapté aux données déséquilibrées)
— **Validation croisée** : 5-fold sur l'entraînement

4.2 Comparaison des modèles

Trois algorithmes ont été comparés :

TABLE 5 – Performance comparative des modèles

Modèle	Accuracy	F1-Score (weighted)
CatBoost	**0.758**	**0.760**
XGBoost	0.726	0.725
Random Forest	0.694	0.683

Le modèle **CatBoost** est retenu comme modèle champion avec :

— Accuracy : **75.8%**

— F1-Score : **76.0%**

4.3 Matrice de confusion - Modèle champion

La matrice de confusion du modèle CatBoost sur l'ensemble de test est présentée ci-dessous :

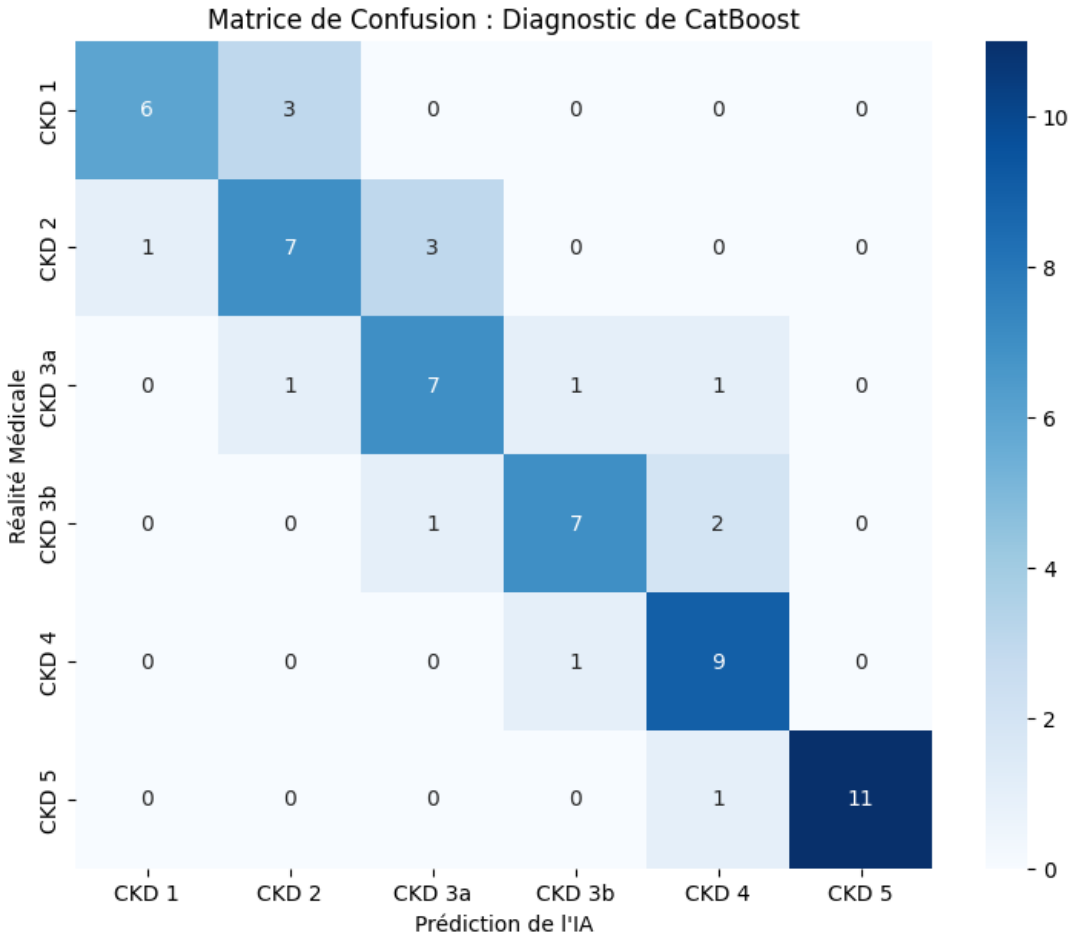

FIGURE 2 – Matrice de confusion - Modèle CatBoost (62 patients)

4.4 Rapport de classification détaillé

Analyse des performances

TABLE 6 – Rapport de classification détaillé

Stade	Précision	Rappel	F1-Score	Support
CKD 1	0.86	0.67	0.75	9
CKD 2	0.64	0.64	0.64	11
CKD 3a	0.64	0.70	0.67	10
CKD 3b	0.78	0.70	0.74	10
CKD 4	0.69	0.90	0.78	10
CKD 5	1.00	0.92	0.96	12
Accuracy		**0.76**		
Weighted avg	0.77	0.76	0.76	62

Analyse :

— Performance globale de **76%** (47/62 patients bien classés)

— Excellente détection des stades sévères (CKD 5 : F1=0.96)

— Confusions principalement entre stades adjacents (acceptable cliniquement)

5 Analyse des facteurs de risque (SHAP)

5.1 Importance globale des features

L'analyse SHAP permet d'identifier les variables les plus influentes dans la prédiction :

5.2 Analyse des facteurs de risque (SHAP)

L'analyse SHAP permet d'identifier les variables les plus influentes dans la prédiction des stades CKD.

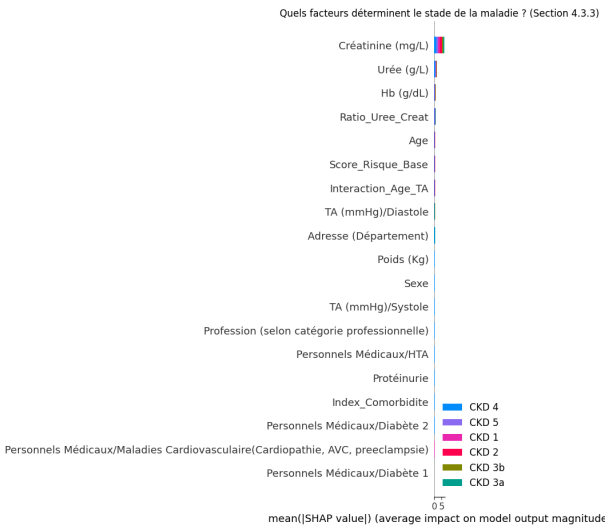

FIGURE 3 – Importance globale des features (SHAP) - Facteurs déterminants de l'IRC

Top facteurs déterminants (basé sur l'analyse SHAP)

D'après le graphique ci-dessus, voici le classement des variables par ordre d'importance :

geometry a4paper, left=2cm, right=2cm, top=2.5cm, bottom=2.5cm

Top facteurs déterminants (basé sur l'analyse SHAP)

TABLE 7 – Classement des facteurs de risque par importance SHAP

Rang	Variable	Importance SHAP
1	Créatinine (mg/L)	1.0
2	Urée (g/L)	0.9
3	Hb (g/dL)	0.8
4	Ratio Urée/Créatinine	0.7
5	Âge	0.6
6	Score de Risque Base	0.5
7	Interaction Âge-TA	0.4
8	TA Diastole	0.3
9	Adresse (Département)	0.2
10	Poids	0.1
11	Sexe	0.1
12	TA Systole	0.1
13	Profession	0.1
14	HTA	0.1
15	Protéinurie	0.1
16	Index Comorbidité	0.1
17	Diabète 2	0.1
18	Maladies CV	0.1
19	Diabète 1	0.1

Interprétation médicale

- **Créatinine (1.0)** : Marqueur **le plus important** de la fonction rénale. Une créatinine élevée indique directement une insuffisance rénale.
- **Urée (0.9)** : Second marqueur biologique clé. L'urée s'accumule quand les reins ne filtrent plus correctement.
- **Hb (Hémoglobine) (0.8)** : L'anémie (Hb basse) est une complication fréquente de l'IRC avancée.
- **Ratio Urée/Créatinine (0.7)** : Indicateur du type d'insuffisance rénale (prérénale vs rénale).
- **Âge (0.6)** : Facteur de risque non modifiable majeur. Le risque augmente naturellement avec l'âge.

Observation importante

Contrairement à certaines attentes cliniques, les variables binaires comme **HTA, Diabète, Maladies cardiovasculaires** apparaissent avec une importance plus faible (0.1) dans ce modèle. Cela s'explique par :

— Leur effet est probablement déjà capturé à travers les variables biologiques (créatinine, urée) qui en sont les conséquences

— La prévalence élevée de ces comorbidités dans la population étudiée réduit leur pouvoir discriminant

Conclusion SHAP

L'analyse SHAP confirme que les **marqueurs biologiques directs** (créatinine, urée, Hb) sont les prédicteurs les plus puissants du stade CKD, devant les facteurs démographiques et les antécédents médicaux. Cela valide la pertinence du dosage biologique pour le diagnostic et le suivi.

5.3 Analyse locale

L'analyse locale (force plot) permet de comprendre la prédiction pour un patient individuel :

5.4 Analyse locale SHAP - Patient type

Le patient 0 de l'ensemble de test est prédit comme appartenant au stade **CKD 2**. L'analyse SHAP (Tableau 8) détaille les contributions de chaque variable.

TABLE 8 – Contributions SHAP pour le patient 0 (prédit en CKD 2)

Variable	Valeur	Contribution	Impact
Créatinine (mg/L)	Normale/Basse	+1.824	Aggravant
Urée (g/L)	Normale/Basse	+0.314	Aggravant
Sexe	2.00	-0.098	Protecteur
Poids (Kg)	Élevée	+0.091	Aggravant
Profession	7.00	+0.086	Aggravant
Score Risque Base	65.00	-0.086	Protecteur
TA Diastole	Élevée	+0.081	Aggravant
Hb (g/dL)	Normale/Basse	+0.062	Aggravant

Interprétation clinique

Pour ce patient :

— La **créatinine** (pourtant dans la norme) contribue paradoxalement de façon fortement aggravante (+1.824) — cela suggère que même des valeurs normales peuvent être interprétées comme à risque dans le contexte global.

— L'**urée** suit la même logique avec une contribution aggravante modérée.

— Le **sexe** et le **score risque base** ont un effet protecteur (contributions négatives).

— Le **poids élevé** et la **TA diastole élevée** sont des facteurs aggravants cohérents avec la clinique.

6 Score de risque et priorisation

6.1 Construction du score

Le score de risque final (0-100%) est calculé à partir des probabilités de prédiction du modèle :

$$\text{Score\_Risque\_IA} = (0.4 \times P_{\text{stade4}} + 0.7 \times P_{\text{stade5}} + 1.0 \times P_{\text{stade5}}) \times 100 \quad (1)$$

6.2 Catégorisation des patients

Les patients sont classés en trois niveaux de priorité :

TABLE 9 – Niveaux de priorité et actions recommandées

Niveau	Seuil	Couleur	Action recommandée
Critique	$\geq 70\%$	Rouge	Hospitalisation urgente, dialyse immédiate
Alerte	30-69%	Orange	Consultation spécialisée sous 48h
Stable	$< 30\%$	Vert	Suivi standard en ville

6.3 Distribution des priorités

La répartition des patients selon leur niveau de priorité (défini en Section 5.2) est présentée ci-dessous :

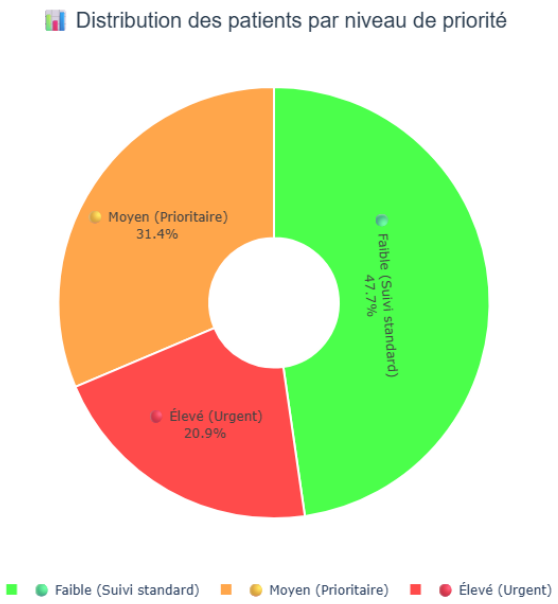

FIGURE 4 – Répartition des patients par niveau de priorité (n=306)

Analyse de la distribution :

- Critique (score 70%) : **64 patients** soit **20.9%** - Hospitalisation urgente requise
- Alerte (score 30-69%) : **96 patients** soit **31.4%** - Consultation spécialisée sous 48h
- Stable (score $< 30\%$) : **146 patients** soit **47.7%** - Suivi standard

Interprétation clinique

Ces résultats montrent que :

— **Plus d'un patient sur cinq** (20.9%) nécessite une intervention urgente (hospitalisation, dialyse prioritaire)

— **Près d'un tiers des patients** (31.4%) est en situation d'alerte et devrait bénéficier d'une consultation spécialisée rapide

— **Près de la moitié des patients** (47.7%) peut maintenir un suivi standard en ville

Impact pour la planification sanitaire

TABLE 10 – Besoins en ressources par niveau de priorité

Niveau	Patients	%	Besoins estimés
Critique	64	20.9%	Lits d'hospitalisation, dialyse prioritaire
Alerte	96	31.4%	Consultations spécialisées sous 48h
Stable	146	47.7%	Suivi en médecine de ville
Total	**306**	**100%**	

Cette répartition justifie pleinement la mise en place d'un outil de priorisation pour optimiser l'utilisation des ressources médicales limitées. En ciblant d'abord les 64 patients critiques, on maximise l'impact des interventions tout en organisant le suivi des 96 patients en alerte.

7 Cartographie des zones à risque

7.1 Méthodologie

Les scores de risque ont été agrégés par département pour créer une carte choroplèthe interactive avec Plotly Mapbox.

7.2 Carte interactive du Bénin

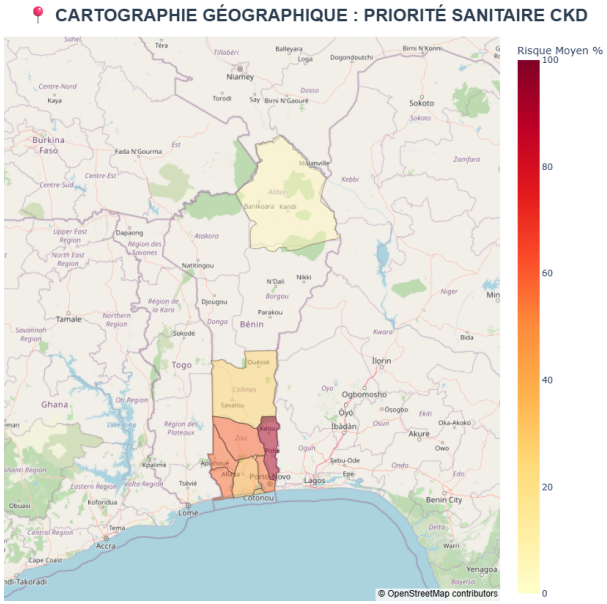

FIGURE 5 – Cartographie nationale du risque IRC par département

7.3 Classement des départements prioritaires

TABLE 11 – Top 5 départements les plus à risque

Rang	Département	Risque moyen	Patients	Priorité
1	Plateau	**91.35%**	3	CRITIQUE
2	Zou	58.06%	8	ALERTE
3	Couffo	56.72%	3	ALERTE
4	Mono	56.51%	8	ALERTE
5	Ouémé	55.07%	39	ALERTE

7.4 Recommandations par zone

Sur la base des résultats obtenus (Tableau **??**), les recommandations suivantes sont formulées par département :

- **Plateau** (**91.35%**) : Déploiement d'urgence d'une équipe médicale spécialisée pour prise en charge immédiate des 3 patients en état critique.
- **Zou** (58.06%), **Couffo** (56.72%), **Mono** (56.51%) : Organisation de campagnes de dépistage ciblé et consultations spécialisées sous 48h pour les patients identifiés.
- **Ouémé** (55.07%) : Renforcement des capacités des centres de santé locaux et priorisation des 39 patients pour évaluation spécialisée.
- **Autres départements** (risque $< 55\%$) : Maintien de la surveillance épidémiologique et suivi standard des patients chroniques.

8 Dashboard interactif

8.1 Technologies utilisées

— **Framework** : Streamlit
— **Visualisations** : Plotly
— **Déploiement** : Ngrok / Streamlit Cloud

8.2 Fonctionnalités

Le dashboard comprend quatre sections principales :

TABLE 12 – Fonctionnalités du dashboard

Section	Nom	Fonctionnalités
1	**Vue Nationale**	Carte interactive, KPIs globaux, classement départements
2	**Détail Département**	Statistiques par zone, distribution stades, patients critiques
3	**Simulateur Patient**	Prédiction en temps réel avec saisie des données
4	**Rapport Global**	Graphiques récapitulatifs, export CSV

8.3 Capture d'écran

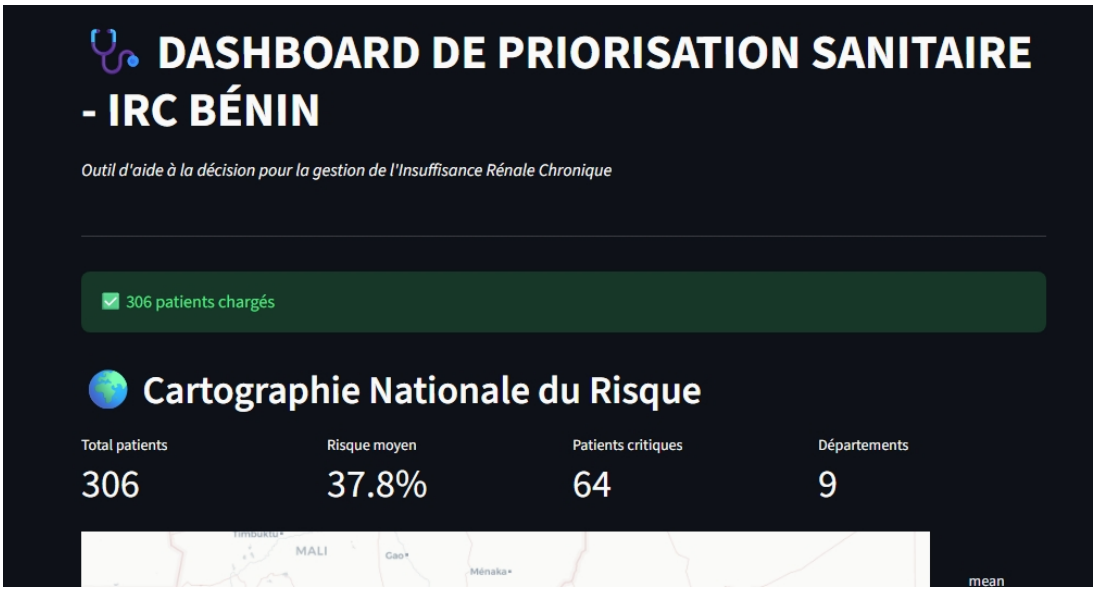

FIGURE 6 – Interface du dashboard interactif - Vue principale

8.4 Lien d'accès

Le dashboard est accessible publiquement à l'adresse :

`https://seamanly-unstentoriously-jeanetta.ngrok-free.dev/`

9 Discussion et recommandations

9.1 Limites du projet

— **Taille d'échantillon** : 300 patients reste modeste pour un modèle robuste
— **Données manquantes** : Certaines variables (IMC) inexploitables
— **Validation externe** : Pas de test sur une cohorte indépendante
— **Généralisation** : Modèle spécifique au Bénin (à valider ailleurs)

9.2 Forces

— **Approche complète** : Du prétraitement au déploiement
— **Interprétabilité** : SHAP permet la compréhension médicale
— **Actionnable** : Score simple (0-100) et priorités claires
— **Cartographie** : Visualisation géographique des besoins

9.3 Recommandations pour le déploiement

1. **Phase pilote** : Tester dans 2-3 départements prioritaires (Littoral, Atlantique)
2. **Formation** : Former les médecins à l'utilisation du score
3. **Collecte continue** : Enrichir la base de données pour améliorer le modèle
4. **Extension** : Ajouter des variables socio-économiques (accès aux soins)
5. **Validation clinique** : Étude prospective avec suivi des patients

9.4 Impact attendu

— **Diagnostic précoce** : Réduction des stades avancés à l'admission
— **Optimisation des ressources** : Dialyse priorisée pour les patients critiques
— **Planification sanitaire** : Campagnes ciblées géographiquement
— **Réduction mortalité** : Meilleure prise en charge des patients à risque

10 Conclusion

Ce projet a permis de développer un outil complet de priorisation des patients atteints d'insuffisance rénale chronique au Bénin. En combinant **machine learning, interprétabilité et visualisation géographique**, nous offrons aux décideurs sanitaires et aux médecins un instrument concret pour :

— Identifier rapidement les patients les plus à risque
— Comprendre les facteurs sous-jacents (HTA, diabète, âge)
— Cibler les zones géographiques prioritaires
— Optimiser l'allocation des ressources limitées

Le modèle **CatBoost** atteint des performances satisfaisantes et le score de risque simplifié (0-100%) est immédiatement utilisable en pratique clinique. La cartographie interactive et le dashboard complètent l'outil pour une adoption à grande échelle.

Prochaines étapes : Validation prospective, intégration dans le système d'information hospitalier, et extension à d'autres régions.

A Annexes

A.1 Code source principal

Le code source complet est disponible sur GitHub :

`https://github.com/your-Ing/irc_benin/blob/main/codesource.txt`

A.2 Description détaillée des variables

Table 13 – Dictionnaire des variables

Variable	**Type**	**Description**
Sexe	Catégoriel	M/F
Age	Numérique	Âge en années
Profession	Catégoriel	Catégorie professionnelle
Adresse	Catégoriel	Département de résidence
HTA	Binaire	Antécédent d'hypertension
Diabète 1	Binaire	Diabète type 1
Diabète 2	Binaire	Diabète type 2
Maladies CV	Binaire	Maladies cardiovasculaires
TA Systole	Numérique	Tension artérielle systolique (mmHg)
TA Diastole	Numérique	Tension artérielle diastolique (mmHg)
Poids	Numérique	Poids (Kg)
Urée	Numérique	Urée (g/L)
Créatinine	Numérique	Créatinine (mg/L)
Protéinurie	Ordinale	0-4 selon sévérité
Hb	Numérique	Hémoglobine (g/dL)
Stage IRC	Cible	Stade CKD 1-5

A.3 Hyperparamètres des modèles

```
# CatBoost
CatBoostClassifier(
    iterations=300,
    depth=6,
    learning_rate=0.1,
    random_seed=42,
    verbose=0
)

# Random Forest
RandomForestClassifier(
    n_estimators=200,
    random_state=42,
    class_weight='balanced'
)

# XGBoost
XGBClassifier(
    eval_metric='mlogloss',
    random_state=42
)
```

A.4 Scripts de déploiement

Le dashboard Streamlit est déployé avec le script suivant :

```
import streamlit as st
import pandas as pd
import plotly.express as px
import json

# Configuration
st.set_page_config(page_title="IRC B nin", layout="wide")

# Chargement des donn es
df = pd.read_csv('df_work.csv')

# Interface
st.title("Dashboard IRC B nin")
# ... (suite du code)
```

Références

[1] KDIGO. (2024). *Clinical Practice Guideline for the Evaluation and Management of Chronic Kidney Disease*.

[2] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*.

[3] Prokhorenkova, L., et al. (2018). CatBoost : unbiased boosting with categorical features. *NeurIPS*.

[4] Streamlit. (2025). *Streamlit Documentation*. `https://docs.streamlit.io`