Rapport de Projet : Prédiction de la Maladie Rénale Chronique

Randriamamonjy Tokiniaina

18 décembre 2025

Table des matières

Chapitre 1

Introduction

Ce rapport présente les résultats du projet de machine learning consistant à analyser le dataset *Chronic Kidney Disease* disponible sur Kaggle et à prédire la présence d'une maladie rénale chronique (CKD) à l'aide de plusieurs modèles de classification.

L'objectif principal était de :

1. Charger et explorer les données,
2. Nettoyer et préparer le dataset,
3. Effectuer une analyse exploratoire détaillée,
4. Comparer les performances de 8 modèles de machine learning différents.

Le travail a été réalisé par **Randriamamonjy Tokiniaina**.

Chapitre 2

Description du Dataset

Le dataset contient 400 observations et 25 variables décrivant des patients. La variable cible est `classification` :

— `ckd` : patient atteint de maladie rénale chronique (250 cas, 62,5%),

— `notckd` : patient sain (150 cas, 37,5%).

Les variables incluent des mesures cliniques (âge, pression artérielle, hémoglobine, créatinine sérique, etc.) et des variables catégorielles (hypertension, diabète, etc.).

Chapitre 3

Analyse Exploratoire des Données (EDA)

3.1 Dimensions et Structure

Le dataset comporte 400 lignes et 25 colonnes après suppression de la colonne `id` non pertinente.

3.2 Valeurs Manquantes

Une heatmap a révélé des valeurs manquantes dans plusieurs colonnes. L'imputation a été effectuée avec la médiane pour les variables numériques et le mode pour les variables catégorielles.

3.3 Distribution de la Variable Cible

La classe `ckd` représente 62,5% des cas, indiquant un léger déséquilibre.

3.4 Visualisations Principales

1. Histogrammes des variables numériques clés (âge, pression artérielle, hémoglobine, etc.).
2. Matrice de corrélation montrant des corrélations fortes entre certaines variables (ex. hémoglobine et volume cellulaire emballé).
3. Boxplots comparant les niveaux d'hémoglobine, de glucose sanguin et de créatinine sérique entre les deux classes : différences significatives observées.
4. Countplots croisés avec hypertension et diabète : forte association avec la présence de CKD.

Chapitre 4

Prétraitement des Données

1. Suppression de la colonne `id`.
2. Nettoyage des variables catégorielles (suppression des espaces et tabulations).
3. Conversion des colonnes `pcv`, `wc`, `rc` en numériques.
4. Imputation des valeurs manquantes (médiane/mode).
5. Encodage des variables catégorielles avec `LabelEncoder`.
6. Normalisation des variables numériques avec `StandardScaler`.
7. Séparation train/test (80%/20%) avec stratification.

Chapitre 5

Modélisation et Résultats

Huit modèles de classification ont été entraînés et évalués :

1. Régression Logistique
2. Arbre de Décision
3. Random Forest
4. XGBoost
5. SVM
6. KNN
7. Naive Bayes
8. Gradient Boosting

Une validation croisée (5-fold) a été utilisée sur l'ensemble d'entraînement. Le déséquilibre des classes a été pris en compte.

5.1 Résultats Globaux

Les résultats obtenus sur l'ensemble de test sont les suivants (triés par AUC sur le test) :

Modèle	CV AUC (Train)	Accuracy	Precision	Recall	F1-Score	AUC (Test)
Random Forest	0.999792	1.0000	1.00	1.00	1.000000	1.000000
Logistic Regression	1.000000	0.9750	1.00	0.96	0.979592	1.000000
SVM	1.000000	0.9875	1.00	0.98	0.989899	1.000000
Naive Bayes	1.000000	0.9750	1.00	0.96	0.979592	1.000000
Gradient Boosting	0.999792	1.0000	1.00	1.00	1.000000	1.000000
XGBoost	0.998542	0.9875	1.00	0.98	0.989899	0.999333
KNN	0.991875	0.9625	1.00	0.94	0.969072	0.999333
Decision Tree	0.973333	0.9625	1.00	0.94	0.969072	0.970000

TABLE 5.1 – Performances des 8 modèles de classification

Les modèles Random Forest et Gradient Boosting atteignent une précision parfaite (Accuracy = 1.000, AUC = 1.000). La plupart des modèles obtiennent des scores très élevés, démontrant une excellente capacité discriminative sur ce dataset.

Chapitre 6

Conclusion

Le projet a permis de construire des modèles très performants pour prédire la maladie rénale chronique à partir de données cliniques. Les variables les plus discriminantes sont l'hémoglobine, la créatinine sérique et le glucose sanguin.

Les modèles Random Forest et Gradient Boosting se distinguent par leurs performances parfaites sur l'ensemble de test.

Ce travail démontre une bonne maîtrise des étapes classiques d'un projet de machine learning : exploration, prétraitement, modélisation et évaluation.

Réalisé par : Randriamamonjy Tokiniaina