

Chronic Kidney Disease (CKD) Classification Using Machine Learning

1. Introduction

Chronic Kidney Disease (CKD) is a long-term medical condition characterized by a gradual loss of kidney function over time. Early detection is critical because CKD often progresses silently and may only become clinically apparent at advanced stages, when treatment options are limited and costly. The objective of this capstone project was to build and evaluate machine learning models capable of classifying whether a patient has CKD based on routinely collected clinical and laboratory features.

This project applies Python-based data science and machine learning techniques to a structured healthcare dataset, with a focus on understanding the data, preparing it appropriately, training classification models, and evaluating their performance in a medically meaningful way.

2. Problem Statement and Objectives

Problem Statement

Diagnosing CKD using traditional clinical methods can be time-consuming and may depend heavily on expert interpretation of multiple test results. An automated classification system can support clinicians by providing faster, data-driven predictions.

Objectives

The main objectives of this project were:

- To explore and understand a CKD dataset containing both numerical and categorical clinical variables.
 - To clean and preprocess the data to make it suitable for machine learning.
 - To build and compare two different classification models for predicting CKD.
 - To evaluate the models using appropriate performance metrics and interpret the results.
-

3. Dataset Description

The dataset used in this project contains patient-level medical information relevant to kidney function. It includes a mix of numerical and categorical features such as:

- Demographic information (e.g., age)
- Laboratory test results (e.g., blood and urine indicators)
- Clinical observations and conditions

The target variable is a binary categorical label indicating whether a patient has Chronic Kidney Disease or not.

4. Methodology

4.1 Data Exploration and Understanding

Exploratory Data Analysis (EDA) was performed to:

- Understand the structure and size of the dataset
- Identify numerical and categorical variables
- Detect missing values and inconsistencies
- Examine class distribution of the target variable

This step was necessary to gain insight into the data quality and to inform preprocessing decisions.

4.2 Data Cleaning and Preprocessing

Data preprocessing was a critical step due to the presence of missing values and categorical variables.

The following actions were taken:

- **Handling missing values:** Missing numerical values were handled using appropriate imputation strategies (such as mean or median), while categorical missing values were filled using the most frequent category.
- **Encoding categorical variables:** Since machine learning models require numerical inputs, categorical features were converted into numerical representations using encoding techniques.
- **Feature-target separation:** The dataset was split into input features (X) and the target variable (y).
- **Train-test split:** The data was divided into training and testing sets to ensure that model evaluation was performed on unseen data.

These steps were necessary to ensure model stability, reduce bias, and improve generalization performance.

5. Model Development

Two different classification models were implemented and trained on the processed dataset.

5.1 Model 1: Logistic Regression

Logistic Regression was used as the baseline classification model in this project. It was selected because it is a widely used and well-understood algorithm for binary classification problems, especially in the medical domain.

Logistic Regression estimates the probability that a patient has Chronic Kidney Disease based on a linear combination of the input features. Its strengths include simplicity, interpretability, and efficiency, making it suitable as a first model for clinical prediction tasks.

The model was trained on the training dataset and evaluated on the test dataset to establish a performance benchmark for CKD classification.

5.2 Model 2: Random Forest Classifier

The Random Forest model was implemented as the second classification approach. It was chosen to compare against Logistic Regression because of its ability to capture complex, non-linear relationships within the data.

Random Forest is an ensemble learning method that builds multiple decision trees and combines their predictions to improve accuracy and reduce overfitting. This makes it particularly effective for datasets with mixed numerical and categorical features, such as medical records.

The Random Forest model was trained using the same training dataset and evaluated on the same test dataset to ensure a fair and consistent comparison with the Logistic Regression model.

6. Model Evaluation and Results

The performance of both models was evaluated on the test dataset using standard classification metrics relevant to medical decision-making.

Evaluation Metrics Used

- **Accuracy:** Overall correctness of predictions
- **Precision:** Proportion of correctly identified CKD-positive cases among predicted positives
- **Recall (Sensitivity):** Ability to correctly detect patients with CKD
- **F1-score:** Harmonic mean of precision and recall
- **ROC-AUC:** Ability of the model to distinguish between CKD and non-CKD cases

Results Summary (Test Set)

Logistic Regression:

- Accuracy: **96.6%**
- Precision: **96.7%**
- Recall: **96.5%**
- F1-Score: **96.5%**
- ROC-AUC: **0.8259**

Random Forest Classifier:

- Accuracy: **96.4%**
- Precision: **96.8%**
- Recall: **96.0%**
- F1-Score: **96.4%**
- ROC-AUC: **0.7689**

Interpretation of Results

Both models demonstrated strong predictive performance, indicating that clinical and laboratory features are highly informative for Chronic Kidney Disease classification. Logistic Regression slightly outperformed Random Forest in terms of recall and ROC-AUC, suggesting that linear relationships in the dataset are particularly effective for distinguishing CKD cases.

In a healthcare context, recall is especially important because failing to identify a patient with CKD can delay treatment. The high recall values achieved by both models indicate that they are suitable for supporting early CKD detection.

7. Discussion

The comparison of the two models highlights the importance of model selection in healthcare applications. While simpler models offer interpretability and ease of implementation, more advanced models may provide better predictive performance.

Key observations from this project include:

- Data preprocessing significantly impacts model performance.
 - Handling missing values appropriately is crucial in medical datasets.
 - Evaluation metrics beyond accuracy are necessary for imbalanced and high-stakes classification problems like disease detection.
-

8. Conclusion

This capstone project demonstrated the application of machine learning techniques to the classification of Chronic Kidney Disease. By systematically exploring the data, preprocessing it, and comparing two classification models, the project achieved reliable predictive performance.

The findings show that machine learning can serve as a valuable decision-support tool in healthcare, particularly for early disease detection. Future work could include hyperparameter tuning, testing additional models, and validating the system on external datasets.

10. Tools and Technologies

- Python
- Pandas, NumPy
- Scikit-learn
- Matplotlib / Seaborn
- Jupyter Notebook