Chronic Kidney Disease Prediction

Project Description

This project uses machine learning to predict the likelihood of **chronic kidney disease** from patient medical data. By applying multiple classification algorithms to a clinical dataset, the model assists healthcare professionals with early identification and better decision-making.

Dataset

- **Source**: The dataset was sourced from the Kaggle Chronic Kidney Disease Dataset.
- **Contents**: It includes **400 patient records** with **25 features** and **1 target**

Programming Language & Libraries

This project was built using **Python** with the following key libraries:

Library	Purpose
pandas	Data manipulation and preprocessing
numpy	Numerical operations
matplotlib and seaborn	Data visualization
Scikit-learn (sklearn)	Model training and evaluation

Dataset Overview

- **Features**: 25 input features covering various clinical measurements.
- **Target Variable**: The class variable, where 1 = CKD and 0 = Not CKD.

Key Feature Categories:

- **Numerical**: age, blood_pressure, blood_urea, serum_creatinine, sodium, potassium, haemoglobin, etc.
- **Categorical**: red_blood_cells, pus_cell, pus_cell_clumps, bacteria, appetite, hypertension, diabetes_mellitus, coronary_artery_disease, pedal_edema, anemia.

Data Preprocessing Steps

The raw data was cleaned and prepared through these steps:

1. The id column was dropped.
2. Missing numerical values were filled with the **mean**, and missing categorical values were filled with the **mode**.
3. Inconsistent labels (e.g., \tyes, ckd\t) were cleaned and mapped to a consistent format.
4. String-based numeric columns were converted to a proper float data type.
5. A correlation analysis was performed to understand feature relationships.

Machine Learning Algorithms & Evaluation

All models were trained on a **75% training and 25% test split** of the dataset.

Algorithm	Accuracy	Precision	Recall	F1 Score	Confusion Matrix
Random Forest	98.0%	0.969	1.000	**0.984**	[[35, 2], [0, 63]]
Decision Tree	97.0%	0.969	0.984	0.976	[[35, 2], [1, 62]]
Support Vector Machine	95.0%	0.953	0.968	0.961	[[34, 3], [2, 61]]
Gaussian Naive Bayes	95.0%	1.000	0.921	0.959	[[37, 0], [5, 58]]
K-Nearest Neighbors	76.0%	0.882	0.714	0.789	[[31, 6], [18, 45]]

Best Model: Random Forest Classifier

The Random Forest model achieved the highest accuracy and a perfect recall score, which is critical for medical diagnosis to minimize **false negatives**.

Submitted by :
Hari prasath K
Deepakumar M