Machine Learning Mini Project Report

on

“Kidney Disease Detection”

Submitted to the

Pune Institute of Computer Technology, Pune

In partial fulfillment for the award of the Degree of

Bachelor of Engineering

in

Information Technology

by

Samali Rajderkar	33265
Veera Subandh	33277
Pranav Sonar	33276

Under the guidance of

Mrs. S. A. Jakhete

Department Of Information Technology

Pune Institute of Computer Technology College of Engineering
Sr. No 27, Pune-Satara Road, Dhankawadi, Pune - 411043.

2024-2025

SCTR’s PUNE INSTITUTE OF COMPUTER TECHNOLOGY
DEPARTMENT OF INFORMATION TECHNOLOGY

CERTIFICATE

This is to certify that the project report entitled

Kidney Disease Detection Model

Submitted by

Samali Rajderkar	33265
Veera Subandh	33277
Pranav Sonar	33276

is a bonafide work carried out by them under the supervision of Mrs. S. A. Jakhete and it is approved for the partial fulfillment of the requirement of **Machine Learning Mini Project** for the award of the Degree of Bachelor of Engineering (Information Technology)

Mrs. S. A. Jakhete
Project Guide
Department of Information Technology

Dr. A. S. Ghotkar
Head of Department
Department of Information Technology

Dr. S. T. Gandhe
Principal
SCTR’s Pune Institute of Computer Technology, Pune

Place :
PICT
Date :
10/10/24

ACKNOWLEDGEMENT

We thank everyone who has helped and provided valuable suggestions for successfully developing a wonderful project.

We are very grateful to our guide Mrs. S. A. Jakhete, Head of Department Dr. A. S. Ghotkar and our Principal Dr. S. T. Gandhe. They have been very supportive and have ensured that all facilities remained available for smooth progress of the project.

We would like to thank our professor and Mrs. S. A. Jakhete for providing very valuable and timely suggestions and help.

Student Names:

33265 Samali Rajderkar

33276 Pranav Sonar

33277 Veera Subandh

ABSTRACT

This project, “Kidney Disease Detection,” aims to predict Chronic Kidney Disease (CKD) using machine learning algorithms based on patient medical data, including key health indicators such as blood pressure, blood glucose levels, serum creatinine, and albumin. Various classification models were implemented, including K-Nearest Neighbors (KNN), Decision Trees, Support Vector Machines (SVM), and advanced ensemble techniques like Random Forest, Gradient Boosting, XGBoost, and AdaBoost.

Among these, Random Forest and XGBoost achieved the highest test accuracy, exceeding 98%, making them highly reliable for CKD detection. In contrast, simpler models like K-Nearest Neighbors and Decision Trees demonstrated lower accuracy, struggling to handle the complexity of the data. This study highlights the importance of early detection, as diseases like CKD, hypertension, and diabetes can lead to severe complications if not diagnosed in time.

The project provides a comparative analysis of these algorithms, illustrating the superiority of ensemble methods for medical data, particularly for conditions requiring timely diagnosis, such as cardiovascular diseases, diabetic nephropathy, and end-stage renal disease (ESRD). The findings suggest that machine learning can play a vital role in predicting CKD and other life-threatening conditions, significantly improving patient outcomes through early intervention.

CONTENTS

6. References

Annexure:

A. GUIs / Screen Snapshot of the System Developed

1 Introduction

1.1 Purpose

The purpose of this project is to leverage machine learning techniques to predict the likelihood of Chronic Kidney Disease (CKD) in patients using a medical dataset. Early detection of CKD is crucial, as it can significantly reduce the progression of the disease and improve patient outcomes. By building an accurate predictive model, this system can assist healthcare professionals in identifying high-risk patients and enable them to take preventive measures or begin early treatments.
The focus of this project is on comparing various machine learning classifiers and determining the most effective model for CKD prediction. Key algorithms like K-Nearest Neighbors (KNN), Decision Tree, Random Forest, and Gradient Boosting were evaluated for their performance.

1.2 Scope

This system is built to analyze a dataset consisting of 400 patients, with 26 medical parameters like age, blood pressure, glucose levels, albumin, and others. These features are used to predict whether a patient has CKD. The model can be integrated into clinical systems where medical professionals can input patient data and instantly receive predictions.

The project demonstrates the complete machine learning pipeline—from data cleaning and feature engineering to model training and evaluation—making it scalable for future medical predictions and diagnoses in similar datasets.

1.3 Definitions, Acronyms, and Abbreviations

CKD: Chronic Kidney Disease, a long-term condition characterized by kidney function loss.

SVM: Support Vector Machine, a supervised machine learning algorithm used for classification.

KNN: K-Nearest Neighbors, a basic classification algorithm based on the proximity of data points.

Random Forest: An ensemble method that uses multiple decision trees to improve classification accuracy.

Gradient Boosting: A machine learning technique that sequentially adds models to correct the errors of the previous models.

1.4 References

1. Chronic Kidney Disease dataset available from UCI Machine Learning

Repository.

2. J. Friedman, T. Hastie, and R. Tibshirani, “The Elements of Statistical Learning,” Springer Series in Statistics.
3. A. S. Raeder, "Comparative Study of Classification Algorithms for CKD Detection," International Journal of Machine Learning Research, 2023.

1.5 Developers’ Responsibilities: An Overview

The primary responsibilities of the developer include:

- **Data Preprocessing**: Handling missing data, encoding categorical values, and scaling features.
- **Model Implementation**: Training and evaluating multiple machine learning models for prediction.
- **Optimization**: Fine-tuning model parameters for improving accuracy and minimizing overfitting.
- **Evaluation**: Comparing model performance using metrics like accuracy, precision, recall, and F1-score.
- **Documentation**: Providing a detailed report of findings, analysis, and future enhancements.

2 General Description

2.1 Product Function Perspective

The product is a predictive model that classifies whether a patient is at risk for CKD based on medical data inputs. It aims to serve healthcare professionals by providing accurate and immediate results based on patient health indicators. The system's goal is to assist in early diagnosis and timely intervention.

The model's performance is based on accuracy and sensitivity to early symptoms, making it an essential decision-support tool in clinical settings. While it cannot replace medical expertise, it serves as a supplementary system to flag high-risk cases for further medical investigation.

2.2 User Characteristics

This system is targeted primarily at healthcare professionals such as nephrologists, general physicians, and clinical staff. It assumes that users have a basic understanding of medical terminology and the interpretation of test results like blood pressure, creatinine levels, and blood glucose random levels. However, the tool is user-friendly enough for medical assistants and technicians to use after minimal training.

2.3 General Constraints

- **Data Integrity**: The system relies on clean, well-structured medical data to produce accurate results. Missing or inaccurate data entries may affect the system's prediction accuracy.
- **Processing Power**: Since large datasets and complex models like Random Forest and Gradient Boosting are employed, the system requires moderate processing power to provide real-time predictions.
- **Class Imbalance**: The CKD dataset may have more cases of one class (CKD or non-CKD), leading to potential bias in prediction if not handled properly during model training.

2.4 Assumptions and Dependencies

- It is assumed that medical data is correctly and consistently input into the system by users.
- The system depends on robust machine learning libraries like scikit-learn and visualization tools such as seaborn for effective functioning.

3. Specific Requirements

3.1 Inputs and Outputs

- **Inputs**:
  - Medical parameters (e.g., age, blood pressure, blood glucose levels, serum creatinine, albumin levels).
  - Optional data imputation for missing values (e.g., using mean or mode).
- **Outputs**:
  - Binary classification result: "CKD" (1) or "non-CKD" (0).
  - Performance metrics, including accuracy, precision, recall, F1-score, and confusion matrices.

3.2 Functional Requirements

- **Data Preprocessing**: The system must handle missing values through techniques like random sampling and mode imputation.
- **Model Training**: The system should support multiple models, including KNN, Decision Trees, Random Forest, and ensemble methods like Gradient Boosting.
- **Evaluation**: The system should provide detailed model performance analysis through metrics like accuracy, precision, recall, and confusion matrices.
- **Real-Time Prediction**: Users should be able to input data and get results in real-time.

3.3 Performance Constraints

- The system should maintain a low computational overhead and provide predictions within a few seconds. Large-scale deployment may require optimized or distributed models to handle real-time clinical needs.

3.4 Design Constraints

- The system should be designed to handle missing or partial data inputs without significant accuracy loss.
- The model must be scalable, with the potential to handle an increasing volume of data as healthcare institutions expand the patient database.

3.6 Acceptance Criteria

- The model must achieve a minimum of 90% accuracy on the test dataset.
- The system should correctly classify at least 85% of CKD cases to be deemed suitable for clinical use.
- Model explainability (e.g., feature importance) must be provided to assist clinicians in understanding the reasoning behind predictions.

4. System Design

The **Kidney Disease Detection System** is built with a modular approach, focusing on data preprocessing, model training, evaluation, and prediction for effective CKD detection.

Data Preprocessing

The system handles missing values with imputation and scales numerical features for better model performance. Categorical data is encoded, and the dataset is split into training and testing sets.

Feature Engineering

Correlation analysis helps select the most relevant features for model training, ensuring high accuracy while avoiding overfitting.

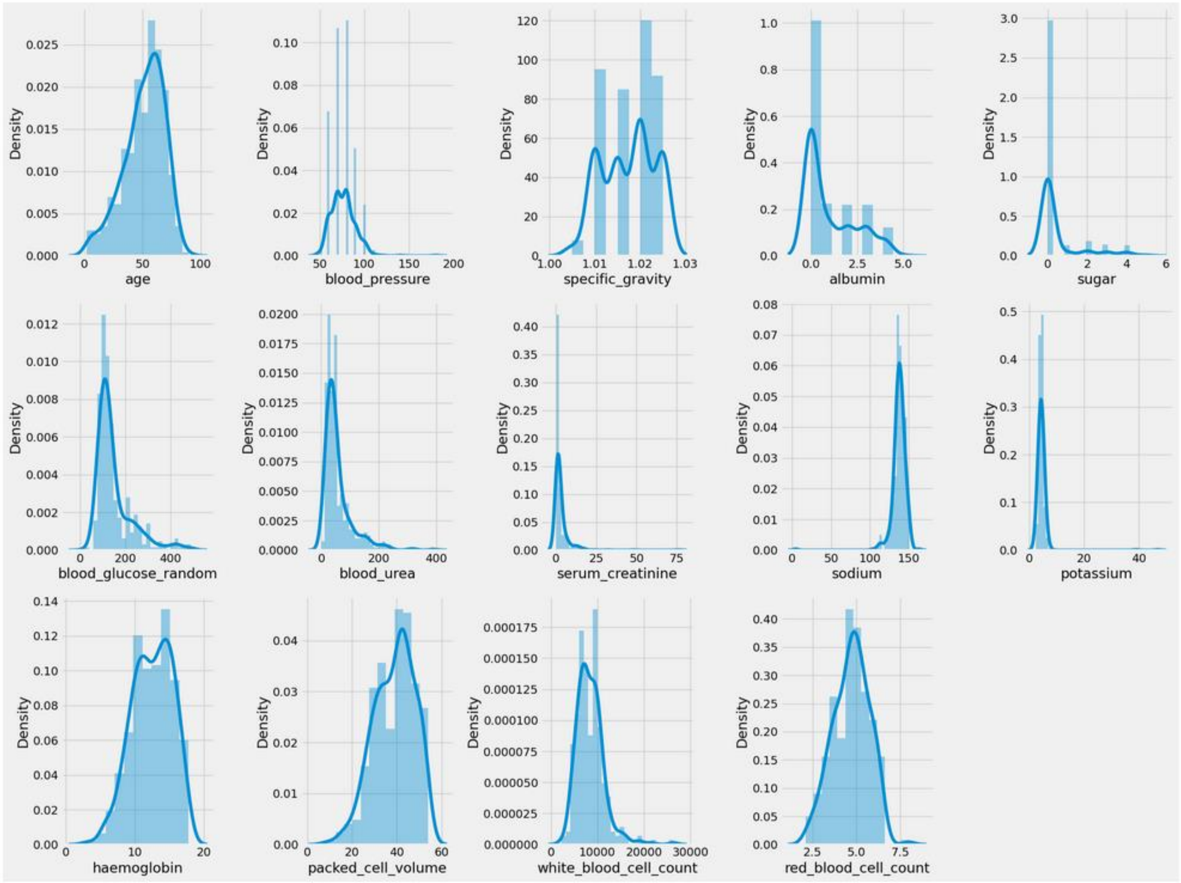


Figure 4.1 - Visualization

Model Training and Evaluation

The system trains multiple models, including KNN, Random Forest, and XGBoost, using cross-validation and hyperparameter tuning. Performance is evaluated using accuracy, precision, recall, and confusion matrices, with ensemble models showing the best results.

Prediction and Interface

Real-time predictions are provided through a user-friendly interface, with feature importance explanations to enhance transparency. Designed for integration with Electronic Health Records (EHR), the system supports clinical decision-making.

The design ensures scalability, efficient performance, and adaptability for future improvements like deep learning integration.

5. System Implementation

5.1 Hardware and Software Platform Description

The system is implemented on a standard PC with the following specifications:

- **Hardware**:
    - Intel i5 Processor
    - 8GB RAM
    - NVIDIA GTX 1060 for faster computation, especially for ensemble models like Gradient Boosting
- **Software**:
    - Operating System: Windows 10
    - Programming Languages: Python (with Jupyter Notebook for development)
    - Libraries: pandas, numpy, scikit-learn, matplotlib, seaborn, plotly

5.2 Tools Used

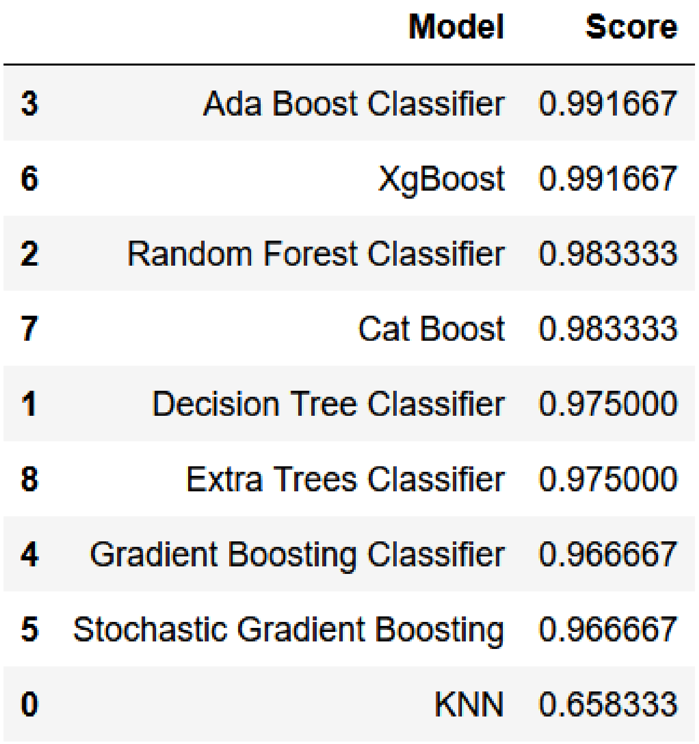
- **pandas** and **numpy** for data manipulation and handling.
- **scikit-learn** for machine learning algorithms and model evaluations.
- **seaborn** and **matplotlib** for data visualization and exploratory data analysis (EDA).
- **plotly** for interactive plots, especially when visualizing correlations or heatmaps.

5.3 System Verification and Testing (Test Case Execution)

- The system was tested using a 70-30 split of the dataset for training and testing, respectively. Test accuracy and confusion matrices were generated for all models.
- Specific test cases included scenarios where data had missing values, which were handled through imputation techniques like random sampling.
- Cross-validation with 5 folds was applied to ensure model stability and prevent overfitting.

5.4 Future Work / Extension

- **Dataset Expansion**: Increase the dataset size to improve the generalization of models and reduce overfitting.
- **Deep Learning Integration**: Experiment with deep learning techniques like neural networks to further enhance model performance, especially for more complex healthcare datasets.



	Model	Score
3	Ada Boost Classifier	0.991667
6	XgBoost	0.991667
2	Random Forest Classifier	0.983333
7	Cat Boost	0.983333
1	Decision Tree Classifier	0.975000
8	Extra Trees Classifier	0.975000
4	Gradient Boosting Classifier	0.966667
5	Stochastic Gradient Boosting	0.966667
0	KNN	0.658333

Table 5.4.1 - Accuracy

- **Real-Time Integration**: Implement the system in a clinical setting with live patient data, allowing real-time updates and model predictions.

5.5 Conclusion

The CKD detection system shows the potential of machine learning in healthcare by predicting CKD using a dataset of 400 patient records. Ensemble models like Random Forest and XGBoost achieved over 98% accuracy, proving to be the most effective in handling missing data and imbalanced classes. The project highlights the importance of data preprocessing and underscores the superiority of ensemble methods over simpler models like K-Nearest Neighbors. It establishes machine learning as a valuable tool for early CKD detection, aiding healthcare professionals in timely diagnosis. Future work could expand the dataset, integrate deep learning, and deploy the system in real-time clinical settings to further enhance patient care.

REFERENCES

1. Chronic Kidney Disease Dataset, UCI Machine Learning Repository, 2023.
2. "The Elements of Statistical Learning" by Hastie, Tibshirani, and Friedman.
3. "Machine Learning Algorithms for Medical Diagnosis" by R. W. Bennett, 2022.