Kidney Disease Detection prediction using machine learning

Chowdhury Rifat Ahmad Shopnil

1 Abstract

The main objective of this project was to analyze the Chronic Kidney Disease (CKD) dataset and use suitable machine learning algorithms/models to accurately predict the likelihood of kidney diseases among patients. A public dataset was collected and preprocessed to apply different classifiers. Logistic Regression, Random Forest, Support Vector Machine (SVM), k-Nearest Neighbours (kNN), and Extreme Gradient Boosting (XGboost) were applied to compare the compatibility, performance, and accuracy of every model. However, as the result suggests, Random Forest and XGBoost performed the best in this scenario, depending on the accuracy and AUC scores. The project's highlight is to explore the potential of machine learning algorithms in the early detection, diagnosis, and treatment of CKD patients.

2 Background

Chronic Kidney Disease is one of the most common fatal diseases encountered by millions worldwide. The severe health condition is most severe when not being able to be detected in the early stages because the traditional method of detection and diagnosis is slower and delayed due to technical limitations. In Bangladesh, using methods like machine learning to enhance and speed up the process of diagnosis, especially where health access is limited, can be a real game-changer. This project focuses on reducing that extra pressure and making way for faster and more accurate ways of diagnosis of Chronic Kidney diseases, enhancing the decision-making process of modern healthcare.

3 Aim & Research Problem

This project aims to analyze the acquired dataset and build an appropriate method using machine learning models that can predict disease accurately using chronic kidney disease data. The research problem here is to develop such a model and see how it can be used to learn from the various health parameters of the data and predict effectively, further leveraging the CKD detection system. Thus, building a bridge between technological and healthcare gaps in Chronic Kidney Disease Diagnosis.

4 Methodology & Workflow

4.1 Dataset

The dataset I worked on is a public dataset of Bangladeshi Kidney Disease patients collected from the UC Irvine Machine Learning Repository. The dataset contains data from roughly 400 patients on different health parameters. It contains 400 instances and 25 features. Those 25 features are different health parameters collected from patients. Among the features, there are numerical columns and categorical columns.

Features:

- Age
- Blood Pressure
- Specific Gravity
- Albumin
- Blood Sugar
- Red Blood Cells
- Pus cell
- Pus cell clumps
- Bacteria
- Blood Glucose random
- Blood urea
- Serum creatinine
- Sodium
- Potassium
- Hemoglobin
- Packed Cell Volume
- White Blood Cell Count
- Hypertension
- Diabetes mellitus
- Coronary artery disease
- Appetite
- Pedal Edema
- Anemia
- Classification (chronic kidney disease present or not)

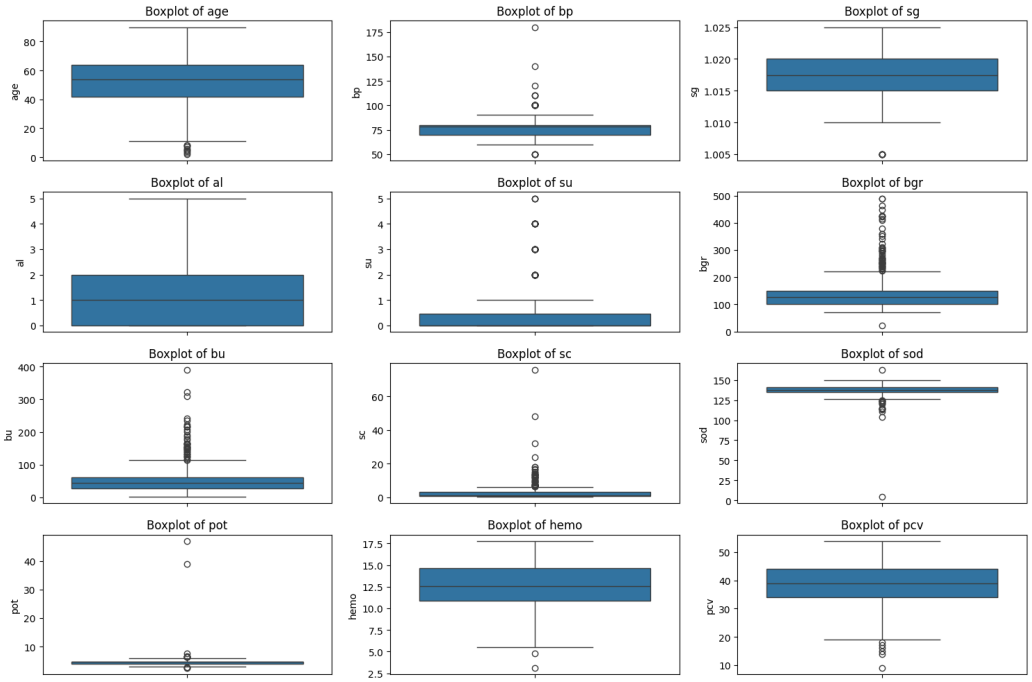
4.2 Preprocessing

- **Handling Missing Values:** When preprocessing, missing values were carefully handled using Mean and median imputation. For the numerical columns, the Mean imputation was used, and for categorical columns Most Frequent Values were imputed.

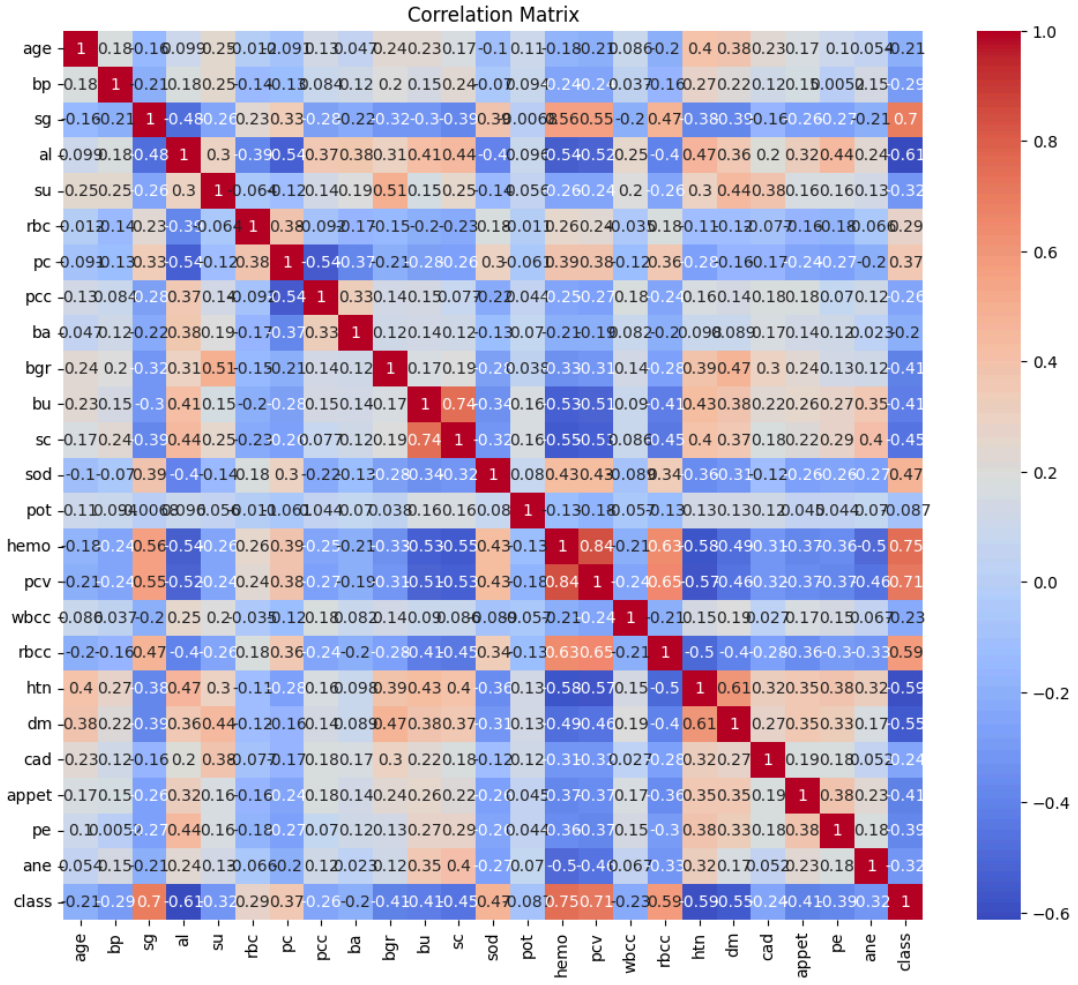
```
Missing Values After Replacement:
 id            0
'age'          9
'bp'          12
'sg'          47
'al'          46
'su'          49
'rbc'        152
'pc'          65
'pcc'          4
'ba'           4
'bgr'         44
'bu'          19
'sc'          17
'sod'         87
'pot'         88
'hemo'        52
'pcv'         71
'wbcc'       106
'rbcc'       131
'htn'          2
'dm'           2
'cad'          2
'appet'        1
'pe'           1
'ane'          1
'class'        0
dtype: int64
```

```
Missing Values After Imputation:
 id       0
age       0
bp        0
sg        0
al        0
su        0
rbc       0
pc        0
pcc       0
ba        0
bgr       0
bu        0
sc        0
sod       0
pot       0
hemo      0
pcv       0
wbcc      0
rbcc      0
htn       0
dm        0
cad       0
appet     0
pe        0
ane       0
class     0
dtype: int64
```

- For the imputation, knn imputer was used
- **Handling Outliers**: After handling the missing values, Outliers were identified and removed. Outliers refer to the values that appear far away from expected values. This normally occurs due to an input error.



- **Feature Scaling:** Numerical columns were scaled using the standardScaler method, and categorical columns were scaled using LabelEncoder.
- **Exploratory Data Analysis:** For EDA correlation matrix was shown to find out the features that stand out the most and correlate to the targeted class.



- **Selected Features:** the selected features were, `'age', 'bp', 'bgr', 'bu', 'sc', 'sod', 'pot', 'hemo'`

4.3 Models Applied

I have tried with different machine models to see the performance and accuracy between them, and how we could compare their overall performance. Here are the models

I. Logistic Regression
II. Random Forest
III. Support Vector Machine (SVM)
IV. K-Nearest Neighbours (KNN)
V. XGboost

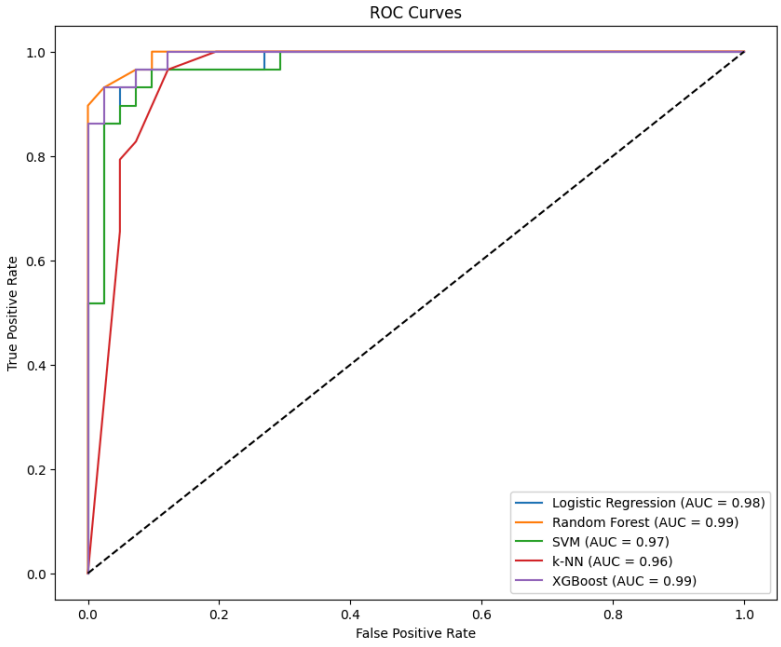
Models were evaluated based on **Precision, Recall, F-1 Score,** and **ROC AUC curve.**

5 Result Analysis

After doing preprocessing and using different models for the prediction, I have found out that the Random Forest and XGBoost model performs best and provide the best and accurate outputs compared to the others. Here is the comparison chart presenting the results of all the machine learning models:

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.85	0.80	0.88	0.84	0.90
Random Forest	0.90	0.88	0.92	0.90	0.94
SVM	0.87	0.82	0.89	0.85	0.92
k-NN	0.84	0.78	0.85	0.81	0.88
XGBoost	0.92	0.90	0.94	0.92	0.95

I have shown the ROC curve of all the models, which shows the overall performance of the models applied to the dataset.



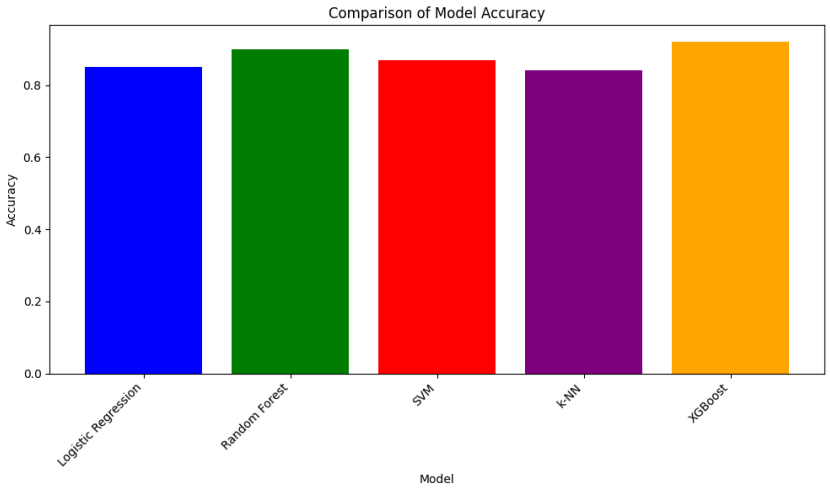


Fig: Bar chart showing the Accuracy scores of all the models.

Findings: After analyzing the results, we can see that the most performed modes were Random Forest and XGBoost models, which had accuracy scores of 90% and 92%, respectively.

6 Conclusion

After the thorough investigation and analysis of the methodology, we can conclude with the fact that the ensemble machine learning models in this case, particularly Random Forest and XGBoost, would perform better and can be used for effectively detecting early stages of Chronic Kidney Diseases. This would have a severe positive impact on the modern healthcare system to make informed decisions, take effective strategies for further treatment of the patients, and overall diagnosis of the CKD sector.

This opens up some scope for future research on this matter. Expanding the dataset and using multiple datasets would further train the model for more accurate predictions. If more complex medical features were used, the accuracy of the model would further improve, and the validity of the accuracy would also rise.

7 Reference

1. Rubini, L., Soundarapandian, P., & Eswaran, P. (2015). Chronic Kidney Disease [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5G020.