Forecasting Chronic Kidney Disease with Machine Learning: Predictive Analytics Approach

Avishek Roy Sparsho [1], Anik Khan [1]

1. School of Data Science, Brac University, 66 Mohakhali, Dhaka-1212

Abstract- Chronic Kidney Disease (CKD) is a long-term condition that develops gradually and can have serious consequences if not treated. It typically requires ongoing medical care, such as kidney replacement surgery or dialysis, to manage and can be fatal in its advanced stages. Early detection and treatment of CKD can help prevent or slow the progression of the disease. Four machine learning techniques were evaluated using a chronic kidney disease dataset: Logistic regression (LR), decision tree, random forest, k-nearest neighbors (KNN), and Naive Bayes classifiers. The goal was to identify the most effective method for predicting chronic kidney disease by comparing the performance of these models.

Keywords - Prediction; kidney disease; machine learning;

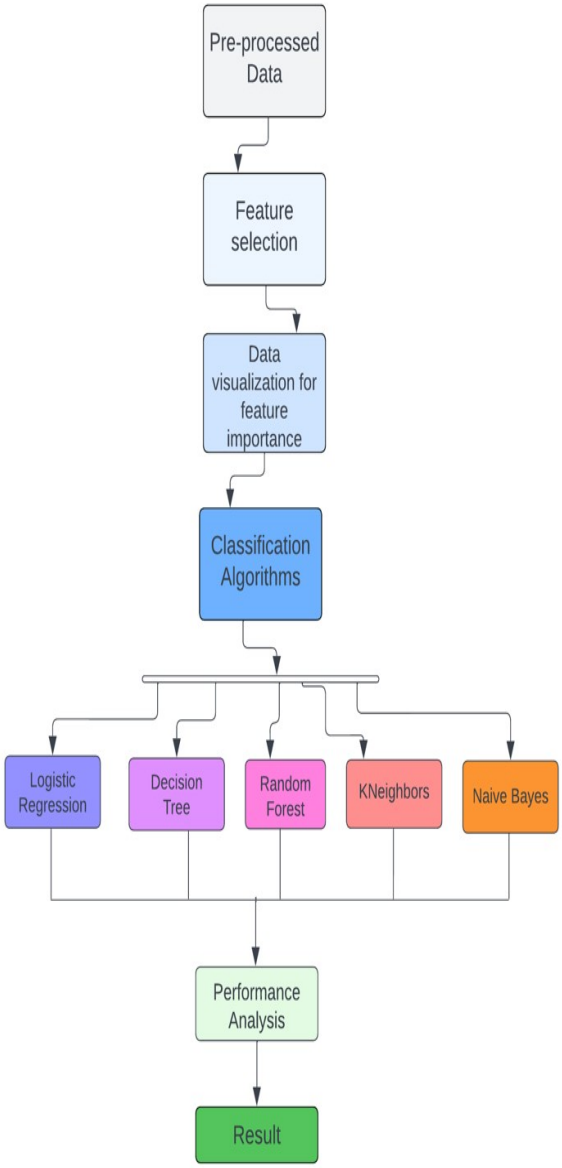
I. INTRODUCTION

Chronic kidney disease (CKD) is a condition in which the kidneys are damaged and cannot filter blood as effectively as healthy kidneys. This can lead to the build-up of waste products and excess fluids in the body, which can cause a range of health problems. CKD is typically a progressive disease, meaning it gets worse over time. It is often a result of other conditions such as diabetes or high blood pressure, which can damage the kidneys. It can also be caused by infections, inherited conditions, or the use of certain medications. Symptoms of CKD may not appear until the disease is advanced. In this paper, we present machine learning models for predicting chronic kidney disease (CKD), including decision tree, logistic regression, K-nearest neighbors, Naive Bayes, and random forest models. These models use machine learning techniques to analyze data and make predictions about the likelihood of a person developing CKD. The paper is structured into several sections. The section II provides an overview of the research method and explains the concepts used in the study. Section III presents the results of the research. The final section offers a conclusion based on the findings of the study.

II. METHODOLOGY

In this research, a workflow diagram is used to illustrate the stages of chronic kidney disease prediction. The diagram will represent the steps that will be taken in the study to predict the stages of the disease. The purpose of the diagram is to provide a visual representation of the research process and to help readers understand the sequence of events that will be followed. The performance of several classification algorithms will be compared in this proposed method in order to forecast the phases of chronic kidney disease. These algorithms include K Nearest Neighbors, Naive Bayes, decision trees, random forests, logistic regression, and decision trees. Comparing these algorithms is done in order to ascertain which one, depending on the study's data is best at forecasting the disease's phases. The comparison's findings will help guide further

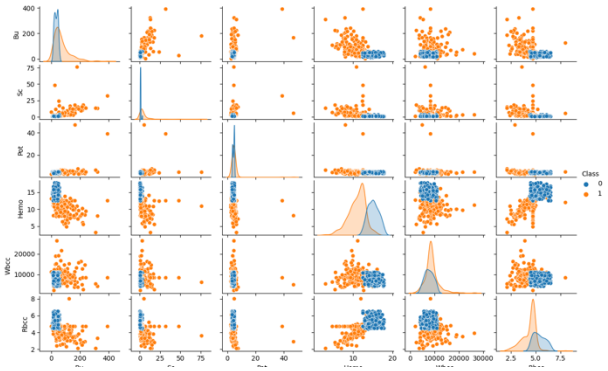
study on the subject and might shed light on the best strategy for identifying the phases of chronic kidney disease.

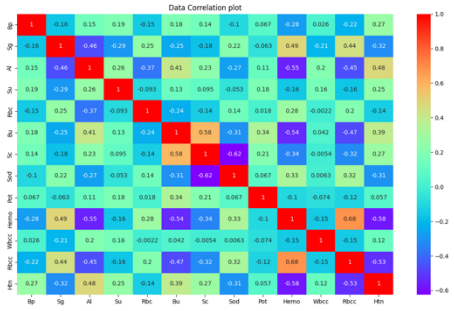


2.1. Data mining and preprocessing

The dataset used for this research was a chronic kidney disease dataset consisting of 14 attributes. The dataset was obtained from Kaggle [1] and had already been processed, so no additional data was required. No values needed to be removed or replaced and no columns or rows needed to be dropped. This allowed researchers to focus on the analysis of the data rather than the preparation of the dataset.

Table: Variable Description Used in Analysis





Sl No.	Attribute Symbols and Description	Distinct Values of Attribute
1.	**Bp -** Blood pressure	Multiple values between 50 & 180

2.	**Sg –** specific gravity	Multiple values between 1.005 & 1.02
3.	**Al-** albumin	Multiple values between 0 & 5
4.	**Su-** sugar	Multiple values between 0 & 5
5.	**Rbc -** Red blood cell	0,1
6.	**Bu-** Blood Urea	Multiple values between 1.5 & 391
7.	**Sc –** Serum Creatinine	Multiple values between 0.4 & 76
8.	**Sod -** Sodium	Multiple values between 4.5 & 163
9.	**Pot-** Potassium	Multiple values between 2.5 & 47
10.	**Hemo -** Hemoglobin	Multiple values between 3.1 & 17.8
11.	**Wbcc –** White Blood Cells	Multiple values between 2200 & 26400
12.	**Rbcc –** Red Blood Cells	Multiple values between 2200 & 26400
13.	**Htn-** Hypertension	0,1
14.	**Class**	0,1

2.2. Data Splitting & Visualization

For the purpose of this study, the dataset was divided into two segments: a training set and a testing set. The training set consisted of 70% of the dataset, while the remaining 30% was used as the testing set. The training set was used to

build the model, while the testing set was used to evaluate the model's performance. This common practice allows researchers to assess the accuracy and reliability of the model.

Logistic regression: Logistic regression is a statistical technique used to predict the probability of a binary outcome, such as the presence or absence of disease. In the context of this study, it could be used to predict the probability that a person has chronic kidney disease based on certain risk factors. This method is particularly well-suited for classification tasks, as it can predict categorical dependent variables based on a set of independent variables. In this case, the dataset contains 13 independent variables, making logistic regression a suitable choice for this analysis. It works by using the logistic function to transform the output of a linear equation into a value between 0 and 1, which represents the probability of the binary outcome.

Decision Tree Classification: Decision tree classification is a machine-learning technique that involves building a model in the form of a tree structure, with the root node at the top and the leaf nodes at the bottom. The algorithm works by dividing the data into smaller subsets based on the feature that results in the greatest reduction in impurity, creating branches that lead to decision nodes. At each decision node, the algorithm further divides the data into smaller subsets until it reaches a leaf node, at which point it predicts the class of the data point based on the majority class in that leaf node. This algorithm can be effective for predicting chronic kidney disease because it is resistant to extreme values and missing values, which may be present in a medical dataset. It is able to handle both numerical and categorical data.

Random Forest Classification: Random forest classification is a machine-learning method that involves creating a forest of decision trees, each of which is trained on a randomly selected subset of the data. The final prediction is based on the class of the majority of the decision trees in the forest. This algorithm can be used to predict whether or not a person has chronic kidney disease based on their risk factors. It is particularly well-suited for this task because it is able to handle categorical data, such as the Bu, Sc, Hemo, and Wbcc values present in the dataset. The accuracy of the algorithm increases with the number of trees in the forest, making it a robust and reliable method for predicting chronic kidney disease.

K-nearest neighbors: K-nearest neighbors (KNN) is the classification method for classifying unknown examples by searching the closest data in pattern space.KNN predicts the class by using the Euclidean distance defined as follows:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

The Euclidean distance is used to determine the distance between data points in the pattern space (a space defined by the features of the data) in order to identify the k data points that are closest to a new, unknown data point. The class label of the unknown data point is then determined by a majority voting process, which involves counting the number of times each

class label appears among the k nearest neighbors and assigning the label that appears most frequently to the unknown data point. **Naive Bayes classification:** Naive Bayes classification is a machine learning algorithm that uses Bayes' theorem to make predictions about the likelihood of a specific class based on the presence or absence of certain features. It is often used for text classification tasks, but it can also be applied to other types of classification problems, including the prediction of chronic kidney disease. This algorithm is known to perform well on datasets with a large number of features, which may be the case in a prediction task involving chronic kidney disease. It is a probabilistic classifier that makes predictions based on the probability of an object belonging to a particular class. One of the advantages of Naive Bayes is that it is simple and effective, making it a good choice for building fast machine-learning models that can make quick predictions.

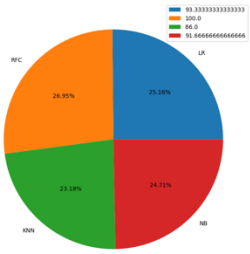
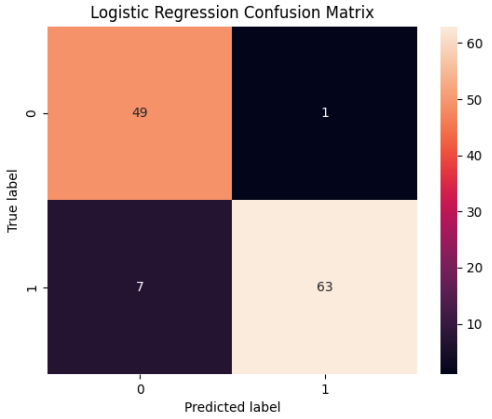


Figure:1

Figure 1 shows the average accuracy of four classifiers and the highest accuracy among them. From the experimental results, it can be seen that the random forest classifier gives the highest accuracy than the others with 99.16% while Logistic, Decision Tree, and KNN can produce average accuracy of 94.16%, 97.5%, and 86.0% respectively. The accuracy of each class is also important because if the classifier predicts incorrectly, it may be a detriment to the patient. Therefore, the sensitivity and specificity value are used in the experiments for evaluating the performance of the proposed methods.

III. RESULT ANALYSIS

In the experiments, the performance of five different machine learning techniques were evaluated and compared. The training and testing of these techniques were carried out using the proposed method.



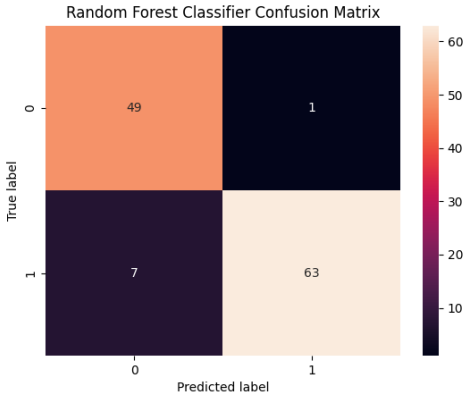
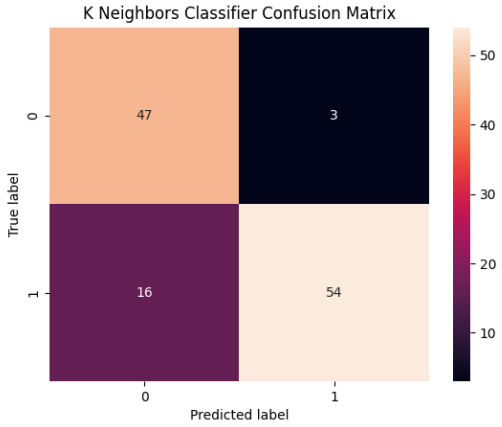


Figure: 2

Figure 2 shows the confusion matrix of Logistic regression and decision tree classifiers.



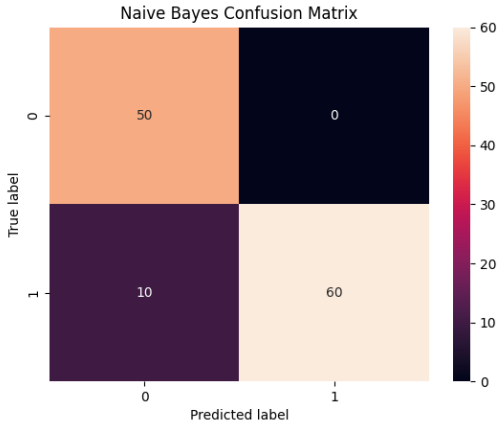


Figure: 3

Figure 3 shows the confusion matrix of KNN and Random forest classifiers.

IV. CONCLUSION

The result from this project suggests that the random forest classifier, decision tree classifier, and logistic regression can be used to predict chronickidney disease more precisely. Their accuracy is 99.16 percent, 94.16%, 97.5%, and 86.0% percent respectively. By developing a web application that uses these build on this work. This will enhance outcomes and increase the precision and effectiveness of healthcare professionals' ability to identify kidney diseases. It is hoped that this will encourage people to make positive changes in their life and seek early treatment at minimal cost and risk for chronic kidney disease.

References

1. "Chronic kidney disease." *Kaggle*, https://www.kaggle.com/datasets/abhia1999/chronic-kidney-disease Accessed 28 April 2023.

methods and a larger dataset than the one used in this study, future research may