Early Detection Of Chronic Kidney Disease (CKD) Using Machine Learning Algorithms and Stacking Model

Submitted By

Susmoy Kormoker

ID: 213-35-762

Department of Software Engineering

Daffodil International University

Supervised By

Dr.Md.Fazla Elahe

Assistant Professor and Associate Head

Department of Software Engineering

Daffodil International University

This report is presented in partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering.

Summer – 2025

APPROVAL

APPROVAL

This thesis titled on "**Early Detection Of Chronic Kidney Disease (CKD) Using Machine Learning Algorithms and Stacking Model**", submitted by **Student Name: Susmoy Kormoker (ID: 213-35-762)** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and approval as to its style and contents.

BOARD OF EXAMINERS

Fazla Elahe

Chairman

Dr. Md. Fazla Elahe
Assistant Professor & Associate Head
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Marzia

Internal Examiner 1

Dr. Marzia Ahmed
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

8/13/09/2025

Internal Examiner 2

Dr. Shabnom Mustary
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

13.09.25

External Examiner

Mohammod Abul Kashem
Professor
Department of Computer Science and Engineering
Dhaka University of Engineering & Technology, Gazipur.

DECLARATION

Declaration

I acknowledge that I have done this thesis under the supervision of Dr. Md. Fazle Elahe, Assistant Professor and Associate Head, Department of Software Engineering, Daffodil International University. I also assert that this thesis is my original work for the degree of B.Sc. in Software Engineering and that neither the whole work nor any part has been submitted for another degree in this or any other university.

Submitted by:

Susmoy

Susmoy kormoker

ID: 213-35-762

Department of Software Engineering

Daffodil International University

Certified by:

Fazla Elahe

Dr. Md. Fazle Elahe

Assistant Professor & Associate Head

Department of Software Engineering

Daffodil International University

ACKNOWLEDGEMENT

First, I express my heartiest thanks and gratefulness to almighty Allah for the divine blessing that makes it possible to complete the final year thesis successfully.

Then I am really grateful to my research supervisor, **Dr. Md. Fazla Elahe,** who guided me throughout the whole research activities.

I wish to express my special thanks to **Dr. Imran Mahmud**, Head of the Faculty, for providing all the necessary facilities for the research purpose. I am also thankful to all the lecturers in, Department of Software Engineering who sincerely guided me through my difficulties. I am thankful to my friend who supported me throughout this venture.

Finally, I must acknowledge with due respect the constant support of my parents.

Abstract

CKD is a long-term disease that impacts a lot of people all over the world, including in Bangladesh, and results in high illness and death rates. CKD is an illness that peaks in a slow reduction day by day. Early detection is vital since timely treatment may halt the advancement of CKD, improve the quality of patients lives, reduce medical costs, and prevent the danger of future health issues. In the past ten years machine learning models have appeared as transformative tools in medical testing, Utlizing large data and difficult algorithms to detect symptoms unknown by the doctor. This research project implemented and accessed several machine learning techniques for detecting chronic kidney disease early. Targeting on their respective efficiency, assets, and limitations. Machine learning model converts health assessment by applying large data and cutting-edge algorithms to detect structures that possibly alternatively bypass healthcare professionals. This research project used the dataset from the UC Irvine Machine Learning Repository created by the University of California in 1987, and this dataset is publicly available in their website. In this research evaluator evaluated and differentiated the performance of Decision Trees, Logistic Refression, and XGboost. All the model are using k-fold cross validation for spiting the data into training and test part. A stacking model also implemented in this research with the combination of Decision Trees, Logistic Regression and XGBoost. Analysis result showed that stacking model named CKDML762 had the best performance generating a perfect accuracy, precision, recall, f1 score. This outcome indicates that the stacking model CKDML762 correctly stored all the cases in training and testing sets. Integrating machine learning algorithms into the CKD holds great potential changing clinical work. Medical experts can imrpove testing accuracy and facilitate timely measures by utilizing proposed algorithm such as CKDML762

Keywords: CKD, chronic kidney disease, CKDML762, stacking model, decision trees, logistic regression, XGBoost, 5-fold cross-validation, ROC.

Table of content

List of Figures

List of Tables

Chapter 1

1.0 Introduction

CKD (chronic kidney disease) is a serious health issue all over the world. If this disease is not treated timely, it can lead to kidney damage and other harmful difficulties. CKD is one such disease that is very common in the world and impacts millions of people in Bangladesh, causes to a lot of illness and death. Recognizing CKD early can help with early treatment, that would be able to improve the quality of life of the patient, slow the disease down substantially, and save healthcare expenses [1]. kidenys filters over water and waste harmful cells of body. The disease is long-term since the damage of the kidney is very slow. [2]. During CKD disease heart can be also face some issue. CKD generally happens beacuse of diabetes, high blood pressure, heart disease and also depends on age and gender [3]. You understand CKD symptoms if you have such as abdominal discomfort, back pain, diarrhea, fever, nosebleeds, rash or vomiting these kind of issues. CKD is increasing day by day with the rate 6.23 percent per year which is paincable[4]. CKD can be identified using a few tests: (i) estimated glomerular filtration rate (eGFR). (ii) Urine test. (iii) Blood pressure. Generally these process takes time and sometimes people not able to recognize they bearing CKD disease. For that reason someone's kidney can damage badly and may causes the deadth. So needs to find out CKD early and gets the proper treatment within early stage. By the uses of machine learning algoritms we can detect CKD at the early stage, which is crucial for this disease. Machine learning techniques is the new invention of detecting or predecting disease in medical fields. This research study compares machine learning algorithms for early detection of CKD, analyzing their accuracy, strengths and shortcomings.

1.1 Background & Motivation

[5] Shows that, CKD is a disease that is associated with persistent losses of kidney functionality. The kidneys, which filter excess fluids and waste from the blood, can gradually lose function due to factors like hypertension, diabetes, and genetic susceptibility. Early diagnosis is important to avoid development of diseases, better patient outcomes, and lowering healthcare expenditures. The World Health Organization (WHO) predicts that millions of individuals worldwide are affected by CKD. However, due to its clinical condition in its initial phases, it often goes misdiagnosed [1].

[6] Indicate that machine learning models have transformed medical diagnosis by utilizing massive data and advanced algorithms to identify trends that clinicians and physicians may not be aware of. Using machine learning for healthcare purposes can improve disease diagnosis accuracy and efficiency, including CKD also. These procedures combine blood tests, case histories, and demographic information to identify at-risk

individuals with greater accuracy than previous methods. Implementing machine learning to detect CKD early is a promising way to enhance patient care.

1.2 Problem Statements

Actually, machine learning algorithms are the future of every field, including the medical field, but early CKD detection remains tough. Early detection of symptoms in CKD is challenging due to its multifaceted nature and complex pathophysiology. Early stages of CKD are often asymptomatic, leading patients to be unaware of the condition until considerable kidney damage occurs [7]. Conventional diagnostic tests are sensitive to variability and late detection of decrease in kidney function. As a result, not all traditional test methods are capable of providing timely detection.

[8] indicates that machine learning models provide a variety of solutions in the healthcare arena. Although they overcome their obstacles. These factors include data amount and quality, feature selection, and model interpretation. Different machine learning approaches provide varying levels of performance, computing expenses, and difficulty. An algorithm may excel on certain datasets but fall short on others. There is need to provide a comparative analysis of machine-learning methods in order to find out the best methods of CKD detection at the early stages.

Chapter 2

2.0 Literature Review

One of the documents defines chronic kidney disease (CKD) as a universal health condition that afflicts millions of individuals worldwide, including those in the United States, placing a significant burden on healthcare systems [16]. Early detection of CKD is crucial for preventing disease progression and reducing consequences, including renal failure. Early detection of chronic kidney disease is difficult due to its mild and silent character. Nowadays machine learning (ML) can improve diagnostic accuracy by analyzing complex patient data patterns. This literature review covers current techniques of diagnosing CKD, the importance of machine learning in medical diagnostics, and research on machine learning applications for CKD detection. It concludes with an analysis of knowledge gaps.

The creatinine level in the blood is one of the most important indicators of kidney function. Creatinine is basically a waste part of the body that is cleaned by the kidney. Serum creatinine is structured by body mass, sex, weight, and age. When kidney function declines, serum creatinine levels rise. Although this is a routine test, the sensitivity has significant limitations [17].

GFR is the estimation of the ability of the kidneys to filter blood which is the normal amount of blood to be filtered by the kidneys per minute. GFR test done by serum creatinine score, age, gender, and body weight. While GFR is more reliable than creatinine in determining kidney function, it may have limits in the early stages of CKD. GFR can vary, and small decreases in renal function may not be clinically obvious until extensive damage has occurred [18].

Urine is one of the most common tests during the kidney disease prediction. Because this test process is easier than other processes and cost-effective. But it also has some early detection issues found by [18].

Medical diagnostics now rely heavily on machine learning (ML), a type of artificial intelligence [5]. ML algorithms excel at analyzing large amounts of data to identify patterns and make predictions based on complicated relationships. This is especially important in healthcare, as medical data can be complex and multivariate. ML may analyze clinical, biochemical, and genetic records to uncover similarities associated with specific diseases, resulting in better specialized diagnostic tools. Machine learning is capable of enhancing the diagnosis and management of diseases to offer more precise and efficient as well as flexible solutions.

The importance of using machine learning algorithms for kidney disease diagnosis because it's strategic and imperative [6]. Machine learning algorithms can effectively treat CKD, a complex disease with multiple risk factors, including age, diabetes, hypertension, and genetics. ML models can identify early warning signals of CKD by analyzing big datasets such as patient records, laboratory test results, and clinical histories, which are difficult to detect by standard diagnostic methods. Machine learning models may combine data from gadgets like watches and home surveillance systems to deliver real-time updates on kidney health and identify changes over time.

A lot of research has been done based on early detection of CKD. Many ML algorithms have been used, like SVM, KNN, RF, decision trees, and linear regression.

[19] Used machine learning approaches, including Random Forests and Support Vector Machines, to predict CKD in 400 patient datasets from the United States. This study reported that the Random Forest algorithm yielded a very good performance with an accuracy rate of 97%. The scientists found that Random Forests can manage imbalanced datasets and provide feature relevance rankings for physicians to better understand CKD risk factors.

Our research is based on machine learning studies for early diagnosis of chronic kidney disease. [9] utilized RFE, Chi-Square, KNN, ANN, SVM, NB, and logistic regression for feature selection. Chi-Square-selected features in logistic regression led to a maximum accuracy of 98% [9]. [10] reached 96.5% accuracy in predicting CKD using SVM. [11] use support vector machine (SVM) algorithm to forecast early CKD and achieved a 99 percent accuracy. [12] found that logistic regression outperformed other machine learning algorithms, with an AUC of 0.870. [13] employed transfer learning to accurately identify congenital kidney anomalies in kidney ultrasound pictures. [14] used SVM and decision tree approaches, with SVM reaching an accuracy of 96%. [15] used a decision tree algorithm on a diabetic dataset to predict CKD risk and achieved 92% accuracy.

[8] investigated a neural network model for predicting CKD in the USA using laboratory test results and patient demographics. The neural network accurately detected chronic kidney disease (CKD) with an AUC of 0.97. The researchers discovered that the deep learning models like the neural networks, can describe complex nonlinear interactions between variables. As a result, they are ideal for illness detection at an early stage.

[17] study found that SVM outperformed decision trees, k-NN, and other machine learning models in detecting CKD early. SVM had the highest accuracy, precision, and recall. SVM-based models are highly computational and often require substantial parameter adjustment for optimal performance. Feature selection was crucial for this model, as values from blood pressure and serum creatinine had a greater impact.

2.1 Research Gap

To date, numerous studies have been conducted on the basis of the early detection of CKD using machine learning algorithms. But it may have some causes or issues found at analysis of current paper. Understanding those issues can be very useful for future work & further improvement in this field.

In machine learning, generalization is the most common feature or implementation. The work criteria of generalization is how ML algorithms success in new data or unseen data, even the part of test data not only train data. In CKD prediction if the data is not in generalized form sometimes it will predict wrong value with new patient data [20]. Generalizing the features with proper diverse data can solve the issues.

In machine learning techniques, performance depends on data. If data quality is good, it will provide better performance. Otherwise the performance will be average or predict a bad way. So most of the cases datasets need to be diverse, complete, and standard. Especially when machine predicts dangerous diseases like CKD. The test report of a patient needs to be 100 percent accurate [20]. So it is important that in research data quality should be very clear and have real patient data throughout from hospitals and many healthcare centers.

Sometimes machine learning models let's say, SVM predict disease with only information of (yes/no). Then it is called the interpretability problem because without knowing any health issues reports like (bp, sc, al..) it is not trustworthy. Then those models called black boxes [20]. For that reason doctor's are not believe in those models if they can't provide the proper documentation report. Make sure that machine learning models are conscious about that matter and solve it properly.

Make sure that machine learning medical tools or machines are user-friendly and easy to use, especially for doctors. If machines are not easy to use or are not understood by the doctor, it will be so problematic because they can't use them properly. If this happens, then those machines have no value for medical purposes [20]. So future research must be careful about this and implement the model by following the rules for medical purposes.

Chapter 3

3.0 Methodology

3.1 Dataset Description

This research project uses a dataset from the UC Irvine Machine Learning Repository. It is one of the oldest and most respected institutes in the world, invented by the University of California in 1987. This dataset has 400 patient records, which are usable for detecting chronic kidney disease. This dataset collect from nearly 2 months of period. It has samples of 250 CKD and 150 no CKD. There are 25 features in this dataset among with 24 predicted features and 1 target feature (CKD or Not CKD). Let's familiar with these features:

Features No	Features	Category
F1	Age (yrs)	Numerical
F2	Blood Pressure (mm/Hg)	Numerical
F3	Specific Gravity	Numerical
F4	Albumin	Numerical
F5	Sugar	Numerical
F6	Blood Glucose Random (mgs/dL)	Numerical
F7	Blood Urea (mgs/dL)	Numerical
F8	Serum Creatinine (mgs/dL)	Numerical
F9	Sodium (mEq/L)	Numerical
F10	Potassium (mEq/L)	Numerical
F11	Hemoglobin (gms)	Numerical
F12	Packed Cell Volume	Numerical
F13	White Blood Cells (cells/cmm)	Numerical
F14	Red Blood Cells (millions/cmm)	Numerical
F15	Red Blood Cells: normal	Binary
F16	Pus Cells: normal	Binary
F17	Pus Cell Clumps: present	Binary

F18	Bacteria: present	Binary
F19	Hypertension: yes	Binary
F20	Diabetes Mellitus: yes	Binary
F21	Coronary Artery Disease: yes	Binary
F22	Appetite: poor	Binary
F23	Pedal Edema: yes	Binary
F24	Anemia: yes	Binary
F25	Chronic Kidney Disease	Target Variable

Table 3.1: Dataset Description

This dataset has 14 numerical features, 10 binary features, and 1 target feature. The target variable predict predict the CKD or Not CKD. If result the is 1, then yes [CKD] or the result is 0, then no [Not CKD].

3.2 Correlation of features

A dataset with proper meaningful features is very important for a good research project. It can handle the issues of the data and works at the final result. This research project has 25 features as discussed in previous sections and also has a proper relation with all features, which can make a good impact on the result. Let's show it with the use of a correlation matrix.

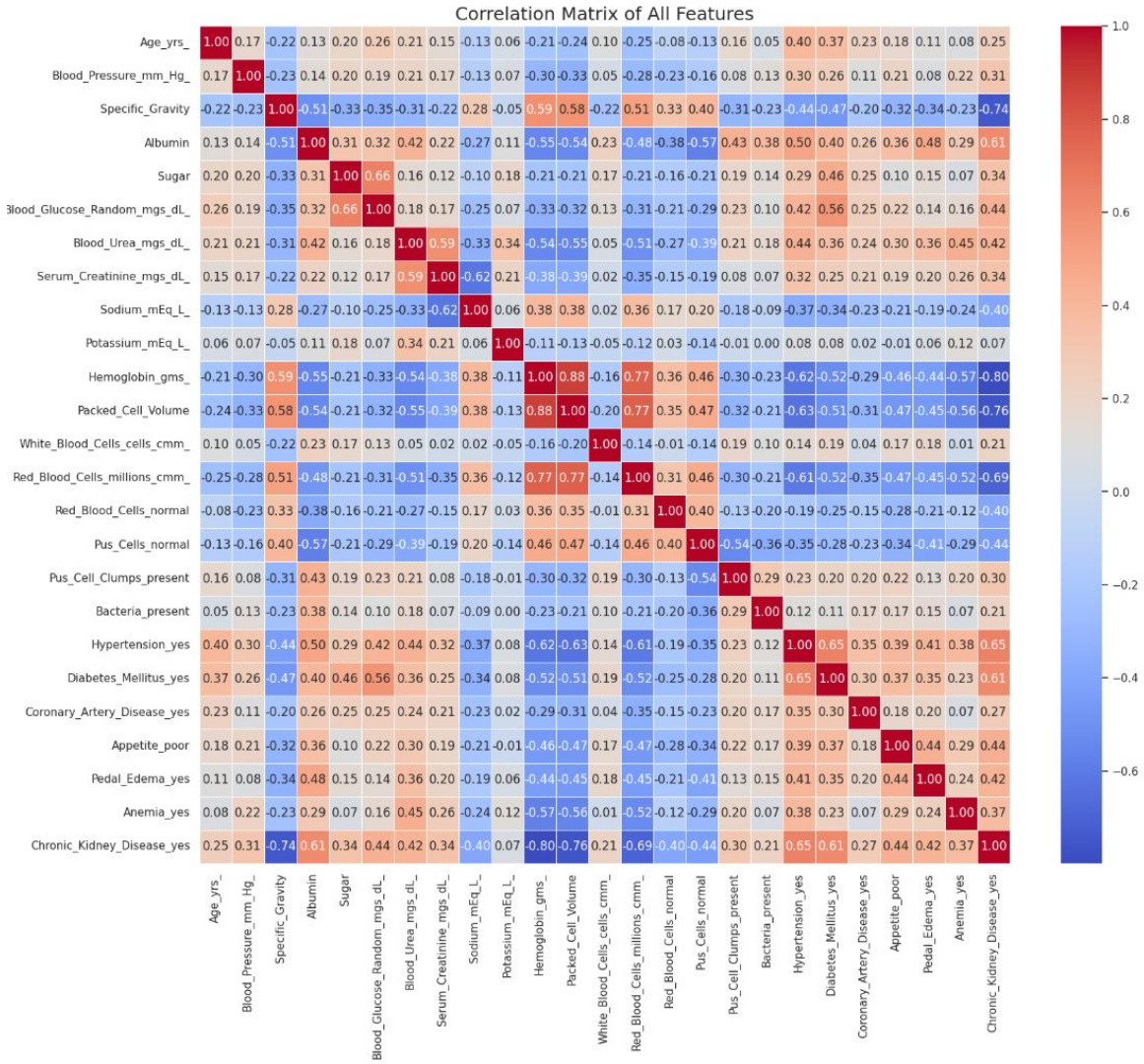


Figure 3.1: Correlation of all features in CKD datasets

3.3 Methodology Diagram (Possible Method Diagram)

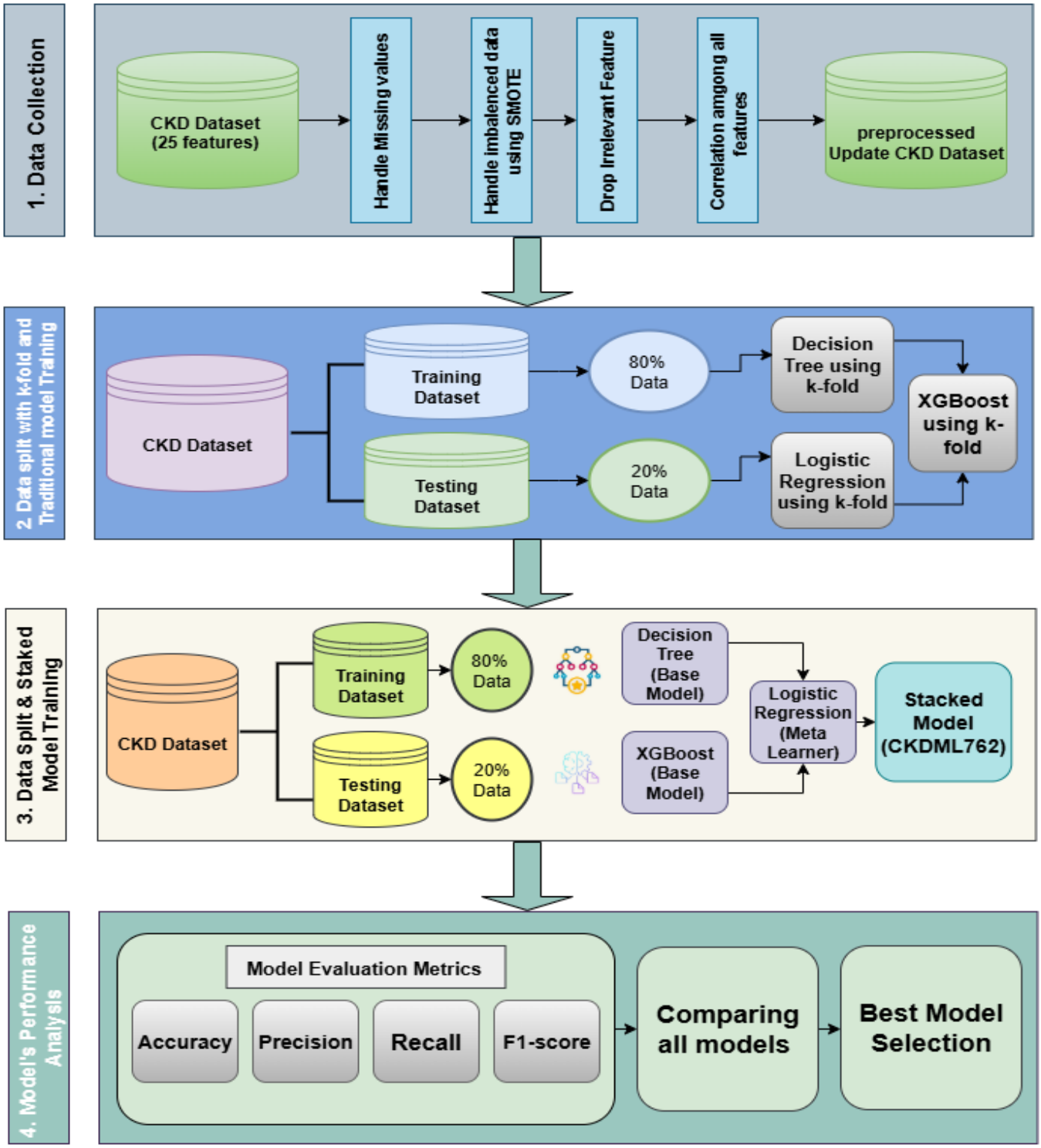


Figure 3.2: Methodology Diagram of CKD

3.4 Data Preprocessing

Most of the cases the values of the dataset are not preprocessed. But it is important to preprocess the data for understanding machine learning model well and providing the proper output as we want. This research dataset also have some issues which can bad impact on the result. To prevent those issues or problems data needed to be preprocessing data by using some techniques. So implement those techniques can be crucial for getting the proper outcome of CKD prediction. This research project uses some techniques and after that data are preprocessed and easy to understandable for machine learning models.

3.5 Handle Missing Values

This research project dataset has some missing values, which are not impactful for this research. So must needed to be handled the missing value. For numerical features handle missing values by the help of mean imputation, and for the categorical or binary features, handle missing values with the help of mode imputation. This research applies those techniques for both types of data and perfectly handles the missing value with Python code.

	Age (yrs)	Blood Pressure (mm/Hg)	Specific Gravity	Albumin	Sugar	Blood Glucose Random (mgs/dL)	Blood Urea (mgs/dL)	Serum Creatinine (mgs/dL)	Sodium (mEq/L)	Potassium (mEq/L)	...	Pus Cells: normal	Pus Cell Clumps: present	Bacteria: present	Hypertension: yes	Diabetes Mellitus: yes	Coronary Artery Disease: yes
0	48.0	80.0	1.020	1.0	0	121.000000	36.0	1.2	136.0	4.7	...	1	0	0	1	1	0
1	7.0	50.0	1.020	4.0	0	90.316581	18.0	0.8	140.0	4.0	...	1	0	0	0	0	0
2	62.0	80.0	1.010	2.0	3	423.000000	53.0	1.8	135.0	4.8	...	1	0	0	0	1	0
3	48.0	70.0	1.005	4.0	0	117.000000	56.0	3.8	111.0	2.5	...	0	1	0	1	0	0
4	51.0	80.0	1.010	2.0	0	106.000000	26.0	1.4	140.0	4.0	...	1	0	0	0	0	0

rows × 25 columns

Figure 3.3: Dataset after handle missing value

3.6 Handle imbalance data using SMOTE

When a dataset have categories by k fold cross validation or some other techniques it's must needed to be balanced. Otherwise the output will be same always like for CKD detection everytime detect Yes even if have no issue it kidney. Then the model will be recognize as a poor model. To recover this need use Synthetic minority over sampling technique (SMOTE), because by using this technique possible to make imbalance to balanced. This project dataset have some imbalacd data also, which are problematic to detect CKD correctly. By using SMOTE solve this issue so that machine learning models can perform well and detect CKD correct way.

3.7 Data Splitting

This project split the data for preprocessing with the implementation of k-fold cross validation. Consider k = 5 and divide the dataset into 5 parts, at every part, 4 sets are training data and 1 set is testing data. By doing this, the dataset is preprocessed initially with 80 percent of training data and 20 percent of testing data.

3.8 Model Selection

In the purpose of CKD detection need to fix machine learning model, so that model can detect CKD early. This project uses Decison tree using k-fold cross validation, Logistic regression using k-fold cross validation, XG boost using k-fold cross validation. Decision tree works like if else means branch by branch and moving parent nodes to leaf nodes. It has non linear relationship so that it can work well in medical purposes. Logistic regression have a sequece of output 0 to 1, it is using in medical purposes for a long period. XG boost is like update form of decison tree, cause many decison tree works together here. It is comaprely fast & give secure result which can be usable for predict CKD. In k-fold cross validation needs to divided dataset into 5 parts, 1 part will be testing part other will be remains as training part. So that the result gives proper output even if have issues like overfitting, imbalenced data these kind of issues. This research mix up 3 models using k-fold cross validation which is more suitable for prediction because it will cover every part of the dataset in all models individually.

3.9 Model Evaluation

The algorithms used in this research for detecting CKD are briefly discussed in this point. This process is done by different types of classifier algorithms, which are barely usable for detecting CKD. This project implements decision tree, logistic regression, XG boost, k-fold cross validation. These all are traditional machine learning algorithms and effectively work well in medical section. This project creates a stacking model named CKD762 with a combination of the top 3 models (Decision tree, Logistic regression, Gradient boosting classifer). So that the model can gives prediction more accurately with the combination form of those 3 models. It gradually works better based on the previous study.

3.9.1 Decision Tree

A decision tree is an excellent supervised learning model that is used properly for classification tasks [20]. For the purpose of CKD detection project this model is flexible and utilized for machine learning tasks. This one creates structure like a tree, with each leaf node connecting with parent nodes. It has procedures like entropy & the Gini index.

Entropy = - P (p1) $\log_2$ (p1) - P (p2) $\log_2$ (p2)
Gini Index = 1 - (p1/n) ² - (p2/n)²

By using these laws, one needs to find out information gain (IG). Higher information gain means it is better than other nodes.

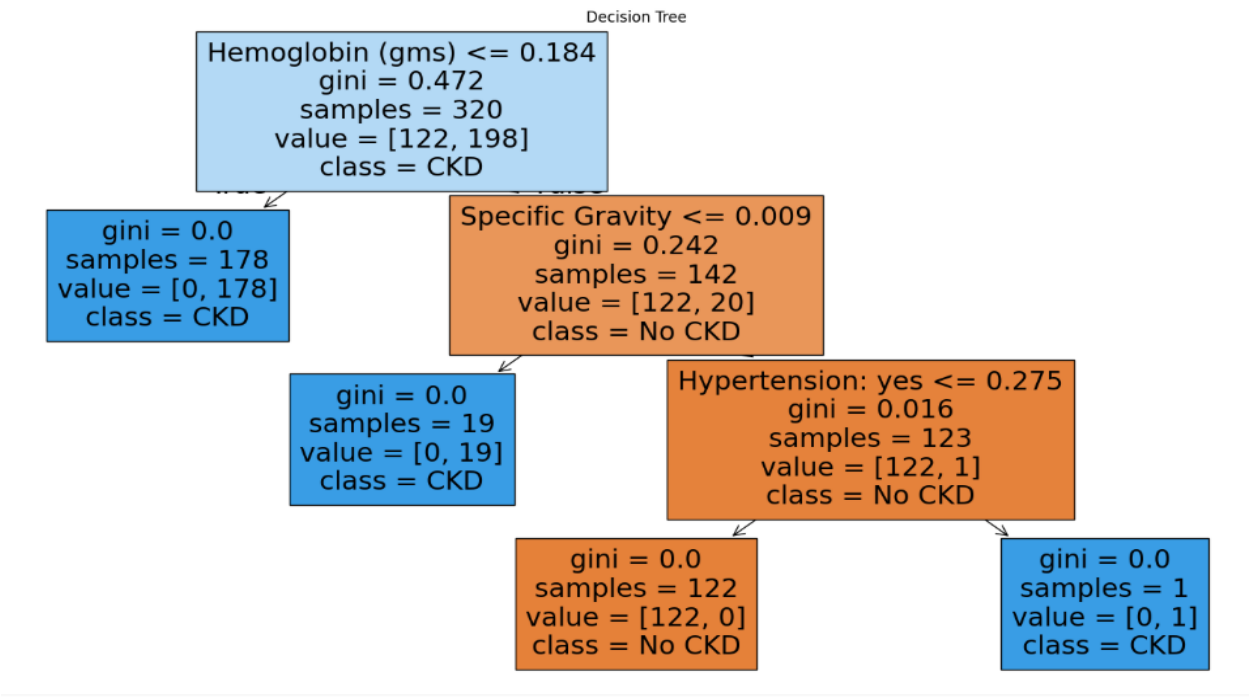


Figure 3.4: Decision Tree for CKD

3.9.2 Logistic Regression

Logistic Regression is also a type supervised machine learning model and contains proper ouput based on binary or classification types of data [18]. It is provides better output when one or more dependent variables depend on independent variable [18]. This research project uses this traditional model to find the proper output for CKD detection. This model has law to determine CKD perfectly which is essention for predictions. It is usable for multiple variable so that findings the correlation between features indicates better solution.

Y (cap) = W1X1 + W. [law of logistic regression]
It has a log loss function to determine better which is indicated by (E).
E = - y log (y cap) - (1-y) log (1 - y cap)
It has discrete numbers 1 and 0. 1 signs positive class(+) class and 0 signs negative class (-). In positive class for reducing the error the values closely need 1, and in negative class values closely need 0. This is a way of working at logistic regression.

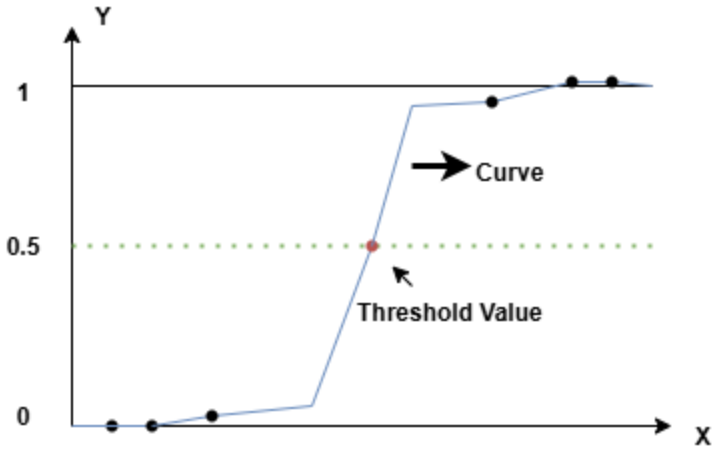


Figure 3.5: Logistic regression

3.9.3 XGBoost

XGBoost's full form is extreme gradient boosting is a traditional machine laeaing learning model which is fast compare to other traditional ML models. It can works with multiple decision trees together. So that it can handle complex data and provide good result for detecting CKD. Gradient determines amount of loss and boosting determines improve the model based on previous model result. This model uses this model for improving result which is efficient.

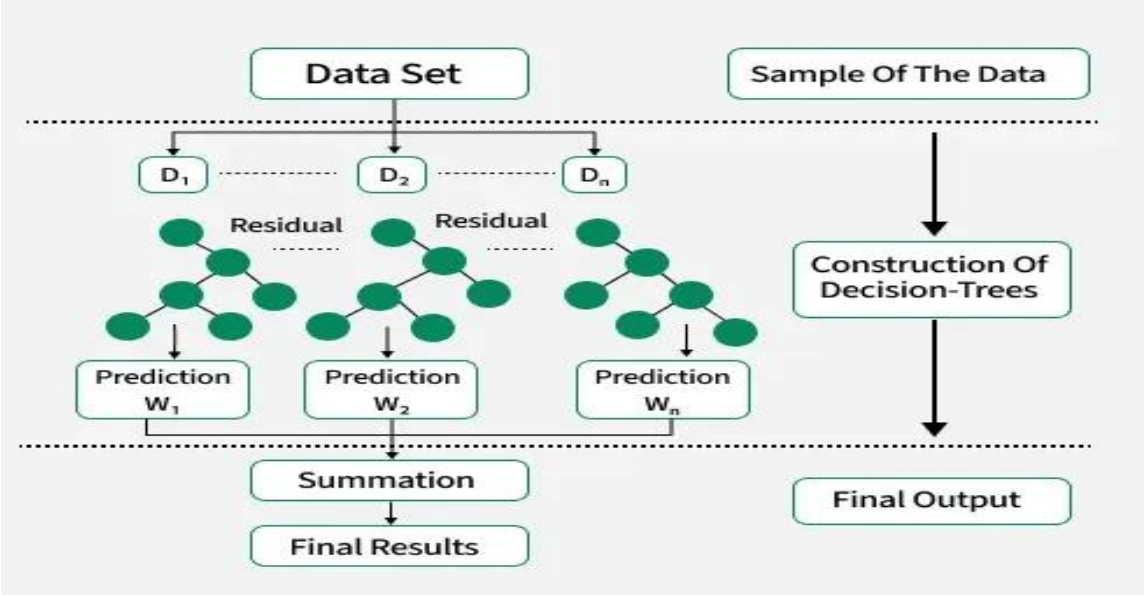


Figure 3.6: XGBoost (Source from geeks for Geeks)

3.9.4 K-Fold Cross-Validation

K-fold cross validation is mainly used for dividing the dataset into 5 parts. When we train and test data, sometimes the same types of data are stored in the training part. Then the machine provides wrong output in test part, so that the machine recognize as a poor machine. To improve this problem need to use k fold validation, because it divides the data into 5 parts and 4 parts as training set and 1 part as a test part. By doing this all types of value are stored in both training and testing part. After that machine will perform well, meanings the out closely remains same in both way. This project use K-fold cross validation with all three traditional model (decision tree, logistic regression, XGBoost) for getting the proper output in both training and testing data.

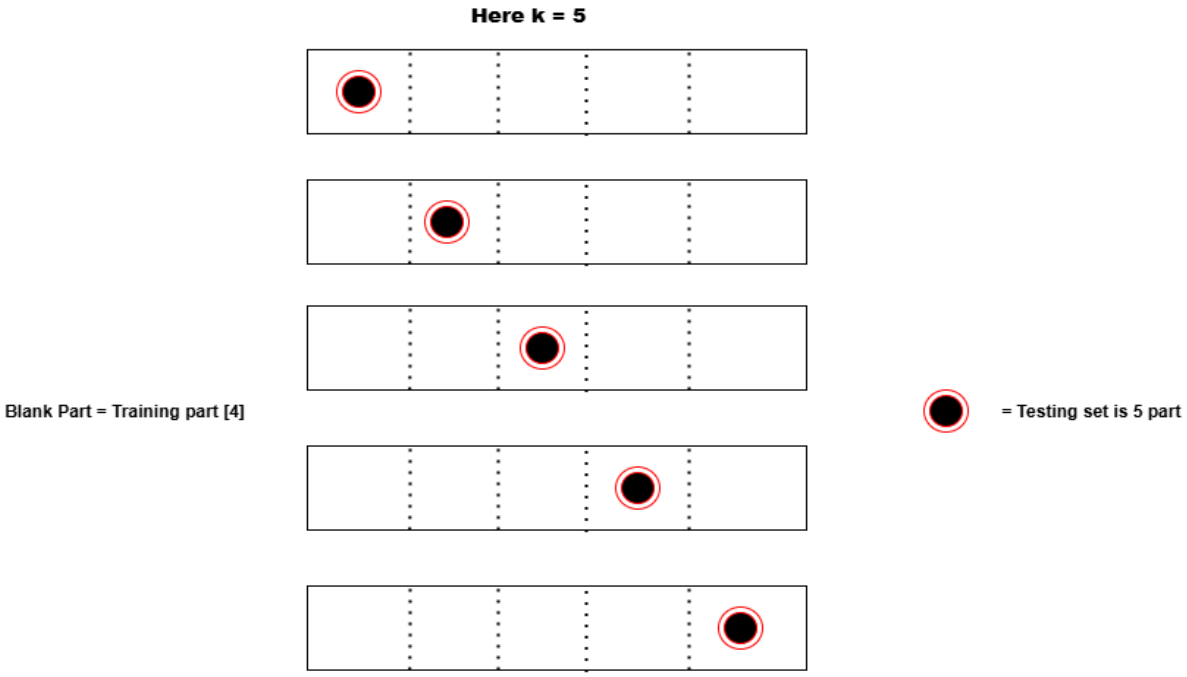


Figure 3.7: K-fold Cross-validation

3.9.5 Stacking Model CKDML 762

This research project creates a hybrid or stacking model by using decision tree, logistic regression, and XGBoost traditional models named CKDML762. This is a new implementation model for this project. The reason that creating a stacking model with combination of traditional machine learning models can provide better performance and detect the CKD more accurately. This model can predict CKD early with a good percentage of accuracy; thus, this model can be suitable and easy to use with proper implementation.

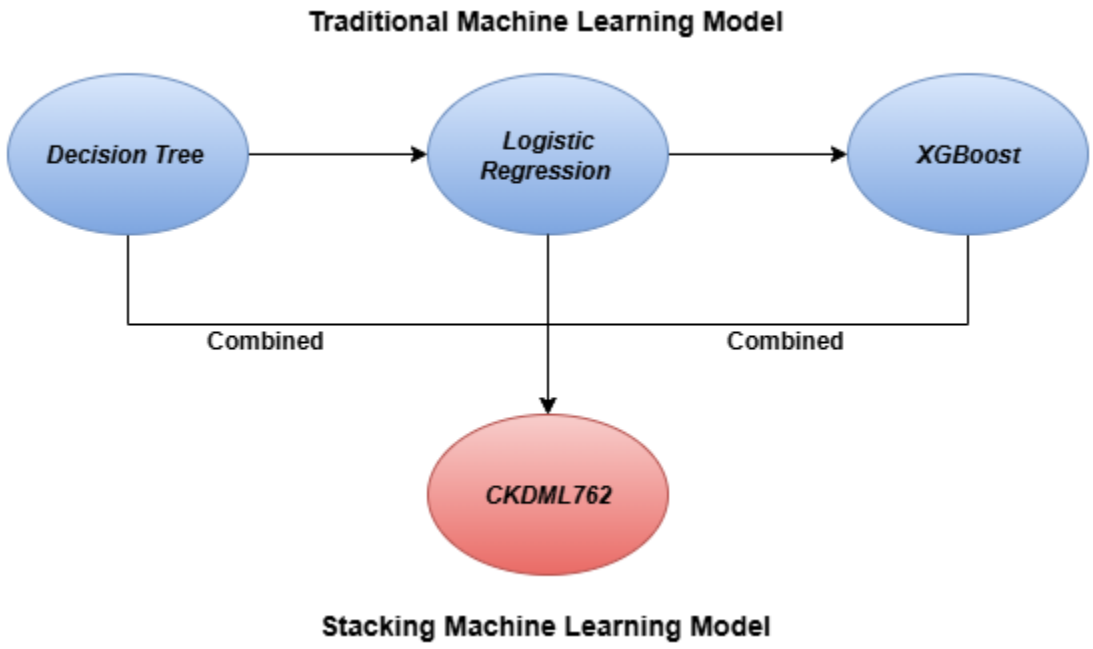


Figure 3.8: CKDML762 Stacking Model

Chapter 4

4.0 Results and Discussion

This research project aims to predict Chronic Kidney Disease (CKD) using various traditional machine learning algorithms, including decision trees using k-fold cross-validation, logistic regression using k-fold cross-validation, and XGBoost using k-fold cross-validation. By combining all those 3 models, this project creates a stacking model named CKDML762. Let's discuss the models step by step.

4.1 Decision Tree using k-fold cross-validation

Here implement the decision tree by splitting the data 5 times with k-fold. By doing this the training and test data provide the same level of accuracy, which is recognizes as a good machine learning model. With this procedure, 80% is contained as training data and 20% is contained as testing data. Let's analyze the result of decision tree.

Confusion evaluation metrics result of decision tree (training data):

Evaluation Metrics Measures	Results (in percentage)
Accuracy	100%
Precision	100%
Recall	100%
F1 Score	100%

Table 4.1: Confusion evaluation result for training data

For training data, the decision tree model predicts CKD so accurately with a 100% accuracy rate.

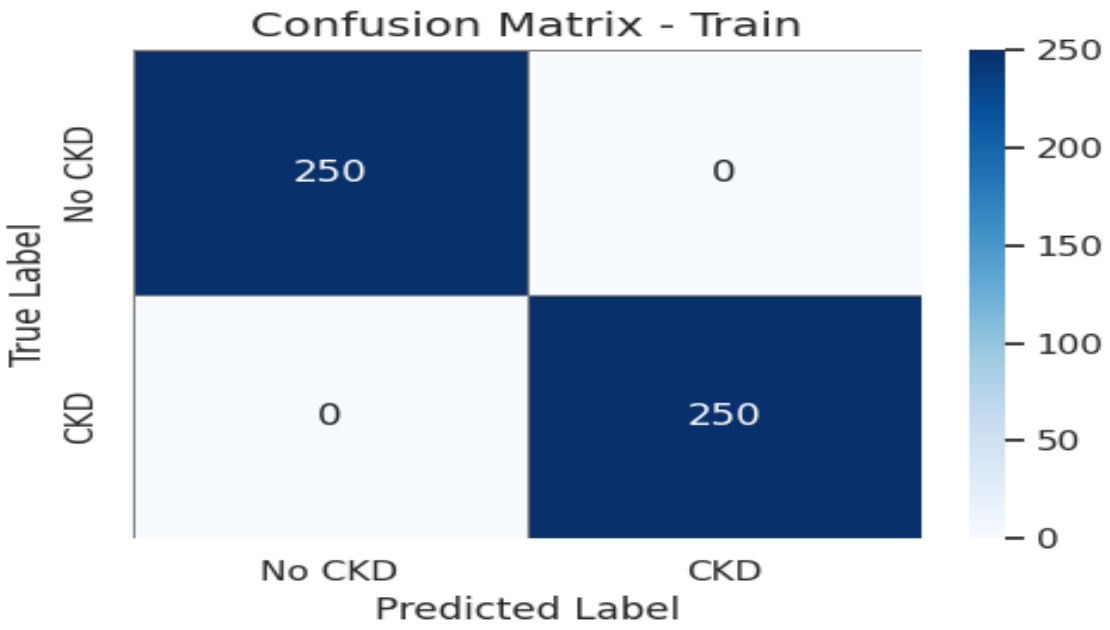


Figure 4.1: Confusion matrix (training) for decision tree

Confusion evaluation metrics result of decision tree (testing data):

Evaluation Metrics Measures	Results (in percentage)
Accuracy	99%
Precision	99%
Recall	100%
F1 Score	99%

Table 4.2: Confusion evaluation result for testing data

For testing data, the decision tree model predicts CKD with a 99% accuracy rate which indicates that this model is working well. The training and testing accuracy is so close, which is the nature of a good model.

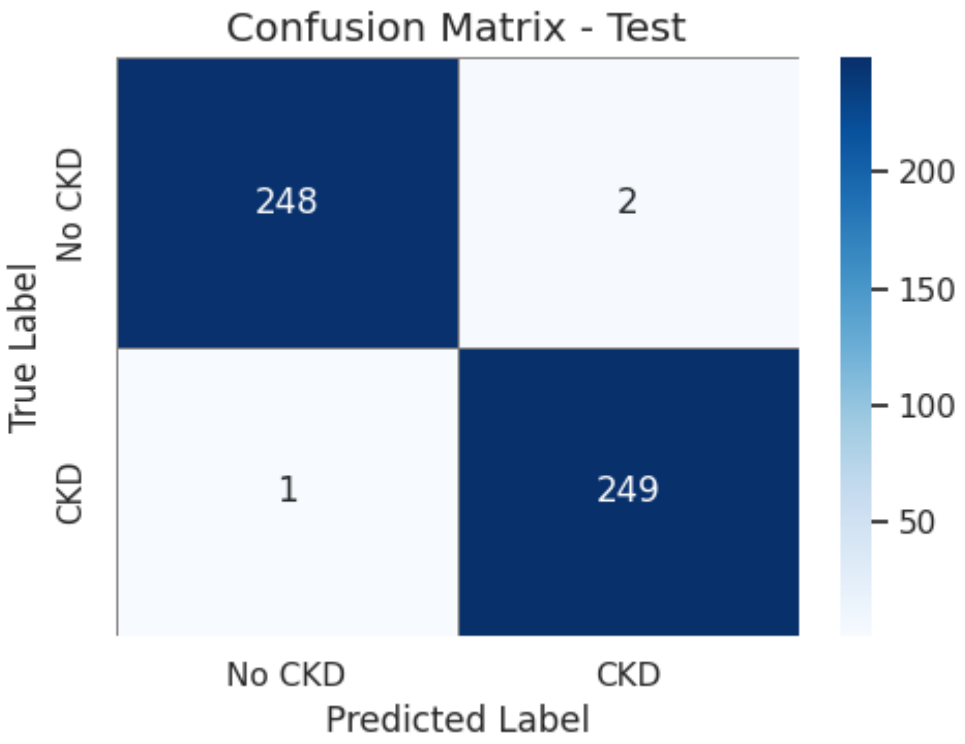


Figure 4.2: Confusion matrix (training) for decision tree

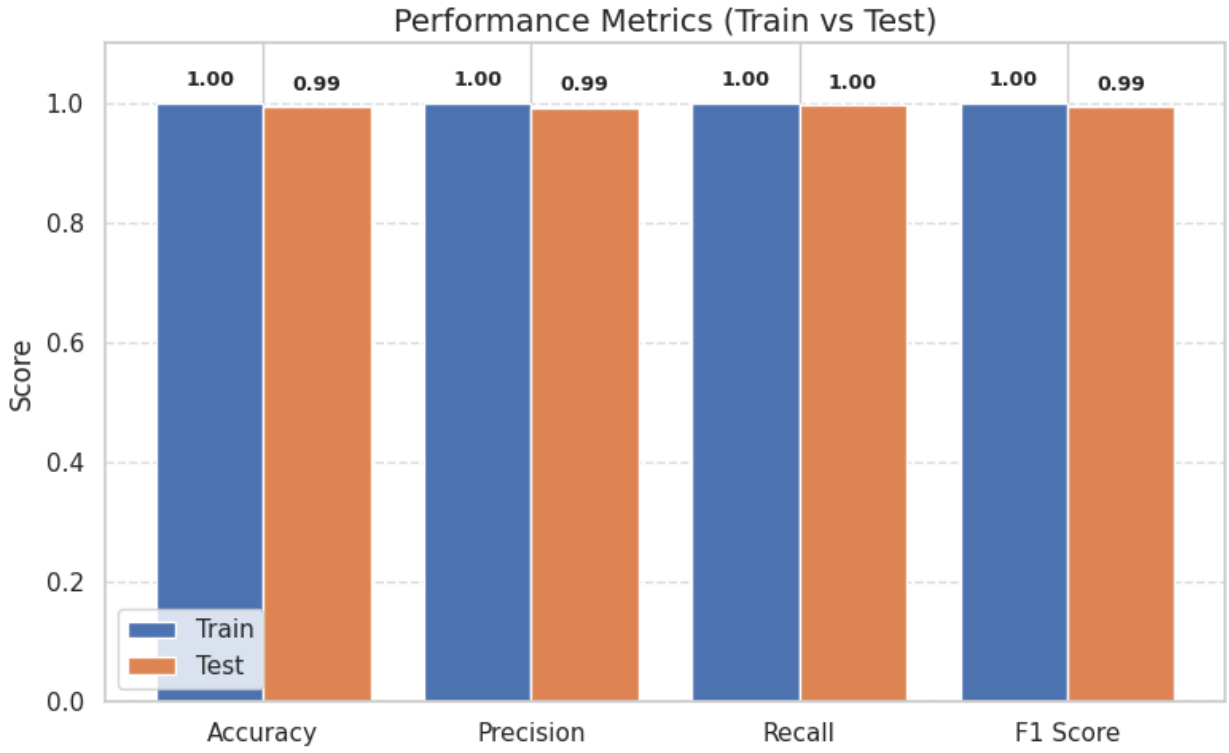


Figure 4.3: Performance Metrics train vs. test for decision tree using k-fold

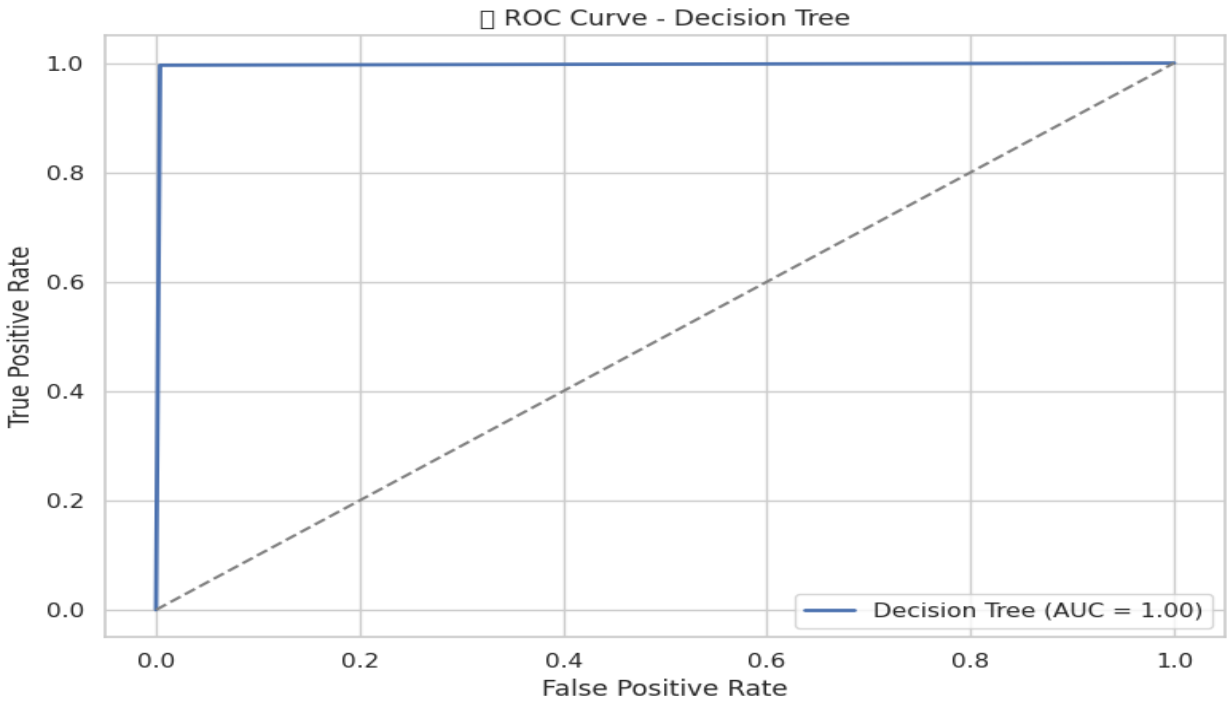


Figure 4.4: ROC Curve Decision Tree

ROC (receiver operating characteristic) showing score 1, which indicates the decision tree is a perfect classifier model for predicting CKD. In this graph x- axis means FPR and y-axis means TPR.

4.2 Logistic Regression using k-fold cross-validation

Here implement the logistic regression by splitting the data 5 times with k-fold. By doing this the training and test data provide the same level of accuracy, which is recognizes as a good machine learning model. With this procedure, 80% is contained as training data and 20% is contained as testing data. Let's analyze the result of logistic regression.

Confusion evaluation metrics result of logistic regression (training data):

Evaluation Metrics Measures	Results (in percentage)
Accuracy	99%
Precision	100%
Recall	98%
F1 Score	99%

Table 4.3: Confusion evaluation result for training data

For training data, the logistic regression model predicts CKD with a 99% accuracy rate, which indicates that performance is better.

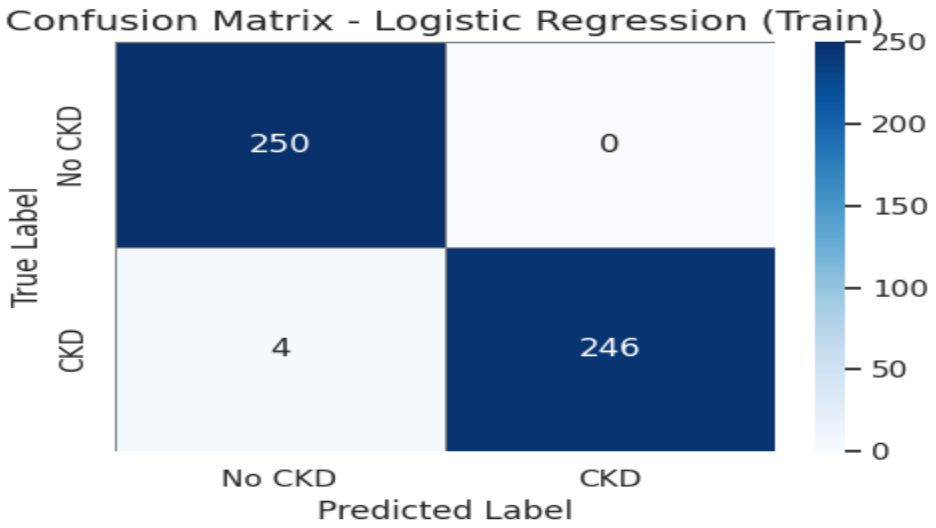


Figure 4.5: Confusion matrix for training data (logistic regression)

Confusion evaluation metrics result of logistic regression (testing data):

Evaluation Metrics Measures	Results (in percentage)
Accuracy	98%
Precision	100%
Recall	96%
F1 Score	97%

Table 4.4: Confusion evaluation result for testing data

For testing data, the logistic regression model predicts CKD with a 98% accuracy rate, which indicates that this model is working well. The training and testing accuracy is so close, which is the nature of a good model.

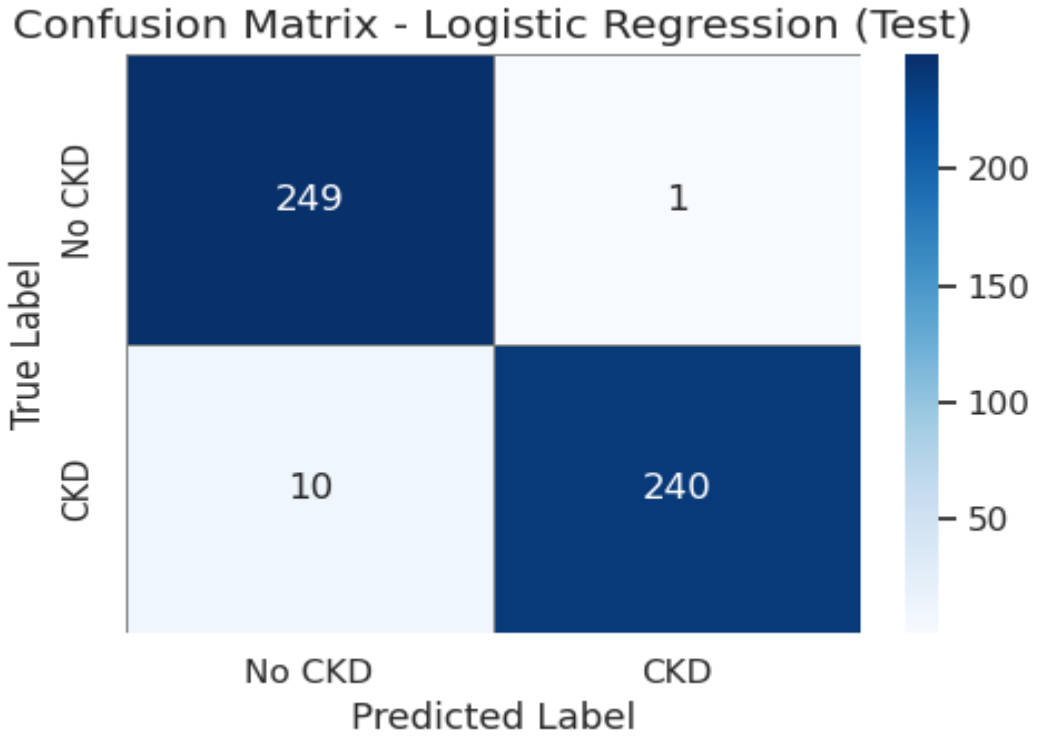


Figure 4.6: Confusion matrix (training) for logistic regression

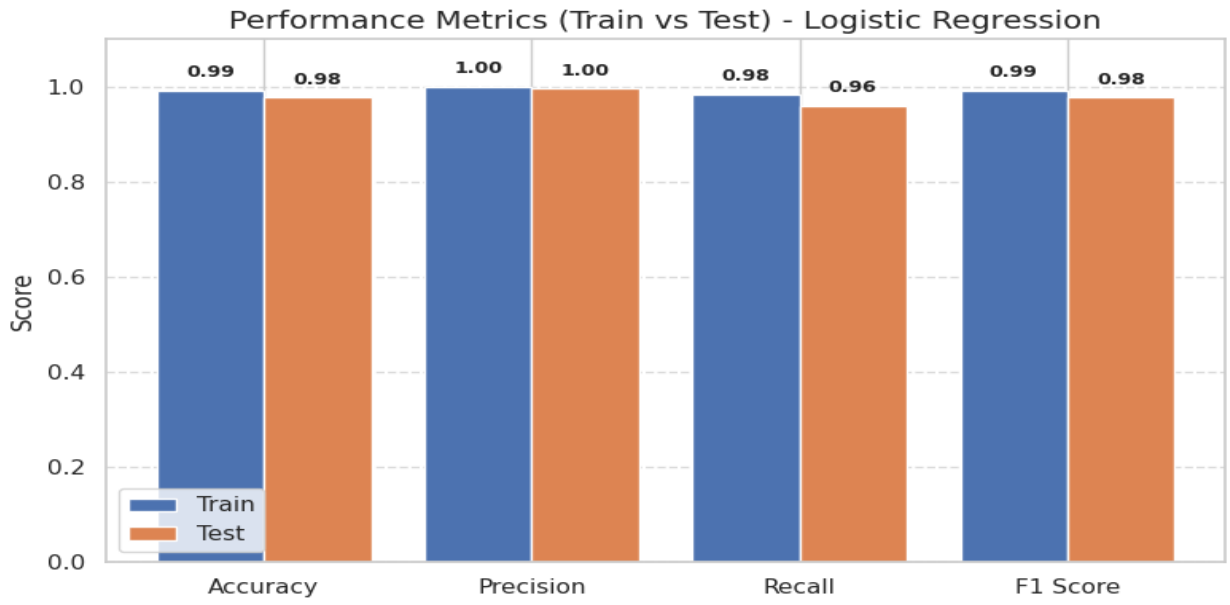


Figure 4.7: Performance Metrics train vs. test logistic regression using k-fold

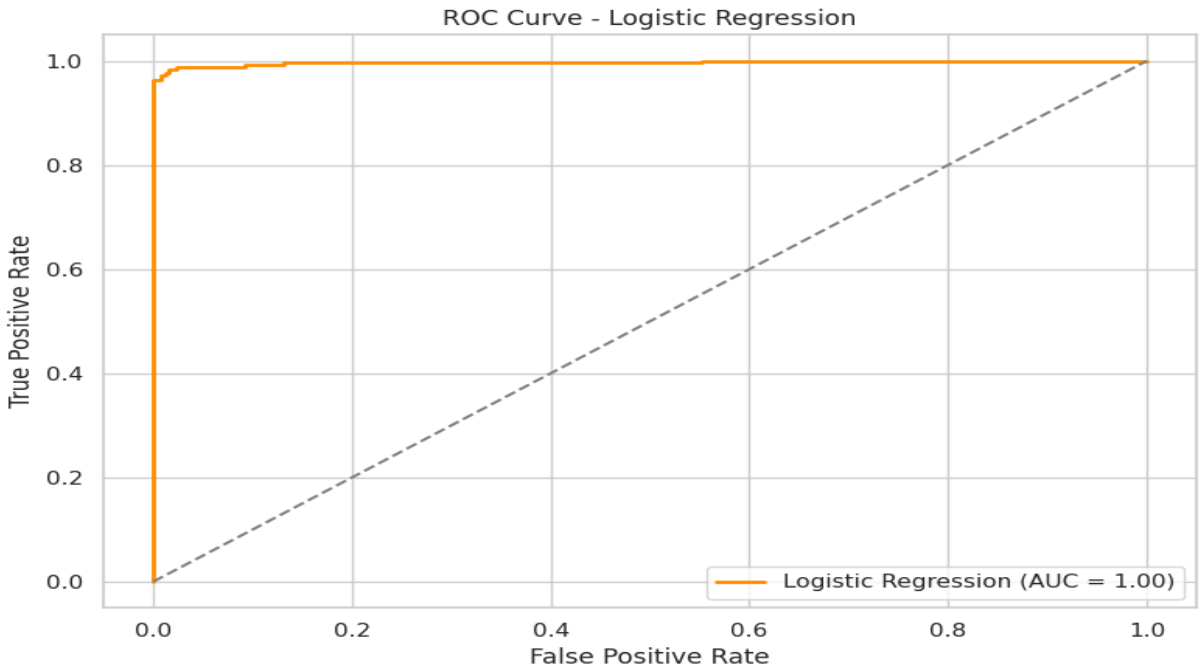


Figure 4.8: ROC Curve Logistic regression

In logistic regression roc curve score is which is better to detect CKD efficiently.

4.3 XGBoost using k-fold cross-validation

Here implement the XGBoost by splitting the data 5 times with k-fold. By doing this the training and test data provide the same level of accuracy, which is recognizes as a good machine learning model. With this procedure, 80% is contained as training data and 20% is contained as testing data. Let's analyze the result of XGBoost.

Confusion evaluation metrics result of XGBoost (training data):

Evaluation Metrics Measures	Results (in percentage)
Accuracy	100%
Precision	100%
Recall	100%
F1 Score	100%

Table 4.5: Confusion evaluation result for training data

For training data, the XGBoost modell predicts CKD with a 100% accuracy rate, which indicates that performance is better.

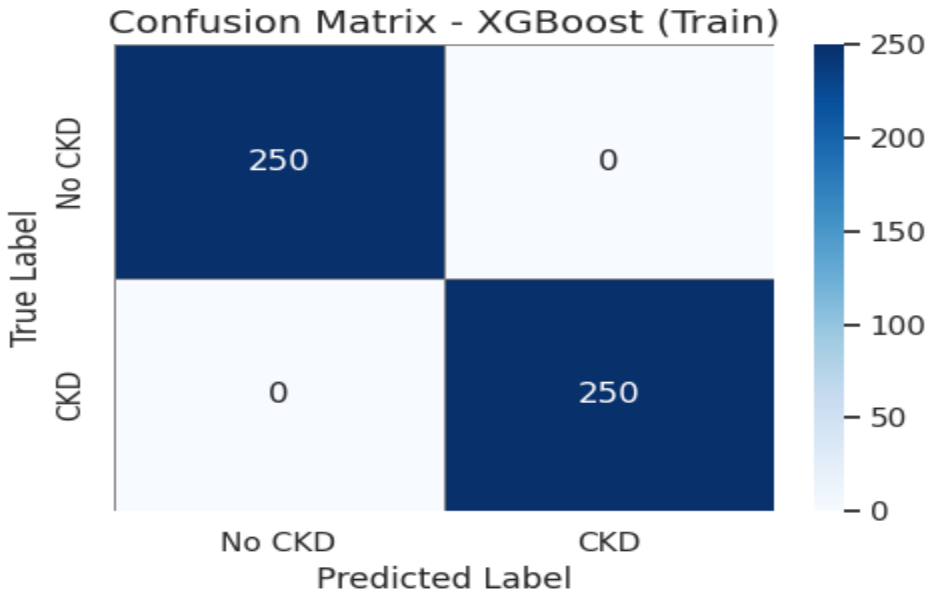


Figure 4.9: Confusion matrix for training data (XGBoost)

Confusion evaluation metrics result of XGBoost (testing data):

Evaluation Metrics Measures	Results (in percentage)
Accuracy	100%
Precision	100%
Recall	99%
F1 Score	100%

Table 4.6: Confusion evaluation result for testing data

For testing data, the decision tree model predicts CKD with a 100% accuracy rate which indicates that this model is working well. The training and testing accuracy is so close, which is the nature of a good model.

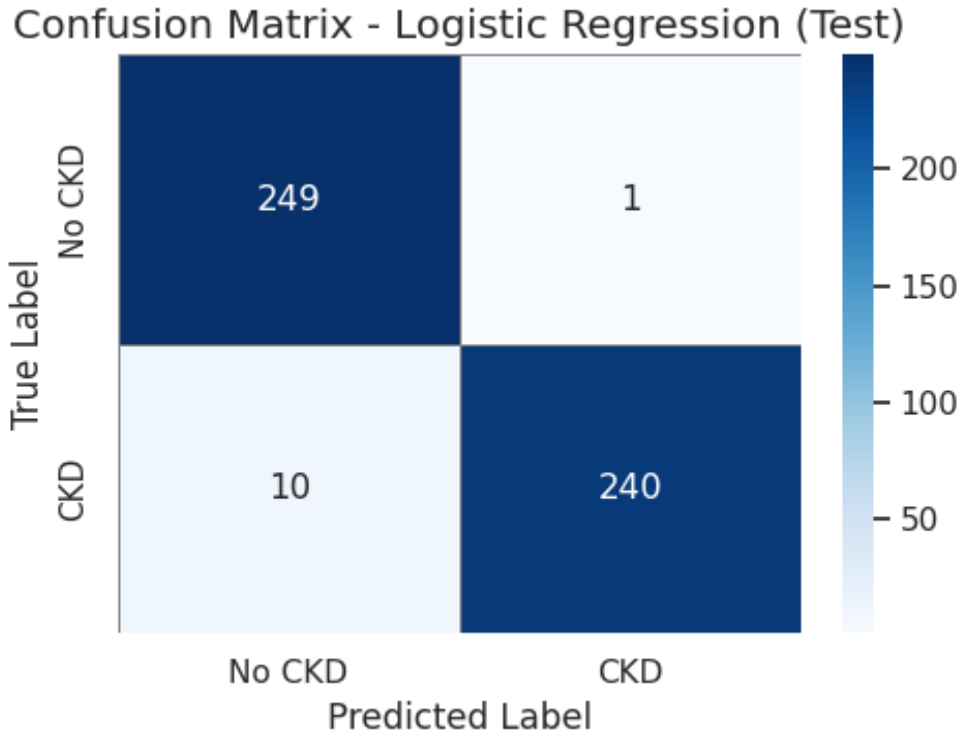


Figure 4.10: Confusion matrix for testing data (XGBoost)

Figure 4.11: Performance Metrics Train vs. test XGBoost using k-fold

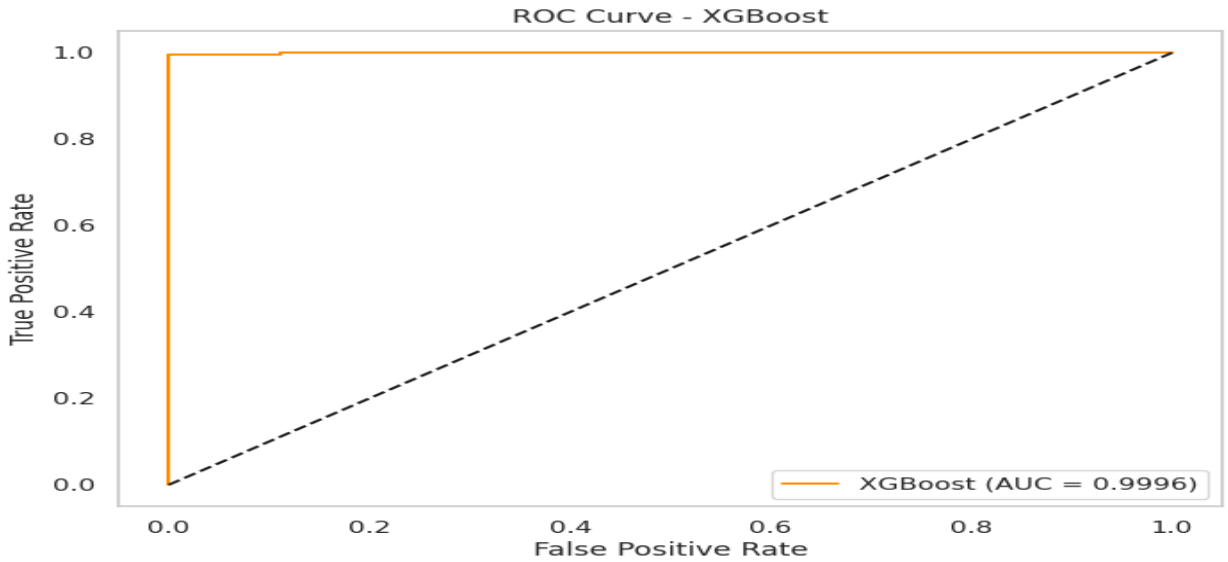


Figure 4.12: Roc Curve XGBoost

The XGBoost models also provide the curve score closely to 1, so the graph remains the same.

4.4 Stacking Model CKDML762

In this section, this project created a stacking model with the combination decision tree classifier, logistic regression, and gradient boosting classifier named CKDML762. The decision tree and XGBoost collaborate as base models. The linear regression model collaborates as a meta-learner in this model. The base learners analyze the result, and meta learner fixed the result with combination. For example, decision trees provide 98% accuracy, logistic regression provides 95% accuracy, and XGBoost provide 92% accuracy. With the combination of all accuracy stacking model will provide 95% accuracy. So it takes care of every model that is found as a good model or perfect model detection. Let's analyze the result of stacking model CKDML762.

Confusion evaluation metrics result of CKDML762:

This model also uses accuracy, precision, recall, and f1-score to evaluate results. Here the model works so accurately for detecting CKD in both test cases and training cases. The model provides the same result for both, which is called a good model.

Evaluation Metrics Measures	Results (in percentage)
Accuracy	98.75%
Precision	98.11%
Recall	100%
F1 Score	99.05%

Table 4.7: Confusion evaluation result for CKDML762

Here table 4.7 shows the result of stacking model in both cases. Have a good accuracy rate with 0.9875%. Best thing about this model is that it has a 100% recall rate. Because we recall is important matrices for the medical field. Here CKDML762 ensures that this model can predict the actual CKD paitents so accurately with 100% recall rate.

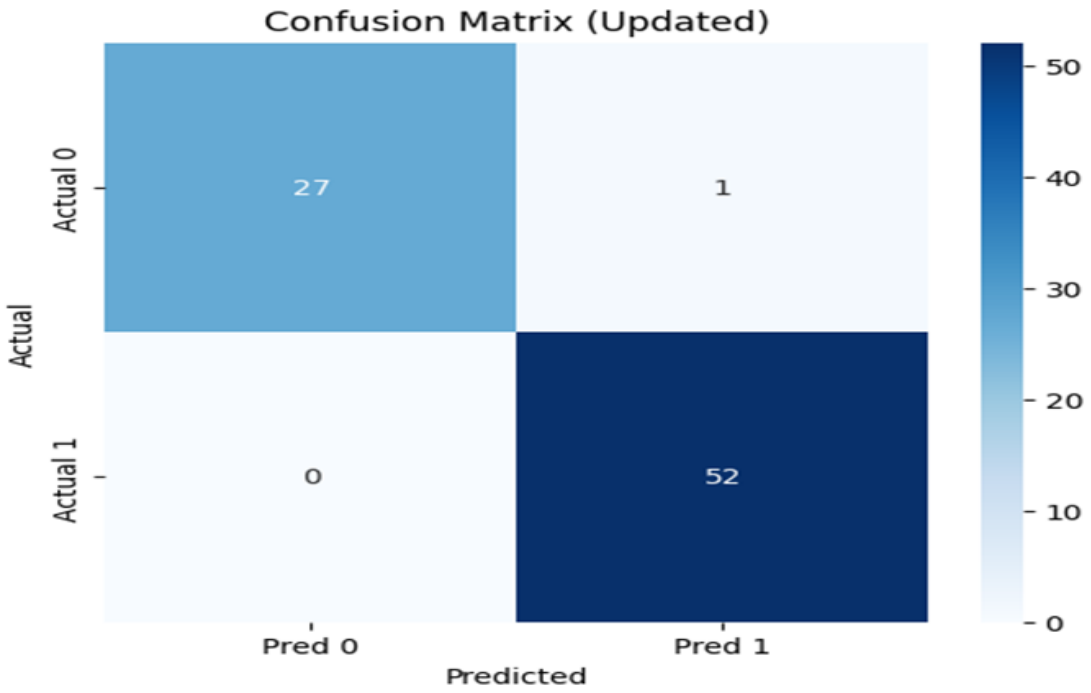


Figure 4.13: Confusion matrix for CKDML762

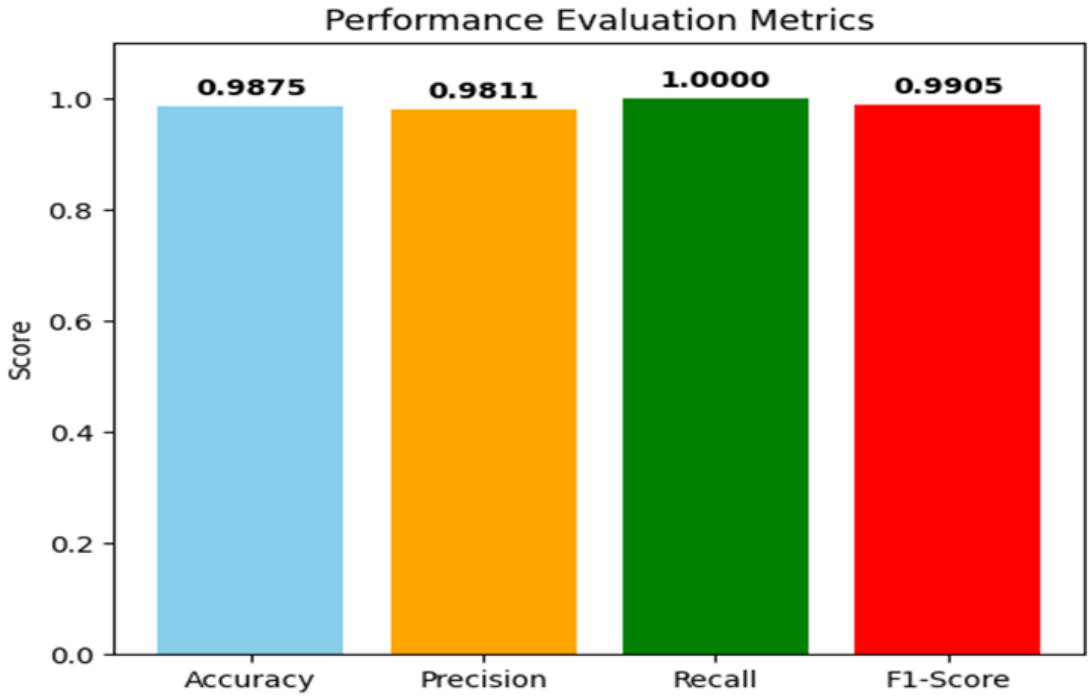


Figure 4.14: Score table of evaluation metrics

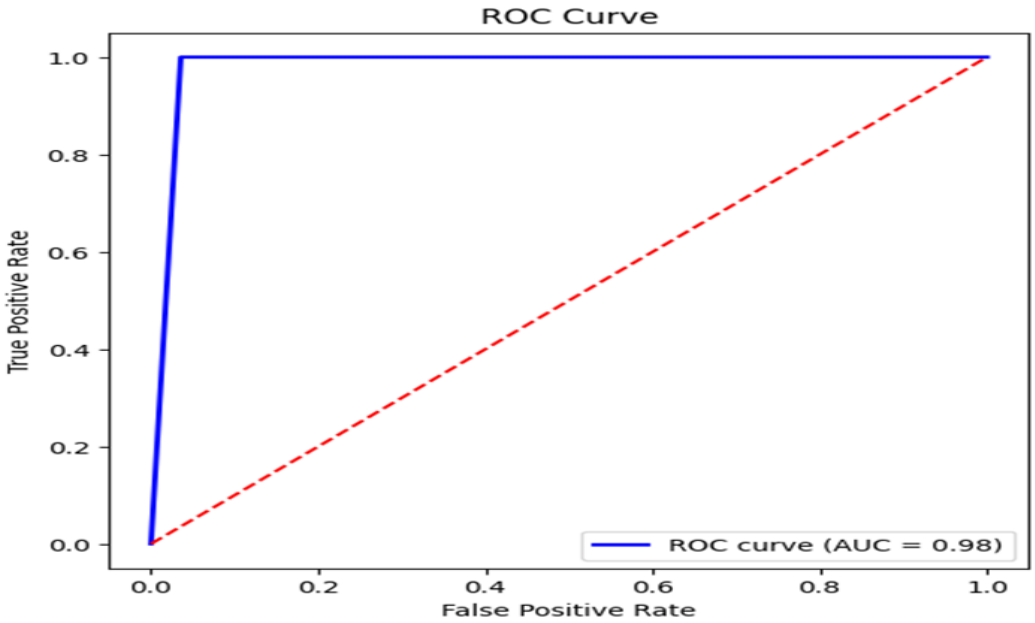


Figure 4.15: ROC curve of CKDML762

Figure 4.15 shows the ROC (receiver operating characteristic) curve of CKDML762 and provides the AUC (area under curve) score, which is close to 1, which indicates a good model.

4.5 Discussions

After showing all models CKD detecting accuracy rate and performance, this project will discuss the measurement here. As per the result, every model works properly with good percentage fo accuracy rate, even though the stacking model CKDML762 works a little better compared to the other 3 models. Because it provides the same level of accuracy in both test and train data. So that it can predict CKD so well with the actual result. Even if we add new data or unseen data it will work better and predict CKD accurately. These techniques will remain data clean and meaningful (have correlation including all features also). This will be the perfect stacking model to predict CKD early and more accurately.

Comparing all models:

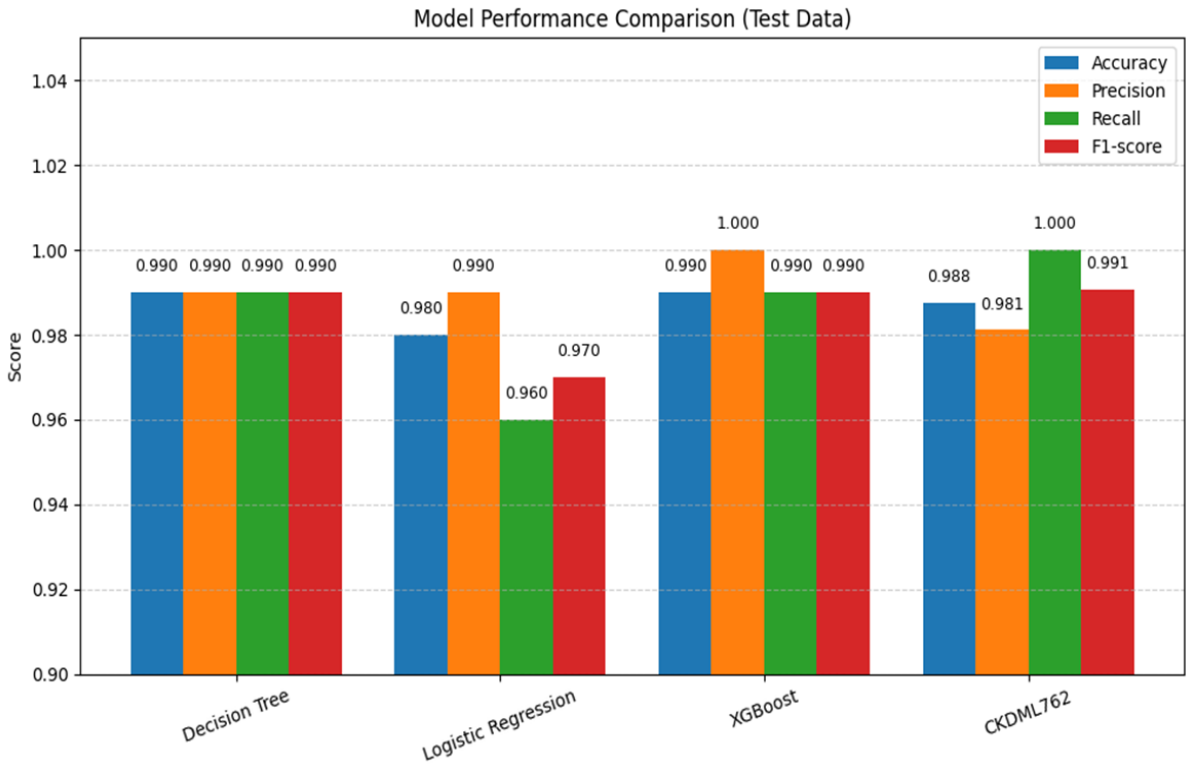


Figure 4.16: Comparison of all models

The 4.16 figure shows the comparison between decision tree, logistic regression, XGBoost, and CKDML762. Decision tree and XGboost provides the same type of result, they provides 99% accuracy rate in test case and 100% in training cases. This is good for training data but maybe causes of overfitting when it comes with unseen or new data. For this reason, those models may not be able to predict CKD accurately with new patient data. At the point of logistic regression, it has good accuracy of 98.75% but lacks in recall score, which is 96%. In medical machine recall, it is crucial for predicting disease. Otherwise, the machine can predict a CKD patient as a non-CKD patient. Come to the point of stacking model CKDML762, it has good accuracy rate in both cases, 98.75%. Also have the good recall rate of 100%, precision rate of 98%, f1-score rate of 99.05%. In this stacking model, there is no overfitting issue it can predict CKD as well as the unseen data. The best thing about this model is have 100% rate of recall means it can predict the real CKD patients accurately without any error, compared to logistic regression (which has this issue). In the medical field, predicting the CKD patient accurately with a machine is so important. This research stacking model CKDML762 can do it properly can be used for detect CKD early for the patients. So compare to all models, stacking model (CKDML762) is the best model compare to all of this for predict CKD early.

Chapter 5

5.0 Conclusion

CKD is an innovative matter that impacts millions of people all over the world, including Bangladesh. This research project shows various machine learning algorithms for detecting CKD early and their different types of causes and challenges. The machine learning model is the invention of modern time, which can change the medical diagnostic, utlizing large data and difficult algorithms to find patterns almost invisible to doctors. This project used the CKD dataset from the UC Irvine Machine Learning Repository, a site invented by the University of California. This dataset is publicly available on their website. This project uses classification algorithms such as decision tree using 5-fold cross-validation, logistic regression using 5-fold cross-validation, XGboost using 5-fold cross-validation, and the stacking model CKDML762 with the combination of decision tree, logistic regression, and XGboost. Decision tree, XGboost are the based model and logistic the meta-learner in this stacking or hybrid model. This investigation setup engaged dataset into python with the machine learning library. The investigation displayed that the CKDML762 stacking model performed superbly with the perfect scores of accuracy, precision, recall, f1 score in both training and testing data. Subsequently the decision tree and XGboost founded the overfiting issue because of it maybe predict error in new or unseen data, logistic regression have the recall issue compare to stacking model, recall is crucial for medical field. So stacking perfoms the best at overall conditions and it will be suitable to use .These findings suggested that this model correctly predicted the CKD early and created a good impact in clinical fields.

5.1 Limitations of the research

Though machine learning models stand for a very great possibility in the early detection of chronic kidney disease, also have some limitations that need to be mentioned in this research. First of all there could be some fundamental biases in the dataset used in this research that may lead to generalizations of the results. For example if dataset is biased throughout the area-based populations, then this research predictive accuracy will be not same as for the mixed population. In addition some of the documents were unavailable in this dataset, which can causes bias like sampling biased, non-response biased, sampling noise, which can destroy the performance of the model. As for some neglected features, the model may fail sometime to predicts the CKD accurately.

Technically, the used algorithm and metrics of the project can be another factor possibly impacting the choice of results taken up by the research. Wheres decision tree, logistic regression, XGboost all have their own structure, the actual CKD detention counts on the good quality of the data and how well the relation between features in dataset including

the generalized form also. Notably overfitting maybe causes a serious problem for detecting CKD based on those model.

May occurs the lackings issue in stacking model when it faces the diverse dataset between all over the world because it has no real-world implementation. Clinical features may have some issue in dataset need to add those features.

5.2 Future Implementation of research

To resolve the limitations of this research project and further advanced field of machine learning in CKD detection, a countless path for future work and optimization and be followed. First the of all, the dataset needs to be assorted with including more areas with diverse populations and information from healthcare all over the world. By dining this the machine learning model can understood the patter more wisely and able predict the can be able predict their dataset to detect CKD. These could consists of collaboration across all over the world and make the dataset large that will be considered a more meaningful dataset. This process can be done by collection of data from various hospital means raw data.

Another reason, if we disccused more about machine learning algorithms techniques like such as deep learning, that maybe provide even more better performance without overfitting. Let's mention these types of model CNNs, ANNs,these model can perform better and also we can make hybrid model by combining these deep learning models. As per this research make hybrid model by combining traditional machine learning model like this method.

Sequential discussion will be essential to notice the outcomes pf machine learning model uses in the scenario of real world for detection of CKD properly. The influence of these models on patient outcomes, medical expenses, and level of total care will be helpful related to their capability and robustness in patient care. These attemps clarify machine learning merging at all levels to later become scheduled for early CKD detection, which will eventually gives better result for patient care and may improve the disease of the kidney. After all, it is possible to find the disease early with the proper implementation of ML technology and it is actually crucial to decrease the death percentage causes for kindney damge.

Needed to implement the stacking model CKDML762 in a clinical way means the real-world implementation with more diverse dataset. By implementing this CKD can be predict so early in real world. That will be best solution for the CKD patients at costs, recoginze, health cooncious, reducing danger of kidney, reduce other disease for kidney. That model will be the perfect solution for predict CKD in early stages.

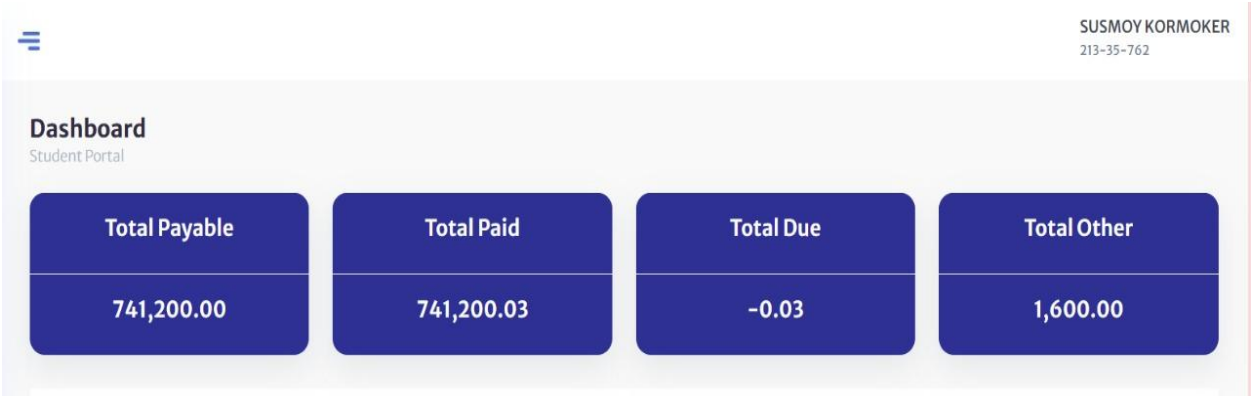
Chapter 6

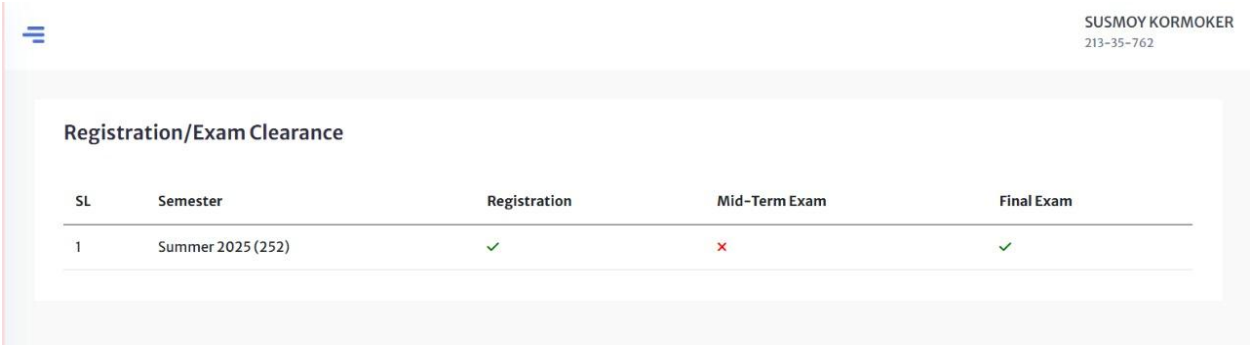
6.0 References

1. Debal, D. A., & Sitote, T. M. (2022). Chronic kidney disease prediction using machine learning techniques. Journal of Big Data, 9(1), 109.
2. Zhang, Q. L., & Rothenbacher, D. (2008). Prevalence of chronic kidney disease in population-based studies: systematic review. BMC public health, 8(1), 117.
3. McClellan, W. M., Warnock, D. G., Judd, S., Muntner, P., Kewalramani, R., Cushman, M., ... & Howard, G. (2011). Albuminuria and racial disparities in the risk for ESRD. Journal of the American Society of Nephrology, 22(9), 1721-1728.
4. de Souza, W., de Abreu, L. C., Silva, L. G. D., & Bezerra, I. M. P. (2019). Incidence of chronic kidney disease hospitalisations and mortality in Espírito Santo between 1996 to 2017. PLoS One, 14(11), e0224889.
5. Chittora, P., Chaurasia, S., Chakrabarti, P., Kumawat, G., Chakrabarti, T., Leonowicz, Z., ... & Bolshev, V. (2021). Prediction of chronic kidney disease-a machine learning perspective. IEEE access, 9, 17312-17334.
6. Dritsas, E., & Trigka, M. (2022). Machine learning techniques for chronic kidney disease risk prediction. *Big Data and Cognitive Computing*, *6*(3), 98.
7. Ebiaredoh-Mienye, S. A., Swart, T. G., Esenogho, E., & Mienye, I. D. (2022). A machine learning method with filter-based feature selection for improved prediction of chronic kidney disease. Bioengineering, 9(8), 350.
8. Islam, M. A., Akter, S., Hossen, M. S., Keya, S. A., Tisha, S. A., & Hossain, S. (2020, December). Risk factor prediction of chronic kidney disease based on machine learning algorithms. In 2020 3rd international conference on intelligent sustainable systems (ICISS) (pp. 952-957). IEEE.
9. Poonia, R. C., Gupta, M. K., Abunadi, I., Albraikan, A. A., Al-Wesabi, F. N., Hamza, M. A., & B, T. (2022, February). Intelligent diagnostic prediction and classification models for detection of kidney disease. In Healthcare (Vol. 10, No. 2, p. 371). MDPI.
10. Kaur, C., Kumar, M. S., Anjum, A., Binda, M. B., Mallu, M. R., & Al Ansari, M. S. (2023). Chronic kidney disease prediction using machine learning. Journal of Advances in Information Technology, 14(2), 384-391.
11. Pinto, A., Ferreira, D., Neto, C., Abelha, A., & Machado, J. (2020). Data mining to predict early stage chronic kidney disease. Procedia Computer Science, 177, 562-567.
12. Xiao, J., Ding, R., Xu, X., Guan, H., Feng, X., Sun, T., ... & Ye, Z. (2019). Comparison and development of machine learning tools in the prediction of chronic kidney disease progression. Journal of translational medicine, 17(1), 119.
13. Zheng, Q., Tastan, G., & Fan, Y. (2018, April). Transfer learning for diagnosis of congenital abnormalities of the kidney and urinary tract in children based on ultrasound imaging data. In 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018) (pp. 1487-1490). IEEE.

14. Tekale, S., Shingavi, P., Wandhekar, S., & Chatorikar, A. (2018). Prediction of chronic kidney disease using machine learning algorithm. International Journal of Advanced Research in Computer and Communication Engineering, 7(10), 92-96.
15. Padmanaban, K. A., & Parthiban, G. (2016). Applying machine learning techniques for predicting the risk of chronic kidney disease. Indian Journal of Science and Technology, 9(29), 1-6.
16. Jena, L., Patra, B., Nayak, S., Mishra, S., & Tripathy, S. (2020). Risk prediction of kidney disease using machine learning strategies. In Intelligent and Cloud Computing: Proceedings of ICICC 2019, Volume 2 (pp. 485-494). Singapore: Springer Singapore.
17. Ghosh, S. K., & Khandoker, A. H. (2023). A machine learning driven nomogram for predicting chronic kidney disease stages 3–5. Scientific Reports, 13(1), 21613.
18. Prasad, M. L., Kiran, A., & Shaker Reddy, P. C. (2024). Chronic kidney disease risk prediction using machine learning techniques. Journal of Information Technology Management, 16(1), 118-134.
19. Ebiaredoh-Mienye, S. A., Swart, T. G., Esenogho, E., & Mienye, I. D. (2022). A machine learning method with filter-based feature selection for improved prediction of chronic kidney disease. Bioengineering, 9(8), 350.
20. El Sherbiny, M. M., Abdelhalim, E., Mostafa, H. E. D., & El-Seddik, M. M. (2023). Classification of chronic kidney disease based on machine learning techniques. Indones. J. Electr. Eng. Comput. Sci, 32(2), 945-955.

Account Clearance





Plagiarism Report

213-35-762

ORIGINALITY REPORT

23%	19%	15%	13%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	4%
2	Submitted to Daffodil International University Student Paper	1%
3	Submitted to King Abdulaziz University Student Paper	1%
4	Submitted to South Bank University Student Paper	1%
5	de Carvalho Pereira Ferreira, Inês. "Churn Prediction in Digital Marketing", Universidade do Porto (Portugal), 2024 Publication	1%
6	www.ijfmr.com Internet Source	1%
7	assets.researchsquare.com Internet Source	1%

	Student Paper	
13	Submitted to University of North Texas Student Paper	<1 %
14	Submitted to University of Hertfordshire Student Paper	<1 %
15	medium.com Internet Source	<1 %
16	www.mdpi.com Internet Source	<1 %
17	iare.ac.in Internet Source	<1 %
18	Submitted to Birla Institute of Technology and Science Pilani Student Paper	<1 %
19	www.coursehero.com Internet Source	<1 %
20	export.arxiv.org Internet Source	<1 %
21	scholar.colorado.edu Internet Source	<1 %