

**PREDICTION OF CHRONIC KIDNEY DISEASE USING DIFFERENT
MACHINE LEARNING METHODS**

BY

**MD. MUBTASIM FUAD
ID: 171-15-8815**

This Report Presented in Partial Fulfillment of the Requirements for the
Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Md. Mahfujur Rahman
Lecturer (Senior Scale)
Department of CSE
Daffodil International University

Co-Supervised By

Mohammad Jahangir Alam
Lecturer
Department of CSE
Daffodil International University



**DAFFODIL INTERNATIONAL UNIVERSITY
DHAKA, BANGLADESH
SEPTEMBER 2022**

APPROVAL

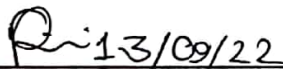
This Project titled “**Prediction of Chronic Kidney Disease using Different Machine Learning Methods**”, submitted by Md. Mubtasim Fuad, ID No: 171-15-8815 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 13/09/2022.



BOARD OF EXAMINERS

Dr. S M Aminul Haque
Associate Professor & Associate Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Chairman



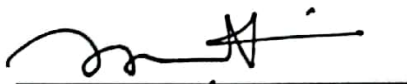
Fahad Faisal
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



Fourcan Karim Mazumder
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner




Dr. Mohammad Shorif Uddin
Professor
Department of Computer Science and Engineering
Jahangirnagar University

External Examiner

DECLARATION

I hereby confirm that I completed this project under the supervision of **Md. Mahfujur Rahman, Lecturer (Senior Scale) in the Department of Computer Science and Engineering** at Daffodil International University. I further affirm that the entire work or any portion of this work, has not been submitted elsewhere for the purpose of receiving any kind of degree or certification.

Supervised by:



Md. Mahfujur Rahman
Lecturer (Senior Scale)
Department of CSE
Daffodil International University

Co-Supervised by:



Mohammad Jahangir Alam
Lecturer
Department of CSE
Daffodil International University

Submitted by:

Mubtasim

Md. Mubtasim Fuad
ID: 171-15-8815
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

At first, I would like to express my sincere gratitude and thanks to the almighty God, whose divine favor made it possible for me to successfully complete the project for the final year.

I would like to express sincere appreciation and thanks to **Md. Mahfujur Rahman, Lecturer (Senior Scale)** in the Computer Science and Engineering Department at Daffodil International University, Dhaka. My supervisor's knowledge and expertise in the discipline of "*Machine Learning*" gave me the confidence I needed to complete the task. His patience, intellectual direction, encouragement, frequent and active supervision, helpful advice, informative counsel, and reading many substandard drafts and editing them made this effort possible.

I would like to thank **Dr. Touhid Bhuiyan, Professor and Head, Department of CSE** for assisting me in the completion of my project. I would also wish my gratitude to the officers and other faculty members at DIU.

I want to express my thanks to all my classmates from Daffodil International University who took part in the discussion while completing the work.

Lastly, the devotion & continual assistance of my parents during my undergraduate studies must be acknowledged.

ABSTRACT

Medical science uses the term "Chronic Kidney Disease" (CKD) to refer to a set of conditions that lead to kidney damage or a low Glomerular Filtration Rate (GFR). Medical advancements in recent years have allowed doctors to apply a wide range of techniques in the treatment of this illness. Recently, AI and ML have been increasingly adopted as a useful method for improving healthcare and medical research. The use of Machine Learning to detect the early symptoms of kidney condition can be a helpful approach as the disease may lead to a life-threatening condition. Different machine learning techniques, programs, and algorithms can be applied together to predict the steady progress of Chronic Kidney Disease. An appropriate result is produced by a machine-learning algorithm, and the algorithm with the highest performance among all others is chosen as the best one. Our web based system could allow doctors to determine the formation of the disease as soon as they receive the dialysis report. Also, the report analysis can help to figure out which elements in the human body are the root cause of Chronic Kidney Disease. Complex and dynamic algorithms such as Naive Bayes, Random Forest, KNN, Decision Tree, AdaBoost & XGBoost etc. are implemented in order to achieve optimal results in this system.

TABLE OF CONTENTS

CONTENTS	PAGE
Approval	ii
Declaration	iii
Acknowledgements	iii
Abstract	iv
List of Figures	xi
List of Tables	xii
 CHAPTER 1	
INTRODUCTION	1-6
1.1 Introduction	1-2
1.2 Motivation	3
1.3 Rationale of the Study	3
1.4 Research Questions	3-5
1.5 Expected Outcome	5
1.6 Report Layout	5-6
 CHAPTER 2	
BACKGROUND	7-12
2.1 Overview	7
2.2 Related Research Works	7-9
2.3 Comparative Analysis of the Study	9-11

2.4 Scope of the Problem	11-12
2.5 Challenges	12

CHAPTER 3

RESEARCH METHODOLOGY AND SYSTEM DESIGN 13-27

3.1 Introduction	13
3.2 Research Topic	13
3.3 Unsupervised Machine Learning Techniques	14
3.4 Supervised Machine Learning Techniques	14
3.5 Classification Techniques	15
3.6 Algorithm Specifications	15-16
3.6.1 Logistic Regression	16
3.6.2 K-Nearest Neighbors	16-17
3.6.3 Gaussian Naïve Bayes	17
3.6.4 Support Vector Machine	17
3.6.5 Perceptron	18
3.6.6 Decision Tree	18
3.6.7 Stochastic Scholar Gradient	19
3.6.8 Random Forest	19-20
3.6.9 XGBoost (Extreme Gradient Boosting)	20
3.6.10 AdaBoost (Adaptive Boosting)	20-21
3.7 Working Procedure of the System	21

3.7.1 Data Collection	22
3.7.2 Dataset from the UCI Repository	22
3.7.3 Data Pre-Processing	22
3.7.4 Data Normalization	23
3.7.5 Apply Different Algorithms	23
3.7.6 Analyze Mode	23
3.7.8 Choosing the Best Algorithm	24
3.7.9 Model for Web Implementation	24
3.7.10 Implementation of a Web Interface	25
3.7.11 Execution of the Model	25
3.7.12 Values for the Input Field	25
3.7.13 Predicted Outcome	25
3.8 Architecture of the System	26
3.8.1 User Segment	27
3.8.2 Web Insider	27
3.8.3 Machine Learning Model	27
 CHAPTER 4	
EXPERIMENTAL RESULTS & DISCUSSION	28-46
4.1 Introduction	28
4.2 Experimental Results	28
4.2.1 Data Gathering	29
4.2.2 Dealing with Null Values	30

4.2.3 Data Utilization	30-31
4.2.4 Feature Importance	31-32
4.3 Predicted Results & Discussion	32
4.3.1 Confusion Matrix	33-36
4.3.2 Classification Report	36-37
4.4 Analysis of the Result	37
4.4.1 Cross Validation Score	38
4.4.2 Accuracy	39
4.4.3 AUC (Area Under the Curve) Score	40
4.4.4 Jaccard Similarity Index	41
4.4.5 ROC (Receiver Operating Characteristic) Curve	42-43
4.4.6 Error & Misclassification	43-44
4.4.7 Standard Deviation	44
4.5 Web Implementation	45
4.6 User Interface of the Website	45
4.7 Analysis of the Website Output	46

CHAPTER 5

IMPACT ON SOCIETY AND SUSTAINABILITY 47-48

5.1 Introduction	47
5.2 Impact on Society	47
5.3 Ethical Aspects	48
5.4 Sustainability	48

CHAPTER 6

CONCLUSION AND IMPLICATION FOR FUTURE WORK 49-51

6.1 Introduction 49

6.2 Implications for Future Research 49

6.3 Recommendations 50

6.4 Conclusion 50-51

REFERENCES 52-53

LIST OF FIGURES

FIGURES	PAGE NO
Figure 3.1: A Proposed System Design for the Prediction of CKD	21
Figure 3.2: Architecture of the System	26
Figure 4.1: Cross Validation Score Chart	38
Figure 4.2: Accuracy Score Chart	39
Figure 4.3: AUC (Area Under the Curve) Score Chart	40
Figure 4.4: Jaccard Index Chart	41
Figure 4.5: Receiver Operating Characteristic Curve	42
Figure 4.6: User Interface of the Website	45
Figure 4.7: Negative Web Output	46
Figure 4.8: Positive Web Output	46

LIST OF TABLES

TABLES	PAGE NO
Table 4.1: Accuracy of Best 3 Algorithms	28
Table 4.2: Data Gathering & Null Values	29
Table 4.3: Description of the Dataset	31
Table 4.4: Feature Importance of each Attribute	32
Table 4.5: Specification of a Confusion Matrix	33
Table 4.6: Confusion Matrix for each Algorithm	34
Table 4.7: Classification Report using 4 Different Metrics	37
Table 4.8: Cross Validated, Accuracy, AUC & Jaccard Score	43
Table 4.9: Errors & Misclassifications	44
Table 4.10: Standard Deviation	44

CHAPTER 1

INTRODUCTION

1.1 Introduction

A renal illness, persistent for an extended length of time is referred to as having Chronic Kidney Disease. This prevalent condition is frequently associated with aging. Anyone can be affected by it, although people of color, particularly those from South Asia and the Caribbean, are more likely to be affected with this condition. It is possible that CKD could worsen over time and the kidneys will eventually stop working totally, although this situation rarely happens. Despite having CKD, many people live their lives to the fullest. Many scientists from different parts of the world have spent time trying to figure out what causes this renal illness.

Kidney illness has emerged as a major problem in modern medicine. The most dangerous kidney illnesses include Chronic Kidney Disease, Glomerulonephritis, Kidney Stones, Urinary Tract Infections, and Polycystic Kidney Diseases. High blood pressure is a major contributor to the development of Chronic Kidney Disease. The filtration units of the kidneys (glomeruli) are likely to be damaged from high blood pressure. In the kidneys, small blood arteries called glomeruli are responsible for blood filtration. Over time, the increased pressure hurts the renal arteries, and kidney function gets worse. Eventually, the kidneys will lose their ability to function. Dialysis treatment would be necessary in such a case. Dialysis is a method of purifying the blood by removing impurities and extra fluid. Kidney failure is treatable with dialysis, but it is not a cure. The effectiveness of a kidney transplant for a given patient depends on many factors. In many cases, diabetes is the direct cause of CKD. It is one of the many illnesses that can lead to increased blood sugar levels. Consistently elevated blood sugar levels damage renal blood vessels. Because of this, the kidneys are not able to filter the blood as they normally would.

Kidney failure can occur if the body is subjected to excessive amounts of pollution. Chronic Kidney Disease is more deadly than previously thought, as demonstrated by studies conducted by researchers from a variety of geographical areas. Male patients with Chronic Kidney Disease used to have a greater mortality risk than female patients, as shown in several studies. As of 2020, the World Health Organization reported 10,841 deaths in Bangladesh were caused by kidney disease. This is 1.51 percent of total deaths in the country [1]. CKD patients who are exposed to COVID-19 are at an increased risk of getting a potentially fatal condition, which creates difficulty in getting access to dialysis as well as other forms of medical care [2]. Machine learning algorithms can be broken down into numerous different categories. Some examples of popular approaches include reinforcement learning and other types of ML techniques such as unsupervised, supervised & semi-supervised. Supervised ML Algorithms like Decision Tree, KNN, Perceptron, Naive Bayes, Random Forest, AdaBoost & XGBoost are utilized to determine chronic renal disease in this research. Anybody can use this approach to calculate their own personal risk of developing severe kidney disease. However, a web-based implementation procedure may be the most effective solution for the pathologist and doctors if they wish to determine the precise probability of being affected by CKD. The objective of this work is to construct a model using a suitable dataset, diagnose CKD at any stage using an appropriate algorithm. In this study, the best algorithm is selected among the most important Machine Learning algorithms, and then the data is transferred to a web-based application where patients, doctors, and pathologists may access the findings and use the values to predict the likelihood of kidney illness. The near future of the modern era promises to be more advanced technologically. Knowing probable information from web implementation, this study will help people better understand the causes of kidney illness and take the required steps to treat the condition.

1.2 Motivation

Very few individuals in Bangladesh know how to keep their kidneys healthy. Nobody knows if they have kidney illness or not. It is estimated that kidney problems are responsible for 1.2 million fatalities per year around the world. There are around 18 million people living in Bangladesh, and each year approximately 35,000 to 40,000 CKD patients progress to kidney failure. In a study, researchers found that persons over the age of 40 had a significantly increased risk of developing kidney illness. There are not many studies on CKD prediction that have demonstrated greater efficiency. Aside from that, we observed that the increased focus of the modern world is mostly on recommendation systems. As a result, people rely on the system to recommend just the best options for them. A system designed to provide recommendations must be capable of making decisions independently. To make decisions without consulting anybody else, classified data is required. All these factors motivated us to carry out the study, in which we would use classification and prediction techniques to facilitate the prevention of CKD.

1.3 Rationale of the Study

There are various kidney disease prediction works. However, the use of so many classification algorithms on a dataset related to Chronic Kidney Illness is not very common. Even though there have been a lot of studies conducted on Chronic Kidney Disease, their conclusions aren't all that different from one another. Therefore, we use 10 different classification methods to determine which ones produces the most accurate findings, with a view to apply the technique in the future.

1.4 Research Questions

This research raises several questions. A bunch of questions has been taken from various people for making this research more reliable.

What prompted this research towards chronic kidney disease prediction?

Chronic Kidney Disease is a big health concern all over the globe. The situation of a CKD patient deteriorates with time. It progresses through several phases, with the final one being complete kidney failure. A person eventually dies as a result. But it should be noted that CKD can be treated effectively with the right medication and by adhering to several rules and regulations, it can be identified at an early stage. For this reason, CKD was the focus of the study.

Why should we use machine learning? How dependable is it?

One of the most popular methods for doing prediction work is Machine Learning. A model can give precise predictions if it has been trained on a reasonably large dataset. Using a machine learning technique on a medical dataset, it is possible to make accurate predictions about Chronic Kidney Disease. An era of global modernization is taking place in the present day across the world. If one imagines a time roughly ten years ago, when Artificial Intelligence and Machine Learning were not quite as advanced as they are now, one will recall that these core points were nothing more than a name associated with some mathematical logic. However, AI is currently essential to the functioning of half of all technological systems on Earth. Consequently, with sufficient training and improved accuracy, this area can become even more trustworthy than it already is.

Why do people typically make use of a web interface?

To enter values into a machine, a web interface is a helpful tool. It offers general users a very flexible approach to access the resource. Our application can be accessed from any location with an internet connection and a web browser. An interface that is based on the web can also generate results more quickly. Using this web interface, one can determine the general status of his or her kidneys at home. This makes it a practical tool for predicting chronic kidney disease.

Why do we need 10 different algorithms to do the same thing?

Ten algorithms were tested in search of one that would work best with the CKD dataset. The best-suited algorithm, with the lowest error rate and maximum accuracy rate, was found by comparing 10 different algorithms. If only one method was used, it would be difficult to determine which one is optimal because no one can predict which algorithm will work best with a given dataset.

1.5 Expected Outcome

Throughout the time of this study, the main idea or what was anticipated to happen has changed a few times. It helped to make clear what the results of this study are. The expected outcome of this study may help in the early detection and prevention of Chronic Kidney Disease, as well as in the identification of its root cause. By using the system, doctors or researchers might also be able to figure out at what aged people get Chronic Kidney Disease (CKD). For medical research, an in-depth computational and algorithmic analysis of CKD could be made public. This study's final approach is to create a web-based system that suggests prospective patients who should be contacted by a doctor.

1.6 Report Layout

To make the results of the study easier to understand for the researcher, it is divided into six chapters.

In Chapter 1, an important introduction to the overall research project is given. In a nutshell, this relates to information about CKD. This chapter outlines the goal of the study, its motivation, relevant research questions, expected results, overall management details and economic factors.

In Chapter 2, background information for this study is discussed in depth, such as how data values are classified, how machine learning systems work, and what other research has been done on similar topics. This chapter also describes the scope of the problem statement and the perceived difficulties with comparative analysis.

In Chapter 3, the research methods, proposed system, and system architecture are described. Each implemented algorithm's details, from their mathematical foundations to their current state, are outlined in this chapter.

In Chapter 4, the full analysis of the results of each step is given. Best accuracy score, best algorithm, cross validated score, Jaccard score, classification report, and confusion matrix are described in this chapter. The ROC-AUC curves for each algorithm are also given here. The chapter concludes with a discussion of error statistics, including topics such as standard deviation, misclassification, MAE, and MSE.

In Chapter 5, "Ethical Aspects," the impact of this research on society is explained. This is the most important part of any influential and effective research work. The final section of this chapter discusses the long-term feasibility of this study.

In Chapter 6, we get a glimpse of the study's future growth through a brief description of its extension. In this final section of the research report, the most important findings are summarized for the reader's convenience.

CHAPTER 2

BACKGROUND

2.1 Overview

For a long time, chronic kidney illness has been a big concern. Therefore, the history of this disease includes a great deal of traumatic events and a great number of deaths. There has been a significant amount of effort put into the prevention of CKD to save lives. In this chapter, the basics of this disease as well as its historical context have been addressed. This chapter has covered some of the related work that has been done on this domain. Finally, some comparison analysis has been presented to show how significantly the work that is being presented here has been affected by previous work.

2.2 Related Research Works

The study of Lambert et al. [3] tries to predict CKD using just nominal parameters, which are both numeric and nominal in nature. The CFS approach was utilized in this research to identify and categorize essential characteristics as either CKD or not CKD. Both the classification and the prediction of this method contain this approach. To select characteristics, CFS can be used with nominal, numerical, or nominal + numerical data. The outcome of the CFS is compared to 3 distinct ranking methods for the purpose of function selection. These methods include the information gain, the gain of the ratio, and the relief methodology. The selection accuracy based on correlation and minimal sequential optimization (CFS-SMO) approach produced results with an accuracy of 95.25% (for numerical), 98.5% (for nominal), and a percentage of 98.5 for the combination of nominal & numerical.

The results of this experiment demonstrated that the CFS selection successfully extracted features from the dataset of CKD, and that SMO recognized renal disease as a suitable benchmark condition. Thus, the CFS-SMO is viewed as a viable technique for accurately diagnosing renal illness and guiding doctors toward the best course of action.

Nusinovici et al. [4] evaluate the success of ML algorithms for predicting Cardiovascular Disease, Chronic Kidney Disease, Hypertension, and Diabetic Mellitus, as well as the accuracy of simple clinical predictions, in a prospective observational research. Five additional ML models were compared to the basic logistic regression model, including a neural network with a single hidden layer, support vector machine, gradient boost, random forest and KNN classifier. Logistic regression had the largest area under the operating curves of CKD & DM for disease prediction (0.905 [0.88, 0.93] and 0.768 [0.73, 0.81] respectively). Among the most successful models, the one that combines gradient boosting, neural networks and logistic regression was discovered to be the best.

To find the best set of features for initial diagnosis of chronic conditions, the proposed research algorithm, named as ITLBO (Improved Teacher Learner Based Optimization), uses the distance requirement of Chebyshev for determining the fitness function & frequently used control parameters such as size of the population and generational time. When applied to CKD data sets, the stated function selection strategy led to a 36% reduction. This is incredibly impressive when compared to the traditional TLBO method, which led to a 25% reduction in features. The ideal feature subsets determined by both TLBO and ITLBO methods are validated by the analysis of SVM, Gradient Boosting approaches, and the Convolutional Neural Network method. This ITLBO method is different from the one described in the paper of Balakrishnan et al. [5], experimental results reveal that the three algorithms produce better overall classification accuracy for the created feature subset.

The work by Ali et al. [6] examines the problems with the wide usage of automated prediction systems by examining the diagnosis of Chronic Kidney Disease (CKD) in developing nations. This research offers a more practical approach to group-based feature selection by introducing a cost-sensitive grouped feature ranking mechanism. This research is groundbreaking because it shows for the first time that cost-sensitive ensemble ranks for non-cutting groups have a chance of achieving the desired results of cheap cost and high accuracy. In the experiment, they used eight different methods of comparison selection and seven different popular classification algorithms to demonstrate the effectiveness of the method. This paper concludes that the practicality of automatic CKD systems can be enhanced by including the cost factor in the goal space of solution formulations. As a result, an affordable and reliable approach for detecting CKD has been identified.

Segal et al. [7] analyzed commercial health insurance claims for 10 million people, drawn from 550,000 individual profiles. The inclusion criteria were patients older than 18 who were diagnosed with stage 1–4 of CKD. A total of 240 predictive factors have been collected and organized into six distinct sets of features. Using a technique called feature embedding, they were able to acquire temporal features on these three primary data components (diagnostics, procedures, and medications) by applying the Word2Vec algorithm. They used the gradient booster technique (XGBoost) for their investigation.

2.3 Comparative Analysis of the Study

A method for detecting chronic renal disease from the test samples of saliva was suggested in a previous study. It uses deep learning and a new detection method. In order to deal with the difficulties presented by the data, they turned to a combination of CNN & SVM. The accuracy of the CNN-SVM network was 97.67% on average, with a specificity of 97.53% and a sensitivity of 97.50%.

Overall, the average accuracy of the base CNN model was 96.51%. Comparing the suggested system's performance to that of other algorithms revealed that it outperformed several popular data classification methods. In contrast, the writers of this work achieved an AUC (Area under the Curve) score of 98.21% and a Jaccard index of 96.14% in the XGBoost Classifier, both of which are improvements over prior research [8].

The paper's data was collected from the ML repository of UCI (University of California, Irvine) that included many blanks. K-Nearest Neighbor (KNN) imputation was used to establish the similar work in Qin J.'s paper. Models were established using six different machine learning techniques in this work. With a higher diagnostic accuracy of 99.75%, Random Forest surpassed the other models of machine learning [9]. The current models' errors were analyzed, and then a new model was provided that combines Decision Tree, AdaBoost, and XGBoost, and it achieved 98.60% accuracy in average over the course of 10 simulations. Both UCI data and additional data acquired from several hospitals in Bangladesh are included in the authors' study. The total number of records counted was around 10321. To discover the most accurate algorithm for this dataset, the authors of this research tested ten different ones.

In this article, the prediction of chronic renal disease serves as an example of a health care service that can be provided through the cloud computing environment. This study could provide a smart model for the prediction of CKD based on cloud-IoT with the use of two methodologies, Neural Networks & Linear Regression. As shown in the results, the intelligent hybrid model predicts CKD with an accuracy of 97.8 percent. Around ten algorithms are included in the authors' planned study. The optimal algorithm was proposed after evaluating the accuracy of all ten algorithms. To anticipate the outcome, the best algorithm was used to build a model, and then that model was used to create a user-friendly interface [10].

One strategy put forth in the prior research was to give priority to establishing a correct diagnosis rather than determining the most effective treatment. The major objective is to compare the two algorithms for predicting chronic renal disease and choose the one that provides the most accurate results. The Random Forest method and the BPNN (Back Propagation Neural Networks) were utilized as two of the data mining strategies. Back Propagation achieves 98.40% accuracy, whereas the Random Forest method achieves 88.7%. The authors have used 10 different algorithms in the study. Four of these algorithms had an accuracy greater than 95%. For example, XGBoost Classifier, Decision Tree, AdaBoost Classifier and Random Forest Algorithm has the accuracy of 98.55%, 98.11%, 98.11%, and 97.09% respectively [11].

In the past, a method for determining the presence of CKD based on clinical data was given. This method included data preparation, a technique to manage missing values, collaborative filtering, and feature extraction. The additional tree classifiers and the random forest classifiers are shown to produce the greatest accuracy (100.00%) out of the 10 ML methods examined. The study takes into account the real-world challenges of data collection and emphasizes the importance of using domain knowledge while applying machine learning to anticipate CKD. The authors of this research have selected the most effective model to predict CKD at every stage. The writers have also built an easy-to-use interface where individuals may enter the information needed to predict CKD and receive a result. The paper's interface can be adapted for use at other hospitals, and it can be turned into a website in order to be accessed online in the later [12].

2.4 Scope of the Problem

Kidney ailment has become an increasing public health concern in recent years. It continues to increase daily at a frightening rate. There is not an appropriate model that could be applied to make the prediction of it.

Because of this, there are several possibilities to deal with this issue by examining the symptoms of kidney disease to establish whether a patient has chronic kidney problems or not. Identifying the relevant dataset and using multiple ML techniques, including a specific model with training and testing, are necessary to resolve this issue. To determine if there is a link between the attributes of the dataset, we must examine the correlations between them. This type of automated diagnosis will improve healthcare quality in Bangladesh. If such a system exists in different hospitals, renal patients can start receiving various treatments at an earlier stage. If the disease can be identified in its early stages by this method, it will have a lower risk of progressing to the chronic stage, which would be a significant step forward from our point of view. If patients are not allowed to progress to the chronic stage, doctors can quickly cure them with correct treatment. When such a predictive system is developed because of our work, there is a significant opportunity to improve the well-being of people.

2.5 Challenges

Before the kidneys can begin the process of producing urine, they are required to go through several intricate processes of excretion and reabsorption, which can take many hours to complete. This process must work properly to maintain the body's chemical balance. However, in people with chronic kidney disease, this process is slowed down, sometimes to a stop. Finding out if someone has a kidney disease might be challenging without knowing how much liquids they consume. The most difficult aspect of this study has been identifying the specific criteria that need to be met in order to diagnose chronic renal illness. Getting rid of all the null values from the dataset is a challenging task. This could be a lengthy and time-swallowing job also. Finding an appropriate algorithm was also a significant challenge. Multiple algorithms were used to train the dataset, and the one with the highest accuracy in detecting CKD was selected for further analysis. One of the most difficult jobs for the team was developing a user-interface for this model so that anyone may enter data and make a CKD prediction at any time.

CHAPTER 3

RESEARCH METHODOLOGY & SYSTEM DESIGN

3.1 Introduction

The research methodology is extremely important before beginning any research. It is necessary to identify a problem before developing strategies to deal with it. This chapter presents background information about the study's topic. Furthermore, the algorithms that were applied to solve the issue were explained, and the methodology was outlined using a graphic diagram for the ease of comprehension. Moreover, an architecture of the system was shown for clearer understanding.

3.2 Research Topic

The primary objective of a research project is to identify a problem that can be used as a topic for the study and then to discover a solution to that particular problem. One of the most prevalent diseases of old age is chronic kidney disease. It's a chronic issue that may worsen with time. Kidney disease causes a high number of deaths annually. When CKD gets to its last stage, the kidneys work less than 15% of the time. Kidney disease is treatable if it can be diagnosed in its early stages. That's why research into nephrology and the development of a model to predict chronic kidney disease (CKD) at different ages are so crucial. Individuals will have the ability to access their report from the comfort of their own houses via a user-friendly web interface.

3.3 Unsupervised Machine Learning Techniques

In unsupervised learning, models are trained on unlabeled datasets and then allowed to make their own predictions about the data. Since we only have access to input data and no associated output data is used in unsupervised learning, it cannot be used to solve a classification or regression problem. Unsupervised learning finds underlying structure of a dataset, groups similar data, and reduces it. As unsupervised techniques cannot be applied for classification problems, we have used only supervised ML techniques to predict the value of target attribute in our thesis.

3.4 Supervised Machine Learning Techniques

With the help of supervised machine learning and categorization, it is easier to sort things into several categories. An ML system is one that can make judgments on its own, without human interference, by continuously collecting and analyzing data. It is feasible to develop a system that continually gets better by gaining knowledge from one's previous experiences, making analytical judgments, and utilizing various other methods. There is a wide range of options available when it comes to machine learning techniques. Techniques of supervised machine learning are utilized quite frequently in this work. Predictions can be made from labelled data of the past using supervised machine learning methods. In order to make output value predictions, the training method generates an inferred function by analyzing a large dataset. After that, the system can be trained as much as it needs to be. The learning algorithm might detect mistakes and modify the model if it compares its results to the original one. A variety of supervised learning methods are applied to the problem of chronic kidney disease prediction in this thesis.

3.5 Classification Techniques

Data classification is a method of data analysis that uses input data to determine which classifications within that data are most relevant. It has surpassed all other machine learning methods in terms of popularity and adoption. In the context of supervised learning, these models, known as classifiers, can make predictions about predefined classes. All the predictions are done separately. Classifier does not provide any intermediate values. To determine if an image shows a dog or a cat, a classifier may be created. There are two possible outcomes: "dog" or "cat." It is not possible to use a classifier to get a value in between. To use a classification learning technique, the data must be labeled. In classification learning, two unique types of datasets are used. The first category is referred to as "Training Data," while the other one is "Test Data." The model is developed with the help of training data, and its accuracy is then verified with test data. The classification procedure consists of two parts. During the training phase, a classifier is built using a suitable method and training data, and then tested in a real-world environment. A classifier is formed by combining a classification method and training dataset. Basically, a classifier is just a collection of rules which can be applied to different situations in different ways. The prediction phase involves applying the model created during the learning phase to a set of unknown data to make a prediction about the class of the target attribute. Here, we utilize the test data to determine if the model's predictions are accurate.

3.6 Algorithm Specifications

Ten of the top Supervised ML techniques are used to accomplish this study. An algorithm, in general, can be thought of as a structured set of instructions that tell computer software how to convert a set of input data into useful data. Statistics are facts and knowledge that can benefit humans, robots, or a machine to perform a specific operation.

Machine learning algorithms follow a similar logic and mathematical procedure. Each Machine Learning Algorithm uses a unique set of mathematical transformations. Furthermore, this research project includes the most popular machine learning algorithms having the relevant algorithmic processes that are part of the complete system design.

3.6.1 Logistic Regression

One type of classification method that can provide a prediction about the outcome is logistic regression. Logistic regression has one of two possible states, either 0 or 1. Since this system predicts CKD or Not CKD, the algorithm can easily understand and anticipate the outcome. By integrating features, this model may solve more challenging issues. The Y-axis has values from zero to one. This is due to the fact that the sigmoid function utilizes these 2 points as minimum and maximum, which is perfect for classifying data sets into two different groups. After computing the value for the sigmoid function of X, the system obtains a value of probability in between zero and one. The observation must be fitted into one of the two categories.

The formula for the sigmoid function is:

$$\text{sigmoid}(x) = \frac{1}{1 + e^{-x}}$$

3.6.2 K-Nearest Neighbors

KNN algorithm, a basic supervised classification strategy, may also be used to address regression and classification challenges. at the same time. KNN can be easily understood and implemented. The foundational principle of KNN is the Euclidean distance. Since the dataset is divided into two classes, KNN is used in this classification.

The formula for the KNN algorithm can be written as follows:

$$distance = \sqrt{(x1 - x2)^2 + (y1 - y2)^2}$$

3.6.3 Gaussian Naïve Bayes

The Naive Bayes variant known as Gaussian Naive Bayes allows both Gaussian normal distributions and continuous data. Study of the Naive Bayes classifiers includes more than one supervised algorithms. Bernoulli, Multinomial, and Gaussian Classifiers are three types of machine learning techniques based on the Bayes theorem. The Gaussian-Naïve Bayes classification method is simple and very helpful. When working with continuous data, it's common to assume that the values of each class are distributed based on a Gaussian distribution. The complete formulization procedure for the Gaussian Naive Bayes algorithm is derived in the given equation.

$$P(X | Y = c) = \frac{1}{\sqrt{2\pi\sigma_c^2}} e^{\frac{-(x-\mu_c)^2}{2\sigma_c^2}}$$

3.6.4 Support Vector Machine

Since the SVM Classifier is an ML technique, it's important to keep monitoring how it performs. It can be applied to solve classification and regression-related problems. Although it can be applied to a variety of tasks, classification is the most common use of it. It was found that the developers of this technology failed in their attempt to plot data items on a graphing calculator in n-dimensional space. This study evaluated two-dimensional data charting based on CKD or Not CKD.

3.6.5 Perceptron

Perceptron is an artificial neural network model used in binary classification systems. A node or neuron takes inputs from a set of data rows and predicts classification. In this approach of prediction, a feature vector along with a prediction function are used, which are both linear classification methods. In this portion, we discussed about the process of making a threshold function that takes an input x and returns a binary value as an output $f(x)$. The input binary value x is converted to the output binary value $f(x)$. The activation function for the perceptron algorithm can be stated as follows:

$$f(x) = \begin{cases} 0, & w \cdot x + b \leq 0 \\ 1, & otherwise \end{cases}$$

3.6.6 Decision Tree

Decision trees can be used to solve classification and regression challenges in supervised learning. Classification problems are the most common use of a decision tree. The internal nodes represent the attributes of the dataset, the branches provide the rules for making decisions, and the leaf nodes reflect the results. To predict the values of the target attribute from the training dataset, the Decision Tree can be used to build a learning model with simple decision logic. To form a decision tree entropy is needed. It is a metric in information theory that quantifies the impureness or uncertainty of a set of observations. It informs a decision tree on how to split up the data. If we know the value of entropy, it is simple to calculate the information gain of a node. The value of entropy can be given by the formula:

$$Entropy = -p \log_2 p - q \log_2 q$$

3.6.7 Stochastic Scholar Gradient

For building linear & regressive classifiers using convex loss functions, like the linear SVM & Logistic Regression, the Stochastic Gradient Descent approach provides a very simple and efficient approach. SGD has been around for a while in the field of machine learning, but it has recently gained a lot of popularity in the field of learning large-scale datasets. A basic SGD learning process is offered by the class SGD Classifier, which supports a number of loss functions used for classification and misclassification problems. A linear SVM-like SGD Classifier that was trained using the hinge loss can be treated as equivalent to the Support Vector Machine. The algorithm for SGD classifiers includes a series of retraining instances such as $(x_1, y_1) \dots (x_n, y_n)$ where x and y values have some predefined constraints. The system needs to take into consideration the sign of $f(x)$ to provide accurate estimations for binary classification problems. This system is required to build a scoring function that is linear in nature based on the model parameters. Function, $f(x) = w^T x + b$ gives the value such that intercept $b \in R$ and model parameters, $w \in R^m$ are met. The normalized training error equation is used to determine model parameters.

$$E(w, b) = \frac{1}{n} \sum_{i=1}^n L(y_i, f(x_i)) + \alpha R(w)$$

Where L describes the loss function used for the evaluation of a model and R represents the normalization term which is responsible for prohibiting the complexity. A positive linear combination, $\alpha > 0$ is responsible for the regulation of the normalization strength.

3.6.8 Random Forest

Random decision making forests, also known as random forests, are a supervised type of learning technique used for classification, regression, and other purposes by constructing multiple decision trees. When used for classification, a random forest's output is the

category supported by most of its individual trees. In the case of regression problems, the average prediction across all trees is given back. The following equation describes the parameters used for the prediction of future outcomes when we use Random Forest algorithm to a specific problem.

$$\sigma = \sqrt{\frac{\sum_{b=1}^B (f_b(x') - \hat{f})^2}{B - 1}}$$

The sample number "B" is a free parameter. Cross-validation or observing the out-of-bag inaccuracy can aid in determining the best count of B trees: the mean error for the prediction of each training x' sample are the trees which contains no x' in their validation set. When several trees seem to be fitted, training and test mistakes start to get worse.

3.6.9 XGBoost (Extreme Gradient Boosting)

The XGBoost method utilizes a gradient boosting framework, and it is an ML tool for making choices. It uses a sequence of categorized and regressed (CART) trees as base learners at the start, and then boosts tree performance by producing a collection of trees that minimize a regularized target function. Split-wisdom discovery in distinct trees, cache-friendly approximate division algorithms, and efficient out-of-core gradient boosting computing were incorporated into the algorithm to produce a fast & accurate prediction-making.

3.6.10 AdaBoost (Adaptive Boosting)

AdaBoost classification, short for "adaptive classification boosting," is a group learning method. It constructs a robust classifier by combining the outputs of several less effective ones. In this approach, a bad classifier improves based on its past misclassifications. We must think about a dataset of n-sample. At first, we give 1/n weight to each sample. This

dataset is used to construct a mediocre classifier. The error, ε is computed for the entire classification process by this classifier. Total error in classifying data samples measures the influence of a classificatory, α . We use α in the equation to reweight the samples in the dataset, resulting in a new dataset.

$$\alpha = \frac{1}{2} \ln \left(\frac{1 - \varepsilon}{\varepsilon} \right)$$

3.7 Working Procedure of the System

After considering each algorithm, the required system architecture can be proposed. Figure 3.1 shows a system diagram that helps understand its procedure.

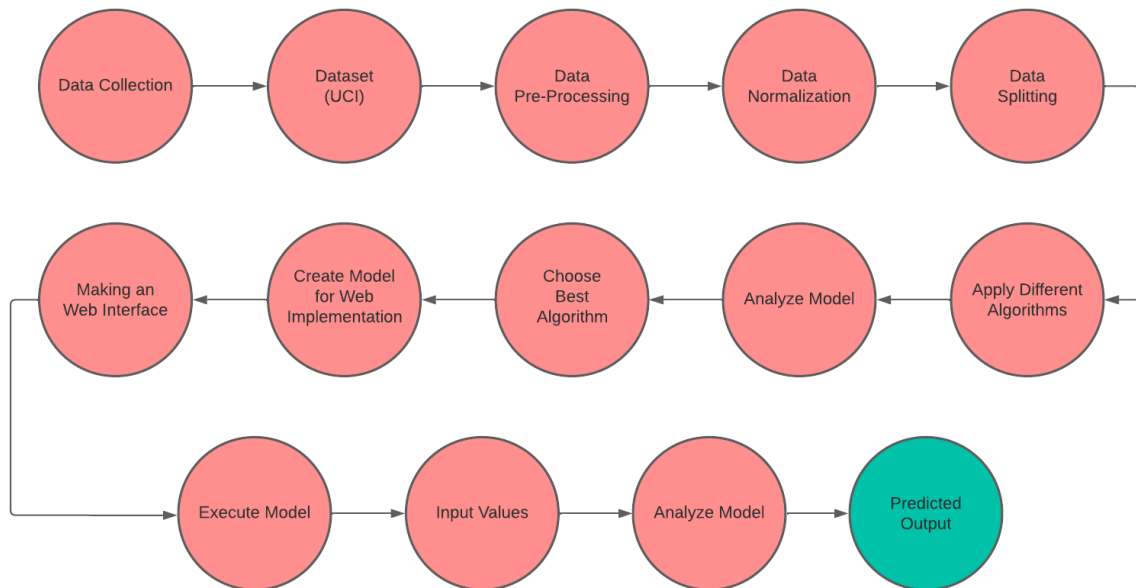


Figure 3.1: A Proposed System Design for the Prediction of CKD

3.7.1 Data Collection

There was a requirement for real-world data for the system to perform CKD analysis. The information used in this analysis came from the University of California, Irvine (UCI). I hope to do future research in Bangladesh by collecting data from different Kidney Hospitals in Bangladesh including the National Institute of Kidney Diseases & Urology, Bangladesh. The first step in this research was to collect the information into a single comma-separated values (CSV) file for readability and comprehension.

3.7.2 Dataset from the UCI Repository

When we combined all the data sets into one CSV file, the total number for the rows was just right for using a variety of Machine Learning Algorithms. To make accurate predictions, machine learning algorithms require a massive amount of information. There are some blanks in the raw data set, which includes 25 columns and 10321 rows.

3.7.3 Data Pre-Processing

Qualitative information and missing values must be converted from the dataset so that the algorithms could be applied easily. Initially, the data values that are qualitative in nature was transformed into numerical form. Following that, we addressed the issue for missing values. The average value was used to fill in all the blanks. The data was separated into a set of dependent and independent variables, X and Y respectively.

3.7.4 Data Normalization

Normalization is the process of transforming a set of numbers into a uniform scale while retaining their original ranges. The next step, which improved the accuracy significantly, was to standardize the independent attribute values (X). For this case, Z-Score normalization was used as a standardization technique.

3.7.5 Apply Different Algorithms

Ten distinct algorithms were tested and utilized in an effort to determine which one provided the most accurate results. The ten algorithms are as follows: Decision Tree (DT), Support Vector Machine (SVM), Random forests (RF), K-nearest Neighbors (KNN), Stochastic Gradient Descent (SGD), Gaussian Naive Bayes (GNB), Adaptive Boosting (AdaBoost), Perceptron Algorithm (PA), eXtreme Gradient Boosting (XGBoost), Logistic Regression (LR). A variety of analytical findings were made using these different algorithms.

3.7.6 Analyze Model

The data was turned into tables after being measured using the Cross Validated Score, Jaccard Score, Accuracy Score, Confusion Matrix, Area Under the Curve (AUC), Misclassification, Mean Squared Error (MSE), and Mean Absolute Error (MAE). The confusion matrix provides a concise overview of the degree to which the facts can be accurately predicted. The percentage of correctness of the predicted data can be found in the Cross Validated Score, Jaccard Score, Accuracy Score, and Area Under the Curve (AUC). The error for an algorithm is then measured in terms of misclassification, MAE, and MSE.

3.7.8 Choosing the Best Algorithm

All required outcomes were measured and observed in tables to figure out the best algorithm to use. The selected algorithm has the highest precision and lowest mean squared error of all possible algorithms applied to that dataset. Developing an appropriate algorithm is the first step in making effective use of the dataset. Many algorithms can be used as models, and the best one can be selected later. In this study, we used a number of analytical criteria to find the best approach. These included Cross Validation Score, Jaccard score, Accuracy Score, Area Under the Curve (AUC), and others. Based on the results of this research, the Decision Tree Classifier is the most effective approach when applied to the CKD dataset. It got the highest scores for all of the things listed above. Once an algorithm has been chosen, the procedure can move forward.

3.7.9 Model for Web Implementation

The authors have developed a web-based interface after choosing the most appropriate method. A "Pickle" was utilized to establish communication between the interface and the top algorithm. Here, pickle is a package used in Python for serializing items. Machine learning algorithms can also be serialized and saved to a file using the pickling method. The writers here have saved the chosen model as a "model.pkl" file, a serialized model format. A model cannot be built without first having access to an object of the chosen algorithm. Using the fit() method, the training dataset is employed to train the model. Once the model has been trained using the proper algorithm, it is ready for usage. The model is saved to a file in the next step, and then it is loaded as a different object called pickled model. The prediction accuracy score and test data are then calculated using the loaded model.

3.7.10 Implementation of a Web Interface

Python's "flask" module was used to create a web interface with its assistance. Creating a user-friendly interface also required having a fundamental understanding of HTML and CSS. In the website's backend, a pickle file was linked.

3.7.11 Execution of the Model

After building the model, it must be stored in a folder. To save the model, Pickle's dump() function is used. This serializes the object and saves it as model.pkl. This model can then be stored or committed to Git and run on unknown test data without retraining the model from scratch. To deserialize a "pickled" model, it is provided to Pickle's load() method. After that, the predict() function of the original model can be executed to get predictions as a form of an array.

3.7.12 Values for the Input Field

Once the model has been developed, a simple interface for predicting CKD is built using flask. The interface will allow its users to input data and predict the outcome. Flask is just a Python interface for creating web apps. Flask's framework is clearer than Django's and has less fundamental code to develop a simple web-App.

3.7.13 Predicted Outcome

The website's user interface makes it simple for anyone to input information that can be used to anticipate CKD. If a patient was found to have CKD, it will simply display a positive outcome. Also, the website will show a negative outcome and reassure the patient that everything is well if the patient does not have any symptoms of CKD.

3.8 Architecture of the System

System architecture is essential to simplify machine learning approach and online implementation for the whole project. In Figure 3.2, an overview of the proposed system's basic architecture is shown.

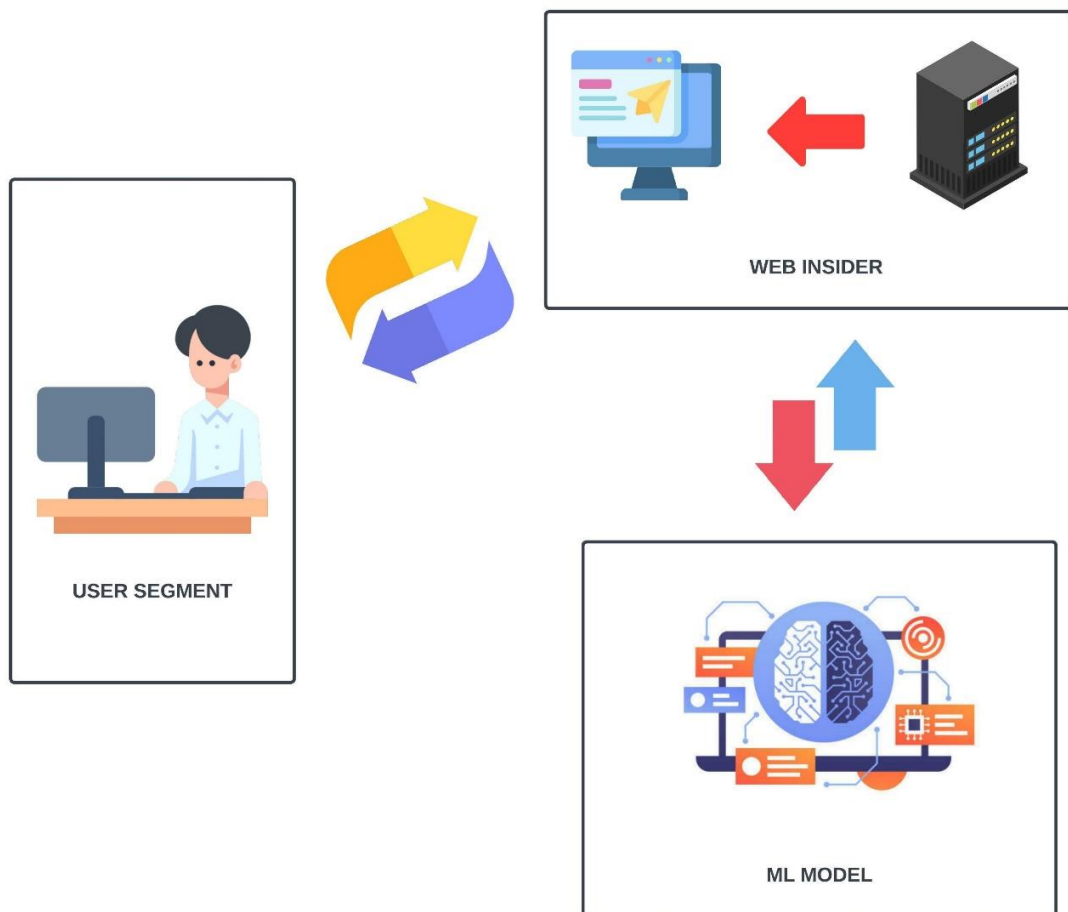


Figure 3.2: Architecture of the System

3.8.1 User Segment

In this user segment, we see a user checking their own kidney health with the use of a machine. A user on a laptop will input the values for the attributes of the diagnosis report. When this is done, the system will provide the user with a general overview of their kidney health. The design of the user segment is quite basic. All users will find the interface to be intuitive and simple to navigate. The user only needs to enter the value and then press the predict button for the machine to accurately predict the current state of the kidney.

3.8.2 Web Insider

The "flask" library in Python was used to create the web interface. It was also necessary to have a basic understanding of HTML and CSS to design a functional and attractive user interface. In the website's backend, the pickle file was linked. The website's user interface makes it easy for anyone to input information that can be used to anticipate Chronic Kidney Disease. The website will indicate a positive result if a patient has Chronic Kidney Disease and suggest seeing a doctor. Furthermore, the website will show a negative result and reassure the patient that everything is fine if the patient has no symptoms of CKD.

3.8.3 Machine Learning Model

By evaluating and contrasting the data in the tables, we were able to determine the optimal algorithm. In that dataset, the selected algorithm has the highest rate of accuracy and the lowest rate of error. After the top algorithm was isolated, a web-based user interface needed to be developed. To link the user interface to the top algorithm, a "pickle" was employed. Serializing ML algorithms and saving them in a file can also be done with the help of pickling process. In this scenario, the chosen model was serialized into a "mode.pkl" file.

CHAPTER 4

EXPERIMENTAL RESULTS & DISCUSSION

4.1 Introduction

The results are crucial to the success of any project or research. Considering that the end result is the sum total of any effort. Detailed tabular results are presented in this chapter. This chapter shows how to understand the CKD dataset by diving into the specifics of data collection, data utilization & feature importance. Then, the results of various algorithms are presented in a confusion matrix table. Precision, Recall, and F1-Score are tabulated in a classification report. Multiple graphs depict the accuracy results, Jaccard scores, cross-validated scores, AUC scores, and ROC curves. The information is also presented as a table for ease of reference. There's a table detailing the standard deviation, too. Last but not least, a table displaying the errors and incorrect classifications was provided.

4.2 Experimental Results

Each algorithm demonstrated its own accuracy & scores after ML model was successfully implemented; these are required to extract the best method to forecast Chronic Kidney Disease. Because of this, the Experimentation outcomes provide an analytical section in which all feasible ratings for every given algorithmic application or technique can be examined. Table 4.1 shows the accuracy of best three algorithms.

Table 4.1: Accuracy of Best 3 Algorithms

Algorithm	Accuracy (in percentage)
Decision Tree	98.60
AdaBoost	98.50
XGBoost	98.16

4.2.1 Data Gathering

The model was trained using the dataset collected from the online repository of University of California, Irvine (UCI). The model was trained using around 10321 data samples. This CKD dataset has a total of 24 variables per sample, 13 of which are nominal (categorical) and 11 of which are numeric, including a goal variable (class). Each class contains two nominal values: CKD and not CKD. There are numerous blanks in the dataset. A brief summary of the data set is provided in Table 4.2.

Table 4.2: Data Gathering & Null Values

Attribute	Scale	Missing Values (%)	Data Type
Age	age in years	2.27	Numeric
Potassium	in mEq / L	0	Numeric
Serum Creatinine	in mgs / dl	0	Numeric
Blood Pressure	in mm / Hg	0	Numeric
Sodium	in mEq / L	0.72	Numeric
Bacteria	(Present, Not Present)	0.97	Categorical
Pus Cell Clumps	(Present, Not Present)	0.97	Categorical
Appetite	(Good, Poor)	0.25	Categorical
Sugar	(0, 1, 2, 3, 4, 5)	0	Categorical
Albumin	(0, 1, 2, 3, 4, 5)	0	Categorical
Specific Gravity	(1.005, 1.010, 1.015, 1.020, 1.025)	0	Categorical
Hemoglobin	in gms	0	Numeric
White Blood Cell Count	in cells / cumm	0	Numeric
Blood Urea	in mgs / dl	0	Numeric
Packed Cell Volume	-	17.88	Numeric
Blood Glucose Random	in mgs / dl	11.06	Numeric
Red Blood Cell Count	in millions / cmm	0	Numeric
Coronary Artery Disease	(Yes, No)	0.5	Categorical
Diabetes Mellitus	(Yes, No)	0	Categorical
Hypertension	(Yes, No)	0	Categorical
Pus Cell	(Normal, Abnormal)	16.2	Categorical
Red Blood Cells	(Normal, Abnormal)	0	Categorical
Anemia	(Yes, No)	0.25	Categorical
Pedal Edema	(Yes, No)	0.25	Categorical
Class	(CKD, Not CKD)	0	Categorical

4.2.2 Dealing with Null Values

To deal with missing value, the affected rows or columns can simply be removed. Columns can be dropped in their whole if more than 1/2 of their rows are null. There's also the option to remove rows where one or even more columns has a null value. In our case, we have taken the mean of all values in a column.

4.2.3 Data Utilization

Each category (nominal variable) was coded independently, which made managing the data in a computer system much easier. A value of 1 was recorded for normal red blood cells and a value of 0 was recorded for irregular white blood cells. Pcc and ba were classed as either 1 or 0 depending on their presence. Therefore, 1 denotes a positive response and 0 indicates a negative response; these 1s and 0s were used to code the yes/no options. Value was determined by assigning a number of 1 to a good appetite and a value of 0 to a bad one. Although al, su, and sg were originally specified as nominal types, their values were decided by the relationship to the numbers. A factorization technique was used to transform all the categorical variables. The samples were labeled with random numbers between 1 to 10321. There were many empty spaces in this dataset. Patients may fail to take necessary pre-diagnosis steps for several different reasons. If the sample diagnostic categories are unknown, then an appropriate imputation approach must be used to fill in the missing values. After encoding categorical variables, the main CKD dataset's missing values were filled. After that, the description of the data for each of the 25 attributes was taken out for the purpose of developing a better interpretation of the dataset. Table 4.2 displays data from the dataset including the count, minimum, maximum, mean, standard deviation, and the quartiles.

Table 4.3: Description of the Dataset

Column Name	Count	Max	Min	Mean	25%	50%	75%	Std
Age	10321	90	2	51.51	42	54	64	16.95
Blood Pressure	10321	1400	0	79.62	70	76	80	70.39
Pus Cell	10321	1	0	0.76	0.76	1	1	0.39
Bacteria	10321	1	0	0.06	0	0	0	0.23
Sugar	10321	5	0	0.4	0	0	0	1.03
Red Blood Cells	10321	1	0	0.88	1	1	1	0.32
Blood Urea	10321	391	1.5	57.73	27	44	64	49.63
Serum Creatinine	10321	76	0.4	3.04	0.9	1.4	3.07	5.31
Sodium	10321	1436	104	144.03	135	137.53	141	87.07
Potassium	10321	7.6	1.4	4.43	3.9	4.63	4.8	0.73
Specific Gravity	10321	1.03	1.01	1.02	1.02	1.02	1.02	0.01
Hemoglobin	10321	17.8	3.1	12.46	10.8	12.53	14.6	2.83
Pus Cell Clumps	10321	1	0	0.12	0	0	0	0.32
Packed Cell Volume	10321	54	9	38.75	34	38.75	44	8.09
White Blood Cell Count	10321	26400	2200	8403.41	7000	8406	9400	2534.28
Albumin	10321	5	0	1.02	0	1	2	1.27
Red Blood Cell Count	10321	58	2.1	4.85	4.5	4.71	5.1	2.81
Hypertension	10321	1	0	0.37	0	0	1	0.48
Diabetes Mellitus	10321	1	0	0.35	0	0	1	0.48
Coronary Artery Disease	10321	1	0	0.09	0	0	0	0.28
Appetite	10321	1	0	0.79	1	1	1	0.4
Blood Glucose Random	10321	490	22	148.4	101	127	150	74.87
Pedal Edema	10321	1	0	0.19	0	0	0	0.39
Anemia	10321	1	0	0.15	0	0	0	0.36
Class	10321	1	0	0.62	0	1	1	0.48

4.2.4 Feature Importance

The title "feature importance" indicates to the technique of assigning a value to input features depending on their predictive ability for a target attribute. It is used to describe a set of techniques for ranking the importance of various features used as inputs to the predictive model. The feature importance score can be used to enhance a predictive model which can also be used for the better understanding of a dataset and the model. Table 4.4 shows the feature importance for different algorithms.

Hemoglobin clearly has the maximum value in Table 4.4. In this sense, the significance of Hemoglobin as a feature is much higher than that of other features.

Table 4.4: Feature Importance of each Attribute

Attribute	Algorithms				
	XGBoost Classifier	Decision Tree	Random Forest	Logistic Regression	AdaBoost Classifier
Age	.00524	.00529	.01393	.07557	.01
Hemoglobin	.36743	.7749	.4042	-3.32391	.36
Red Blood Cell Count	.07197	.07877	.26318	-4.10634	.27
Albumin	.00522	.00133	.00416	.03239	0
Sodium	.01034	.00334	.0105	-0.00714	.01
Blood Pressure	.00507	.00177	.00714	.04297	0
Appetite	.00483	.00033	.00182	-0.0092	0
Specific Gravity	.00804	0	.00528	-0.01694	0
White Blood Cell Count	.05247	.03789	.07531	.05221	.28
Hypertension	.38551	.07549	.12988	2.98574	.01
Sugar	.01002	.00071	.00292	.00864	0
Red Blood Cells	.00626	.00044	.00104	-0.02106	0
Diabetes Mellitus	.01044	0	.00402	.15624	.01
Potassium	.00671	.00482	.01041	.02034	0
Pus Cell	.00649	.00034	.00369	.0285	0
Pedal Edema	.00352	0	.00229	-0.06433	0
Blood Glucose (Random)	.00663	.00316	.01468	-0.04264	.02
Pus Cell Clumps	.00907	.00041	.00194	-0.00889	0
Packed Cell Volume	.00684	.00494	.01401	-0.08325	0
Anemia	.0043	.00034	.00221	.0338	0
Coronary Artery Disease	0	0	.00159	.07053	0
Bacteria	0	.00047	.00152	.03948	0
Serum Creatinine	.00713	.00079	.01145	-0.01995	.02
Blood Urea	.00649	.00445	.01284	-0.07015	.01

4.3 Predicted Results & Discussion

In this study, CKD was given a positive number and Not CKD a negative one. The confusion matrix is used to evaluate machine learning methods. For several kinds of algorithms, the confusion matrix template is shown in Table 4.6.

4.3.1 Confusion Matrix

A confusion matrix must be generated in order to validate the results from an implementation perspective. A confusion matrix is a matrix of $N \times N$ order with N target classes, can be used to evaluate the efficacy of a given classification model. The use of To assess the accuracy of machine learning models, the Confusion Matrix is used. The evaluation is done by contrasting actual target values with the anticipated ones. In this approach, both the successes and failures of the algorithmic model may be noticed. From binary classification, it's easier to calculate Precision, Recall, and Accuracy. In addition, micro average or macro average must be applied to these values for multi-class classification. Before addressing these, it's important to understand four basic blocks utilized in evaluating measurements. True positives (TP), False positives (FP), False Negative (FN) & True Negatives (TN) are used as evaluation measures for a Confusion Matrix. Confusion Matrices can be built with the use of these four values, as shown in Table 4.5. Table 4.6 displays the confusion matrix for each of the algorithmic approach.

Table 4.5: Specification of a Confusion Matrix

Confusion Matrix		
Predicted Class	Actual Class	
	TP (True Positive)	FP (False Positive)
	FN (False Negative)	TN (True Negative)

Table 4.6: Confusion Matrix for each Algorithm

Used Method	CM (in value)			CM (in percentage)	
		CKD	Not CKD	CKD	Not CKD
Logistic	CKD	1224	58	59%	3%
	Not CKD	53	730	3%	35%
KNN	CKD	1155	87	56%	4%
	Not CKD	122	701	6%	34%
Naive Bayes	CKD	1128	46	55%	2%
	Not CKD	149	742	7%	36%
SVM	CKD	1220	60	59%	3%
	Not CKD	57	728	3%	35%
Perceptron	CKD	1169	69	57%	3%
	Not CKD	108	719	5%	35%
SGD	CKD	1214	47	59%	2%
	Not CKD	63	741	3%	36%
Random Forest	CKD	1255	22	61%	1%
	Not CKD	22	766	1%	37%
Decision Tree	CKD	1261	14	61%	1%
	Not CKD	16	774	1%	37%
AdaBoost	CKD	1268	22	62%	1%
	Not CKD	9	766	0%	37%
XGBoost	CKD	1260	21	61%	1%
	Not CKD	17	767	1%	37%

- **True Positive (TP)**

"Positive tuples" are those that the classifier correctly labeled. The acronym represents it with the letter TP. Around 62% of the results from XGBoost, AdaBoost, and Decision Tree were TP (True Positive) values. Before that, 59% and 61% respectively went to the Logistic Regression and Random Forest methods respectively.

- **True Negative (TN)**

Negative tuples refer to the misclassified positive tuples. The letter TN might be used to indicate such situations. Decision Tree has the highest True Negative (TN) score at 37%, followed by Perceptron (35%), SGD (36%), and Random Forest (37%) in that order. And 35% using logistic regression.

- **False Positive (FP)**

It is these tuples with negative labels that the classifier has incorrectly identified as positive that we are focusing now. Such a relationship can be denoted with FP. In this analysis phase, FP values were the lowest (1%) when using the Decision Tree, AdaBoost & XGBoost Classifiers. Comparably, Naïve Bayes & SGD both produced 2% of FP values, whereas Logistic Regression produced 3%.

- **False Negative (FN)**

The classifier made an error and assigned a negative value to these positive tuples. The FN symbol represents this concept. AdaBoost Classifier achieved a 0% rate of False Negative values. Next on the list are Random Forest, XGBoost & Decision Tree with 1% False Negative values, followed by Logistic Regression having 3% FN values.

- **Precision**

It is possible to use precision as a statistic for evaluating how precise the given results are (i.e., what proportion of tuples that have been classified as positive are in fact positive). This indicates, the ratio of positive values to the total number of positive outcomes. The precision measuring formula is displayed by the given equation.

$$Precision = \frac{TP}{TP + FP}$$

- **Recall**

To calculate recall value, we divide all predicted positive samples by the total of True Positive and False Negative values. Recall measures how well a model can identify true positives. The proportion of identified positive samples grows as recall rises.

The math formula for measuring Recall is shown in the next equation.

$$\text{Recall (Sensitivity)} = \frac{TP}{TP + FN}$$

- **F1-Measure**

Harmonic mean between precision and recall determines the F1 score. Because the F1 score is derived by averaging the Precision and Recall scores, it follows that both Precision and Recall are given the same amount of weight in determining the F1 score. Precision and Recall must be high for a model to receive a high F1 score. Low Precision and Recall scores will result in a low F1 score. If one of a model's Precision or Recall scores is low while the other is high, the model will receive an F1 score that is considered as medium. Equation for the F1-Measure is given as below:

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- **Accuracy**

The accuracy of a classifier is measured by how many tuples from a specific test set are correctly categorized by the classifier. From the given equation, it may be easier to see how to measure accuracy.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

4.3.2 Classification Report

In machine learning, one way to measure how well a system is doing is through a "classification report." It is used to show the precision, recall, F1 Score, and validity of a trained classification model. Simply said, it measures how well a machine learning model performs at making classifications. The success of a machine learning approach can be gauged by looking at its displayed metrics, which include Recall, F1-measure, Precision,

and Accuracy. It offers a more comprehensive overview of the effectiveness of the trained model. Knowledge of all measures presented in this study is necessary for understanding classification reports generated by machine learning models. Classification results for all used methods are summarized in Table 4.7, where Precision, Recall, F1-Score, and Accuracy are presented as percentages.

Table 4.7: Classification Report using 4 Different Metrics

Algorithm	Class	Precision	Recall	F1-Score	Accuracy (%)
Logistic	Not CKD	0.93	0.93	0.93	94.62
	CKD	0.95	0.96	0.96	
KNN	Not CKD	0.85	0.89	0.87	89.88
	CKD	0.93	0.90	0.92	
GNB	Not CKD	0.83	0.94	0.88	90.56
	CKD	0.96	0.88	0.92	
SVM	Not CKD	0.93	0.92	0.93	94.33
	CKD	0.95	0.96	0.95	
Perceptron	Not CKD	0.87	0.91	0.89	91.43
	CKD	0.94	0.92	0.93	
SGD	Not CKD	0.93	0.93	0.93	94.48
	CKD	0.95	0.96	0.96	
Random Forest	Not CKD	0.98	0.98	0.98	97.82
	CKD	0.95	0.95	0.95	
Decision Tree	Not CKD	0.98	0.98	0.98	98.60
	CKD	0.99	0.99	0.99	
AdaBoost	Not CKD	0.99	0.97	0.98	98.50
	CKD	0.98	0.99	0.99	
XGBoost	Not CKD	0.98	0.97	0.98	98.16
	CKD	0.98	0.99	0.99	

4.4 Analysis of the Result

Now that we've calculated all the metrics, including Recall, F1-measure, Precision & Accuracy, etc., we can analyze the results. We will compare the performance of various algorithms and determine which ones are the most and least effective.

4.4.1 Cross Validation Score

Performance of a machine learning model is measured with the use of cross-validation, a statistical method. To begin Cross Validation, data is randomly shuffled and split into k separate groups. Next, we fit k models to $\frac{k-1}{k}$ of the data, and then we evaluate $\frac{1}{k}$ of the data. The final score is computed by taking the average of the results of each evaluation, and the model that is produced as a consequence is then implemented after being fitted to the entire dataset. The results of the cross validation for each algorithm are displayed in Figure 4.1.

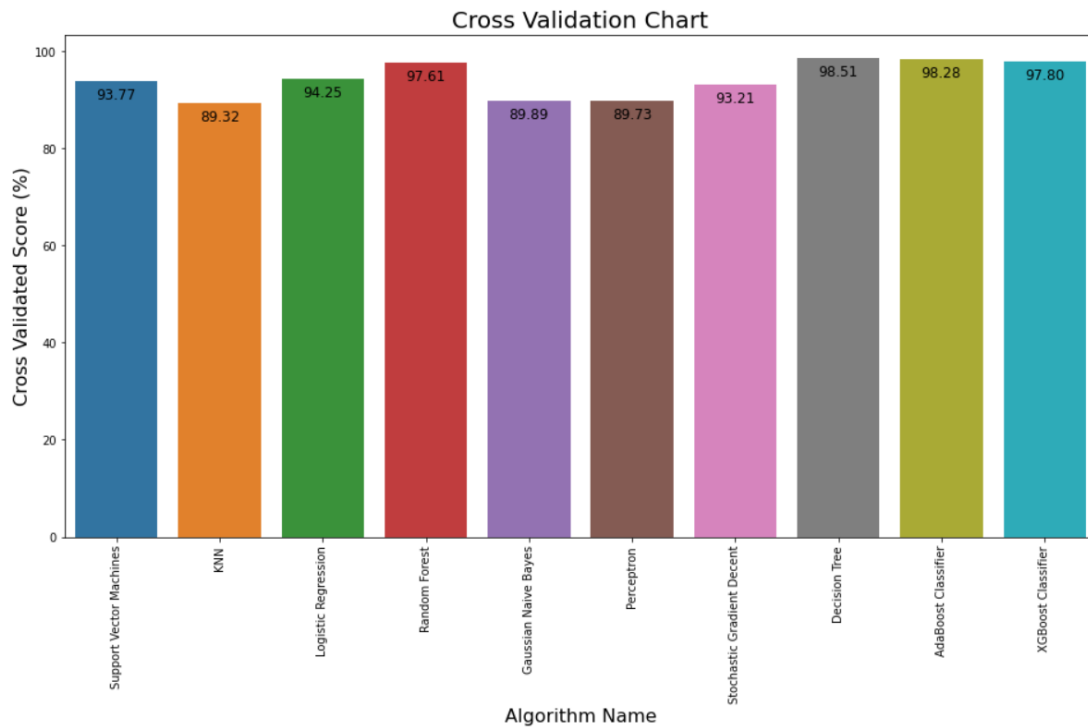


Figure 4.1: Cross Validation Score Chart

4.4.2 Accuracy

When evaluating an algorithm's effectiveness, accuracy is the yardstick of choice. How well it works, depends on the information it is given. The performance can be evaluated by its level of accuracy, which can be calculated using a probabilistic method. Among these techniques, Decision Tree Classifier has the highest accuracy, whereas K-Nearest Neighbor Classifier has the lowest. When it comes to making accurate predictions and classifications, Decision Tree is the most popular choice. In a Decision Tree, each internal node represents an attribute test, each branch provides an outcome, and the leaf nodes store a class label. The accuracy chart (Figure 4.2) and percentages (Table 4.8) for all the prediction algorithms used in this model are presented below.

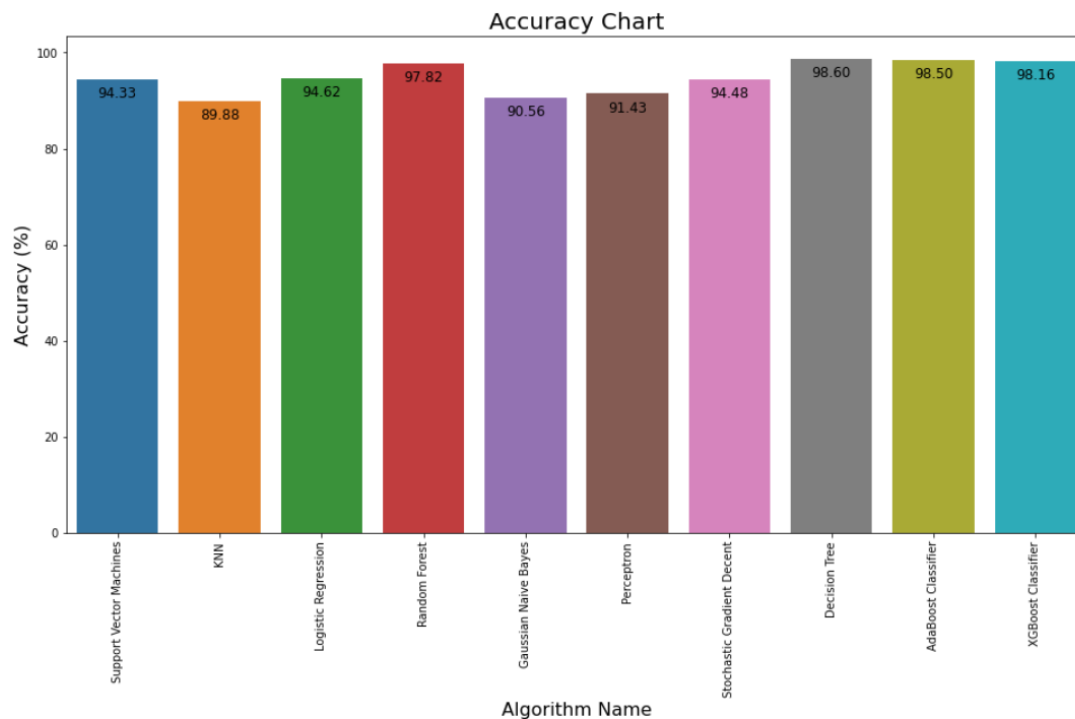


Figure 4.2: Accuracy Score Chart

4.4.3 AUC (Area Under the Curve) Score

This efficiency metric, when combined with machine learning programs, can be used to estimate how well a system will perform at different levels of classification. The AUC can be calculated by contrasting the model's performance on a percentile of randomly selected positive cases with that of a percentile of randomly selected negative ones. There are four possible values for this number, with one being the most likely. The numbers fall between 0 and 1, with 0 representing the absolute minimum. The outcomes of the AUC score for each algorithm are shown in Figure 4.3.

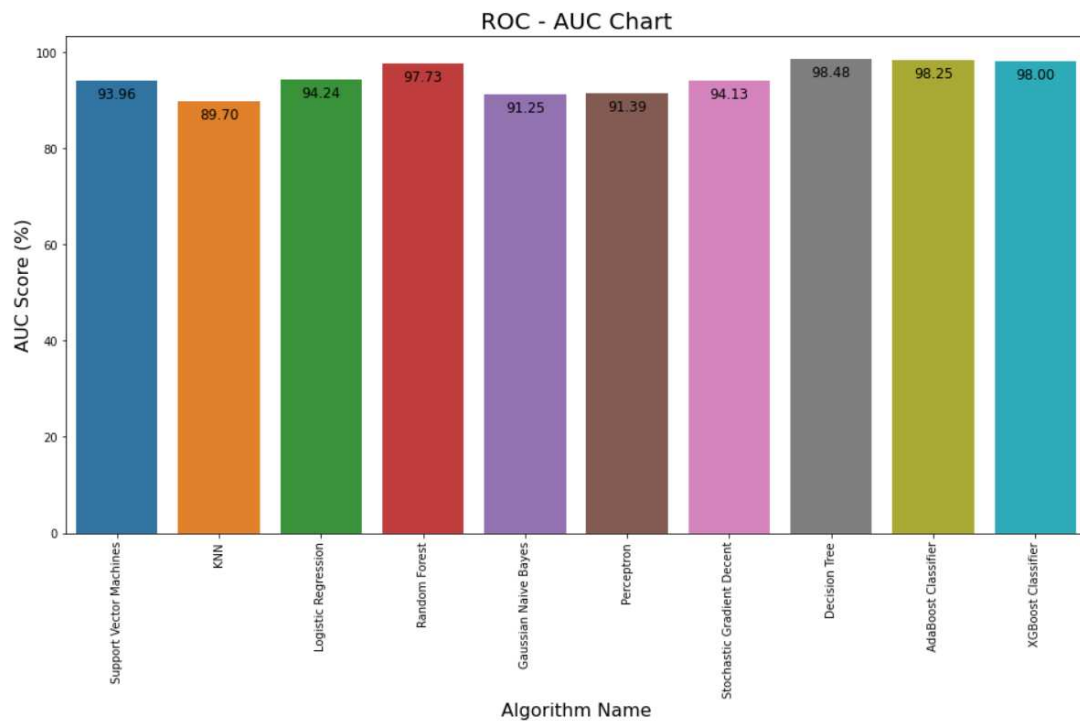


Figure 4.3: AUC (Area Under the Curve) Score Chart

4.4.4 Jaccard Similarity Index

One way to measure the extent to which two samples are alike or dissimilar can be done with a Jaccard score. The field of data science makes extensive use of Jaccard Similarity. By using Jaccard similarity index, which represents the ratio of intersection size to the union size, the similarity between two finite sets can easily be compared. The value of Jaccard Similarity index lies between 0 and 1. The formula for determining Jaccard Index can be given as,

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

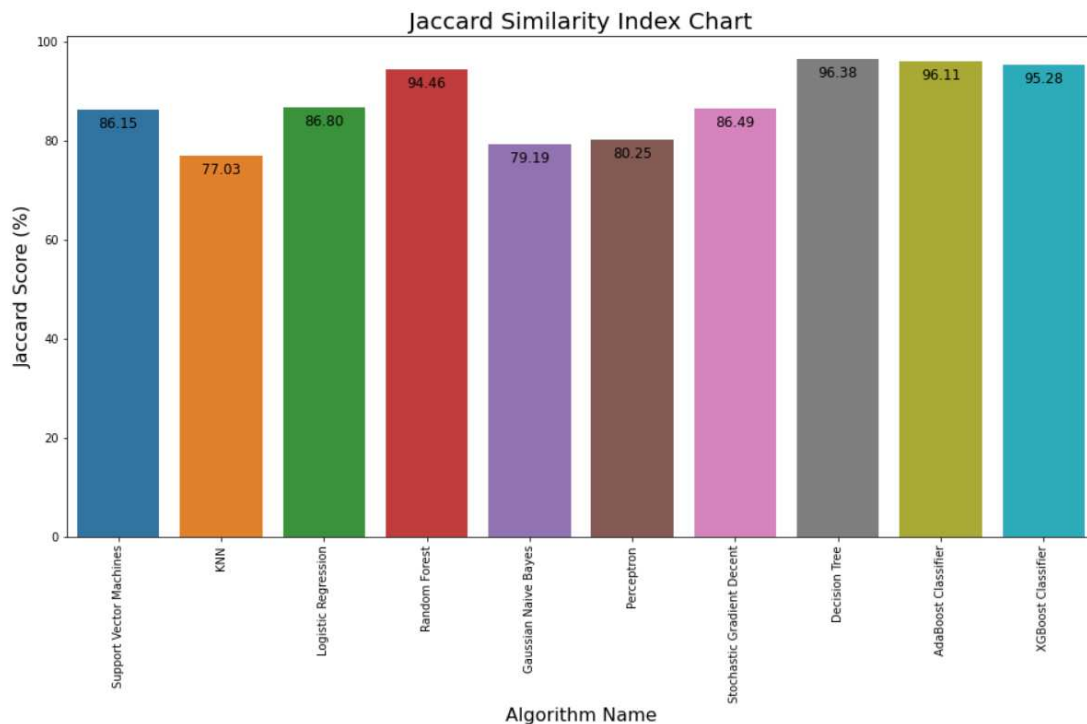


Figure 4.4: Jaccard Index Chart

4.4.5 ROC (Receiver Operating Characteristic) Curve

ROC analysis is used to evaluate the performance of a diagnostic test and the accuracy that a statistical model has. It classifies people as diseased or not diseased. ROC curve analysis is a basic graphical method for displaying medical diagnostic test accuracy. This ROC curve for each algorithm is displayed in Figure 4.5.

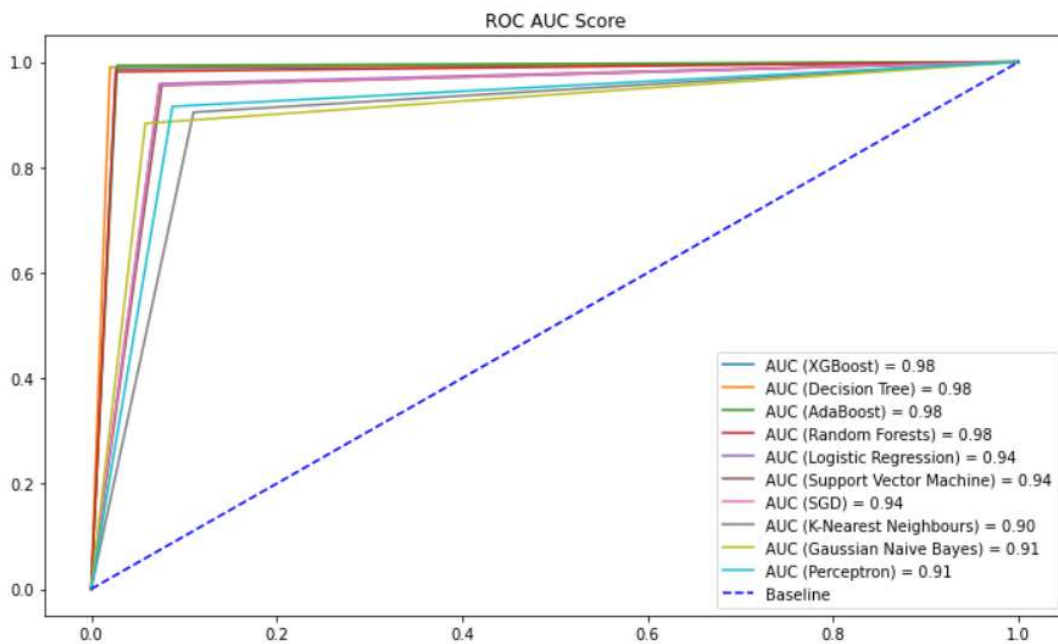


Figure 4.5: Receiver Operating Characteristic Curve

Overall, Decision Tree Classifier performed the best in terms of accuracy (98.60%), Jaccard Score (96.38%), Cross Validated Score (98.51%), and Area Under Curve (AUC) (98%). Table 4.7 provides a concise summary of the accuracy.

Table 4.8: Cross Validated, Accuracy, AUC & Jaccard Score

Algorithm	Accuracy Score (%)	Jaccard Score (%)	Cross Validated (%)	AUC Score (%)
SVM	94.33	86.15	93.77	93.96
KNN	89.88	77.03	89.32	89.70
Logistic Regression	94.62	86.80	94.25	94.24
Random Forest	97.82	94.46	97.61	97.73
Gaussian Naïve Bayes	90.56	79.19	89.89	91.25
Perceptron	91.43	80.25	89.73	91.39
SVG	94.48	86.49	93.21	94.13
Decision Tree	98.60	96.38	98.51	98.48
AdaBoost Classifier	98.50	96.11	98.28	98.25
XGBoost Classifier	98.16	95.28	97.80	98.00

4.4.6 Error & Misclassification

Errors become problematic when attempting to determine the accuracy of any algorithm. Accuracy measures for a machine-learning model include mean square error and mean absolute error after misclassification. When an inappropriate attribute is used, misclassification occurs. If the rate of inaccuracy is the same across all possible divisions of a variable, then the variable is treated as misclassified. Absolute error is a term used to describe the degree of measurement error. MAE is the mean of a measurement's absolute errors. The formula for MAE is denoted by the given equation.

$$MAE = \frac{1}{n} \sum_{i=1}^n |x - x_i|$$

Mean Squared Error (MSE) indicates the distance between a regression line and a set of points. The mathematical formula for the MSE can be shown as,

$$MSE = \frac{1}{n} \sum_{i=1}^n |y - y_i|^2$$

The algorithms' misclassification, mean squared error, and mean absolute error are displayed in Table 4.9. When compared to other popular methods, XGBoost's error rate was lower on all levels (1.45%).

Table 4.9: Errors & Misclassifications

Algorithm	MSE (%)	MAE (%)	Misclassification (%)
SVM	5.67	5.67	5.67
KNN	10.12	10.12	10.12
Logistic Regression	5.38	5.38	5.38
Random Forest	2.18	2.18	2.18
GNB	9.44	9.44	9.44
Perceptron	8.57	8.57	8.57
SGD	5.52	5.52	5.52
Decision Tree	1.40	1.40	1.40
AdaBoost Classifier	1.50	1.50	1.50
XGBoost Classifier	1.84	1.84	1.84

4.4.7 Standard Deviation

The standard deviation can also be calculated using the data in this study. The variability of a dataset is measured by its standard deviation from the mean. The square root of the variance of each data point is used to calculate the standard deviation. When the data points are further from the mean, the standard deviation rises.

Table 4.10: Standard Deviation

Algorithm	S.D.
Logistic Regression	1.24
Random Forest	0.09
Decision Tree	0.15
AdaBoost Classifier	0.1
XGBoost Classifier	0.11

4.5 Web Implementation

The development of machine learning models has been successfully implemented, and the right technique has been chosen and implemented. To conclude this chapter, the System Architecture outlined in the previous chapter will be demonstrated by displaying the implementation of the stated model via the Web Interface.

4.6 User Interface of the Website

According to Chapter 3, the "Flask" method will be used to make online predictions for chronic kidney disease. As mentioned in Chapter 3, we'll be developing a web interface using the Flask framework. A website that is both extremely effective and fully functional has been developed exclusively for the aim of ensuring that this task is carried out successfully [16]. Figure 4.6 illustrates this web-based user interface.

Chronic Kidney Disease Prediction

Age	Blood Pressure	Specific Gravity	Albumin
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	Red Blood Cell Count	Hypertension	Diabetes Mellitus
Coronary Artery Disease	Appetite	pedal Edema	Anemia

Predict

Diagnosis Result :

Figure 4.6: User Interface of the Website

4.7 Analysis of the Website Output

Based on the methodology used and the data obtained, we can only make two classifications: CKD and Not CKD. The entirety of this section will necessitate two separate output analyses. Figure 4.7 shows the outcomes of randomly inputting all data necessary to identify CKD using a trained model. Similarly, in Figure 4.8, random data from a test case was used to anticipate CKD using a previously trained model. After all of these steps are taken, it is determined that the trained model performed accurately.

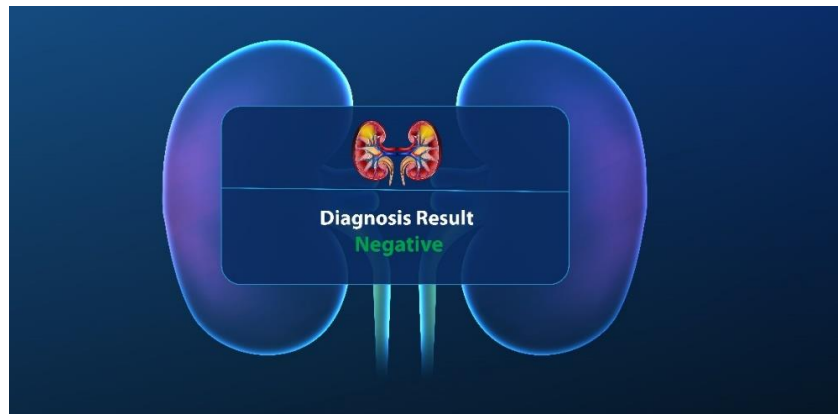


Figure 4.7: Negative Web Output

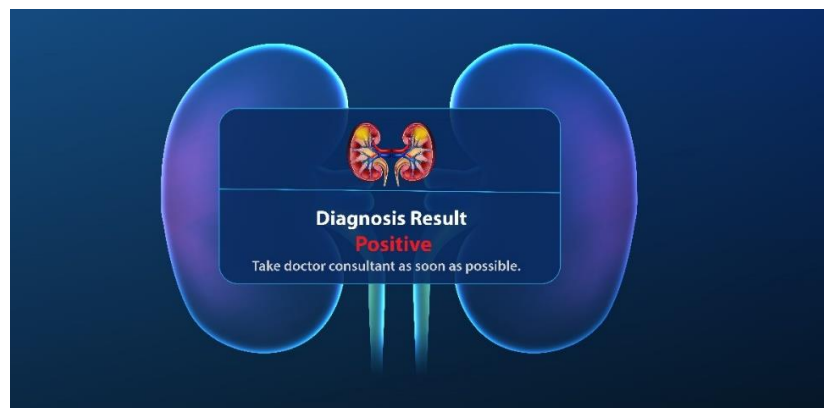


Figure 4.8: Positive Web Output

CHAPTER 5

IMPACT ON SOCIETY & SUSTAINABILITY

5.1 Introduction

When a project work is implemented, its effects on the community should be discussed and studied. In this chapter, the consequences of the project have been broken down into three different sections. Within the section titled "Impact on Society," it has been addressed how this work will have a positive impact on the society. After that, I have analyzed the ethical implications. For this, the ethical aspects have been discussed in a way that helped people understand how this work will benefit both the doctors and the patients. And at the end, the sustainability of the project has been discussed. The last section discusses how this work can expand in the future to benefit more individuals.

5.2 Impact on Society

The findings of this research will have a considerable effect on society. Modern life is so hectic that taking the time to visit the doctor to make sure a patient isn't getting a chronic illness might be a hassle. Users can now submit the information necessary to make an accurate prediction of chronic renal disease through a user-friendly interface. Sometimes, it is unsafe to visit a hospital for medical diagnostic procedures especially when the global spread of the COVID-19 occurs. If anyone is able to see their report at home, it would be really useful. In future, I have a plan to make a survey on patients and medical professionals to assess the efficacy of this study. The results of that survey will undoubtedly have a favorable effect on society at large and on the doctor-patient relationship.

5.3 Ethical Aspects

In this era of technology, people could get their diagnostic needs met without visiting a hospital. People could learn the fundamentals of chronic renal disease without leaving their home. As a result, they would be capable of making predictions on their own. Since a user interface has been made, the model will improve eventually because all of the data that has been collected and saved in the database will be added to the training set. Since the project is based on machine learning, it seems that we could not rely completely on this model right now. A machine cannot precisely anticipate anything. It will take time if we want to get a better accuracy from a ML model. When the dataset has a million records, the model would be significantly more robust, and more reliable predictions can be made. It may not be necessary to visit the hospital for a diagnosis in the future if AI and the IoT could be integrated with our project. Individuals will have the ability to diagnose chronic renal disease at home before consulting a physician about their condition. If a portable blood and urine testing device could be developed, then people can more accurately test for chronic kidney disease at their own houses.

5.4 Sustainability

An individual's risk of developing chronic kidney disease may be assessed using a website, making this study highly sustainable. Currently, the website predicts kidney disease only through the use of machine learning algorithms. Deep learning, AI, and the IoT can be put together to improve this project in the future. Therefore, various tasks can be accomplished with this project in the future if the sustainability is ensured. A smartphone app can be developed for the prediction as well. Till now, the study is only confined to the prediction of CKD, but this ML project and the web-based method can be put together to predict numerous diseases of the kidney as well as other organs.

CHAPTER 6

CONCLUSION & IMPLICATION FOR FUTURE WORK

6.1 Introduction

The chapter includes the future scope and conclusion part of my work such as how this project may help the company grow in a productive way in the future, and what methods could be applied to develop a superior machine in the future. This chapter ends with a clean and clear summary, which is given therefore at end of the chapter. There is a list of references at the conclusion of this chapter for further reading.

6.2 Implications for Future Research

No matter how challenging it is, the idea can be used as the foundation for a web application for any hospital specializing in the treatment of kidney disease. The authors of this project have created a simple interface. At some point in the near future, an IoT-based website could be built and made accessible to the general public online. Chronic Kidney Disease (CKD) can be easily predicted at home by filling out a web form with all of the required information. The person can decide if he or she has CKD or not. Since this would be an IoT-based website, and users will be entering new information each time, the site will keep track of that information. Eventually, the model will be able to improve its accuracy by collecting knowledge from newly acquired data. Incorporating Artificial Intelligence and Neural Network algorithms into a Deep Learning technique may make it more effective over time.

6.3 Recommendations

For every diagnostic test, there is a base level, and if an abnormal phase is identified, this system would be able to become aware of that specific situation. Here in Chronic Kidney Disease, the main factors that help predict the disease are Urine Albumin, Hemoglobin, Serum Creatinine, Creatinine Clearance, etc. The typical range for Serum Creatinine is 0.7-1.3 mg/dL for male patients and 0.6-1.1 mg/dL for female. Creatinine Clearance normal ranges for men and women are 97-137 mL/min and 88-128 mL/min, respectively. When the quantity of protein found in the urine goes up, it could be a sign of a problem with the kidneys. If someone wants to know how serious their condition is, finding out if their urine has protein, is a must. The normal range is between 0 and 8 mg/dL. An abnormal increase in the protein levels presented in the urine might indicate to a problem with the kidneys. A healthy Hgb level for women is 12-16 g/dL, whereas a good Hgb level for men is 14-18 g/dL. It may be an early sign of chronic renal disease if the ranges fluctuate, as this may occur in some circumstances. As a result, there is still time to take preventative measures such as changing the way of eating, working out, drinking more water, etc. Getting a doctor's advice is required at such time.

6.4 Conclusion

It is possible that a patient with CKD could be cured if their condition were diagnosed and treated promptly. The presence of chronic kidney disease can be determined using a series of laboratory testing (CKD). These tests could take a long time and a large amount of money in actual scenario. Any stage of chronic renal disease can be predicted by an ML model that has been trained on an appropriate dataset. The author has created a user interface for this project in which users can receive their CKD diagnosis result by filling out a predetermined form.

Ten different ML algorithms were trained on the UCI dataset and then evaluation was done according to their accuracy, AUC score, Cross Validated score and Jaccard index. This is done in order to discover the model that is the best fit for the dataset. Accuracy, Jaccard index, and cross validation are the three score values under which the efficiency of all the algorithms are tested. When compared to other classifiers, Decision Tree achieves superior performance. Almost always, it was the best decision to choose Decision Tree classifier. The web-application for this project could be made available to any hospital specialized in the treatment of kidney disease. Online CKD reports will be available 24/7, so patients won't have to take time out of their busy day to get them.

REFERENCES

- [1] World Health Rankings, deaths due to renal disease in Bangladesh, accessible at <<<https://www.worldlifeexpectancy.com/bangladesh-kidney-disease>>>, last visited on 09-01-2022 at 11:08 AM.
- [2] Center for Disease Prevention & Control, accessible at <<<https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/people-with-medical-conditions.html>>>, last visited on 09-01-2022 at 11:14 AM.
- [3] Lambert, J. R., Perumal, E., & Arulanthu, P., “Identification of Nominal Attributes for Intelligent Classification of Chronic Kidney Disease using Optimization Algorithm” in the 2020 ICCSP (International Conference on Communication and Signal Processing) IEEE, pp. 0119-0125, July 2020.
- [4] Nusinovici, S., Yan, M. Y. C., Tham, Y. C., Cheng, C. Y., Ting, D. S. W., Li, J., & Sabanayagam, C., “Logistic regression was as good as machine learning for predicting major chronic diseases” in the Journal of clinical epidemiology, 122, pp. 56-69, 2020.
- [5] Balakrishnan, S., “Feature Selection Using Improved Teaching Learning Based Algorithm on Chronic Kidney Disease Dataset” in Procedia CS, 171, pp. 1660-1669, 2020.
- [6] Ali, S. I., Hussain, M., Lee, S., Bilal, H. S. M., Hussain, J., & Satti, F. A., “Ensemble feature ranking for cost-based non-overlapping groups: A case study of chronic kidney disease diagnosis in developing countries” in IEEE-Access, 8, pp. 215623-215648, 2020.
- [7] Segal, Z., Ehrenberg, B., Koren, G., Kalifa D., Maor, G., Radinsky, K., & Elad, G., “Machine learning algorithm for early detection of end-stage renal disease” in BMC-Nephrology, 21(1), pp. 1-10, 2020.
- [8] Navaneeth, B., and Suchetha, M., “A dynamic pooling based convolutional neural network approach to detect chronic kidney disease”, Biomedical Signal Control & Processing Operation, 62, p. 102068, 2020.
- [9] Qin, J., Feng, C., Liu, C., Chen, B., Liu, Y., & Chen, L., “A machine learning methodology for diagnosing chronic kidney disease”, IEEE-Access, 8, pp. 20991-21002, 2019.
- [10] Abdelaziz A., Riad, A. M., Mahmoud, A. N., & Salama, A. S., “A machine learning model for predicting of chronic kidney disease-based internet of things and cloud computing in smart cities” in the Springer Cham, pp. 93-114, 2019.
- [11] Snegha, J., Bhavani, S., Charanya, R., Tharani, V., & Preetha, S. D. “Chronic Kidney Disease Prediction Using Data Mining” in the 2020 ic-ETITE (International Conference on Emerging Trends in Information Technology and Engineering) IEEE, pp. 1-5, February 2020.
- [12] Herath, D., & Ekanayake, I. U., “Chronic Kidney Disease Prediction Using Machine Learning Methods” in the 2020 MERCon (Moratuwa Engineering Research Conference) IEEE, pp. 260-265, July 2020.
- [13] Shahbaz, M., Yashfi, S. Y., Sakib, N., Islam, M. A., Pantho, S. S., & Islam, T., “Risk Prediction of Chronic Kidney Disease Using Machine Learning Algorithms” in the 2020 11th ICCCNT (International Conference on Computing, Communication and Networking Technologies) IEEE, pp. 1-5, July 2020.

- [14] Thangavelu, M., & Harimoorthy, K., “Multi-disease prediction model using improved SVM-radial bias technique in healthcare monitoring system” in the Journal of Humanized Computing & Ambient Intelligence, 12(3), pp. 3715-3723, 2021.
- [15] Ding, C., Ren, L., Chen, G., Li, X., Xue, W., & Li, Y., “Prediction of Chronic Kidney Disease Using Adaptive Hybridized Deep Convolutional Neural Network on the Internet of Medical Things Platform”, IEEE-Access, 8, pp. 100497-100508, 2020.
- [16] Live Website, Chronic Kidney Disease Prediction, accessible at <<<https://mubtasim.herokuapp.com>>>, last visited on 09-11-2022 at 2:19 PM.

Prediction of Chronic Kidney Disease Using Different Machine Learning Methods

ORIGINALITY REPORT

11 %
SIMILARITY INDEX

7 %
INTERNET SOURCES

5 %
PUBLICATIONS

5 %
STUDENT PAPERS

PRIMARY SOURCES

1 dspace.daffodilvarsity.edu.bd:8080 **3** %
Internet Source

2 Submitted to Daffodil International University **3** %
Student Paper

3 doctorpenguin.com **<1** %
Internet Source

4 Ebrahime Mohammed Senan, Mosleh Hmoud Al-Adhaileh, Fawaz Waselallah Alsaade, Theyazn H. H. Aldhyani et al. "Diagnosis of Chronic Kidney Disease Using Effective Classification Algorithms and Recursive Feature Elimination Techniques", Journal of Healthcare Engineering, 2021 **<1** %
Publication

5 "Proceedings of Data Analytics and Management", Springer Science and Business Media LLC, 2022 **<1** %
Publication
