Chronic Kidney Disease Prediction Using Machine Learning Approach

BY

Prima Rani Chanda
ID: 172-15-9713
AND
Sharon Kumar Das
ID: 172-15-9991

This Report Presented in Partial Accomplishment of the Specifications for the Degree of Bachelor of Science in Computer Science and Engineering.

Supervised By

Shah Md. Tanvir Siddiquee
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised By

Mr. Narayan Ranjan Chakraborty
Assistant Professor
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

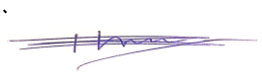
DHAKA, BANGLADESH

3rd JUNE 2021

APPROVAL

This Project titled **"Chronic Kidney Disease Prediction Using Machine Learning Approach",** submitted **by Prima Rani Chanda** and **Sharon Kumar Das** to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial accomplishment of the requirements for the degree of B.Sc. in Computer Science and Engineering and certified as to its method and contents. The presentation has been held on Thursday 3rd June 2021.

BOARD OF EXAMINERS



Chairman

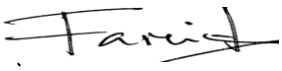
Dr. Touhid Bhuiyan
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Gazi Zahirul Islam
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Raja Tariqul Hasan Tusher
Senior Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



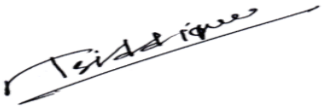
External Examiner

Dr. Dewan Md. Farid
Associate Professor
Department of Computer Science and Engineering
United International University

DECLARATION

We hereby certify that, this project has been executed by us under the supervision of **Shah Md. Tanvir Siddiquee**, Assistant Professor, Department of CSE Daffodil International University. We also certify that this project or any part of this project has been submitted elsewhere for award of any degree.

Supervised by:



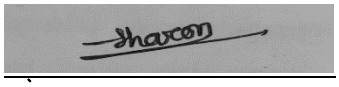
Shah Md. Tanvir Siddiquee
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised by:

Mr. Narayan Ranjan Chakraborty
Assistant Professor
Department of CSE
Daffodil International University

Submitted by:

Prima Rani Chanda.
ID: 172-15-9713
Department of CSE
Daffodil International University



Sharon Kumar Das.
ID: 172-15-9991
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

We are highly appreciative and thankful to **Shah Md. Tanvir Siddiquee,** Assistant Professor, Department of CSE, Daffodil International University, Dhaka. Wisdom & highest credit to our supervisor in the realm of "*Data Mining*" to bring out this project. His expertise supervision, boundless endurance, steady reassurance, faithful and active guidance, helpful advice, expert judgment, understanding several substandard drafts, and fixing them at all frames have made it possible to achieve this project.

We would like to express our heartiest gratitude to our supervisor **Shah Md. Tanvir Siddiquee** and co-supervisor **Mr. Narayan Ranjan Chakraborty** and Department head **Dr. Touhid Bhuiyan** Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

We would also like to express our heartiest gratitude to other faculty members and the staffs of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

Machine learning is currently playing a very important role in various sectors. It also has a significant position in the healthcare sector for its working efficiency. Machine learning classification algorithms are popular in medical science for predicting various complex diseases. Kidney disease is a familiar name to all of us nowadays. Moreover, this disease has become a major public health problem for people worldwide. People are affected with kidney disease when they do not follow a proper diet in their daily life, lack of proper health awareness including drinking a very little amount of water. As a result, the disease later turned into a terrible disease, which is called CKD. Countless people all over the world are suffering from this disease and they are dying due to a lack of proper awareness and treatment. The treatment of this disease is extremely expensive and difficult. When a patient is affected with CKD, their kidneys stop working completely. There is no limit to the suffering of patients when the kidneys stop their activity. In this research, we have tried to predict CKD by applying various classification algorithms of machine learning. Accurate data is very important to do this work so that we have done this work with the proper dataset. Our goal is to identify and predict chronic kidney disease. To predict CKD, we used four popular machine learning classification algorithms. Which are respectively SVM, Decision Tree, Random Forest, KNN. We have basically completed our work in two steps. In the first step we have completed the training of the data and in the next step completed the testing. After trained by the machine learning algorithm, we got our desired result. In that case, the Decision tree algorithm has shown the best performance among the four classification algorithms.

TABLE OF CONTENTS

CONTENTS PAGE

List of Figure

List of Table

CHAPTER 1

Introduction

1. 1 Introduction

At present, kidney disease is a problem that can take a terrible and deadly form in a very short time. Countless people around the world are constantly dying for this kidney disease and its severity. Chronic kidney disease, commonly known as a terrible disease, which mainly affects the kidneys. CKD is a group of diseases that affect the kidneys and reduce their ability to keep us healthy. Chronic kidney disease (CKD) affects 10% of the global population, and millions of people die per year due to a lack of affordable and costly treatment [24]. Chronic kidney disease was ranked 27th on the list of causes of the total number of deaths globally in 1990, but climbed to 18th in 2010, according to the 2010 Global Burden of Disease report [24]. CKD is a global health major crisis. According to the WHO, there were nearly 58 million people deaths globally in 2005, with 35 million of those deaths due to infectious illness [24]. Most two common causes of CKD are mainly extreme blood pressure and diabetes, basically up to two-thirds of patients. When our blood sugar levels are much high, diabetes damages multiple main organs of our body, including some main organs like the heart, nerves, blood vessels, and also our eyes [7]. Although modern medical advances have made significant contributions to reducing human mortality as a result of many lethal chronic diseases, eradicating illness grasp is still a long way off.

In the current context, it is very important to solve this problem because lots of people are constantly dying due to kidney disease and this terrible disease is being able to take the lives of people in a very short time. Mainly salt and minerals in our blood, such as calcium, phosphate, potassium, and also sodium are balanced by our kidneys, which both extract wastes from the blood and eliminate them by urination [10]. When a patient suffers from kidney disease, all these activities are completely stopped. Not all of their body fluids can be properly excreted. That is why they mainly need a waste disposal medium like dialysis. which is basically an artificial process of removing different types of liquid waste from our body. This is an extremely cumbersome and expensive approach. A suffering patient has

to remove unwanted waste from their body through this artificial treatment several times a month. Which are very expensive and difficult processes for patients.

Advanced technology is being used in different sector of medical sciences, including the diagnosis of kidney disease. However, if its initial symptoms are not detected, it can later become a fatal situation. Later, a patient has to face death which is very undesirable occurrence for us. So, in that case a patient has to face different types of difficulties. If the problem is not treated properly after the first stage of the diagnosis, it can become serious at the last stage. The treatment of this disease is also very expensive in the medical field. Therefore, a proper approach is most needed which will easily diagnosis the disease from progressing to chronic diseases without any hassle and will bring benefits through proper prediction.

In this case, we think that various techniques of data mining will play vital role mainly in this regard. Machine learning approaches of the data mining are playing a significant role in predicting different diseases through various classification algorithm. Using various classification algorithms, it is possible to predict many complex diseases by accurate analysis and detailed data. By which we think that it is possible to bring different facilities. Above all, machine learning techniques will play a pivotal role in this sector.

Our goal is to bring these detailed datasets to the right output through proper training and testing. In our work, we have used the approach of machine learning. Basically, we have used the classification algorithms of machine learning. In this case, we have completed training and testing through four classification algorithms. We verify the correct classification algorithm for the dataset. Also, preprocessed the dataset using proper methods and we have completed the training of our machine through the classification algorithm. We have tried to bring maximum accuracy from our dataset through these classification algorithms.

1.2 Motivation

The number of kidney patients in Bangladesh and other countries is increasing day by day. Huge number of patients in various clinical institutes and medical hospitals is innumerable, which shows the increase in the severity of this disease. Moreover, a search of various medical sites and institutions reveals a list of countless patients with this disease. From which the severity of this disease is going to be estimated. In Bangladesh, it is identified as a big public health issue. Several types of kidney disease can be seen all over the world. Chronic kidney disease affects overall 37 million of people in the US [24].

The phrase "CKD" refers to long-term kidney damage that can worsen over time. In that situation, their kidneys can stop functioning if the damage is serious. Kidney failure, recognize as ESRD, it's a terrible condition in which both kidneys stop working. We may require dialysis or a kidney transplant to survive if the kidneys fail [24]. CKD was shown to be prevalent in the elderly community (aged > 65 years) in three databases, the Kidney Advanced Diagnosis Program, the National Health and Nutrition Examination Survey with the greatest share in those aged 80 years and older. Countless patients with the managed ESRD is also estimated to have almost doubled [23].

We are inspired to do this by seeing the horrible situation around us. We have seen, many people around us suffer from kidney disease. Our research aim is to estimate the likelihood of developing CKD. This research also reveals the age range of the patients, with the bulk of them being between the ages of 46 and 64. Analyzing all of the data using machine learning techniques would have the highest level of precision, allow us to create an application based on them. The findings of this research have revealed that the patients have CKD or don't have CKD.

Also, we are motivated to do this work by thinking about different types of kidney patients and their last terrible consequences. The disease can be identified in its early stages and proper action must be taken. The main purpose of this study is, to predict ckd using machine learning approaches regarding chronic kidney disease by various classification algorithms. If these patients go to the chronic stage, they will face death. So, we have done this research to predict ckd and for human welfare. Through this work, we want to predict whether the

patients have chronic kidney disease or not from our huge dataset. Finally, we can say that, we are motivated to do this research work, thinking about the suffering of helpless people.

1.3 Rationale of the Study

There is no doubt that, CKD is a lethal disease because only a kidney patient can understand the severity of its end consequences. This disease is not given much importance in the early stages, due to which the disease later takes the form of chronic disease. As a result, a patient has to die. In order to predict this disease, different types of researchers are presenting their results by analyzing different approaches through their research. But after so much research on it, we still don't see any far-reaching results. We see different patients around us suffering from different diseases in the hospitals and not getting proper treatment. Since the treatment of this kidney disease is very expensive and they do not know much about the severity of this disease. As a result, they are facing chronic-kidney disease in the future. If such a prediction system can be invented that can predict their future onset of chronic diseases by observing the various symptoms of their early stages. It would be extremely beneficial for such patients will be able to reduce the amount of suffering of the patients. However, there are many classification techniques but collecting kidney disease data is very difficult. There is no other way to deal with confidential data than to create more automated applications or to make Machine Learning approaches in Kidney Disease even more effective. All of this has given us an interest in doing this work. Researchers have long been interested in the feasibility of an integrated method for disease classification. While it has not received much publicity in terms of public adoption, it has the potential to be adopted.

chronic kidney disease is very dangerous because the signs of this disease frequently appear late in the disease's progression. In that case, If, such a system discovers the patients can understand which type of test to perform, and once the test is completed, the patient can assess their own CKD risk level. An automated system can verify the risk percentage of patients in public hospitals on a regular basis, which can be checked by experienced physicians.

Patients are also tested on a variety of other medical diseases. The machine will review this and mainly notify the doctor if the patient is at risk of developing CKD, hopefully reducing the risk of ESRD by catching it early through proper diagnosis.

The vast majority of people who pursue medical treatment in Bangladesh's public medical facilities must suffer excruciating pain throughout this process. Basically, an automated predicting system will undoubtedly save such individuals a significant amount of money and time.

If such a prediction system is created, the doctors will be able to give proper treatment to the patients and the patients who are at high risk will get the right treatment in different government hospitals with extreme care. As a result, patients will survive CKD in the initial stages, they will return to recovery with proper care after diagnosis and they will be free from chronic disease.

The rationale for our work is that patients with different types of kidney disease can be relieved of the severity of the disease and the suffering of the public will be alleviated somewhat if our proposal can be transformed into an automated system in the future, then patients in the different types of government hospitals will be able to get proper treatment at an early stage. Besides, physicians can predict whether the disease will lead to chronic disease, and as a result, they will be able to give proper treatment to a kidney patient and the suffering of the people will be greatly reduced.

1.4 Research Questions

- What is CKD?
- Does it predict Chronic Kidney Disease by ML algorithm?
- Can we show the perfect accuracy for CKD prediction?
- Can we apply various types of classifiers in the dataset?
- Have we been able to show good performance between different types of algorithms?
- Have we compared the accuracy of the applied algorithms?

1.5 Expected Output

❖The expected result of our work is, that we have been able to distinguish between different machine learning algorithms. Also, we have properly implemented various types of machine learning techniques.

❖We have been able to choose the right algorithms to train the detailed dataset properly.

❖ In addition, we have been able to select the correct classification method for prediction by comparing the trained algorithms and find out the best accuracy through it. To determine which classification algorithm give the better result for predictions.

❖We have found out whether CKD or Not CKD by using the proper method through training and testing.

❖To build a high-accuracy predictive model.

❖Finally, we have been able to determine the best algorithms by training the various types of ML classifiers using the accurate way. Moreover, we are finding out the best accuracy, precision and recall value from the trained dataset.

❖Also, we have found out which algorithm gives the best result in this case and shown which classification algorithm predicts well.

1.6 Report Layout

In this, Chapter 1 we have given a complete description of our work. We have also discussed the motivation of our work in detail. An important part of this chapter is Rational Study. We have described it properly. Also, we have revealed the outcome of our work in this part. Moreover, we have raised some research-related important questions in this segment.

Chapter 2 provides a detailed discussion about the background study. We have divided this chapter into some sub-sections. At the beginning of this portion, we have given a descriptive introduction. In the next part, we have discussed the work of different researchers. which is especially related to our work. We have created a comparative summary through research and analysis. In addition to these, we have highlighted the areas

in which our work can be used successfully and what kind of challenges we have to face to do this work are also discussed in detail in the last part of this chapter.

In chapter 3, we have tried to present the methodology of our research through theoretical explanations. Like other chapters, we have divided this chapter into different sub-sections. Firstly, we gave a brief introduction. Then we discuss the subject of our research work and what types of equipment we have used to complete our work properly. Moreover, we have elaborated the methodology of our work by statistical analysis. In the proposed methodology section, we have given a theoretical explanation of the machine learning classifier used in our work.

In chapter 4, we have analyzed the results obtained from the trained algorithms. In this case, we have explained the outcomes by some tables and figures. Finally, we have found out which algorithm can perform better for CKD prediction.

As revealed in chapter 5, We mentioned that how our work will impact society. We have also written about the ethical aspects of this work. In the last section, we have briefly discussed the sustainability plan of our work.

We have made a summary of our research study, in the last chapter 6 of our report. Where the importance of the work has been highlighted. An important segment of this chapter is the conclusion part. This part gives a great overview of the whole work. In the last part, we have discussed the further plan of our work.

CHAPTER 2

Background Study

2.1 Introduction

Data mining has a huge field, one of which is machine learning. The use of this machine learning is being used in different ways in different research areas. Basically, there are some classification algorithms of machine learning through which researchers are being used for various types of statistical analysis and research including the prediction of various diseases. Huge amounts of data are utilized and analyzed to find the right output. This field seems like a great discovery in this present era. Through data mining, various fields including knowledge discovery in databases seem to be an easy way for more detailed research. It will play an important role in pioneering research. In this chapter, we will mainly analyze various related topics and highlight the doctrines of different researchers. We will briefly highlight other researcher various works. Besides, we will do a comparative analysis of the work used by different researchers. In the next sections, we will discuss what kind of problems we had to face to get the work done and how we have done this work by overcoming the various challenges. Moreover, In the following sections reflect the similar work that has already been performed by several experts in this area in the past. In addition to providing a good overview, from this chapter mainly we will explain what's the shortcomings of this works were, and finally will define by the purpose of the study and its difficulties.

2.2 Related Work

Various types of data mining techniques are currently playing very important role in the research sector. Especially in some cases of medical science such as some diseases prediction. In this case, many researchers are doing important and unforgettable work using the techniques of Data mining for CKD prediction. In research field, various techniques of data mining are playing a vital role in predicting many complex diseases.

Researchers have investigated in their research work, for predicting major diseases, various classification methods and ML algorithms are used [1]. The aim of their study is, develop the decision-making supporting method for predicting CKD. They have used different

types of machine learning algorithms in their research work, including SVM and KNN. Based on these algorithms, they have come up with the best accuracy and through this, they have been able to decide whether to have chronic disease or not. In their research, they have claimed that KNN has better accuracy than SVM [1].

Vijayarani & Dhayanand explores two of the most popular algorithms of ML, which are respectively SVM and ANN, using these two classification algorithms to determine kidney disease [2]. They have collected about 5,084 data from different types of medical lab hospitals for kidney functional tests where they have collected their data in six attributes and completed their research [2]. Their analysis on this detailed data set shows that ANN are doing well in predicting huge data.

Tangri et al. have revealed that by different analysis of the model by using routinely obtaining lab tests that can rightly predict the progress of both kidney's failure to the sufferings patient with CKD stages which is 3-5 [3]. Mainly their research is on approximately 3449 patients where 386 victims with kidney damages 11% and the similarly 4942 victims 1177 also the failure that means 24% were included in the growth and validation cohorts, respectively [3].

Anderson et al. said their research work that, the progress of renal failure to the ESRD is an expensive as well significant occurrence of clinical morbidity, it's occurs few often in older adults than cardiovascular mortality [4]. They have also claimed that the number of patients with ESRD who are over basically 65 year's age and it's increasing doubled in last almost 25 year, and the rising portion group over the last decade are those over 75 years old [4].

The increasing progress of omics methods, specifically tailored for the discovery towards primary diagnostic with proper follow-up, has made a significant move forward mainly in this work [5].

Fisher & Taylor have analyzed in their research population was split into two independent and distinct samples at random [6]. The multivariable model which is called logistic regression model was built using Sample data. They have collected 5,978 raw data from the different clinics and medical labs [6].

Elhoseny et al. worked for chronic kidney disease, in their paper present the Density-dependent feature selections which is mainly DFS with the D-ACO algorithm, that is basically an intelligent predicting purpose and the classifier methods using for the healthcare [7]. Preprocessing, feature-selection and lastly classification are total three steps of proposed the system that's called D-ACO [7].

Xiao et al. are basically used their work Logistic regression, k-nearest neighbor, XGBoost, RR, SVM, lasso regression, ANN, RF and lastly Elastic Net, were among the nine predictive models developed and compared [8]. After training the data set by all these algorithms, the researchers claimed in their paper which is mainly accuracy of LR is totally better from other classifications. [8].

This research work we have seen that sampling algorithms can increase the efficiency of classification algorithms, and that learning rate is an authentic parameter that has a major impact on multilayer perceptron [9]. In this case, the researchers have used the sampling model in their research paper and have used the two sub-methods called under-sampling and over sampling associated with it. Moreover, they have used the multilayer perception method of the neural network [9].

Researchers have proposed an authentic approach that is Kidney Disease Prediction Monitoring and Application. In this case, they have divided the data of their collected data sets into two parts, one of which is the training dataset and the other one is the test dataset [10]. In their work, they have used the ten most popular algorithms of machine learning. In these ten algorithms they have used to train datasets and Gaussian naive Bayes and decision tree algorithms provide 100% accuracy [10].

Devika et al. compares mainly the accuracy, execution time also precision, of naive bayes algorithm, KNN, and RF classifiers for CKD prediction [11]. Finally, the output of this study shows that, the RF classifier outperforms the Naive Bayes and KNN classifiers [11].

This research paper, we have seen that researchers have collected several clinical data and applied the classification algorithms of machine learning on it [12]. They have used four machine learning algorithms and compared those to experiment with which gave the best result and predicted CKD [12].

Researchers have collected CKD datasets, was retrieved from ML library at the California university, UCI, that contains significant collection of missing output. To fill in the missing outputs, KNN method was used, which chooses multiple full samples of the most close measurements set missing all data in each missing sample [13]. They trained this critical dataset and show that RF algorithm performed better [13].

Suman & Krishan talked about the bad effects of the kidney disease. The Kidney disease data is analyzed using ML classification methods which are mainly DT, naive bayes and ANN [14].

The primary focus of their research is on predicting whether a patient has CKD or not. These models were tested on a recently acquired CKD dataset with 400 raw data records and also roughly 25 attributes, which was downloaded from UCI collection repository [15]. In the contrast, such models with the Multiclass DF classification worked best, with best percent of accuracy for reduced dataset, with approximately fourteen attributes [15].

Sharma et al. said in their research work that, Machine Learning, subdomain of AI, have been extensively using medical experts, physicians by detection and prognosis also diagnosis of different terrible diseases and several health problems in current history [16]. They trained different algorithm and find out best accuracy. They conduct that, decision-tree gives best accuracy [16].

Using only comorbid conditions dataset from NHI (Taiwan), they investigate mainly the feasibility of prediction developing model to estimate the occurrence of RRT3 and duration for first diagnosis of CKD is 6-12 months [17]. A total of 23,948 patients were included in their data collection [17]. Four techniques were used to preprocess the data in order to see whether preprocessing could enhance the performance [17].

Maurya et al. Their aim is to use a machine learning algorithm to recommend an appropriate diet schedule for CKD patients based on medical examination records using a classification algorithm [18]. Dietary advice will be given to patients based on their potassium region, which is determined based on blood potassium levels, in order to slow down the development of CKD [18].

In their research different tests are among the attributes used in the UCI of CKD dataset used in this analysis. The key goal of their work is to measure and also evaluate the efficiency of decision tree classifier [19].

The Clinical Research Data Warehouse at the University of Chicago had access to demographics, position information, vital signs, lab results, interventions, prescriptions, nurse documentation, and diagnostic guidelines. Researchers have used their detailed data to predict kidney disease through statistical analysis [20].

In their article researchers uses two ensemble methods for improving model classification performance: subspace approaches of three learning base KNN, Nave Bayes, and DT classification algorithm [21].

2.3 Comparative Analysis

In this section, we have compared our work of different approaches of researchers with our work and have made a summary analysis of our work**.** Kidney disease is currently the name of a terrible disease. Which is constantly caused by some bad effects on our daily living habits. We have reviewed the works of various researchers in detail in the above section. They have comparatively used many approaches and tried to solve this problem with different techniques. They have used different approaches to machine learning and have achieved the expected results. Following the discussion of numerous aspects of research activities by various research teams, it appears to us that research on CKD is growing than before. This argument has been supported by several positive results. While sufficient resources are lacking, there is hope that this area will become more resourceful with each passing day.

We have properly analyzed the doctrines of various researchers and in some cases have been inspired to bring more innovation to our work. Machine learning classifier has brought incomparable success in the medical science for predicting terrible diseases. We basically tried to simply predict kidney disease using some famous ML classifier. Where we have tried to predict kidney disease by proper classification methods. In our work, we have tried to predict CKD with four algorithms of machine learning. In this case, we have collected the data from different types of patients and completed the training by classifying it in different ways by preprocessing the data in different steps. Then we comparatively

analyzed the best result by both classification algorithms and completed the task by predicting the best result.

2.4 Scope of the Problem

In recent times kidney problems is a very significant major health issue for the public. It is rising at an alarming rate every day. There is no proper model that can be used to explain it. As a result, there is a lot of potential opportunities for this challenge to be solved by evaluating kidney disease symptoms to determine if a patient has CKD or not. The aim of solving this issue is to identify the proper data and using various ML techniques, including particular model with training and also testing. We'll look at the relationships between the dataset attributes to see if they're intertwined in progression of CKD. In Bangladesh, an automatic diagnostic system will speed up the healthcare process. For a better symptom analysis algorithm, the machine will recommend medical tests to users, saving time and money in large hospitals. If there is such a prediction system in different types of hospitals, for that kidney patients can get various treatments from their initial stage. Identifying the disease in its initial stages through that will reduce the possibility of going in chronic stage, which we think it is a breakthrough scope. Doctors can easily cure patients through proper treatment if they refrain from going to the chronic stage. There is a huge scope of human welfare through our work when such a predicting system is created.

2.5 Challenges

In the early stages we had to face various problems to done the work properly. We are facing lots of challenges to do the work. Mainly we are facing various problems in pre-processing the data for our proposal work. After getting the idea of this work, we have tried our best to get the data from different sources, but in this case, we are facing some problem to collecting the proper data. In the purpose of Bangladesh, we are making maximum efforts to liaise with various hospitals and Kidney Disease Institutes. But in the current catastrophic epidemic situation, we have been unable to go to different sources and collect data. In that case, we have tried to gather information through various mediums to move our work forward. Moreover, we have faced various challenges in preprocessing this huge dataset. In this catastrophic situation, we have faced many problems at different times. The most difficult aspect of this work is gathering data on kidney disease in Bangladesh. We used a detailed dataset that had been thoroughly pre-processed. Alternatively, In Bangladesh, locating this kind of dataset is very difficult to collect. The data of the same patient is not matching from all sides. As a result, finding multiple test outcomes for the same individual is difficult. Aside from that, the experiments are often performed in secrecy. But, in most cases, all of the necessary data tests are not completed. As a result, the data is incomplete. So, this type of evidence, inadequate or skewed preparation occurs, resulting in lower accuracy. The type of data used in the model's preparation is often crucial. Since most patients with kidney disease present at a late stage, the model is trained with data in which the majority of the class meaning is labeled as CKD. As a result, the algorithm will be unable to distinguish the initial type of CKD, which will make it difficult to construct an effective model.

CHAPTER 3

Research Methodology

3.1 Introduction

This portion, we have presented our work through theoretical knowledge. So that, we will be able to clear concepts about our research work. We have divided our work of this chapter into different parts. We have also given brief descriptions about what kind of instrumentations are required for data training, preprocessing and also describe the core methods of our research. As we know that to research the Data mining field data plays a very important role. In a word, we can say that to work in this field data is like a heart. We also discussed the data in detail in this segment. Moreover, in this part, we have done statistical analysis in various ways through training and testing of the dataset. The methods of analysis and training on the algorithms of machine learning techniques are also discussed in this section. We have to go through many steps to get the output using the machine learning algorithms. In this case, we have got the final result by following different processes. So, through a proper figure below we have highlighted the methodology part of our research.

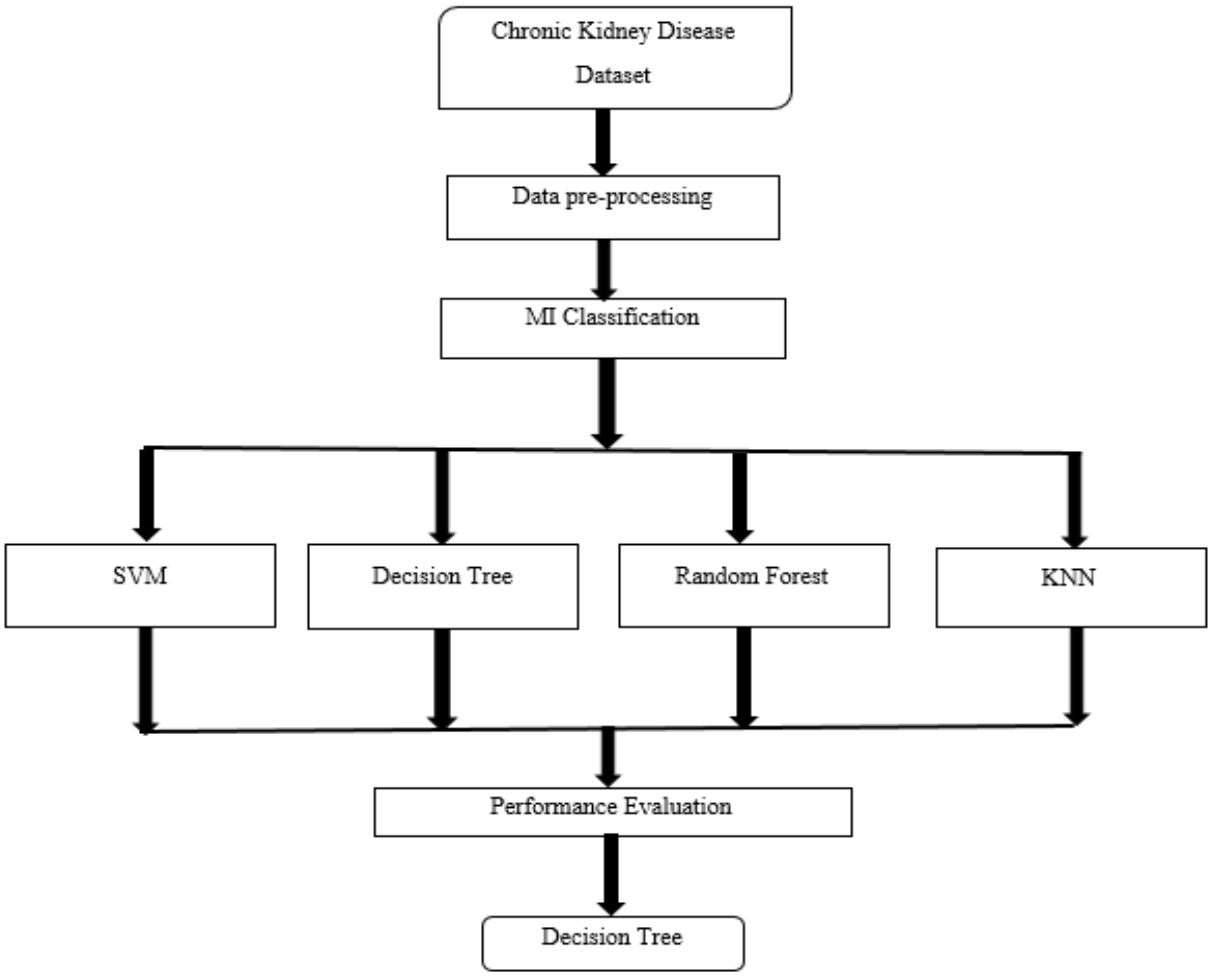


Figure 3.1.1: Workflow of our proposed model

By the above-proposed model, we have expressed a clear workflow of our research. Where at first, we collect the data. Then the next step is to preprocess the data in different steps. Then we completed the training on the classification algorithms of machine learning. In this case, we have used four algorithms to get better results in our research work. After that compared the performance of these classification algorithms and selected the Decision tree classifier that gave the best results.

3.2 Research Subject and Instrumentation

The most important thing to do a research work is that research subject. Because the process of work depends on the subject of research. It is necessary to choose the right subject and

proceed accordingly with a very efficient analysis. In order to do research work, such a researcher has to determine its exact field. Similarly, the research topic reveals the importance of their work and also the reflection of their working method. The subject of a research is very essential for working flow. We have done this research with several patients who have been suffering from kidney disease for a long time. In a word, we have worked to predict CKD. Which means “Chronic kidney disease”.

We did this work thinking about the suffering of different kidney patients. We are using different types of equipment to accomplish this work properly. We have to use the necessary equipment to preprocess the data and make it suitable for machine training and to get accurate results through proper training and testing.

Hardware or Software Requirements

- Operating System (Windows 7 or above)
- Hard Disk (minimum 512 GB)
- Ram (more than 1 GB)
- Web Browser (preferably Microsoft Edge)

Tools for Development:

- Windows- 10
- Python- (3.7)
- Anaconda3
- jupyter notebook
- sklearn
- Pandas
- Numpy
- Seaborn
- Matplotlib

3.3 Data Collection Procedure

Data collection is an important task for various types of research work in machine learning. It's also the foremost vital work for building machine learning models. Appropriate data collection is most important for any research work. Data collection is basically a process of collecting important information from different sources to find answers to research problems and to test estimates and evaluate its results. It is also an important task of making data useful for research through judicial analysis by collecting data from different mediums and using the information needed to solve problems in different strategies. The role of data collection is invaluable in determining and validating different results by combining data from different mediums as required. By selecting the right data it plays a leading role by analyzing all the collected data to solve problem in the research. Which in turn makes the research criteria more accurate and robust. Also, data is the heart of the machine learning approaches. Collecting the right data in the right way for research work is a challenging matter that determines the criteria for research work.

In the current catastrophic situation, we have not been able to collect real-time data, so we have collected basically global data. At this time, we have failed to go to different hospitals, clinics, and different institutions to collect data. So, in this case, we have tried to get the data from different research data sites to collect the right data. Finally, we have collected suitable datasets for our research from an online global platform. The owner of the dataset has made these data available for use by various researchers. In this case, we did our research with 400 data. So, we used data from 400 patients for our research work.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
1	age	bp	sg	al	su	rbc	pc	pcc	ba	bgr	bu	sc	sod	pot	hemo	pcv	wbcc	rbcc	htn	dm	cad	appet	pe	ane	class
2	48	80	1.02	1	0	0	0	1	1	121	36	1.2	?	?	15.4	44	7800	5.2	0	0	1	1	1	1	ckd
3	7	50	1.02	4	0	0	0	1	1	?	18	0.8	?	?	11.3	38	6000	0	1	1	1	1	1	1	ckd
4	62	80	1.01	2	3	0	0	1	1	423	53	1.8	?	?	9.6	31	7500	0	1	0	1	0	1	0	ckd
5	48	70	1.005	4	0	0	1	0	1	117	56	3.8	111	2.5	11.2	32	6700	3.9	0	1	1	0	0	0	ckd
6	51	80	1.01	2	0	0	0	1	1	106	26	1.4	?	?	11.6	35	7300	4.6	1	1	1	1	1	1	ckd
7	60	90	1.015	3	0	0	0	1	1	74	25	1.1	142	3.2	12.2	39	7800	4.4	0	0	1	1	0	1	ckd
8	68	70	1.01	0	0	0	0	1	1	100	54	24	104	4	12.4	36	0	0	1	1	1	1	1	1	ckd
9	24	?	1.015	2	4	0	1	1	1	410	31	1.1	?	?	12.4	44	6900	5	1	0	1	1	0	1	ckd
10	52	100	1.015	3	0	0	1	0	1	138	60	1.9	?	?	10.8	33	9600	4	0	0	1	1	1	0	ckd
11	53	90	1.02	2	0	1	1	0	1	70	107	7.2	114	3.7	9.5	29	12100	3.7	0	0	1	0	1	0	ckd
12	50	60	1.01	2	4	0	1	0	1	490	55	4	?	?	9.4	28	0	0	0	0	1	1	1	0	ckd
13	63	70	1.01	3	0	1	1	0	1	380	60	2.7	131	4.2	10.8	32	4500	3.8	0	0	1	0	0	1	ckd
14	68	70	1.015	3	1	0	0	0	1	208	72	2.1	138	5.8	9.7	28	12200	3.4	0	0	0	0	0	1	ckd
15	68	70	?	?	?		0	0	1	1	98	86	4.6	135	3.4	9.8	0	0	0	0	0	0	0	1	ckd
16	68	80	1.01	3	2	0	1	0	0	157	90	4.1	130	6.4	5.6	16	11000	2.6	0	0	0	0	0	1	ckd
17	40	80	1.015	3	0	0	0	1	1	76	162	9.6	141	4.9	7.6	24	3800	2.8	0	1	1	1	1	0	ckd
18	47	70	1.015	2	0	0	0	1	1	99	46	2.2	138	4.1	12.6	0	0	0	1	1	1	1	1	1	ckd
19	47	80	?	?	?		0	0	1	1	114	87	5.2	139	3.7	12.1	0	0	0	1	1	0	1	1	ckd
20	60	100	1.025	0	3	0	0	1	1	263	27	1.3	135	4.3	12.7	37	11400	4.3	0	0	0	1	1	1	ckd
21	62	60	1.015	1	0	0	1	0	1	100	31	1.6	?	?	10.3	30	5300	3.7	0	1	0	1	1	1	ckd
22	61	80	1.015	2	0	1	1	1	1	173	148	3.9	135	5.2	7.7	24	9200	3.2	0	0	0	0	0	0	ckd
23	60	90	?	?	?		0	0	1	1	?	180	76	4.5	?	10.9	32	6200	3.6	0	0	0	1	1	ckd
24	48	80	1.025	4	0	0	1	1	1	95	163	7.7	136	3.8	9.8	32	6900	3.4	0	1	1	1	1	0	ckd
25	21	70	1.01	0	0	0	0	1	1	?	?	?	?	?	?	0	0	0	1	1	1	0	1	0	ckd
26	42	100	1.015	4	0	0	1	1	0	?	50	1.4	129	4	11.1	39	8300	4.6	0	1	1	0	1	1	ckd
27	61	60	1.025	0	0	0	0	1	1	108	75	1.9	141	5.2	9.9	29	8400	3.7	0	0	1	1	1	0	ckd
28	75	80	1.015	0	0	0	0	1	1	156	45	2.4	140	3.4	11.6	35	10300	4	0	0	1	0	1	1	ckd
29	69	70	1.01	3	4	0	1	1	1	264	87	2.7	130	4	12.5	37	9600	4.1	0	0	0	1	0	1	ckd

Figure 3.3.1: Sample of Data Set

3.4 Data Preprocessing

Preprocessing a dataset is a technique of ML that is basically used to efficiently format raw data as required. To use data for a specific purpose, it is very important to preprocess the collected data. Because preprocessing data is very important in order to get accurate and improved results by arranging random data in a certain way and preparing it accurately. We preprocessed the data after collecting the dataset because in the real-time dataset are always some lost or contains garbage values. As a result, we are trying to smooth out the messy values. Convert the data through a string to numerical values. The data set we worked with contains data on kidney patients at different stages. Our dataset has 26 columns where the name of all columns are 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. In this case, KFT test is shown in the last column of the

dataset. KFT test is mainly Kidney Functional Test. How many patients are in CKD and how many are not in CKD. Moreover, the data in the last column is basically divided into two parts. In this case, the Kidney function test results of 250 patients in the first row have CKD and the results of the remaining 150 patients that have not CKD. Since there were lots of problems in the collected dataset. So, In this case, we have cleaned the data set properly with a different type of process. We have cleaned unnecessary symbol in numeric values and some missing and Garbage values

3.5 Statistical Analysis

Statistical analysis is a very important task for research work. It mainly enriches a research work through different analyses. Science our data set contains 400 data and after collecting the data we have preprocessed it different ways. So, when working with raw data, the success of the work usually, depends on preprocessed data. Which means, the more accurate data we will pre-processed, our machine will perform accurately and give better performance. It's the initial challenging part for Research-Based criteria. In the above section, we have described the part of data pre-processing in detail.

We have completed our data coding portion in order to arrange the data according to our favor. Usually, we have arranged the dataset in our own way to work properly and simply. At the beginning of the coding part, first of all, we have input some methods of Pandas. Then we have used some functions to replace symbols from raw data with null values and made it suitable for training and testing.

After successfully selecting features, we are prepared to partition data, which is done through training our proposed machine. We have divided the whole data by 7:3 ratio. The first seven portion of the data is used for our machine training and the last rest of the portions of our data used for testing. In other words, out of the total 400 data, 280 data has been used for training and 120 data for tasting. If we look at the ratio, it stands at 70% data used for training and 30% for testing. So, in this stage, the whole data is fully fit and suitable for classification. So, now we've trained our data set with a proper classification algorithm through proper training and analysis. In this stage, we are using four famous ML

algorithms to predict CKD properly. For these, Sklearn has a built-in classifier. We simply imported it and also fit it properly.

This is the last step of our CKD classification prediction process. Our model is currently being prepared for the testing of all other symptoms as the input data. After that, we can see from this four algorithm Decision tree algorithm show better output performance. That means Out of all the algorithms, Decision tree performs the best accuracy result. In this case, decision tree classifier will be able to give the best and accurate prediction result for predicting CKD. We have tried to show the whole process through a flowchart,

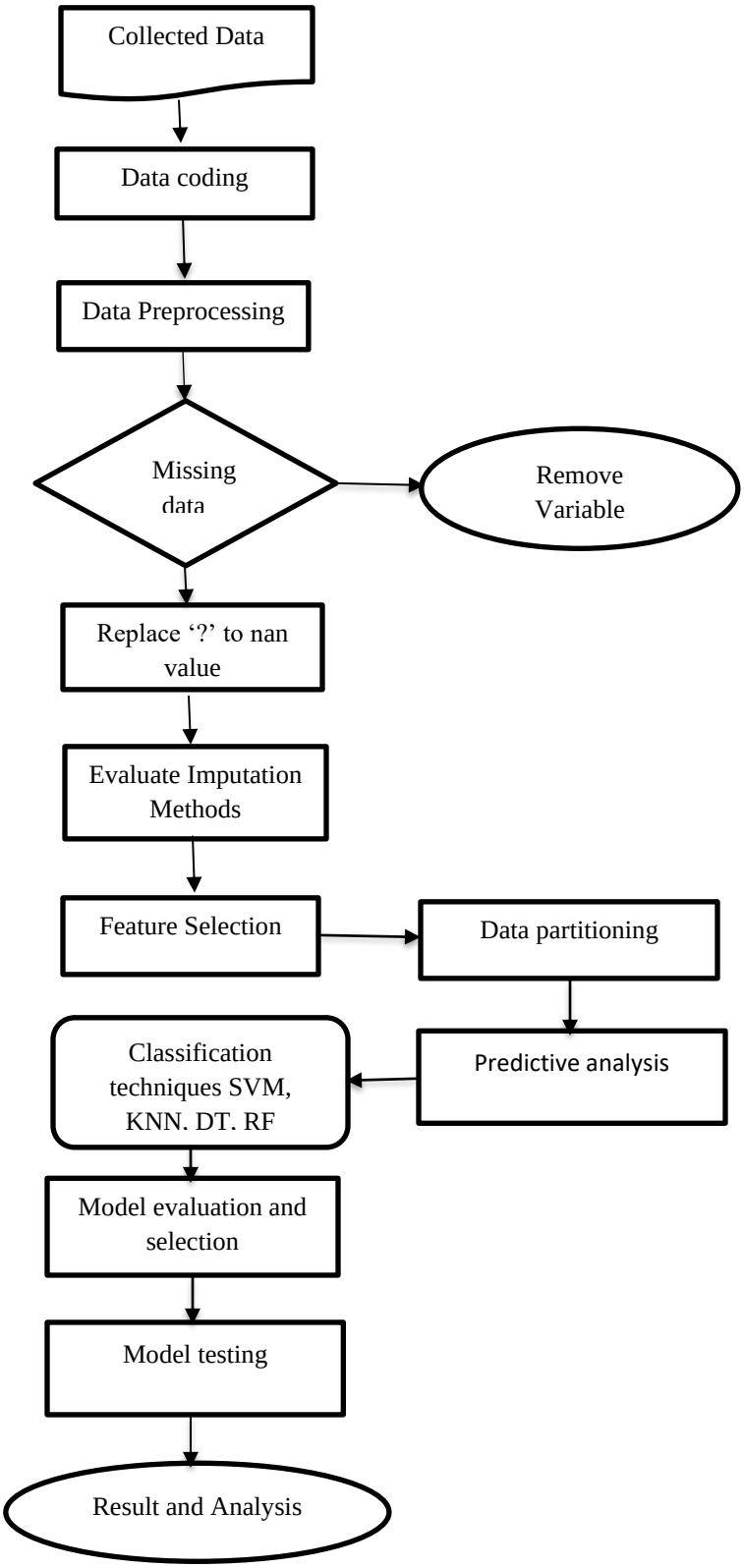


Figure 3.5.1: Flowchart of CKD analysis & prediction

3.6 Proposed Methodology

3.6.1 Classification Algorithms

The Classification algorithm is known as the Supervised Learning method that uses trained data by determine the specific category of new findings. Mainly aim of ML classifier is to predict the target class by evaluating the training dataset. Determine the exact borders for each goal class. In general, stronger boundary states can be obtained by using the training dataset, which can then be used to evaluate each target class. In this process is called classification.

At present ML classifiers are leading vital role in various types of research work. Machine learning algorithms are different categories. Different types of researchers train their datasets through machine learning algorithms according to the benefits of their research. In our work, we have used various classification algorithms of machine learning. These classification algorithms are popular for different tasks at this time. These algorithms are also widely used in prediction work from a variety of functions. We have used four popular algorithms of machine learning for our work. Through these classification algorithms, we have been able to successfully find out which algorithm gives the best results to predict chronic kidney disease by training the dataset in the proper classification method.

3.6.2 Support Vector Machine

SVM is very famous supervised classifier technique that analyzing the data for mainly classification and also regression analysis in ML. SVMs, which are built on statistical learning models, is one of the most reliable prediction approaches. It is very popular in research areas for prediction in different types of research. SVM algorithm generally creates a basic model that's adds an examples in one of the two categories and rendering a non-probabilistic of binary linear classification, given training examples of a collections, per labeled belongs to 1 of 2 categories. SVM is the probabilistic classification setting, such as Platt scaling. SVM maps teaching examples to points in space to widen the distance between the two groups as far as possible. The minimum distance between a pattern and a decision surface is known as the margin. As a result, if we constrain the margins of the function class from underneath, we can control complexity. This is accomplished by support vector learning, which recognizes that when the margin is maximized, threat is

reduced. By it, we are attempting to solve a text classification problem or other prediction problem. It is possible to match the full margin hyperplane in a function space S using the kernel function for SVM. The function space S is a non-linear projection $\Phi : R\ N \rightarrow$ S from the actual input space, which is usually much larger in dimensionality than the initial input space.

3.6.3 Decision Tree

In ML classification algorithm DT is most popular one from other classifiers. Internal nodes represent dataset attributes, branches represent judgment laws, and each left node represents the outcome in this classifier. Mainly decision trees have core two nodes. Decision Node and the Leaf Node are the two nodes that make up a Decision tree. A leaf node is the output of such decisions and doesn't have any other branches, while Decision core nodes are used to make a decision and have several branches. The test or decision is made based on the characteristics of the dataset. We use the CART algorithm that stands for Regression Tree algorithm and Classification to construct a tree. A decision tree essentially poses a question and divides the tree into different subtrees depending on the answer (Yes/No). Machine learning is the compact of various algorithm. So, finding the proper classifiers for dataset mainly problem is an important issue to consider when we build a ML predictive model. DT are designed to imitate human reasoning abilities when making decisions, making them very simple to comprehensive. Predicting the dataset class such a DT it begins basically the tree that means in the main root-node of each tree. It also checks exact value of root attribute with the records of root value.

3.6.4 Random Forest

It is a robust, easily useable ML techniques that, in most cases, produces outstanding results without any hassle. Because of its simplicity and adjustability, it is perhaps one of the most popular algorithms. It is usually used for mainly classification and also regression works.

It also known learning algorithm which is supervised. It creates "forest" out of an ensemble of decision trees, which are normally educated using the "bagging" technique. The bagging method's basic premise is that combining different learning models improves the outcome [28]. To get a more precise and reliable forecast, random forest creates several decision trees and combines them. Leo Breiman proposed the Random Forest (RF) classification system, which is an easy, highly reliable, and noise-resistant machine learning classification method. Bagging and function selection at random are mixed [28] This algorithm basically works on the dataset by following few steps:

Step 1: At first start with the selection of random samples from the given dataset.

Step 2: Following that, this algorithm would build a decision tree for each sample. The prediction outcome from each decision tree would be extracted.

Step 3: Voting will take place in this stage with each expected outcome.

Step 4: At last, selecting the maximum voted prediction result as the main prediction output.

3.6.5 K-Nearest Neighbor

The KNN is built on the supervised learning methodology and most popular and simplest techniques of ML algorithms. This classifier assumes new case or data and existing cases are identical places in new cases in that category. It's more similar to by the existing categories. K-NN algorithm records all the relevant data and classifies new data points depending on the resemblance. This ensures as new data occurs then it can be quickly sorted into a good suite group by using K- NN algorithm. The K-NN algorithm is a non-parametric classification algorithm, which means it makes no assumptions about the underlying results. It's also known as a lazy learner algorithm because it doesn't learn from the training set right away; instead, it saves the dataset and operates on it when it comes time to classify it. During the training process, the KNN algorithm simply stores the dataset. The KNN algorithm of ML works on a dataset by following some steps. KNN operates by calculating the distances between a sample and all of the examples in the results, choosing the K nearest examples to the query, and either voting for the most

frequent label or averaging the labels (in case of regression). The scikit-learn module is used to import the k-nearest neighbor algorithm. Since it makes highly precise forecasts, the KNN algorithm will contend with the most accurate ones. As a result, the KNN algorithm can be used in applications that need high precision but don't need a human-readable approach. The distance calculation affects the accuracy of the forecasts. Both classification and regression predictive some problems can be solved with KNN. However, in the field, it is most often used in classification issues [29].

3.7 Performance Evaluation Metrics

After completing the data preprocessing part, the dataset is divided into train and test subsets for training and evaluating the machine learning algorithm. For that "performance metrics" is come. The metrics used to evaluate the model's output are referred to as performance evaluating metrics. To assess and comparison the each performance of Supervised ML models, a wide range evaluation metrics are usable.

3.7.1 Accuracy

Accuracy refers to the percentage of accurate predictions made by our model. One way to calculate how much a classification algorithm correctly classifies a data point is called accuracy [30]. Accuracy is important for analyzing performance of each classifier.

$$\text{Accuracy} = \frac{\text{Number of correct prediction}}{Total\ number\ of\ prediction}$$

However, as we all know, evaluating a model solely based on its performance metrics is insufficient. If there are an equivalent number of instances of the search target class, the classifier's accuracy would improve. Fortunately, there are nearly equal numbers of every class in this dataset. Here's the proof that equality.

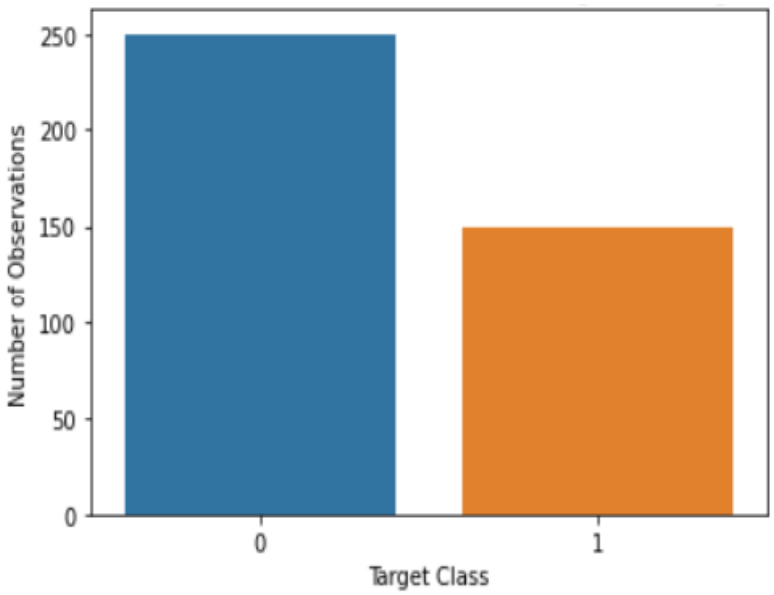


Figure 3.7.1.1: Total number of records for each class

In the figure above we can see that there are two target classes denoted by 0 and 1 value. Here positive result refers to 0 which means CKD and not CKD refers to 1 which means a negative result. So that the positive and negative test results are basically represented by this above figure.

3.7.2 Confusion Matrix

Confusion matrix basically the method of summarization and a classifier algorithm results. If we've an uneven value of observation of a particular classes or dataset as more then two classes, that purpose classification accuracy for each class can be deceptive. Calculating a confusion matrix will help us to see what the classification model is doing correctly and where it is going wrong.

Table 1: Details of the Confusion Matrix

		Actual	
		Positive (1)	Negative (0)
Predicted	Positive (1)	**TP**	**FP**
	Negative (0)	**FN**	**TN**

• TP: True positive. When all of the real data points are positive and the expected outcomes are positive as well.

• FP: False positive. In this case, the actual points of data are mainly negative but it also predicted such as positive.

• FN: False negative. Here the actual points of data are mainly positive but it also predicted such as negative.

• TN: True negative. Here the actual points of all data are mainly negative and it also predicted such as negative.

As a result, the accuracy formula is,

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

3.7.3 Precision

Precision is positive predictive values, it's the fractions of the saturated instance between the each retrieved instance, while recall is the fractions of relevant each instances which were retrieved with pattern detection, info retrieval, and lastly classification. It indicates how many of the records listed as positive values are accurately positive. The proportion of all expected true records (TP+FP) with positive records basically which is denoted by (TP). Precision denoted the accuracy with the result have achieved. Thus, if the model accurately captures only one positive event, the model is 100 percent accurate.

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

3.7.4 Recall

It indicates how many of the positive records are expected to be positive. This is the proportion of all the positive records (TP+FN) to the real positive records (TP). The recall is not the case of correctly capturing the record, but rather of correctly capturing both positive matters denoted as positive. It works by following a formula to get results by training the dataset by algorithm.

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}$$

3.7.5 F1-Score

It's basically harmonic meaning with both recall and the precision. It is also called F1-measures. It is calculated from the test score obtained from precision and recall value. F1-score is very important for testing a model to get the most accurate results. While the

model's high precision and low recall result in an incredibly precise model, there are also a considerable number of cases that are difficult to categorize. The model's success is accelerated by F1-Score.

$$\text{F1 Score} = \quad 2 * \frac{1}{\frac{1}{Precision} + \frac{1}{Recall}}$$

CHAPTER 4

Experimental Results

4.1 Introduction

We have basically shown here, the experimental results of our work and analyzed the outcomes. we will usually analyze the result of the proposed model of our work and explain the result of our work with proper logic. We have already discussed in detail how to preprocess our data, here we will also discuss how we have completed the training and testing of our dataset. Also, we will discuss the desired results obtained from the trained dataset. Since we have collected our dataset from the website, we have had to modify this dataset in various ways. After receiving the dataset, we have stored that dataset file in a csv file. Also there has some missing value in the file, which mainly we are filled in using the Pandas tool to get value. As a result, the dataset became useful and the preprocessing was completed. We have divided our dataset into two parts to building the model properly.

- Training Dataset
- Testing Dataset

We will analyze our training and testing data through a figure. We have built our model in a 7: 3 ratio. So, we have considered three portions of the ten parts as the tasting dataset and we have considered the rest among the portion as the trained dataset. Following this feature, we have basically divided this dataset into different portions to build the model.

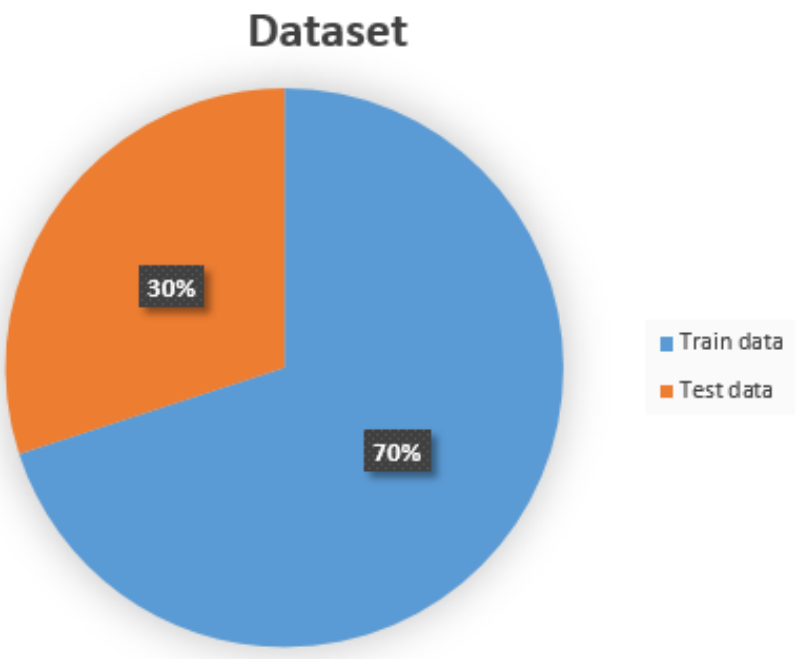


Figure 4.1.1: Dataset chart ratio

The dataset for our research we have collected from an online medium where the total number of data is 400 and 26 attributes. Whereas in this huge amount of data, 25 out of 26 attributes are for the input variable and lastly one attribute is for the output variable. Moreover, 12 of these variables are numerical. Moreover, others are basically nominal. Also, the output of the 26 attributes contains the result of each kidney patient for their kidney functional test (KFT). So, the last column shows two types of output variable and which is CKD or Not CKD.

We have used a variety of ML algorithms to predict CKD. We are using 4 most popular algorithms for our work. These classification algorithms are SVM, DT, RF, and KNN algorithm respectively. We used these algorithms primarily to predict chronic kidney disease. By training the dataset through these algorithms, we get Accuracy, precision, recall value, and F-1 score.

4.2 Experimental Results

In this section, we will mainly show the outputs of our trained algorithm and analysis the result of this algorithm. Moreover, we have got accuracy for each classifiers and precision values, Recall and also f1 Score separately for each algorithm. We will show it, particularly through different tables and figures.

For Support Vector Machine (SVM) Classifier

Table 2: Expected Result of SVM

accuracy	**95.83%**
precision	**0.95**
recall	**0.93**
F1-**s**core	**0.94**

Confusion Matrix:

[[70 2]

[3 45]]

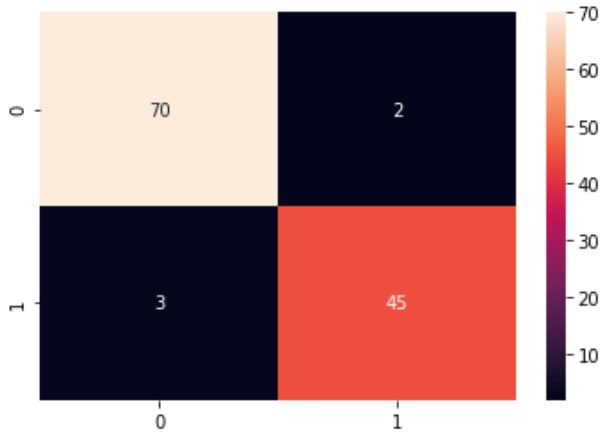


Figure 4.2.1: Confusion Matrix for SVM

For Decision Tree Classifier:

Table 3: Expected Result of Decision Tree

Accuracy	**96.66%**
Precision	**1.0**
Recall	**0.91**
F1-Score	**0.95**

Confusion Matrix:

[[72 0]

[4 44]]

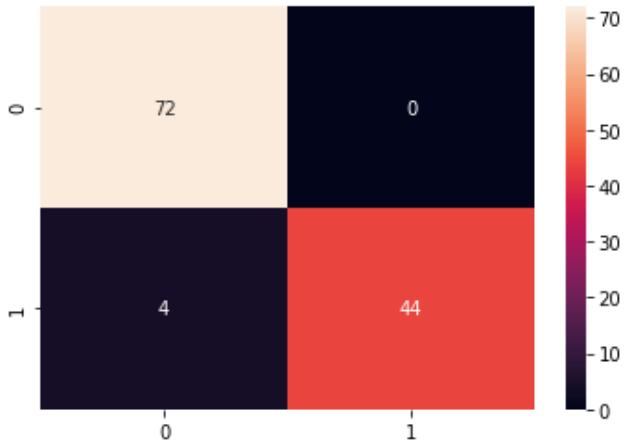


Figure 4.2.2: Confusion Matrix for Decision Tree

For Random Forest Classifier:

Table 4: Expected Result of Random Forest

Accuracy	**95.83%**
Precision	**0.97**
Recall	**0.91**
F1-Score	**0.94**

Confusion Matrix:

[[71 1]

[4 44]]

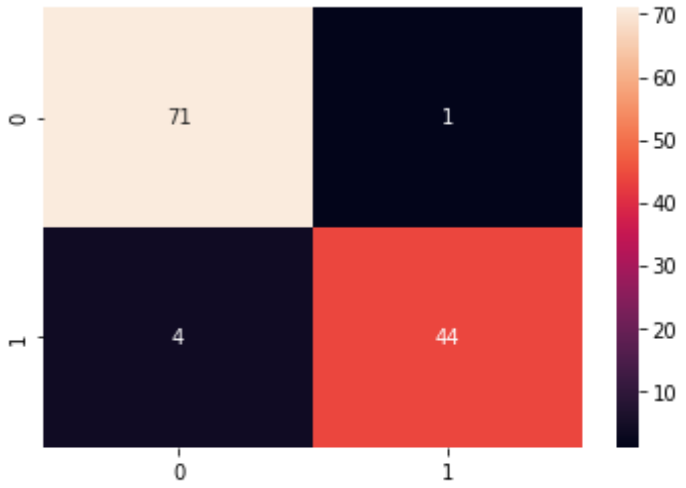


Figure 4.2.3: Confusion Matrix for Random Forest

K-Nearest Neighbor:

Table 5: Expected Result of KNN

Accuracy	**95.0%**
Precision	**0.93**
Recall	**0.93**
F1-Score	**0.93**

Confusion Matrix:

[[69 3]

[3 45]]

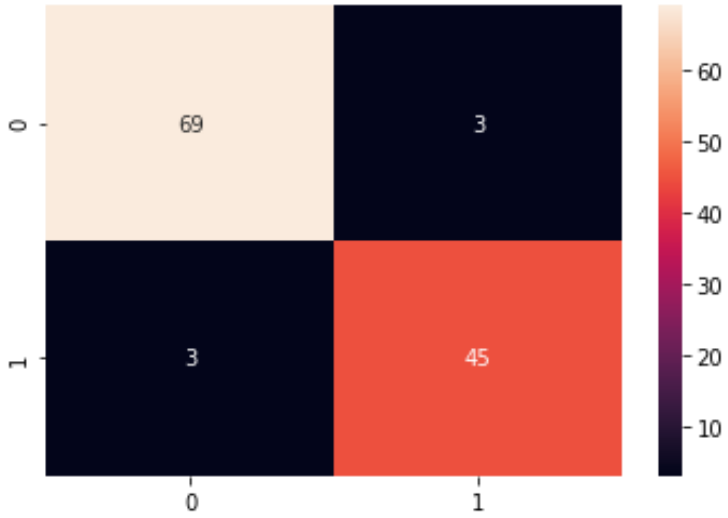


Figure 4.2.4: Confusion Matrix for KNN

4.3 Descriptive Analysis & Discussion

This segment, we will verify and also compare particular results of each classifiers trained by the collected dataset. In the above chapter, we have shown the results of each algorithm through different tables and figures. At this stage, we will discuss the outcomes that have come from the classification algorithms through a bar chart. Here, we have highlighted the accuracies of four algorithms in the following chart. From the output mainly we get separately Accuracy, precision, recall, F1-score, and confusion matrix from each of the algorithms. The output variable has only two types, CKD and not CKD, in the data used in this experiment.

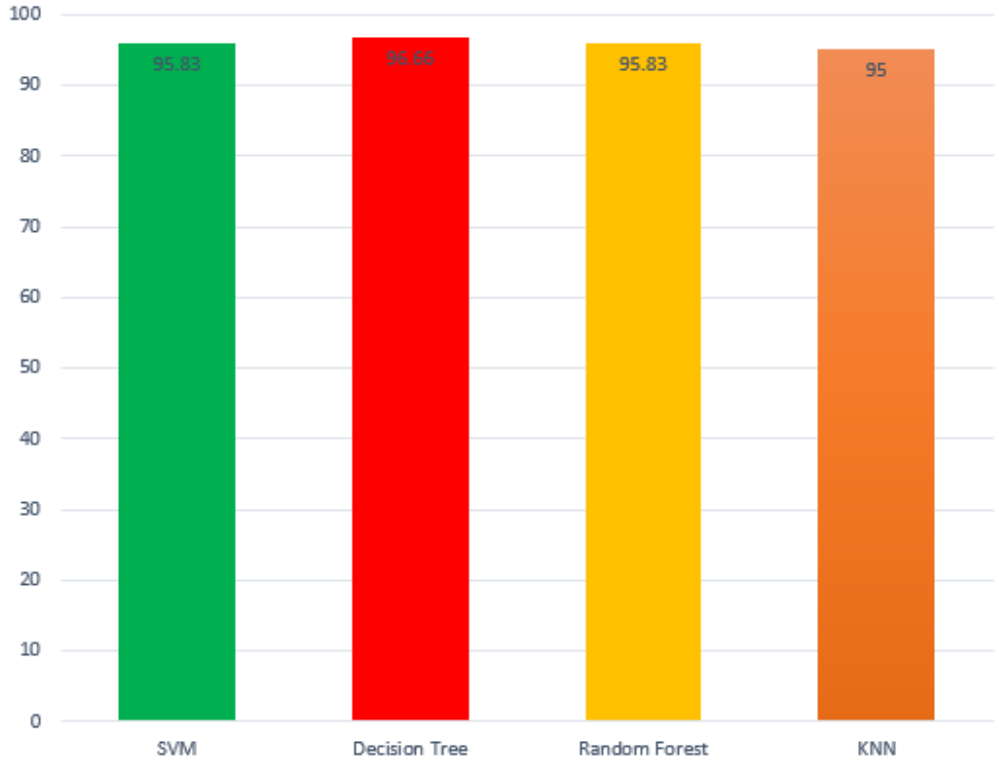


Figure 4.3.1: Accuracy differences Bar chart

After properly training the classification algorithms of machine learning, we get good performance from almost all the algorithms. But through analyzing the result efficiently.

we can see that the Decision tree classification algorithm has given the best accuracy performance. This algorithm has shown better results than other algorithms. From this algorithm, we got maximum accuracy which is 96.66%. Based on which we can say that the Decision Tree classifier has given the best outcome for prediction. In that case, If our dataset was equipped with more patient data, our accuracy level would be more developed. Because the more accurate data is processed, the algorithm will better perform. So, we want to say that in our research experiments, the Decision tree algorithm has given the best accuracy among the other algorithms of ML to predict CKD.

CHAPTER 5

Impact on Society, Environment & Sustainability

5.1 Impact on Society

Every year, 35,000-40,000 CKD patients, out of a total population of 18 million, in Bangladesh develop kidney failure, according to the Kidney Foundation [25]. Chronic nephropathy (40 percent), diabetes (34 percent), and hypertension (15 percent) are the leading causes of ESRD in our country. people with these diseases should seek care as soon as possible. Doctors in our country and various doctors around the world tell patients to be aware of the first stage of kidney disease and take proper treatment. But in this case, due to the unawareness of the patients and the cost of additional expensive treatment, they initially neglected the disease. As a result, they have to face chronic diseases. In such a situation, it is very important to come up with such a system so that the patients can save themselves from costly expenses of treatment and get proper treatment in a short time. Then we think that it will be very beneficial for society and the common people. We think that our work will play a very important role in society. If a prediction system can be developed that can quickly monitor the early stages of kidney disease and predict Chronic kidney disease. Then it will be possible to reduce the suffering of society and the helpless people to a great extent. They will be freed from expensive treatment. By which we think that it will always have a great impact on the helpless people around us and our society.

Moreover, we think that our work will have a positive impact on society. Because our work will help alleviate the suffering of people from different walks of life. Every time we see the people around us being in a miserable situation, we will be able to reduce a lot through this work. We did this mainly for those people who were suffering from Chronic Kidney Disease.

If we can create such a prediction system according to our work, we will be able to save kidney patients from costly treatments. we will try to reduce the suffering of their difficult treatment. So that we think it will bring far-reaching results for the society and the country. If a system can be created according to our proposal, it will have a positive impact on society. Our work will not be affected society in any way. Rather it will bring far-reaching

results for society and the country. So that the patients suffering from kidney disease can get rid of the terrible consequences like chronic kidney disease through proper treatment very quickly.

5.2 Ethical Aspects

Everyone is acquainted with the severity of kidney disease and we are also aware of what kind of suffering this problem affects a patient's family and the patient. We all know the severity of this disease and the tragic consequences for the patient. Long stretches of serious illness, multiple hospitalizations, and an enormous amount of burden put on patients and their families. Over than 100,000 people in the USA are hoping for a kidney transplant, but only around 21,000 donated kidneys were available last year [27]. The moral aspect of our work is that we do it for the benefit of the people.

We have completed the training and testing through ML algorithms and find out which classifier will give best results to predict CKD. In this case, we have analyzed the huge data and completed the training.

In doing this we have tried our best to get the work done by adopting some ethical aspects. we have done the work with honesty from all sides. We have failed to properly collect data by visiting various institutes or hospitals for the current catastrophic situation. But despite that, we have done our work with honesty.

We have set our highest goals so that we do not have any bad effects. We do not use any type of document or information without proper reference. We have tried to work honestly with proper references in all the cases where we have collected and used the information of our work from different sites or research papers. We do not use any data without reference from any type of document. We have done this from within the policy.

Also, our work will not cause any harm. we have done this in the interest of human welfare. Considering the dire consequences for kidney patients, we have done this to alleviate their suffering.

So, above all, we have maintained 100% ethical aspects in our work. We did not do things that would be a threat to people or they would face obstacles in their daily lives.

Lastly, we would like to say that we are doing this research work honestly and it is not our intention to do anything against the policy. So, we have tried our best to do something unique with our efforts.

5.3 Sustainability Plan

The number of kidney patients in different hospitals and clinics is constantly increasing. It is too much difficult for a patient to bear the whole cost of its treatment. So far still now no solution has been found. Through which the suffering of the patients can be reduced easily by low cost. According to our proposal, if it is possible to create such a system in the future, the suffering of patients can be reduced. In our research, we have made appropriate use of machine learning algorithms. In this case, we have got a better outcome from Decision tree classifier. So, we hope that better results will come with more amounts of accurate data. It will be possible to predict CKD by using a proper prediction system according to a better sustainable plan. Which will certainly reduce the suffering of the patients. It is possible to create a predicting system that works in the right way for proper planning and implementation. In that situation, we need the help of the government and cooperation with experienced doctors and mostly needed an accurate sustainable plan.

CHAPTER 6

Conclusion & Future Work

6.1 Summary of the Study

Presently kidney patients in our country are rapidly increasing. Huge kidney patients in the different hospitals around us is constantly increasing. In the early stages of this disease, common symptoms are seen, for which people forget about the horrible consequences of this disease. After a long time, the disease becomes severe. As a result, patients develop chronic kidney disease. It's increasing rapidly day by day. As a result, identifying the risk factors will help to minimize the number of patients who are affected by this disease. It's difficult to claim that kidney disease has no solution at the end stage, but it is possible to extend life expectancy through proper medical treatment and care. Kidney failure will make the patient believe he or she is dying. It is not possible to bring them back to normal life from this last stage. We are inspired to do this by considering such affected people. The current cost of kidney disease is extremely difficult for patients to bear. Our goal is to create a more advanced and capable (ML) application that can properly predict the CKD status. Mainly the whole procedure has been separated into two parts in this process. A training dataset was created in one section, while a test dataset was used to test another section. In this study, four of the most important and efficient machine learning approach procedures are used that is considered for predicting major kidney issue. So, we did this work for the patients in the early stage who do not have to suffer from CKD. ML algorithms are capable to predict CKD.

6.2 Conclusions

In the context of our country, kidney disease is a terrible disease at present. If a person has been suffering from this disease for a long time, later it may take the form of CKD. Treating this incurable disease is very difficult and also expensive for patients. Many lives are being lost prematurely without proper treatment for this terrible disease. We have done this work for the general people so that they could diagnose the disease at an early stage and then get a prediction before going to the chronic stage. In a word, we have done this research for human welfare. In this case, various techniques of machine learning are being used very popularly in the medical sector for diagnosing various diseases and predicting diseases. Different techniques of machine learning are also playing an important role in diagnosing complex diseases through proper training of the machine. In our research purpose, our collected dataset has different types of patient diagnosis data. We have trained our dataset using classification algorithms. We have done our work using four popular classification algorithms of ML through various analyzes. After that, we got our expected results. We have got accurate accuracy by training detailed datasets using ML techniques. In that case, Decision Tree algorithm has given the best accuracy. We have tried our best to get better accuracy. However, we got good results from these four algorithms but the decision tree classifier showed the best performance. We think that our work will be more robust if the predicting system can be created properly according to our research. Then we will be able to reduce the suffering of the people and our work will also play a leading role in various researches. In conclusion, we have done our best to do our work properly and our main purpose in doing this research is to predict the CKD. Finally, we would like to say that our work will be used for the benefit of many people in the future and it will make a groundbreaking contribution to predicting chronic disease.

6.3 Implication for Further Study

Machine learning techniques are currently being used in various fields in the medical sector. Therefore, we think that it is possible to bring a lot of improvements in the medical field in the future by using the machine learning approach properly. In our work since we have worked with 400 data and could not work by collecting more data for this pandemic situation. So, in the future, if we can do the work by going to the different clinics, institutes,

and hospitals and collecting more data, the quality of work will be more efficient. Because if we use more data in our work then we will get better output from our trained dataset. We also trained the machine so that it can predict CKD properly. But using this work in the future, it is possible to predict more complex diseases. Several diseases can be predicted using some ML techniques with accurate data. Which we think will be very beneficial for people in future. Moreover, in the future, we can implement our work through a huge field like deep learning. Then the work will be more accurate and better. We can also add more categories for our work and use other machine learning classification algorithms, so that we can determine which algorithm will give a better result.

Moreover, through advanced planning in the future, we can make this work useful for the public by using proper mobile or web-based applications. We think that if such a system is invented then kidney disease and various complex diseases can be predicted very easily. Through which human welfare will come and we will be able to reduce the suffering of different patients easily.

REFERENCES

[1] Sinha, P., & Sinha, P. (2015). Comparative study of chronic kidney disease prediction using KNN and SVM. *International Journal of Engineering Research and Technology*, *4*(12), 608-12.

[2] Vijayarani, S., Dhayanand, S., & Phil, M. (2015). Kidney disease prediction using SVM and ANN algorithms. *International Journal of Computing and Business Research (IJCBR)*, *6*(2), 1-12.

[3] Tangri, N., Stevens, L. A., Griffith, J., Tighiouart, H., Djurdjev, O., Naimark, D., ... & Levey, A. S. (2011). A predictive model for progression of chronic kidney disease to kidney failure. *Jama*, *305*(15), 1553-1559.

[4] Anderson, S., Halter, J. B., Hazzard, W. R., Himmelfarb, J., Horne, F. M., Kaysen, G. A., ... & High, K. P. (2009). Prediction, progression, and outcomes of chronic kidney disease in older adults. *Journal of the American Society of Nephrology*, *20*(6), 1199-1209.

[5] Mihai, S., Codrici, E., Popescu, I. D., Enciu, A. M., Albulescu, L., Necula, L. G., ... & Tanase, C. (2018). Inflammation-related mechanisms in chronic kidney disease prediction, progression, and outcome. *Journal of immunology research*, *2018*.

[6] Fisher, M. A., & Taylor, G. W. (2009). A prediction model for chronic kidney disease includes periodontal disease. *Journal of periodontology*, *80*(1), 16-23.

[7] Elhoseny, M., Shankar, K., & Uthayakumar, J. (2019). Intelligent diagnostic prediction and classification system for chronic kidney disease. *Scientific reports*, *9*(1), 1-14.

[8] 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), 1-13.

[9] Rabby, A. S. A., Mamata, R., Laboni, M. A., & Abujar, S. (2019, July). Machine learning applied to kidney disease prediction: Comparison study. In *2019 10th international conference on computing, communication and networking technologies (ICCCNT)* (pp. 1-7). IEEE.

[10] Rabby, A. S. A., Mamata, R., Laboni, M. A., & Abujar, S. (2019, July). Machine learning applied to kidney disease prediction: Comparison study. In *2019 10th international conference on computing, communication and networking technologies (ICCCNT)* (pp. 1-7). IEEE.

[11] Devika, R., Avilala, S. V., & Subramaniyaswamy, V. (2019, March). Comparative study of classifier for chronic kidney disease prediction using naive Bayes, KNN and random forest. In *2019 3rd International conference on computing methodologies and communication (ICCMC)* (pp. 679-684). IEEE.

[12] Charleonnan, A., Fufaung, T., Niyomwong, T., Chokchueypattanakit, W., Suwannawach, S., & Ninchawee, N. (2016, October). Predictive analytics for chronic kidney disease using machine learning techniques. In *2016 management and innovation technology international conference (MITicon)* (pp. MIT-80). IEEE.

[13] Qin, J., Chen, L., Liu, Y., Liu, C., Feng, C., & Chen, B. (2019). A machine learning methodology for diagnosing chronic kidney disease. *IEEE Access*, *8*, 20991-21002.

[14] Bala, S., & Kumar, K. (2014). A literature review on kidney disease prediction using data mining classification technique. *International Journal of Computer Science and Mobile Computing*, *3*(7), 960-967.

[15] Gunarathne, W. H. S. D., Perera, K. D. M., & Kahandawaarachchi, K. A. D. C. P. (2017, October). Performance evaluation on machine learning classification techniques for disease classification and forecasting through data analytics for chronic kidney disease (CKD). In *2017 IEEE 17th international conference on bioinformatics and bioengineering (BIBE)* (pp. 291-296). IEEE.

[16] Sharma, S., Sharma, V., & Sharma, A. (2016). Performance based evaluation of various machine learning classification techniques for chronic kidney disease diagnosis. *arXiv preprint arXiv:1606.09581*.

[17] Dovgan, E., Gradišek, A., Luštrek, M., Uddin, M., Nursetyo, A. A., Annavarajula, S. K., ... & Syed-Abdul, S. (2020). Using machine learning models to predict the initiation of renal replacement therapy among chronic kidney disease patients. *Plos one*, *15*(6), e0233976.

[18] Maurya, A., Wable, R., Shinde, R., John, S., Jadhav, R., & Dakshayani, R. (2019, January). Chronic kidney disease prediction and Recommendation of Suitable Diet plan by using Machine Learning. In *2019 International Conference on Nascent Technologies in Engineering (ICNTE)* (pp. 1-4). IEEE.

[19] Maurya, A., Wable, R., Shinde, R., John, S., Jadhav, R., & Dakshayani, R. (2019, January). Chronic kidney disease prediction and Recommendation of Suitable Diet plan by using Machine Learning. In *2019 International Conference on Nascent Technologies in Engineering (ICNTE)* (pp. 1-4). IEEE.

[20] Koyner, J. L., Carey, K. A., Edelson, D. P., & Churpek, M. M. (2018). The development of a machine learning inpatient acute kidney injury prediction model. *Critical care medicine*, *46*(7), 1070-1077.

[21] Jongbo, O. A., Adetunmbi, A. O., Ogunrinde, R. B., & Badeji-Ajisafe, B. (2020). Development of an ensemble approach to chronic kidney disease diagnosis. *Scientific African*, *8*, e00456.

[22] Inc, N. K. (2021 , 02 15). Retrieved from https://www.kidney.org/kidneydisease/global-facts-about-kidney-disease

[23] Nitta, K., Okada, K., Yanai, M., & Takahashi, S. (2013). Aging and chronic kidney disease. *Kidney and Blood Pressure Research*, *38*(1), 109-120.s

[24] (2020, September 3). Retrieved from https://builtin.com/data-science/random-forest-algorithm?fbclid=IwAR0yemQU4X7jtFu1azaqBcEpPR8XTNwc-4sS6QEFfpwrfssSXMKEx2B0PFY

[25] 2020, 09 16). Retrieved from https://pharm.ucsf.edu/kidney/need/statistics?fbclid=IwAR0DQBp8P9kSjncGv7sdaohGZRLa5x5A0tfNbYk5I2VVHMVfxQo_VZSKtvg

[26] Hickey, K. M. (1972). Impact of kidney disease on patient, family, and society. *Social Casework*, *53*(7), 391-398.

[27] (2021, 1 14). Retrieved from https://deepai.org/machine-learning-glossary-and-terms/accuracy

[28] (2020, 9). Retrieved from https://www.kidneyfund.org/kidney-disease/kidney-failure/treatment-of-kidney-failure/kidney-transplant/

[29] 2018, 5 15). Retrieved from https://www.kidneyfund.org/kidney-disease/chronic-kidney-disease-ckd/?fbclid=IwAR1NYen0No8sUVhBkQe6XoxSiS9SyBmBysaUINyjnhL1DK-n-pM2Z6iR8QY

[30] (2021, 5 3). Retrieved from https://www.thedailystar.net/city/news/18m-kidney-patients-bangladesh-every-year1703665#:~:text=According%20to%20Kidney%20Foundation%2C%20some,140%20nephrologist

PLAGIARISM REPORT

13%	10%	8%	11%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	2%
2	Submitted to Daffodil International University Student Paper	1%
3	Arkadip Ray, Avijit Kumar Chaudhuri. "Smart healthcare disease diagnosis and patient management: Innovation, improvement and skill development", Machine Learning with Applications, 2021 Publication	1%
4	Submitted to Liverpool John Moores University Student Paper	1%
5	Submitted to Universidad Del Magdalena Student Paper	<1%
6	Submitted to Wright State University Student Paper	<1%
7	pisrt.org Internet Source	<1%