ANALYSIS AND PREDICTION OF CHRONIC KIDNEY DISEASE

Thilina Duminda Nakkawita

(168249N)

Degree of Master of Science

Department of Computer Science and Engineering

University of Moratuwa
Sri Lanka

April 2020

ANALYSIS AND PREDICTION OF CHRONIC KIDNEY DISEASE

Thilina Duminda Nakkawita

(168249N)

Thesis/Dissertation submitted in partial fulfilment of the requirements for the degree
Master of Science

Department of Computer Science and Engineering

University of Moratuwa
Sri Lanka

April 2020

DECLARATION

I declare that the thesis is purely based on my own work, and it does not include course materials from any other university or college diploma without acknowledgement, and as per my knowledge this does not include Materials that are published or written by other persons, unless otherwise not indicated in the text.

In addition, I give the non-exclusive right to the University of Moratuwa to replicate or distribute my article in whole or in part in print, electronic or other media. I reserve the right to use this content in whole or part in future works.

……………...   2020-05-30 ……………………………

Signature   Date

I confirm that, as per my knowledge, the above statement of the candidate is correct and that the project report can be used to evaluate the MSc research project.

……………...   ……………………………

Signature of the supervisor   Date

ABSTRACT

In Sri Lanka, chronic kidney disease has become a significant public health problem over the past two decades. Since there are few signs or symptoms in the early stages, it is difficult to identify whether people have the CKD disease, because. Due to this reason, they do not get treatments. If the disease is detected at an early stage, CKD can be cured. Sri Lanka currently lacks comprehensive and systematic surveillance procedures to identify and monitor all aspects of CKD in the general population.

The disease can be identified in the early stages if there is a proper dataset to analyze. Based on the data a predictive model can be developed and this will help doctors diagnose if a patient has early-stage CKD. CKD can prevent if this detects early and provide necessary treatments.

As part of my research; I have developed a computerized system to capture and track aspects of CKD in Sri Lanka, including a predictive model to detect CKD in its early stages. The predictive model was developed using different types of data mining classification algorithms. In the healthcare sector, data mining is mainly used for disease detection. Broad data mining techniques exist for predicting diseases, such as classification, clustering, association rules, summaries, and regression. Additionally, the tool was developed to perform several analyses based on the collected data.

ACKNOWLEDGEMENTS

I would like to take this opportunity to share my sincere gratitude for the people who have been instrumental in the completion of this dissertation.

To Dr. Shehan Perera, I cannot say Thank You enough for all of your support, guidance, and encouragement: you have motivated and enlightened me more than you know throughout this project. It would not have been possible without your help.

To all of the other individuals who have helped and contributed on this project, I Thank You for your incredible assistance and involvement on this dissertation.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLE

LIST OF ABBREVIATIONS

AGA	**A**ssistant **G**overnment **A**gents
API	**A**pplication **P**rogramming **I**nterface
CKD	**C**hronic **K**idney **D**isease
CKDu	**C**hronic **K**idney **D**isease **U**nknown
CRF	**C**hronic **R**enal **F**ailure
eGFR	**E**stimated **G**lomerular **F**iltration **R**ate
GFR	**G**lomerular **F**iltration **R**ate
PTH	**P**ara **T**hyroid **H**ormone

1. INTRODUCTION

Chronic kidney disease (CKD) is a gradual loss of kidney function over time. Eventually, patients suffering from CKD will sustain permanent renal failure.

Nowadays, CKD is very common; this can be undiscovered & unknown until the sickness is in critical stage. Usually, people don't recognize that they have kidney disease until their kidney is functioning at 25% of its normal efficiency.

1.1 Chronic Kidney Disease (CKD)

There are five stages of kidney disease that are mainly based on the estimated, or measured Glomerular Filtration Rate (GFR) [1]. GFR is used to determine how well the kidneys are functioning by estimating how much blood passes through the glomeruli per minute. Glomeruli works as a filter in the kidney to filter waste from the blood.

Different levels in CKD [2]

The disease has five major stages. In the initial stages the kidneys are still functioning correctly but in the later stages it will stop functioning.

Stage 1 (GFR more significant than 90 mL/min)

In this stage there is minor damage to the kidney hence it does not show any symptoms. Most of the time if the GFR more significant than 90 mL/min means that the kidneys are healthy and functioning well. However, it's possible that a patient in Stage 1 kidney disease may have a normal eGFR. Even so, there are other noticeable signs of their kidney damage.

Stage 2 (GFR between 60 and 89 mL/min)

Not unlike Stage 1, Stage 2 presents with mild kidney damage, and there are usually no symptoms.

Stage 3A (GFR within 45 and 59 mL/min) & Stage 3B Moderate CKD (GFR within 30 and 44 mL/min)

This means the kidney is damaged to a certain extent and does not work properly.

Stage 4 (GFR between 15 and 29 mL/min)

This means that the kidneys are damaged, and this should consider as very critical as it is one stage before the kidney dies.

Stage 5 (GFR less than 15 mL/min)

Stage 5 is known as “end-stage kidney disease”. A patient with stage 5 CKD will likely not survive unless they receive artificial kidney filtering (known as "dialysis") or they undergo a kidney transplant.

1.2 Chronic Kidney Disease (CKD) in Sri Lanka

In Sri Lanka CKD becomes a significant health issue in the past two decades. World Health Organization (WHO) reports approximately 15% of the population in Anuradhapura, Badulla and Polonnaruwa Districts within the ages of 15-70 are affected by CKDu.

According to the sources, in early 1990's the outbreak of CKD in the North Central Province in Sri Lanka was firstly recognized [4].

Agriculture is the main income source in this area. However, because the disease is endemic the causes are believed to be environmentally induced.

People who are involved in paddy farming are profoundly affected by CKD. Drinking well water and being under treatment for hypertension can be identified as the significant predictors of kidney disease.

1.3 Risk Factors of CKDu

In the past few years, much research has been done to determine the prevalence of CKD in the north-central provinces.

An article written by K. Wanigasuriya [5] has mentioned several risk factors for CKDu. The case controlled study conducted in Anuradhapura helped to identify the subsequent main risk factors of CKD: being a farmer, using pesticides, drinking ground water reception and within the field, having member in the family with renal dysfunction, having taken Ayurveda treatment within the past, and a history of snakebites.

Additionally, the following items were listed as moderate risk factors:, patients older than 60, exposure to nephrotoxic drugs, exposure to poison, alcohol consumption, working-age patients, treatment costs, general health and diet, environmental factors, population at lower elevation, use of water and fish products and occupation.

1.4 Prediction by Data Mining

Presently, with the recent improvements in technology over the past 2 decades computers and databases are able to collect vast amounts of data. Data mining [6] is the system of analyzing large sets of facts to generate new information. Data mining strategies can be classified as unsupervised and supervised gaining knowledge of. The unsupervised getting to know technique is not guided by using the variable and does no longer create a hypothesis before analysis.

A model will be constructed based on the results. In Data Science, we can use unsupervised technique called as clustering analysis to gain some valuable insights from our data when we apply a clustering algorithm.

The supervised learning technique requires the development of a model that's utilized in the preliminary analysis. Supervised learning techniques utilized in medical and clinical research are classification, regression and association rules.

The main objective of this research is to predict kidney diseases by using data mining techniques. To obtain essential information from medical databases, data mining techniques have been very useful. By combining automatic learning with statistical analysis, beneficial information can be derived from medical datasets.

Machine learning methods that coordinate the various statistical analyses and databases help us extract hidden models and relationships from both massive and multiple variable datasets. To ensure the accuracy of the selected classifier, the available test phases are checked. Early-stage kidney disorders can be identified by applying the different data mining techniques for the desired classification methods.

2. RESEARCH PROBLEM & OBJECTIVES

Despite the recognized importance of (CKD) Sri Lanka currently lacks a comprehensive and systematic surveillance program to capture and track all aspects of CKD.

World Health Organization reports millions of people worldwide suffer from severe kidney problems, and the number is increasing annually. Therefore, a procedure for detecting early-stage kidney disease is a necessity. Data mining is becoming more and more popular today in healthcare and has been very useful in obtaining essential information from medical databases. The mostly used data mining technique is Classification. One of the goals of this research is to use a classification technique to accurately predict CKD within potential patients

However, in Sri Lanka a model to analyze sets of patient data to predict CKD does not exist. The only way to identify that a patient has CKD is by conducting several medical tests. This increases the medical expenses that are covered by the Sri Lankan government. After taking the tests, patients are categorized as either “CKD” or “non-CKD”. Unfortunately, even if a patient does not show signs of CKD, the cost of these test falls to the patient or the government. These extra costs can be minimized by implementing a system to predict which patients are required to perform the further tests. An accurate predictive model will limit the total number of patients that take many tests, cutting down on overall medical costs. Since there is no electronic system currently in place to gather patient information, building a predictive model was a challenge.

The main goal of this system was the development, validation, and analysis of predictive models using demographic, clinical, and laboratory data. The construction of this model was divided into three categories: - building a centralized system to collect the patient data; performing several analyses and, making a predictive data model based on the collected data.

2.1 Centralized System to Collect Data

Currently there are no centralized systems to collect patient data. As a part of the research, a centralized system (Figure 2.1) was developed to accomplish this. Data was collected through both web and mobile-based applications.

Figure 2.1: Centralized System

Several researchers are willing to research with a centralized system. The problem, however, is that there are no proper datasets for them to use. But the data that collects from this system, can be used by these researchers. The system provides an API through which the researchers can access the data. The API does not return sensitive data (e.g., name of the patient), but it does provide data that is necessary for additional analysis and research.

2.2 Analysis of Patient Data

The dataset collected through the centralized system can be linked with the Microsoft Power BI tool which provides interactive visualizations with self-service business intelligence capabilities.

This tool is integrated directly with the centralized system, allowing users to generate different types of reports and perform their own analyses.

Listed below are some examples of analyses that can be performed with the proper dataset.

2.2.1. Analysis of patient data according to AGA Divisions

The end user can get a better idea of how the disease has speared in each area. They can analyze data within the period to get a better understanding about which region will have the highest threat of Kidney disease in the next few years (Figure 2.2). The graph below shows the distribution in each AGA division.

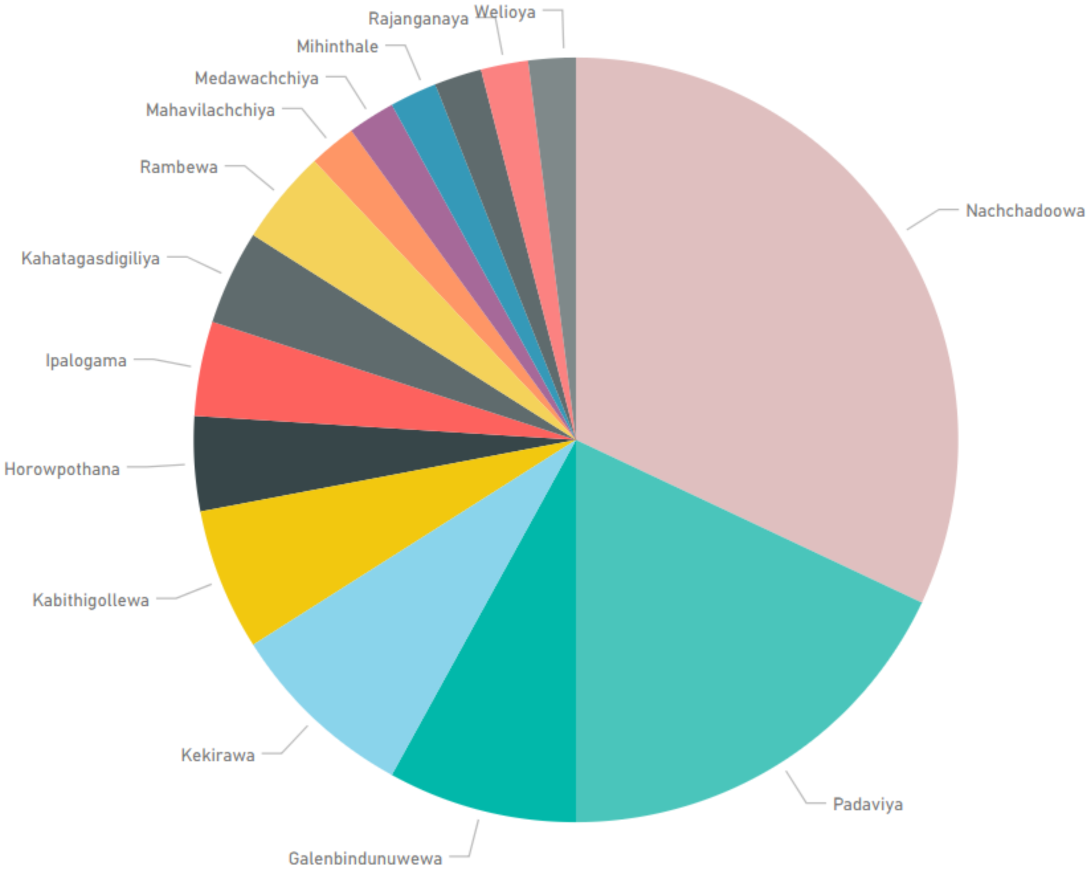


Figure 2.2: Patient data according to AGA Divisions

2.2.2. Analysis of patient data according to the age of patients

With the collected data, the system can generate a graph to show the distribution of CKD by age.

2.2.3. Analysis of the impact of associated medical conditions of the patient

If a patient has Alcoholic liver disease, this can be used to analyze how many kidney patients have Alcoholic liver disease and perform some prediction (Figure 2.3).

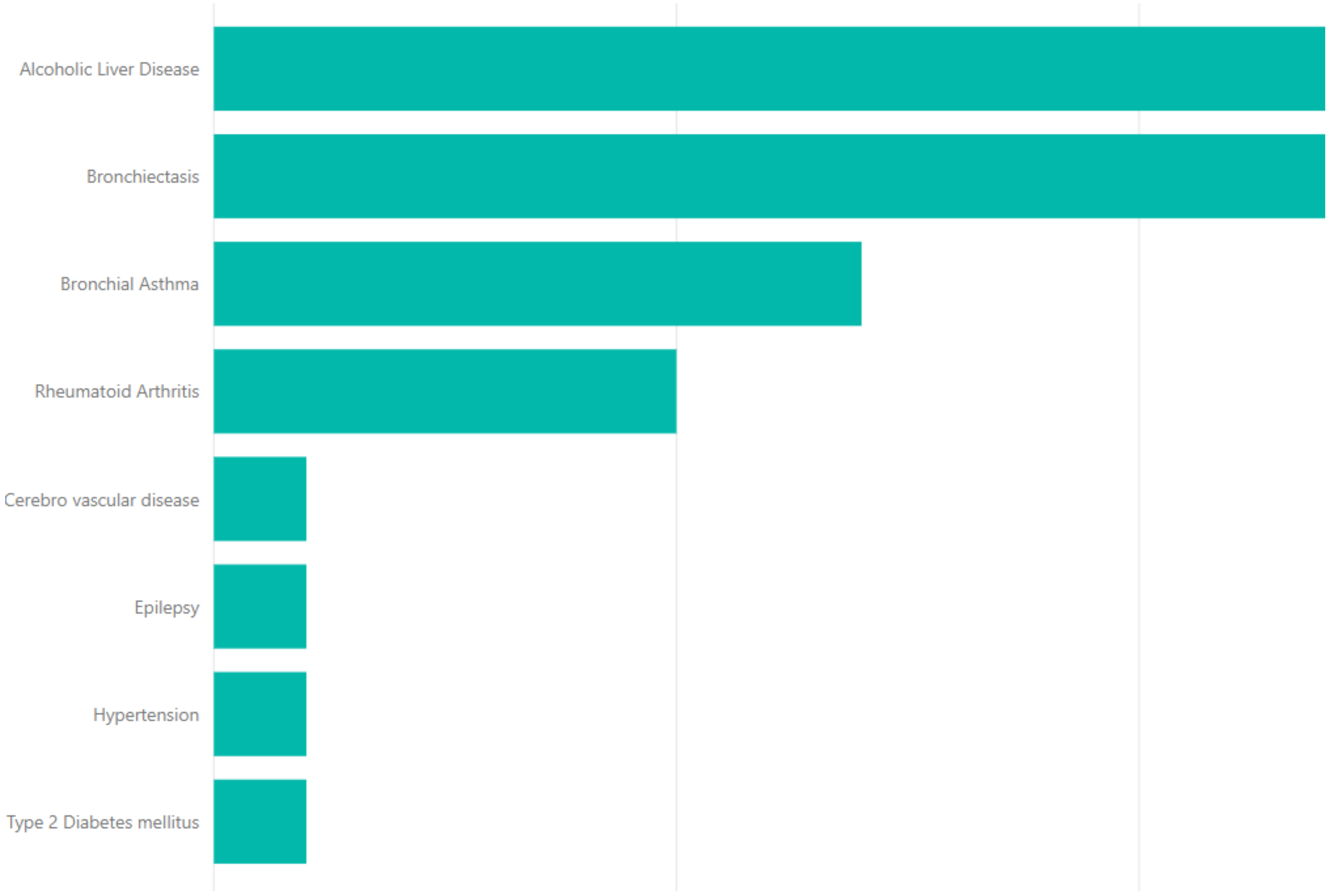


Figure 2.3: Impact of associate medical conditions of the patient

2.2.4. Analyzing patient data according to the lab data

Important information called “Laboratory Data”, is recorded in the centralized system and is useful into performing data analysis and predictions for CKDu patients Figure (2.4).

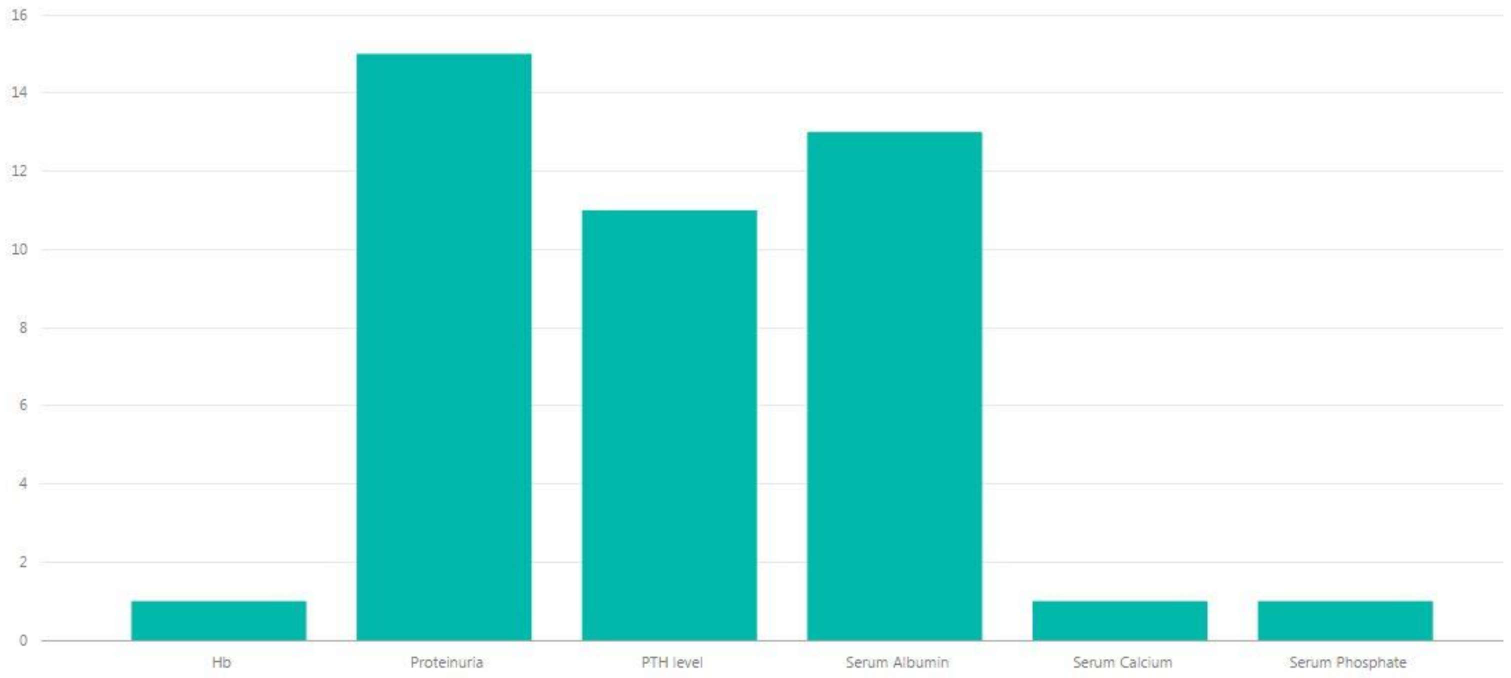


Figure 2.4: Analyzing patient data according to the lab data

Additionally, the following analysis can be done with the provided dataset;

- Analyze patient data according to "Proteinuria"

- Analyze patient data according to "Serum albumin".

- Analyze patient data according to "PTH level"

2.3 Building a Predictive Model

Unfortunately, the costs associated with CKD treatments, examinations, and screening tests are very high. And while the government bears those costs on behalf of the patient, there are situations where patients are found to be free of CKD after additional testing. The purpose of this centralized system and data analysis tool is to identify CKD in patients at a very early stage so that patients that do not have the disease are not required to undergo additional testing and treatment. Alternately, patients that do show early signs of CKD will be notified as early as possible so that they can receive the appropriate treatment plan. This predictive model will help those patients with CKD by identifying it much earlier than normal methods and it will cut down on the unnecessary medical costs associated with screening individuals that do not have CKD. (Figure 3.5)

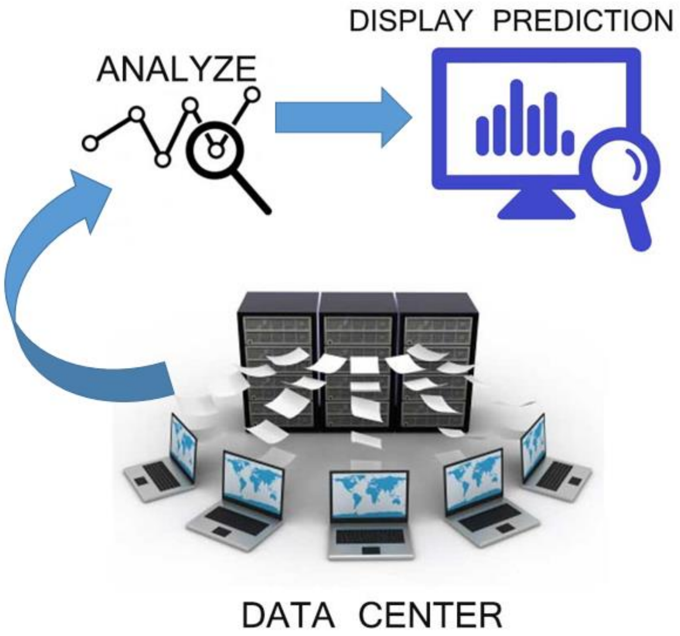


Figure 2.5: Prediction with the dataset

2.4 Summary

Currently, in Sri Lanka, there is no proper system to capture patient data, analyze and perform CKD prediction via data mining. As a solution to this problem, a system was built to capture patient data; Power BI was integrated to interpret and perform predictions based on the data collected.

3. LITERATURE REVIEW

Data mining and analysis are widely used to extract useful information from raw data. Today, data mining has become an essential area of healthcare for detecting unknown information in medical data sets and using analytics to predict disease. In this chapter, we discuss how the authors used predictive analytics model to predict disease based on its cause and develop its progression model for chronic kidney disease (CKD).

3.1 About the CKD

During the past few years, the widespread presence of Chronic Kidney Disease was present in some geographic areas of Sri Lanka. In the research conducted by Mr Senaka Rajapakse, has explained the CKDu, Epidemiology of CKDu, clinical characteristics & histopathological findings in CKDu, including the Causative factors of CKDu.

In the article published by the National Kidney Foundation, has explained the symptoms and the causes of CKD. They have mentioned that diabetic & high blood pressure is responsible for up to 2/3 of the cases. Glomerulonephritis, Inherited diseases, such as polycystic kidney disease, & Repeated urinary infections etc are other conditions that affects the kidney.

Further, they have listed the signs which we can notice if we have CKD. Feeling more tired, lacking energy, lack of concentration, loss of appetite, need to urinate more frequently, especially at night, and they have few of the symptoms listed in this article.

3.2 Analytics

3.2.1 Descriptive Analysis

An analytic study has done by Hubert Kouame Yao [7]. This study was carried out in the Internal Medicine Department of the University Hospital of Treichville. This study describe the current profile of CKD in our working conditions. This is a descriptive retrospective study of patients admitted for CKD during the period from January 2010 to December 2014 in the Internal Medicine Department of the university hospital of Treichville in Abidjan.

During the study, they collected 252 cases of CKD out of 3573 patients recorded, yielding a prevalence of 7.05%. Out of the CKD patients studied, 67.1% were known to have hypertension, 7.9% had diabetes, and 8.7% were HIV positive. 5.6% of the cases are hypertension and diabetes. On clinical examination, hypertension was observed in 211 cases (83.7%). It was Grade-1 in 14.7%, Grade-2 in 16.3% of cases and Grade-3 in 52.8%. The aetiology was dominated by hypertension, seen in 59.9% of cases, chronic glomerulonephritis in 25% of cases, HIV infection in 9.1% of cases, and diabetes in 4.8% of cases.

As a conclusion, they have mentioned the risk factors like hypertension, HIV infection, and diabetes. Their results showed the importance of early diagnosis and nephrology monitoring of the disease to slow down the progression of CKD.

Lenildo de Moura et al. a study [8] was conducted in which 60,202 people over the age of 18 were evaluated who independently reported a medical diagnosis of chronic renal failure or kidney disease.

The prevalence of CKD was 1.4% according to the result. Similarity between the genders: 1.4% for men and 1.5% for women. Southern Brazil presents the indicator most frequently. In people diagnosed medically with end-stage renal disease, the prevalence of dialysis is 7.4%, which is higher in men.

The results revealed characteristics of CKD. Data is authorized to plan public policies aimed at preventing disease and promoting health.

Knight T, Schaefer et al., have done a research [9]. The methodology of this study is a retrospective, pair-wise cohort study that examined the use of adult medical resources (MRU) and costs in the American database of claims for reimbursement from private payers with diagnosis for ADPKD.

At the end of the study, they noted that compared with the general population, patients with ADPKD with normal kidney function were associated with a huge economic burden on the medical system. Any delayed treatment in the advanced stages of CKD can offset potential medical costs.

According to Winnipeg researchers, CBC News [10] can easily predict kidney failure with a simple equation.

According to the Dr. Navdeep Tangri’s study will apply its equation more broadly, claiming that it can reliably predict patients' renal failure, and regardless of location, gender, his age and general health. [11]

Tangri, is a professor at the University of Manitoba's School of Medicine, first developed a reliable and straightforward method in 2011 to predict the likelihood of future development in CKD patients.

De Nicola L and other members of the SIN-TABLE CKD research team studied the prognosis of patients with CKD receiving renal ambulatory care in Italy. [12]

The prognosis for patients with chronic renal failure (CKD) who are not on dialysis under routine renal monitoring is rarely done. From 2003 to death or until June 2010, they followed 1,248 patients with 3 to 5 stages of chronic renal failure in 25 ambulatory renal failure clinics in Italy and received care for more than a year kidney disease. The method of cumulative risk of ESRD or death before end-stage renal disease was estimated using the competitive risk method.

The results of this study are an estimated ratio of ESRD and deaths (per 100 patient-years) From stage 3 to stage 5, the risk of ESRD and death gradually increases. In the 4th and 5th stages of CKD, ESRD occurs more frequently than death, while in the 3rd stage of CKD, the risk of RDD is opposite.

According to the article, in patients on continuous treatment at the Italian Kidney Clinic, the incidence of stages 4 and 5 of CKD is higher than death. However, the opposite is true in phase 3.

The target blood pressure (CRI) level of the CRI of the Italian nephrology research group De Nicola L et al. [13].

The reason for conducting this study was to determine whether age changes the prognosis of CKD patient in the treatment of renal disease and to prospectively follow patients with Crohn's disease who have received kidney disease for more than one year at the clinic. The incidence of ESRD was estimated by a competitive risk approach and the interaction between ages, which was defined as initiation of dialysis or transplantation or death without ESRD, and median risk factors in the Cox model The number followed for 62.4 months.

3.2.2 Predictive Analysis

3.2.2.1 Using SVM

Dr. Vijayarani and MS Dhayanand have examined six different attributes of kidney disease in GFR, namely, the glomerular filtration rate is a measurement attribute for predicting kidney disease [14]. They implemented and compared two naive Bayesian classification techniques and SVM (Support Vector Machine). Their experimental results show that SVM is more precise than Naive Bayes.

3.2.2.2 Using Random Forest Classification techniques

Dr. S. Ramya and N. Radha have developed a system to predict kidney failure by applying four classification techniques to test data from patient medical reports [15]. They have 1000 records with 15 attributes. They also compared these four techniques, such as backpropagating neural networks, radial basic functions, and random forests. Their results show that radial basic function (RBF) has better accuracy in predicting CKD.

3.2.2.3 Using Naïve Bayes classification techniques

Narendra Kamila & Lambodar Jena used various classification techniques such as Naive Bayes, Perceptron multilayer, Support Vector Machine, J48, Conjunctival Rules and Decision Tables [16] to analyze all the data on chronic kidney disease. They use Weka software. They use 25 different attributes for classification. Their research shows that for the prediction of chronic renal failure, the multilayer perceptron is relatively more precise than other technologies (99.75%).

3.2.2.4 Using Cox proportional hazards

According to the research done by Tangri [17] the main goal of this study was to develop and validate predictive models of CKD progress. They used three datasets (demographic, clinical and laboratory data) to build and validate predictive models from two independent cohorts of stage three to five CKD patients in Canada. The model used the Cox proportional hazards regression method, and C statistics and a complete discriminatory improvement, a correction map and the applicability of Akaike's information criteria, and an improvement in classification. of net weight (NRI) were used.

The study concluded that the use of routine laboratory test models can accurately predict the progression of kidney failure in 3-5 patients with CKD.

Debora C. Cerqueira et al. Conducted a study in an interdisciplinary pre-dialysis program [18].

The purpose of this study was to build a model that predicts ESRD in children and adolescents with CKD participating in an interdisciplinary pre-dialysis management program (steps 2-4).

In their study, they selected 147 CKD patients (stages 2-4) who were treated with CKD between 1990 and 2008. The main result is to enter phase 5 of CKD. Cox's proportional hazard model was used to build a prediction model and evaluated by statistics.

The study concluded that a predictive model of CKD progression may help in early identification of subgroups of patients at high risk of developing accelerated renal failure.

Tangri N et al. has researched "A Dynamic Predictive Model for Progression of CKD [19]”.

This study includes the development of predictive models based on population, clinical and chronological laboratory data (of a group of stage 3 to 5 patients with chronic renal failure).

They studied 3,004 patients who attended the OPD clinic in Sunnybrook, Toronto from April 1, 2001 to December 31, 2009 (they performed 344 renal failure events, followed up for an average of 3 years, and performed an average of 5 times Clinical visits). Canadian hospital.

The results of this study were treated with renal failure, defined as the initial process of dialysis treatment or kidney transplantation.

In the static model, the eight predictors are included. The latest measurement models available include age and the last five variables, which are predictors that change over time. The researchers used Cox's proportional hazard model to calculate the time for kidney failure and compared the discriminating power, calibration, fit of the model, and classification of the net weight of the model.

The limitation of the study was that the data set they used came from a single renal failure clinic. Nor can they include time-dependent changes in proteinuria.

Mendonça AC, a predictive model of idiopathic nephrotic syndrome with progressive chronic renal failure has been studied [20].

They concluded that a predictive model of CKD can help early identify a subset of NSI patients at high risk for renal dysfunction.

The study [21] done by Adler Perotte shows that the adoption of electronic medical records continues to increase, clinical documents as well as laboratory and demographic data can be integrated into risk prediction models.

The study cohort consisted of 2,908 primary care patients who had visited at least 3 times before January 1, 2013 and had progressed to stage III of Crohn's disease in their history.

3.2.2.5 Using MATLAB

Jamshid Norouzi et al. Have studied "the use of an integrated fuzzy intelligent expert system to predict the progression of renal failure in chronic renal disease" [22].

A comparison between the forecast and actual data shows that the ANFIS model can accurately estimate the changes in the GFR (mean absolute error less than 5%) for all subsequent cycles.

Despite the high degree of uncertainty in the human body and the dynamic nature of the CKD process, the results of the research remain. Their model can accurately predict long-term GFR changes.

3.2.2.6 Using the beta coefficients

Taylor GW wrote in a Fisher article in Fisher MA a report on predictive models of chronic kidney disease, including periodontal disease [23], and they mentioned that among 7 million Americans with moderate to chronic kidney disease severe And almost 75% have not been diagnosed.

The results of the research mentioned indicate that the estimated probability of chronic renal failure in older whites than in older Hispanics is different from that of individuals with virtually no probability (0%), and there are no 12 factors extremely high risk (98%) of individual smokers, diabetes is higher than 10 years megaloproteinuria, high density lip-o-protein, low in-come and hospitalization in the last year.

As a result, they summarized the research paper because U.S.-based population studies show the importance of considering several risk factors, including periodontal conditions, as it improves the ability to identify chronic kidney disease. in high-risk populations and ultimately mitigates them.

.

3.3 Summary

In this chapter, we discussed the various researches and studies conducted for CKD descriptive and predictive analysis.

In the researches, different approaches and techniques have used. But, some of the researchers have used similar attributes for their predictive models such as age, gender, eGFR, proteinuria, hematuria, and hypertension. They have used various classification techniques such as Naïve Bayes, Cox proportional hazards regression methods, Multilayer Perceptron, Support Vector Machine, J48, and Decision Table to validate their model.

Out of the above techniques, most of the time Random Forest Classification techniques, Naïve Bayes classification techniques and Cox proportional hazards regression methods have shown high accuracy in the results.

4. RESEARCH MODEL / METHODOLOGY

The primary objective of the research is to build a predictive model. As the first step, CKD patient data needs to be collected. To achieve that, I developed a Web/Mobile based system that was able to perform critical analyses on the data collected.

4.1 Build a System to Collect Data

Currently, hospitals maintain all of their patient records in files. This information, however, can be entered into the centralized system that was developed. Since the system is mobile-friendly, doctors can quickly enter patient data while he/she is consulting with patients.

Patient records are maintained for each stage of the CKD. The system included the data of all patients (CKD & non-CKD) which will be used when performing the analysis and prediction.

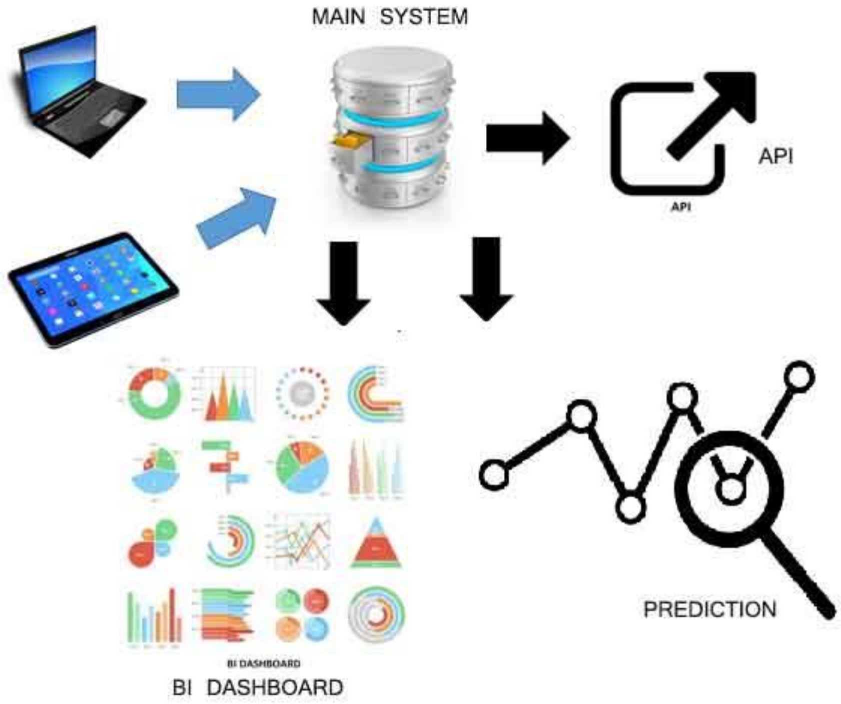
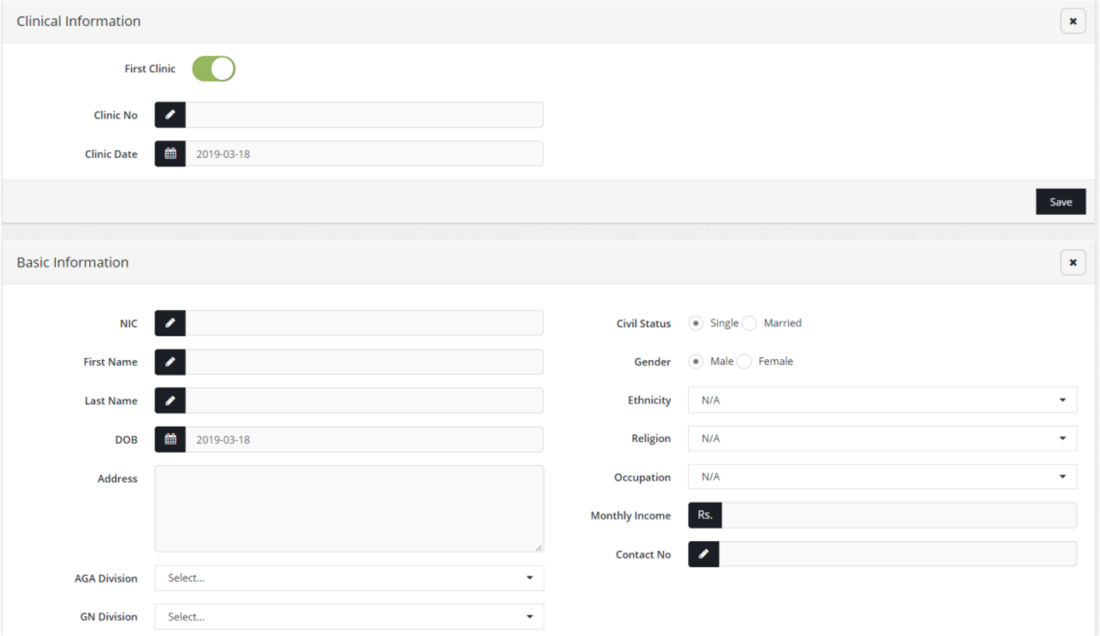


Figure 4.1: System

The information is captured through a form developed in the system. The patient form consists of several sections. The figures below (4.2 & 4.3) show the information that can be collected for each patient.



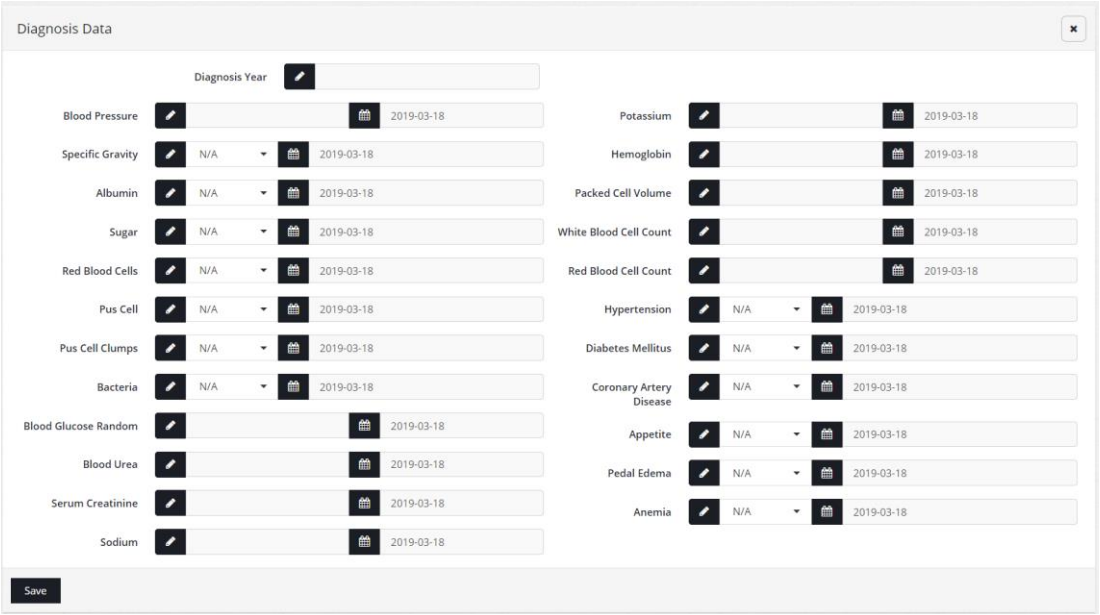
Clinical Information

First Clinic	●
Clinic No	
Clinic Date	2019-03-18

Basic Information

NIC		Civil Status	● Single ○ Married
First Name		Gender	● Male ○ Female
Last Name		Ethnicity	N/A
DOB	2019-03-18	Religion	N/A
Address		Occupation	N/A
		Monthly Income	Rs.
		Contact No	
AGA Division	Select...		
GN Division	Select...		

Figure 4.2: Patient entry form – Basic Information



Diagnosis Data

Diagnosis Year

Blood Pressure		2019-03-18	Potassium		2019-03-18
Specific Gravity	N/A	2019-03-18	Hemoglobin		2019-03-18
Albumin	N/A	2019-03-18	Packed Cell Volume		2019-03-18
Sugar	N/A	2019-03-18	White Blood Cell Count		2019-03-18
Red Blood Cells	N/A	2019-03-18	Red Blood Cell Count		2019-03-18
Pus Cell	N/A	2019-03-18	Hypertension	N/A	2019-03-18
Pus Cell Clumps	N/A	2019-03-18	Diabetes Mellitus	N/A	2019-03-18
Bacteria	N/A	2019-03-18	Coronary Artery Disease	N/A	2019-03-18
Blood Glucose Random		2019-03-18	Appetite	N/A	2019-03-18
Blood Urea		2019-03-18	Pedal Edema	N/A	2019-03-18
Serum Creatinine		2019-03-18	Anemia	N/A	2019-03-18
Sodium		2019-03-18			

Figure 4.3: Patient entry form - Diagnosis data entry form

4.2 Dataset

The collected dataset contained 500 total patients. The dataset consisted of both CKD patients (in all stages) & Non-CKD patients. Out of the 500 patients, 248 were in the early stages of CKD and 152 were non-CKD patients.

4.3 Analyze the Collected Data

Based on the data collected, critical analysis can be done, and predictions can be made. For the analysis, the system uses an external BI tool like PowerBI and the custom BI tool in the system.

There are several analysis reports that can be generated from the system. From the reports, we can gather information about how the CKDu diseases spreads within the defined geographical area. The reports indicate which area has the highest concentration of CKD patients. They also display the CKD distribution filtered by several factors including, but not limited to age, occupation, residence, and dietary habits.

4.3.1. Analyze patient data according to the patient’s AGA division

For each patient, the AGA division of the patient is recorded. With this data, we can perform an analysis to determine which AGA division has the highest number of CKD patients.

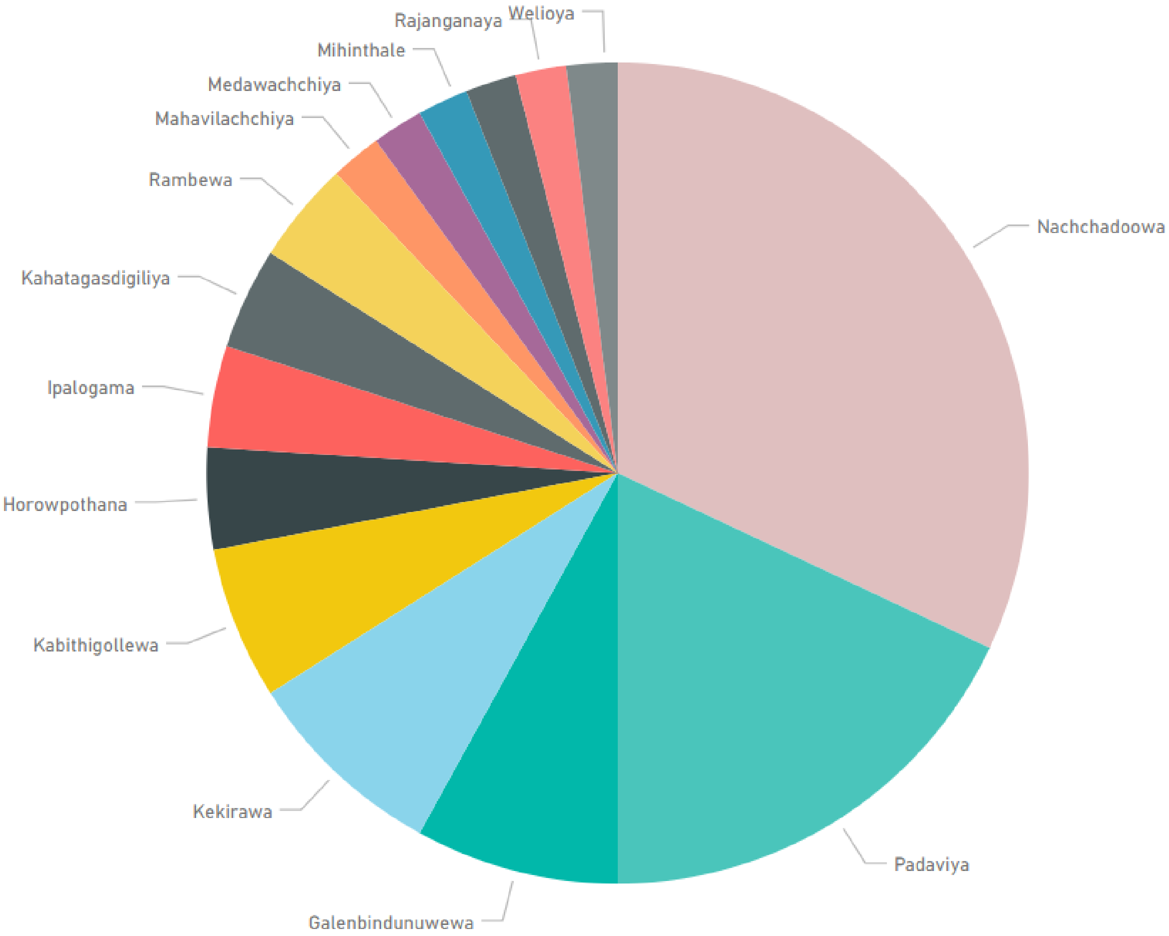


Figure 4.4: Analyze patient data according to AGA

4.3.2. Analyze patient data according to the medical condition of the patient.

Each patient was recorded with his/her medical conditions. With this data, we can analyze which medical conditions have the highest number of CKD patients & observe the association between these medical conditions and with CKD.

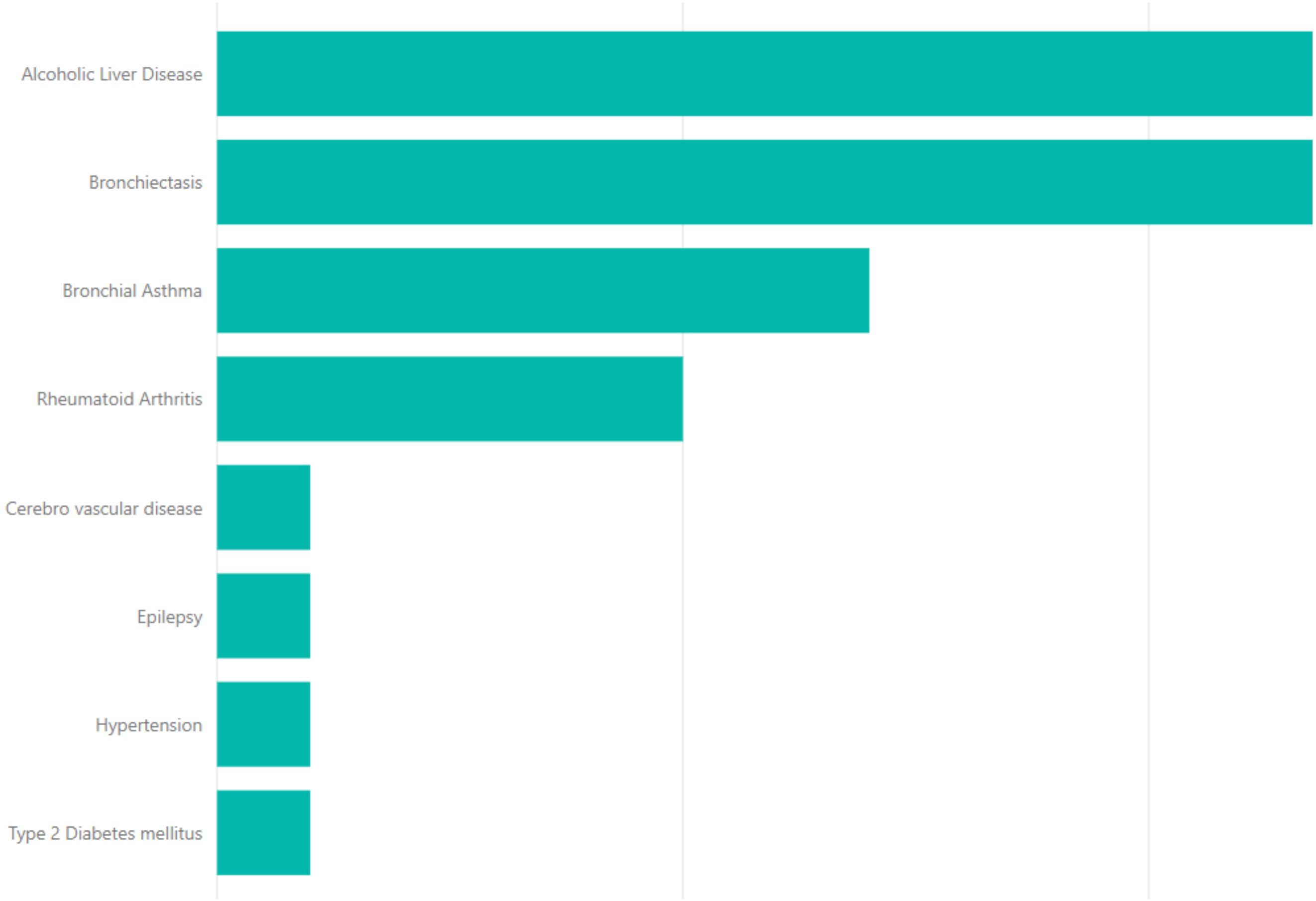


Figure 4.5: Analyze patient data according to Medical Condition

4.3.3. Analyze patient data according to "Proteinuria."

Proteinuria defines the availability of abnormal protein level in the urine which indicates a damage to the kidneys. We can determine the association between Proteinuria & CKD by analyzing the data collected.

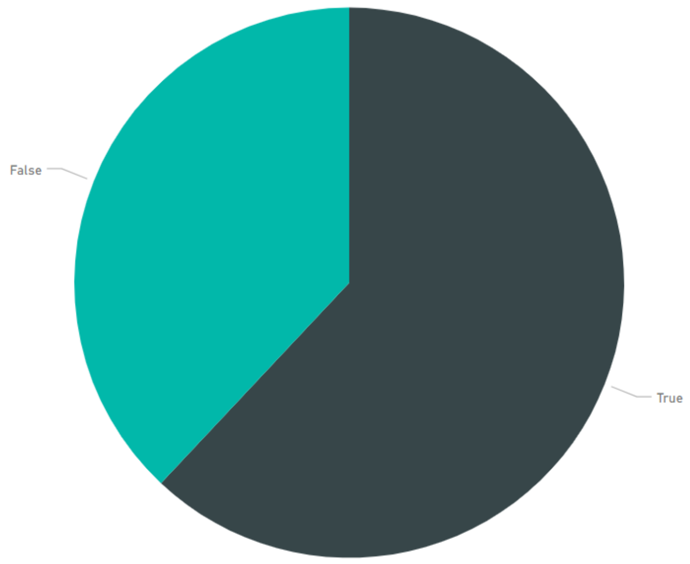


Figure 4.6: Analyze patient data according to Proteinuria

4.3.4. Analyze patient data according to "Serum albumin."

Serum albumin is a type of globular protein which can be found in vertebrate blood. We can determine the association between the Serum albumin & CKD by analyzing the data collected.

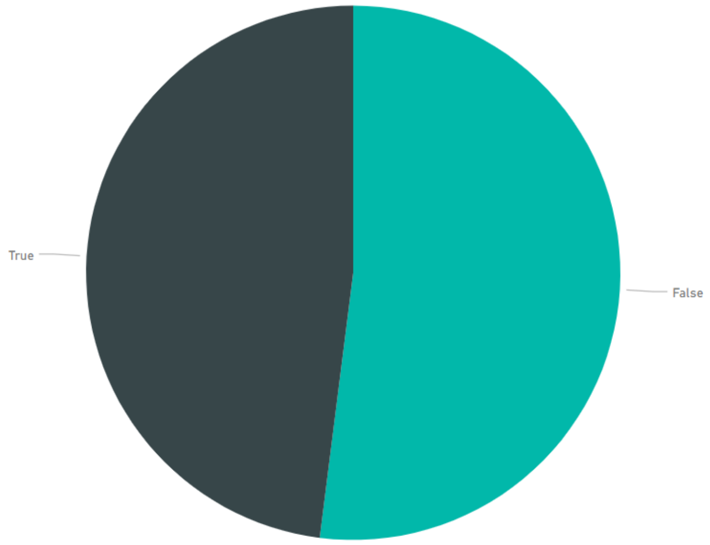


Figure 4.7: Analyze patient data according to Serum albumin

4.3.5. Analyze patient data according to "PTH level."

The PTH level is a measurement of the amount of parathyroid hormone in the blood. We can determine the association between the PTH Level & CKD by analyzing the data collected.

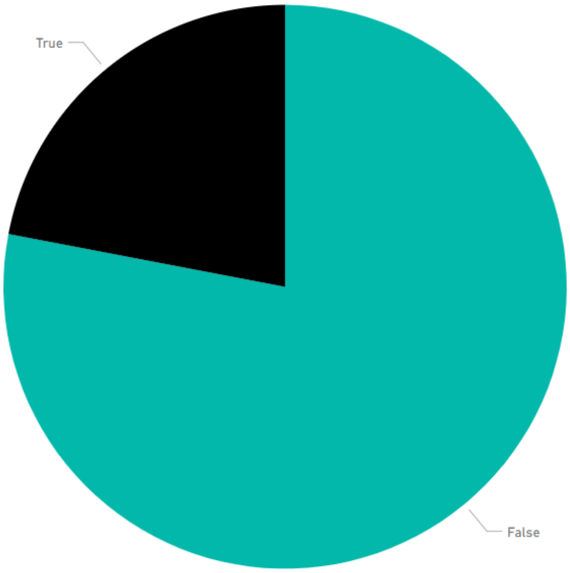


Figure 4.8: Analyze patient data according to PTH Level

4.3.6. Analyze patient data according to the age range.

From this analysis, we can determine which age range is the most susceptible to having CKD.

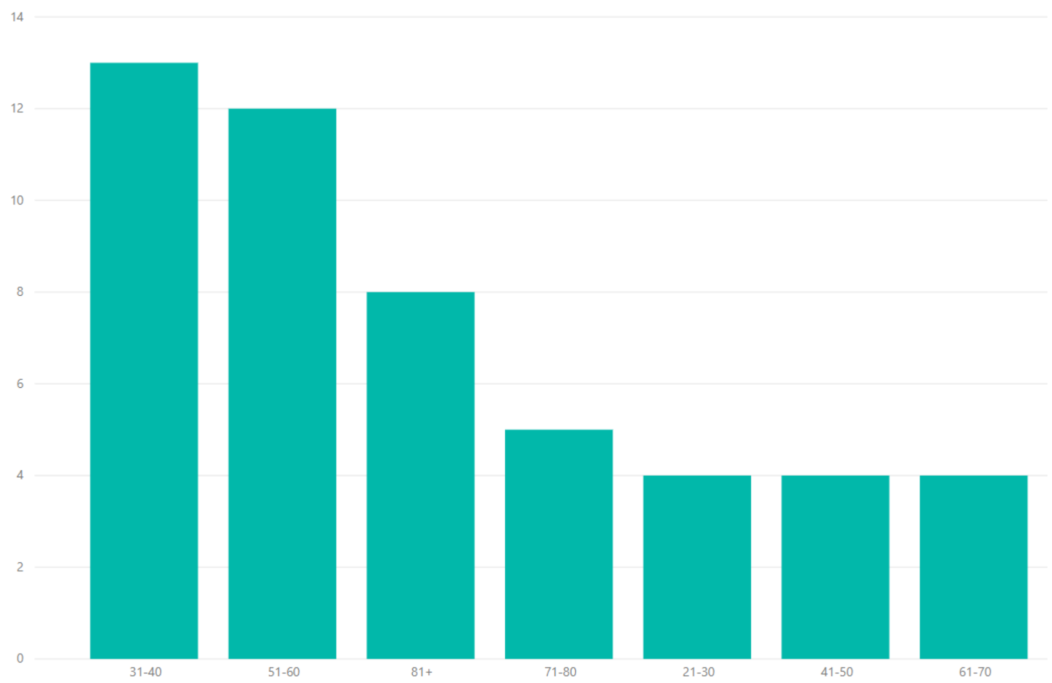


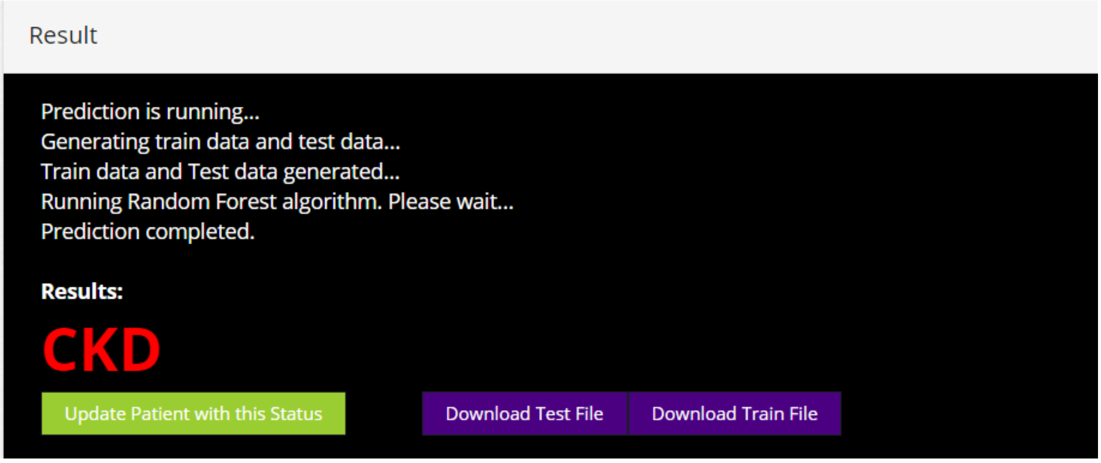
Figure 4.9: Analyze patient data according to the age range

4.4 Prediction Based on Analysis

One of the main objectives of this project was the development of a predictive model that uses machine learning to accurately predict which patients are likely to develop CKD.

In the developed system there are both CKDu and Non-CKDu patient datasets. To help identify whether a person has CKD or not, a WEKA tool that is integrated with the system can be used to analyze the dataset. Once we enter the patient medical information the system predicts the percentage/likelihood of the patient having CKD.

The dataset also contains patient information for several stages of CKD. For the predictive model, we only considered only the data present in the early stages of CKD (Stages 1 to 3) since the probability of having CKD is 100% in the last stages (Stages 4 & 5).



Result

```
Prediction is running...
Generating train data and test data...
Train data and Test data generated...
Running Random Forest algorithm. Please wait...
Prediction completed.
```

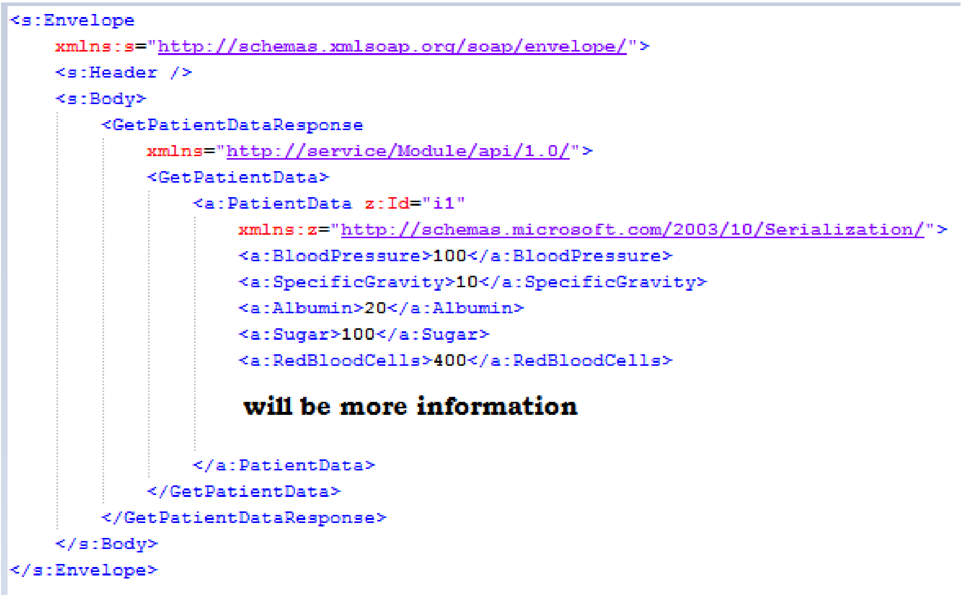
Results:

CKD

Figure 4.10: Prediction tool result

The API built into the system gives outside researchers access to the data that is collected (excluding sensitive patient data) so that they can perform their own research and analysis on the dataset.

During data collection process there are pieces of data collected for "medical information". From the system, that information will be shared via the API so that other researchers can access the data via an XML output.



```
<s:Envelope
    xmlns:s="http://schemas.xmlsoap.org/soap/envelope/">
    <s:Header />
    <s:Body>
        <GetPatientDataResponse
            xmlns="http://service/Module/api/1.0/">
            <GetPatientData>
                <a:PatientData z:Id="i1"
                    xmlns:z="http://schemas.microsoft.com/2003/10/Serialization/">
                    <a:BloodPressure>100</a:BloodPressure>
                    <a:SpecificGravity>10</a:SpecificGravity>
                    <a:Albumin>20</a:Albumin>
                    <a:Sugar>100</a:Sugar>
                    <a:RedBloodCells>400</a:RedBloodCells>

                     will be more information

                </a:PatientData>
            </GetPatientData>
        </GetPatientDataResponse>
    </s:Body>
</s:Envelope>
```

Figure 4.11: XML Response

4.5 Classification Techniques

In the Literature Review, I discussed the various techniques used by different researchers. Based on those results, Random Forest and J48 have shown a higher accuracy when compared to other classifications. The models of those studies were developed based on demographics, clinical and laboratory data collected from the patients. For these reasons I decided to consider Random Forest and J48 classifications find the best algorithm out of the available classification techniques and I chose the clinical and laboratory data of patients as the dataset.

4.6 Best Algorithm

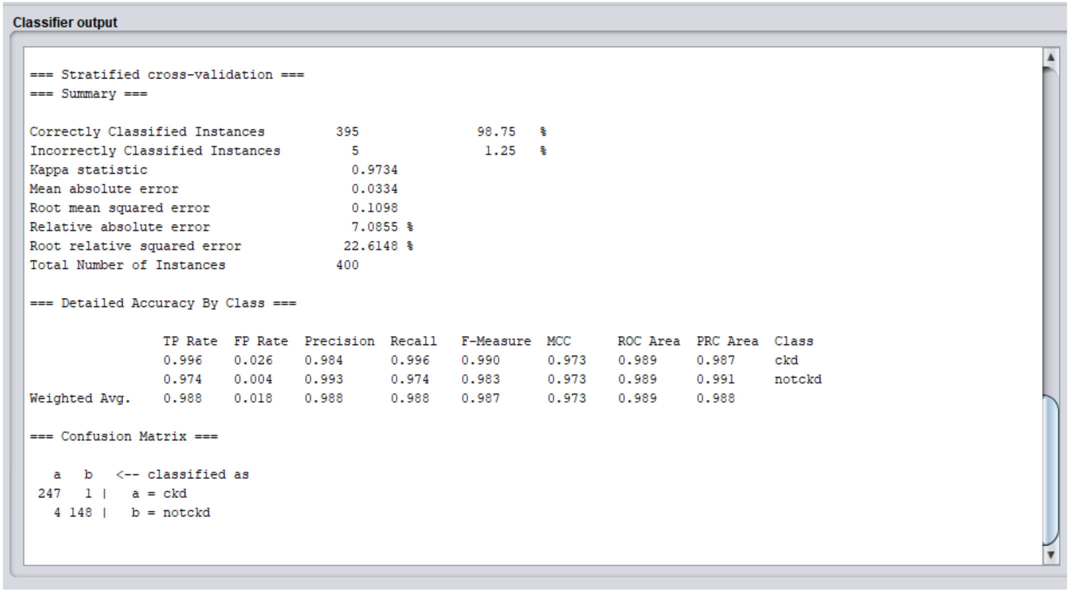
The biggest challenge I faced was to find the most accurate algorithm to perform the CKD prediction analysis. To find the best algorithm from among available data mining algorithms, I compared the results of the trained model using the WEKA tool.

4.6.1 J48 Algorithm

J48 is a classifier that used to build decision trees in WEKA [24]

Results of J48 Algorithm

The figure below (4.12) shows the results using the J48 algorithm with cross-validation – 10 Folds.



```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         395               98.75   %
Incorrectly Classified Instances         5                1.25   %
Kappa statistic                          0.9734
Mean absolute error                      0.0334
Root mean squared error                  0.1098
Relative absolute error                  7.0855 %
Root relative squared error             22.6148 %
Total Number of Instances              400

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.996    0.026    0.984      0.996    0.990      0.973    0.989     0.987     ckd
                 0.974    0.004    0.993      0.974    0.983      0.973    0.989     0.991     notckd
Weighted Avg.    0.988    0.018    0.988      0.988    0.987      0.973    0.989     0.988

=== Confusion Matrix ===

   a   b   <-- classified as
 247   1 |   a = ckd
   4 148 |   b = notckd
```

Figure 4.12: Results using J48 Algorithm

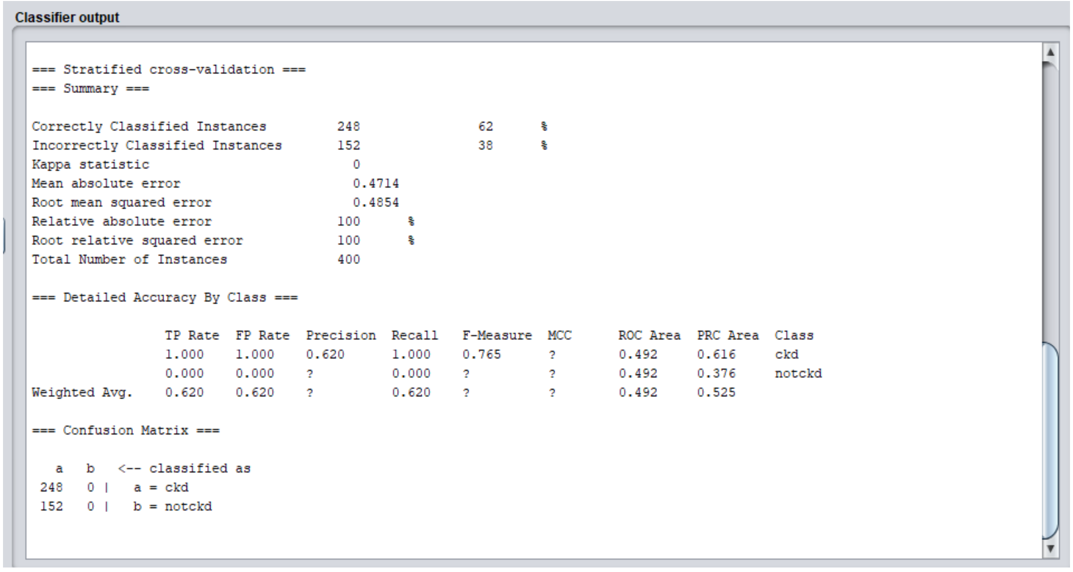
With the J48 algorithm, the Correctly Classified Instances (With Cross-Validation - 10 Folds) is 98.75%.

4.6.2 Zero Algorithm

ZeroR algorithm is based on the target and it ignores all the predicts [25]

Results of ZeroR Algorithm

The figure below (4.13) shows the results using the ZeroR algorithm with cross-validation – 10 folds.



```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         248               62      %
Incorrectly Classified Instances       152               38      %
Kappa statistic                          0
Mean absolute error                      0.4714
Root mean squared error                  0.4854
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances              400

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    1.000    0.620      1.000    0.765      ?        0.492     0.616     ckd
                 0.000    0.000    ?          0.000    ?          ?        0.492     0.376     notckd
Weighted Avg.    0.620    0.620    ?          0.620    ?          ?        0.492     0.525

=== Confusion Matrix ===

   a   b   <-- classified as
 248   0 |   a = ckd
 152   0 |   b = notckd
```

Figure 4.13: Results using ZeroR Algorithm

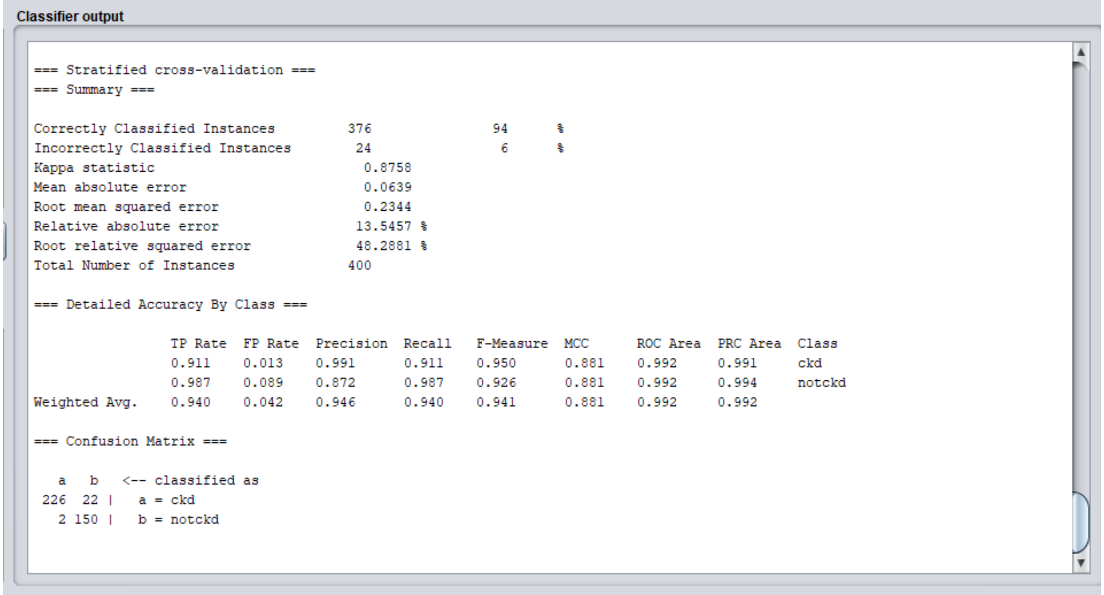
With the ZeroR algorithm, the Correctly Classified Instances (With Cross-Validation - 10 Folds) is 62%.

4.6.3 Naive Bayes Algorithm

This algorithm is based on Baysian theory [26]

Results of Naïve Bayes Algorithm

The figure below (4.14) shows the results using the Naïve Bayes algorithm with cross-validation – 10 folds.



```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         376               94      %
Incorrectly Classified Instances        24                6      %
Kappa statistic                          0.8758
Mean absolute error                      0.0639
Root mean squared error                  0.2344
Relative absolute error                 13.5457 %
Root relative squared error             48.2881 %
Total Number of Instances              400

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.911    0.013    0.991      0.911    0.950      0.881    0.992     0.991     ckd
                 0.987    0.089    0.872      0.987    0.926      0.881    0.992     0.994     notckd
Weighted Avg.    0.940    0.042    0.946      0.940    0.941      0.881    0.992     0.992

=== Confusion Matrix ===

   a   b   <-- classified as
 226  22 |   a = ckd
   2 150 |   b = notckd
```

Figure 4.14: Results using Naïve Bayes Algorithm

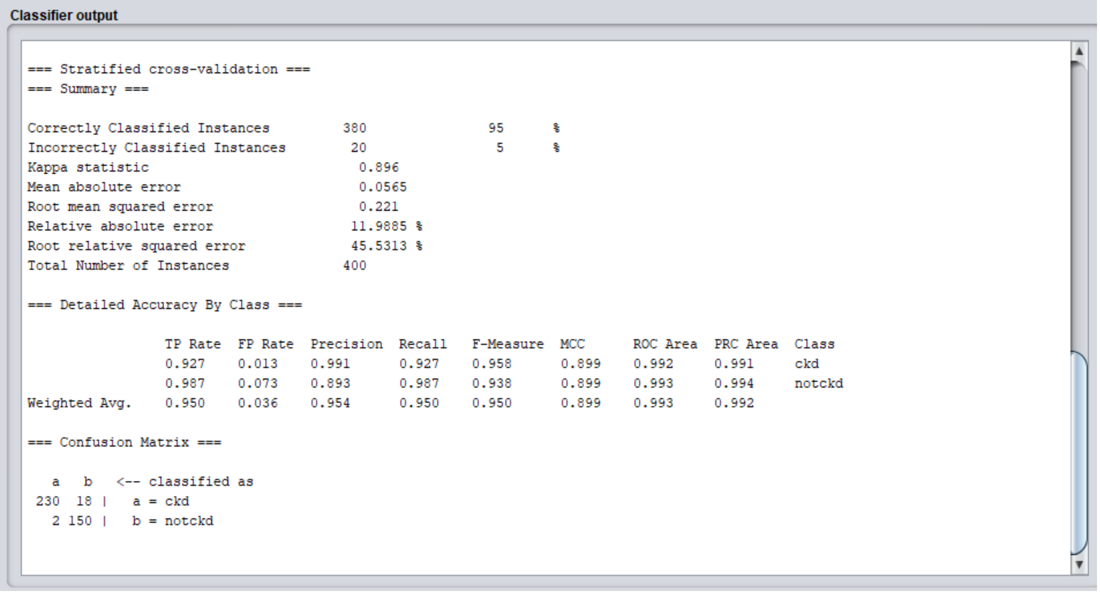
With the Naïve Bayes algorithm, the Correctly Classified Instances (With Cross-Validation - 10 Folds) is 94%.

4.6.4 Hoeffding Tree Algorithm

A Hoeffding tree is an increment based decision tree [27]

Results of Hoeffding Tree Algorithm

The figure below (4.15) shows the results using Hoeffding Tree algorithm with cross-validation – 10 folds.



```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         380               95      %
Incorrectly Classified Instances        20                5      %
Kappa statistic                          0.896
Mean absolute error                      0.0565
Root mean squared error                  0.221
Relative absolute error                 11.9885 %
Root relative squared error             45.5313 %
Total Number of Instances              400

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.927    0.013    0.991      0.927    0.958      0.899    0.992     0.991     ckd
                 0.987    0.073    0.893      0.987    0.938      0.899    0.993     0.994     notckd
Weighted Avg.    0.950    0.036    0.954      0.950    0.950      0.899    0.993     0.992

=== Confusion Matrix ===

   a   b   <-- classified as
 230  18 |   a = ckd
   2 150 |   b = notckd
```

Figure 4.15: Results using Hoeffding Tree Algorithm

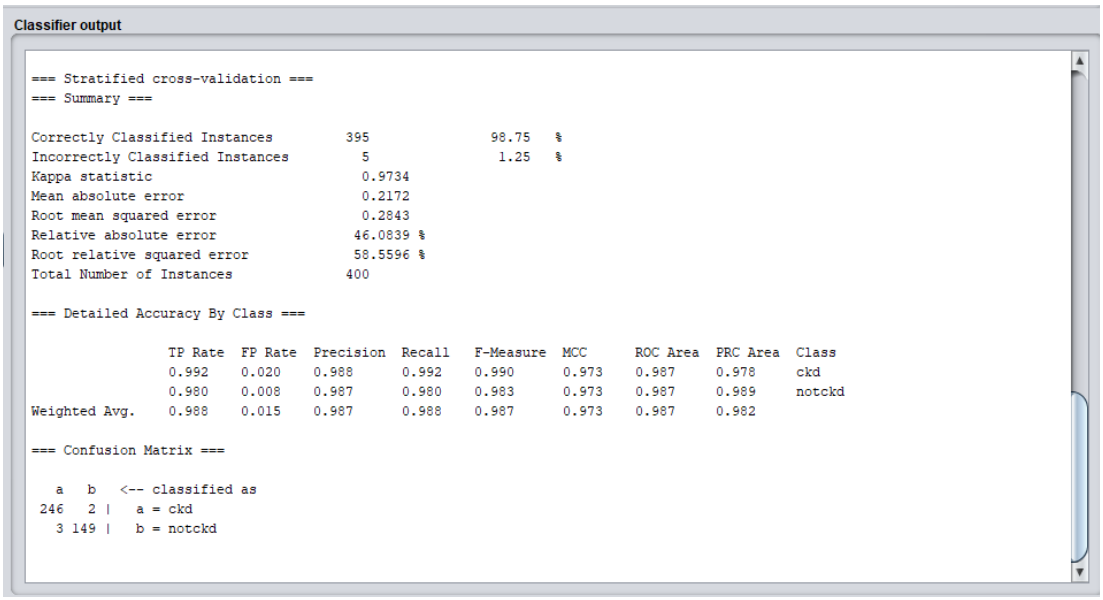
With the Hoeffding Tree algorithm, the Correctly Classified Instances (with Cross-Validation - 10 Folds) is 95%.

4.6.5 Decision Table Algorithm

Decision tables, such as decision trees or neural networks, are structures used for predicting classification. [28]

Results of Decision Table Algorithm

The figure below (4.16) shows the results using the Decision Table algorithm with cross-validation – 10 folds.



```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         395               98.75   %
Incorrectly Classified Instances         5                1.25   %
Kappa statistic                          0.9734
Mean absolute error                      0.2172
Root mean squared error                  0.2843
Relative absolute error                 46.0839 %
Root relative squared error             58.5596 %
Total Number of Instances              400

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.992    0.020    0.988      0.992    0.990      0.973    0.987     0.978     ckd
                 0.980    0.008    0.987      0.980    0.983      0.973    0.987     0.989     notckd
Weighted Avg.    0.988    0.015    0.987      0.988    0.987      0.973    0.987     0.982

=== Confusion Matrix ===

   a   b   <-- classified as
 246   2 |   a = ckd
   3 149 |   b = notckd
```

Figure 4.16: Results using Decision Table Algorithm

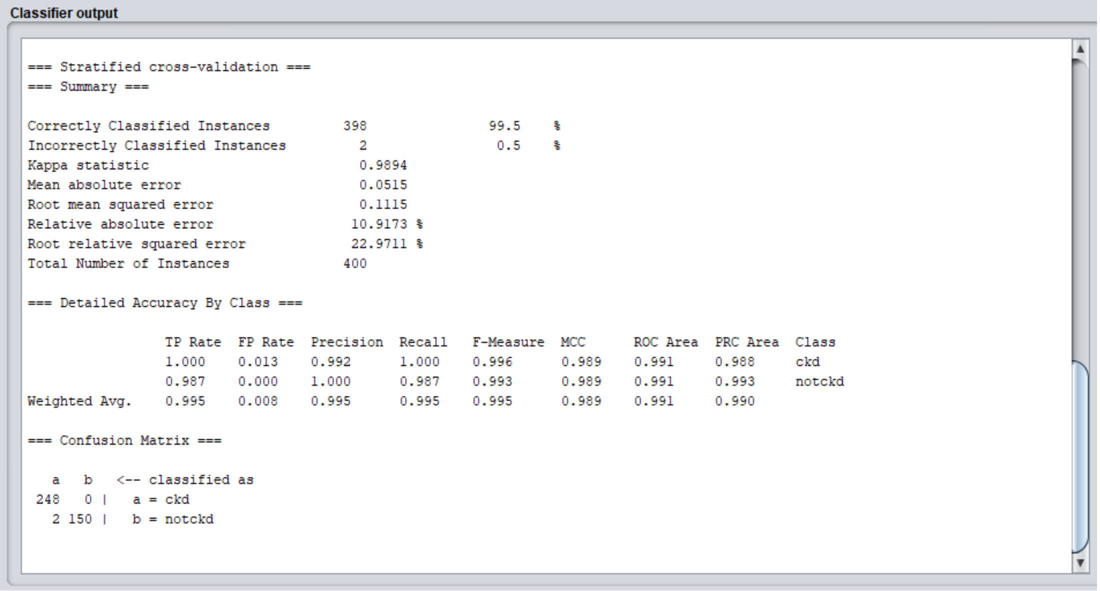
With the Decision Table algorithm, the Correctly Classified Instances (with Cross-Validation - 10 Folds) is 98.75%.

4.6.6 Random Forest Algorithm

Random forest algorithm is a supervised classification algorithm. [29]

Results of the Random Forest Algorithm

The figure below (4.17) shows the results using the Random Forest algorithm with cross-validation. – Ten folds



```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         398               99.5    %
Incorrectly Classified Instances         2                0.5    %
Kappa statistic                          0.9894
Mean absolute error                      0.0515
Root mean squared error                  0.1115
Relative absolute error                 10.9173 %
Root relative squared error             22.9711 %
Total Number of Instances              400

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.013    0.992      1.000    0.996      0.989    0.991     0.988     ckd
                 0.987    0.000    1.000      0.987    0.993      0.989    0.991     0.993     notckd
Weighted Avg.    0.995    0.008    0.995      0.995    0.995      0.989    0.991     0.990

=== Confusion Matrix ===

   a   b   <-- classified as
 248   0 |   a = ckd
   2 150 |   b = notckd
```

Figure 4.17 Results using Random Forest Algorithm

With the Random Forest algorithm, the Correctly Classified Instances (with Cross-Validation - 10 Folds) is 99.5%.

4.7 Comparison of the Results

It is critical to identify the best algorithm for the prediction model because when dealing with patient information and medical records you must make sure your predictions are as accurate as possible. By thoroughly analyzing the output/results of the algorithms which were discussed in the previous section, we can determine which algorithm will be best suited for our prediction & analysis tool.

The summary of the results is listed below (Table 4.1).

Table 4.1: Summary of WEKA results

Algorithm	**With Cross-Validation**	**Precision**	**Recall**
J48	98.75%	0.984	0.996
ZeroR	62%	0.620	1

Naive Bayes	94%	0.911	0.911
Hoeffding Tree	95%	0.911	0.927
Decision Table	98.75%	0.988	0.992
Random Forest	99.5%	0.922	1

Considering the results from Table 4.1 we can determine that the Random Forest algorithm is the most accurate with a cross-validation value of 99.5%.

Because it is delivering the most accurate results when compared to the other algorithms, I have decided to use the Random Forest algorithm in our prediction model.

4.8 Summary

In this chapter we discussed the model & the methodology of the research. The system was developed to collect the patient data, analyze the data and predict whether or not a patient has CKD. Furthermore, we discussed the analysis, which can be done through the system and the prediction model, which recognize the CKD patient by performing a data mining without facing the CKD screening test which is highly expensive. Additionally, the system includes an API which makes available medical information related to the CKD patients so that researchers with access to the API can perform additional studies and research on the dataset.

5. SYSTEM/SOLUTION ARCHITECTURE AND IMPLEMENTATION

The goal of my research was to build a predictive data model for accurately diagnosing patients with CKD, however a database for CKD patients and their medical information did not exist. To help create a working dataset I developed a web/mobile-based interface to collect relevant patient information. The system is composed of three main components: centralized system to collect data; analysis; prediction. In addition to predicting the probability of CKD in a patient based on the data, the dataset can also be used to generate various reports for continued patient analysis.

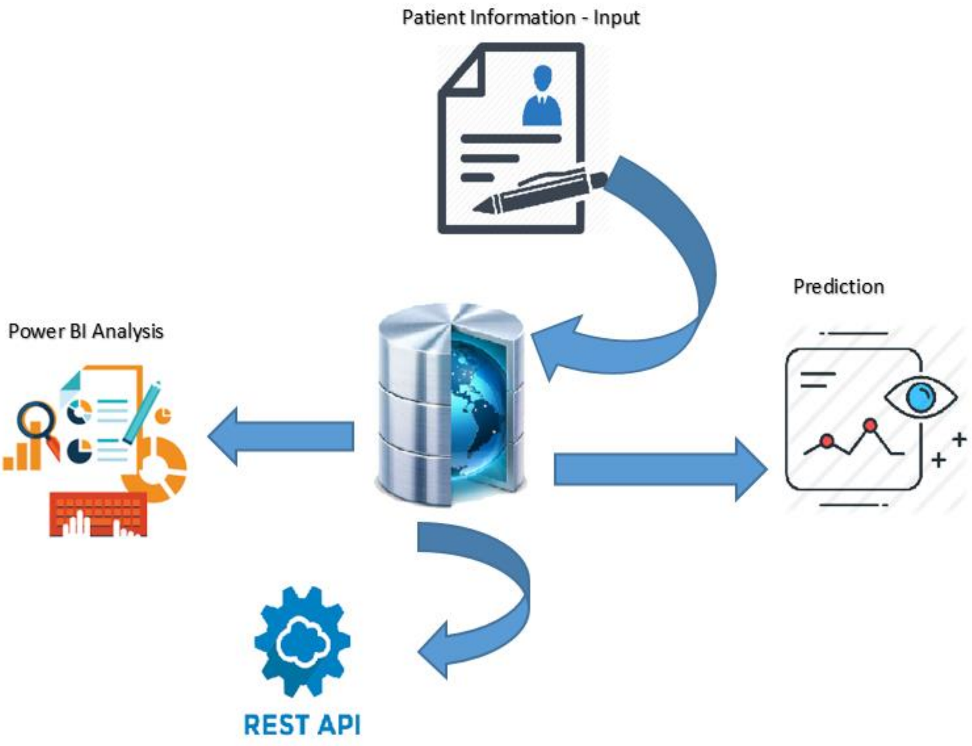
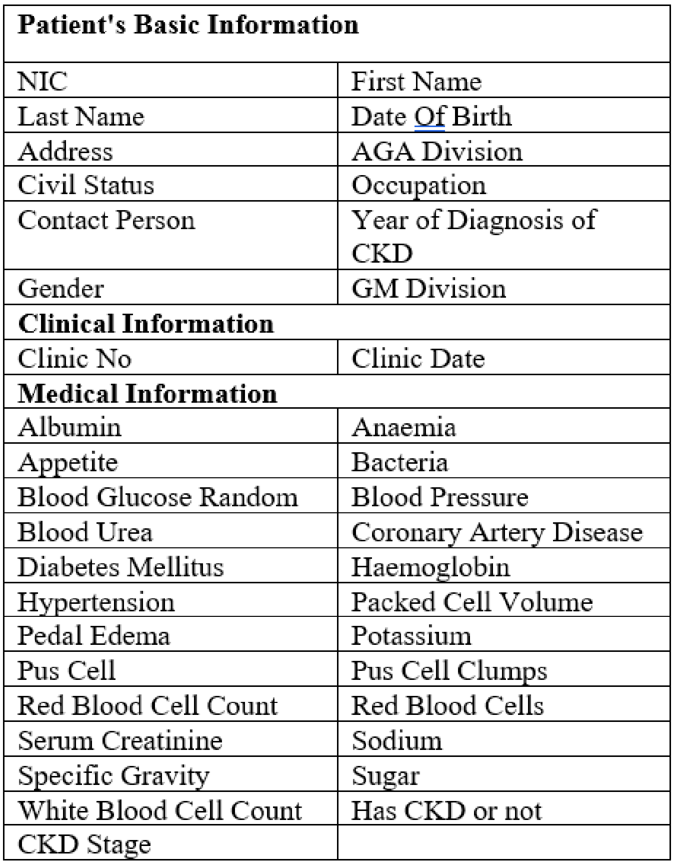


Figure 5.1: System Architecture

5.1 Patient Database

The table below lists all of the patient information that was collected by the system.

Table 5.1: Attributes of CKD Patients



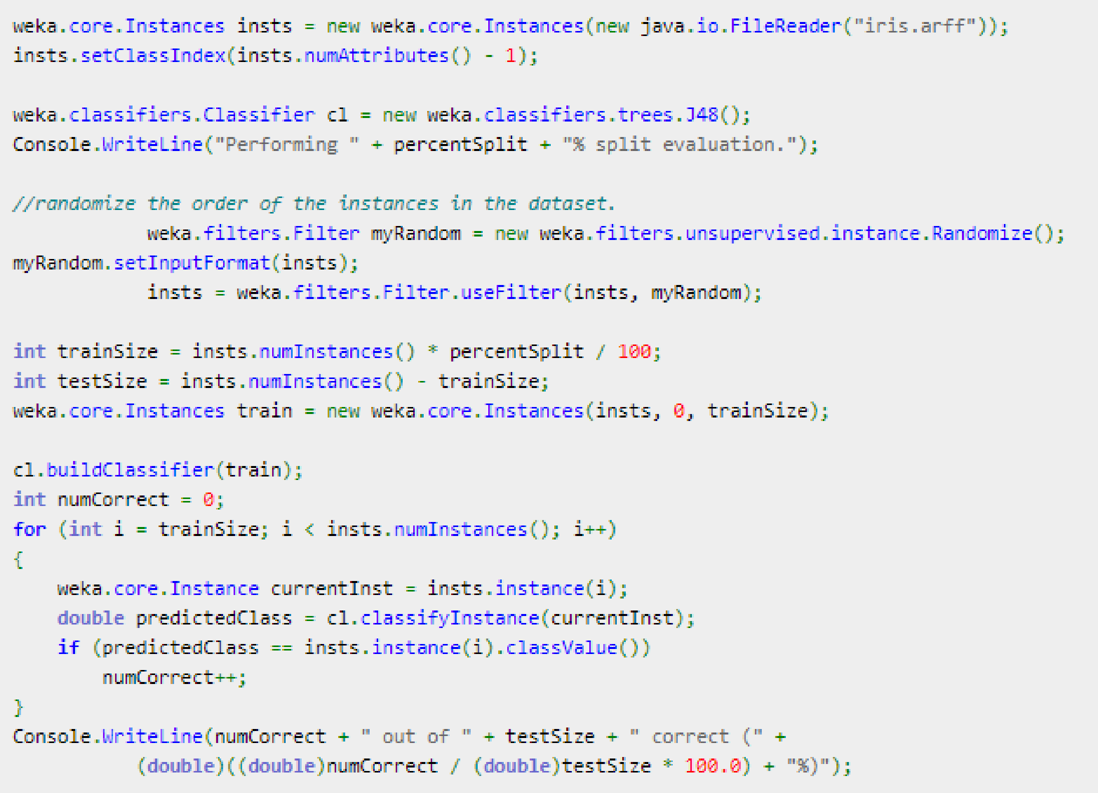
Patient's Basic Information	
NIC	First Name
Last Name	Date Of Birth
Address	AGA Division
Civil Status	Occupation
Contact Person	Year of Diagnosis of CKD
Gender	GM Division
Clinical Information	
Clinic No	Clinic Date
Medical Information	
Albumin	Anaemia
Appetite	Bacteria
Blood Glucose Random	Blood Pressure
Blood Urea	Coronary Artery Disease
Diabetes Mellitus	Haemoglobin
Hypertension	Packed Cell Volume
Pedal Edema	Potassium
Pus Cell	Pus Cell Clumps
Red Blood Cell Count	Red Blood Cells
Serum Creatinine	Sodium
Specific Gravity	Sugar
White Blood Cell Count	Has CKD or not
CKD Stage	

5.2 Prediction Tool

To integrate the WEKA prediction tool with the centralized system, a plugin was used. Based on the medical information that is collected from each patient, the tool can predict the percentage chance of each patient having CKD.

As previously discussed, the prediction tool uses The Random Forest algorithm for its calculations.

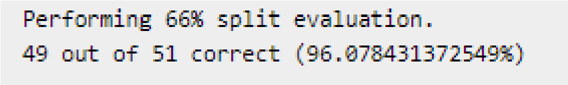
Sample Code:-



```
weka.core.Instances insts = new weka.core.Instances(new java.io.FileReader("iris.arff"));
insts.setClassIndex(insts.numAttributes() - 1);

weka.classifiers.Classifier cl = new weka.classifiers.trees.J48();
Console.WriteLine("Performing " + percentSplit + "% split evaluation.");

//randomize the order of the instances in the dataset.
            weka.filters.Filter myRandom = new weka.filters.unsupervised.instance.Randomize();
myRandom.setInputFormat(insts);
            insts = weka.filters.Filter.useFilter(insts, myRandom);

int trainSize = insts.numInstances() * percentSplit / 100;
int testSize = insts.numInstances() - trainSize;
weka.core.Instances train = new weka.core.Instances(insts, 0, trainSize);

cl.buildClassifier(train);
int numCorrect = 0;
for (int i = trainSize; i < insts.numInstances(); i++)
{
    weka.core.Instance currentInst = insts.instance(i);
    double predictedClass = cl.classifyInstance(currentInst);
    if (predictedClass == insts.instance(i).classValue())
        numCorrect++;
}
Console.WriteLine(numCorrect + " out of " + testSize + " correct (" +
           (double)((double)numCorrect / (double)testSize * 100.0) + "%)");
```

Figure 5.2: Sample WEKA code

The results of the sample code are as follows;



```
Performing 66% split evaluation.
49 out of 51 correct (96.078431372549%)
```

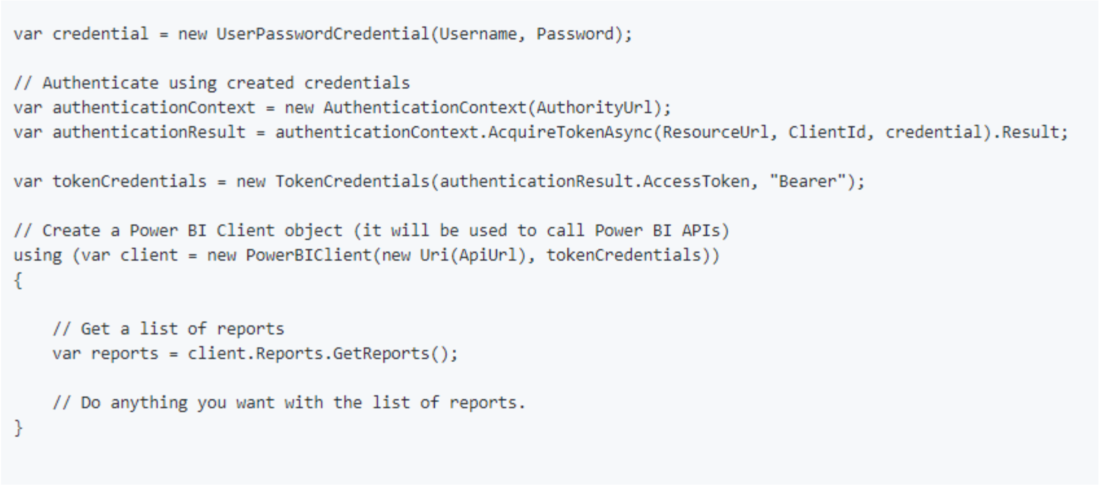
Figure 5.3: Weka Results

5.3 Patient Data Analysis

The analysis tool that is built within the primary system is also integrated with Power BI. This Power BI tool is integrated into the order using the available plugin which supports the same framework as the centralized system that we developed.

The sample code below shows how the Power BI tool was integrated with the system using this plugin.

Sample Code:



```
var credential = new UserPasswordCredential(Username, Password);

// Authenticate using created credentials
var authenticationContext = new AuthenticationContext(AuthorityUrl);
var authenticationResult = authenticationContext.AcquireTokenAsync(ResourceUrl, ClientId, credential).Result;

var tokenCredentials = new TokenCredentials(authenticationResult.AccessToken, "Bearer");

// Create a Power BI Client object (it will be used to call Power BI APIs)
using (var client = new PowerBIClient(new Uri(ApiUrl), tokenCredentials))
{

    // Get a list of reports
    var reports = client.Reports.GetReports();

    // Do anything you want with the list of reports.
}
```

Figure 5.4: BI sample code

5.4 Summary

In this chapter I discussed the architecture of the centralized system and how the various components of the solution were developed. Data was collected through the patient entry form and was used for different analyses using both the integrated Power BI tool and the custom analysis tool. To incorporate the Power BI with the system, the Microsoft Power BI plugin was used. A plugin was also used to connect with the WEKA tool. The system includes a predictive data model which can identify if the patient has CKD. This model uses the Random Forest algorithm for its classifications.

6. SYSTEM EVALUATION (DATA AND ANALYSIS)

In this chapter I discuss the techniques which I have used to evaluate the system.

A lot of information is captured for each patient via the patient entry form. However, the majority of the information collected is not directly used by the prediction model. Some of these fields that are not used include, but are not limited to, Name, NIC, Address, Contact No… etc. This requires the implementation of a data cleanup process to filter out the information that is not directly used by the prediction tool.

6.1 Data Pre-processing

The dataset consists of following attributes related to medical information. (Table 6.1)

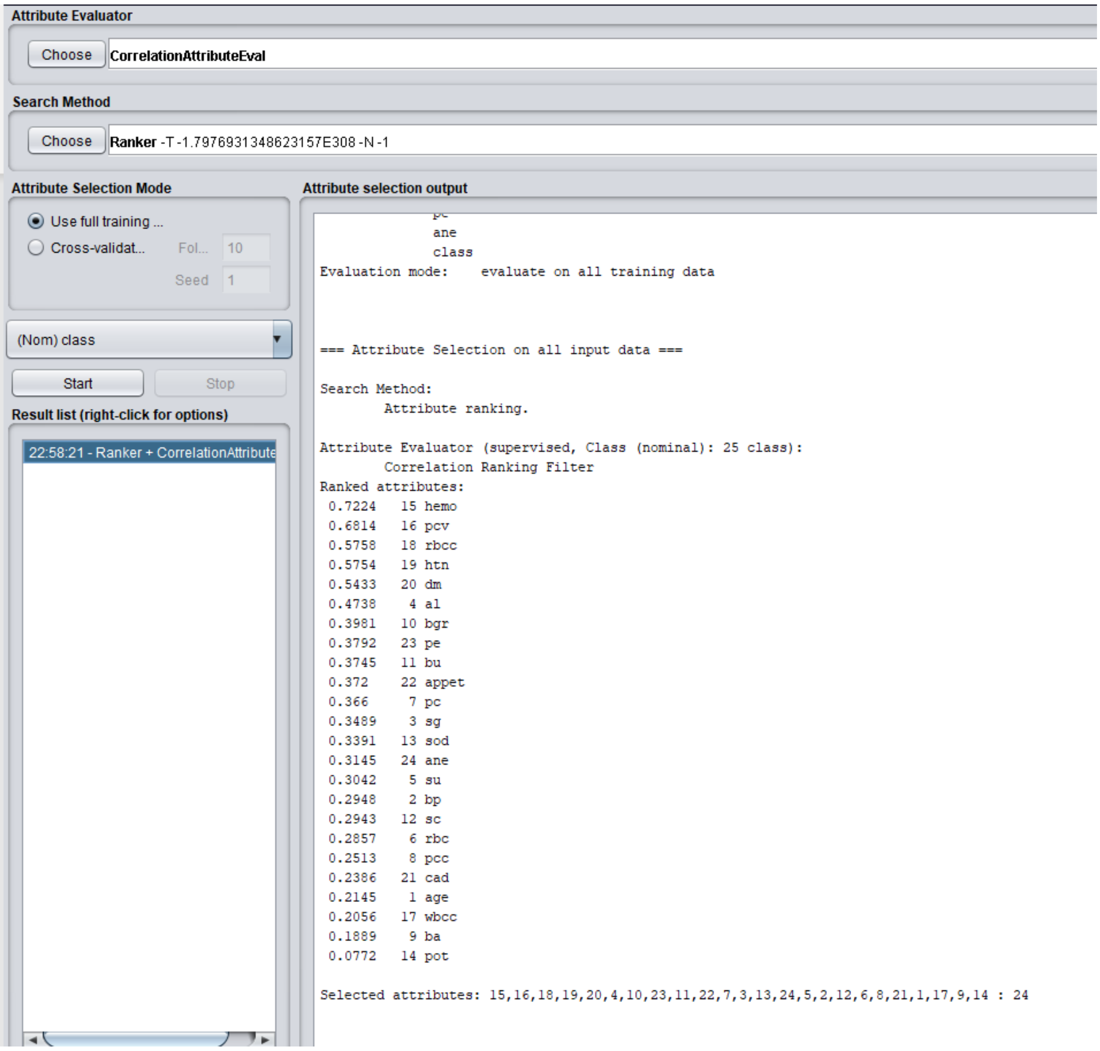
Table 6.1: Medical information in the dataset

Age	Blood Pressure
Albumin	Sugar
Pus Cell	Pus Cell Clumps
Blood Glucose Random	Blood Urea
Sodium	Potassium
Packed Cell Volume	White Blood Cell Count
Hypertension	Diabetes Mellitus
Appetite	Pedal Edema
Specific Gravity	Bacteria
Red Blood Cells	Anaemia
Serum Creatinine	Haemoglobin
Red Blood Cell Count	Coronary Artery Disease

The data model with above attributes has a 99.5% accuracy rate. To select the optimal subsets of data I have used Correlation based feature selection.

Correlation based feature selection

I used “CorrelationAttributeEval” as the attribute evaluator and “Ranker” as the search method. (Figure 6.1)


Choose
Choose
Attribute Selection Mode
Fol... 10
Seed 1
Start
Stop
Result list (right-click for options)
Attribute selection output

Attribute Evaluator
CorrelationAttributeEval

Search Method
Ranker -T -1.7976931348623157E308 -N -1

● Use full training ...
○ Cross-validat...

(Nom) class

22:58:21 - Ranker + CorrelationAttribute

```
              pc
              ane
              class
Evaluation mode:    evaluate on all training data



=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 25 class):
        Correlation Ranking Filter
Ranked attributes:
 0.7224   15 hemo
 0.6814   16 pcv
 0.5758   18 rbcc
 0.5754   19 htn
 0.5433   20 dm
 0.4738    4 al
 0.3981   10 bgr
 0.3792   23 pe
 0.3745   11 bu
 0.372    22 appet
 0.366     7 pc
 0.3489    3 sg
 0.3391   13 sod
 0.3145   24 ane
 0.3042    5 su
 0.2948    2 bp
 0.2943   12 sc
 0.2857    6 rbc
 0.2513    8 pcc
 0.2386   21 cad
 0.2145    1 age
 0.2056   17 wbcc
 0.1889    9 ba
 0.0772   14 pot

Selected attributes: 15,16,18,19,20,4,10,23,11,22,7,3,13,24,5,2,12,6,8,21,1,17,9,14 : 24
```

Figure 6.1: Correlation based feature selection

According to the results, out of 25 attributes, 15 attributes show more than 0.3 of correlations. Therefore, I have selected those 15 attributes as the optimal subset of attributes to predict CKD.

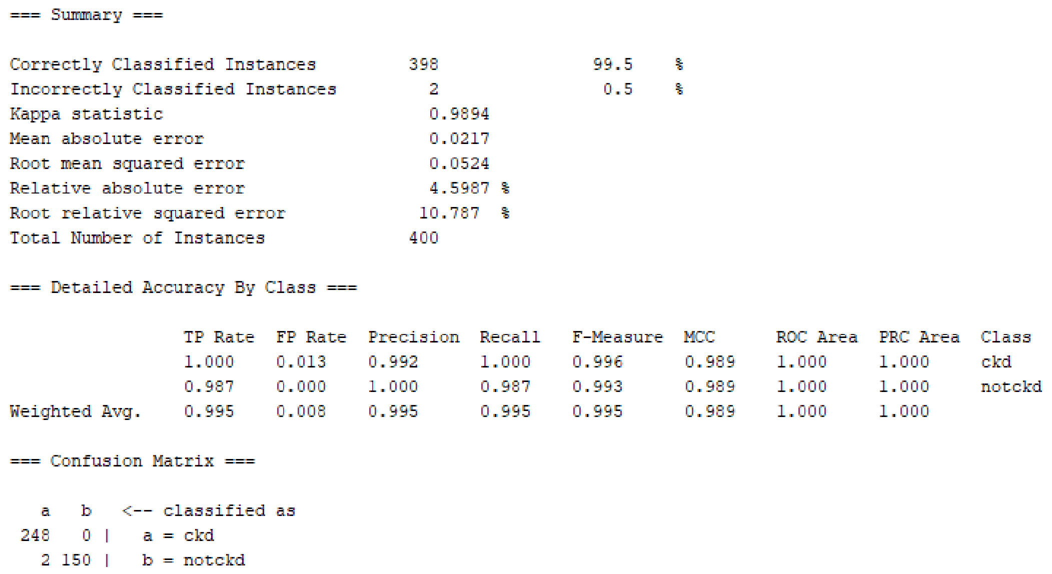
The attributes which are selected for the Final prediction model are listed below (Table 6.2).

Table 6.2: Selected attributes for the predictive data model

Haemoglobin	Packed Cell Volume
Red Blood Cell Count	Hypertension
Diabetes Mellitus	Albumin
Blood Glucose Random	Appetite
Pus Cell	Pedal Edema
Blood Urea	Specific Gravity
Sodium	Anaemia
Sugar	

The data model with these attributes also has a 99.5% accuracy rate. To validate the accuracy of the data model I have used the WEKA tool.

The screen below (6.2) shows the accuracy of the data model using the WEKA tool.

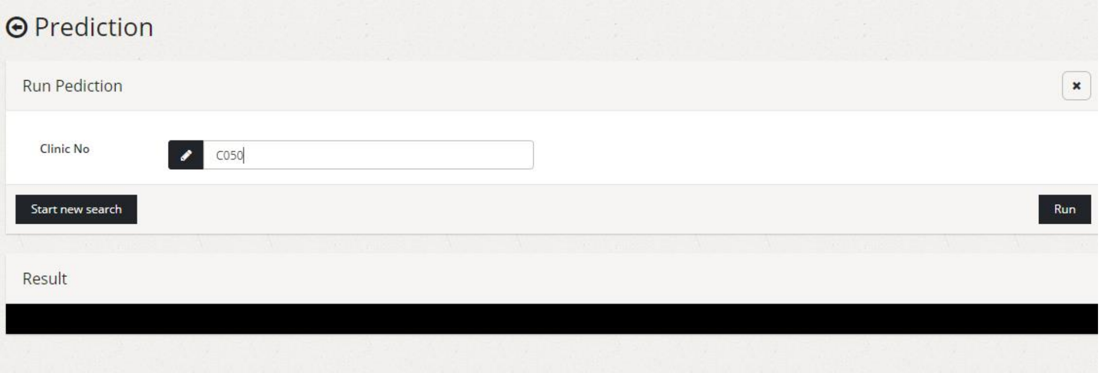


```
=== Summary ===

Correctly Classified Instances         398               99.5    %
Incorrectly Classified Instances         2                0.5    %
Kappa statistic                          0.9894
Mean absolute error                      0.0217
Root mean squared error                  0.0524
Relative absolute error                  4.5987 %
Root relative squared error             10.787  %
Total Number of Instances              400

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.013    0.992      1.000    0.996      0.989    1.000     1.000     ckd
                 0.987    0.000    1.000      0.987    0.993      0.989    1.000     1.000     notckd
Weighted Avg.    0.995    0.008    0.995      0.995    0.995      0.989    1.000     1.000

=== Confusion Matrix ===

   a   b   <-- classified as
 248   0 |   a = ckd
   2 150 |   b = notckd
```

Figure 6.2: Accuracy of the model

6.2 Results / Outcome of the Prediction Tool

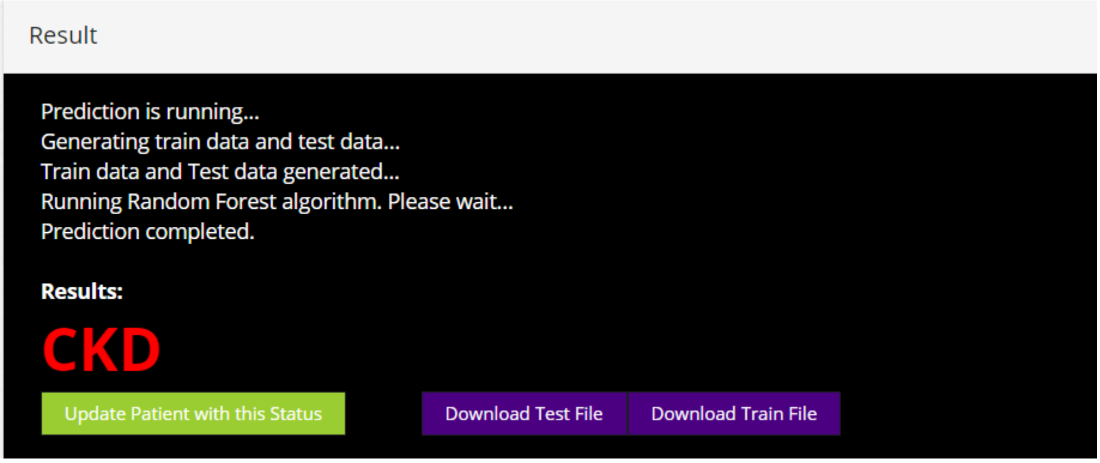
From the tool the system user can determine if the patient has CKD or not by selecting the particular patient and running the analysis (Figure 6.2).



Clinic No C050

Figure 6.2 Screenshot of Prediction Tool

Once a system user provides the patient record number and runs the tool, the predictive model creates two datasets: a train dataset and a test dataset. It processes and analyzes each data set using the Random Forest algorithm and delivers the results as either "CKD" or "Non-CKD" for the patient. (Figure 6.3).



```
Prediction is running...
Generating train data and test data...
Train data and Test data generated...
Running Random Forest algorithm. Please wait...
Prediction completed.

Results:
CKD
```

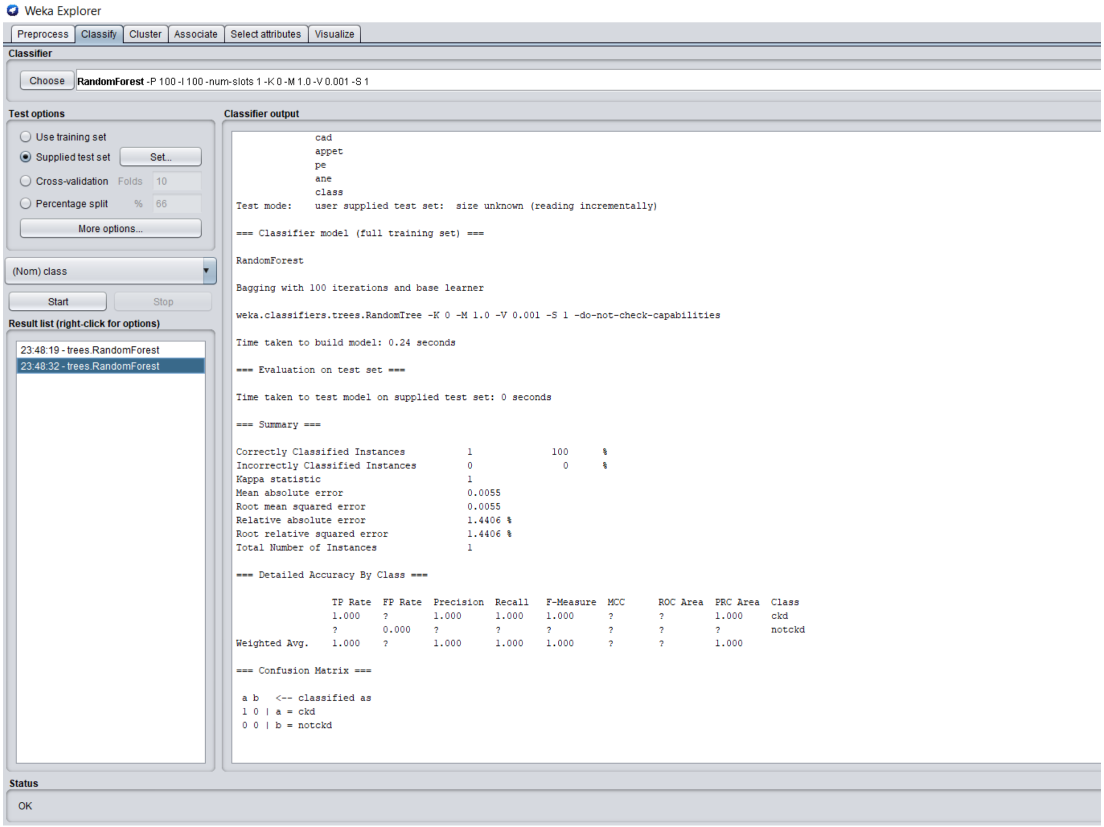
Figure 6.3: Prediction Results

6.3 Validate the Results of the Prediction Tool

In the prediction tool results, there are two buttons to download the Train and Test files. This will download “train.arff” file and “test.arff” files. The files have 25 attributes.

Using the WEKA tool, we can manually train the model and then run the classification using the Random Forest algorithm. Then we can compare the results with the Prediction tool and determine the accuracy of the results.

The figure below (6.4) shows the results of the Weka Tool for CKD patients.



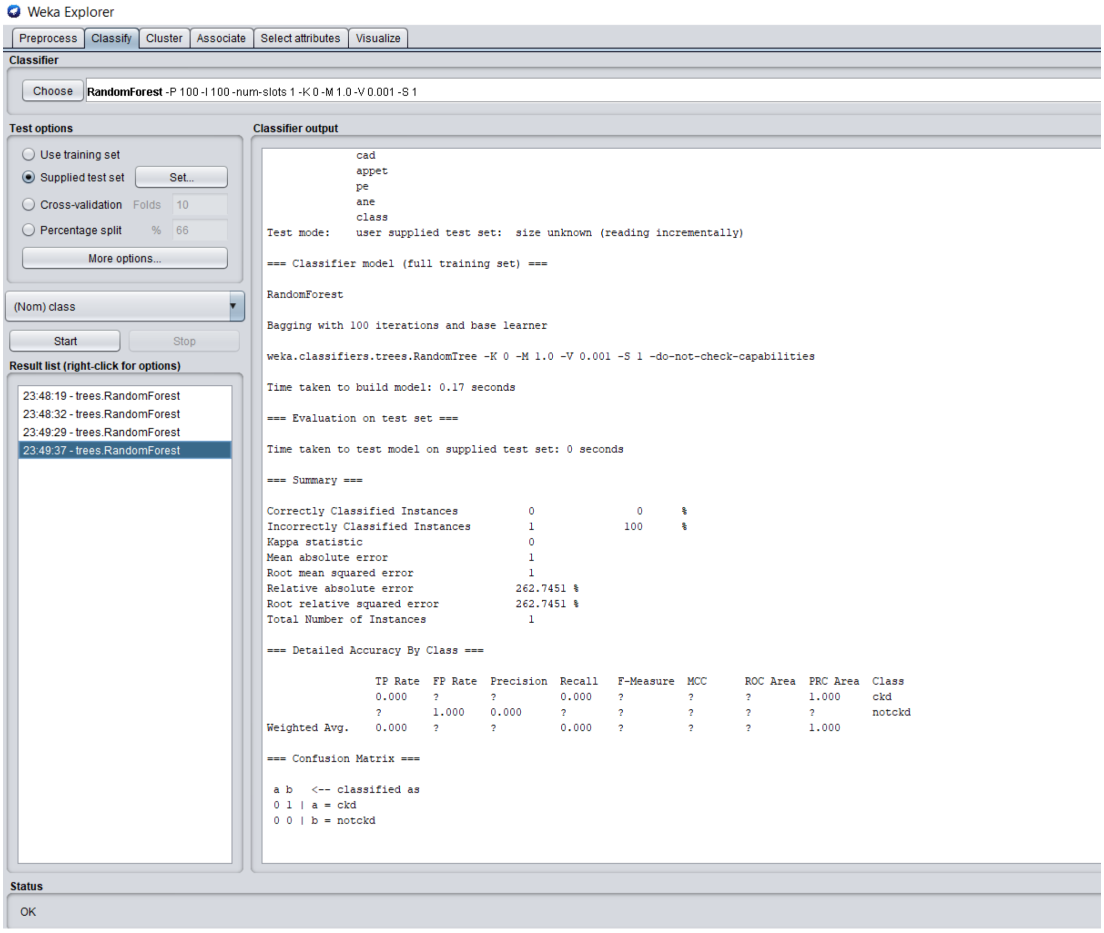
RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1

```
              cad
              appet
              pe
              ane
              class
Test mode:    user supplied test set:  size unknown (reading incrementally)

=== Classifier model (full training set) ===

RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 0.24 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0 seconds

=== Summary ===

Correctly Classified Instances           1              100      %
Incorrectly Classified Instances         0                0      %
Kappa statistic                          1
Mean absolute error                      0.0055
Root mean squared error                  0.0055
Relative absolute error                  1.4406 %
Root relative squared error              1.4406 %
Total Number of Instances                1

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    ?        1.000      1.000    1.000      ?        ?         1.000     ckd
                 ?        0.000    ?          ?        ?          ?        ?         ?         notckd
Weighted Avg.    1.000    ?        1.000      1.000    1.000      ?        ?         1.000

=== Confusion Matrix ===

 a b   <-- classified as
 1 0 | a = ckd
 0 0 | b = notckd
```

Figure 6.4: Weka Results for CKD Patients

Figure 6.4 shows the 'Correctly Classified Instances' is 100% for the above scenario, indication that the patient does have CKD.

The figure below (6.5) shows the results of the Weka Tool for non-CKD patients.



RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1

○ Use training set
● Supplied test set
○ Cross-validation Folds 10
○ Percentage split % 66

(Nom) class

23:48:19 - trees.RandomForest
23:48:32 - trees.RandomForest
23:49:29 - trees.RandomForest
23:49:37 - trees.RandomForest

```
              cad
              appet
              pe
              ane
              class
Test mode:    user supplied test set:  size unknown (reading incrementally)

=== Classifier model (full training set) ===

RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 0.17 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0 seconds

=== Summary ===

Correctly Classified Instances           0                0      %
Incorrectly Classified Instances         1              100      %
Kappa statistic                          0
Mean absolute error                      1
Root mean squared error                  1
Relative absolute error                262.7451 %
Root relative squared error            262.7451 %
Total Number of Instances                1

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    ?        ?          0.000    ?          ?        ?         1.000     ckd
                 ?        1.000    0.000      ?        ?          ?        ?         ?         notckd
Weighted Avg.    0.000    ?        ?          0.000    ?          ?        ?         1.000

=== Confusion Matrix ===

 a b   <-- classified as
 0 1 | a = ckd
 0 0 | b = notckd
```

OK

Figure 6.5: Weka Results for non-CKD Patients

In this scenario, the 'Correctly Classified Instances' is 0%, indication that the patient does not have CKD. For medical data modelling, anything that does not accountable for false negatives consider as critical. The recall identifies as the better measure than precision. Precision is a measurement of how relevant a positive result is and Recall (red cells) is a measurement of the proportion of correct positive results. It is also known as the True Positive Rate or Sensitivity. Let’s assume, from our prediction tool return a result indicating the patient is a non-CKD patient. But if the person is a CKD patient, then it is a critical error as we are dealing with the life of the person. Hence Precision/Recall is a more meaningful measurement in a medical application

6.4 Patient Data Analysis Tool

From this tool, the system user can also generate several graphs based on the collected patient data (Figure 6.6).

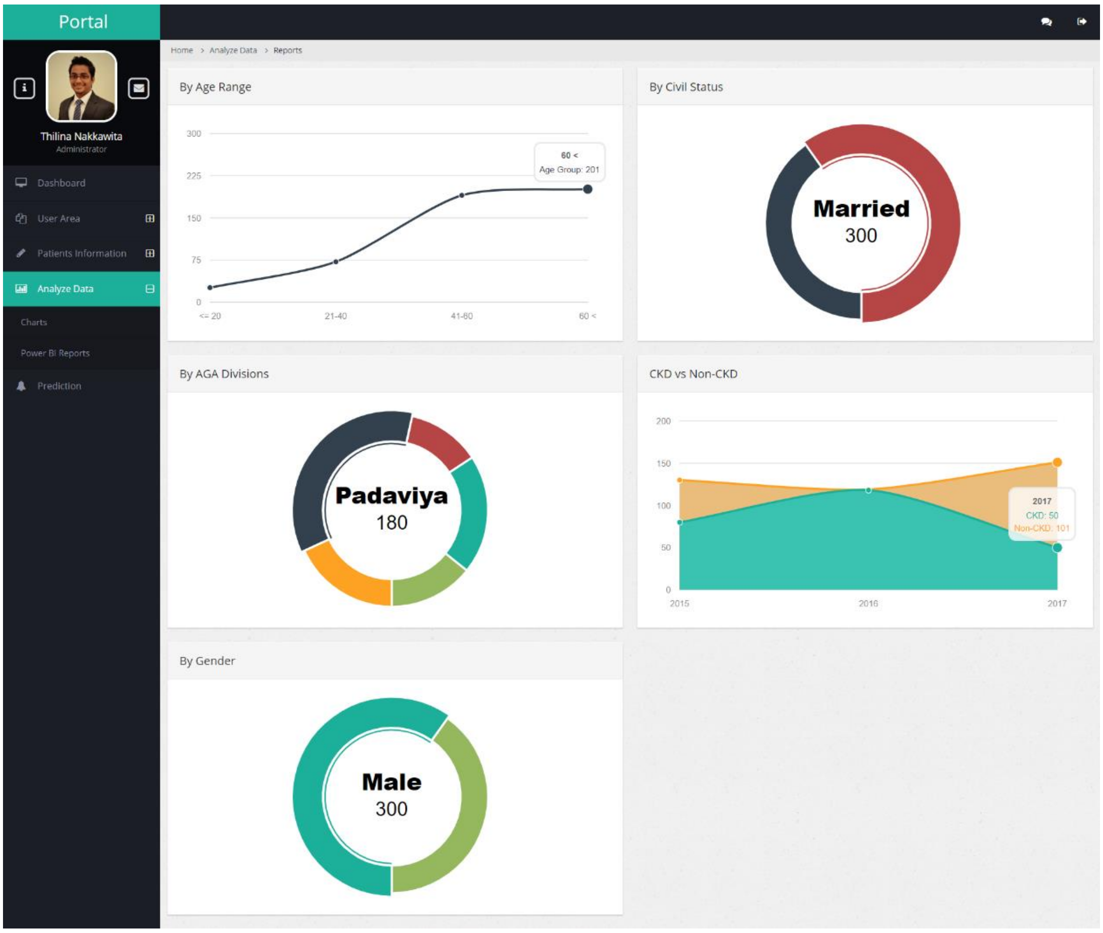
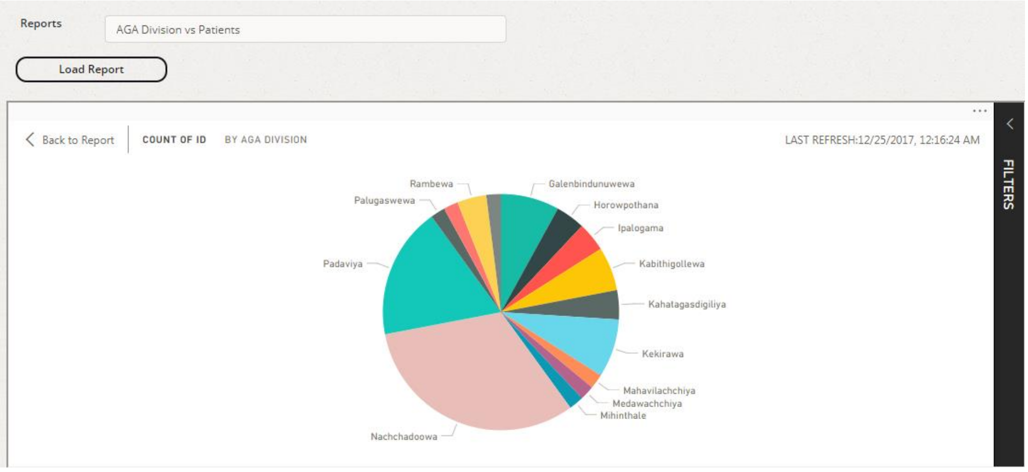


Figure 6.6: Analysis Tool

Also, POWER BI integration is available within the system. The user can generate several reports using this tool (Figure 6.7).



Reports AGA Division vs Patients

Figure 6.7: Power BI Chart

6.5 Summary

In this chapter I discussed the various algorithms which were considered to find out which was the best algorithm to use in the prediction tool. I ran some tests for each algorithm and analyzed the results of each algorithm independently. The Random Forest algorithm showed a 99.5% accuracy; therefore, I selected this algorithm for the predictive model.

Furthermore, I discussed how the prediction tool works, and how it was developed into the system and is used to validate the results generated by the centralized system. Finally, I discussed the patient data analysis tool, included in the system.

7. CONCLUSION

The main objective of the research was to gather CKD patient information, build a predictive model, and analyze the patient data.

Currently, to diagnose CKD in a patient a CKD screening test must be performed. These screening tests are costly for both the patient and the Sri Lankan government. To help cut down on costs, the goal of my research was to develop a data collection and analysis tool capable of accurately predicting the probability that a patient does, or does not, have CKD. By providing a probability percentage of having CKD for each patient we can filter out the patients that do not have CKD and therefore they do not need to undergo a CKD screening test.

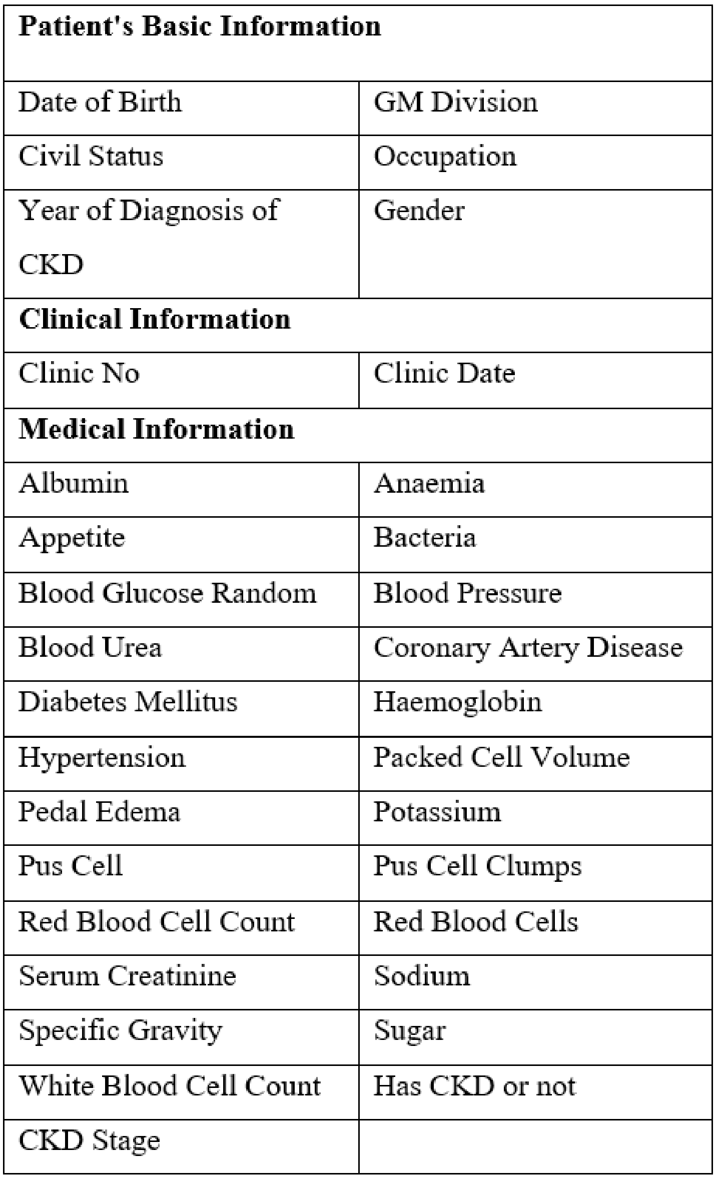
To gather patient information, I have developed a mobile and web-based patient entry form. The interface can be used in hospitals, especially in areas where CKD is the most prevalent.

I have also developed a predictive data model that can accurately diagnose whether a patient has CKD. This model helps to identify patients in the early stages of CKD, allowing doctors to more effectively determine which patients need additional medical testing.

The predictive model uses only the necessary information from a larger group of data collected for each patient in its analysis and uses the best available algorithm.

The data used by the prediction model to produce the accurate results are below.

Table 7.1: Attributes of patient record used by prediction model



Patient's Basic Information	
Date of Birth	GM Division
Civil Status	Occupation
Year of Diagnosis of CKD	Gender
Clinical Information	
Clinic No	Clinic Date
Medical Information	
Albumin	Anaemia
Appetite	Bacteria
Blood Glucose Random	Blood Pressure
Blood Urea	Coronary Artery Disease
Diabetes Mellitus	Haemoglobin
Hypertension	Packed Cell Volume
Pedal Edema	Potassium
Pus Cell	Pus Cell Clumps
Red Blood Cell Count	Red Blood Cells
Serum Creatinine	Sodium
Specific Gravity	Sugar
White Blood Cell Count	Has CKD or not
CKD Stage	

Since several algorithms can be used when building the predictive model, the main challenge was to select the best suitable algorithm.

To choose the best algorithm, I had to compare the results generated by each algorithm with the WEKA tool.

The comparison of the results of each algorithm is listed below.

Table 7.2: Summary of WEKA results

Algorithm	**With Cross-Validation**	**Precision**	**Recall**
J48	98.75%	0.984	0.996
ZeroR	62%	0.620	1
Naive Bayes	94%	0.911	0.911
Hoeffding Tree	95%	0.911	0.927
Decision Table	98.75%	0.988	0.992
Random Forest	99.5%	0.922	1

According to the summary above, we can determine that the 'Random Forest' algorithm is the best algorithm as it produces a 99.75% accuracy rate. Several researchers in previous studies were not able to report an accuracy as high as 99.75%. Therefore, this model can be considered as the most accurate model built for CKD prediction.

Also, in the above table, I have listed the Precision and Recall for each algorithm. These values are important, especially when building a medical application.

For medical data modelling, anything that does not consider for false-negatives very critical. The Recall is a better measure than Precision. Precision is a measurement of how relevant a positive result is and Recall (red cells) is a measurement of the proportion of correct positive results. It is also known as the True Positive Rate or Sensitivity. Let's assume, for example, that from our prediction tool we receive result indicating that a patient does not have CKD. If however, the patient truly does have CKD, then it is a critical error as we are dealing with the patient's life. Hence, Recall is a more meaningful measurement in a medical application.

The prediction tool helps to identify which patients that visit the clinic or hospital have CKD. Also, in the system I developed, the users of the system can generate several reports using the integrated POWER BI application for further analysis. Additionally, since other researchers can access the data model via an API, they have the ability to do their own tests and studies with the data that has been collected.

7.1 Future Works

This system was built only for CKD patient analysis and if produces predictions based on the collected data. In the future, we can develop similar data collection and analysis tools for other diseases as well.

Also, currently, the system collects and analyzes patient data only from local hospitals. If this can be expanded to collect patient data internationally, for example, then we will be able to investigate data from a larger geographical context and build a predictive model which could help people all over the world.

REFERENCES

[1]. "Glomerular filtration rate: MedlinePlus Medical Encyclopedia", *Medlineplus.gov*, 2019. [Online]. Available: https://medlineplus.gov/ency/article/007305.htm. [Accessed: 22- Mar- 2019].

[2]. "Stages of Chronic Kidney Disease (CKD)", *Kidneyfund.org*, 2019. [Online]. Available: http://www.kidneyfund.org/kidney-disease/chronic-kidney-disease-ckd/stages-of-chronic-kidney-disease. [Accessed: 22- Mar- 2019].

[3]. C. Nordqvist, "Chronic kidney disease: Symptoms, causes, and treatment", *Medical News Today*, 2019. [Online]. Available: http://www.medicalnewstoday.com/articles/172179.php. [Accessed: 22- Mar- 2019].

[4]. Ruvini Takshala Rubasinghe, Sunethra Kanthi Gunatilake, S. Sunil Samaratunga. *Chronic Kidney Disease (CKD) in Sri Lanka - Current Research Evidence Justification: A Review*. Sabaragamuwa University Journal, 2015.

[5]. K. Wanigasuriya. *Aetiological factors of Chronic Kidney Disease in the North Central Province of Sri Lanka: A review of evidence to-date. Journal of the college of community physician of Sri Lanka*, vol. 17, 2012.

[6]. "What is Data Mining, Predictive Analytics, Big Data", *Statsoft.com*, 2019. [Online]. Available: http://www.statsoft.com/textbook/data-mining-techniques. [Accessed: 22- Mar- 2019].

[7]. Hubert Kouame Yao, Serge Didier Konan, Sindou Sanogo, Sery Patrick Diopoh, Amadou Demba Diallo, *Prevalence and risk factors of chronic kidney disease in Cote D'Ivoire: An analytic study conducted in the*

department of internal medicine, Saudi Journal of Kidney Diseases and Transplantation, 2018, P. 153-159

[8]. Lenildo de Moura, Silvânia Suely Caribé de Araújo Andrade, Deborah Carvalho Malta, Cimar Azeredo Pereira, José Eduardo Fogolin Passos, *Prevalence of self-reported chronic kidney disease in Brazil: National Health Survey of 2013,* Revista Brasileira de Epidemiologia, 2015, vol 18

[9]. Tyler Knight, Caroline Schaefer, Holly Krasa, Dorothee Oberdhan, Arlene Chapman, Ronald D Perrone, '*Medical resource utilization and costs associated with autosomal dominant polycystic kidney disease in the USA: a retrospective matched cohort analysis of private insurer data*', ClinicoEconomics and Outcomes Research, 2015

[10]. "Kidney failure in patients predicted with a simple equation | CBC News", *CBC*, 2019. [Online]. Available: http://www.cbc.ca/news/canada/manitoba/winnipeg-researcher-makes-predicting-kidney-failure-easy-with-simple-equation-1.3399857. [Accessed: 22- Mar- 2019].

[11]. Andrew S. Levey et al. Navdeep Tangri, Morgan E. Grams. Multinational *Assessment of Accuracy of Equations for Predicting Risk of Kidney Failure*. Journal of the American Medical Association, vol. 315, 2016.

[12]. Zoccali C Borrelli S Cianciaruso B Di Iorio B Santoro D Giancaspro V Abaterusso C Gallo C Conte G Minutolo R De Nicola L, Chiodini P. *Prognosis of CKD patients receiving outpatient nephrology care in Italy*. Clinical Journal of the American Society of Nephrology, 2011.

[13]. Chiodini P Borrelli S Zoccali C Postorino M Iodice C Nappi F Fuiano G Gallo C Conte G De Nicola L, Minutolo R. *The effect of increasing age on*

the prognosis of non-dialysis patients with chronic kidney disease receiving stable nephrology care. Kidney International, vol. 82, 2012.

[14]. Dr S. Vijayarani and Mr S Dhayanand. *Data Mining Classification Algorithms for Kidney Disease Prediction*. International Journal of Cybernetics and informatics (IJCI), vol. 4, 2015.

[15]. S. Ramya and Dr N. Radha. *Diagnosis of Chronic Kidney Disease Using Machine Learning Algorithms*. International Journal of Innovative Research in Computer and Communication Engineering, vol. 4, 2016.

[16]. Narendra Ku. Kamila Lambodar Jeena. *Distributed Data Mining Classification Algorithms for Prediction of Chronic Kidney Disease*. International Journal of Engineering Research in management and Technology, vol. 4, 2015.

[17]. Grith J Tighiouart H Djurdjev O Naimark D Levin A Levey AS Tangri N, Stevens LA. *A predictive model for progression of chronic kidney disease to kidney failure*. 2011.

[18]. Vanessa R. Silva Juliana O. Magalhes Isabella P. Barcelos Mariana G. Duarte Sergio V. Pinheiro Enrico A. Colosimo Ana Cristina Simes e Silva Eduardo A. Oliveira Debora C. Cerqueira, Cristina M. Soares. *A Predictive Model of Progression of CKD to ESRD in a Predialysis Pediatric Interdisciplinary Program*. Clinical Journal of the American Society of Nephrology, vol. 9, 2014.

[19]. Hiebert B Wong J Naimark D Kent D Levey AS Tangri N, Inker LA. *A Dynamic Predictive Model for Progression of CKD*. American Journal of Kidney Diseases, 2016.

[20]. Fres BP Faria LD Pinto JS Nogueira MM Lima GO Resende PI Assis NS Simes E Silva AC Pinheiro SV Mendona AC, Oliveira EA. *A predictive model of progressive chronic kidney disease in idiopathic nephrotic syndrome. Pediatric Nephrology,* vol. 30, 2015.

[21]. Jamie S Hirsch David Blei Nomie Elhadad Adler Perotte, Rajesh Ranganath. *Risk Prediction For Chronic Kidney Disease Progression Using Heterogeneous Electronic Health Record Data And Time Series Analysis*. Journal of the American Medical Informatics Association, 2015.

[22]. Seyed Ahmad Mirbagheri Mitra Mahdavi Mazdeh Jamshid Norouzi, Ali Yadollahpour and Seyed Ahmad Hosseini. *Predicting Renal Failure Progression in Chronic Kidney Disease Using Integrated Intelligent Fuzzy Expert System*. Computational and Mathematical Methods in Medicine, vol. 2016, 2016.

[23]. Taylor GW Fisher, MA. *A prediction model for chronic kidney disease includes periodontal disease*. Journal of Periodontology, vol. 80, 2009.

[24]. “Decision Trees - J48”, *cl.lingfil.uu.se*, 2018. [Online]. Available: http://stp.lingfil.uu.se/santinim/ml/2016/Lect03/Lab02DecisionTrees.pdf. [Accessed: 14- Oct- 2018].

[25]. “ZeroR Algorithm”, *ingsistemastelesup*.files.wordpress.com, 2018. [Online]. Available: https://ingsistemastelesup.files.wordpress.com/2017/03/zeror-algorithm.pdf. [Accessed: 14- Oct- 2018].

[26]. "Naive Bayes Classifier", *Statsoft.com*, 2018. [Online]. Available: http://www.statsoft.com/textbook/naive-bayes-classifier. [Accessed: 14- Oct- 2018].

[27]. "Hoeffding Tree ", *weka.sourceforge.net*, 2018. [Online]. Available: http://weka.sourceforge.net/doc.dev/weka/classifiers/trees/HoeffdingTree.html. [Accessed: 14- Oct- 2018]

[28]. B.G. Becker. Visualizing decision table classifiers. Proceedings IEEE Symposium on Information Visualization (Cat. No.98TB100258), 1998.

[29]. "How the random forest algorithm works in machine learning", *Dataaspirant*, 2018. [Online]. Available: http://dataaspirant.com/2017/05/22/random-forest-algorithm-machine-learing/. [Accessed: 14- Oct- 2018].