Sudan University of Science and Technology

College of Graduate Studies

Prediction of Chronic Kidney Disease Using Data Mining Techniques

توقع مرض الفشل باستخدام تقنيات التنقيب في البيانات

A dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of Master in Information Technology

Submitted by:
Rania Karamalla Ahmed Karamalla,

Supervised by:
Dr. Tallat Mohyeldin Wahbi

January 2021

الآيـة

قال تعالى:

﴿ اقْرَأْ بِاسْـمِ رَبِّكَ الَّذِي خَلَقَ *خَلَقَ الْإِنْسَـانَ مِنْ عَلَقٍ *اقْرَأْ وَرَبُّكَ الْأَكْرَمُ *الَّذِي عَلَّمَ بِالْقَلَمِ *عَلَّمَ الْإِنْسَانَ مَا لَمْ يَعْلَمْ ﴾

صدق الله العظيم

العلق (الآية 1-5)

DEDICATION

This dissertation is dedicated to:

My incredible *mam*

Dear *husband* and lovely *sons*

My dear *brother* and *sisters*

All my *family*, my *teachers*, my *friends* and *colleagues*

Acknowledgments

First of all, I would like to express my sincere gratitude to Dr. Tallat Mohyledin Wahbi for his patience and great support during all phases of the research. Also I would like to thank a lot Dr. Hisham Abdullah Mansur for his support and valuable feedback. Also nothing could be done without the perfect Data Mining lectures of Dr. Howaydah is the actual beginning of this research. I am also thankful to my sister Dr. Reem KaramAllah who gave me all knowledge about the CKD from medical perspective which I used in this research. Without her reviews and valuable discussions I couldn't make that incredible merge between data science and medicine. To conclude, I cannot forget to thank my family and friends for all the unconditional support in this very intense academic year.

ABSTRACT

Recently renal failure disease has spread widely all over the world, especially in Sudan, as indicated by the WHO reports. Therefore, it was necessary to use all available scientific methods to contribute in studying the factors that lead to the disease and predict it in its early stage, to decrease its wide spread. In this research, data mining techniques were used to study and determine the factors that lead to Chronic Kidney Disease in its early stages, and to build models to predict the disease using the selected features. Data used in this research was collected from a Medical Center for Renal Failure Treatment in India. WEKA machine learning software was used in this research for all data mining operations like data exploration, feature selection, and model development. Supervised machine learning algorithms, such as Naïve Bayes, Random Forest, C4.5 Tree and Neural Networks, were used to select the important features and develop the models. Several models were built using several algorithms, each of which gave high accuracy and acceptable interpretation to the physicians.

The research motivates other researchers to start working intensively in this field by forming research groups from data scientists and physicians to solve such problems using real patients' data.

المستخلص

في الآونة الأخيرة تفشى مرض الفشل الكلوي بصورة كبيرة جداً في العالم، وبالذات في السودان كما توضحه تقارير منظمة الصحة العالمية. فكان لابد من استعمال كافة الوسائل العلمية للتقليل من مخاطر هذا المرض بدراسة العوامل المؤثرة عليه، والتنبؤ به في مراحله الأولية. بهذا البحث تم إستخدام تقنيات التنقيب في البيانات لدراسة وتحديد العوامل المؤثرة على مرض الفشل الكلوي بمراحله الأولية، وبناء نماذج للتنبؤ بالمرض باستعمال العوامل المؤثرة عليه. حيث تم استعمال بيانات مرضى الفشل الكلوي في مراحله الأولية والتي جمعت من مركز لعلاج الفشل الكلوي بالهند. كما تمت جميع عمليات تحليل ودراسة البيانات وبناء النماذج باستعمال برنامج تعلم الآله WEKA من خلال خوارزميات التصنيف التي يتيحها البرنامج وذلك مثل Naïve Bayes, Random Forest, C4.5 Tree and Neural Networks. وقد تم بناء عدة نماذج باستعمال عدة خوارزميات حيث أعطت كل منها دقة عالية وتفسيراً مقبولاً للأطباء بما حددته الخوارزمية من العوامل المؤثرة على المرض.

البحث يفتح آفاقاً للباحثين في مجال علوم البيانات لتكوين مجموعات بحثية من مختلف التخصصات الطبية مع الباحثين في مجال علوم البيانات لإجراء مثل هذه الدراسة.

Table of Contents

List of tables

List of figure

List of Abbreviations

CKD	Chronic Kidney Disease
DM	Data Mining
HIT	Health Information Technology
HIM	Health Information Management
MLP	Multi Layer Perceptron algorithm
ESKD	End Stage Kidney Disease

CHAPTER ONE

INTRODUCTION

CHAPTER I: INTRODUCTION

1.1 Introduction

In this chapter introduction about the basic knowledge required for this research will be provided. The chapter starts by introducing the Health Informatics and Health IT, and shows the difference between these terminologies and their types. Then basic introduction about the kidney functions, and the Chronic Kidney Disease is provided. The significance of the study is then presented, after that the problem is stated and followed by the identifying the research scope and objectives. The chapter also shows the thesis organization and brief about each chapter.

1.1.1 Health Informatics

Health informatics is an evolving specialization that links information technology, communications and healthcare to improve the quality and safety of patient care. Health informatics (also called health care informatics, medical informatics, nursing informatics, clinical informatics, or biomedical informatics) is the information engineering which is applied to the field of health care, by managing patient health care information. It is a multidisciplinary field that uses health information technology (HIT) and health information systems (HIS) to improve health care via any combination of higher quality, higher efficiency and new opportunities. (Nadir et al. 2017) The disciplines involved include information science, computer science, social science, behavioral science, management science, and others. The United States National Library of Medicine (NLM) defines health informatics as "the interdisciplinary study of the design, development, adoption and application of IT-based innovations in health care services delivery, management and planning". (NICHSR, 2020)

Health informatics deals with the resources, devices, and methods required to optimize the acquisition, storage, retrieval, and use of information in health and bio-medicine. Health informatics uses different tools like computers, clinical guidelines, formal medical terminologies, and information systems. (O'donoghue & Herbert 2012)(Mettler & Raptis 2012) It is applied to many areas such as nursing, clinical medicine, dentistry, pharmacy, public health, occupational therapy, physical therapy, biomedical research, and alternative medicine. (Daniel et al. 2013) all of which are

designed to improve the overall effectiveness of patient care delivery by ensuring that the data generated is of a high quality. (Daniel et al. 2013)

1.1.2 Health Information Technology

Health IT (Health Information Technology) is the area of IT involving the design, development, creation, use and maintenance of information systems for the healthcare industry. Automated and interoperable healthcare information systems will continue to improve medical care and public health, lower costs, increase efficiency, reduce errors and improve patient satisfaction, while also optimizing reimbursement for ambulatory and inpatient healthcare providers. Today, the importance of health IT results from the combination of evolving technology and changing government policies that influence the quality of patient care. (searchhealthit, 2020)

Healthcare information systems capture, store, manage, or transmit information related to the health of individuals or the activities of an organization that work within the health sector. Figure (1.1) shows the difference between the Health informatics, Health Information Technology and Health Information Systems.

Though the concept of health IT includes the use of technology in healthcare, health informatics is not synonymous with health IT. Instead, informatics is "the science, the how and why, behind health IT," according to the Centers for Disease Control and Prevention. (usfhealthonline, 2020)

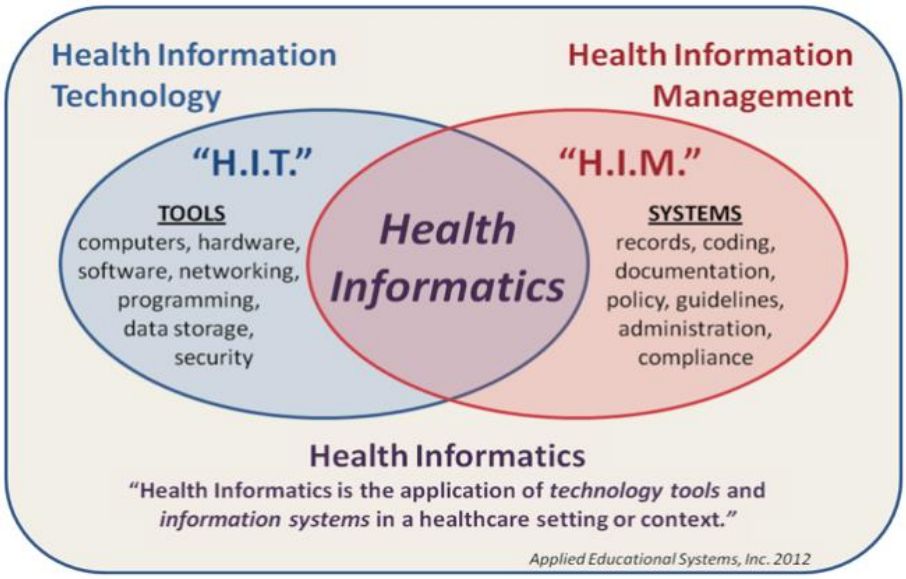

Figure 1.1 Health Informatics. (libguides, 2020)

1.1.3 Types of health information technology

The health information technology could be divided into four main categories as follows (healthit, 2020):

1. *Electronic health records* (EHRs): EHRs allow doctors to better keep track of your health information and may enable them to see it when you have a problem even if their office is closed. EHRs also make it easier for your doctor to share information with specialists, so that specialists who need your information have it available when it's needed.

2. *Personal health records* (PHRs): A PHR is a lot like an EHR, except that you control what kind of information goes into it. You can use a PHR to keep track of information from your doctor visits, but the PHR can also reflect your life outside the doctor's office and your health priorities, such as tracking what you eat, how much you exercise, and your blood pressure. Sometimes, your PHR can link with your doctor's EHR.

3. *Electronic prescribing* (E-prescribing): A paper prescription can get lost or misread. E-prescribing allows your doctor to communicate directly with your pharmacy. This means you can go to the pharmacy to pick up medicine without having to bring the paper prescription.

4. *Privacy and security*: all of these electronic systems can increase the protections of your health information. For example, electronic information can be encrypted so that only authorized people can read it. Health IT can also make it easier to record and track who has accessed your information.

Although there are many specialized hospitals and health centers of Kidney in Sudan they are working separately, no central database for patients' record. Moreover, most of them are not using computers at all. The scattered data of all these patients if collected in a single database and properly analyzed, will definitely lead to valuable results. In this paper a review will be provided about using Information Technology in health care as general, then more details will be provided about work related to implementation of data science in Kidney failure specifically.

1.1.4 Kidneys and Kidney Function

The kidneys are two bean-shaped organs in the renal system. They help the body pass waste as urine. They also help filter blood before sending it back to the heart. The kidneys perform many crucial functions, including (healthline, 2020):

- Maintaining overall fluid balance.
- Regulating and filtering minerals from blood.
- Filtering waste materials from food, medications, and toxic substances.
- Creating hormones that help produce red blood cells, promote bone health, and regulate blood pressure.

1.1.5 Chronic Kidney Disease (CKD)

Chronic Kidney Disease (CKD) means kidneys are damaged and can't filter blood the way they should. (niddk, 2020) Chronic kidney disease, also called chronic kidney failure, describes the gradual loss of kidney function. Kidneys filter wastes and excess fluids from human blood, which are excreted in urine. When chronic kidney disease reaches an advanced stage, dangerous levels of fluid, electrolytes and wastes can build up in the body. In the early stages of chronic kidney disease, few signs or symptoms may appear. Chronic kidney disease may not become apparent until kidney function is significantly impaired.

The treatment of chronic kidney disease focuses on slowing the progression of the kidney damage, usually by controlling the underlying cause. Chronic kidney disease can progress to end-stage kidney failure, which is fatal without artificial filtering (dialysis) or a kidney transplant. More details about Chronic Kidney Disease and its stages and types will be provided in chapter two the Literature Review.

1.2 Significance of Study

According to the latest WHO data published in 2017 Kidney Disease Deaths in Sudan reached 5,905 or 2.21% of total deaths. In 2016 chronic kidney disease affected 753 million individuals globally, including 417 million females and 336 million males. In 2015 it resulted in 1.2 million deaths, up from 409,000 in 1990. The causes that contribute to the greatest number of deaths are high blood pressure at 550,000, followed by diabetes at 418,000. Diagnosis of CKD is largely based on history, examination and urine dipstick combined with the measurement of the serum creatinine level. (worldlifeexpectancy, 2020)

1.3 Problem Statement

Although there are many specialized hospitals and health centers of Kidney in Sudan they are working separately, no central database for patients' record. Moreover, most of them don't use computers at all. Collection of all kidney failure patients' data in a centralized database, and proper analysis using data mining methods will definitely lead to valuable results. Although there is no data available about CKD Sudanese patients, the researcher depends on a dataset that was collected and prepared by: Dr.P.Soundarapandian.M.D.,D.M, L.Jerlin Rubini, and Dr.P.Eswaran from India. The dataset is collected and prepared for early stage chronic kidney disease patients in India.

After reviewing the collected literature, and intensive discussion with urologists, the problem could be stated as follows:

1. Can we identify more precisely the factors that lead to chromic kidney disease by analyzing the laboratory test results of patients?
2. Can we early predict the Kidney failure using the laboratory test results of CKD patients?
3. Can we provide better understanding about these factors and the relations between them, and their effect on the disease?

1.4 Objectives of Study

The goal of this research is to precisely determine the factors that lead to kidney failure disease, and provide better understanding about the relations between these factors and the disease using Data mining algorithms. To achieve this goal, the following objectives have been specified:

1. Use feature selection methods to identify the factors (features) that lead to chromic kidney disease by analyzing the laboratory test results of patients.
2. Develop data mining model for early prediction of Chronic Kidney Disease using the selected set of features.
3. Provide better understanding for urologists about the relation between the laboratory test results and the Chronic Kidney Disease.

1.5 Thesis Organization

This thesis is organized into six chapters to achieve the research objectives, and depict the methodologies followed to achieve these goals. These six chapters are organized as follows:

Chapter 1, Introduction: this chapter provides basic understanding about the Chronic Kidney Disease, and terminologies used in the research. The chapter also shows the significance of the study, then the problem is clearly stated. According to the stated problems, the research objective is defined.

Chapter 2, Literature Review: this chapter presents a comprehensive review about all areas related to this research. The chapter starts by giving a brief introduction about data mining and the process which will be followed in this research. Then introduction about the algorithms that are used in this research is also provided. Then essential knowledge about Chronic Kidney Disease is provided. After that various applications, techniques and researches about data mining in health informatics is provided. The last part is focusing in the related work about using data mining methods to study Chronic Kidney Disease from various perspectives. Then summary of all literature reviewed throughout this chapter is summarizes in a simple and clear tables.

Chapter 3, Research Methodology: this chapter describes the methodologies used and steps followed to achieve the research objectives. The chapter also provides more details about steps the dataset. Also the chapter shows the data mining process which is used in the research from the problem definition to model deployment.

Chapter 4, Data exploration: this chapter provides more description about the dataset. Then each attribute is studied separately, to provide the relation between the attribute and the class CKD. Also basic statistics about these attributes is provided to show initial findings and results.

Chapter 5, Classification Models: this chapter provides the details of experiments developed for feature selection and classification models. Also the chapter provides the evaluation of each model. The most important part of the chapter is the feedback and discussion of results from urologists, which represents a real health informatics model.

Chapter 6, Conclusion: this chapter presents the research conclusion by showing the discussion of results from both IT and Medicine perspectives. The chapter also provides suggestions and recommendations for future work.

CHAPTER TWO

LITERATURE REVIEW

CHAPTER II: LITERATURE REVIEW AND RELATED WORK

2.1 Introduction

Data mining techniques and their applications have developed rapidly during the last two decades. This chapter provides a review of the application of data mining techniques in big data, especially in health care, through a survey of literature from 2000 to 2020. Keyword indices and article abstracts and conclusions were used to classify more than 8 articles, from many academic journals and research centers. Because this research concerns about application of data mining in big data, this review started by providing a brief introduction about data mining and big data to give clear vision about these two different disciplines. From this review we found that data mining could be used to solve many types of problems in big data in health care, like prediction of CKD kidney failure and many others. Also there is no standard technique that could be used for a specific problem. Application of data mining is a huge research area and still need more researches.

2.2 Introduction to Data mining

This section is divided into three parts, the first one gives an introduction of data mining, the second is about Research Roadmap which is a complete blueprint for conducting a data mining project, the third is about details of some classification algorithms.

2.2.1 What is Data Mining?

Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. (kdd, 2020)

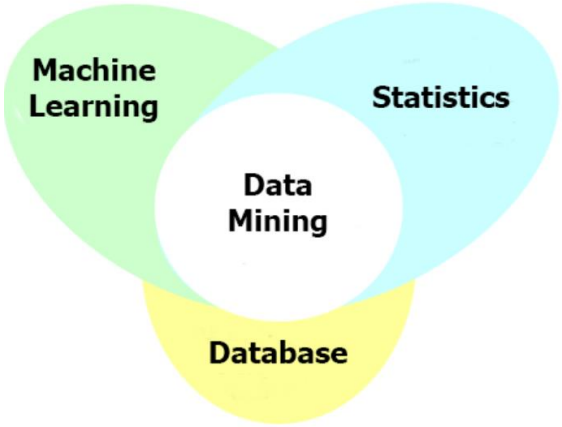

Figure 2.1 Data Mining Intersection. (researchgate, 2020)

Another definition of data mining is the process of finding anomalies, patterns and correlations within large data sets to predict outcomes. Using a broad range of techniques, you can use this information to increase revenues, cut costs, improve customer relationships, reduce risks and more. (sas, 2020)

2.2.3 Some Classification Algorithms

2.2.3.1 Random forests

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees as shown in figure 2.2. Random decision forests correct for decision trees' habit of overfitting to their training set. The first algorithm for random decision forests was created by Tin Kam Ho. (Ho & Tin Kam 1995)

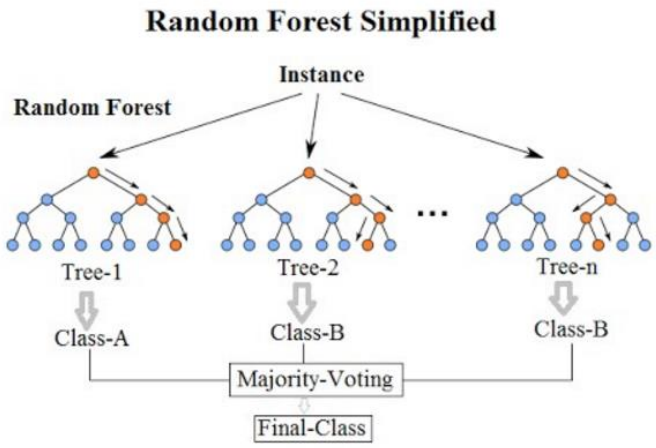

Figure 2.3 Random Forest Algorithm Diagram. (researchgate, 2020)

2.2.3.2 J48

C4.5 (J48) is an algorithm used to generate a decision tree developed by Ross Quinlan mentioned earlier. C4.5 is an extension of Quinlan's earlier ID3 algorithm. (researchgate, 2020)

Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. Decision trees learn from data to approximate a sine curve with a set of if-then-else decision rules. The deeper the tree, the more complex the decision rules and the fitter the model. Decision tree builds classification or regression models in the form of a tree structure. It breaks down a data set into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed to nodes and leaf nodes. A decision node has two or more branches. Leaf node represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data. (researchgate, 2020)

2.2.3.3 Bayes Net

Bayesian networks are a type of Probabilistic Graphical Model that can be used to build models from data and/or expert opinion. They can be used for a wide range of tasks including prediction, anomaly detection, diagnostics, automated insight, reasoning, time series prediction and decision making under uncertainty. (bayesserver, 2020)

Bayesian networks are ideal for taking an event that occurred and predicting the likelihood that any one of several possible known causes was the contributing factor. For example, a Bayesian network could represent the probabilistic relationships between diseases and symptoms. Given symptoms, the network can be used to compute the probabilities of the presence of various diseases. (Pearl & Judea 2000)

2.2.3.4 Trees.LMT

A logistic model tree (LMT) is a classification model with an associated supervised training algorithm that combines logistic regression (LR) and decision tree learning. (Niels & Mark & Eibe 2003)

Logistic model trees are based on the earlier idea of a model tree: a decision tree that has linear regression models at its leaves to provide a piecewise linear regression model (where ordinary decision trees with constants at their leaves would produce a piecewise constant model).
In the logistic variant, the LogitBoost algorithm is used to produce an LR model at every node in the tree; the node is then split using the C4.5 criterions. Each LogitBoost invocation is warm-started[vague] from its results in the parent node. Finally, the tree is pruned.
The basic LMT induction algorithm uses cross-validation to find a number of LogitBoost iterations that does not overfit the training data. (Sumner et al. 2005)

2.2.3.5 Simple Logistec

2.2.3.5.1 What is Logistic Regression?

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is binary. Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables. (statisticssolutions, 2020)

2.2.3.5.2 Logistic VS. Linear Regression

Logistic regression gives you a discrete outcome but linear regression gives a continuous outcome. A good example of a continuous outcome would be a model that predicts the value of a house. That value will always be different based on parameters like it's size or location. A discrete outcome will always be one thing (you have cancer) or another (you have no cancer).

Logistic regression uses the concept of predictive modeling as regression; therefore, it is called logistic regression, but is used to classify samples; Therefore, it falls under the classification algorithm.

Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems. (javatpoint, 2020)

2.2.3.6 MLP

Artificial neural networks (ANN) or connectionist systems are computing systems vaguely inspired by the biological neural networks that constitute animal brains. (Chen et al. 2019)

A multilayer perceptron (MLP) is a deep, artificial neural network. It is composed of more than one perceptron. They are composed of an input layer to receive the signal, an output layer that makes a decision or prediction about the input, and in between those two, an arbitrary number of hidden layers that are the true computational engine of the MLP. MLPs with one hidden layer are capable of approximating any continuous function.

Multilayer perceptron is often applied to supervised learning problems, they train on a set of input-output pairs and learn to model the correlation (or dependencies) between those inputs and outputs. Training involves adjusting the parameters, or the weights and biases, of the model in order to minimize error. Backpropagation is used to make those weigh and bias adjustments relative to the

error, and the error itself can be measured in a variety of ways, including by root mean squared error (RMSE).

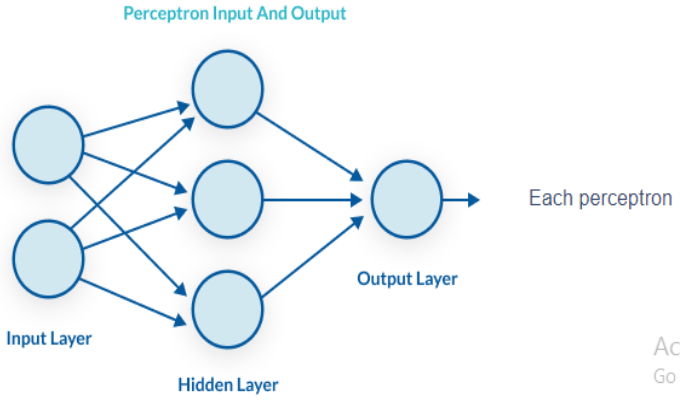

Figure 2.4 Multilayer perceptron Diagram. (linkedin, 2020)

In machine learning, the perceptron is an algorithm for supervised learning of binary classifiers. A binary classifier is a function which can decide whether or not an input, represented by a vector of numbers, belongs to some specific class. (Freund, Y & Schapire 1999)

2.3 Chronic Kidney Disease (CKD)

Chronic Kidney Disease (CKD) means kidneys are damaged and can't filter blood the way they should. (niddk, 2020) Chronic kidney disease, also called chronic kidney failure, describes the gradual loss of kidney function. Kidneys filter wastes and excess fluids from human blood, which are excreted in urine. When chronic kidney disease reaches an advanced stage, dangerous levels of fluid, electrolytes and wastes can build up in the body. In the early stages of chronic kidney disease, few signs or symptoms may appear. Chronic kidney disease may not become apparent until kidney function is significantly impaired.

The treatment of chronic kidney disease focuses on slowing the progression of the kidney damage, usually by controlling the underlying cause. Chronic kidney disease can progress to end-stage kidney failure, which is fatal without artificial filtering (dialysis) or a kidney transplant.

2.3.1 Chronic Kidney Disease (CKD) Stages

There are five stages of CKD—based on two GFR tests at least 90 days apart. Stages 1 and 2 occur only in people whose kidneys are not normal. They may have been born with just one kidney. They may have kidney cysts, urine that backs up into the kidneys, or protein in their urine. *Most people find out they have CKD at stage 3, 4, or 5*. In the early stages, the risk of heart disease is higher than the risk of kidney failure. So, to feel your best, you need to protect your kidneys *and* your heart!

You may have kidney disease and not feel it! A GFR is a test of how well your kidneys work.

GFR is glomerular filtration rate. Glomeruli are kidney filters that clean water and wastes out of your blood. A GFR is about the same as your percent kidney function. So, a GFR of 58 means you have about 58% function.

The GFR equation for calculating CKD grades contains the patient's age, gender, race, and the result of a blood serum creatinine test. (lifeoptions, 2020)

Figure 2.5 CKD Stages. (lifeoptions, 2020)

2.3.2 Causes of CKD

Chronic kidney disease occurs when a disease or condition impairs kidney function, causing kidney damage to worsen over several months or years.

Diseases and conditions that cause chronic kidney disease include: (lifeoptions, 2020) (mayoclinic, 2020)

- Type 1 or type 2 diabetes.

- High blood pressure.
- Glomerulonephritis an inflammation of the kidney's filtering units (glomeruli).
- Interstitial nephritis, an inflammation of the kidney's tubules and surrounding structures.
- Polycystic kidney disease.
- Prolonged obstruction of the urinary tract, from conditions such as enlarged prostate, kidney stones and some cancers.
- Vesicoureteral reflux, a condition that causes urine to back up into your kidneys.
- Recurrent kidney infection, also called pyelonephritis.

2.3.3 Risk factors

Factors that may increase your risk of chronic kidney disease include:

- Diabetes.
- High blood pressure.
- Heart and blood vessel (cardiovascular) disease.
- Smoking.
- Obesity.
- Being African-American, Native American or Asian-American.
- Family history of kidney disease.
- Abnormal kidney structure.
- Older age.

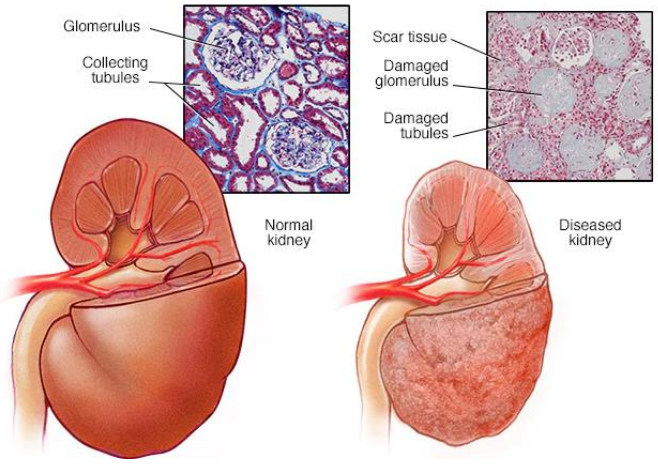

Figure 2.6 Normal kidney vs. diseased kidney. (mayoclinic, 2020)

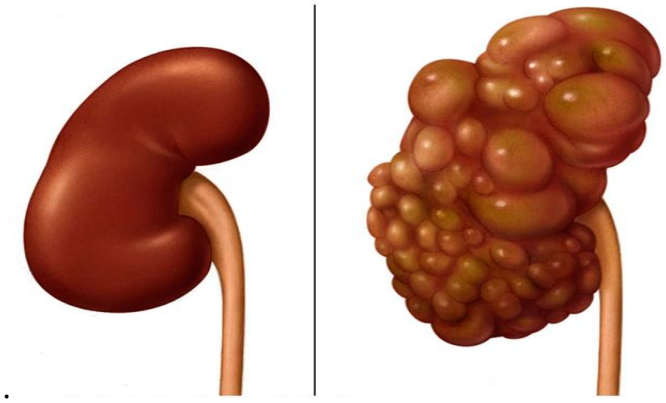

Figure 2.7 Polycystic kidney. (mayoclinic, 2020)

Mohamed Elhafiz Elsharif and Elham Gariballa outline causes of end-stage renal disease in Sudan in a single-center experience. The study was conducted in May 2009. The population examined here consisted of 224 patients on regular HD in Gezira Hospital for renal disease. They found that the etiologies were dominated by unknown causes (53.57%). The leading cause of ESRD for those who were younger than 40 years was glomerular disease (glomeruli, the tiny units within the kidney where blood is cleaned. When they damaged they causes glomerular disease), hypertension for those between 40 and 60 years and obstruction for those who were older than 60 years. (Mohamed E & Elham G 2011)

Hans-Joachim Anders et al. provide recommendations for the management of patients with immune-mediated kidney diseases based on the available evidence, similar circumstances with other infectious organisms and expert opinions. Such recommendations may help to minimize Severe acute respiratory syndrome coronavirus 2 (SARSCoV-2) is a new member of the coronavirus family, first described in the city of Wuhan, China in December 2019. (Hans-Joachim A et al)

2.4 Health Informatics

Applications of data mining methods in the Health informatics

Data mining could be used to solve many types of problems in Health Informatics, like prediction of disease, prediction of patient's health status, diagnosis of patient and many others. There is no standard technique that could be used for a specific problem. Signal processing, image

processing, and data mining tools have been developed for effective analysis of medical information, in order to help clinicians in making better diagnosis for treatment purposes.
In the literature reviewed, a lot of researchers provide many interesting articles.

Daniele et al. presented deep learning for health informatics based on a knowledge discovery in databases, supported by data mining technique (NN) neural network. They mainly focus on key applications of deep learning in the fields of translational bioinformatics, medical imaging, pervasive sensing, medical informatics, and public health. They have outlined how deep learning has enabled the development of more data-driven solutions in health informatics by allowing automatic generation of features that reduce the amount of human intervention in this process. (Daniele R et al. 2017)

Claudio and Alexis have developed an expert system based on rules for the pre-diagnosis of hypertension, diabetes mellitus type 2 (DMT2) and metabolic syndrome (MS), diseases that are considered NCD. The expert system offers to patients a tool for self-evaluation and to create awareness of their pathologies; this system is developed as a complementary and supportive tool, and in no case intends to replace the diagnosis of a physician.
The user interface for data collection is done through a web application which involves the use of Apache, as a Web server, MySQL as a database and PHP as a language for the implementation of the web system. (Claudio & Alexis 2020)

J. Jeffery Reeves et al. outlined the design and implementation of EHR-based rapid screening processes, laboratory testing, clinical decision support, reporting tools, and patient-facing technology related to COVID-19. An Electronic Health Record (EHR) is an electronic version of a patient's medical history. (cms, 2020) They concluded that the EHR is an essential tool in supporting the clinical needs of a health system managing the COVID-19 pandemic. (Jeffery, J et al. 2020)

2.5 Data mining for CKD

How Data mining is Useful for CKD

Data mining and analytics techniques can be used for predicting CKD (Chronic Kidney Disease) by utilizing historical patient's data and diagnosis records.

Shaik MD et al, study identifies the big data analytics for health on kidney disease. This research involves in providing a detailed study of segmenting and detection of tumor based on computed

tomography images. Many CT image segmentation algorithms are available; it can be divided into edge, texture, and -. The main thing in this work is to determine the kidney tumor by applying the algorithm on the test result acquired from the patient medical report and segmentation of the tumor location. They analyze images of tumor however they did not analyze the patient's complete record data that did not give us a chance to devise other factors that could diagnose the disease. (Shaik, MD et al. 2017)

Tabassum et al, study identifies analysis and prediction of chronic kidney disease by using data mining techniques, researchers have the scope to predict the Chronic Kidney Disease. This helps doctors to diagnose and suggest the treatment at an early stage. It also helps the patients to know about their health condition at an earlier stage and follow necessary diet and prescriptions. However, they did not explain the most important factors that lead to kidney failure from patient record they compare between different accuracies of algorithms.

Future suggestions analyze all the features of the patient's registry to determine which ones are most effective and which lead to the disease**.** (Tabassum et al. 2017)

Basma et al, study identifies the performance of data mining techniques to predict in healthcare case study: chronic kidney failure disease they presented an overview on the evolution of big data in healthcare system, and they applied three learning algorithms on a set of medical data. The objective of this research work is to predict kidney disease by using multiple machine learning algorithms that are Support Vector Machine (SVM), Decision Tree (C4.5), and Bayesian Network (BN), and chose the most efficient one. They used several learning algorithm C4.5, SVM and NB, to predict patients with chronic kidney failure disease (ckd), and patients who are not suffering from this disease (notckd). Simulation results showed that C4.5 classifier proved its performance in predicting with best results in terms of accuracy and minimum execution time. (Basma et al. 2016)

Table 2.1 Summarized of the papers

Author	Title	Year	Topic	Methodology/ Tool	Discussion
Mohamed Elhafiz Elsharif	Causes of end-stage renal disease in Sudan: A single-center experience	Apr-17	Kidney Disease	Medicine research.	This paper used to provide medical back ground about the research.
Hans-Joachim Anders et al.	Recommendations for the management of patients with immune-mediated kidney disease during the severe acute respiratory syndrome coronavirus 2 pandemic	May-20	Kidney Disease and coronavirus 2	Medicine research.	This paper used to provide medical back ground about CKD relation with Corona virus.
Daniele Rav`I et al.	Deep Learning for Health Informatics	Jan-17	Health Informative	(NN) neural network	Only NN was used to analyze data, if other algorithms were used, more understanding could be gained.
Claudio Urrea and Alexis Mignogna	Development of an expert system for pre-diagnosis of hypertension, diabetes mellitus type 2 and metabolic syndrome	2020	Health Informative	Apache, as a Web server, MySQL as a database and PHP as a language for the implementation of the web system	The dataset contains data for 72 patients only. If the dataset is bigger enough better results could be achieved.
J. Jeffery Reeves et al.	Rapid response to COVID-19: health informatics support for outbreak management in an academic health system	Feb-2020	Health Informative	(EHR) is an electronic version of a patient's medical history.	Although the EHR was mentioned, the research did not mention which language or data base used.
Shaik MD et al	Big Data Analytics for Health on Kidney Disease	April - June 17	Datamining of CKD	Fuzzy C and For feature extraction GLCM and for classification of normal images and Tumor images SVM algorithm	Image processing was done to classify the Tumor image.
Tabassum et al	Analysis and Prediction of Chronic Kidney Disease using Data Mining Techniques	Sep - 17	Datamining of CKD	Artificial Neural Network (ANN), EM, and (C4.5).	Three algorithms were used. However, no medical discussion provided to link the achieved results with the medicine rules.
Basma Boukenze, H.M.a.A.H.,	Performance Of Data Mining Techniques To Predict In Healthcare Case Study: Chronic Kidney Failure Disease	Jun - 16	Datamining of CKD	Support Vector Machine (SVM), Decision Tree (C4.5), and Bayesian Network (BN)	High accuracy and good performance were achieved, however feature selection was not used, moreover no discussion from domain expert.

CHAPTER THREE

RESEARCH METHODOLOGY

CHAPTER III: METHODOLOGY

3.1 Introduction

This chapter describes the methodology that will be followed to achieve research objectives. The methodology is composed of several steps starting by problem definition up to model deployment. The first research objective is to determine the factors that lead to chronic Kidney Disease. While the second objective is to design a prediction model to give the doctors better understanding about the relations between these factors and assist in patient classification. The research uses: Early stage of Indians Chronic Kidney Disease (ckd) dataset.

This chapter is composed of eight parts, the first one is introduction. The second part shows the research road map. The third part defines the problem, while the fourth and fifth parts presents the data preparation and data exploration tasks. Then on the sixth part the methodology of modeling will be shown. After that the information about models' evaluation will be provided. The last section is about models' deployment and how these results will be discussed with the domain experts.

3.2 Research Roadmap

The research roadmap is an organized steps that guide the researcher to achieve the objectives. The researcher depends on the Data Mining Map which is proposed by Dr. Saed Sayad in his web site. (saedsayad, 2020) Figure 3.1 shows the original data mining map which was proposed by Dr. Saed Sayed.

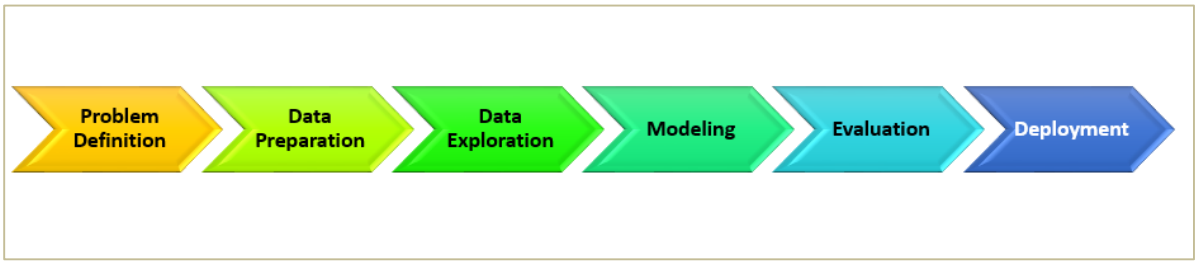

Figure 3.1 Data Mining Map (saedsayad, 2020)

The research road map is created by merging *Dr. Saed Sayed data mining map*, to the *special research contents* like the literature review and initial comparison between algorithms. Figure 3.2 shows the road map for this research.

The research road map is composed of seven phases:

1. Problem Definition.
2. Data Preparation.
3. Data Exploration.
4. Modeling.
5. Evaluation.
6. Deployment.

Next section provides more details about each phase.

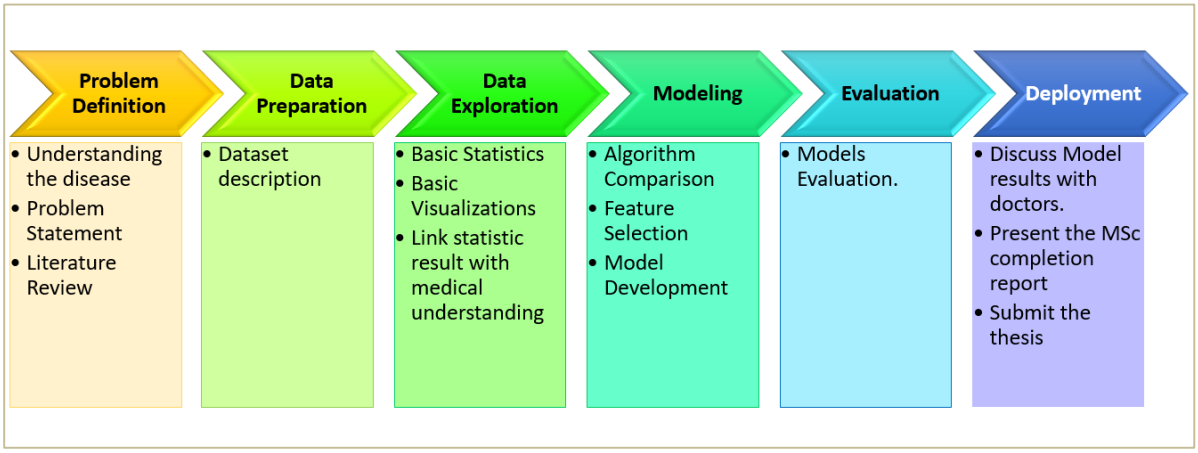

Figure 3.2 Research Road Map (sas, 2020)

3.4 Data Preparation

The data set used for this research is patients' data during their early stage of Chronic Kidney Disease. The data set is composed of 25 attributes, 24 of them are independent attributes describing patient's status like ***age***, or a laboratory results of the patient, the last attribute is the class which describes the patient's status whether he or she

is suffering from the disease (***ckd***), or not (***notckd***). The dataset is composed of 400 instances. This data is used because it is the most appropriate data found about the topic, because no suitable data was found in the local hospitals and kidney treatment centers.

3.5 Data Exploration

Any data mining research should start by data exploration phase which will provide basic understanding about the dataset, which provides preliminary results. In some cases, these initial results provide much information and useful findings. In chapter four Data Exploration, basic statistics and visualization about each field is provided. Also in that chapter a brief link between the statistics observation and their medical interpretation is provided.

3.6 Modeling

This is the main part of the research in which all research questions will be answered by conducting the appropriate experiments, and running the selected algorithms. As shown in figure 3.2 the research road map, the modeling phase contains three main tasks: algorithm comparison, feature selection and model development. In the next sections more details about each step will be provided.

3.6.1 Algorithms Comparison

Many algorithms are available in Weka, each algorithm can work in specific data mining category (classification, clustering, association or regression). Even in the same category some algorithm can provide better results than others according to the dataset. So the best way to determine the most suitable algorithm is by testing all available algorithms, then compare between their results to select the most suitable one for our research. This comparison was done using Weka Experimenter, by running 25 algorithms for our dataset. More details about how to do the experiment and compare between results is provided in chapter five: The feature Selection and Model development. Table 3.1 shows the list of available algorithms that will be used in the experiment.

3.6.2 Attribute Selection

Because of the negative effect of irrelevant attributes on most machine learning schemes, it is common to precede learning with an attribute selection stage that selects the most relevant attributes. The best way to select relevant attributes is manually, based on an experts' decision about the problem, and what the attributes actually mean. However, automatic methods can also be useful. Reducing the dimensionality of the data by deleting irrelevant attributes improves the performance of learning algorithms, it also speeds them up. More over dimensionality reduction yields a more compact, more easily interpretable representation of the target concept, focusing the user's attention on the most relevant variables. The attribute selection and classification models are done in one step in this research. The result of each attribute selection process is a set of attributes, according to the classification model only these selected attributes were found to have relation with the class.

Table 3.1 Available Algorithms in Weka

#	Algorithm
1	bayes.NaiveBayesMultinomialText
2	functions.SGDText
3	trees.RandomForest
4	trees.J48
5	bayes.BayesNet
6	trees.LMT
7	functions.SimpleLogistic
8	meta.AdaBoostM2
9	rules.PART
10	meta.FilteredClassifier
11	functions.SMO
12	rules.DecisionTable
13	trees.RandomTree
14	trees.REPTree
15	rules.JRip
16	lazy.IBk
17	trees.HoeffdingTree
18	bayes.NaiveBayes
19	bayes.NaiveBayesUpdateable
20	trees.DecisionStump
21	lazy.KStar
22	rules.OneR
23	functions.VotedPerceptron
24	misc.InputMappedClassifier
25	rules.ZeroR

3.6.3 Classification Models

Because the class field is available in the dataset, this will be supervised learning, and because the class field is categorical that contains two distinct values (ckd, notckd) the model will be classification model. As stated in the previous section the attribute selection and model development will be done in one step using the Wrapper Method. The classification model starts by identifying the selected attributes, then this newly composed set will be used to develop the classification model. The model will provide better understanding about the important features and their relation with the class. These results will then be presented to urologists to cross check and more discussion.

5.5.1 Introduction

Because of the negative effect of irrelevant attributes on most machine learning schemes, it is common to precede learning with an attribute selection stage that strives to eliminate all but the most relevant attributes. The best way to select relevant attributes is manually, based on a deep understanding of the learning problem and what the attributes actually mean. However, automatic methods can also be useful. Reducing the dimensionality of the data by deleting irrelevant attributes improves the performance of learning algorithms. It also speeds them up, although this may be outweighed by the computation involved in attribute selection. More important, dimensionality reduction yields a more compact, more easily interpretable representation of the target concept, focusing the user's attention on the most relevant variables.

When selecting a good attribute subset, there are two fundamentally different approaches:

(1) **Filter method**: this method makes an **independent** assessment based on general characteristics of the data. It is called the filter method because **the attribute set is filtered** to produce the most promising subset before learning commences. Filter type methods select variables regardless of the model. They are based only on general features like the correlation with the variable to predict. Filter methods suppress the least interesting variables. The other variables will be part of a classification or a regression model used to classify or to predict data. These methods are particularly effective in computation time and robust to over fitting. However, filter methods tend to select redundant variables because they do not consider the relationships between variables. Therefore, they are mainly used as a pre-process method.

(2) **Wrapper method**: this method evaluates the subset using the machine learning algorithm that will ultimately be employed for learning. It is called *wrapper* method because the **learning algorithm is wrapped into the selection procedure**. Wrapper methods evaluate subsets of variables which allows, unlike filter approaches, to detect the possible interactions between variables. That means the feature selection and model development are done in the same step.

In this research the second approach will be used because it gives better accuracy, and the two steps – feature selection and model development - are done in one experiment.

3.7 Evaluation

After each model there will be evaluation section that evaluates the model accuracy by comparing the number of correctly and incorrectly classified instances, and the percentage of each. Then a confusion matrix will be presented to show the model accuracy. For testing 10-fold cross validation will be used because the number of instances is not that much. Also the precision and recall results for each model will be discussed.

3.8 Deployment

The results achieved and models developed will then be presented to urologists. Their feedback and comments will be presented in the last chapter of this research Results Discussion and Future Works.

CHAPTER FOUR

DATA EXPLORATION

CHAPTER IV: DATA EXPLORATION

4.1 Introduction

Data Science is about explaining the past to predict the future, by means of data analysis. (saedsayad, 2020) The goal of this chapter is to provide data exploration because it is the first step in any data mining project which gives basic understanding about the problem and the dataset. Many questions could be answered like how much data dose we have? What are the available features? What can we understand from the basic statistics like means, standard deviation, minimum and maximum values? Even we can start getting some basic knowledge behind these numbers. That is very clear when we noticed that the percentage of patients who are suffering from the disease have high Blood Glucose Random (***bgr***), and high Sugar (***su***).

This chapter starts by reviewing the basic knowledge about data visualization and exploration. After that some information about the dataset itself is shown, like the data source and the number of attributes and instance. Then a brief summary about each attribute is provided. The last section of the chapter present basic statistics about each attribute, and its relation with the ***class*** attribute.

4.2 Basic Concepts about Data Exploration

Data Exploration is about describing the data by means of statistical and visualization techniques to be able to predict the future. Data is explored in order to get better understanding about the problem. As shown in figure 4.1 data exploration could be divided into two types Univariate Analysis, and Bivariate Analysis, below in this introduction brief about each type will be provided. (saedsayad, 2020)

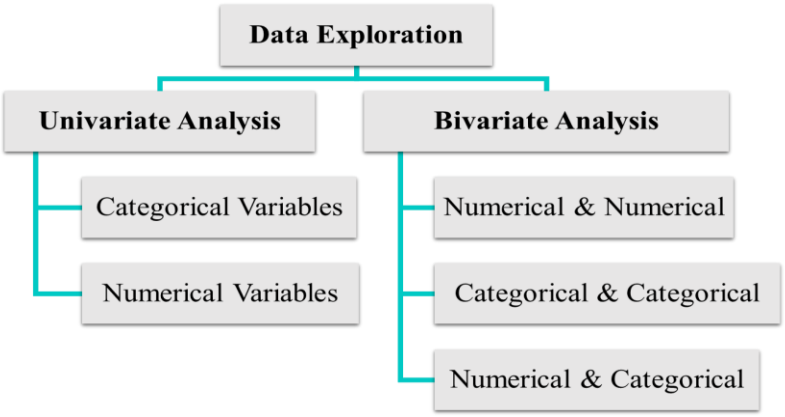

Figure 4.1 Data Exploration Types (saedsayad, 2020)

(1) **Univariate Analysis**

Univariate analysis explores features one by one. Attributes could be either *categorical* or *numerical*. There are different suitable statistical and visualization techniques to explore each type of variables. If needed numerical variables can be transformed to categorical by a process called binning or discretization. The vice versa process is encoding. Finally, proper handling of missing values is an important issue in mining data. Below is some more information about how to explore the categorical and numerical variables.

i. Categorical Variables

A categorical or discrete variable is one that has two or more values. There are two types of categorical variable, **nominal** and **ordinal**. A nominal variable has no real ordering to its categories. For example, gender is a categorical variable having two categories (male and female) with no real ordering to the categories. An ordinal variable has a clear ordering. For example, temperature as a variable with three *orderly* categories (low, medium and high). The **frequency table** is a way of counting how often each category of the variable in question occurs. It may be enhanced by the addition of percentages that fall into each category. Pie chart is one of the most suitable visualizations that could be used with categorical data as shown in figure 4.2.

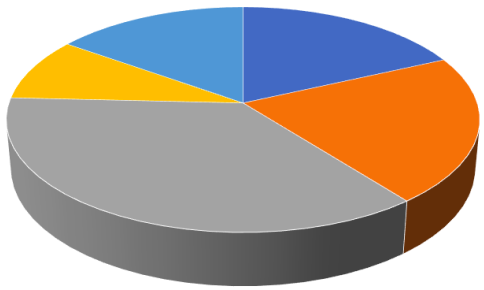

Figure 4.2 Pie Chart (saedsayad, 2020)

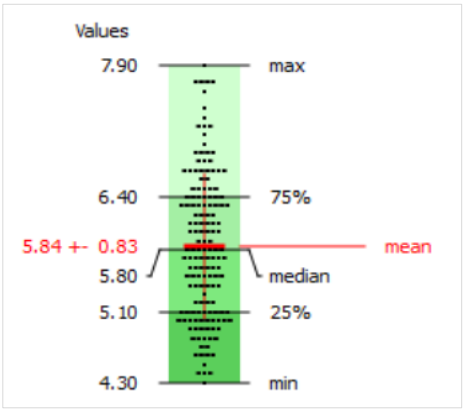

Figure 4.3 Box Plot (saedsayad, 2020)

i. Numerical Variables

The numerical or continuous variable is one that may take on any value within a finite or infinite interval (e.g., blood pressure, weight, temperature, blood glucose .etc.). There are two types of numerical variables, **interval** and **ratio**. An interval variable has values whose differences are interpretable, but it does not have a true zero. A good example is temperature in Centigrade degrees. Data on an interval scale can be added and subtracted but cannot be meaningfully multiplied or divided. For example, we cannot say that one day is twice as hot as another day. In contrast, a ratio variable has values with a true zero and can be added, subtracted, multiplied or divided (e.g., weight). Box plot chart is one of the most suitable visualizations that could be used with numerical data as shown in figure 4.3.

(2) **Bivariate Analysis**

Bivariate analysis is the simultaneous analysis of two variables (attributes). It explores the concept of relationship between two variables, whether there exists an association and the strength of this association, or whether there are differences between two variables and the significance of these differences. There are three types of bivariate analysis:

i. Numerical & Numerical
ii. Categorical & Categorical
iii. Numerical & Categorical

i. Numerical & Numerical

A scatter plot is a useful visual representation of the relationship between two numerical variables (attributes) and is usually drawn before working out a linear correlation or fitting a regression line. The resulting pattern indicates the type (linear or non-linear) and strength of the relationship between two variables. More information can be added to a two-dimensional scatter plot, for example, we might label points with a code to indicate the level of a third variable. If we are dealing with many variables in a data set, a way of presenting all possible scatter plots of two variables at a time is in a *scatter plot matrix*. Scatter plot chart is one of the most suitable visualizations that could be used with numerical data as shown in figure (4.4)

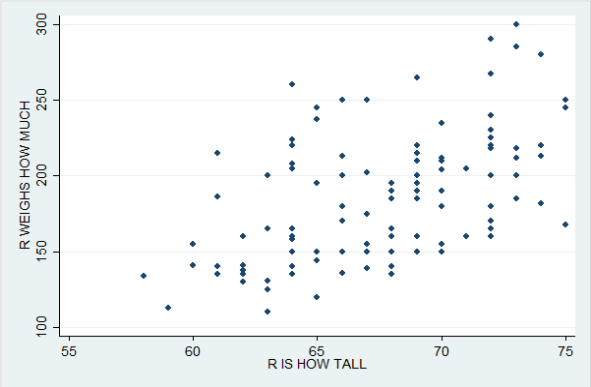

Figure 4.4 Scatter Plot Chart (saedsayad,

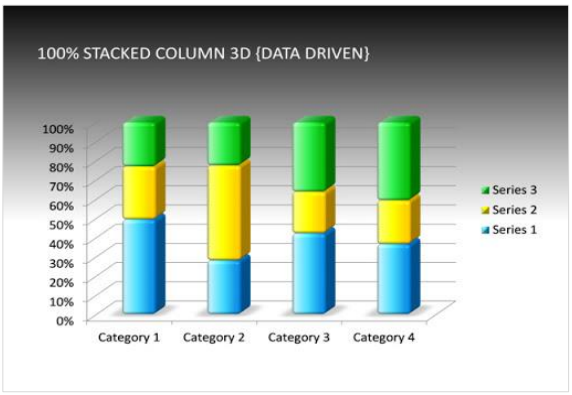

Figure 4.5 Stacked Column (saedsayad,

ii. Categorical & Categorical

As shown in figure (4.5) stacked Column chart is a useful graph to visualize the relationship between two categorical variables. It compares the percentage that each category from one variable contributes to a total across categories of the second variable.

iii. Numerical & Categorical

A line chart with error bars displays information as a series of data points connected by straight line segments. Each data point is average of the numerical data for the corresponding category of the categorical variable with error bar showing standard error. It is a way to summarize how pieces of information are related and how they vary depending on one another. Figure 4.6 shows a sample of line chart.

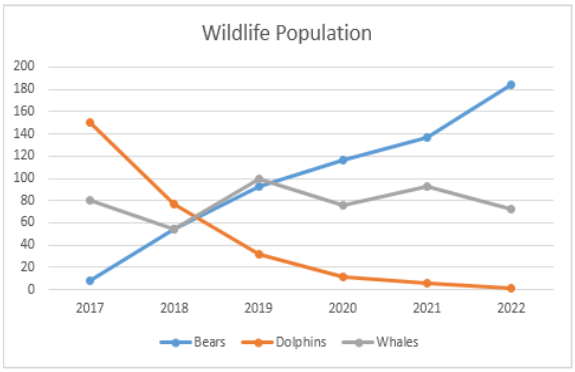

Figure 4.6 Line Chart (saedsayad, 2020)

4.3 Dataset Description

The dataset used in this research was downloaded from (uci, 2020). The data was collected for 400 patients, 250 of them were diagnosed to be in their early stage of Chronic Kidney Disease (***ckd***). While the rest 150 patients are not infected. The data composed of 25 attributes, some of them are numerical and the others are categorical. The class field - which is our target in the proposed classification model – contains the medical status of the patient, whether he is suffering from Chronic Kidney Disease (***ckd***) or not (***notckd***). There is some data missing in different fields, many techniques could be used to deal with it, however the researcher preferred to leave it as is and depends on the classification algorithms to deal with those missing items. Some other researchers deleted the records that contains the missing data, but this may be unrealistic and may lead to some sort of over fitting.

The dataset was prepared by Dr.P.Soundarapandian.M.D.,D.M (Senior Consultant Nephrologist) in India, L.Jerlin Rubini(Research Scholar) from Alagappa University, and Dr.P.Eswaran Assistant Professor from Alagappa University –India.

In the next section brief description about the attributes of the dataset will be provided.

4.3 Dataset Summary

Table 4.1 shows all attributes of: ***The Early Stage of Indians Chronic Kidney Disease*** dataset. The first column (#) in the table shows the attribute ID, it is just a serial number that uniquely identifies the attribute in this research. The second column (Attribute Name) is the name of the attribute in the dataset. The third one is a brief description about this attribute. The last column is the name of the attribute in Arabic, to give the reader a clue about the attribute meaning and its effect on the disease.

The next section provides detailed description and basic statistical analysis about each one of these attributes.

Table 4.1 Summary of CKD dataset attributes

#	Attribut	Type	Description	Arabic Name
1	age	numerical	Patient's age.	العمر
2	bp	numerical	Blodd Pressure. Normal range between 80 - 90	قياس ضغط الدم
3	sg	nominal	specific gravity(a measure of the concentration of solutes in the urine)	الثقل النوعي
4	al	nominal	Albumin -(albumin in blood) (0,1,2,3,4,5)	المنيوم
5	su	nominal	Sugar(blood sugar level)- (0,1,2,3,4,5)	تصنيف جلكوز الدم
6	rbc	nominal	Red Blood Cells-(a red blood cell count) (normal,abnormal)	نوع كريات الدم الحمراء
7	pc	nominal	Pus Cell - (normal,abnormal)	خلية صديد
8	pcc	nominal	Pus Cell clumps - (present,notpresent)	كتل الخلايا الصديدية
9	ba	nominal	Bacteria - (present,notpresent)	بكتريا
10	bgr	numerical	Blood Glucose Random(blood Random glucose testing)	سكر الدم العشوائي
11	bu	numerical	Blood Urea(the amount of urea nitrogen in the blood)	يوريا الدم
12	sc	numerical	Serum Creatinine(measures the level of creatinine in blood)	كرياتينين
13	sod	numerical	Sodium(blood test)	صوديوم
14	pot	numerical	Potassium(blood test)	بوتاسيوم
15	hemo	numerical	Hemoglobin(measures how much hemoglobin red blood cells contain.)	هيموقلوبين
16	pcv	numerical	Packed Cell Volume(a part of the full blood count test)	حجم الخلية المعبأة في ا
17	wbcc	numerical	White Blood Cell Count	كريات الدم البيضاء
18	rbcc	numerical	Red Blood Cell Count	عدد كريات الدم الحمراء
19	htn	nominal	Hypertension -does patient has hypertension or not (yes,no)	مرض ضغط الدم
20	dm	nominal	Diabetes Mellitus -(des patient has diabetes or not) (yes,no)	مرض السكر
21	cad	nominal	Coronary Artery Disease (Does the patient has coronary artery disease) (y,n)	مرض القلب التاجي
22	appet	nominal	Patint's appetite - (good,poor)	الشهية
23	pe	nominal	Pedal Edema - (Does patient has pedal edema or not)(yes,no)	تورم
24	ane	nominal	Anemia - (Does patient has anemia or not)(yes,no)	فقر الدم
25	class	nominal	class - (Does the patient has kidney disease or not)(ckd,notckd)	صنف

4.4 Attributes' Detailed Analysis

1. Class (ckd / not ckd)

As shown in figure 4.7 this attribute is the class which we want to predict in this research. The value is either ckd or notckd, ckd means the patient suffers from kidney failure, while notckd means the patient is well. The dataset composed of 400 instances, 250 of them are sick patients, while the other 150 are not infected. The blue color represents the ***ckd*** and the red represents the ***notckd*** patients. This color representation will be used throughout this chapter.

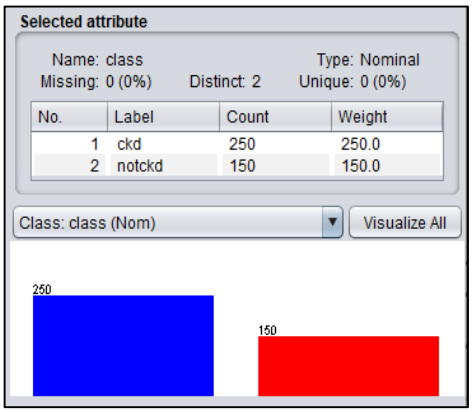

Figure 4.7 Basic Statistics of ***class***

2. Patient's Age (age)

As shown in figure 4.8 the basic statistics the patients' age ranges from 2 to 90 years. The mean is about 51 years. Each column in the bar graph represents 8 years, i.e. the first column represents patients whom ages range from 2 to 10 years. By excluding the first and last columns, we can notice that the ***ckd*** appears in the old patient. That means the relation between the disease and ***age*** is directly proportional, i.e. whenever the patient is old he/ she will be more vulnerable to disease.

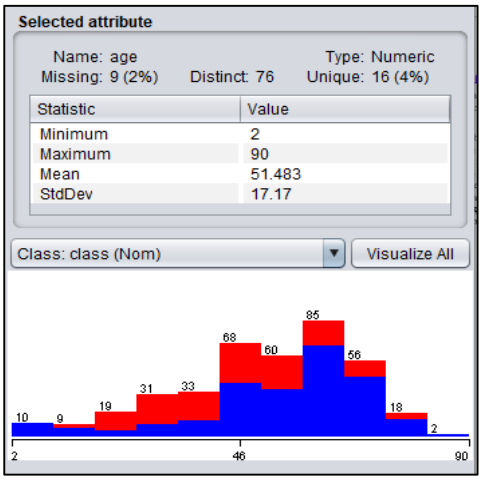

Figure 4.8 Basic Statistics of ***age***

3. Blood Pressure (bp)

The normal blood pressure is less than 120 over 80 (120/80). The ***bp*** in this dataset is taken for the lower measure (denominator or the 80). Figure 4.9 shows the minimum bp is 50 and the maximum bp is 180, where the mean is 76.5. Each column in the bar graph represents 6.5 points in ***bp***, i.e. the first column represents patients whom bp range from 50 to 56.5. It is very clear the blue color appears much with high pressure. That means the relation between the disease and ***bp*** is also directly proportional, i.e. whenever the patient is old he/she will be more vulnerable to disease.

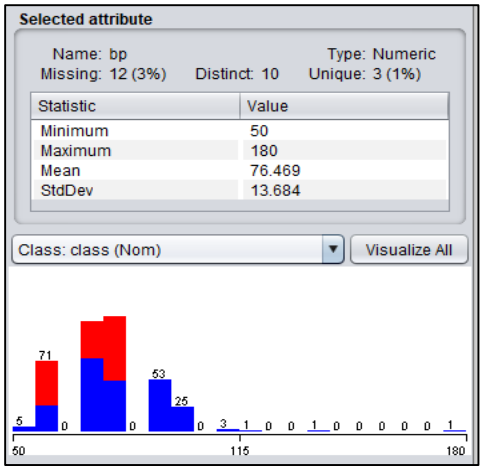

Figure 4.9 Basic Statistics of ***bp***

4. Urinary Specific Gravity (sg)

Urinary specific gravity (***sg***) is a measure of the concentration of solutes in the urine. It measures the ratio of urine density compared with water density and provides information on the kidney's ability to concentrate urine. A urinary specific gravity measurement is a routine part of urinalysis. Ideally, ***sg*** will fall between 1.002 and 1.030. sg above 1.010 indicate mild dehydration. (medscape, 2020) . Figure 4.10 ***sg*** measures and the count for each, it is clear in the figure the blue color - which denotes sick people - is concentrated on the left side, i.e. the lower the scale, the more likely the disease will appear.

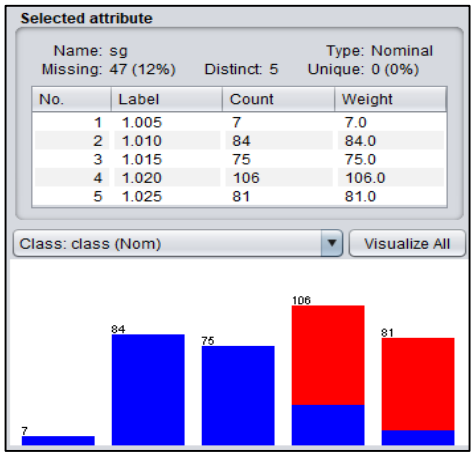

Figure 4.10 Basic Statistics of ***sg***

5. Basic Exploration about (al)

This test measures the amount of the protein albumin in blood. Liver makes albumin, which carries substances such as hormones, medicines, and enzymes throughout the body. This test can help diagnose and evaluate kidney and liver conditions. When the kidney starts to fail, albumin starts to leak into your urine. This causes a low albumin level in your blood. The normal albumin range is 3.4 to 5.4 g/dL. Lower albumin level is indication to malnutrition. (urmc, rochester, 2020).As shown in figure 4.11 High ***al*** rates appears with the ***ckd***.

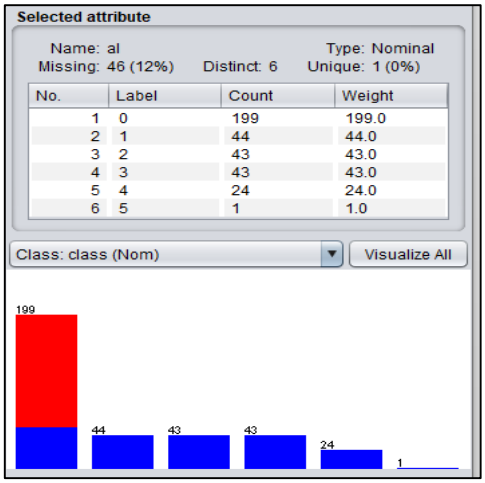

Figure 4.11 Basic Statistics of ***al***

6. Basic Exploration about (su)

Blood sugar concentration, or blood glucose level is the concentration of glucose present in the blood . Glucose is a simple sugar and approximately 4 grams of glucose are present in the blood of a 70-kilogram human at all times (nih, 2020). As shown in figure 4.12 High su rates appears with the ckd.

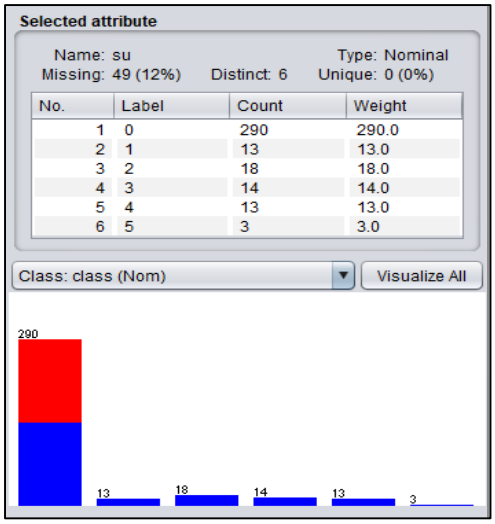

Figure 4.12 Basic Statistics of ***su***

7. Red Blood Cells (rbc)

Red blood cell count is a blood test that counts the number of red blood cells (RBCs) per microliter. The number of RBCs you have can affect how much oxygen your tissues receive. According to the Leukemia & Lymphoma Society: The normal RBC range for men is 4.7 to 6.1 million cells per microliter (mcL). And 4.2 to 5.4 million mcL for women, and 4.0 to 5.5 million mcL for children. (healthline, 2020). Figure 4.13 shows the results of data exploration of the ***rbc***.

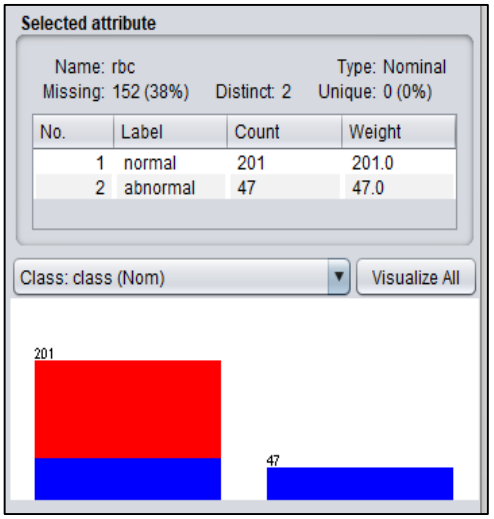

Figure (4.13) Basic Statistics of ***rbc***

8. Basic Exploration about (pc)

Pus cell (nominal) (normal, abnormal) pc Pyuria, which is a urinary condition that is characterized by an elevated number of white blood cells in the urine. Doctors define a high number as at least 10 white blood cells per cubic millimeter (mm3) of centrifuged urine. Pyuria can cause the urine to look cloudy or as if it contains pus.(medicalnewstoday, 2020) That is clear from figure (4.14) that ckd is found when the pc is abnormal (the right blue bar is abnormal).

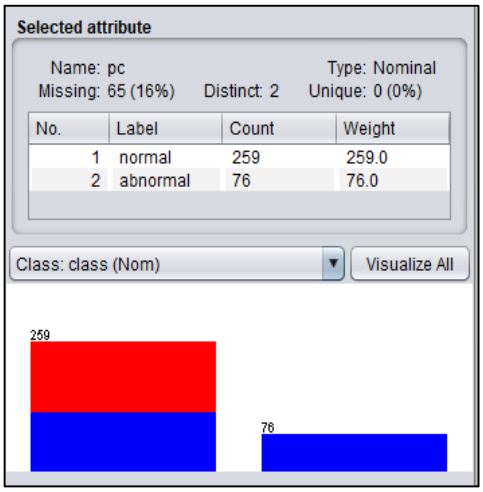

Figure (4.14) Basic Statistics of ***pc***

9. Basic Exploration about (pcc)

Pus cell clumps (nominal) (present, notpresent) pcc is Pyuria, which is urinary condition that is characterized by an elevated number of white blood cells in the urine. Doctors define a high number as at least 10 white blood cells per cubic millimeter (mm3) of centrifuged urine. Pyuria can cause the urine to look cloudy or as if it contains pus.. [49] That is clear from figure (4.15) that ckd is found when the ***pcc*** is present (the left blue bar is present).

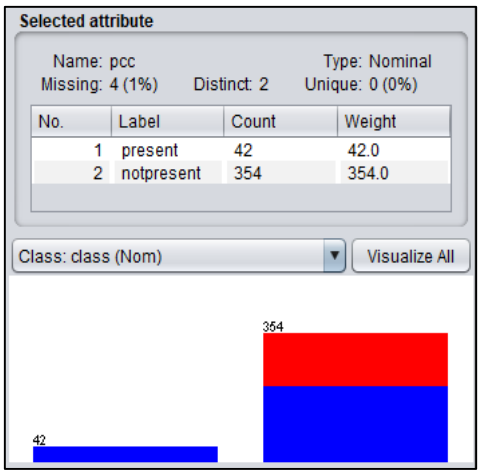

Figure (4.15) Basic Statistics of ***pcc***

10. Bacteria (ba)

Bacteria: Infectious diseases are caused by microorganisms, such as bacteria, viruses, fungi, and parasites. Doctors suspect an infection based on the person's symptoms, physical examination results, and risk factors. First, doctors confirm that the person has an infection rather than another type of illness. A sample is taken from an area of the person's body likely to contain the microorganism suspected of causing the infection. Samples may include, Blood, Sputum, Urine, Stool, Tissue, Cerebrospinal fluid, Mucus from the nose, throat, or genital area (merckmanuals, 2020). Figure 4.16 shows the results of data exploration of the ***ba***.

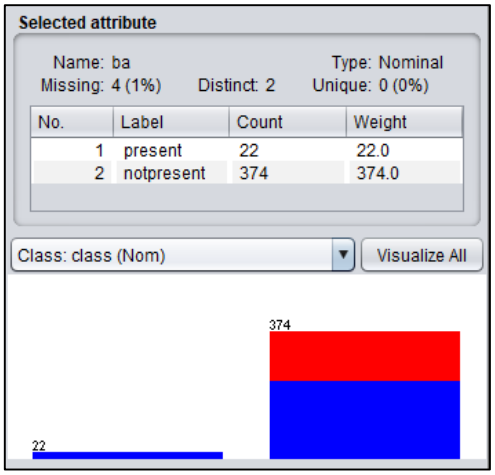

Figure (4.16) Basic Statistics of ***ba***

11. Blood Glucose Random (bgr)

Blood Glucose Random test measures the level of glucose in the blood at any given point in the day. Many blood tests for diabetes involve either fasting or continuous monitoring, but this test does not. It's useful for people who need a speedy diagnosis, such as those with type 1 diabetes who require supplementary insulin as a matter of emergency. (medicalnewstoday, 2020) High blood glucose, also called blood sugar, can damage the blood vessels in your kidneys. As shown in figure (4.17) the ckd appears much with high bgr.

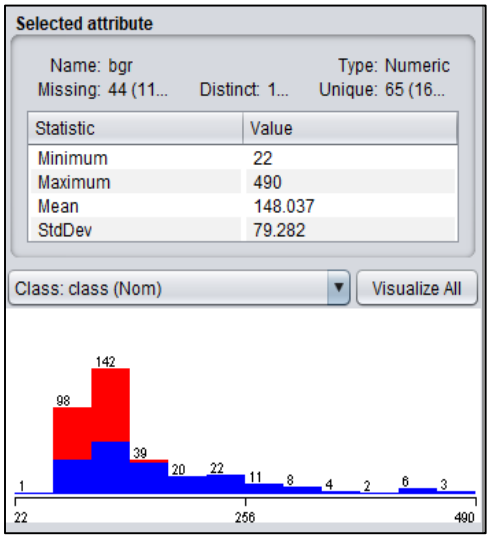

Figure (4.17) Basic Statistics of ***bgr***

12. Blood Urea (bu)

A blood urea nitrogen test is a blood test that is most commonly used to evaluate kidney function. It does this by measuring the amount of urea nitrogen in the blood. Urea nitrogen is a waste product that's created in the liver when the body breaks down proteins. Normally, the kidneys filter out this waste and urinating removes it from the body. BUN levels tend to increase when the kidneys or liver are damaged. Having too much urea nitrogen in the blood can be a sign of kidney or liver problems. (healthline, 2020) And this is very clear with the high number of ckd appears with high rates of bu as shown in figure (4.18).

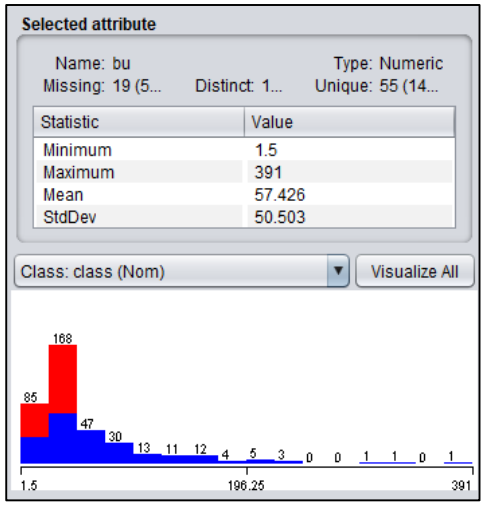

Figure (4.18) Basic Statistics of ***bu***

13. Serum Creatinine (sc)

Serum Creatinine test reveals important information about your kidneys. Creatinine is a chemical waste product that's produced by your muscle metabolism and to a smaller extent by eating meat. Healthy kidneys filter creatinine and other waste products from your blood. The filtered waste products leave your body in your urine. If your kidneys aren't functioning properly, an increased level of creatinine may accumulate in your blood. A serum creatinine test measures the level of creatinine in your blood and provides an estimate of how well your kidneys filter (mayoclinic, 2020). Figure 4.19 shows the results of data exploration of the ***sc***.

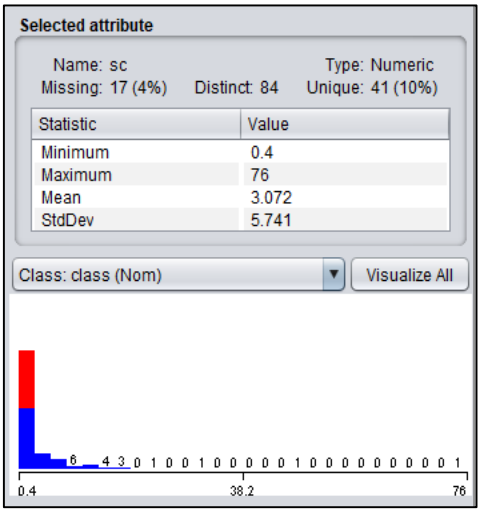

Figure (4.19) Basic Statistics of ***sc***

14. Sodium (sod)

A sodium blood test measures the amount of sodium in your blood. Sodium is a type of electrolyte. Electrolytes are electrically charged minerals that help maintain fluid levels and the balance of chemicals in your body. Sodium also helps your nerves and muscles work properly. Kidneys get rid of the rest in your urine. If your sodium blood levels are too high or too low, it may mean that you have a problem with your kidneys, dehydration, or another medical condition (medlineplus, 2020). Figure 4.20 shows the results of data exploration of the ***sod***.

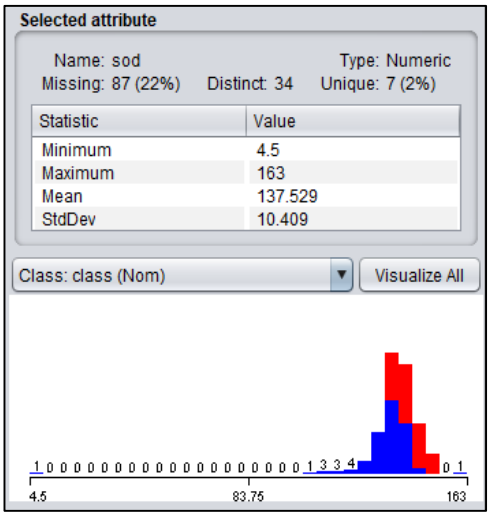

Figure (4.20) Basic Statistics of ***sod***

15. Potasium (pot)

A potassium blood test measures the amount of potassium in your blood. Potassium is a type of electrolyte like sodium. Potassium levels that are too high or too low may indicate a medical problem. A potassium blood test is often included in a series of routine blood tests called an electrolyte panel. The test may also be used to monitor or diagnose conditions related to abnormal potassium levels. These conditions include kidney disease, high blood pressure, and heart disease (medlineplus, 2020). Figure 4.21 shows the results of data exploration of the ***pot***.

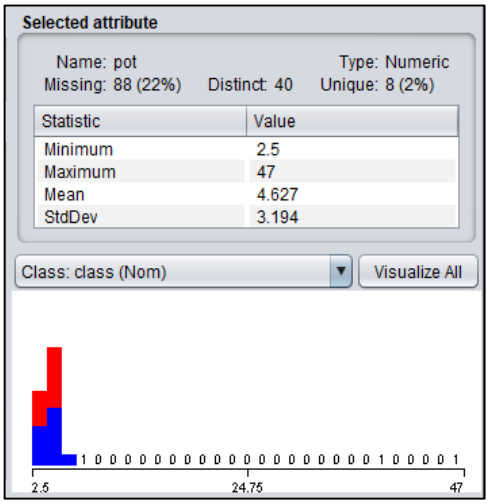

Figure (4.21) Basic Statistics of ***pot***

16. Hemoglobin (Hemo)

The hemoglobin (hgb) test measures how much hemoglobin your red blood cells contain, hgb is a protein produced by your bone marrow that's stored in red blood cells. It helps red blood cells transport oxygen from your lungs to your body through your arteries. It also transports carbon dioxide (CO_2) from around your body back to your lungs through your veins. (healthline, 2020) Typical healthy hgb levels are as follows: For men, Hgb levels below 13 g/dL are considered low. For women, Hgb levels below 12 g/dL are considered low if not pregnant. One of the Possible causes of low hgb include is chronic kidney disease, it is clear from figure (4.22) that ckd is found when the hemo is low.

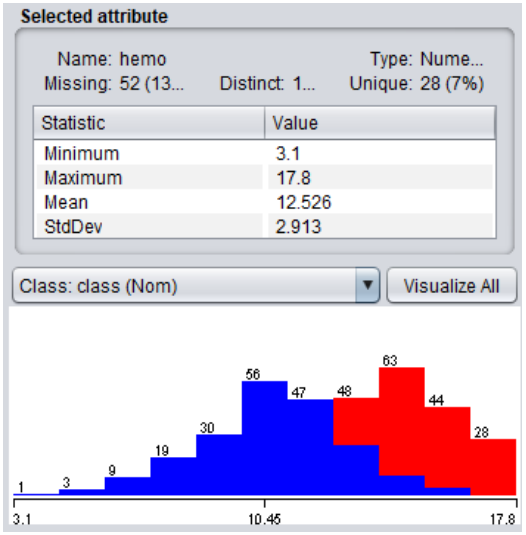

Figure (4.22) Basic Statistics of ***hemo***

17. Packed Cell Volume (pcv)

Packed Cell Volume test is done to diagnose polycythemia, dehydration or anemia in certain patients. It is generally a part of the full blood count test. The pcv test measures how much of the blood consists of cells. If the pcv returns a reading of 50%, it means that 50 ml of the cells are present in exactly 100 ml of blood. If the rbc number increases, then the total reading of the pcv is also up. This number can also increase due to dehydration. There are certain conditions that contribute to the low reading in the pcv, these include Kidney diseases. (portea, 2020) That is clear from figure (4.23) that ckd is found when the pcv is low.

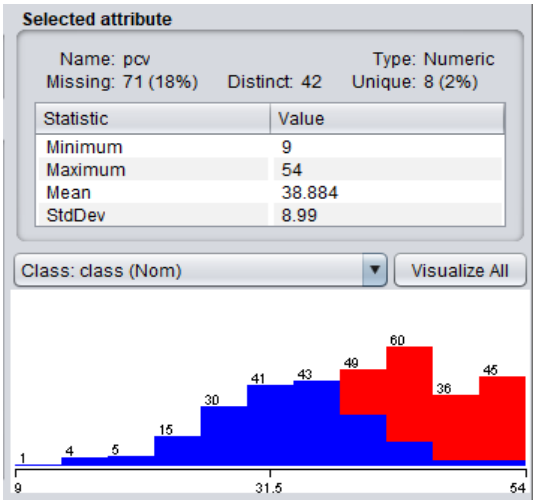

Figure (4.23) Basic Statistics of ***pcv***

18. White Blood Cell Count (wbcc)

White Blood Cell Count is a test that measures the number of white blood cells in your body. A wbcc can detect hidden infections within your body and alert doctors to undiagnosed medical conditions, such as autoimmune diseases, immune deficiencies, and blood disorders. It's also called leukocytes, are an important part of the immune system. These cells help fight infections by attacking bacteria, viruses, and germs that invade the body (healthline, 2020). Figure 4.24 shows the results of data exploration of the ***wbcc***.

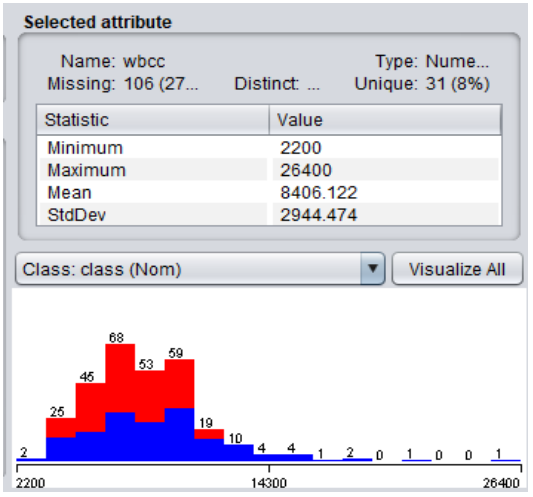

Figure (4.24) Basic Statistics of ***wbcc***

19. Red Blood Cell Count (rbcc)

Red Blood Cell Count is a blood test that is used to find out how many red blood cells you have. The test is important because ***rbcc*** contain hemoglobin, which carries oxygen to your body's tissues. According to the Leukemia & Lymphoma Society: The normal rbcc range for men is 4.7 to 6.1 million cells per microliter (mcL), and 4.2 to 5.4 million mcL for women, and 4.0 to 5.5 million mcL for children. Lower than normal count indicator to ***ckd***. (healthline, 2020) That is clear from figure (4.25) that ckd is found when the ***rbcc*** is low.

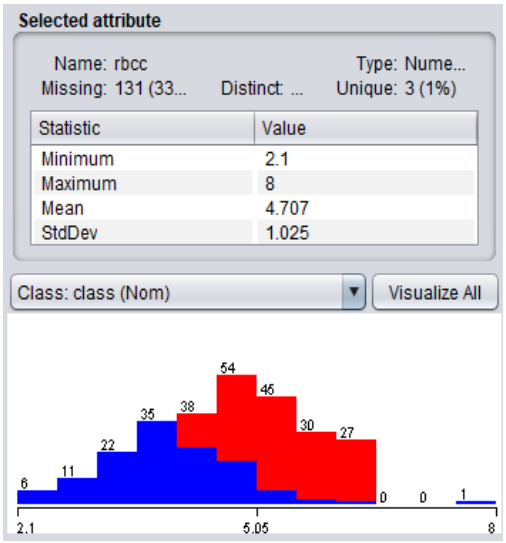

Figure (4.25) Basic Statistics of ***rbcc***

20. Hypertension (htn)

Hypertension: Is the patient suffering from hypertension(High blood pressure)? the answer is yes or no in the(htn) attribute. Patients may have (hypertension) for years without any symptoms. Even without symptoms, damage to blood vessels and your heart continues and can be detected. Uncontrolled high blood pressure increases your risk of serious health problems, including heart attack, stroke and Kidney problems. (mayoclinic, 2020) That is clear from figure (4.26) that ckd is found when the ***htn*** is yes (the left blue bar is yes).

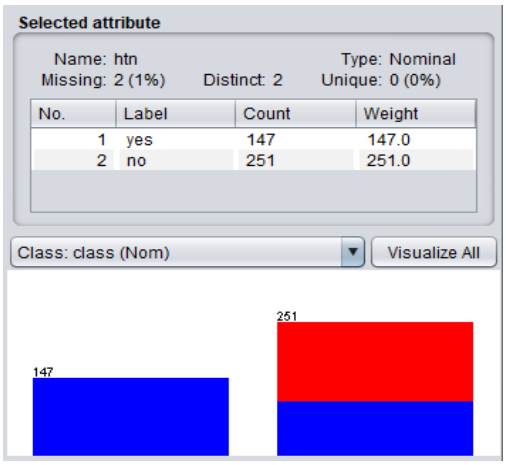

Figure (4.26) Basic Statistics of ***htn***

21. Diabetes Mellitus (dm)

Diabetes Mellitus: Is the patient suffering from diabetes, the answer is yes or no in the(dm) attribute. Diabetes mellitus is a disease that prevents your body from properly using the energy from the food you eat. There are two main types of diabetes: type 1 and type 2 One of the risks of not maintaining a diabetic patient is kidney damage. Diabetes is the leading cause of kidney disease. About 1 out of 4 adults with diabetes has kidney disease. (niddk, 2020) That is clear from figure (4.27) that ckd is found when the ***dm*** is yes (the left blue bar is yes).

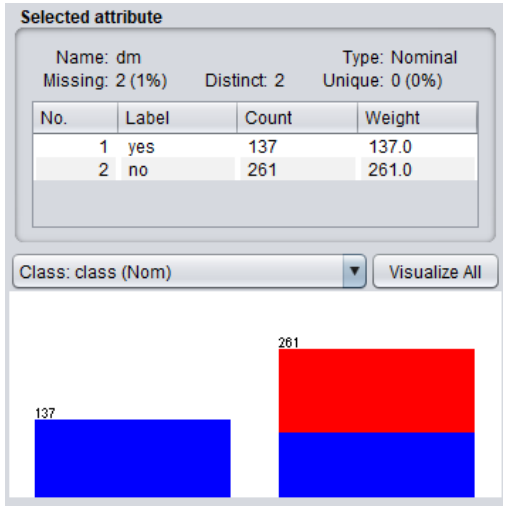

Figure (4.27) Basic Statistics of ***dm***

22. Coronary Artery Disease (cad)

Coronary Artery Disease: Is the patient suffering from cad? The answer is yes or no in the (***cad***) attribute. Coronary artery disease is the narrowing or blockage of the coronary arteries, usually caused by atherosclerosis. Atherosclerosis is the buildup of cholesterol and fatty deposits on the inner walls of the arteries (clevelandclinic, 2020). Figure 4.28 shows the results of data exploration of the ***cad***.

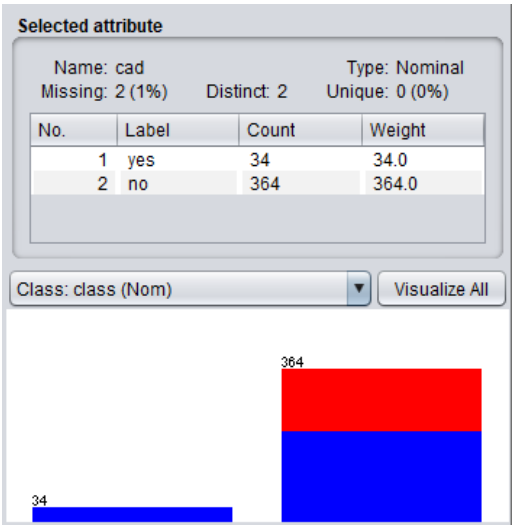

Figure (4.28) Basic Statistics of ***cad***

23. Appetite (appet)

Appetite: Dose the patient suffering from lack of appetite? The answer is poor or good in the(***appet***) attribute. One of the causes of loss of appetite is chronic kidney failure. (medicinenet, 2020) That is clear from figure (4.29) that ckd is found when the ***appet*** is poor (the right blue bar is poor).

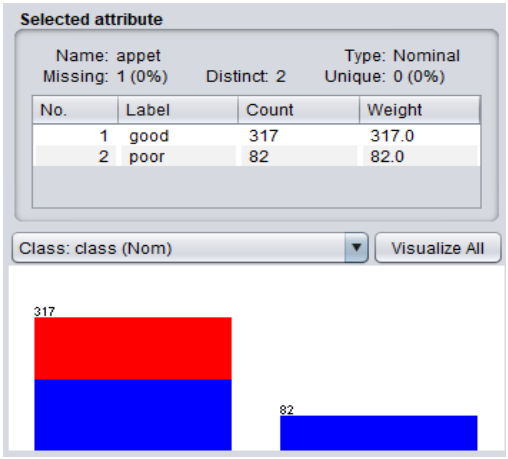

Figure (4.29) Basic Statistics of ***appet***

24. Pedal Edma (pe)

Pedal Edma: Dose the patient suffering from Pedal Edma? The answer is yes or no in the(***pe***) attribute. Edema is swelling caused by excess fluid trapped in your body's tissues. Although edema can affect any part of your body, you may notice it more in your hands, arms, feet, ankles and legs. IT may be a sign of Kidney disease or Kidney damage or Congestive heart failure or many diseases. (mayoclinic, 2020) That is clear from figure (4.30) that ckd is found when the ***pe*** is yes (the left blue bar is yes).

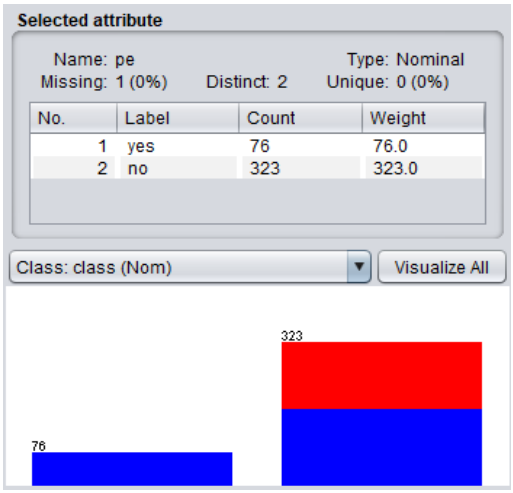

Figure (4.30) Basic Statistics of ***pe***

Anemia (ane)

Anemia: Dose the patient suffering from Anemia, the answer is yes or no in the (ane) attribute. Anemia is a condition in which you lack enough healthy red blood cells to carry adequate oxygen to your body's tissues. Having anemia can make you feel tired and weak. kidney failure can lead to a shortage of red blood cells. (mayoclinic, 2020) That is clear from figure (4.31) that ckd is found when the ***ane*** is yes (the left blue bar is yes).

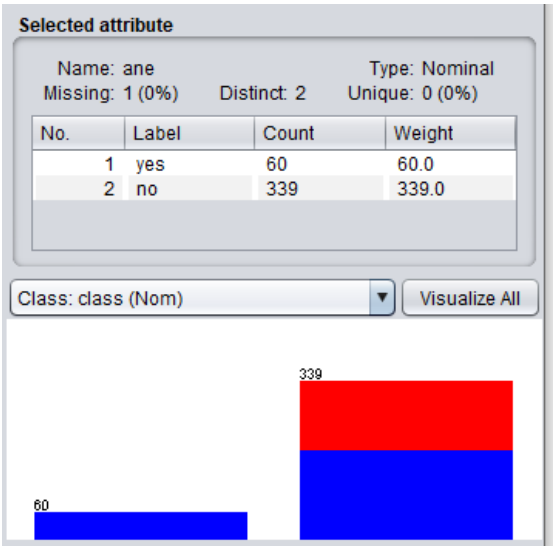

Figure (4.31) Basic Statistics of ***ane***

CHAPTER FIVE

FEATURE SELECTION

AND

CLASSIFICATION MODELS

CHAPTER V: FEATURE SELECTION AND CLASSIFICATION MODELS' DEVELOPMENT

5.1 Introduction

There are two main goals for this chapter. The first goal is to design a *feature selection* model that can determine the factors which can lead to kidney failure, and to identify relations between them. While the second goal is to builds a *classification model* that can helps in early prediction of the disease, and also can provide assistance for doctors to get better understanding about the Chronic Kidney Disease.

This chapter starts by describing the steps used to achieve the research objectives as shown in figure (5.1). Then basic description about the datasets is provided. After that a description about Weka experiment is shown, to depict the steps that were followed to compare between the available classifications algorithms. The result of this comparison is then presented to identify the best algorithms for our problem. Then the top five algorithms plus the neural network were used to develop the final models. Each model is presented by two parts: a set of *selected attributes*, and a *classification model*. By the end of each model an evaluation is presented to show the model's accuracy, then a brief discussion was made to show the meaning behind the numbers. The chapter ends by an overall discussion about the achieved results, and link the models' results with the medical theories to shows how these experiments and model achieved the research objective, and prove the power of data mining in the medical field.

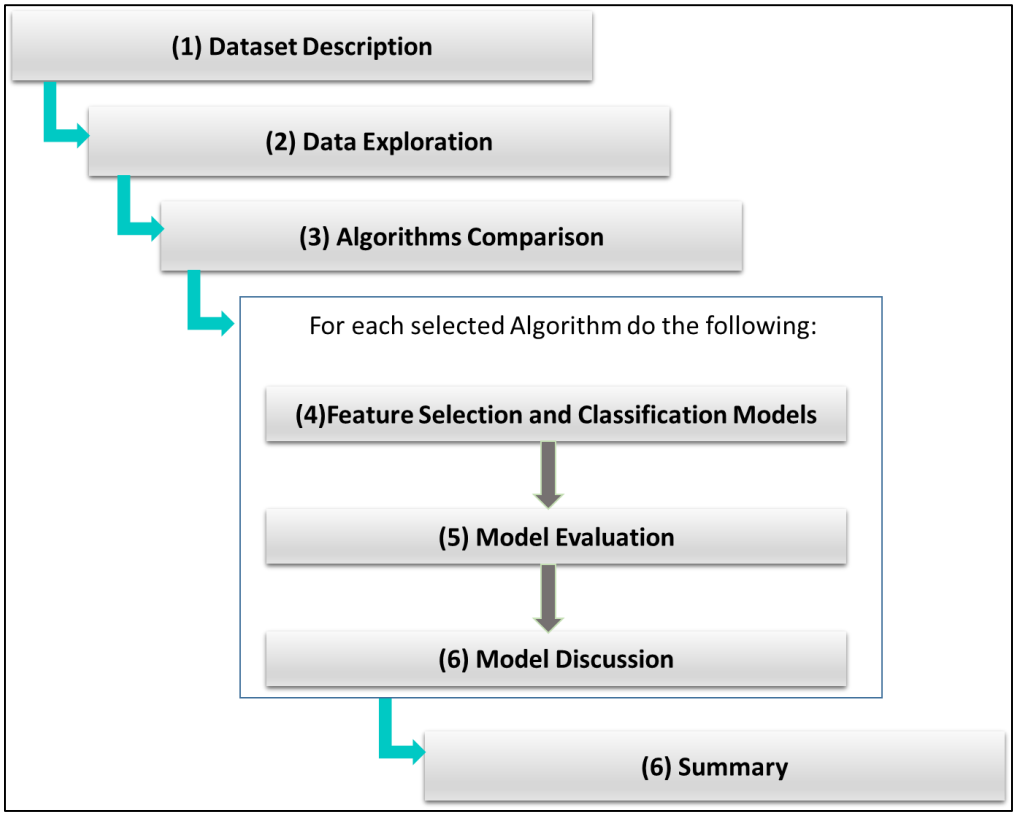

Figure (5.1) Methodology for Feature Selection and Classification Models

5.2 Dataset Description

The data set used for this research is patients' data during their early stage of Chronic Kidney Disease. The data set is composed of 25 attributes, 24 of them are independent attributes describing patient's status like age, or a laboratory results of the patient, the last attribute is the class which describes the patient's status whether he is suffering from the disease (***ckd***), or not (***notckd***). The dataset is composed of 400 instances. Table (5.1) gives brief description about the dataset attributes and the meaning of each. More details about the dataset - like the meaning of each attribute - was provided in chapter 3: Data Exploration.

Table 5.1 Dataset Description

#	Attribute	Type	Description
1	**age**	**Numeric**	Patient's age.
2	**bp**	**Numeric**	Blodd Pressure. Normal range between 80 - 90
3	**sg**	**Nominal**	specific gravity(a measure of the concentration of solutes in the urine) (1.005,1.010,1.015,1.020,1.025)
4	**al**	**Nominal**	Albumin -(albumin in blood) (0,1,2,3,4,5)
5	**su**	**Nominal**	Sugar(blood sugar level)- (0,1,2,3,4,5)
6	**rbc**	**Nominal**	Red Blood Cells-(a red blood cell count) (normal,abnormal)
7	**pc**	**Nominal**	Pus Cell - (normal,abnormal)
8	**pcc**	**Nominal**	Pus Cell clumps - (present,notpresent)
9	**ba**	**Nominal**	Bacteria - (present,notpresent)
10	**bgr**	**Numeric**	Blood Glucose Random(blood Random glucose testing)
11	**bu**	**Numeric**	Blood Urea(the amount of urea nitrogen in the blood)
12	**sc**	**Numeric**	Serum Creatinine(measures the level of creatinine in blood)
13	**sod**	**Numeric**	Sodium(blood test)
14	**pot**	**Numeric**	Potassium(blood test)
15	**hemo**	**Numeric**	Hemoglobin(measures how much hemoglobin red blood cells contain.)
16	**pcv**	**Numeric**	Packed Cell Volume(a part of the full blood count test)
17	**wbcc**	**Numeric**	White Blood Cell Count
18	**rbcc**	**Numeric**	Red Blood Cell Count
19	**htn**	**Nominal**	Hypertension -does patient has hypertension or not (yes,no)
20	**dm**	**Nominal**	Diabetes Mellitus -(des patient has diabetes or not) (yes,no)
21	**cad**	**Nominal**	Coronary Artery Disease (Does the patient has coronary artery disease or not) (yes,no)
22	**appet**	**Nominal**	Patint's appetite - (good,poor)
23	**pe**	**Nominal**	Pedal Edema - (Does patient has pedal edema or not)(yes,no)
24	**ane**	**Nominal**	Anemia - (Does patient has anemia or not)(yes,no)
25	**class**	**Nominal**	class - (Does the patient has kidney disease or not)(ckd,notckd)

5.3 Data Exploration

Any data mining research should start by data exploration phase which will provide basic understanding about the dataset, and even can provide preliminary results. Chapter 3 Data Exploration, provides statistics about each field. In the coming sections a description about the experiments used, and deep discussion about results will be provided.

5.4 Algorithms' Comparison

The target of this section is to determine the most suitable algorithms for our data set. This was done by developing an experiment in Weka experiment, by running 25 algorithms for our dataset.

The experiment run the selected algorithms one after another and registered the results evaluation metrics in the result's file. Because 10-fold cross-validation was used for evaluating the test result, the result file contains 100 results (10 * 10) for each algorithm. Then the average was taken for the 100 results for each algorithm, and these averages were collected in one table. Table 5.2 shows the average result for each algorithm descending ordered by the number of correctly classified instances, and ascending by the mean absolute error.

Table 5.2 Evaluation results of Algorithms comparison Experiment

#	Algorithm	Number of correct	Number of incorrect	Number of Unclassified	Percent of Correct	Percent of Incorrect	Kappa Statistic	Mean Absolute Error
1	bayes.NaiveBayesMultinomialText	40	25	15	0	62.5	0	0
2	functions.SGDText	40	25	15	0	62.5	0	0
3	trees.RandomForest	39.91	0.09	0	99.775	0.225	0.99526316	0.04159762
4	trees.J48	39.57	0.43	0	98.925	1.075	0.97695284	0.02431427
5	bayes.BayesNet	39.48	0.52	0	98.7	1.3	0.97282297	0.0137934
6	trees.LMT	39.48	0.52	0	98.7	1.3	0.97262719	0.02176853
7	functions.SimpleLogistic	39.48	0.52	0	98.7	1.3	0.97262719	0.0218776
8	meta.AdaBoostM2	39.45	0.55	0	98.625	1.375	0.97081997	0.01975878
9	rules.PART	39.44	0.56	0	98.6	1.4	0.96999859	0.02864362
10	meta.FilteredClassifier	39.42	0.58	0	98.55	1.45	0.9690566	0.02928084
11	functions.SMO	39.24	0.76	0	98.1	1.9	0.96035175	0.019
12	rules.DecisionTable	39.2	0.8	0	98	2	0.95694147	0.19477431
13	trees.RandomTree	39.06	0.94	0	97.65	2.35	0.95061001	0.03446911
14	trees.REPTree	38.86	1.14	0	97.15	2.85	0.93891855	0.05801606
15	rules.JRip	38.69	1.31	0	96.725	3.275	0.92997051	0.0393878
16	lazy.IBk	38.34	1.66	0	95.85	4.15	0.91436488	0.04403315
17	trees.HoeffdingTree	38.29	1.71	0	95.725	4.275	0.91201898	0.04356265
18	bayes.NaiveBayes	38.08	1.92	0	95.2	4.8	0.90144847	0.04855553
19	bayes.NaiveBayesUpdateable	38.08	1.92	0	95.2	4.8	0.90144847	0.04855553
20	trees.DecisionStump	36.84	3.16	0	92.1	7.9	0.83688606	0.13505046
21	lazy.KStar	36.7	3.3	0	91.75	8.25	0.82806386	0.12977092
22	rules.OneR	36.69	3.31	0	91.725	8.275	0.82654515	0.08275
23	functions.VotedPerceptron	25	15	0	62.5	37.5	0	0.375
24	misc.InputMappedClassifier	25	15	0	62.5	37.5	0	0.46892265
25	rules.ZeroR	25	15	0	62.5	37.5	0	0.46892265

5.5 Feature Selection and Classification Models

5.5.2 Feature Selection

The wrapper method will be used in this research, that means the feature selection and the prediction model development are done in one step. This chapter starts by describing the datasets, then basic statistical analysis about these datasets is shown. After that an initial comparison between the prediction algorithms is done for each dataset, to select the most appropriate algorithm for each dataset to build the prediction model. Then for each dataset three main tasks are done: the first is the attributes' selection, the second is the power prediction model using the selected features, the third is the model evaluation. After that a summary and discussion about feature selection is done. Finally results discussion and models Comparison is done to depict the knowledge behind these numbers. The methodology followed to achieve the goal is shown in figure (5.1).

5.5.2.2 Summary of Selected Attributes

Table 5.3 shows the summary of attribute selection result for the selected algorithms. The table is in descending ordered by the Total column. The first column shows the attribute ID, the second column is the attribute name. The rest of the columns represent the result of the attribute selection. If the feature is selected by an algorithm, the number "1" is written in the corresponding cell; else the cell is empty. The last column is the percentage, which gives the percentage of the algorithms that selected the corresponding attribute, of the total number of algorithms. The total field is the summation of ones, which represents the number of algorithms that had selected this feature. The higher number in the total field means the attribute was selected by many algorithm, while zero in the total means none of the algorithms selected this attribute. Of course this gives indication about the importance this feature in the classification of Chronic Kidney Disease.

5.5.2.3 Attribute Selection Results and Discussion

Table 5.3 shows the summary of attribute selection result for the selected algorithms. The table is in descending ordered by the total column, this gives indication about the importance of feature in the classification of Chronic Kidney Disease. According

to the attribute selection result, the features could be ranked into five groups. In this section more discussion about feature selection result of each group will be provided from medicine perspective.

5.5.2.3.1 Group 1: which includes *sg* and *hemo*.

Group (1) which consists of ***sg*** and ***hemo***, there is a consensus from all algorithms about the importance of ***sg*** (specific gravity) and ***hemo*** (hemoglobin), that is obvious because both of them were selected by all algorithms. Specific gravity ***sg***, in the context of clinical pathology, is a urinalysis parameter commonly used in the evaluation kidney function and can aid in the diagnosis of various renal diseases. Also the kidney produce a hormone called erythropoietin (EPO), which is responsible for the production of the red blood cells. CKD results in low levels of (EPO) so anemia can happen early in the course of kidney disease and grow worse as of kidney fail and can no longer make EPO.

Therefore, there is consensus from all algorithms about the importance of ***sp*** and ***hemo*** as an early indicators of predictions of early stages of CKD because both of them were selected by all algorithms as shown by table 5.3.

5.5.2.3.2 Group 2 : which includes *al, pcv, rbcc, and dm*

Group 2: which consists of ***al***, ***pcv***, ***rbcc***, and ***dm***. These attributes were selected by three algorithms. Firstly, albuminuria is a sign of kidney disease and means that you have too much albumin in your urine. Albumin is a protein found in the blood. A healthy kidney doesn't let albumin pass from the blood into the urine. The less albumin in your urine, the better. Secondly there's a significant negative correlation between ***pcv*** and duration in months of chronic kidney dysfunction. CKD progression was associated with decreases in ***hb*** and RBC lifespan (***rbcc***). Lastly Diabetes mellitus is the leading cause of chronic kidney disease (CKD) and a major public health issue worldwide. Approximately 20-30% of patients with type 2 diabetes mellitus (T2DM) have renal impairment, classified as moderate-to-severe CKD (glomerular filtration rate (GFR)<60mL/min/1.73m2).

Since they were selected by three out of six algorithms they could aid in the diagnosis of early stages of CKD but to a smaller extent than group 1 as shown by table 5.3.

5.5.2.3.3 Group 3 : which includes *sc, htn, appet* and *pe*

Group (3) consist of ***sc, htn, appet, pe*** , each attribute in this group was selected by two algorithms, also this is expected because. As mentioned before ***sc*** is the most commonly used screening test for renal failure. Creatinine is a waste product in the blood that comes from muscle activity. It is normally removed from blood by the kidney but when kidney function slows down, the creatinine level rises and this mainly with advanced stages of CKD. Regarding ***htn*** it is both an important and a consequence of CKD. ***htn*** is highly prevalent in CKD, particularly with ESKD.

Patients with CKD frequently experience poon appetite related to uremia and other co-morbidities. Appetite may worsen with progression of kidney disease leading to malnutrition.

CKD does not cause any symptoms until most of the kidneys are destroyed once the kidney is severely damaged the patients suffer from swelling around eyes, called pre orbital edema, swelling of legs called pedal edema.

Based on Table 5.3 Summary of Selected attributes, this group was selected by two out of six algorithms. It seems that the prevalent of the mentioned attribute of this group is towards advanced stages of CKD and this is expected since our target is concerning early stages of CKD.

5.5.2.3.4 Group 4 : which includes *age, bp, su, rbc, pc, ba, bgr* and *sod*.

Group 4: which include ***age, bp, su, rbc, pc, ba, bgr and sod***. This group was selected by only one algorithm, according to the obtained results this could be interpreted as the less important attributes, CKD becomes more common with increasing age. After the age of 40, Kidney filtration begins to fall by approximately 1% per year. In addition to the natural aging of the kidneys, many conditions that damages the kidneys are more common in older people including diabetes, high blood pressure and heart disease. ***bp***, ***htn*** and CKD are closely associated with an intermingled cause and effect relationship. (***bp***) typically rises with declines in kidney function, and sustained evolutions in ***bp*** hasten progression of kidney disease. CKD is defined by reduction in GFR, which in turn is associated with increased plasma creatinine and urea concentrations. As CKD progresses, plasma levels of both rise in tandem. Healthy kidneys produce a hormone called EPO. EPO

promotes the bone marrow to make red blood cells, which then carry oxygen throughout the body. When the kidneys are diseased or damaged, they do not make enough EPO. As a result, the bone marrow makes fewer red blood cells, causing anemia. In CKD impaired kidney function results in accumulation of uranic toxins, which effects and contribute to inflammation which results in infection anywhere, like urinary tract infection, gut infection, pneumonia and etc. DM is growing epidemic and the most common cause of CKD and kidney failure. In CKD, the kidneys cannot remove excess salt and fluid so they build up in the body and can cause: 1 High blood pressure 2 Swelling of ankles, feet, hands and puffiness under eyes 3 Shortness of breath.

According to the obtained results this could be interpreted as least important attributes again because all were selected by one out of six algorithms as they associated with later stages of CKD.

5.5.2.3.5 Group 5: which is composed of *pcc, bu, pot, wbcc, cad* and *ane*.

Group (5): consists of ***pcc, bu, pot, wbcc, cad***, and ***ane*** , these attributes were not selected at all. This could be interpreted as an evidence that they may not be directly related to the diagnose of kidney failure. This result is totally true according to discussion with the doctors. When these results presented to doctors they gave/-***pcc***:Clumps of numerous white cells are seen in inflections. Presence of many white cells in urine is called pyuria. Pus cells greater than 10/HPF or presence of clumps is suggestive of urinary tract infections. -***bu***:Ureamia is a clinical syndromes marked by elevated concentration of urea in the blood and associated with fluid, electricity and hormone imbalance and metabolic abnormalities, which develop in parallel with deterioration of renal function. Ureamia more commonly develops with (CKD) especially the later stages of CKD. -***pot***:When kidneys fail they can no longer remove excess potassium, so the level builds up in the body. High potassium in the blood is called hyperkalemia, which may occur in people with advanced stages of CKD. -***wbcc***:Inflammation is a pathogenic factor in renal injury but-weather inflammation is related to renal outcome in (CKD) patients is little known. So elevated levels of ***wbcc*** are risk factors associated with rapid renal progression in advanced CKD patients. -***cad***:The risk of (CAD) increases as (CKD) advances (ESRD) despite adjustment for traditional cardiovascular risk factors. That means that ESRF patients have highest CVD risk among

CKD population. – ***ane***:In CKD anemia may occur at early stages CKD stages 2 and 3 of the KDIGO guidelines. The ***bu*** levels decrease when estimated (GFR) is around 70ml/min/1.73 m2 (men) and 50ml/min/1.73 m2 (women). However, anemia is more common in CKD stage 4 (even earlier in diabetic patients) and worsens as CKD progresses. In advanced stages of CKD and in the dialysis population anemia is present in as high as 90% of patients. It is interesting to notice that the attributes of this group have not detected by any algorithm. This indicates that these attributes have negative role diagnosing early stages of CKD.

Table 5.3 Summary of Selected attributes

Group	#	Attribute	Random Forest	J48	Bayes Net	LMT	Simple Logistic	Neural Network	Total	Attribute %
1	3	sg	1	1	1	1	1	1	6	100%
	15	hemo	1	1	1	1	1	1	6	100%
2	4	al			1	1		1	3	50%
	16	pcv	1		1			1	3	50%
	18	rbcc		1		1	1		3	50%
	20	dm	1	1	1				3	50%
3	12	sc		1	1				2	33%
	19	htn	1			1			2	33%
	22	appet	1		1				2	33%
	23	pe	1	1					2	33%
4	1	age		1					1	17%
	2	bp		1					1	17%
	5	su				1			1	17%
	6	rbc						1	1	17%
	7	pc		1					1	17%
	9	ba		1					1	17%
	10	bgr		1					1	17%
	13	sod				1			1	17%
5	8	pcc							0	0%
	11	bu							0	0%
	14	pot							0	0%
	17	wbcc							0	0%
	21	cad							0	0%
	24	ane							0	0%
Total			7	11	7	7	3	5		

5.5.3 Classification Models

The most critical part for developing the classification model is the selection of the appropriate algorithm, this had been done by a separate experiment as shown in part 4.4 Algorithm Comparison. Table 5.2 shows comparison between 25 algorithms, the table is ordered by ***Number of Correct Classified Instances*** descending, and ascending by ***Mean Absolute Error***. Two algorithms ***NaiveBayesMultinomialText***, and ***SGDText*** achieved 100% accuracy, however they were excluded from the models, because this high result leads to model over fitting. After excluding these two algorithms, the top five and Neural Network (***Multi-Layer Perceptron***) were selected to be used for model development. The first algorithm is ***Random Forest*** which achieved 99.775 accuracy. The second one is ***J48*** which achieved 98.925, then ***Bayes Net*** with 98.7 accuracy, ***LMT*** with 98.7 accuracy, and the last one is ***Simple Logistic*** which achieved 98.7.

In the subsequent parts of this chapter, the design and evaluation of each model will be presented, followed by a discussion about the model.

5.5.3.1 Random Forest Model

The first model was built using Random Forest algorithm, which achieved very high accuracy (99.775) and very low error rate (MAE=0.042) as shown in table 5.2 Algorithm Comparison. Below are more details about the model:

a. **Feature Selection:** The algorithm selected 7 attributes: (***sg, hemo, pcv, htn, dm, appet, pe***). The model succeeded to select the relevant attributes. As per the doctor's feedback which is shown in the discussion part.
b. **Classification Model**: The model shows very high success rate in patients' classification. That is very clear from the number of correctly classified instances in table 5.4, which are 395 out of 400 instances that means the model accuracy is 98.75.
c. **Model Evaluation**: The model evaluation is done using the confusion matrix as shown in table 5.5. The model succeeded to classify all of the notckd patient correctly. Only 5 of the ckd patients were wrongly classified as shown in table 5.5. Table 5.6 gives the details about the model evaluation, as shown in the table the true positive rate (TP Rate) is 0.980 for ***ckd***, and it is 1.00 for ***notckd***, because the model succeeded to classify all

the ***notckd*** patients correctly. Also the model Precision and Recall are very high, Precision is 1.00 for ***ckd***, and 0.968 for ***notckd***, while the Recall is 0.980 for ***ckd*** and 1.00 for ***notckd***.

Table 5.4 Summary of Random Forest

Summary	**Instances**	**%**
Correctly Classified Instances	395	98.75
Incorrectly Classified Instances	5	1.25

Table 5.5 Random Forest Confusion

Predicted		Confusion Matrix	
ckd	**notckd**		
245	5	**ckd**	Actual
0	**150**	**notckd**	

Table 5.6 Details of Random Forest

	TP Rate	FP rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.980	0.000	1.000	0.980	0.990	0.974	1.000	1.000	ckd
	1.000	0.020	0.968	1.000	0.984	0.974	1.000	1.000	notckd
Weighted Avg.	0.988	0.008	0.988	0.988	0.988	0.974	1.000	1.000	

5.5.3.2 J48 Model

The second model was built using J48 algorithm, which also achieved very high accuracy (98.925) and very low error rate (MAE=0.024) as shown in table 5.2 Algorithm Comparison. Although this MAE is less than the Random forest, but the random forest achieved higher number of correctly classified instance. Below are more details about the model:

a. **Feature Selection:** The algorithm selected 11 attributes: (***age, bp, sg, pc, ba, bgr, sc, hemo, rbcc, dm, pe***). The model succeeded to select the relevant attributes. The model provides the highest number of attributes, some attributes like ***age***, are selected only by this algorithm. Also the model succeeded in selecting the relevant attributes as per the doctors' feedback which is shown in the discussion part.

b. **Classification Model**: The model shows very high success rate in patients' classification. That is very clear from the number of correctly classified instances in table 5.7, which are 388 out of 400 instances that means the model accuracy is 97%. One of the most benefits of classification trees, are their clarity. Figure 5.6 shows the J48 tree view. The root node is the ***sc***, because it shows the highest information gain among all the attributes, then ***pe*** was selected to branch the tree, then ***dm,*** then ***hemo,*** then the last attribute which is ***sg***.

c. **Model Evaluation**: The model evaluation is represented by the confusion matrix as shown in table 5.8.The model succeeded to classify 248 out of 250 ***ckd*** patient, and 140 out of 150 ***notckd*** patients. Table 5.9 shows the details of model evaluation, as shown in the table the true positive rate (TP Rate) is 0.992 for ***ckd***, and it is 0.933 for ***notckd***. Also the model Precision and Recall are very high, Precision is 0.961 for ***ckd***, and 0.986 for ***notckd***, while the Recall is 0.992 for ***ckd*** and 0.933 for ***notckd***.

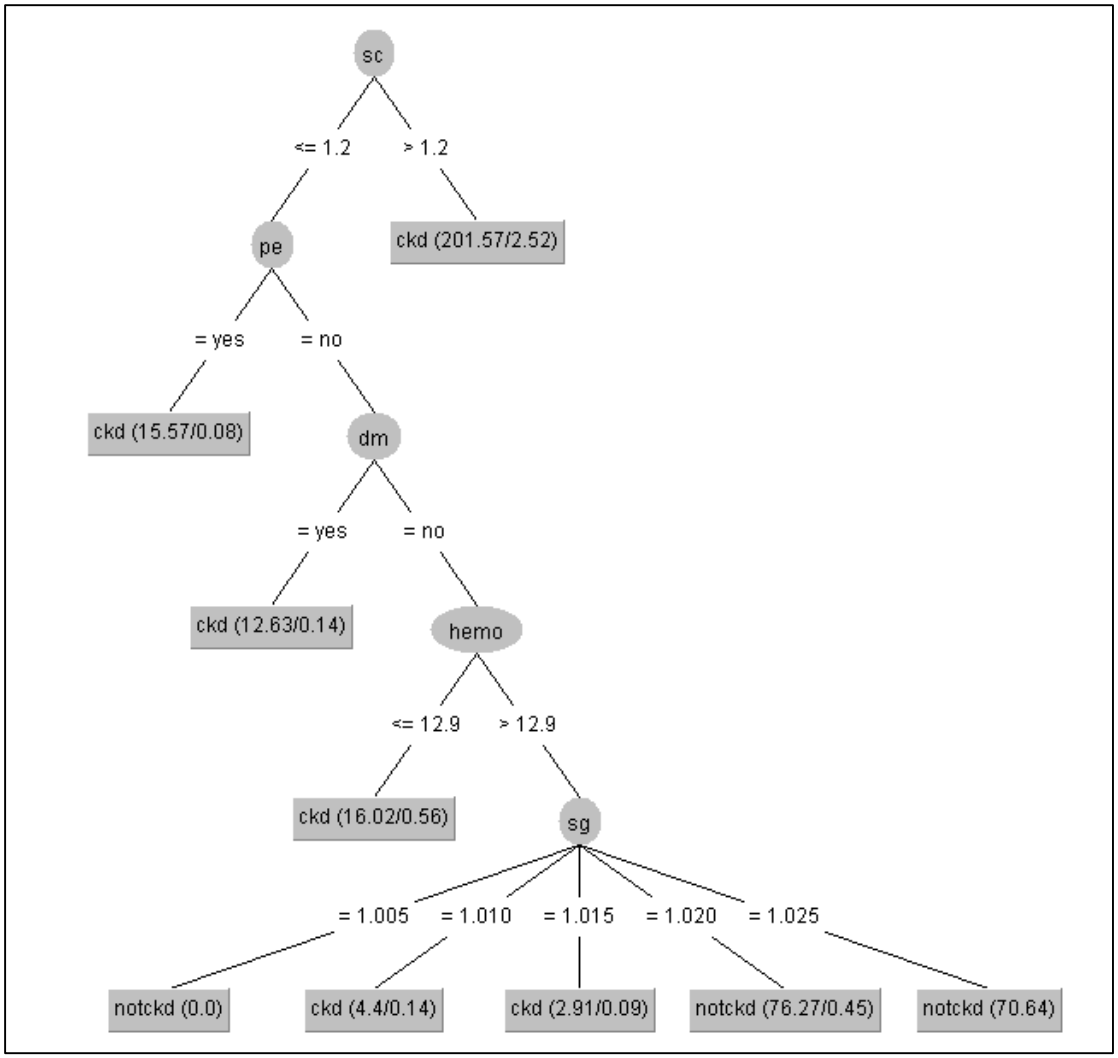

Figure (5.9) J48 Tree

Table 5.7 Summary of J48 Accuracy

Summary	Instances	%
Correctly Classified Instances	388	97
Incorrectly Classified Instances	12	3

Table 5.8 J48 Confusion Matrix

Predicted		Confusion Matrix	
ckd	**notckd**		
248	2	**ckd**	Actual
10	**140**	**notckd**	

Table 5.9 Details of J48 Accuracy

	TP Rate	FP rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.992	0.067	0.961	0.992	0.976	0.936	0.997	0.999	ckd
	0.933	0.008	0.986	0.933	0.959	0.936	0.997	0.992	notckd
Weighted Avg.	0.970	0.045	0.970	0.970	0.970	0.936	0.997	0.996	

5.5.3.3 Bayes Net Model

The third model was built using Bayes Net algorithm, which also achieved high accuracy (98.7) and very low error rate (MAE=0.012) as shown in table 5.2 Algorithm Comparison. Below are more details about the model:

a. **Feature Selection:** The algorithm selected 7 attributes: (***sg, al, sc, hemo, pcv, dm ,appet***). All attributes selected by Bayes Net, were also selected by other algorithms, this gives indication about the success of this model in attribute selection process. ***sg***, are selected only by this algorithm. Also the model succeeded in selecting the relevant attributes as per the doctors' feedback which is shown in the discussion part.

b. **Classification Model**: The model shows very high success rate in patients' classification. That is very clear from the number of correctly classified instances in table 5.7, which are 397 out of 400 instances that means the model accuracy is 99.25%, this is the highest result among the seven models. During algorithm comparison phase the model achieved only 98.7 because at that time the model was built using the full set of attributes (24 attributes), but after attribute selection the model achieved better results, because the irrelevant attributes were eliminated and only the seven selected attributes were used to develop the final model. Figure 5.7 shows graphical representation of the attributes that were used to predict the class.

c. **Model Evaluation**: The model evaluation is represented by the confusion matrix as shown in table 5.11. The model succeeded to classify 248 out of 250 ***ckd*** patient, and 149 out of 150 ***notckd*** patients. Table 5.12 shows the details of model evaluation, as shown in the table the true positive rate (TP Rate) is 0.992 for ***ckd***, and it is 0.993 for ***notckd***. Also the model Precision and Recall are very high, Precision is 0.996 for ***ckd***, and 0.987 for ***notckd***, while the Recall is 0.992 for ***ckd*** and 0.993 for ***notckd***.

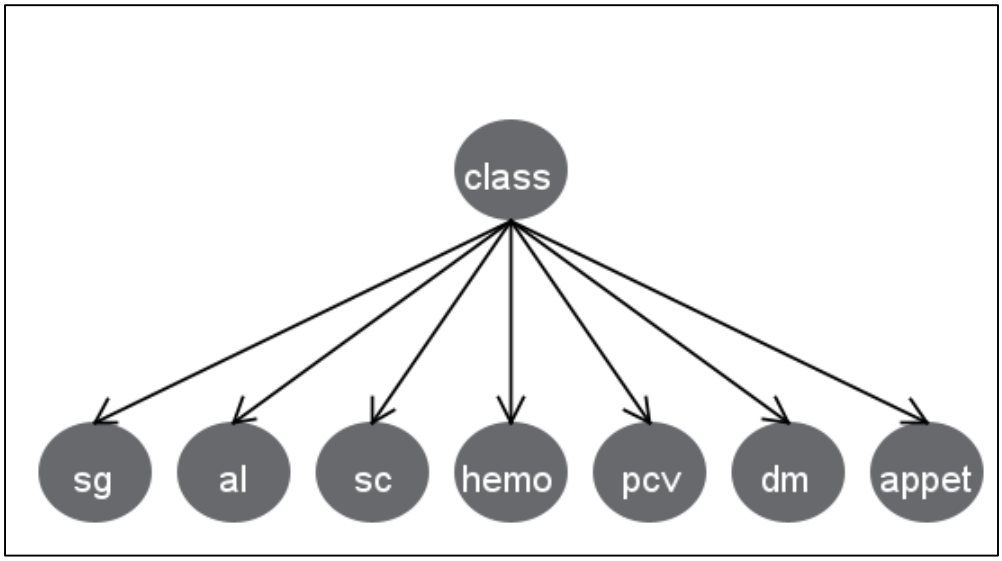

Figure (5.10) Bayes Net

Table 5.10 Summary of Bayes Net Accuracy

Summary	Instances	%
Correctly Classified Instances	397	99.25
Incorrectly Classified Instances	3	0.75

Table 5.11 Bayes Net Confusion

Predicted		Confusion Matrix	
ckd	**notckd**		
248	2	**ckd**	Actual
1	**149**	**notckd**	

Table 5.12 Details of Bayes Net Accuracy

	TP Rate	FP rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.992	0.007	0.996	0.992	0.994	0.984	0.999	1.000	ckd
	0.993	0.008	0.987	0.993	0.990	0.984	0.999	0.999	notckd
Weighted Avg.	0.993	0.007	0.993	0.993	0.993	0.984	0.999	0.999	

5.5.3.4 LMT Model

The fourth model was built using LMT algorithm, which also achieved high accuracy (98.7) same as Bayes Net, and very low error rate (MAE=0.022) as shown in table 5.2 Algorithm Comparison. Below are more details about the model:

a. **Feature Selection:** The algorithm selected 7 attributes: (***sg, al, su, sod, hemo, rbcc, htn***). Two of these attributes – ***sod*** and ***su*** - were only selected by this algorithm, this

reflects the benefit of using different algorithms for the same problem, because not all algorithms can select all relevant attributes. When referring back to doctors, they depicted the importance of ***sod*** and ***su*** in the diagnosis of Chronic Kidney Disease.

b. Classification Model: The model shows very high success rate in patients' classification. That is very clear from the number of correctly classified instances in table 5.13, which are 389 out of 400 instances that means the model accuracy is 97.25%. During algorithm comparison phase the model achieved 98.7 because at that time the model was built using the full set of attributes (24 attributes), but after attribute selection the model achieved lower results.

c. Model Evaluation: The model evaluation is represented by the confusion matrix as shown in table 5.11. The model succeeded to classify 242 out of 250 ***ckd*** patient, and 147 out of 150 ***notckd*** patients. Table 5.15 shows the details of model evaluation, as shown in the table the true positive rate (TP Rate) is 0.968 for ***ckd***, and it is 0.980 for ***notckd***. Also the model Precision and Recall are very high, Precision is 0.988 for ***ckd***, and 0.948 for ***notckd***, while the Recall is 0.968 for ***ckd*** and 0.980 for ***notckd***.

Table 5.13 Summary of LMT Accuracy

Summary	Instances	%
Correctly Classified Instances	389	97.25
Incorrectly Classified Instances	11	2.75

Table 5.14 LMT Confusion Matrix

Predicted		Confusion Matrix	
ckd	**notckd**		
242	8	**ckd**	Actual
3	**147**	**notckd**	

Table 5.15 Details of LMT Accuracy

	TP Rate	FP rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.968	0.020	0.988	0.968	0.978	0.942	0.997	0.999	ckd
	0.980	0.032	0.948	0.980	0.964	0.942	0.997	0.996	notckd
Weighted Avg.	0.973	0.025	0.973	0.973	0.973	0.942	0.997	0.998	

5.5.3.5 Simple Logistic

The fifth model was built using Simple Logistic algorithm, which also achieved high accuracy (98.7) same as Bayes Net, and very low error rate (MAE=0.022) as shown in table 5.2 Algorithm Comparison. Below are more details about the model:

a. Feature Selection: The algorithm selected only 3 attributes: (***sg, hemo, rbcc***). The model failed to select the most relevant attributes. When referring back to doctors, they

claimed that the model failed to identify the relevant attributes of Chronic Kidney Disease.

b. Classification Model: The model shows high success rate in patients' classification. That is very clear from the number of correctly classified instances in table 5.16, which are 387 out of 400 instances that means the model accuracy is 96.75%. During algorithm comparison phase the model achieved 98.7%, the accuracy is lower because three features out of 24 cannot give better accuracy, but still the model provides good information that these factors are highly related to the disease.

c. Model Evaluation: The model evaluation is represented by the confusion matrix as shown in table 5.17. The model succeeded to classify 241 out of 250 ***ckd*** patient, and 146 out of 150 ***notckd*** patients. Table 4.18 shows the details of model evaluation, as shown in the table the true positive rate (TP Rate) is 0.964 for ***ckd***, and it is 0.973 for ***notckd***. Also the model Precision and Recall are very high, Precision is 0.984 for ***ckd***, and 0.942 for ***notckd***, while the Recall is 0.964 for ***ckd*** and 0.973 for ***notckd***.

Table 5.16 Summary of Simple Logistic

Summary	Instances	%
Correctly Classified Instances	387	96.75
Incorrectly Classified Instances	13	3.25

Table 5.17 Simple Logistic Confusion

Predicted		Confusion Matrix	
ckd	**notckd**		
241	9	**ckd**	Actual
4	**146**	**notckd**	

Table 5.18 Details of Simple Logistic

	TP Rate	FP rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.964	0.027	0.984	0.964	0.974	0.931	0.995	0.997	ckd
	0.973	0.036	0.942	0.973	0.957	0.931	0.995	0.992	notckd
Weighted Avg.	0.968	0.030	0.968	0.968	0.968	0.931	0.995	0.995	

5.5.3.6 Multi-Layer Perceptron with one Hidden layer

The sixth model was built using Neural Network (Multi-Layer Perceptron) algorithm. This algorithm was not part of the algorithms' comparison experiment, because it elapsed much time when run as part of the experiment. Two models were developed using the neural network, this model and the next model. For this model one hidden layer was used as per Weka default setting. Figure 5.8 shows the Multi-Layer Perceptron network with

one hidden layer with three neurons, and 0.3 learning rate. The five selected attributes represent the input layer, and the output layer is the class which is represented by two nodes (***ckd***, ***notckd***). Below are more details about the model:

a. **Feature Selection:** The model selected only 5 attributes: (***sg, hemo, al, pcv, rbc***). This is the only model that selected the ***rbc*** attribute. The rest of attributes were also selected by the other algorithms. When referring back to doctors, they claimed that the model failed to identify the relevant attributes of Chronic Kidney Disease.
b. **Classification Model**: The model shows low success rate in patients' classification compared to other models. That is very clear from the number of correctly classified instances in table 5.19, which are 378 out of 400 instances that means the model accuracy is 94.5%.
c. **Model Evaluation**: The model evaluation is represented by the confusion matrix as shown in table 5.20. The model succeeded to classify 248 out of 250 ***ckd*** patient which is very good result, but the model can predict only 130 out of the 150 ***notckd*** patients. Table 5.21 shows the details of model evaluation, as shown in the table the true positive rate (TP Rate) is 0.992 for ***ckd***, and it is 0.867 for ***notckd*** which is very low compared to other models. Also the model Precision and Recall are very high, Precision is 0.925 for ***ckd***, and 0.985 for ***notckd***, while the Recall is 0.992 for ***ckd*** and 0.867 for ***notckd***.

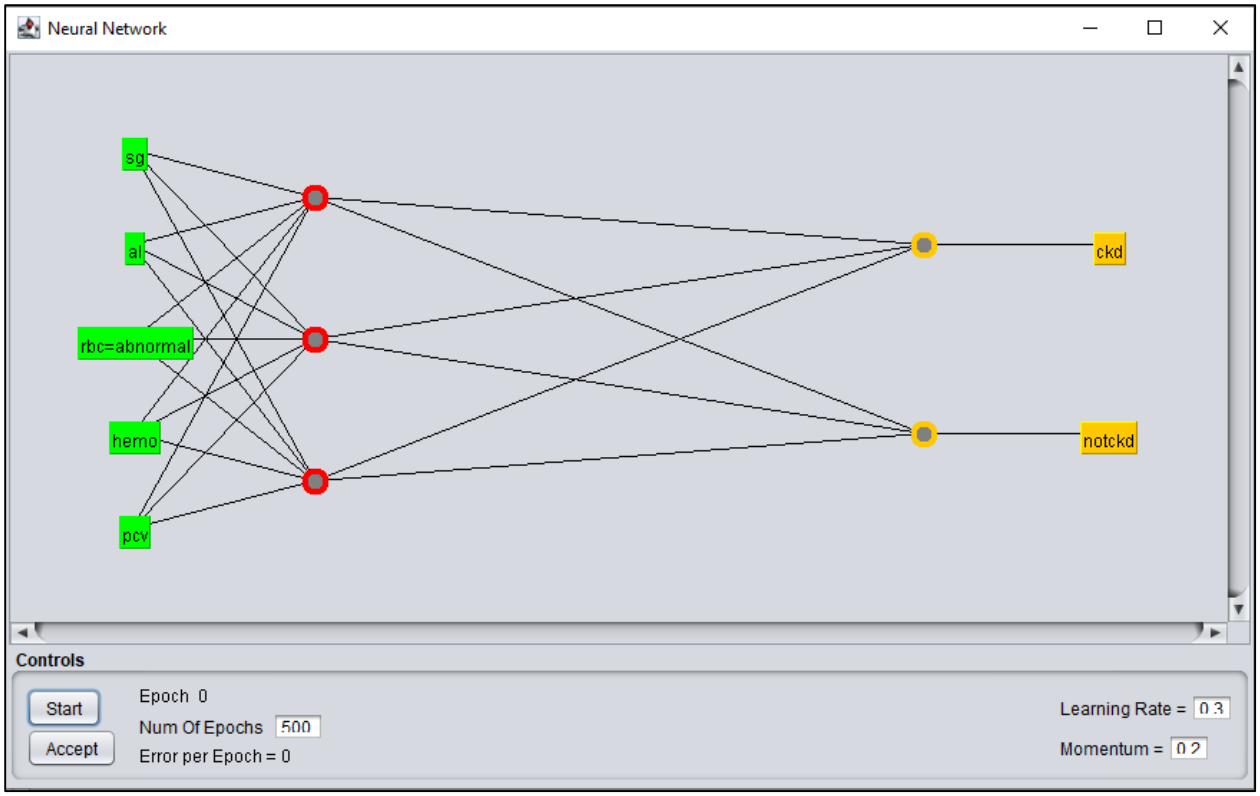

Figure (5.11) MLP – with one hidden layer

Table 5.19 Summary of MLP-1 hidden Accuracy

Summary	Instances	%
Correctly Classified Instances	378	94.5
Incorrectly Classified Instances	22	5.5

Table 5.20 MLP-1 hidden Confusion

Predicted		Confusion Matrix	
ckd	**notckd**		
248	2	**ckd**	Actual
20	**130**	**notckd**	

Table 5.21 Details of MLP-1 hidden

	TP Rate	FP rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.992	0.133	0.925	0.992	0.958	0.884	0.975	0.982	ckd
	0.867	0.008	0.985	0.867	0.922	0.884	0.975	0.970	notckd
Weighted Avg.	0.945	0.086	0.948	0.945	0.944	0.884	0.975	0.978	

5.5.3.7 Multi-Layer Perceptron with Three Hidden Layer

The seventh model - like the previous model - was built using Neural Network (Multi-Layer Perceptron) algorithm. The difference is that for this model three hidden layer were used, with 5 neurons for the first layer, 10 for the second layer, and 20 neurons for the third hidden layer. The learning rate remains the same as the previous model which is 0.3. Figure 5.9 shows the Multi-Layer Perceptron network with the three hidden layers. The five selected attributes represent the input layer, and the output layer is the class which is represented by two nodes (***ckd***, ***notckd***). Below are more details about the model:

a. Feature Selection: The model selected the same 5 attributes which were selected by the previous one: (***sg, hemo, al, pcv, rbc***). So in the feature selection part the two Multi-Layer Perceptron models are identical.

b. Classification Model: The model shows better success rate in patients' classification compared to the one layer neural network model. That is very clear from the number of correctly classified instances in table 5.22, which are 392 out of 400 instances that means the model accuracy is 98% compared to 94.5% for the previous model, this reflects the enhancement in accuracy which is caused by the additional layers.

c. Model Evaluation: The model evaluation is represented by the confusion matrix as shown in table 5.23. The model succeeded to classify 247 out of 250 ***ckd*** patient which is very good result, and 145 out of the 150 ***notckd*** patients. Table 5.25 shows the details of model evaluation, as shown in the table the true positive rate (TP Rate) is 0.988 for ***ckd***, and it is 0.967 for ***notckd*** which is very low compared to other models. Also the model Precision and Recall are very high, Precision is 0.980 for ***ckd***, and 0.980 for ***notckd***, while the Recall is 0.988 for ***ckd*** and 0.967 for ***notckd***.

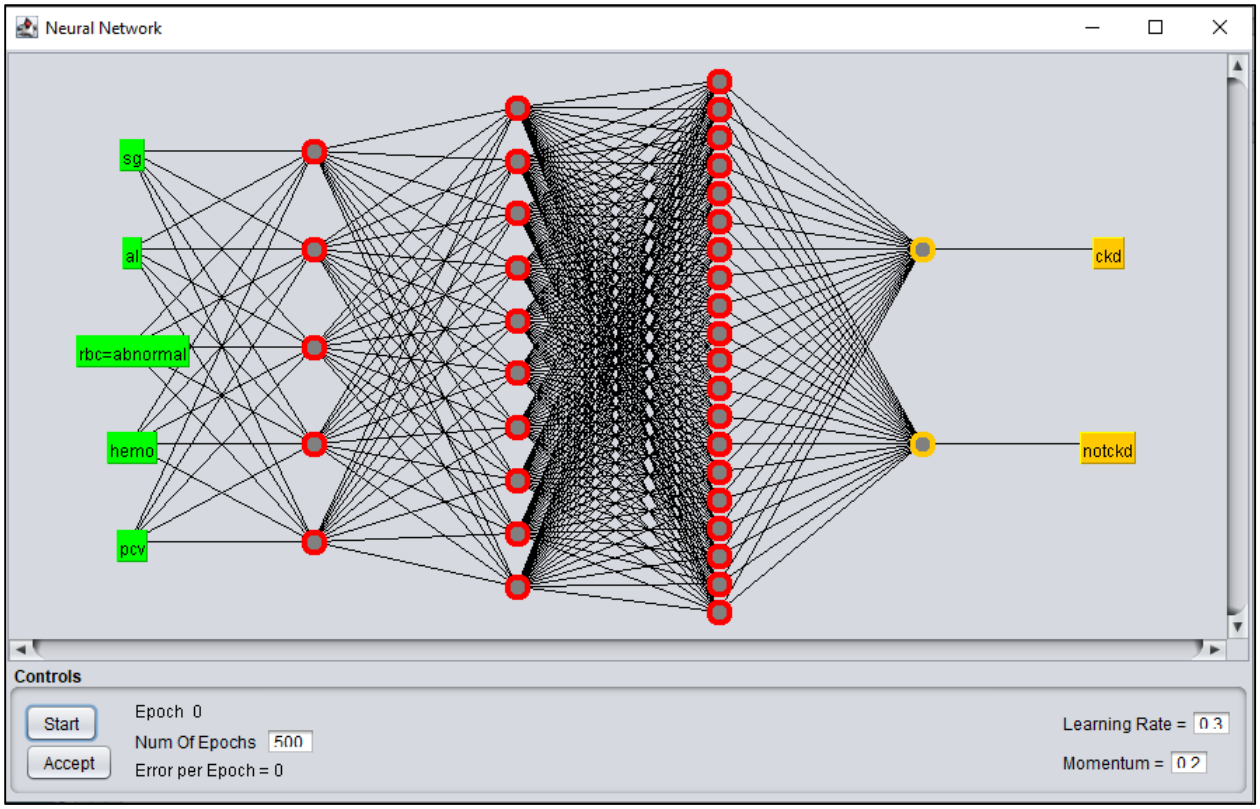

Figure (5.12) MLP – with three hidden layer

Table 5.22 Summary of MLP-3 hidden Accuracy

Summary	Instances	%
Correctly Classified Instances	392	98
Incorrectly Classified Instances	8	2

Table 5.23 MLP-3 hidden Confusion Matrix

Predicted		Confusion Matrix	
ckd	**notckd**		
247	3	**ckd**	Actual
5	**145**	**notckd**	

Table 5.24 Details of MLP-3 hidden Accuracy

	TP Rate	FP rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.988	0.033	0.980	0.988	0.984	0.957	0.988	0.990	ckd
	0.967	0.012	0.980	0.967	0.973	0.957	0.988	0.989	notckd
Weighted Avg.	0.980	0.025	0.980	0.980	0.980	0.957	0.988	0.990	

5.6 Summary and Results' Discussion

The discussion about achieved results and models could be divided into two main parts: the *attribute selection* and *classification model*. Below is the summary and brief discussion about these parts.

5.6.1 Conclusion and Discussion of Attribute Selection

Feature selection results are compatible with the doctors' expectations about the attributes. According to the attribute selection result, the features could be ranked into five groups. Below is summary of observations and discussion of the attribute selection results with the Nephrology and urinary tract specialist.

The first group consists of ***sg*** and ***hemo*** are the top two attributes, both of them were selected by the six algorithms. Among parameters, ***sg*** and ***hemo*** gives the highest measurements (100%) as both of them were selected by all algorithms. Accordingly, they could be considered as highly sensitive parameters during early stages of CKD.

The second group of attribute consist of ***al, pcv, rbcc*** and ***dm*** all of them were selected by three algorithms. Regarding ***al***, ***pcv***, ***rbcc*** and ***dm*** all of them were selected by three algorithms, so they gave considerable value in diagnosing early stages of CKD.

The third group of attribute was selected by two algorithms, this group consist of ***sc, htn, appet, pe*** also this is expected because. Since ***sc***, ***htn***, ***appet*** and ***pe*** show impairment towards advanced stages of CKD, they give low sensitivity in detecting early stages of CKD.

The fourth group is composed of ***age, bp, su, rbc, pc, ba, bgr*** and ***sod*** . This group was selected by only one algorithm; they give the measurements (17%) according to the obtained results this could be interpreted as the less important attributes.

The last group which is composed of ***pcc***, ***bu***, ***pot***, ***wbcc***, ***cad*** and ***ane*** was not selected at all. This shows the negative role in diagnosing early stages of CKD for these attributes.

So the conclusion of this part is that, the feature selection result is compatible with the expectations of the Nephrology and urinary tract specialist.

5.6.2 Classification Model Results and Discussion

Seven models were built using five different classification algorithms, plus Neural Network with two options. The diversity of algorithms enriches the interpretation of the problem, some models successfully selected the attributes that are directly related to the CKD, others give high accuracy, although the models that achieved 100% accuracy were rejected to mitigate the over fitting problem.

Referring to the results achieved and discussed in details in the above section we can say that the research succeeded to answer the research problems.

The last chapter of this research provides the conclusion and the future work which could be done to enhance the research and open the door to other researchers to provide more.

CHAPTER SIX

CONCLUSION AND FUTURE WORK

CHAPTER VI : CONCLUSION AND FUTURE WORK

6.1 Introduction

This chapter provides conclusion of the research by providing a summary of the method followed to achieve the research objectives. This research used data mining to select the attributes that affects the Chronic Kidney Disease and develop a prediction model that can predict the CKD in its early stage. The research consists of two main parts firstly to specify the factors that have direct effect of the disease. Secondly to use suitable data mining algorithms to predict CKD.

Out of 25 classification algorithms; seven were selected, Random Forest, J48, Bayes Net, LMT , Simple Logistic and Neural Network with one and three hidden layers. These algorithms were selected because they showed in the algorithm selection experiment.

6.2 The Proposed Method

Data mining was used to achieve the research objectives. The first objective is precisely identifying the factors that lead to CKD. The method which was followed to achieve this is the feature selection technique. Wrapper method was used to enable the researcher to select the feature and develop the model in one step, because this will guarantee that the results achieved will be more accurate and the experiments were much easier.

The second objective is to predict the disease in its early stages using the selected set of features. Classification models were developed to achieve this objective using the selected set of features. Results from different 25 algorithms were compared, the top five plus the Neural Network algorithms were selected to develop different models.

The third objective is to provide better understanding of the relation between the laboratory tests and the disease, and this was done by discussion the models with the Urologists and

their feedback about the results achieved and the created models, specially the Classification tree model.

6.3 Contribution of the Study

As mentioned earlier in chapter one Introduction. The main goal of this research is to use data mining to identify the factors that lead to chromic kidney disease by analyzing the laboratory test results of patients, and to predict the CKD in its early stages, and to highlight the relation between these tests and the disease. The contribution of this thesis is to show the benefits that could be gained from using data mining in achieving these goals. This contribution could be summarized as follows:

i. Can we identify more precisely the factors that lead to chromic kidney disease by analyzing the laboratory test results of patients?

To identify the factors that lead to CKD, feature selection technique was used. Because six different algorithms were used, there are different feature sets. So the result of the feature selection phase was wrapped up in one table, the attribute which was selected by many algorithms was considered to be the most important one. The result of this phase was presented and discussed with the doctors, who provided their feedback and assure the correctness of the results.

ii. Prediction of Kidney failure in its early stage using patients' laboratory test results

After selecting the most important features among many laboratory tests that directly affect the CKD, the selected set of features was used to develop the classification model to predict the disease it its early stage. Also the results of this part was submitted and discussed with the Urologists, who were impressed by the results and techniques.

iii. Provide better understanding about selected features and the CKD

Some of the developed model has provided a very clear understanding of the relation between the disease and the laboratory tests, and between the tests themselves.

That is very clear from the J48 model (Classification tree), because the model started by ***sc***, as the root node of the tree, and this is totally true because it is one of the most important factors and which shows highest information gain among all the attributes, then ***pe*** was selected to branch the tree, then ***dm***, then ***hemo***, then the last attribute which is ***sg***. Also that totally matches the Urologists understanding.

6.4 Future Work

This research achieved the objectives of the study. However, the data used in the research was downloaded from the internet and it is a real world data of patients from India. The current work has focused only on selecting the attributes and building the classification model. Although, in future work, we are planning to do the following:

- Work closely with the kidney specialized hospitals to collect data of Sudanese patients, because this may reveal other findings.
- Other factors could be added like the food style, area of Sudan, disease history in family and other diseases.
- Build a central data warehouse for patient's data to enhance the analysis efficiency.
- If the patient data is available in a central data warehouse it could be used to predict many diseases in their early stages.
- If the patient data is available in a central data warehouse it could be used to find relations between diseases themselves.
- Use other visualization tools like Power BI or Tableau to give better visualization of the data sets.
- Develop an AI expert system that uses the methods and algorithms used in this research to instantly predict the disease according to his/her laboratory test results.

The above points are proposed to be done by a research group that consists of doctors and data scientists.

6.5 Summary

The objectives of this research are to use data mining techniques to solve three problems. Is to identify precisely the factors that lead to chromic kidney disease by analyzing the laboratory test results of patients. Secondly to predict the Kidney failure using the laboratory test results of CKD patients. Lastly to provide better understanding about these factors and the relations between them, and their effect on the disease.

This chapter presents the summaries of the methods, tools, and processes used in this research to achieve these goals. Moreover, the research describes exactly how the data was explored, and how experiments were done to assist other researcher in the future to go in the same track and achieve better results. This research shows that using data mining in disease prediction gives better results. Also different algorithms show different results, so according to the available dataset results may differ. The research is composed mainly of two parts: the first one is to specify the factors that have direct effect on disease, and the second is to predict the disease in its early stage, this was done using Weka.

The above results show that all research goals have been achieved. Moreover, the research provides ideas for future work to enhance the current work and gives better life of our society by using the data mining tools to predict diseases in their early stages specially the Chronic Kidney Disease.

References

Nadri, H, Rahimi, B, Timpka, T & Sedghi, S 2017, 'The top 100 articles in the medical informatics: a bibliometric analysis', *Journal of Medical Systems.*

NICHSR.com,(2020). [online] Available at: https://hsric.nlm.nih.gov/hsric_public/topic/informatics/ [Accessed 13 Feb 2020].

O'donoghue, J & Herbert, J 2012, 'Data management within mHealth environments: Patient sensors, mobile devices, and databases', *Journal of Data and Information Quality.*

Mettler, T & Raptis, DA 2012, 'What constitutes the field of health information systems? fostering a systematic framework and research agenda', *Health Informatics Journal.*

Daniel, C et al. 2013, 'Biomedical informatics and health research translational informatics in medical informatics', pp. 463–493, ISBN 978-2-8178-0337-1 .

O'Donoghue et al. 2011, 'Modified early warning scorecard: the role of data/information quality within the decision making process', *Electronic Journal Information Systems Evaluation.*

searchhealthit.com(2020).[online]Available at: https://searchhealthit.techtarget.com/definition/Health-IT-information-technology [Accessed 13 Feb 2020].

usfhealthonline.com(2020). [online] Available at: https://www.usfhealthonline.com/resources/key-concepts/what-is-health-informatics [Accessed 13 Feb 2020].

libguides.com, (2020). [online] Available at: https://libguides.eastern.edu/Nurs301 [Accessed 20 Mar 2020].

healthit.com,(2020). [online] Available at: https://www.healthit.gov/sites/default/files/pdf/health-information-technology-fact-sheet.pdf [Accessed 20 Mar 2020].

healthline.com,(2020). [online] Available at: https://www.healthline.com/health/human-body-maps/kidney#kidney-diagram [Accessed 25 Mar 2020].

niddk.com,(2020). [online] Available at: www.niddk.nih.gov/health-information/kidney-disease/chronic-kidney-disease-ckd [Accessed 25 Mar 2020].

worldlifeexpectancy.com,(2020). [online] Available at:https://www.worldlifeexpectancy.com/sudan-kidney-disease [Accessed 1 May 2020].

kdd.org,(2020). [online] Available at:https://www.kdd.org/curriculum/index.html [Accessed 1 May 2020].

researchgate.net, (2020). [online] Available at: https://www.researchgate.net/figure/Intersection-of-data-mining-with-different-disciplines_fig1_334825343 [Accessed 11 May 2020].

sas.com, (2020). [online] Available at: https://www.sas.com/en_us/insights/analytics/data-mining.html [Accessed 11 May 2020].

saedsayad.com,(2020). [online] Available at: https://www.saedsayad.com/data_mining_map.htm [Accessed 11 May 2020].

Ho & Tin Kam 1995, 'Random decision forests', Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal .

researchgate.net(2020). [online] Available at: https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d [Accessed12 May 2020].

researchgate.net(2020). [online] Available at: https://www.researchgate.net/post/What_is_the_algorithm_of_J48_decision_tree_for_classification [Accessed 12 May 2020].

researchgate.net(2020). [online] Available at: https://medium.com/@chiragsehra42/decision-trees-explained-easily-28f23241248 [Accessed 12 May 2020].

bayesserver.com,(2020). [online] Available at: https://www.bayesserver.com/docs/introduction/bayesian-networks [Accessed 12 May 2020].

Pearl & Judea 2000, 'Causality: models, reasoning, and inference', Cambridge University Press. ISBN:978-0-521-77362-1.

Niels Landwehr & Mark & Eibe 2003, 'Logistic model trees'.

Sumner et al. 2005, 'Speeding up logistic model tree induction', pp. 675–683.

statisticssolutions.com,(2020). [online] Available at: https://www.statisticssolutions.com/what-is-logistic-regression [Accessed 25 May 2020].

javatpoint.com,(2020). [online] Available at: https://www.javatpoint.com/logistic-regression-in-machine-learning [Accessed 26 May 2020].

Chen et al. 2019, 'Design and implementation of cloud analytics-assisted smart power meters considering advanced artificial intelligence as edge analytics in demand-side management for smart homes sensors' .

linkedin.com,(2020). [online] Available at: https://missinglink.ai/guides/neural-network-concepts/perceptrons-and-multi-layer-perceptrons-the-artificial-neuron-at-the-core-of-deep-learning/ [Accessed 26 May 2020].

Freund, Y & Schapire 1999, 'R. E, Large margin classification using the perceptron, algorithm'.

lifeoptions.org,(2020). [online] Available at: https://lifeoptions.org/learn-about-kidney-disease/causes-and-stages/[Accessed 1 Jun 2020].

mayoclinic.org,(2020). [online] Available at: https://www.mayoclinic.org/diseases-conditions/chronic-kidney-disease/symptoms-causes/syc-20354521 [Accessed 1 Jun 2020].

Mohamed E & Elham G 2011, 'Causes of End-Stage Renal Disease in Sudan: A Single-Center Experience', Nephrology Department, Gazira Hospital for Renal Diseases, Khartoum, North Sudan .

Hans-Joachim, A et al. , 'Immunonephrology working group of ERA-EDTA'.

Daniele, R et al. 2017, 'Deep learning for health informatics', IEEE Journal OF Biomedical And Healthinformatics, Vol. 21, No. 1.

Claudio & Alexis 2020, 'Development of an expert system for pre-diagnosis of hypertension, diabetes mellitus type 2 and metabolic syndrome', Mignogna University of Santiago of Chile.

cms.gov,(2020). [online] Available at: https://www.cms.gov/Medicare/E-Health/EHealthRecords [Accessed 1 Jun 2020].

Jeffery, J et al. 2020, 'Rapid response to COVID-19: health informatics support for outbreak management in an academic health system' , Department of Surgery, University of California.

Shaik, MD et al. 2017, 'Big Data Analytics for Health on Kidney Disease', Universityof Chennai.

Tabassum et al. 2017, 'Analysis and prediction of chronic kidney disease using data mining techniques' , Meenakshi Institute of Technology, Bengaluru, India.

Basma et al. 2016, 'Performance of data mining techniques to predict in healthcare case study: chronic kidney failure disease'.

saedsayad.com, (2020). [online] Available at: https://www.saeds.ayad.com/data_mining.htm [Accessed 1 Jun 2020].

saedsayad.com,(2020). [online] Available at: https://www.saedsayad.com/data_exploration.htm [Accessed 1 Jun 2020].

uci.edu,(2020). [online] Available at: http://archive.ics.uci.edu/ml//datasets/Chronic_Kidney_Disease [Accessed 1 Jun 2019].

medscape.com,(2020). [online] Available at: https://emedicine.medscape.com /article/2090711-overview [Accessed 3 Jun 2020].

urmc.rochester.edu,(2020). [online] Available at: https://www.urmc.rochester.edu/encyclopedia/content.aspx?contenttypeid=167&cont entid=albumin_blood [Accessed 3 Jun 2020].

nih.gov,(2020). [online] Available at: https://pubmed.ncbi.nlm.nih.gov/18840763/[Accessed 3 Jun 2020].

healthline.com,(2020). [online] Available at: https://www.healthline.com/health/rbc-count [Accessed 3 Jun 2020].

medicalnewstoday.com,(2020). [online] Available at: https://www.medicalnewstoday.com/articles/321964 [Accessed 3 Jun 2020].

merckmanuals.com,(2020). [online] Available at: https://www.merckmanuals.com/home/infections/diagnosis-of-infectious-disease/diagnosis-of-infectious-disease [Accessed 15 Jun 2020].

medicalnewstoday.com,(2020). [online] Available at: https://www.medicalnewstoday.com/articles/323022 [Accessed 25 Jun 2020].

healthline.com,(2020). [online] Available at: https://www.healthline.com/health/bun [Accessed 25 Jun 2020].

mayoclinic.org,(2020). [online] Available at: https://www.mayoclinic.org/tests-procedures/creatinine-test/about/pac-20384646 [Accessed 25 Jun 2020].

medlineplus.gov,(2020). [online] Available at: https://medlineplus.gov/lab-tests/sodium-blood-test [Accessed 1 Jul 2020].

medlineplus.gov,(2020). [online] Available at: https://medlineplus.gov/lab-tests/potassium-blood-test [Accessed 1 Jul 2020].

healthline.com,(2020). [online] Available at: https://www. healthline.com /health/hgb [Accessed 1 Jul 2020].

portea.com,(2020). [online] Available at: https://www.portea.com/labs/diagnostic-tests/packed-cell-volume-test [Accessed 13 Jul 2020].

healthline.com,(2020). [online] Available at: https://www.healthline.com/health/wbc-count [Accessed 13 Jul 2020].

healthline.com, (2020). [online] Available at: https://www.healthline.com/health/rbc-count [Accessed 13 Jul 2020].

mayoclinic.org,(2020). [online] Available at: https://www.mayoclinic.org/diseases-conditions/high-blood-pressure/symptoms-causes/syc-20373410 [Accessed 25 Jul 2020].

niddk.nih.gov,(2020). [online] Available at: https://www.niddk.nih.gov/healthinformation/diabetes/overview/preventing-problems/diabetic-kidney-disease [Accessed 25 Jul 2020].

clevelandclinic.org,(2020). [online] Available at: https://my.clevelandclinic.org/health/diseases/16898-coronary-artery-disease [Accessed 1 Aug 2020].

medicinenet.com,(2020). [online] Available at: https://www.medicinenet.com/loss_of_appetite/symptoms.htm [Accessed 13 Aug 2020].

mayoclinic.org,(2020). [online] Available at: https://www.mayoclinic.org/diseases-conditions/edema/symptoms causes/syc-20366493 [Accessed 14 Sep 2020].

mayoclinic.org,(2020). [online] Available at: https://www.mayoclinic.org/diseases-conditions/anemia/symptoms-causes/syc-20351360 [Accessed 14 Sep 2020].