A MACHINE LEARNING METHODOLOGY FOR DIAGNOSING CHRONIC KIDNEY DISEASE AND SEVERITY ANALYSIS

A PROJECT REPORT

Submitted by

ADWAITH B S

REG NO : TKM19MCA001

to

APJ Abdul Kalam Technological University

In partial fulfillment of the requirements for the award of the Degree of

MASTER OF COMPUTER APPLICATIONS

Thangal Kunju Musaliar College of Engineering
Kerala

DEPARTMENT OF COMPUTER APPLICATIONS

MAY 2022

DECLARATION

I undersigned hereby declare that the project report **A MACHINE LEARNING METHODOLOGY FOR DIAGNOSING CHRONIC KIDNEY DISEASE AND SEVERITY ANALYSIS** , submitted for partial fulfillment of the requirements for the award of degree of Master of Computer Applications of the APJ Abdul Kalam Technological University, Kerala is a bonafide work done by me under supervision of **Prof.NATHEERA BEEVI M**. This submission represents my ideas in my own words and where ideas or words of others have been included, I have adequately and accurately cited and referenced the original sources. I also declare that I have adhered to ethics of academic honesty and integrity and have not misrepresented or fabricated any data or idea or fact or source in my submission. I understand that any violation of the above will be a cause for disciplinary action by the institute and/or the University and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been obtained. This report has not been previously formed the basis for the award of any degree, diploma or similar title of any other University.

Kollam

12/05/2022

ADWAITH B S

DEPARTMENT OF COMPUTER APPLICATIONS
TKM COLLEGE OF ENGINEERING

C E R T I F I C A T E

This is to certify that, the report entitled **A MACHINE LEARNING METHODOLOGY FOR DIAGNOSING CHRONIC KIDNEY DISEASE AND SEVERITY ANALYSIS** submitted by **ADWAITH B S**, to the **APJ Abdul Kalam Technological University** in partial fulfillment of the requirements for the award of the Degree of **Master of Computer Applications** is a bonafide record of the project work carried out by him under our guidance and supervision.This report in any form has not been submitted to any other University or Institute for any purpose.

Internal Supervisor | Head of the Department | External Examiner

ACKNOWLEDGEMENT

First and foremost I thank GOD almighty and my parents for the success of this project. I owe sincere gratitude and heart full thanks to everyone who shared their precious time and knowledge for the successful completion of our project.

I am extremely grateful to **Dr. Fousia M Shamsudeen**,Head of the Department, for providing us with best facilities.

I would like to thank my project guide **Prof. Natheera Beevi M**, Department of Computer Applications, who motivated me throughout this project. who motivated me throughout the project.

I profusely thank all other faculty members in the department and all other members of TKM College of Engineering, for their guidance and inspirations throughout our course of study.

I owe my thanks to our friends and all others who have directly or indirectly helped us in the successful completion of this project..

ADWAITH B S

ABSTRACT

Chronic kidney disease (CKD) is a global health issue with a high rate of death and morbidity, and it is a cause of other diseases. In the early stages of chronic kidney disease, there are no evident symptoms, therefore patients frequently fail to recognise the disease. Detection of CKD at an early stage enables patients to obtain timely treatment to mitigate the disease's progression. Due to their rapid and precise recognition capabilities, machine learning models can aid doctors in achieving this objective. In this paper, I suggest a technique for diagnosing CKD based on machine learning. The CKD data set was taken from the machine learning repository at the University of California, Irvine (UCI), which has a substantial number of missing values. Six machine learning algorithms (logistic regression, random forest, support vector machine, k-nearest neighbour, naive Bayes classifier, and multilayer perceptron) were employed to create models after effectively completing the missing data set. The best result among these machine learning models was achieved by random forest with a diagnostic accuracy of 99 percent. By examining the errors produced by the existing models, I presented a model that combines logistic regression with random forest. This recommended method is 99 percent accurate. Also, the severity of CKD is anticipated based on the individual's clinical data.

Contents

List of Figures

Chapter 1

INTRODUCTION

Chronic kidney disease, commonly known as chronic kidney failure, is characterised by a progressive decline in kidney function. Kidneys filter blood to remove wastes and excess fluids, which are then eliminated as urine. Advanced chronic renal disease can lead to harmful accumulations of fluid, electrolytes, and wastes. In the beginning stages of chronic kidney disease,it may exhibit minimal symptoms.The person not recognise kidney disease until the issue has progressed. The goal of treatment for chronic kidney disease is to reduce the course of kidney damage, typically by addressing the underlying cause. However, addressing the underlying cause may not prevent the progression of kidney disease. Chronic kidney disease can proceed to kidney failure in its last stages, which is fatal without artificial filtration (dialysis) or a kidney transplant. Increasing frequency of diabetic patients, hypertension, heart disease, diabetes, and a family history of kidney failure are high-risk factors for chronic kidney disease. If CKD goes unnoticed and untreated, it can cause hypertension and, in extreme circumstances, renal failure. Globally, it is anticipated that one in five men and one in four women aged 65 to 74 will have chronic kidney disease (CKD).

Due to their rapid and precise recognition capabilities, machine learning models can successfully assist doctors in achieving this goal. A computer software that calculates and deduces task-related information and obtains the features of the relevant pattern is referred to as machine learning. This technique can provide reliable and cost-effective illness diagnosis, therefore it could be a potential method for detecting CKD. With the advancement of information technology, it has evolved into a new type of medical instrument with a wide range of applications due to the rapid development

of electronic health records. Machine learning has already been utilised in the medical area to identify human body state, assess disease-related aspects, and diagnose various diseases.

1.1 Problem Definition

Blood and urine testing can be used to diagnose CKD. These tests check for abnormally high levels of particular compounds in your blood and urine, which indicate that your kidneys aren't functioning properly. If you have a high chance of developing kidney disease (for example, if you have a known risk factor like high blood pressure or diabetes), you may be advised to have regular testing to detect CKD early on. Because of the growing number of chronic renal patients, the paucity of specialised physicians, and the high costs of diagnosis and treatment, particularly in developing countries, computer-assisted diagnostics are needed to assist physicians and radiologists in making diagnostic judgments. Machine learning and deep learning techniques have been used in the early phases of disease prediction and diagnosis, and artificial intelligence approaches have played a role in the health industry and medical image processing.

1.2 Objective

The project's major purpose is to::

- Develop a system to detect CKD and it's seriousness
- Deploying the model with higher accuracy
- A reduction in the kidney disease burden
- Faster detection of CKD
- Longer lives and improved quality of life for people with CKD

.

Chapter 2

LITERATURE SURVEY

A literature review is a complete examination and analysis of literature on a certain topic. When research questions are identified through a literature review, one seeks to answer them by looking for and analysing relevant material. Re-analyzing the study's results can lead to fresh discoveries, which is why literature reviews are important. A literature review summarises and explains the whole and current state of knowledge on a topic as found in academic books and journal articles. At university, there are two types of literature reviews: one that students are asked to write as a stand-alone assignment in a course, and another that is prepared as an introduction to, or preparation for, a larger work, usually a thesis or research report. The type of review you are writing will affect the focus and perspective of your review, as well as the type of hypothesis or thesis argument you make. Reading published literature reviews or the introductory chapters of theses and dissertations in your own subject area is one approach to comprehend the differences between these two forms. Examine the framework of their arguments and how they approach the topics.

2.1 Related Works

Machine Learning and data mining-based algorithms have been employed by many researchers to solve difficulties in the health sector. To classify and predict the patients' CKD status, they used a variety of approaches and methods.

To predict CKD, Charleonnan et al. [1] used four machine learning methods: K-nearest neighbors (KNN), support vector machine (SVM), logistic regression (LR), and decision tree classifiers. These models are built using a CKD dataset acquired from Apollo Hospitals Indians, and their performance is compared to find the best classifier for predicting chronic renal disease. According to the trial data, the SVM classifier has the greatest accuracy of 98.3 percent. Furthermore, after training the dataset, SVM has the highest sensitivity.

Neves et al. have diagnosed CKD using an Artificial Neural Network. Results indicated that the sensitivity value of diagnosis ranged from 93.1 to 94.9 percent, while the specificity value ranged from 91.9 to 94.2 percent [2]. Di Noia et al. have developed a software tool for classifying end-stage kidney disease using artificial neural networks. Their proposed model has achieved 91.37 percent accuracy with a sensitivity of 70.76 percent and a positive predictive accuracy of 70.76 percent [3].

Image registration was used by Hodneland et al. to detect renal morphologic alterations [4]. Vasquez-Morales et al. developed a neural network-based classifier using large-scale CKD data, and the model's accuracy on their test data was 95 percent [5]. Furthermore, the majority of previous studies used the CKD data set from the UCI machine learning library. To diagnose CKD, Chen et al. employed k-nearest neighbour (KNN), support vector machine (SVM), and soft independent modelling of class analogies. KNN and SVM had the greatest accuracy of 99.7 percent [6]. They also employed a fuzzy rule-building expert system, fuzzy optimum associative memory, and partial least squares discriminant analysis to diagnose CKD, with accuracy ranging from 95.5 to 99.6 percent [7].

Polat et al. created an SVM based on feature selection technology; the proposed models lowered computational cost by feature selection; and its accuracy ranged from 97.75 to 98.5 percent[8]. J. Aljaaf et al. utilised MLP neural network (MLP) and attained a 98.1 percent accuracy [9]. Subas et al. used MLP, SVM, KNN, C4.5 decision tree, and random forest (RF) to detect chronic kidney disease (CKD) [10], with the RF achieving a 100 percent accuracy. [11] In the models developed by Boukenze et al., MLP attained the highest degree of accuracy, 99.75 percent. Almansour et al. [12] employed SVM and neural network to diagnose CKD with respective accuracies of 97.75

percent and 99.75 percent. In the models developed by Gunarathne et al., missing values were filled in with zero, and decision forest attained the best performance with an accuracy of 99.1 percent [13].

In[14],the researchers used five different feature selection techniques to compare the accuracy ,sensitivity and specificity, those are chi-square, gain ratio, information gain, symmetrical uncertainty, relief-f on the two classification algorithm that is multilayer preceptor network(MLPN) and radial-basis function networks(RBFN) with an accuracy of 97 percent and 98.5percent respectively.

Chapter 3

METHODOLOGY

3.1 Proposed system

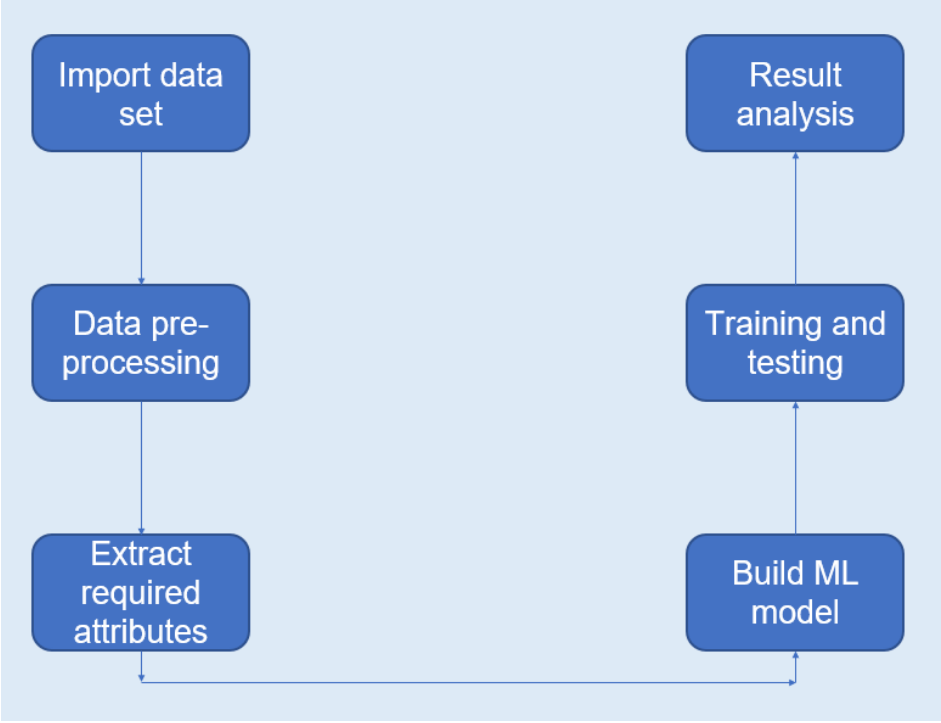

Figure 3.1: Block diagram of the proposed system

By analyzing the misjudgments generated by the previously established models,here proposing an integrated model that combines logistic regression and random forest.The dataset using this sys-

tem was obtained from the University of California Irvine (UCI) machine learning repository.It is a machine learning model,it is trained using this CKD dataset.It will predict chronic kidney disease of person based on the clinical data of the individual person which can also be entered manually.By checking the parameters such as age,hypertension,diabetes melitus,appetite and anemia it will also predict the seriousness of the disease.

The block diagram of the proposed system is shown in Figure 3.1. All of this diagram's components are discussed in the subsections that follow.

3.1.1 Dataset

This study obtained the CKD data set from the UCI machine learning repository. The data set contains 400 samples. Each sample in this CKD data set contains 24 predictive variables or features (11 numerical variables and 13 categorical (nominal) variables) and one categorical response variable (class). Each class has two values: ckd (sample with CKD) and notckd (sample without CKD) (sample without CKD). 250 samples out of 400 belong to the ckd category, while 150 samples belong to the notckd category.

Variables	Explain	Class	Scale	Missing Rate
age	Age	Numerical	age in years	2.25%
bp	Blood Pressure	Numerical	in mm/Hg	3%
sg	Specific Gravity	Nominal	(1.005,1.010,1.015,1.020,1.025)	11.75%
al	Albumin	Nominal	(0,1,2,3,4,5)	11.5%
su	Sugar	Nominal	(0,1,2,3,4,5)	12.25%
rbc	Red Blood Cells	Nominal	(normal,abnormal)	38%
pc	Pus Cell	Nominal	(normal,abnormal)	16.25%
pcc	Pus Cell clumps	Nominal	(present,notpresent)	1%
ba	Bacteria	Nominal	(present,notpresent)	1%
bgr	Blood Glucose Random	Numerical	in mgs/dl	11%
bu	Blood Urea	Numerical	in mgs/dl	4.75%
sc	Serum Creatinine	Numerical	in mgs/dl	4.25%
sod	Sodium	Numerical	in mEq/L	21.75%
pot	Potassium	Numerical	in mEq/L	22%
hemo	Hemoglobin	Numerical	in gms	13%
pcv	Packed Cell Volume	Numerical	-	17.75%
wbcc	White Blood Cell Count	Numerical	in cells/cumm	26.5%
rbcc	Red Blood Cell Count	Numerical	in millions/cmm	32.75%
htn	Hypertension	Nominal	(yes,no)	0.5%
dm	Diabetes Mellitus	Nominal	(yes,no)	0.5%
cad	Coronary Artery Disease	Nominal	(yes,no)	0.5%
appet	appet	Nominal	(good,poor)	0.25%
pe	Pedal Edema	Nominal	(yes,no)	0.25%
ane	Anemia	Nominal	(yes,no)	0.25%
class	Class	Nominal	(ckd,notckd)	0%

Figure 3.2: Dataset

3.1.2 Data Preprocessing

Data Pre-Processing is the phase in which distorted or encoded data is transformed into a form that can be easily analysed by a machine. A dataset is observable as a collection of data objects. A number of features, such as the mass of a physical object or the time at which an event occurred, are assigned to data objects as labels. There may be missing values in the dataset; they can be eliminated or estimated. The most prevalent method for dealing with missing values is to replace them with the mean, median, or mode of the respective feature. As object values cannot be used for analysis, the numeric values of type object must be converted to float64. In the categorical attributes, the null values are replaced with the attribute column's most frequent value. The purpose of label encoding is to convert categorical attributes into numeric attributes by converting each unique attribute value into an integer. This transforms the attributes to int type automatically. Using the "dropna" and "fillna" functions from the Pandas library, which drop columns or rows with missing data and replace NAs with the attribute's mean value, respectively, we remove missing values from the dataset, such as NAs and blanks. We will use two methods to fill null values: random sampling for larger null values and mode sampling for smaller null values.

3.1.3 Extract required attributes

Here, select computationally the features that contribute the most to our prediction variable or output.In this study,I use lasso regression for feature extraction.Similar to linear regression, lasso regression employs a technique called "shrinkage" in which the coefficients of determination are shrunk toward zero. Linear regression provides the observed regression coefficients from the dataset. The lasso regression enables you to reduce or regularise these coefficients to prevent over-fitting and improve their performance across various datasets. This type of regression is utilised when the dataset exhibits high multicollinearity or when variable elimination and feature selection are to be automated.

Lasso is a linear model that uses this cost function:

$$\frac{1}{2N_{training}} \sum_{i=1}^{N_{training}} \left(y_{real}^{(i)} - y_{pred}^{(i)}\right)^2 + \alpha \sum_{j=1}^{n} |a_j|$$

aj is the coefficient of the j-th feature. The final term is called l1 penalty and is a hyperparameter that tunes the intensity of this penalty term. The higher the coefficient of a feature, the higher the value of the cost function. So, the idea of Lasso regression is to optimize the cost function reducing the absolute values of the coefficients

3.1.4 Build machine learning model

Six machine learning models,logistic regression,random forest,support vector machine,k nearest neighbor,naive bayes classifier and multilayer perceptron are used.These models are imported from skleran library.And an integrated model combining logstic regression and random forest is build.Imported VotingClassifier from sklearn.ensemble to combine these two algorithms.

3.1.5 Training and testing

70 percent of the data is taken for training and 30 percent of the data is taken for testing.These machine learning models are trained and tested.And the performance of these machine learning models are found out.

3.1.6 Result analysis

The accuracy of these models are evaluated.The integrated model is taken for the implementation of the system.The parameteres age,hypertension,diabetes melitus,appetite and anemia are taken to the analysis of severity of the disease.Based on the presence of these conditions and the age of the individual,the severity of the disease is predicted.If a person is detected ckd,it shows whether he belongs to mild,moderate or severe case of the disease.

3.2 Graphical user interface

A user interface is created.The front end of the user interface is created using html,css and javascript.Flask is the web framework which is used here and sqlite is the database which is used here.The user and register and login to it.In the system he/she can enter the values of the attributes and the system

will predict if he/she is having ckd or not.In the case of ckd, it shows the seriousness of the disease.

3.2.1 Save machine learning model using Pickle

To save the model, we simply need to pass the model object to Pickle's dump() function. This will serialise the object and convert it into a "byte stream" that can be stored in the model.pkl file.

3.3 Tools used

The tools used for the project:

• Python

• Jupyter Notebook

• Flask

• HTML

• CSS

• JavaScript

• SQlite

3.3.1 Python

Guido Rossum created Python, an object-oriented programming language, in 1989. It is optimised for rapid prototyping of complex applications.Python is widely used in artificial intelligence, natural language generation, neural networks, and other advanced computer science fields. Python is a potent programming language that can be used to create games, GUIs, and web applications. It is an advanced language. Python is an object-oriented programming language that allows users to create and execute programmes by managing data structures or objects. Everything in Python is in fact of the highest quality. Python treats all objects, data types, functions, methods, and classes equally. Programming languages are developed to meet the needs of programmers and users for an effective tool to create applications that have an impact on lives, lifestyles, the economy, and so-

ciety. They contribute to the improvement of life by enhancing productivity, communication, and potency. Languages become extinct and obsolete when they fail to live up to expectations and are supplanted by more potent languages. Python is an artificial programming language that has withstood the test of time and remained relevant across industries, businesses, and among individual programmers. It is a living, thriving, and incredibly useful language that is highly recommended as a primary programming language for those who wish to begin programming.

3.3.2 Jupyter Notebook

The Jupyter Notebook is an open-source web application for creating and sharing documents with live code, equations, visualisations, and text. Jupyter Notebook is administered by the Project Jupyter team. The Jupyter Notebooks project is a spin-off of the IPython project, which previously had its own IPython Notebook project. Jupyter derives its name from the three primary programming languages it supports: Julia, Python, and R. Jupyter ships with the IPython kernel, which enables you to write your programmes in Python, but you can also use more than 100 other kernels. Jupyter notebooks are intended to provide a more user-friendly interface for code used in digitally-supported research or education.. Jupyter, an open-source environment compatible with a variety of programming languages, has gained traction in a variety of fields.It is useful for documenting code, teaching programming languages, and providing students with a space to easily experiment with provided examples. Jupyter Notebooks can be executed in two major environments: Jupyter Notebook and the more recent Jupyter Lab. Jupyter Notebook is widely used and well-documented; it provides a simple file browser and an environment for creating, editing, and executing notebooks. Jupyter Lab is more complex and resembles an Integrated Development Environment in its user interface.

The Jupyter Notebook is not only useful for teaching and learning programming languages like Python, but also for sharing data.

Google and Microsoft each offer their own version of the Notebook, which can be used to create and share documents at Google Colaboratory and Microsoft Azure Notebooks, respectively. JupyterLab incorporates Jupyter Notebook into a browser-based Integrated Development type Editor. JupyterLab can be viewed as an advanced version of Jupyter Notebook. In addition to Note-

books, JupyterLab allows you to run terminals, text editors, and code consoles in your browser.

3.3.3 Flask

Flask is a Python-based web application framework. It is created by Armin Ronacher, who leads an international group of Python enthusiasts called Pocco. . Flask's framework is more explicit than Django's and easier to learn, as it requires less base code to implement a simple web application. Flask relies on the WSGI (Web Server Gateway Interface) toolkit and the Jinja2 template engine. Flask was made to be simple to use and expand. Flask's purpose is to provide a solid foundation for web applications of varying complexity. Flask is useful for a variety of applications. It is particularly useful for prototyping. Flask relies on the following external libraries:the Jinja2 template engine and the Werkzeug WSGI framework.

Flask has a lightweight and modular design, so it is simple to transform it into the required web framework by adding a few modules. It is incredibly simple to deploy Flask in production and it features HTTP request handling and high flexibility. The configuration is even more flexible than Django's, providing a multitude of solutions for all production needs.

3.3.4 HTML

HTML is the standard markup language for documents intended for display in a web browser. It can be aided by Cascading Style Sheets (CSS) and scripting languages such as JavaScript.

Web browsers receive HTML files from a web server or local storage and convert them into multimedia web pages. HTML describes the semantic structure of a web page and originally included hints for the document's appearance.

HTML elements are the fundamental constituents of HTML pages. Images and other objects, such as interactive forms, can be embedded in the rendered page using HTML constructs. HTML enables the creation of structured documents by assigning structural semantics to text elements such as headings, paragraphs, lists, links, and other elements.

3.3.5 CSS

CSS is a style sheet language that is used to describe the presentation of a document written in a markup language such as HTML. In addition to HTML and JavaScript, CSS is a fundamental technology for the World Wide Web.

CSS is designed to separate presentation from content, including layout, colours, and fonts. This separation can improve content accessibility; provide greater flexibility and control in the specification of presentation characteristics; enable multiple web pages to share formatting by specifying the relevant CSS in a separate.css file, which reduces complexity and repetition in the structural content; and enable the.css file to be cached to improve page load speed for pages that share the file and its formatting.

3.3.6 JavaScript

JavaScript, also known as JS, is a programming language that, along with HTML and CSS, is one of the core technologies of the World Wide Web. Over 97 percent of websites use JavaScript for client-side web page behaviour, with third-party libraries frequently incorporated. All major web browsers include a dedicated JavaScript engine for executing code on user devices.

JavaScript is a high-level, often just-in-time, ECMAScript-compliant, compiled programming language.

It features dynamic typing, prototype-based object orientation, and functions with first-class status. Event-driven, functional, and imperative programming styles are supported. APIs for working with text, dates, regular expressions, standard data structures, and the Document Object Model are available (DOM). Initially, JavaScript engines were only used in web browsers, but they are now integral components of many servers and applications.

3.3.7 SQlite

SQLite is a database engine implemented in the C programming language. It is not a standalone app, but rather a library that developers incorporate into their apps. As such, it belongs to the embedded database family. It is the most widely deployed database engine, as several of the most popular web browsers, operating systems, mobile phones, and embedded systems utilise it.

SQLite's design objectives were to enable the programme to be run without installing a database management system or requiring a database administrator.

3.4 Algorithms

3.4.1 Logistic regression

Logistic regression is a supervised learning algorithm for predicting the probability of a target variable. There are only two possible classes based on the dichotomous nature of the target or dependent variable. In other words, the dependent variable is binary, with data coded as 1 (representing success/yes) or 0 (representing failure/no).

As a function of X, a logistic regression model predicts P(Y=1) mathematically. It is one of the simplest machine learning algorithms that can be applied to a variety of classification problems, including spam detection, diabetes prediction, cancer detection, etc.

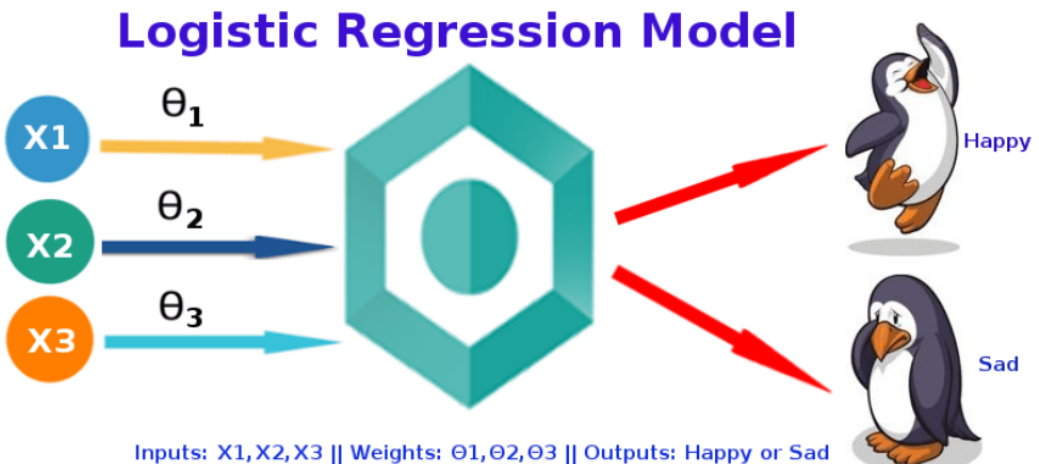

Figure 3.3: Logistic regression model

3.4.2 Random forest

Popular machine learning algorithm that belongs to the supervised learning technique is Random Forest. It is applicable to both Classification and Regression problems in Machine Learning. It is based on ensemble learning, which is the process of combining multiple classifiers to solve a complex problem and improve the model's performance.

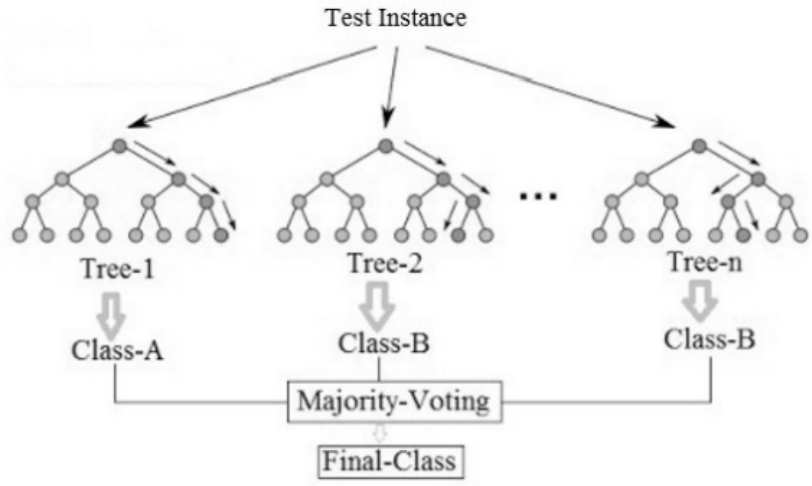

Figure 3.4: Example of random forest with majority voting

3.4.3 Support vector machine

Support Vector Machine, or SVM, is one of the most popular algorithms for Classification and Regression problems in Supervised Learning. In Machine Learning, it is used primarily for Classification problems.

The objective of the SVM algorithm is to generate the optimal line or decision boundary that divides n-dimensional space into classes, so that future data points can be easily classified. This optimal decision boundary is referred to as a hyperplane.

SVM selects the extreme points/vectors that contribute to the formation of the hyperplane. These extreme cases are referred to as support vectors, and the corresponding algorithm is known as the

Support Vector Machine.

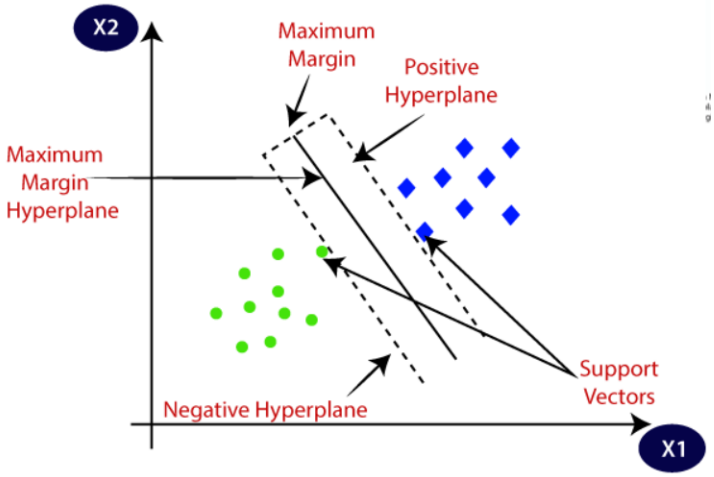

Figure 3.5: support vector machine

3.4.4 K-nearest neighbor

K-Nearest Neighbor is one of the simplest Machine Learning algorithms that utilises the Supervised Learning technique.It assumes the similarity between the new case/data and existing cases and places the new case in the category that is the most similar to the existing categories.The K-NN algorithm stores all available data and classifies a new data point on the basis of similarity. This implies that when new data becomes available, the K- NN algorithm can easily classify it into a suitable category.It can be used for both Regression and Classification, although Classification is its primary application.

K-NN is a non-parametric algorithm, meaning it makes no assumptions about the data it is analysing.It is also referred to as a lazy learner algorithm because it does not immediately learn from the training set. Rather, it stores the dataset and, at the time of classification, performs an action on the dataset. During the training phase, the system simply stores the dataset, and when it receives new data, it classifies it into a category that is highly similar to the original category.

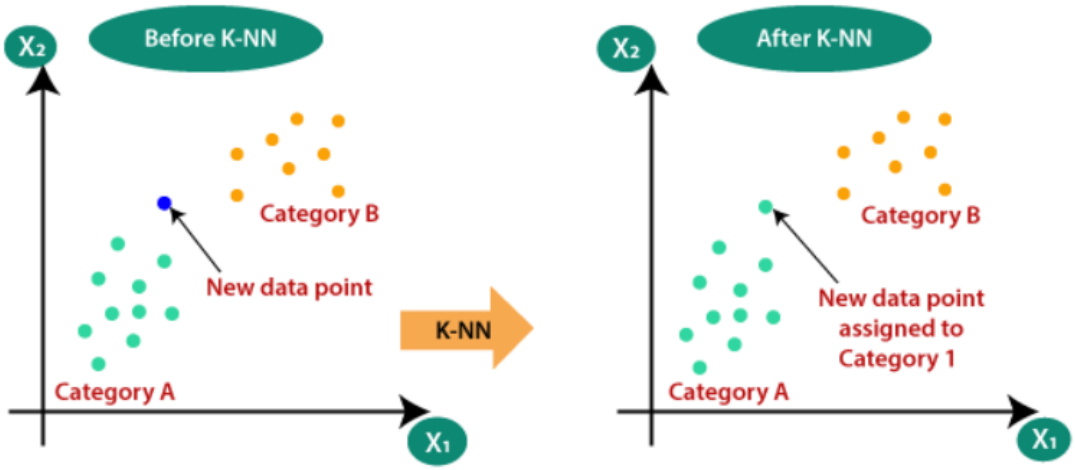

Figure 3.6: K-nearest neighbor

3.4.5 Naive Bayes classifier

Naive Bayes classifiers are a collection of Bayes' Theorem-based classification algorithms. It is not a single algorithm, but rather a family of algorithms that share a common principle, namely that each pair of features being classified is independent from the other.

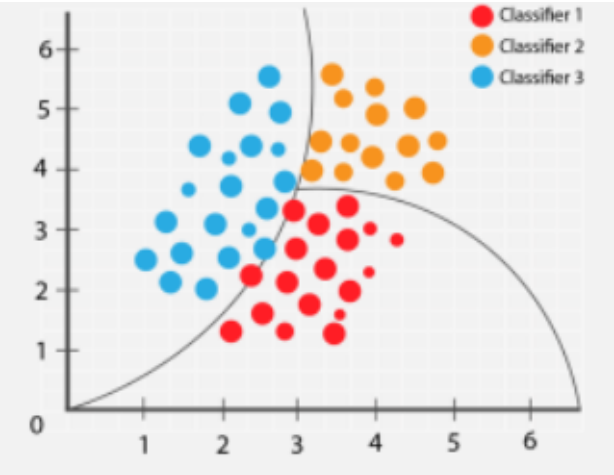

Figure 3.7: Naive Bayes classifier

3.4.6 Multilayer perceptron

A multilayer perceptron (MLP) is a type of feedforward artificial neural network that is fully connected (ANN). Sometimes the term MLP is used loosely to refer to any feedforward ANN, and other times it is used strictly to refer to networks composed of multiple layers of perceptrons. Multilayer perceptrons are sometimes referred to colloquially as ”vanilla” neural networks, particularly when they contain a single hidden layer.

At least three layers of nodes comprise an MLP: an input layer, a hidden layer, and an output layer. Each node, excluding input nodes, is a neuron with a nonlinear activation function. MLP employs backpropagation, a supervised learning technique, for training. MLP differs from a linear perceptron due to its multiple layers and non-linear activation. It can distinguish data that cannot be separated linearly.

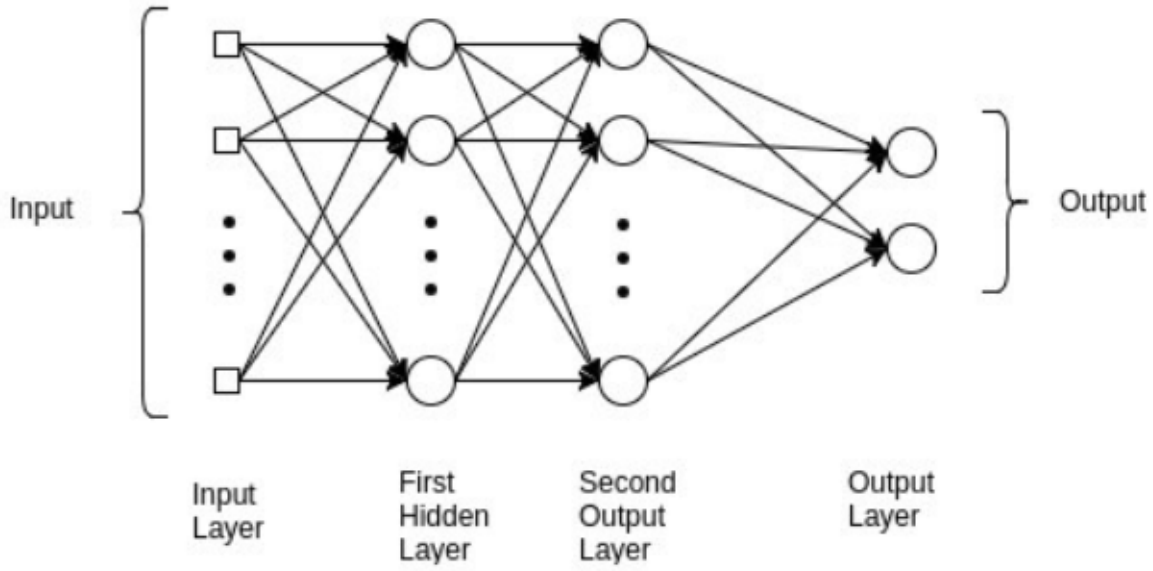

Figure 3.8: Multilayer perceptron

3.4.7 Integrated Model

Logistic regression and random forest are combined to create the intergrated model.Thsese models are combined using voting classifier.A Voting Classifier is a machine learning model that trains on an ensemble of numerous models and predicts an output (class) based on their highest probability of chosen class as the output. It simply aggregates the findings of each classifier passed into Voting Classifier and predicts the output class based on the highest majority of voting. It create a single model which trains by these models and predicts output based on their combined majority of voting for each output class.

Chapter 4

RESULTS AND DISCUSSIONS

4.1 Data preprocessing results

The dataset is successfully preprocessed.

	id	age	bp	sg	al	su	rbc	pc	pcc	ba	bgr	bu	sc	sod	pot	hemo	pcv
0	0	48.0	80.0	1.020	1.0	0.0	NaN	normal	notpresent	notpresent	121.0	36.0	1.2	NaN	NaN	15.4	44
1	1	7.0	50.0	1.020	4.0	0.0	NaN	normal	notpresent	notpresent	NaN	18.0	0.8	NaN	NaN	11.3	38
2	2	62.0	80.0	1.010	2.0	3.0	normal	normal	notpresent	notpresent	423.0	53.0	1.8	NaN	NaN	9.6	31
3	3	48.0	70.0	1.005	4.0	0.0	normal	abnormal	present	notpresent	117.0	56.0	3.8	111.0	2.5	11.2	32
4	4	51.0	80.0	1.010	2.0	0.0	normal	normal	notpresent	notpresent	106.0	26.0	1.4	NaN	NaN	11.6	35

Figure 4.1: Dataset before preprocessing

	age	blood_pressure	specific_gravity	albumin	sugar	red_blood_cells	pus_cell	pus_cell_clumps	bacteria
0	48.0	80.0	1.020	1.0	0.000000	1	1	0	0
1	7.0	50.0	1.020	4.0	0.000000	1	1	0	0
2	62.0	80.0	1.010	2.0	1.386294	1	1	0	0
3	48.0	70.0	1.005	4.0	0.000000	1	0	1	0
4	51.0	80.0	1.010	2.0	0.000000	1	1	0	0

Figure 4.2: Dataset after preprocessing

4.2 Accuracy shown by different ml models

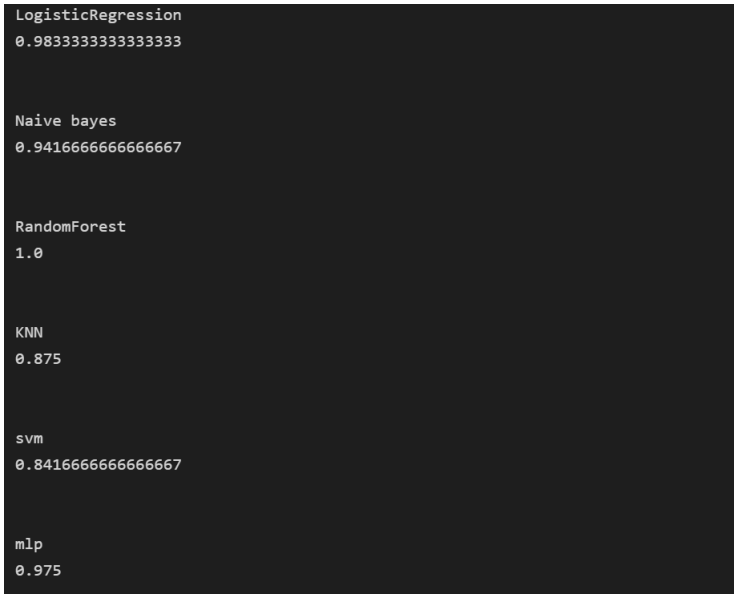

Figure 4.3: Accuracy of different ml models

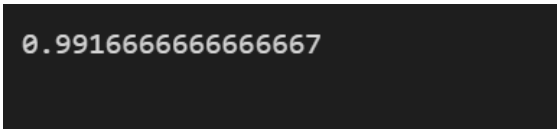

Figure 4.4: Accuracy of integrated models

4.3 Performance metrics

A model's performance is evaluated in order to determine its superiority. Utilized evaluation metrics include Accuracy, Precision, Recall, F1 score, and the Confusion Matrix.

The performance indicators includes:

• True positive(TP) : A true positive is an outcome where the model correctly predicts the positive class.

• True negative(TN) : A true negative is an outcome where the model correctly predicts the negative class.

• False positive(FP): A false positive is an outcome where the model incorrectly predicts the positive class.

• False negative(FN): A false negative is an outcome where the model incorrectly predicts the negative class.

Derived performance metrics includes: • Accuracy : Accuracy represents the number of correctly classified data instances over the total number of data instances.

$Accuracy = (TP+TN)/(TP+TN+FP+FN)$

• Precision : Precision refers to the number of true positives divided by the total number of positive predictions.

$Precision = TP/(TP+FP)$

• Recall : Recall is the fraction of the total amount of relevant instances that were actually the retrieved instance of the classification algorithm. $Recall = TP/(TP+TN)$

• F1 Score : The accuracy of the model in whole dataset. It is the harmonic mean between precision and recall. $F1score = precision * Recall/(precision + Recall)$

• Confusion matrix : A Confusion matrix is a N x N matrix used to evaluate a classification model's performance, where N is the number of target classes. The matrix compares the actual target values against the values predicted by the machine learning model. This provides a comprehensive view of our classification model's performance and the types of errors it makes.

The outcomes of the most common performance matrix such as classification accuracy,precision,recall and f1 score shows that the proposed integrated model gives a better performance.

	precision	recall	f1-score	support
0	0.98	1.00	0.99	44
1	1.00	0.99	0.99	76
accuracy			0.99	120
macro avg	0.99	0.99	0.99	120
weighted avg	0.99	0.99	0.99	120

Figure 4.5: Performance metrics of integrated model

```
Confusion Matrix:
[[44  0]
 [ 1 75]]
```

Figure 4.6: Confusion matrix

4.3.1 ROC Curve

For binary classification problems, the Receiver Operator Characteristic (ROC) curve is an evaluation metric. It is a probability curve that compares the TPR to the FPR at different threshold values. As a summary of the ROC curve, the Area Under the Curve (AUC) quantifies the ability of a classifier to distinguish between classes.In fig we can see that the area under the curve is near to one,which means the integrated model is giving us a good classification.

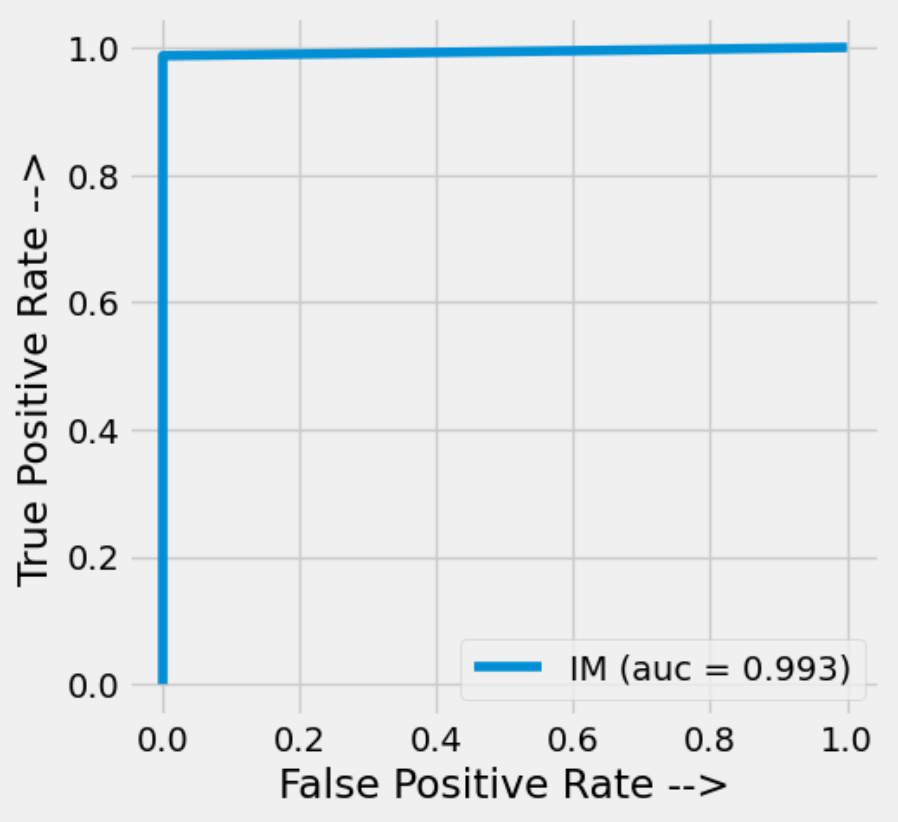

Figure 4.7: ROC Curve

4.4 Disease detection and severity analysis

We can manually enter the values of different features and analysing these values the system predict that if the person is having chronic kidney disease or not.And analysing the factors such as age,hypertension,diabetes melitus,appetite and anemia the system will also predict the seriousness of the disease if the person is having chronic kidney disease.

```
new_val=[48,80,1,121,3.610918,0.78845,2.151762,15.4,44,5.2,1,1,0,0
]
```
✓ 0.1s

Figure 4.8: Entering attribute values

```
disease detected
severe
```

Figure 4.9: Disease detection and severity prediction

4.5 Graphical user interface

A user interface is created.The front end of the user interface is created using html,css and javascript.Flask is the web framework which is used here and sqlite is the database which is used here.The user and register and login to it.SQlite is the database which is used here.In the system he/she can put the values of the attributes and the system will predict if he/she is having ckd or not.In the case of ckd,it shows the seriousness of the disease.

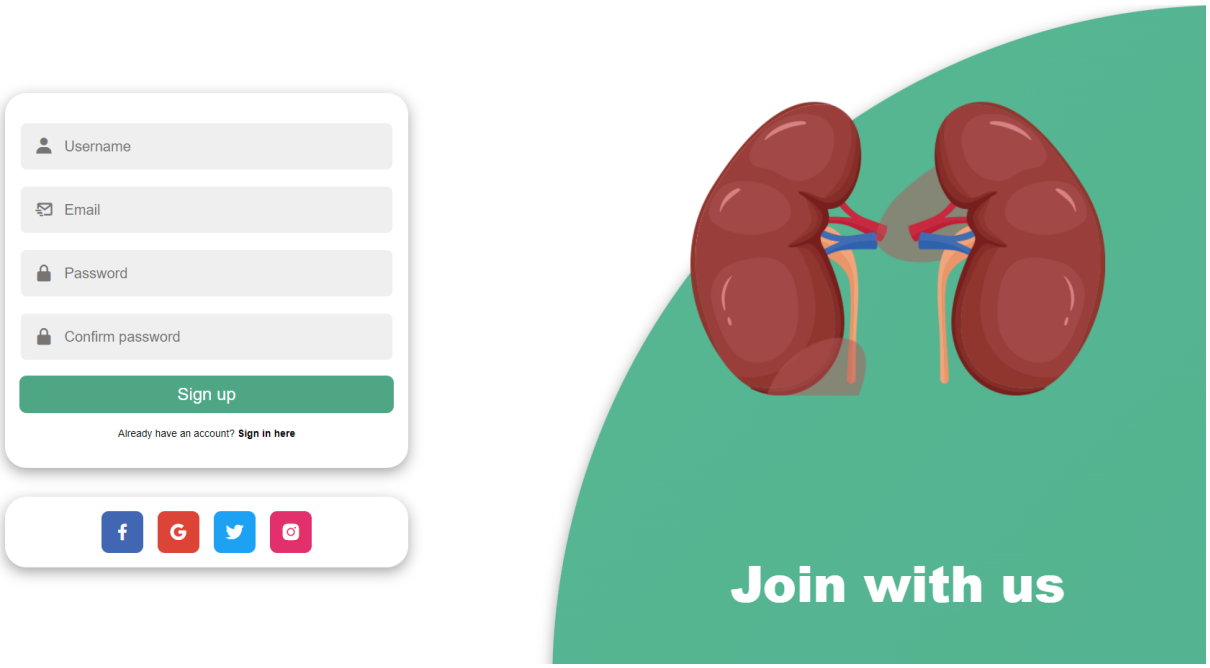

Figure 4.10: User Registration

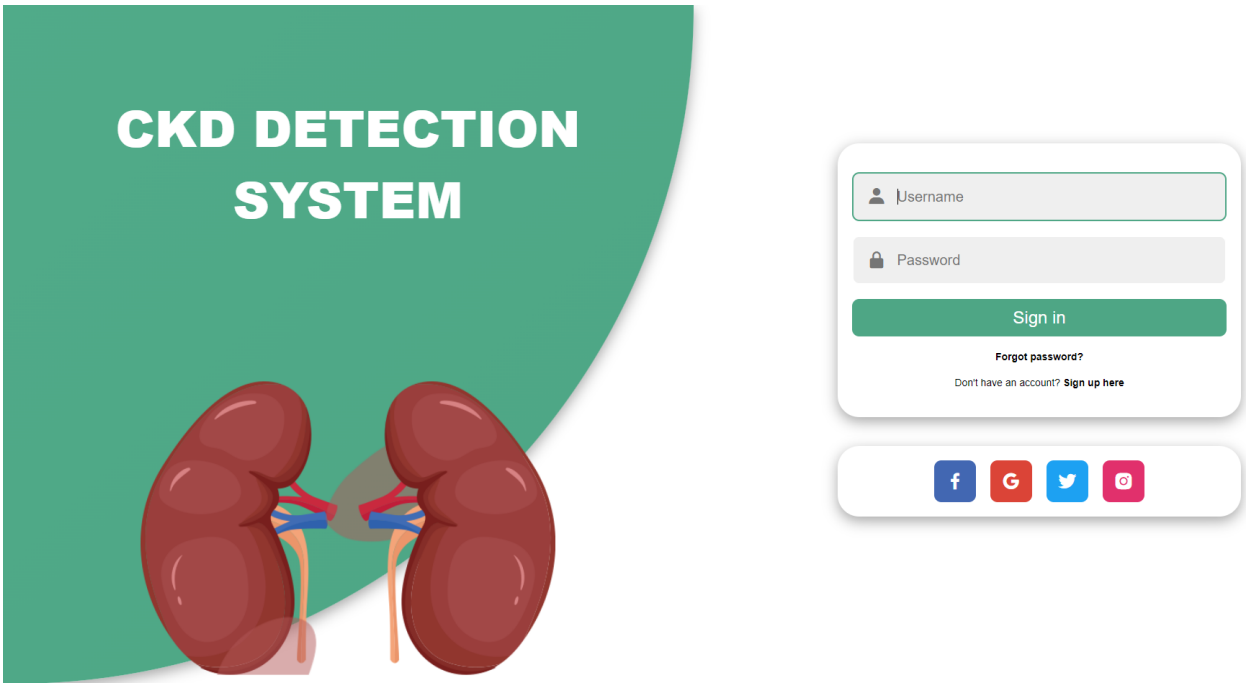

Figure 4.11: User login

Figure 4.12: Input values

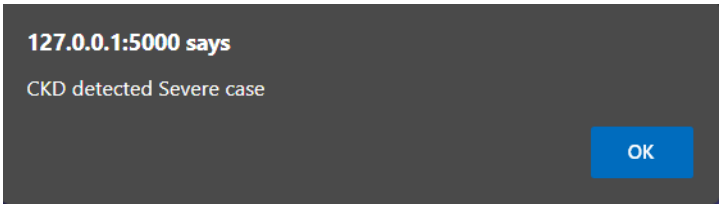

Figure 4.13: Output

Chapter 5

CONCLUSION

The identification of the CKD is a difficult task.The diagnosis process by using lab tests, imaging studies and biopsy might be a time-consuming, invasive, and costly.The proposed CKD diagnostic method is practicable in terms of disease diagnosis and severity.The integrated model could achieve an accuracy of 99 percent.The severity of the disease is diagnosed using the parameters age,hypertension,diabetes melitus,appetite and anemia.This is extraordinarily beneficial in the medical field for the early and accurate diagnosis of chronic kidney disease in patients. The importance of early diagnosis in saving a patient's life lies in the fact that it ensures effective and timely treatment.

5.1 Advantages

The key advantages of the proposed model are:

- No expert knowledge is required to identify the disease.
- Early disease detection.
- Detection of the disease's severity.
- The system provide more accuracy and efficiency for CKD detection.

Chapter 6

Future Scope

This methodology may be applicable to the clinical data of other diseases for the purposes of actual medical diagnosis.I believe that as the quantity and quality of data improves, this model will be more and more perfect.

REFERENCES

[1] Anusorn Charleonnan, Thipwan Fufaung, Tippawan Niyomwong, Wandee Chokchueypattanakit, Sathit Suwannawach, Nitat Ninchawee, "Predictive Analytics for Chronic Kidney Disease Using Machine Learning Techniques," Proc. Management and Innovation Technology International Conference (MITiCON-2016) , IEEE, Oct. 2016, doi:10.1109/MITICON.2016.8025242.

[2] José, N., Rosário Martins, M., Vilhena, J., Neves, J., Gomes, S., Abelha, A., Machado, J., and Vicente, H., A soft computing approach to kidney diseases evaluation. J. Med. Syst. 39:131, 2015. doi:10.1007/s10916-015-0313-4.

[3] Di Noia, T., Claudio, V., Ostuni, F.P., Binetti, G., Naso, D., Schena, F.P., and Di Sciascio, E., An end stage kidney disease predictor based on an artificial neural networks ensemble. Expert Syst. Appl. 40:4438–4445, 2013. doi:10.1016/j.eswa.2013.01.046.

[4] neland, E. Keilegavlen, E. A. Hanson, E. Andersen, J. A. Monssen, J. Rorvik, S. Leh, H.-P. Marti, A. Lundervold, E. Svarstad, and J. M. Nordbotten, "In Vivo detection of chronic kidney disease using tissue deformation fields from dynamic MR imaging," IEEE Trans. Biomed. Eng., vol. 66, no. 6, pp. 1779–1790, Jun. 2019.

[5] G. R. Vasquez-Morales, S. M. Martinez-Monterrubio, P. Moreno-Ger, and J. A. Recio-Garcia, "Explainable prediction of chronic renal disease in the colombian population using neural networks and case-based reasoning," IEEE Access, vol. 7, pp. 152900–152910, 2019.

[6] Z. Chen, X. Zhang, and Z. Zhang, "Clinical risk assessment of patients with chronic kidney disease by using clinical data and multivariate models," Int. Urol. Nephrol., vol. 48, no. 12, pp. 2069–2075, Dec. 2016.

[7] Z. Chen, Z. Zhang, R. Zhu, Y. Xiang, and P. B. Harrington, "Diagnosis of patients with chronic kidney disease by using two fuzzy classifiers," Chemometrics Intell. Lab. Syst., vol. 153, pp. 140–145, Apr. 2016.

[8] H. Polat, H. D. Mehr, and A. Cetin, "Diagnosis of chronic kidney disease based on support vector machine by feature selection methods," J. Med. Syst., vol. 41, no. 4, p. 55, Apr. 2017.

[9] A. J. Aljaaf, D. Al-Jumeily, H. M. Haglan, M. Alloghani, T. Baker, A. J. Hussain, and J. Mustafina, "Early prediction of chronic kidney disease using machine learning supported by predictive analytics," in Proc. IEEE Congr. Evol. Comput. (CEC), Jul. 2018, pp. 1–9.

[10] A. Subasi, E. Alickovic, and J. Kevric, "Diagnosis of chronic kidney disease by using random forest," in Proc. Int. Conf. Med. Biol. Eng., Mar. 2017, pp. 589–594.

[11] B. Boukenze, A. Haqiq, and H. Mousannif, "Predicting chronic kidney failure disease using data mining techniques," in Proc. Int. Symp. Ubiquitous Netw., Nov. 2016, pp. 701–712.

[12] N. A. Almansour, H. F. Syed, N. R. Khayat, R. K. Altheeb, R. E. Juri, J. Alhiyafi, S. Alrashed, and S. O. Olatunji, "Neural network and support vector machine for the prediction of chronic kidney disease: A comparative study," Comput. Biol. Med., vol. 109, pp. 101–111, Jun. 2019.

[13] W. Gunarathne, K. Perera, and K. Kahandawaarachchi, "Performance evaluation on machine learning classification techniques for disease classification and forecasting through data analytics for chronic kidney disease (CKD)," in Proc. IEEE 17th Int. Conf. Bioinf. Bioeng. (BIBE), Oct. 2017, pp. 291–296.

[14] A. K. Shrivas and S. Kumar Sahu, "Classification of Chronic Kidney Disease using Feature Selection Techniques," IJCSE, vol. 6, no. 5, pp. 649–653, 2018.

APPENDIX

Screenshots

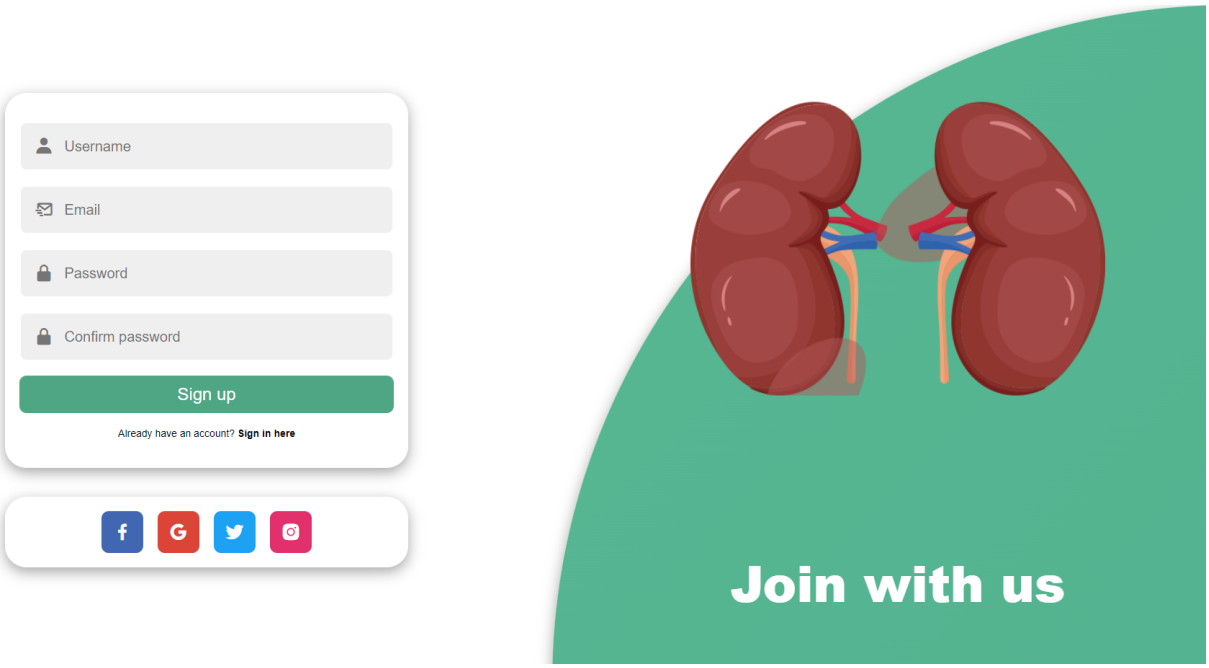

Figure A.1 : User Registration

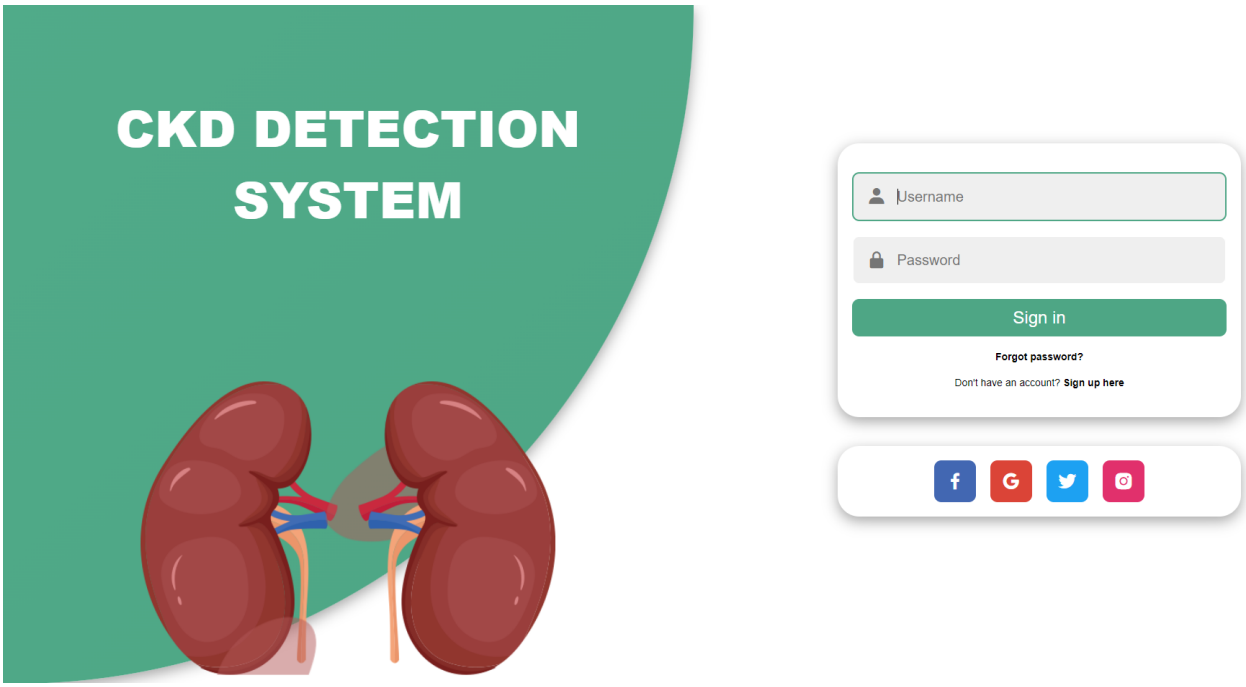

Figure A.2 : User login

Figure A.3 : Input values

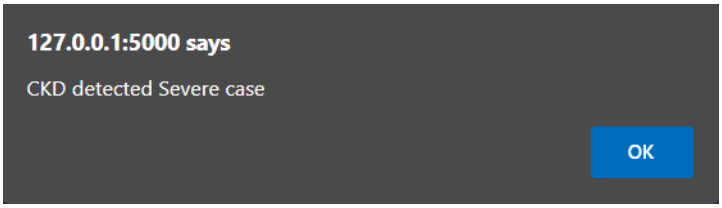

Figure A.4 : Output