A Project Report on

Chronic Kidney Disease Prediction Using AdaBoost Classifier

submitted in partial fulfillment for the award of

Bachelor of Technology

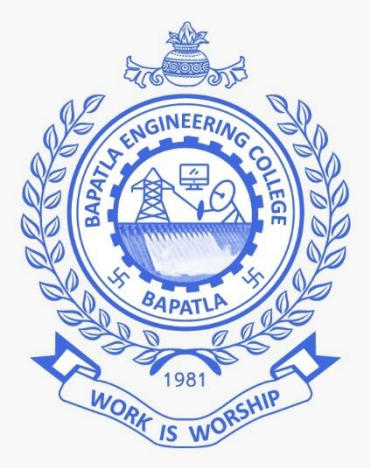
in

Computer Science and Engineering

By

P.Niharika (Y20ACS538) **V.Joshna Florence (Y20ACS583)**

B.Gowtham (Y19ACS416) **S.Mohan (Y20ACS572)**



Under the guidance of

Mr. M M Meera Durga,M.Tech

Assistant Professor

Department of Computer Science and Engineering

Bapatla Engineering College

(Autonomous)

(Affiliated to Acharya Nagarjuna University)

BAPATLA – 522 102, Andhra Pradesh, INDIA

2023-2024

Department of
Computer Science and Engineering

CERTIFICATE

This is to certify that the project report entitled **Chronic Kidney Disease Prediction Using AdaBoost Classifier** that is being submitted by P.Niharika (Y20ACS538), V.Joshna Florence (Y20ACS583), B.Gowtham (Y19ACS416) and S.Mohan (Y20ACS572) in partial fulfillment for the award of the Degree of Bachelor of Technology in Computer Science & Engineering to the Acharya Nagarjuna University is a record of bonafide work carried out by them under our guidance and supervision.

Date:

Signature of the Guide
Mr.M M Meera Durga
Assistant Professor

Signature of the HOD
Dr. M. Rajesh Babu
Associate Professor

DECLARATION

We declare that this project work is composed by ourselves, that the work contained herein is our own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

P.Niharika (Y20ACS538)
V.Joshna Florence (Y20ACS583)
B.Gowtham (Y19ACS416)
S. Mohan (Y20ACS572)

Acknowledgement

We sincerely thank the following distinguished personalities who have given their advice and support for successful completion of the work.

We are deeply indebted to our most respected guide **Mr.M M Meera Durga,** Assistant Professor, Department of CSE, for his valuable and inspiring guidance, comments, suggestions and encouragement.

We extend our sincere thanks to **Dr.M Rajesh Babu**,Associate Professor & Head of the Dept. for extending his cooperation and providing the required resources.

We would like to thank our beloved Principal **Dr.Nazeer Shaik** for providing the online resources and other facilities to carry out this work.

We would like to express our sincere thanks to our project coordinator **Dr. N. Sudhakar,** Prof. Dept. of CSE for his helpful suggestions in presenting this document.

We extend our sincere thanks to all other teaching faculty and non-teaching staff of the department, who helped directly or indirectly for their cooperation and encouragement.

P.Niharika (Y20ACS538)
V.Joshna Florence (Y20ACS583)
B.Gowtham (Y19ACS416)
S.Mohan (Y20ACS572)

Table of Contents

List of Figures

Abstract

Since last three decades, Kidney has been emerged as major chronic diseases, impacting the health of all age human beings and may cause several other diseases if it is untreated and unidentified. No proper treatments are given for chronic kidney disease which is also a major disease in the world. Early prediction and proper medication for betterment is the right way to help the people and physicians.

To cope with such problems and address the challenges of healthcare, a wide range of tools, techniques and frameworks have been offered by Machine Learning as it has the capability of determining and recognizing patterns in complex datasets. By the use of Machine learning techniques data can be analyzed easily to provide better treatments for patients. Analysis of data and prediction of disease can be early done through machine learning. Machine learning plays a vital role in the healthcare industry and helps in prediction of disease from the datasets by using machine learning techniques.

In this project, an efficient Ensemble classifier based advanced Machine Learning approach that is AdaBoost Classifier has been proposed for the diagnosis of associated risk factors, cofactors promoting its progression, complications in prevention and control of CKD. In expansion, procuring data concerning specialists of that specific infection as per the prerequisite facilitates legitimate and proficient determination of disease.

1. Introduction

A persistent decline in kidney function defines chronic kidney disease (CKD), a condition that becomes worse over time. Millions of individuals are impacted by this important worldwide health problem, which is associated with greater rates of sickness, mortality, and medical costs. Understanding the origins, risk factors, symptoms, diagnosis, and treatment of CKD is crucial for early detection, effective treatment, and improved patient outcomes[1].

CKD is a long-term disorder caused by renal failure in both kidneys. Renal damage refers to any kind of kidney disease that has the potential to diminish the capacity of kidney functions, namely the glomerular filtration rate (GFR). Millions of small blood capillaries in the kidneys act as filters to cleanse the blood of waste products. In certain cases, this filtering mechanism fails, and the kidneys lose their capacity to filter out waste products, resulting in renal disease[2].

Machine learning is field concerned with the study of large and numerous variable information. In Health Care discerning, Machine learning guarantees to help doctors to form perfect determination, suggests the leading medicines for the patient's, spot patients at high-risk for pitiable results and particularly progressing patient's physical condition whereas minimizing costs. Machine learning has demonstrated a victory in forecast and conclusion of different basic illness.

In the field of medicine, machine learning has become a potent tool, and its use in predicting chronic kidney disease (CKD) has shown great promise. Machine learning approaches may help with precise prediction, early identification of CKD .

1.1 Problem Statement

The number of patients with kidney disease continues to increase due to the consumption of junk food and lack of water in our body and many other reasons. Diagnosis of kidney infections can be very costly and dangerous if kidney tests are done frequently. For these reasons, many patients are neglecting treatment. Kidney disease is a major chronic disease associated with high blood pressure, diabetes and aging. The main function of the kidneys is to remove waste products and excess water from our body.

In our project ,we are going to apply machine learning techniques like Adaboost, to develop a predictive model capable of accurately identifying individuals at risk of CKD based on clinical, and laboratory data.

1.2 Objectives

a) System is an health care application which is an efficient tool for disease prediction.

b) System is an real time application which is meant for physician and peoples.

c) System is an automation for chronic kidney disease prediction. System makes use of “Machine learning” algorithm CKD prediction.

d) System predicts CKD based on the attributes such as age, sugar, serum creatinine, hypertension and some other parameters.

2. Literature Survey

This section consists of the reviews of various technical and review articles on data mining techniques applied to predict Kidney Disease.

DSVGK Kaladhar, Krishna Apparao Rayavarapu and Varahalarao Vadlapudi et al . described in their research to understand machine learning techniques to predict kidney stones. They predicted good accuracy, Classification tree and Random forest (93%) followed by Support Vector Machines (SVM) (91.98%). Logistic and NN has also shown good accuracy results with zero relative absolute error and 100% correctly classified results [3]. ROC and Calibration curves using Naive Bayes has also been constructed for predicting accuracy of thedata. Machine learning approaches provide better results in the treatment of kidney stones.

J.Van Eyck, J.Ramon, F.Guiza, G.Meyfroidt, M.Bruynooghe, G.Van den Berghe, K.U.Leuvenet al. Explored data mining techniques for predicting acute kidney injury after elective cardiac surgery with Gaussian process & machine learning techniques (classification task & regressiontask) [4].

K.R.Lakshmi, Y.Nagesh and M.VeeraKrishna et al. presented performance comparison of Artificial Neural Networks, Decision Tree and Logical Regression are used for Kidney dialysissurvivability [5]. The data mining techniques were evaluated based on the accuracy measures such as classification accuracy, sensitivity and specificity. They achieved results using 10 fold cross-validations and confusion matrix for each technique. They found ANN shows better results. Hence ANN shows the concrete results with Kidney dialysis of patient records.

Morteza Khavanin Zadeh, Mohammad Rezapour, and Mohammad Mehdi Sepehri et al described in their research by using supervised techniques to predict the early risk of AVF failure in patients [6]. They used classification approaches to predict probability of complication in new hemodialysis patients whom have been referred by nephrologists to AVF surgery.

Abeer Y. Al-Hyari et al .proposed in their research by using Artificial Neural Network (NN), Decision Tree (DT) and Naïve Bayes (NB) to predict chronic kidney disease[7]. The proposed NN algorithm as well as the other data mining algorithms demonstrated high potential in successfulkidney disease.

Xudong Song, Zhanzhi Qiu, Jianwei Mu et al .introduced data mining decision tree classification method, and proposed a new variable precision rough set decision tree classification algorithm based on weighted limit number explicit region [8].

N. SRIRAAM, V. NATASHA and H. KAUR et al .presented data mining approach for parametric evaluation to improve the treatment of kidney dialysis patient [9]. Their experimental result shows that classification accuracy using Association mining between the ranges 50– 97.7% is obtained based on the dialysis parameter combination. Such a decision-based approachhelps the clinician to decide the level of dialysis required for individual patient.

Jicksy Susan Jose, R.Sivakami, N. Uma Maheswari, R.Venkatesh et al . Their research describes an efficient Diagnosis of Kidney Images Using Association Rules [10]. Their approach is divided into four major steps: pre-processing, feature extraction and selection, association rulegeneration, and generation of diagnosis suggestions from classifier.

Koushal Kumar and Abhishek et al .their research describes comparison of all three neural networks [11] such as (MLP, LVQ, RBF) on the basis of its accuracy, time taken to build model, and training data set size.

Konstantina Kourou et.al [12] proposed a study of Machine learning applications in cancer prognosis and prediction. In this paper, they have presented a review of various recent ML approaches that are applied for the prediction of cancer detection. Here they have presented review of newly published content for the work done so far in cancer detection.

P.Swathi Baby et. al [13] proposed a project to diagnosis and prediction system based on predictive mining. Here kidney disease data set is used and analysed using Weka and Orange software. Here the Machine learning algorithms such as AD Trees, J48, K star , Naïve Bayes, Random forest are used for the performance study of each algorithm which gives the Statistical analysis and predicting kidney diseases using the algorithms. Theirobservation shows that the best algorithms K-Star and Random Forest for the used Dataset .

3. System Analysis

System analysis is crucial for organizations seeking to enhance efficiency, optimize processes, or develop new systems. By comprehensively understanding both existing and proposed systems, organizations can make informed decisions that align with their strategic objectives while minimizing risks and maximizing benefits.

3.1 Existing System

The existing system develops the most accurate model for predicting chronic kidney disease (CKD) by assessing various machine learning algorithms predictive capacities.The research compares the performance of three ML algorithms- Random forest,Adaboost and Decision tree classifier.

To enhance prediction accuracy, researchers used cross-validation techniques, such as K-Fold Cross-Validation which has a drawback of increased training time and potential data leakage leading to overestimated performance.

3.1.1 Limitations of Existing System

a. No automation for CKD prediction.

b. No proper medication in emergency.

c. Requires more time for the test report.

d. Understanding the test report is difficult for peoples.

3.2 Proposed System

Chronic Kidney Disease Prediction has become the need of the patients and Physician. Although future events are uncertain, so accurate prediction is not possible.The

proposed system aims to further enhance Chronic Kidney Disease(CKD) by implementing an AdaBoost Classifier trained on preprocessed medical data that can be helpful for doctors to provide better medication and also for patients.

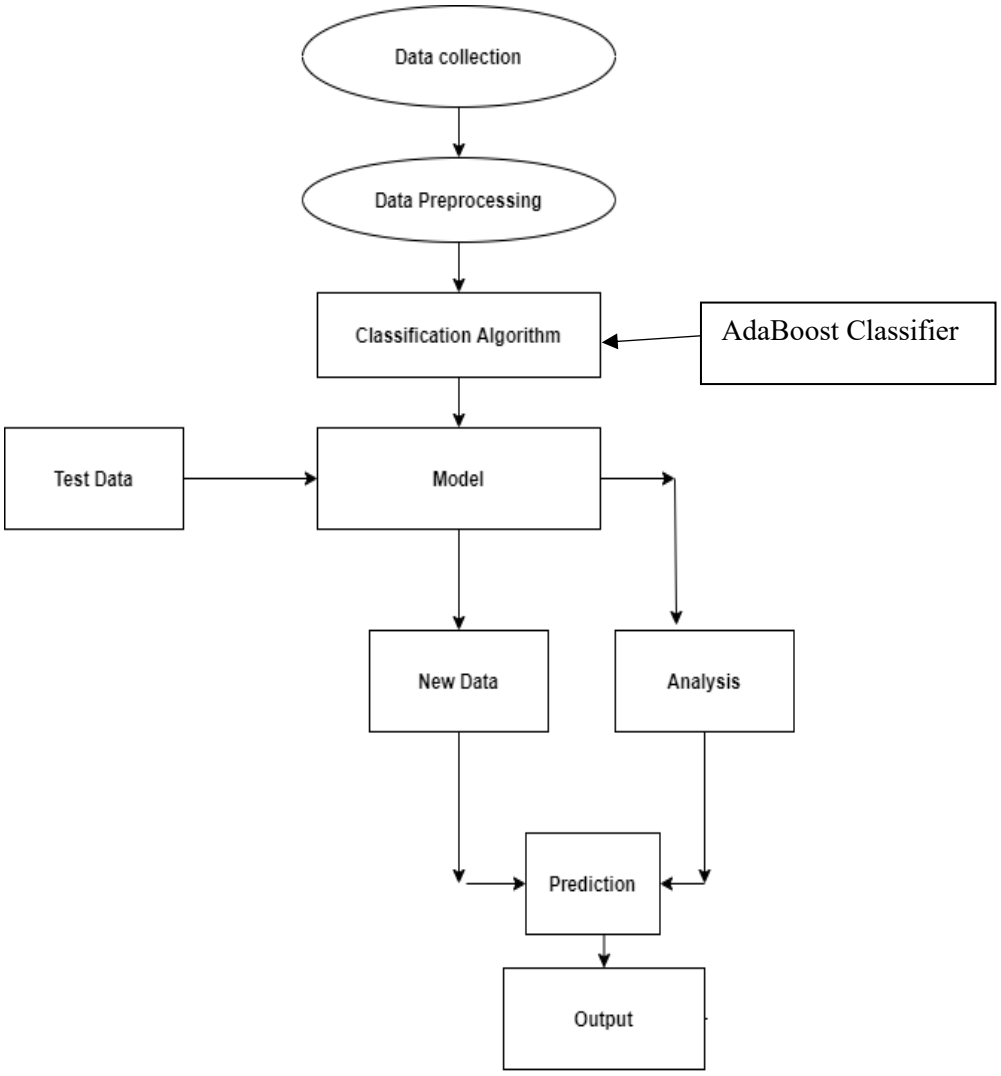


Figure 3.1:Architecture of Proposed System.

As shown in figure 3.1 Architecture of Proposed system for CKD prediction.Firstly,the data set is collected from Kaggle repository. After data collection, the focus shifts to data preprocessing, and to ensure data quality and relevance. Preprocessing techniques

include data cleaning, where missing values and inconsistencies are addressed;and feature encoding,converting categorical data into numerical format using techniques like one-hot encoding or label encoding.

Once the data is preprocessed and ready,it is split into training and testing sets for model training and evaluation.Model selection, training, and tuning occur next, where algorithms and hyperparameters are optimized for best performance. After successful training, the model is evaluated and fine-tuned before being deployed for real-world use.Once deployed, ongoing monitoring and maintenance ensure the model's accuracy and reliability.

The data preprocessing techniques used in the proposed system include:

a) Handling Missing Values:

i. Mean value imputation for missing numerical values.

$$\text{Mean Imputation} = \frac{\text{Sum of available values}}{\text{Number of available values}}$$

ii. Mode imputation for missing categorical values.

$$\text{Mode Imputation} = \text{Mode of available values}$$

b) Handling Categorical Data:

i. Value replacement for inconsistent values.

ii. Label encoding using LabelEncoder.

c) Data Cleaning and Transformation:

i. Column dropping.

ii. Column renaming.

iii. Data type conversion.

3.2.1 Ada Boost Learning Algorithm

AdaBoost, short for Adaptive Boosting, is an ensemble machine learning algorithm that can be used in a wide variety of classification and regression tasks. It is a supervised learning algorithm that is used to classify data by combining multiple weak or base learners (e.g., decision trees) into a strong learner. AdaBoost works by weighting the instances in the training dataset based on the accuracy of previous classifications.

3.2.1.1 Working of Ada Boost

Step1 – Initialize the weights

For a dataset with N training data points instances, initialize N W_{i} weights for each data point with W_{i} = 1/N .

Step2 – Train weak classifiers

Train a weak classifier Mk where k is the current iteration and calculate the weighted error E_1 of the first model:

Step3 – Calculate the error rate and importance of each weak model Mk

Calculate the weighted error E_1 of the first model:

$$E_1 = \frac{\text{Sum of misclassified samples' weights}}{\text{Total sum of weights}}$$

Calculate the importance of each model α_k using formula :

$$\text{Amount of say} = \frac{1}{2} \ln \left(\frac{1 - E_1}{E_1} \right)$$

Step4 – Update data point weight for each data point Wi

Correctly classified samples' weights are decreased:

$$w_i' = w_i \times \exp(-\text{Amount of say})$$

Misclassified samples' weights are increased:

$$w_i' = w_i \times \exp(\text{Amount of say})$$

Step5 – Normalize the Instance weight

We will normalize the instance weight so that they can be summed up to 1 using the formula:

$$\text{Normalized weight} = \frac{\text{Original weight}}{\text{Sum of all weights}}$$

Step6 – Repeat steps 2-5 for K iterations.

We will train K classifiers and will calculate model importance and update the instance weights using the above formula The final model M(X) will be an

ensemble model which is obtained by combining these weak models weighted by their model weight.

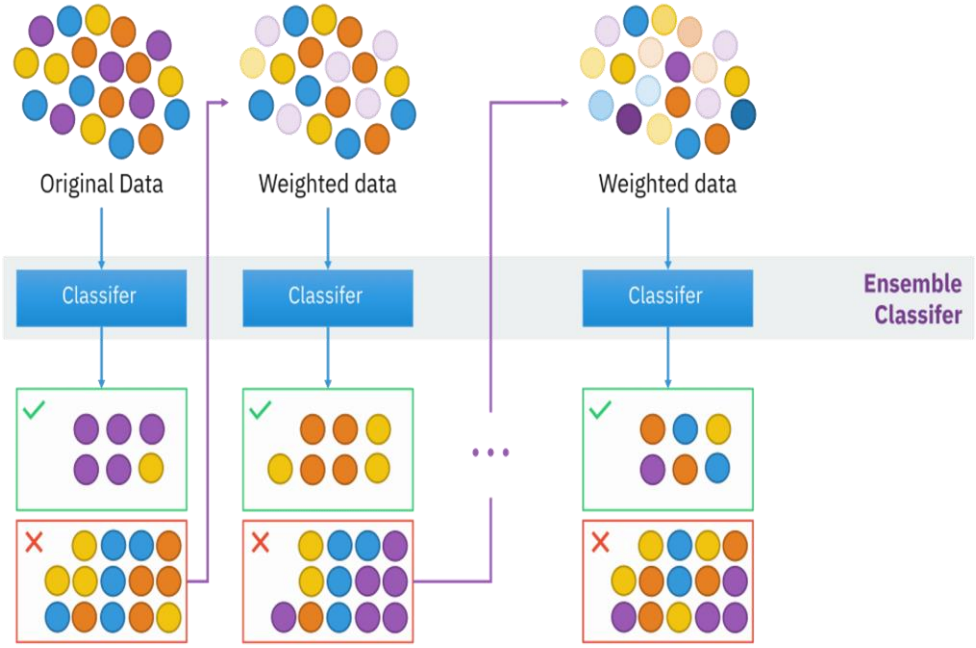


Figure 3.2:Working of AdaBoost Classifier

3.2.2 Advantages of Proposed System

a) Useful to health department to predict the CKD.

b) Useful for the patients to take better recovery.

c) We use data science techniques for accurate results.On click of button output will be generated, no too much time required for CKD prediction.

d) No need to analyze manually.

4. Software Requirements Specification

The presentation of the Software Requirements Specification (SRS) gives a review of the whole SRS with reason, scope, definitions, abbreviations, contractions, references and diagramof the SRS. The point of this report is to assemble, dissect, and give a top to bottom knowledge of the total "Chronic disease prediction" by characterizing the difficult articulation in detail. The point-by-point necessities of the Indian car purchasing conduct – client related capacities are given in this archive.

4.1 Purpose

The Purpose of the Software Requirements Specification is to give the specialized, Functional and non-useful highlights, needed to build up a web application App. The whole application intended to give client adaptability to finding the briefest as well as efficient way. To put it plainly, the motivation behind this SRS record is to give an itemized outline of our product item, its boundaries and objectives.

This archive depicts the task's intended interest group andits UI, equipment and programming prerequisites. It characterizes how our customer, group and crowd see the item and its usefulness.

4.2 Software Architecture

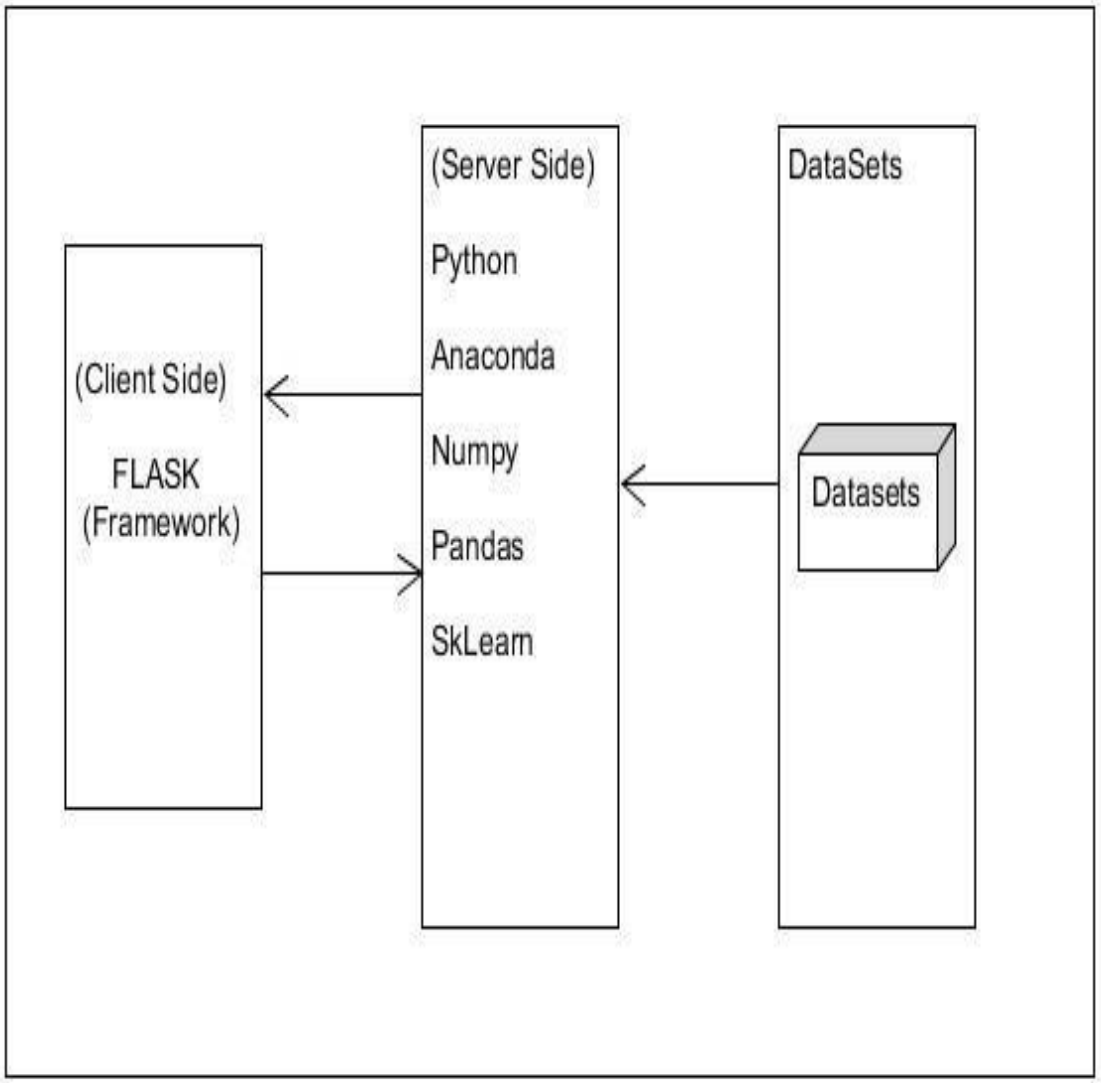


Figure 3.4: Software Architecture

4.2.1 Python:

Python is a deciphered, significant level, broadly useful programming language. Made by Guido van Rossum and first delivered in 1991, Python's plan reasoning accentuates code meaningfulness with its prominent utilization of critical whitespace. Its language develops and object-arranged methodology plan to assist software engineers with composing clear, consistent code for little and huge scope ventures.

Python is progressively composed and trash gathered. It underpins numerous programming standards, including procedural, object-arranged, and practical programming. Python is frequently portrayed as a"batteries included" language because of its thorough standard library.

Python is a multi-worldview programming language. Article arranged programming and organized writing computer programs are completely upheld, and a significant number of its highlights uphold useful programming and angle situated programming (counting by metaprogramming and metaobjects (enchantment methods)).Many different standards are upheld by means of expansions, including plan by agreement and rationale programming.

4.2.2 Flask:

Flask is a miniature web system written in Python. It is delegated a microframework in light of the fact that it doesn't need specific apparatuses or libraries.[3] It has no information base deliberation layer, structure approval, or whatever other segments where prior outsider libraries give normal capacities. In any case, Flask upholds augmentations that can include application includes as though they were executed in

Flask itself. Augmentations exist for object-social mappers, structure approval, transfer dealing with, different open confirmation advancements and a few basic system related devices. Augmentations are refreshed unmistakably more as often as possible than the center Flask program.

4.2.3 NumPy:

NumPy is the principal bundle for logical registering with Python. It contains in addition toother things.

a) Sophisticated (broadcasting) capacities.

b) tools for incorporating C/C++ and Fortran code.

c) useful straight polynomial math, Fourier change, and arbitrary number abilities.

4.2.4 Pandas:

pandas is an open source, BSD-authorized library giving elite, simple to-utilize information structures and information investigation apparatuses for the Python programming language. pandas is a Num FOCUS supported undertaking. This will help guarantee the achievement ofimprovement of pandas as an a-list open-source venture, and makes it conceivable to give to the task.

4.2.5 Sci-kit Learn

Scikit-learn is a popular machine learning library because it is easy to use and provides a wide range of algorithms. It is also well-documented and has a large community of users who can provide support.

4.3 Feasibility Study

The feasibility study helps to find solutions to the problems of the project. The solution isgiven how looks like a new system look like.

4.3.1 Technical Feasibility

The project entitled “Prediction of Chronic disease” is technically feasible because of the below mentioned features. The project is developed in Python. The web server is used to develop “Prediction Chronic disease” is local serve. The local server very neatly coordinatesbetween the design and coding parts. It provides a Graphical User Interface to design an application while the coding is done in python. At the same time, it provides high-level reliability, availability, and compatibility.

4.3.2 Economic Feasibility

In economic feasibility, cost-benefit analysis is done in which costs and benefits are evaluated.Economic analysis is used for the effectiveness of the proposed system. In economic feasibility,the main task is cost-benefit analysis. The system “Prediction of Chronic diseaseusing Data Mining Techniques” is feasible because it does not exceed the estimated cost and the estimatedbenefits are equal.

4.3.3 Operational Feasibility

The project entitled “Prediction of Chronic disease” is technically feasible because of the belowmentioned features. The system predicts the chronic disease prediction based on the historical data, further the details of the patient are added to the Data Base. The performance of the Data miningtechniques are compared based on their execution time and displayed it through a graph.

4.3.4 Behavior Feasibility

The project entitled “Prediction of Chronic disease using deep learning and Machine Learning” isbeneficial because it satisfies the objectives when developed and installed.

4.4 Overview

Following a section of this document will focus on describing the system in terms of product functions. In the next section, we will address specific requirements of the system, which willenclose functional requirements and non-functional requirements.

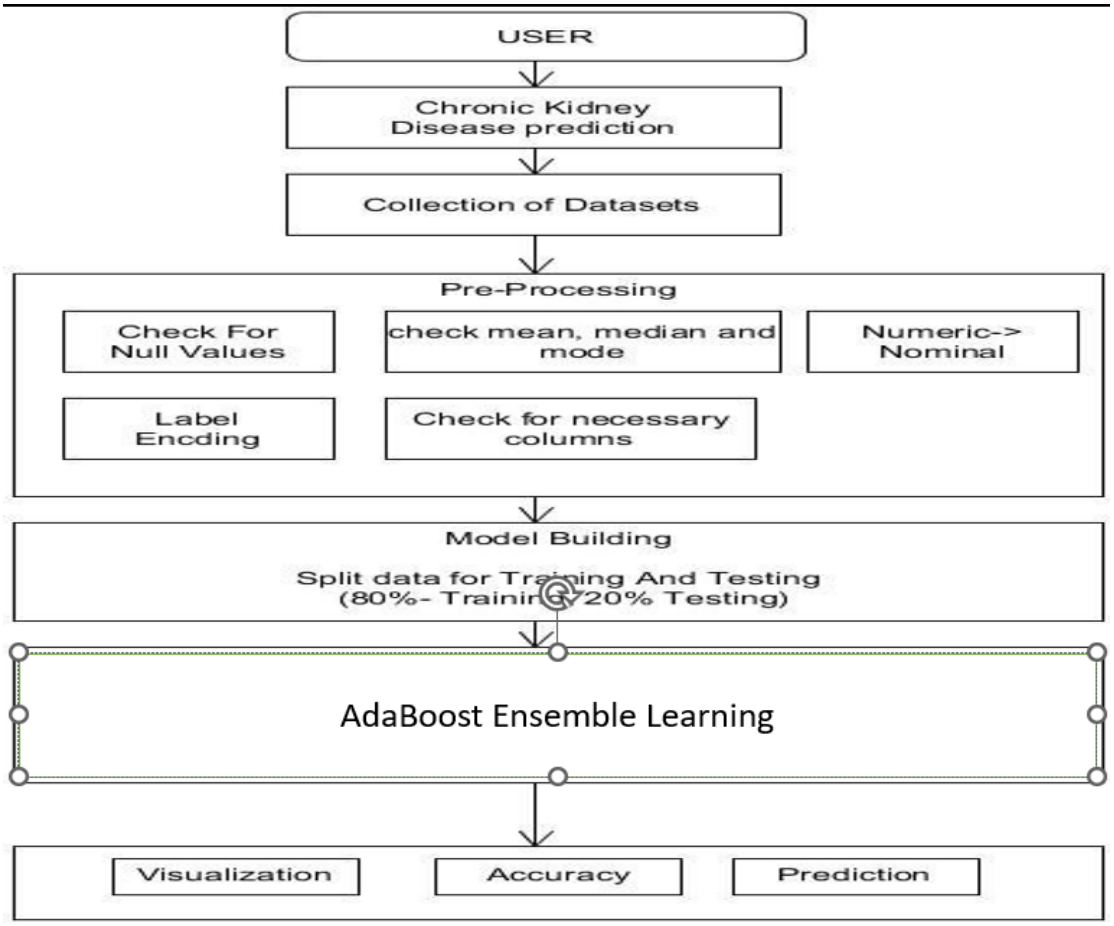


Figure 3.3: Overview

4.5 General description

General Description section, with its focus on product functions and constraints, plays a crucial role in setting the stage for a successful system analysis and project execution. It ensures that all stakeholders are on the same page regarding what the system will do and any factors that might impact its development and operation.

4.5.1 Product Functions

a) Collected datasets of chronic disease prediction from Kaggle.

b) Pre-processing of obtained datasets.

c) Select Attributes which helps in predicting the stock.

d) The selected datasets are trained using AdaBoost.

e) The trained data sets are tested for Accuracy.

f) The obtained result is showed in the graph.

4.5.2 General constraints

a) The system should have enough RAM and Disk Storage Space.

b) The Source code must be written in Python for ML.

c) The results generated have to be entered in to the system and any error or any valueentered out of the boundary will not be understood by the system.

4.6 Specific Requirements

Specific requirements refer to detailed, precise, and unambiguous descriptions of what a software system or component must do to meet its intended purpose. These requirements provide the foundation for software design, development, testing, and validation. Specific requirements can be functional or non-functional:

4.6.1 Functional Requirements

A functional requirement defines a function of a system or its component. A role is describedas a set of inputs, behaviours, and outputs. Functional requirements may be calculations, technical details, data manipulation, and processing.

The Methods of the system are as follows.

Data preprocessing: Dataset will be added to the preprocessing

a) Input: Chronic dataset

b) Process: Preprocessing includes handling missing value and error data.

c) Output: preprocessed dataset.

d) Error handling: If the input file is not a valid one.

a) Input: preprocessed dataset.

b) Process: It will split the data into the train set and test set.

c) Output: Dataset will be displayed as Train set and Test set and it will be tested for thespecific algorithms and performance analysis will be carried out.

4.7 System Requirements

System requirements refer to the specifications and constraints that define the hardware, software, and operational conditions necessary for a system or application to function effectively. These requirements ensure that a given system can be properly installed, configured, and operated within a specific environment, providing a clear understanding of the resources needed for successful implementation.

4.7.1 Hardware Requirements

a) Processor : Intel i3

b) Hard-Disk :500GB

c) RAM :4GB or Above

4.7.2 Software Requirements

a) Operating System : Windows 7 and above

b) Front End : Html, CSS

c) Framework : Flask

d) Language : Python

e) Libraries : Pandas,Numpy,Sklearn,Scikit.

f) Editor : Visual Studio Code

5. System Design

The Software Design will be used to aid in software development for android application by providing the details for how the application should be built. Within the Software Design, specifications are narrative and graphical documentation of the software design for the projectincludes use case models, sequence diagrams, and other supporting requirement information.

5.1 Scope

The design Document is for a primary level system, which will work as a basement for buildinga system that provides a base level of functionality to show feasibility for large-scale production use. The software Design Document, the focus placed on the generation and modification of the documents. The system will be used in conjunction with other pre-existingsystems and will consist largely of a document interaction faced that abstracts document interactions and handling of the document objects. This Document provides the Designspecifications of “Chronic Disease detection”.

5.2 Data Flow Diagram

LEVEL 0 DFD: Here Dataset will be given as input and will be processed for further implementation.

Figure 5.1: Level 0 DFD

LEVEL 1 DFD: Using python libraries and algorithms prediction will be carried out

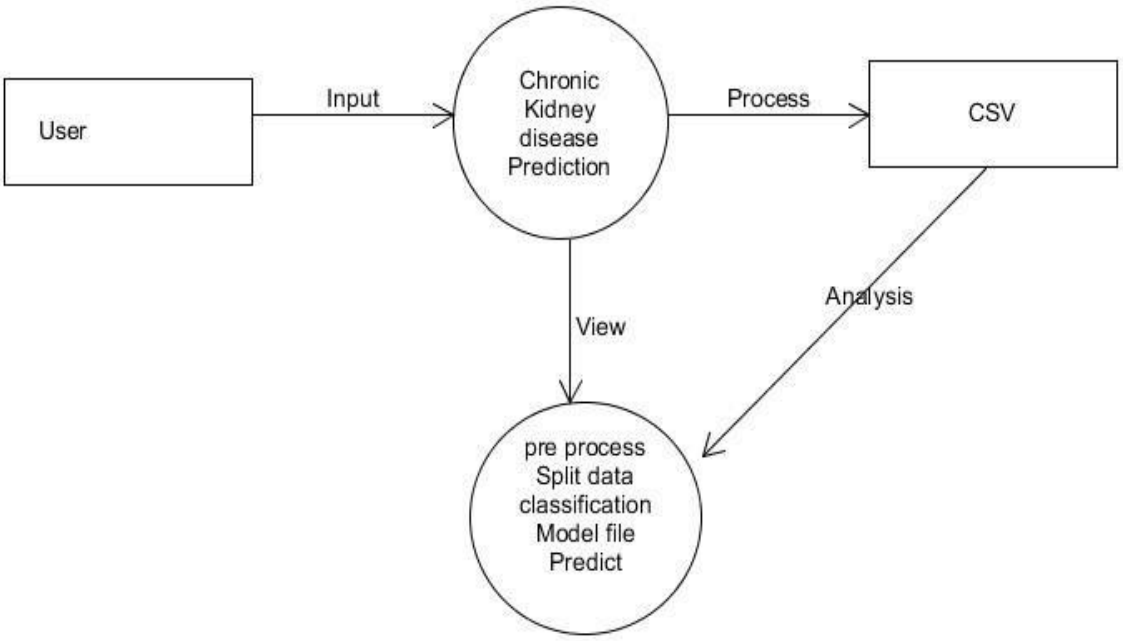
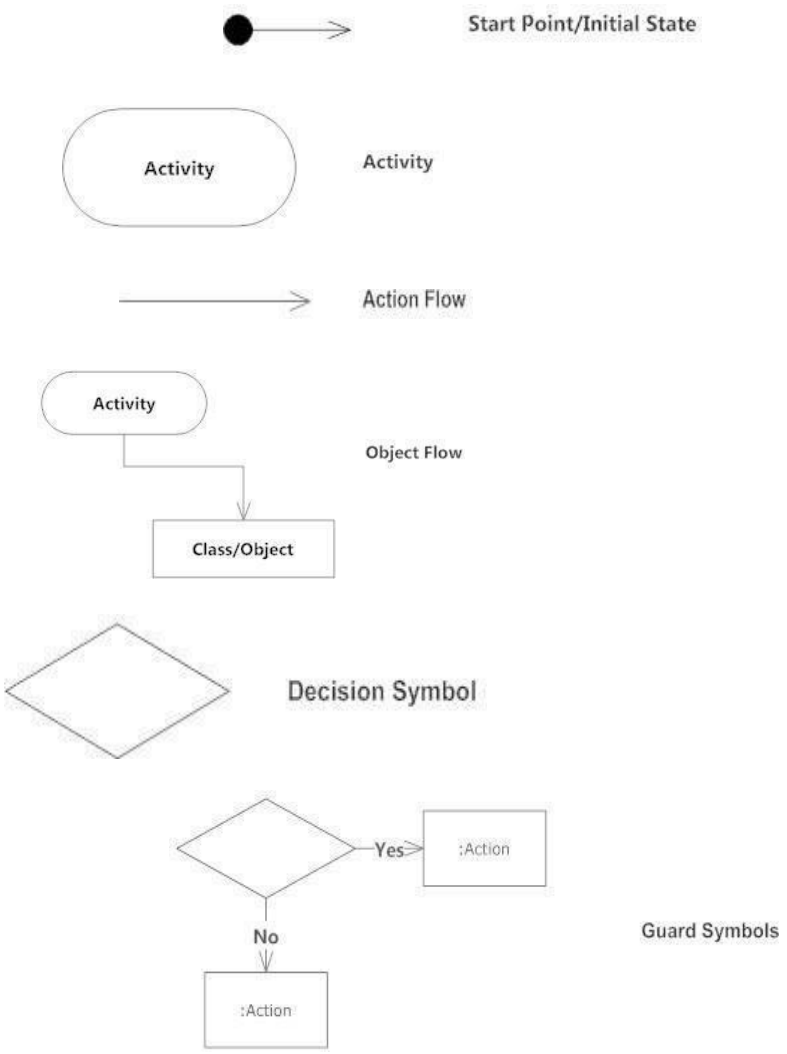


Figure 5.2: Level 1 DFD

5.3 Activity Diagram

An activivity diagram outwardly presents a progression of activities or stream of control in a framework like a flowchart or an information stream chart. Action graphs are regularly utilized in business measuredemonstrating. They can likewise depict the means in an utilization case chart. Exercises demonstratedcan be consecutive and simultaneous. In the two cases, an action outline will have a start (an underlyingstate) and an end (a last state).



Figures 5.3: Symbols

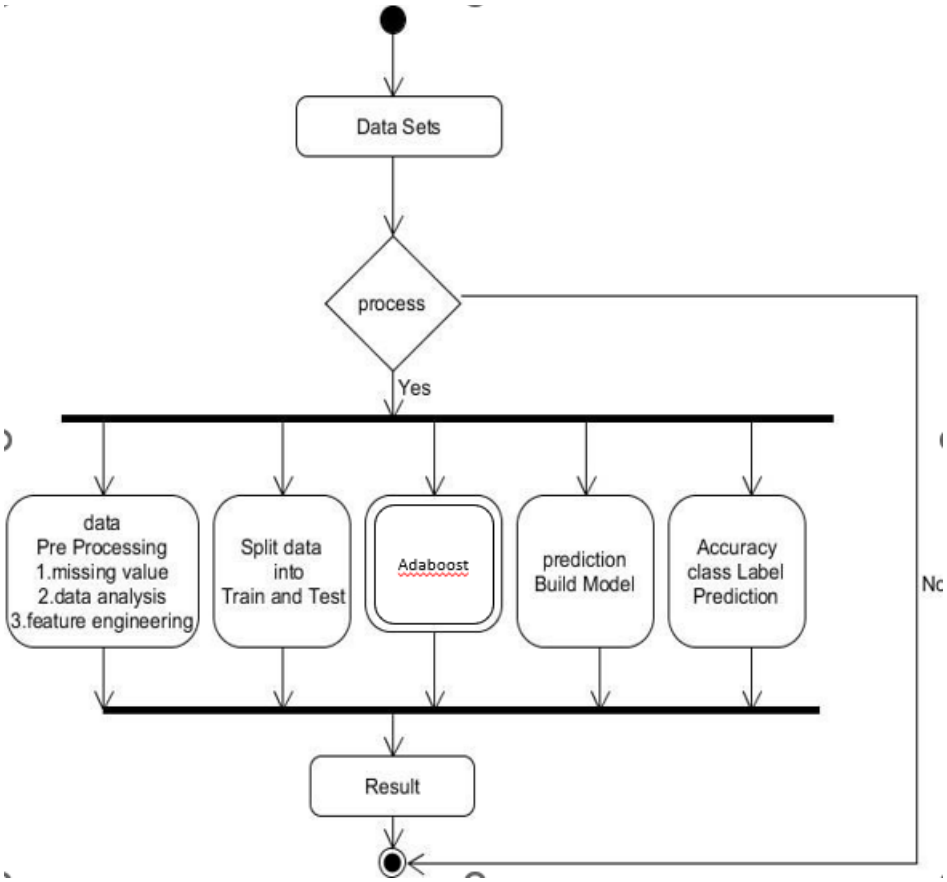


Figure 4.4 : Activity Diagram

5.4 Sequence Diagram

Sequence diagram depict cooperations among classes as far as a trade of messages after some time. They're likewise called occasion charts. A grouping chart is a decent method to envisionand approve different runtime situations. These can assist with anticipating how a framework will act and to find duties a class may need to have during the time spent demonstrating anotherframework.

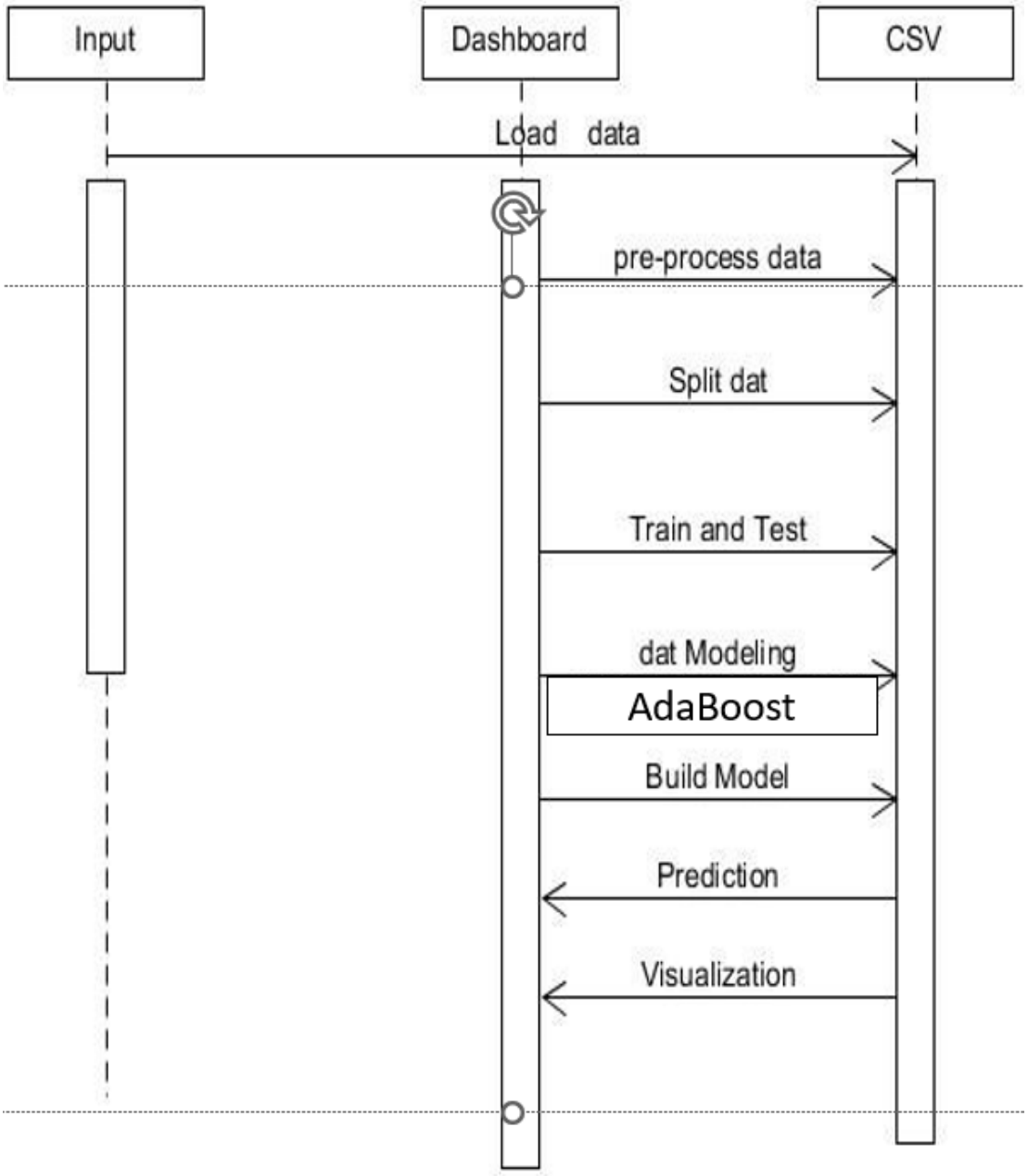


Figure 4.5 : Sequential Diagram

5.5 Use-Case Diagram

The motivation behind use case diagram is to catch the dynamic part of a framework. In any case, this definition is too nonexclusive to even think about describing the reason, as other fouroutlines (action, grouping, cooperation, and Statechart) likewise have a similar reason. We willinvestigate some particular reason, which will recognize it from other four charts.

Use case graphs are utilized to accumulate the prerequisites of a framework including inside and outside impacts. These prerequisites are generally plan necessities. Consequently, when aframework is investigated to accumulate its functionalities, use cases are readied and entertainers are distinguished

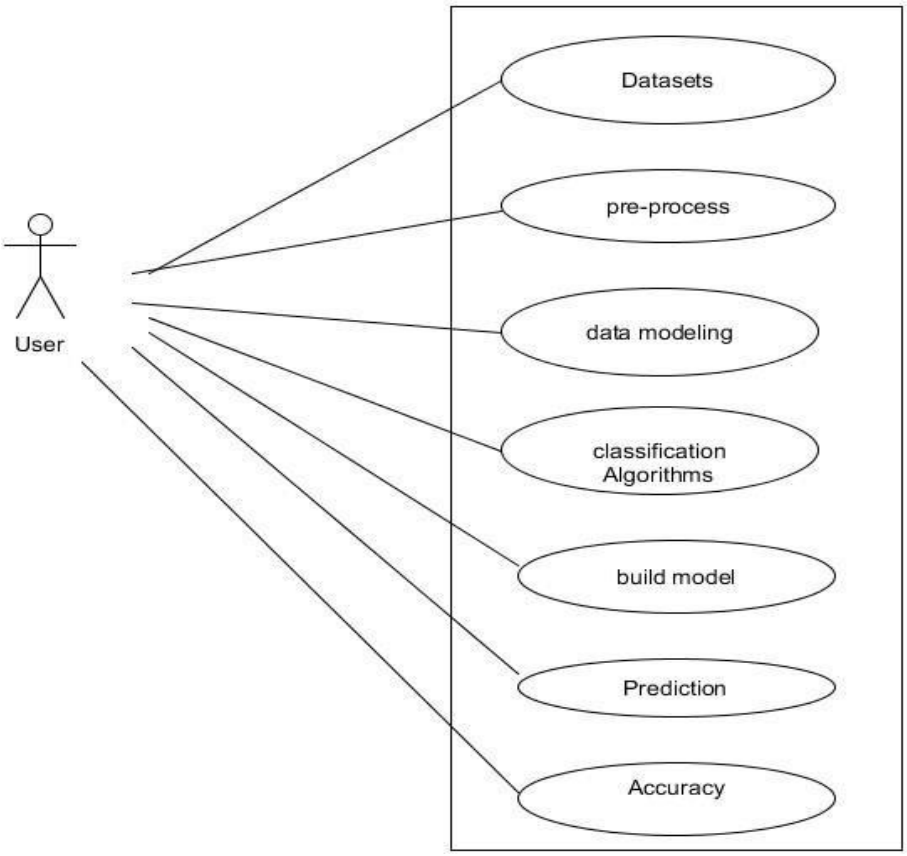


Figure 5.5:Use Case Diagram

6. Implementation

The project is implemented using Python which is an object oriented programming language and procedure oriented programming language. Object oriented programming is an approach that provides a way of modularizing program by creating partitioned memory area of both data and function that canbe used as a template for creating copies of such module on demand.Python is dynamically typed and garbage-collected. It supports multiple programming paradigms, including procedural, object-oriented,and functional programming.

6.1 Machine Learning overview

Machine learning is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of Computer Programs that can change when exposed to new data.

Machine learning involves a computer to be trained using a given data set, and use this training topredict the properties of a given new data. For example, we can train a computer by feeding it 1000 images of cats and 1000 more images which are not of a cat, and tell each time to the computer whether a picture is cat or not. Then if we show the computer a new image, then from the above training, thecomputer should be able to tell whether this new image is a cat or not. The process of training and prediction involves the use of specialized algorithms. We feed the training data to an algorithm, and the algorithm uses this training data to give predictions on a new test data. Adapative Boosting is technique used to predict CKD.

6.1.1 Challenges in Implementing Machine Learning

Most insurers recognize the value of machine learning in driving better decision-making and streamlining business processes. Research for the Accenture Technology Vision 2018 shows that more than 90 percent of insurers are using, plan to use or considering using machine learning or AI in the claims or underwriting process.Some of the challenges insurers typically encounter when adopting machine learning are.

a) **Training requirements**:AI-powered intellectual systems must be trained in a domain, e.g., claims or billing for an insurer. This requires a separate training system, which insurers find hard to provide for training the AI model. Models need to be trained with huge volumes of documents/transactions to cover all possible scenarios.

b) **Right data source:** The quality of data used to train predictive models is equally important as the quantity, in the case of machine learning. The datasets need to be representative and balanced so that they can give a better picture and avoid bias. Thisis important to train predictive models. Generally, insurers struggle to provide relevantdata for training AI models.

c) **Difficulty in predicting returns:** It's not very easy to predict improvements that machine learning can bring to a project. For example, it's not easy to plan or budget a project using machine learning, as the funding needs may vary during the project, based on the findings. Therefore, it is almost impossible to predict the return on investment. This makes it hard to get everyone on board the concept and invest in it.

d) **Data security**:The huge amount of data used for machine learning algorithms has created an additional security risk for insurance companies.

With such an increase in collected data and connectivity among applications, there is a risk of data leaks and security breaches. A security incident could lead to personal information falling into the wrong hands. This creates fear in the minds of insurers.

e) **Overfitting and Underfitting :** Overfitting occurs when a model learns too much from the training data, failing to generalize. Underfitting happens when the model doesn't learn enough. Balancing these is key to building robust models.

6.2 Workflow

A workflow is a systematic approach to organizing and executing tasks in a project. In a machine learning context, it involves multiple stages, from data collection to model deployment, with each step logically leading to the next. Workflows help ensure efficiency, consistency, and repeatability in project execution.

6.2.1 Data Collection

a) Dataset: Collect a dataset related to Chronic Kidney Disease (CKD) from Kaggle. These datasets typically contain a variety of medical indicators such as blood pressure, age, glucose levels, albumin, hemoglobin, serum creatinine, and more, which are used for CKD detection and prediction.

b) Source: Kaggle provides numerous datasets for machine learning projects. Find a suitable CKD dataset that meets your project's needs.

age	bp	sg	al	su	rbc	pc	pcc	ba	bgr	bu	sc	sod	pot	hemo	pcv	wbcc
48	80	1.02	1	0		normal	notpreser	notpreser	121	36	1.2			15.4	44	7800
7	50	1.02	4	0		normal	notpreser	notpresent		18	0.8			11.3	38	6000
62	80	1.01	2	3	normal	normal	notpreser	notpreser	423	53	1.8			9.6	31	7500
48	70	1.005	4	0	normal	abnormal	present	notpreser	117	56	3.8	111	2.5	11.2	32	6700
51	80	1.01	2	0	normal	normal	notpreser	notpreser	106	26	1.4			11.6	35	7300
60	90	1.015	3	0			notpreser	notpreser	74	25	1.1	142	3.2	12.2	39	7800
68	70	1.01	0	0		normal	notpreser	notpreser	100	54	24	104	4	12.4	36	
24		1.015	2	4	normal	abnormal	notpreser	notpreser	410	31	1.1			12.4	44	6900
52	100	1.015	3	0	normal	abnormal	present	notpreser	138	60	1.9			10.8	33	9600
53	90	1.02	2	0	abnormal	abnormal	present	notpreser	70	107	7.2	114	3.7	9.5	29	12100
50	60	1.01	2	4		abnormal	present	notpreser	490	55	4			9.4	28	
63	70	1.01	3	0	abnormal	abnormal	present	notpreser	380	60	2.7	131	4.2	10.8	32	4500
68	70	1.015	3	1		normal	present	notpreser	208	72	2.1	138	5.8	9.7	28	12200
68	70						notpreser	notpreser	98	86	4.6	135	3.4	9.8		
68	80	1.01	3	2	normal	abnormal	present	present	157	90	4.1	130	6.4	5.6	16	11000
40	80	1.015	3	0		normal	notpreser	notpreser	76	162	9.6	141	4.9	7.6	24	3800
47	70	1.015	2	0		normal	notpreser	notpreser	99	46	2.2	138	4.1	12.6		
47	80						notpreser	notpreser	114	87	5.2	139	3.7	12.1		
60	100	1.025	0	3		normal	notpreser	notpreser	263	27	1.3	135	4.3	12.7	37	11400
62	60	1.015	1	0		abnormal	present	notpreser	100	31	1.6			10.3	30	5300
61	80	1.015	2	0	abnormal	abnormal	notpreser	notpreser	173	148	3.9	135	5.2	7.7	24	9200
60	90						notpreser	notpresent		180	76	4.5		10.9	32	6200

6.2.2 Data Pre-Processing

Data preprocessing is a crucial step in any machine learning project, including CKD detection. It involves preparing and transforming raw data into a form that can be used effectively by a machine learning algorithm. The quality of data preprocessing can significantly impact model performance and accuracy.

Data cleaning is the essential part of data pre-processing. It involves identifying and correcting errors, handling missing or inconsistent data, and ensuring the data is consistent and reliable. Label encoding transforms categorical data into numerical form, making it suitable for machine learning algorithms. Together, these steps ensure that the data is properly cleaned and formatted.

6.2.3 Data Splitting

Data splitting is a fundamental step in the machine learning pipeline, ensuring that the model is trained and tested on separate subsets of data. This process helps evaluate how well a model generalizes to new, unseen data, which is crucial for preventing overfitting and underfitting.

Train-Test Split is the most basic form of data splitting. The dataset is divided into two subsets: one for training the model and another for testing it.A typical split ratio is 80/20 or 70/30, where 80% or 70% of the data is used for training, and 20% or 30% is used for testing.This method provides a straightforward way to evaluate a model's performance on unseen data.

6.2.4 Model Selection and Training

AdaBoost (Adaptive Boosting) Classifier: It is an ensemble learning algorithm designed to improve the performance of weak classifiers by combining them to create a stronger overall model. It works by iteratively adjusting the weights of training samples to focus on those that are harder to classify.

Training: Train the AdaBoost classifier on the training data.

6.2.5 Model Evaluation

Model evaluation involves assessing a model's performance to determine how well it meets the objectives of the project. A thorough model evaluation provides insights into the model's strengths and weaknesses, helping to guide further development and optimization. Test the trained AdaBoost classifier on the testing data to evaluate its accuracy, precision, recall, F1-score, and other relevant metrics.

6.2.5.1 Classification Metrics

a) **Accuracy:** The ratio of correctly predicted samples to the total number of samples. While simple, it may be misleading in imbalanced datasets.

b) **Precision:** The ratio of true positives to the total number of predicted positives. It reflects the accuracy of positive predictions.

c) **Recall:** The ratio of true positives to the total number of actual positives. It measures how many actual positives were captured by the model.

d) **F1-Score:** The harmonic mean of precision and recall. It's useful when you need a balanced metric for imbalanced datasets.

6.2.6 Deployment

Deployment is a crucial step in the machine learning lifecycle, where a trained model becomes a functional component of a broader system. Successful deployment requires careful planning, robust infrastructure, and adherence to best practices for testing, monitoring, security, and scalability. By following these guidelines, you can ensure a reliable, scalable, and secure deployment of your machine learning model, providing value to users and stakeholders.

Deployment in the context of machine learning involves taking a trained model and integrating it into a production environment where it can be used to make predictions or classifications in real time or as part of a batch process. This phase transforms a trained model from a static object into a dynamic, operational component that delivers value to end-users, businesses, or other stakeholders.

6.3 GitHub link

- https://github.com/PuttaNiharika/Teamc12.git

7. Testing

Testing is the way toward running a framework with the expectation of discovering blunders. Testingupgrades the uprightness of the framework by distinguishing the deviations in plans and blunders in the framework. Testing targets distinguishing blunders – prom zones. This aides in the avoidance of mistakes in the framework. Testing additionally adds esteems to the item by affirming the client's necessity.

Testing must be intensive and all around arranged. A somewhat tried framework is as terrible as an untested framework. Furthermore, the cost of an untested and under-tried framework is high. The execution is the last and significant stage. It includes client preparation, framework testing so as to guarantee the effective running of the proposed framework. The client tests the framework and changes are made bytheir requirements. The testing includes the testing of the created framework utilizing different sorts of information. While testing, blunders are noted and rightness is the mode.

7.1 Objective Of Testing

Testing in a cycle of executing a program with the expectation of discovering mistakes. An effective experiment is one that reveals an up 'til now unfamiliar blunder. Framework testing is a phase of usage, which is pointed toward guaranteeing that the framework works accurately and productively according to the client's need before the live activity initiates. As expressed previously, testing is indispensable to the achievement of a framework. Framework testingmakes the coherent presumption that if all the framework is right, the objective will be effectively accomplished. A progression of tests are performed before the framework is prepared for the client

acknowledgment test.

7.2 Testing Methods

System testing is a stage of implementation. This helps the weather system works accurately and efficiently before live operation commences. Testing is vital to the success of the system. The candidate system is subject to a variety of tests: online response, volume, stress, recovery, security and usability tests series of tests are performed for the proposed system are ready for user acceptancetesting.

7.2.1 White Box Testing

The test is conducted during the code generation phase itself. All the errors were rectified at the moment of its discovery. During this testing, it is ensured that

a) Exercise all logical decisions on their true or false side.

b) Execute all loops at their boundaries.

7.2.2 Black Box Testing

It is focused around the practical necessities of the product. It's anything but a choice to white boxtesting; rather, it is a reciprocal methodology that is probably going to reveal an alternate class of blunders than White Box strategies. It is endeavored to discover mistakes in the accompanying classes.

a) Incorrect or missing capacities

b) Interface blunders.

c) Errors in an information structure or outside information base access

7.2.3 Unit Testing

Unit testing chiefly centers around the littlest unit of programming plan. This is known as module testing. The modules are tried independently. The test is done during the programming stage itself. Inthis progression, every module is discovered to be working acceptably as respects the normal yield from the module.

7.2.4 Integration Testing

Mix testing is an efficient methodology for developing the program structure, while simultaneouslyleading tests to reveal blunders related with the interface. The goal is to take unit tried modules andmanufacture a program structure. All the modules are joined and tried in general.

7.2.5 Output Testing

Subsequent to performing approval testing, the following stage is yield trying of the proposedframework, since no framework could be valuable on the off chance that it doesn't create the necessary yield in a particular configuration. The yield design on the screen is discovered to be right. The organization was planned in the framework configuration time as indicated by the client needs. For the printed copy likewise, the yield comes according to the predefined prerequisites by the client. Subsequently yield testing didn't bring about any amendment for the framework.

7.3 Validation

Validation is a critical process that ensures software meets stakeholder expectations, fulfills its business goals, and delivers value to users. By conducting thorough validation, organizations can improve software quality, reduce risks, and increase user

satisfaction. The primary goal of validation is to ensure that the software aligns with stakeholder expectations, meets business goals, and is usable, reliable, and compliant with applicable regulations and standards.

7.3.1 Methods of Validation:

a) **User Acceptance Testing (UAT):** Involves stakeholders and end-users testing the software to ensure it meets their needs.

b) **System Testing:** Focuses on testing the software as a whole to ensure that it delivers the expected outcomes.

c) **Beta Testing:** Invites a broader group of users to test the software in a real-world environment.

d) **Simulation and Prototyping:** Helps stakeholders visualize and interact with the software to confirm its functionality and usability.

7.3.2 Importance of Validation:

a) Validation ensures that the software is aligned with user requirements and expectations, reducing the risk of project failure due to unmet needs.

b) It helps identify gaps between the expected and actual functionality, allowing for corrections before the software is deployed.

c) It increases stakeholder confidence and satisfaction by demonstrating that the software meets business objectives.

8. Results

The results section presents the outcomes of our project, focusing on the classification report and the associated output from the machine learning models we developed. The purpose of this section is to provide a detailed analysis of the project's performance and demonstrate how well the model met the project's objectives.

8.1 Classification Report

A classification report is a detailed summary of the performance of a classification model, providing key metrics that help evaluate the effectiveness of the model in categorizing or predicting data into distinct classes. These metrics help stakeholders understand how well the model is performing and where it might need improvement.

8.1.1 Significance of the Classification Report

A classification report helps in several ways:

a) **Evaluating Model Performance:** It provides a comprehensive view of how well the model is classifying data. This can help identify strengths and weaknesses in the model's predictions.

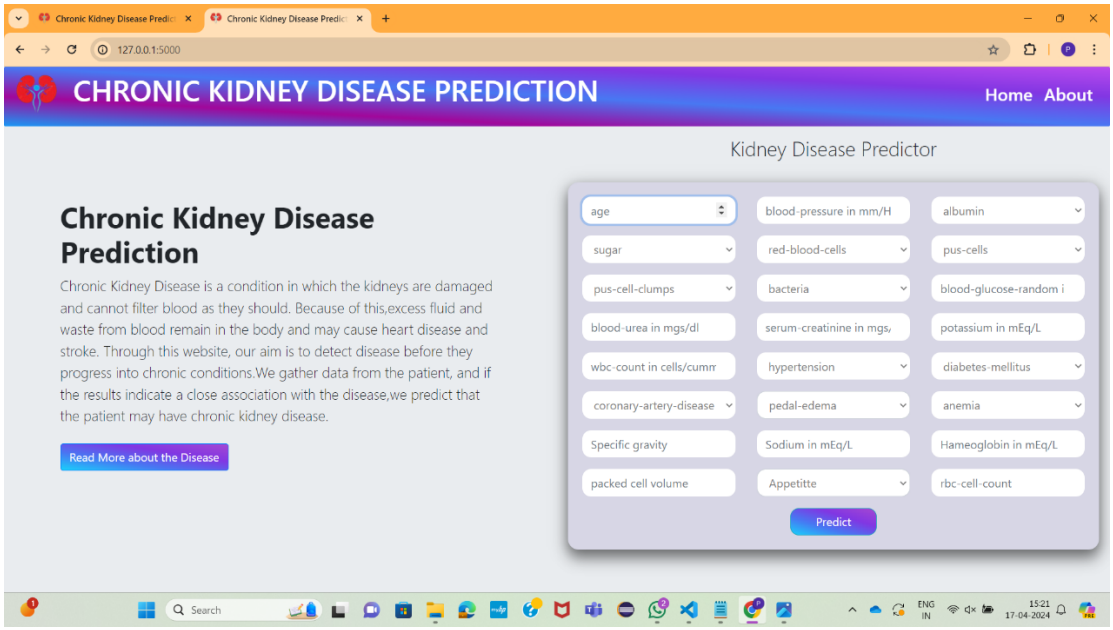
b) **Guiding Model Improvement:** By analyzing precision, recall, and F1-scores, you can determine whether the model is biased towards certain classes, has issues with overfitting or underfitting, or requires additional tuning.

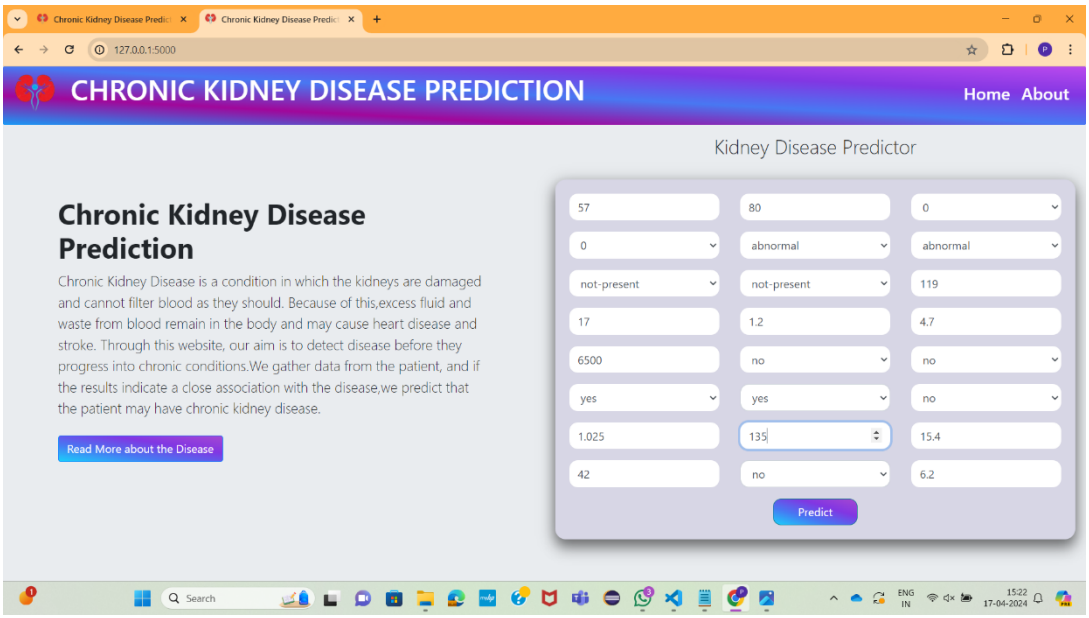
c) **Communicating Results:** The classification report is a concise way to communicate the model's effectiveness to stakeholders, enabling them to understand the project's outcomes and make informed decisions.

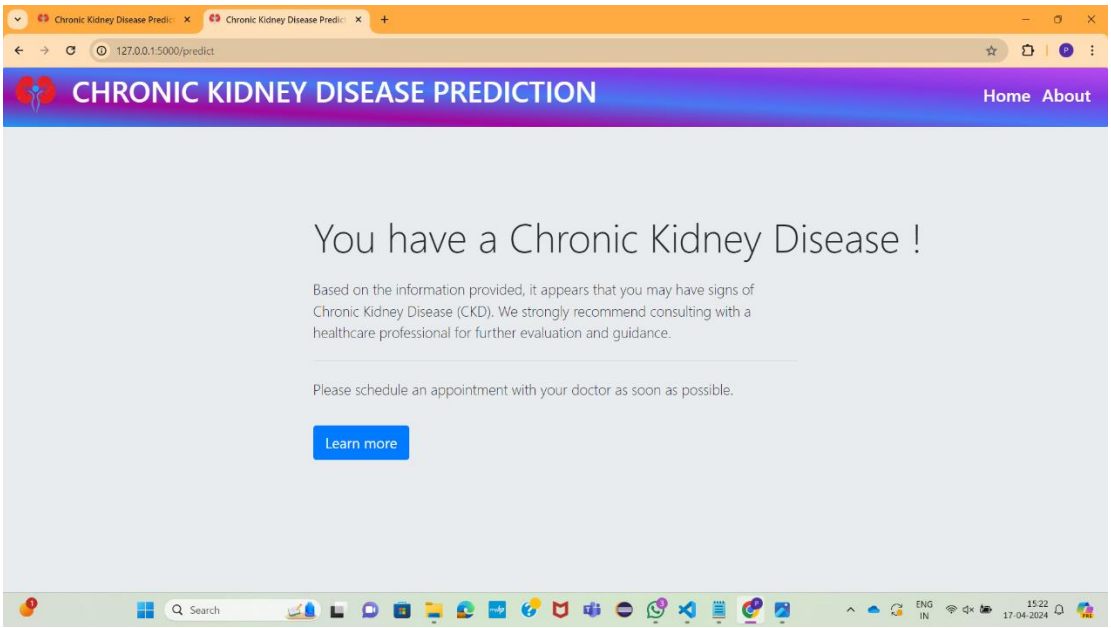
Finally Concluding that Ada Boost Algorithm achieved 98% Accuracy here some of the metrics for the Ada Boost.

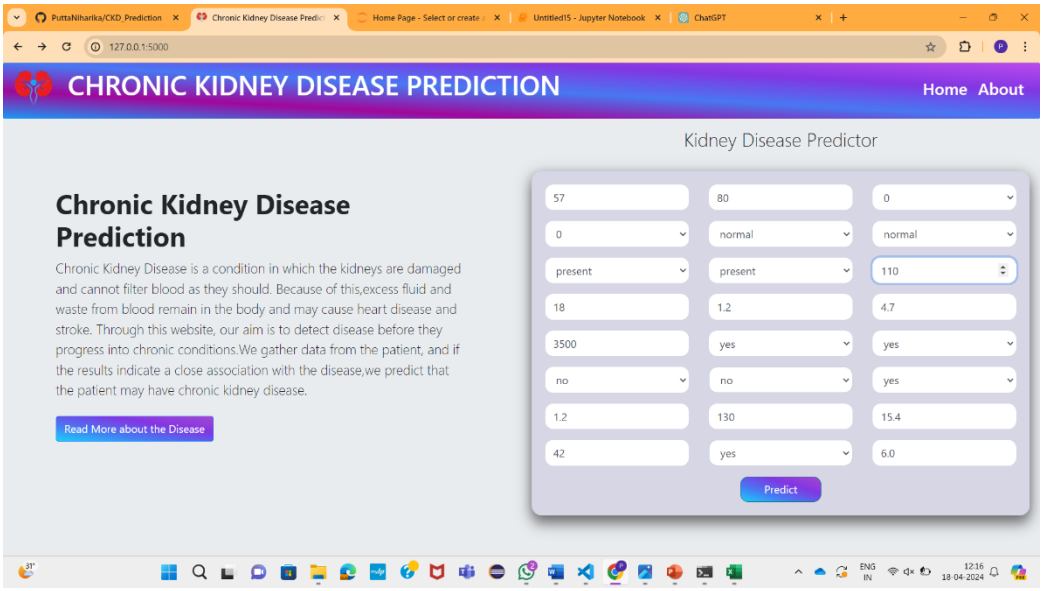
```
Training Accuracy of Ada Boost Classifier is 1.0
Test Accuracy of Ada Boost Classifier is 0.9833333333333333

Confusion Matrix :-
[[72  0]
 [ 2 46]]

Classification Report :-
              precision    recall  f1-score   support

           0       0.97      1.00      0.99        72
           1       1.00      0.96      0.98        48

    accuracy                           0.98       120
   macro avg       0.99      0.98      0.98       120
weighted avg       0.98      0.98      0.98       120
```

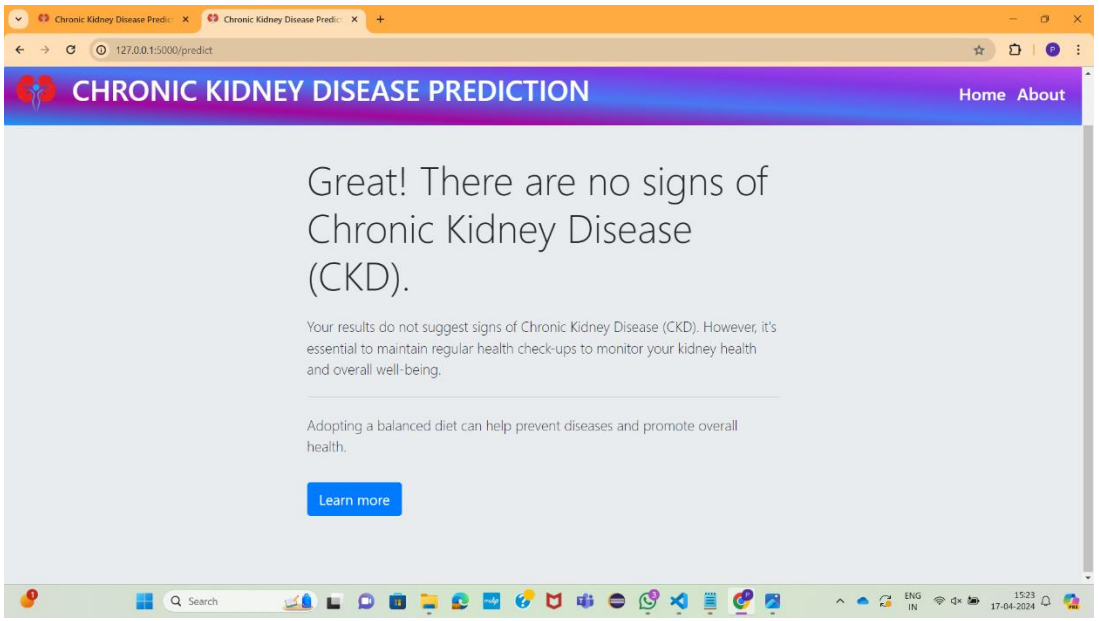
Figure 8.1:Clasification Report

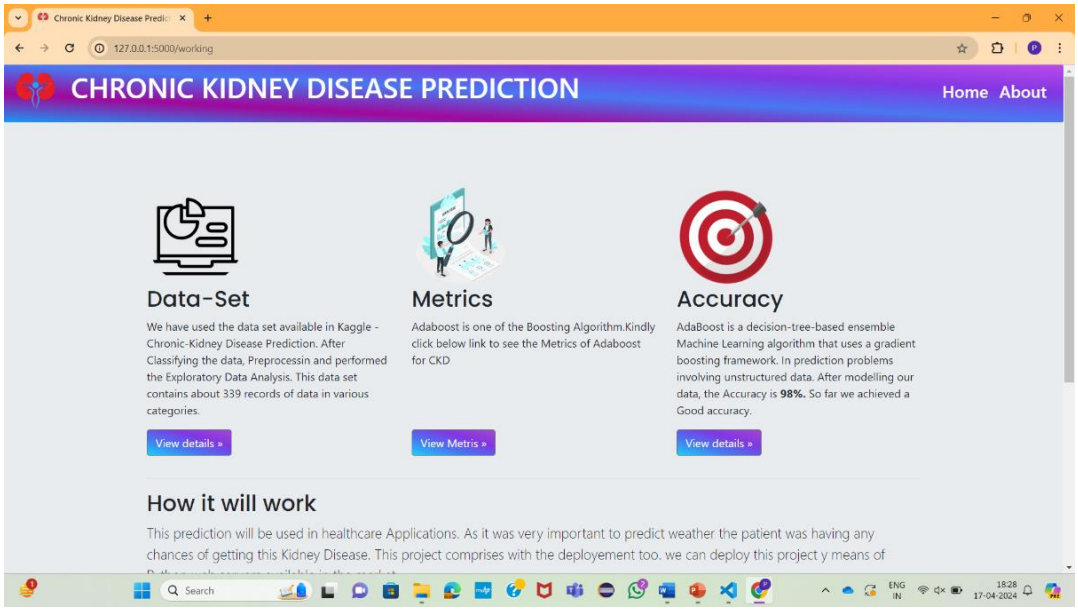


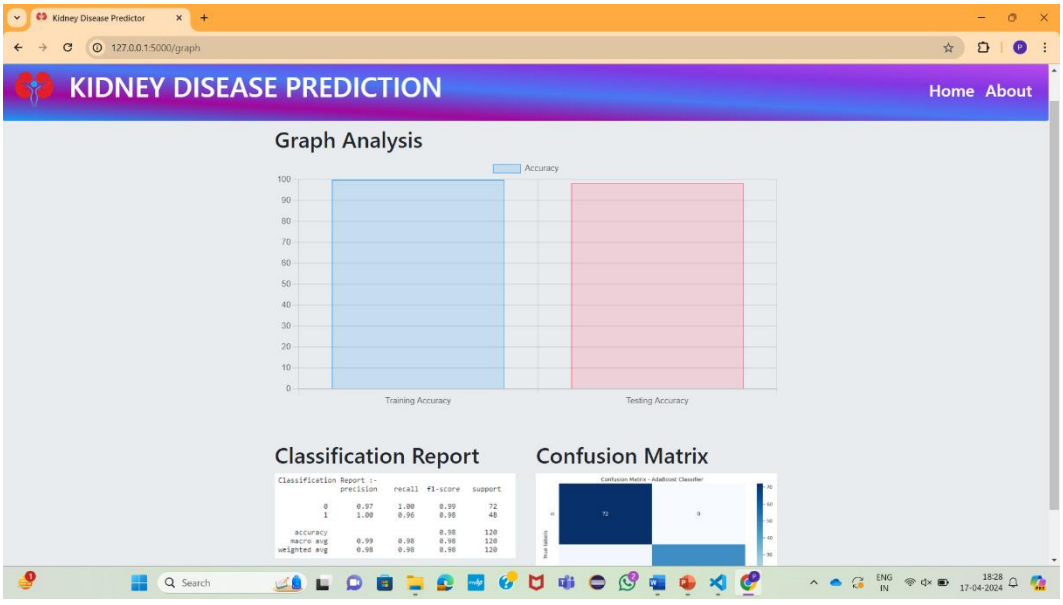

CHRONIC KIDNEY DISEASE PREDICTION
Home About
Kidney Disease Predictor
Chronic Kidney Disease Prediction
Chronic Kidney Disease is a condition in which the kidneys are damaged and cannot filter blood as they should. Because of this,excess fluid and waste from blood remain in the body and may cause heart disease and stroke. Through this website, our aim is to detect disease before they progress into chronic conditions.We gather data from the patient, and if the results indicate a close association with the disease,we predict that the patient may have chronic kidney disease.
Read More about the Disease
57
80
0
0
abnormal
abnormal
not-present
not-present
119
17
1.2
4.7
6500
no
no
yes
yes
no
1.025
135
15.4
42
no
6.2
Predict


CHRONIC KIDNEY DISEASE PREDICTION
Home About
You have a Chronic Kidney Disease !
Based on the information provided, it appears that you may have signs of Chronic Kidney Disease (CKD). We strongly recommend consulting with a healthcare professional for further evaluation and guidance.
Please schedule an appointment with your doctor as soon as possible.
Learn more


CHRONIC KIDNEY DISEASE PREDICTION
Home About
Kidney Disease Predictor
Chronic Kidney Disease Prediction
Chronic Kidney Disease is a condition in which the kidneys are damaged and cannot filter blood as they should. Because of this,excess fluid and waste from blood remain in the body and may cause heart disease and stroke. Through this website, our aim is to detect disease before they progress into chronic conditions.We gather data from the patient, and if the results indicate a close association with the disease,we predict that the patient may have chronic kidney disease.
Read More about the Disease
57
80
0
0
normal
normal
present
present
110
18
1.2
4.7
3500
yes
yes
no
no
yes
1.2
130
15.4
42
yes
6.0
Predict


CHRONIC KIDNEY DISEASE PREDICTION
Home About
Great! There are no signs of Chronic Kidney Disease (CKD).
Your results do not suggest signs of Chronic Kidney Disease (CKD). However, it's essential to maintain regular health check-ups to monitor your kidney health and overall well-being.
Adopting a balanced diet can help prevent diseases and promote overall health.
Learn more


CHRONIC KIDNEY DISEASE PREDICTION
Home About
Data-Set
We have used the data set available in Kaggle - Chronic-Kidney Disease Prediction. After Classifying the data, Preprocessin and performed the Exploratory Data Analysis. This data set contains about 339 records of data in various categories.
View details »
Metrics
Adaboost is one of the Boosting Algorithm.Kindly click below link to see the Metrics of Adaboost for CKD
View Metris »
Accuracy
AdaBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. In prediction problems involving unstructured data. After modelling our data, the Accuracy is 98%. So far we achieved a Good accuracy.
View details »
How it will work
This prediction will be used in healthcare Applications. As it was very important to predict weather the patient was having any chances of getting this Kidney Disease. This project comprises with the deployement too. we can deploy this project y means of


KIDNEY DISEASE PREDICTION
Home About
Graph Analysis
Accuracy
Training Accuracy
Testing Accuracy
Classification Report
Confusion Matrix

9. Conclusions and Future Work

The application of Machine Learning techniques for predictive analysis is very important in the health field because it gives us the power to detect chronic diseases earlier and therefore save people's lives through the anticipation of cure.

In this project, CKD prediction has been accomplished using the ensemble model from the CKD dataset as they can reduce the risk factors and improve the outcome in terms of efficiency and accuracy. We collected diagnostic data set with 26 CKD attributes of 400 patients for the study. Based on these attributes we applied basic machine learning algorithms ensembled method AdaBoost.

After the comparative analysis among all the models, it is evident that the ensembled model Ada Boost accuracy supersedes over other models with the accuracy of 98.33%.

9.1 Future Work

This study used a supervised machine-learning algorithm, feature selection methods. It is better to see the difference in performance results using unsupervised or deep learning algorithms models. In future we can use more number of datasets and other parameters which are affecting chronic disease. We can use deep learning approach for better result, we can add the health recommendation module as a future enhancement to the application where user can get the health recommendation based on their disease status or health status.

10. References

[1] M. Agrawal, N. Mohan and V. Jain, "Chronic Kidney Disease Prediction Using Random Forest, Decision Tree and Ada Boost Classifier," 2023 4th International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, 2023, pp. 1589-1593, doi: 10.1109/ICOSEC58147.2023.10276324.

[2] S. Samet, M. R. Laouar and I. Bendib, "Predicting and Staging Chronic Kidney Disease using Optimized Random Forest Algorithm," 2021 International Conference on Information Systems and Advanced Technologies (ICISAT), Tebessa, Algeria, 2021, pp. 1-8, doi: 10.1109/ICISAT54145.2021.9678441.

[3] DSVGK Kaladhar, Krishna Apparao Rayavarapu and Varahalarao Vadlapudi, on ” used Statistical and Data Mining Aspects on Kidney Stones: A Systematic Review and Meta-analysis”, Open Access Scientific Reports, Volume 1 • Issue 12 • 2012.

[4] J.Van Eyck, J.Ramon, F.Guiza, G.Meyfroidt, M.Bruynooghe, G.Van den Berghe, K.U.Leuven, on ” used Data mining techniques for predicting acute kidney injury after elective cardiac surgery”, Springer, 2012.

[5] K.R.Lakshmi, Y.Nagesh and M.VeeraKrishna, on ” done performance on comparison of three data mining techniques for predicting kidney disease survivability”, International Journal of Advances in Engineering & Technology, Mar. 2014.

[6] Morteza Khavanin Zadeh, Mohammad Rezapour, and Mohammad Mehdi Sepehri, on ” used Data Mining for funding performance in Identifying the Risk Factors of Early Arteriovenous Fistula Failure in Hemodialysis Patients”, International journal of hospital research, Volume 2, Issue 1,2013, pp 49-54.

[7] Abeer Y. Al-Hyari, on ” chronic kidney disease prediction system using classifying data mining techniques”, library of university of Jordan, 2012. Manish Kumar, International Journal of Computer Science and Mobile Computing, Vol.5 Issue.2, February- 2016, pg. 24-33 © 2016, IJCSMC All Rights Reserved 31.

[8] Xudong Song, Zhanzhi Qiu, Jianwei Mu, on ” Study on Data Mining Technology and its Application for Renal Failure Hemodialysis Medical Field”, International Journal of Advancements in Computing Technology(IJACT) ,Volume4, Number3, February 2012.

[9] N. SRIRAAM, V. NATASHA and H. KAUR, on ” data mining approaches for kidney dialysis treatment”, journal of Mechanics in Medicine and Biology, Volume 06, Issue 02,June 2006.

[10] JicksySusan Jose, R.Sivakami, N. Uma Maheswari, R.Venkatesh, on ” An Efficient Diagnosis of Kidney Images using Association Rules”, International Journal of Computer Technology and Electronics Engineering (IJCTEE),Volume 2, Issue 2,april 2012.

[11] Konstantina Kourou et.al, “Machine learning applications in cancer prognosis and prediction” Computational and structural bi technology Journal, Elsevier.

[12] P.Swathi Baby, T. Panduranga Vital ,”Statistical Analysis and Predicting Kidney Diseasesusing Machine Learning Algorithms” International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-018, Vol. 4 Issue 07, July-2015,206-210.

[12] Unsupervised Learning of Disease Progression Models Xiang Wang.