**A Project Report on**

**Chronic Kidney Disease Prediction Using Ada Boost Classifier**

submitted in partial fulfillment for the award of

**Bachelor of Technology**

in

**Computer Science & Engineering**

By

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

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



Under the guidance of

**Prof. M M Meera Durga.**

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 & Engineering**



**CERTIFICATE**

This is to certify that the project report entitled **Chronic Kidney Disease Prediction Using Ada Boost 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 Signature of the HOD**

**Mr.M M Meera Durga Dr.M Rajesh Babu**

**Asst.Prof. Assoc.Prof. & Head**

**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, asst.Prof.,** Department of CSE, for his/her valuable and inspiring guidance, comments, suggestions and encouragement.

We extend our sincere thanks to **Dr.M Rajesh Babu**,Assoc. Prof. & 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

[Table of Contents iv](#_Toc164281662)

Table of Figures………………………………………………………………………………………………………….vi

[Abstract vii](#_Toc164281663)

[1. INTRODUCTION 1](#_Toc164281664)

[1.1 Problem Statement 2](#_Toc164281665)

[1.2 Existing System 2](#_Toc164281666)

[1.2.1 Limitations of Existing System 3](#_Toc164281667)

[1.3 Proposed System 3](#_Toc164281668)

[1.3.1 Background techniques 4](#_Toc164281669)

[1.3.2 Advantages of Proposed System 5](#_Toc164281670)

[1.3.3 Applications 5](#_Toc164281671)

[1.4 Objectives 5](#_Toc164281672)

[1.5 Scope of the Project 6](#_Toc164281673)

[2. REQUIREMENT SPECIFICATION 7](#_Toc164281674)

[2.1 Descriptions Of The Software Used 7](#_Toc164281675)

[2.2 Literature review done in connection with the work 9](#_Toc164281676)

[3. SOFTWARE REQUIREMENT SPECIFICATION 13](#_Toc164281677)

[3.1 Introduction 13](#_Toc164281678)

[3.2 Purpose 13](#_Toc164281679)

[3.2.1 Scope 13](#_Toc164281680)

[3.3 Software Architecture 14](#_Toc164281681)

[3.3.1 Python: 15](#_Toc164281682)

[3.3.2 Flask: 15](#_Toc164281683)

[3.3.3 NumPy: 16](#_Toc164281684)

[3.3.4 Pandas: 16](#_Toc164281685)

[3.3.5 Sci-kit Learn 16](#_Toc164281686)

[3.4 Feasibility Study 17](#_Toc164281687)

[3.4.1 Technical Feasibility 17](#_Toc164281688)

[3.4.2 Economic Feasibility 17](#_Toc164281689)

[3.4.3 Operational Feasibility 17](#_Toc164281690)

[3.4.4 Behavior Feasibility 18](#_Toc164281691)

[3.5 Overview 18](#_Toc164281692)

[3.6 General description 19](#_Toc164281693)

[3.6.1 Product Functions 19](#_Toc164281694)

[3.6.2 General constraints 19](#_Toc164281695)

[3.7 Specific Requirements 20](#_Toc164281696)

[3.7.1 Functional Requirements 20](#_Toc164281697)

[3.7.2 Product Functions 22](#_Toc164281698)

[3.8 System Requirements 22](#_Toc164281699)

[3.8.1 Hardware Requirements 22](#_Toc164281700)

[3.8.2 Software Requirements 22](#_Toc164281701)

[5. SYSTEM DESIGN 23](#_Toc164281702)

[5.1 Introduction 23](#_Toc164281703)

[5.2 Scope 23](#_Toc164281704)

[5.3 Data Flow Diagram 23](#_Toc164281705)

[5.4 Activity Diagram 24](#_Toc164281706)

[5.5 Sequence Diagram 26](#_Toc164281707)

[5.6 Use Case Diagram 28](#_Toc164281708)

[6. IMPLEMENTATION 29](#_Toc164281709)

[6.1 Introduction 29](#_Toc164281710)

[6.2 Machine Learning overview 29](#_Toc164281711)

[6.2.1 Challenges in Implementing Machine Learning 30](#_Toc164281712)

[6.3 Architecture 32](#_Toc164281713)

[6.4 Ada Boost Learning Algorithm 34](#_Toc164281714)

[6.4.1 Working of Ada Boost 34](#_Toc164281715)

[6.4.2 PseudoCode 37](#_Toc164281716)

[7. TESTING 38](#_Toc164281717)

[7.1 Introduction 38](#_Toc164281718)

[7.2 Objectives Of Testing 38](#_Toc164281719)

[7.3 Testing Methods 39](#_Toc164281720)

[7.3.1 White Box Testing 39](#_Toc164281721)

[7.3.2 Black Box Testing 39](#_Toc164281722)

[7.3.3 Unit Testing 40](#_Toc164281723)

[7.3.4 Integration Testing 40](#_Toc164281724)

[7.3.5 Output Testing 40](#_Toc164281725)

[7.4 Validation 41](#_Toc164281726)

[8. RESULTS 42](#_Toc164281727)

[8.1 Classification Report: 42](#_Toc164281728)

[9. CONCLUSION AND SCOPE FOR FUTURE WORK 45](#_Toc164281729)

[9.1 Conclusion 45](#_Toc164281730)

[9.2 Scope For Future work 45](#_Toc164281731)

[10. REFERENCES 46](#_Toc164281732)

**Table of Figures**

[**Figure 1.1.1 Architecture of Proposed system** 4](#_Toc164255474)

[**Figure 3.1: Software Architecture** 13](#_Toc164255475)

[**Figure 4.1: Level 0 DFD** 22](#_Toc164255476)

[**Figure 4.2: Level 1 DFD** 23](#_Toc164255477)

[**Figure 4.3: Symbols** 24](#_Toc164255478)

[**Figure 5.1:Architecture** 31](#_Toc164255479)

[**Figure 5.2:working of Ada Boost** 35](#_Toc164255480)

[**Figure 7.1:Clasification Report** 41](#_Toc164255481)

# Abstract

The large amount of data that are generated from healthcare are too complex to analyzed manually Legitimately analyzing clinical records approximately patients wellbeing expect the plausibility of an event of various infections. In expansion, Procuring data concerning specialists of that specific infection as per the prerequisite facilitates legitimate and proficient determination of disease. Nowadays doctors are more focused on corona disease by forgetting other diseases which are present. 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. By the use of Machine learning techniques data can be analysed 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 like Adaboost.

# INTRODUCTION

Healthcare industry generates terabytes of data every year. The medical documents maintained are a pool of information regarding patients. The task of extracting useful information or quality healthcare is tricky and important. By analysing these voluminous data, we can predict the occurrence of the disease and safe guard people. Thus, an intelligent system for disease prediction plays a major role in controlling the disease and maintaining the good health status for people by providing accurate and trustworthy disease risk prediction.

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.

Chronic Kidney disease is worldwide health disease with higher burden with regard to the wellbeing within the show circumstance. Chronic Kidney infection is characterized as a glomerular filtration rate(GFR)<60mL/min or Kidney harm or both for at slightest a period of 3 months. End-stage renal illness is completely connected with mortality. Chronic Kidney is recognized with research facility tests. Major downside of this disease is, most of the time CKD is recognized at its last stage and which too leads to kidney failure. Within the early stages of chronic kidney illness, there will be few signs or side effects. CKD may not ended up clear until kidney work is altogether disabled.

Chronic kidney malady can be advance to conclusion organize of kidney failure, which is fatal without dialysis or a kidney transplantation. CKD is complicated illness by influencing the parts of the body by causing anemia, cardiovascular disease, Decreased Immune system, harm to central nervous system. It is exceptionally critical to urge check-up patient within short period of time.

## 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

## Existing System

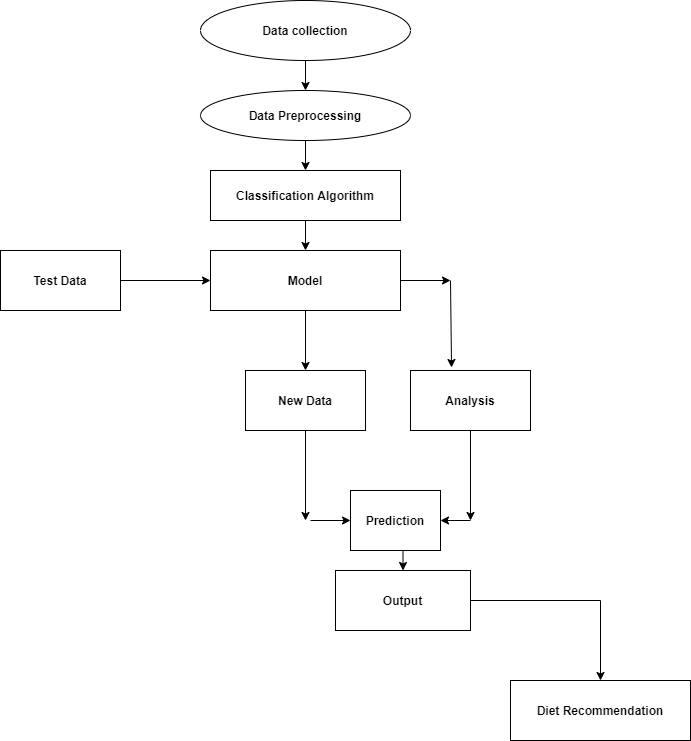
The Health care need more support for its development and developing countries like India. Previously prediction of CKD was done by checking and testing the various attributes which is depended on the kidney disease. Result of the test will test more time and some hospitals will not give proper test report because of amount. Patients are getting high risk for their treatment in critical condition .Another method is to check whether the patient had high blood pressure, a history of cardiovascular disease, the patient’s relative who have kidney disease and dialysis of kidney takes more time to analysis the working of kidney. Suppose we are available data set which contains the attributes which affects the CKD and that data is collected and build the model which helps patients that predicts the CKD.

### Limitations of Existing System

* + No automation for CKD prediction.
  + No proper medication in emergency.
  + Requires more time for the test report.
  + Understanding the test report is difficult for peoples.

## 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) is predicted 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 1.1.1 Architecture of Proposed system**

As shown in fig 1.1 Architecture of Proposed system for CKD prediction .Firstly,the data set which fits the model which uses Ada Boost learning algorithm, then model will be able to provide CKD Prediction which will be beneficial for patients to make better decisions on their health.

### Background techniques

**Ensemble Techniques:** Ensemble algorithm is supervised learning algorithms.The Ensemble algorithm solves the problem, by using tree representation.

Every internal node of the tree corresponds to each leaf node corresponds to a class label and attribute.

**Ada Boost:**

The Adapative Boosting algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems.

### Advantages of Proposed System

* + - Useful to health department to predict the CKD.
    - Useful for the patients to take better recovery.
    - We use data science techniques for accurate results.
    - On click of button output will be generated, no too much time required for CKD prediction.

▪ No need to analyze manually.

### Applications

* + - Proposed system can be used in medical department for the prediction of CKD .
    - Proposed system can be used by patients to know the if the CKD is present or not by inputting data such as “age”, “blood pressure”, “serum cretanine”, “sugar” and “bacteria” etc.

## Objectives

* + - System is an health care application which is an efficient tool for

disease prediction.

* + - System is an real time application which is meant for physician

and peoples.

* + - System is an automation for chronic kidney disease prediction. System makes use of “Machine learning” algorithm CKD prediction.
    - System predicts CKD based on the attributes such as age, sugar, serum creatinine, hyper tension and some others.

## Scope of the Project

* System is an health care application which is an efficient tool for disease prediction.
* System is an real time application which is meant for physician and peoples.
* System is an automation for chronic kidney disease prediction.

# REQUIREMENT SPECIFICATION

## Descriptions Of The Software Used

**Reasons for choosing python**

**Limitations of C:**

* + - C developers are forced to contend with manual memory management.
    - Ugly pointer arithmetic.
    - C is structured programming language.
    - Programmers require complete knowledge of best programming technique.

**Limitations of C++:**

* + - C++ can be thought as an Object Oriented layer on top of C.
    - It involves manual memory management.
    - Ugly pointer arithmetic.
    - Ugly syntactical constructs.

**Limitations of JAVA/J2EE:**

* + - Java programmers must use java front to back during development cycle.
    - It is not appropriate for many graphical or numerical intensive applications.
    - .NET provides solution to all the above-mentioned problems.

**Limitations of SQL:**

* + - SQL is most commonly used database.
    - It has a lot of capabilities(ex. For loop and functions)
    - Data warehouse function for decision support and integration .

## Literature review done in connection with the work

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. ROC and Calibration curves using Naive Bayes has also been constructed for predicting accuracy of the data. 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.Leuven et al. Explored data mining techniques for predicting acute kidney injury after elective cardiac surgery with Gaussian process & machine learning techniques (classification task & regression task).
* 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 dialysis survivability. 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. 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. The proposed NN algorithm as well as the other data mining algorithms demonstrated high potential in successful kidney 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.
* N. SRIRAAM, V. NATASHA and H. KAUR et al .presented data mining approach for parametric evaluation to improve the treatment of kidney dialysis patient. 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 approach helps 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. Their approach is divided into four major steps: pre-processing, feature extraction and selection, association rule generation, and generation of diagnosis suggestions from classifier.
* Divya Jain et al presented effect of diabetes on kidney using C4.5 algorithm with Tanagra tool. The performance of classifier is evaluated in terms of recall, precision and error rate.
* Koushal Kumar and Abhishek et al .their research describes comparison of all three neural networks such as (MLP, LVQ, RBF) on the basis of its accuracy, time taken to build model, and training data set size.
* In 2015, Konstantina Kourou et.al [1] 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.
* In 2015, P.Swathi Baby et. al [2] 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. Their observation shows that the best algorithms K-Star and Random Forest for the used Dataset ,where Build the models are less time(0 sec and 0.6 sec) and the ROC values are 1.
* In 2015, Konstantina Kourou et.al [1] 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. In 2015 P.Swathi Baby et. al [2] 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. 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. Their observation shows that the best algorithms K-Star and Random Forest for the used Dataset ,where Build the models are less time(0 sec and 0.6 sec) and the ROC values .

# SOFTWARE REQUIREMENT SPECIFICATION

## Introduction

The presentation of the Software Requirements Specification (SRS) gives a review of the whole SRS with reason, scope, definitions, abbreviations, contractions, references and diagram of 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.

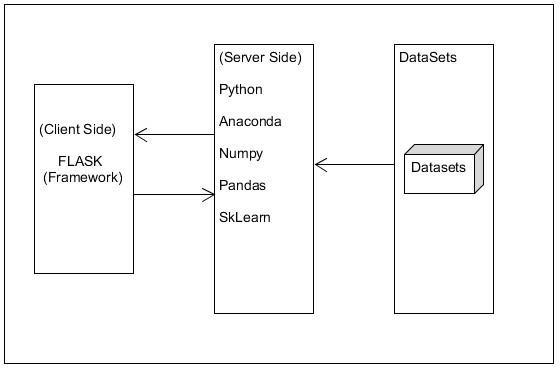
## Purpose

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 and its UI, equipment and programming prerequisites. It characterizes how our customer, group and crowd see the item and its usefulness.

### Scope

The Scope of this framework is to presents a survey on information digging strategies utilized for the expectation of Chronic disease prediction. It is obvious from the framework that information mining strategy, similar to grouping, is profoundly productive in expectation of Indian car.

## Software Architecture



**Figure 3.3.1: Software Architecture**

### 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.

### 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.

### NumPy:

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

* a amazing N-dimensional cluster object
* Sophisticated (broadcasting) capacities
* tools for incorporating C/C++ and Fortran code.
* useful straight polynomial math, Fourier change, and arbitrary number abilities.

### 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 of improvement of pandas as an a-list open-source venture, and makes it conceivable to give to the task.

### 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.

## Feasibility Study

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

### 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 coordinates between 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.

### 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 estimated benefits are equal.

### Operational Feasibility

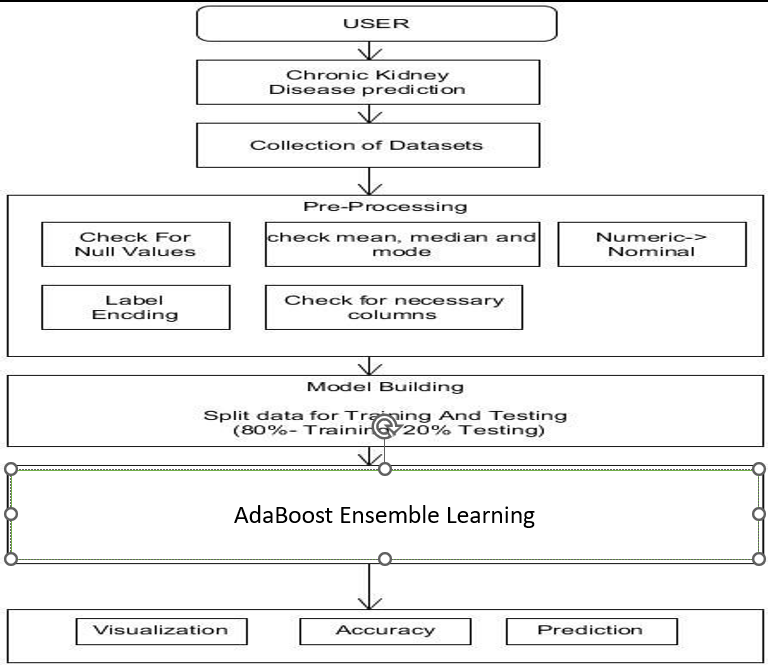
The project entitled “Prediction of Chronic disease” is technically feasible because of the belowmentioned features. The system predicts the chronic disease prediction based on the historical data, further the details of the patient are added to the Data Base. The performance of the Data mining techniques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

### Behavior Feasibility

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

## 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 will enclose functional requirements and non-functional requirements.



**Figure 3.3: Overview**

## General description

### Product Functions

* Collected datasets of chronic disease prediction from Kaggle
* Pre-processing of obtained datasets
* Select Attributes which helps in predicting the stock
* The selected datasets are trained using AdaBoost
* The trained data sets are tested for Accuracy
* The obtained result is showed in the graph
* The results are compared with the machine learning

### General constraints

* The system should have enough RAM and Disk Storage Space.
* The Source code must be written in Python for ML.
* The results generated have to be entered in to the system and any error or any value entered out of the boundary will not be understood by the system.

## Specific Requirements

### Functional Requirements

A functional requirement defines a function of a system or its component. A role is described as 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

* **Input**: Chronic dataset.
* **Process**: Preprocessing will find missing value and also does feature remove.
* **Output**: preprocessed dataset.
* **Error handling:** If the input file is not a valid one **.**
* **Input**: preprocessed dataset.
* **Process**: It will select only important data which is required.
* **Output**: Selected data will be displayed **Splitting of the Data:** Training data and Test Data.
* **Input**: Feature selected data.
* **Process**: It will split the data into the train set and test set.
* **Output**: Dataset will be displayed as Train set and Test set and it will be tested for the specific algorithms and performance analysis will be carried out.



**Data Collection**

**Data Preprocessing**

**Ada Boost Classifier**

**( AdaBoost)**

**Comparison Study**

**Prediction**

**Analysis**

**New Data**

**Model**

**Test data**

Figure 3.3: Functional requirements

**The model that functions based on :**

* + Collecting Data The data is collected from previous chronic disease records in Kaggle datasets.
  + Then pre-processing the data pre-processing is adding the data.
  + Performing data mining algorithms The data mining algorithms include (AdaBoost).
  + The algorithm helps in predicting the result based on the parameters.
  + The analysis help in the prediction of Disease.

### Product Functions

* Uploading Data Uploading data sets.
* Perform Prediction is done by each algorithm based on the constraints.
* Comparison Study Prediction results and its stages of each algorithm is represented through graph.

## System Requirements

### Hardware Requirements

|  |  |  |
| --- | --- | --- |
|  | * PROCESSOR | : Intel i3 |
|  | * HARD-DISK | :500GB |
|  | * RAM | :4GB or Above |

### Software Requirements

* + - * OPERATING SYSTEM : Windows 7 and above
      * FRONT END : Html, CSS
      * FRAMEWORK : Flask
      * LANGUAGE : Python version 3.7
      * LIBRARIES :Pandas,Numpy,Sklearn,Scikit.
      * EDITOR : Visual Studio Code

# SYSTEM DESIGN

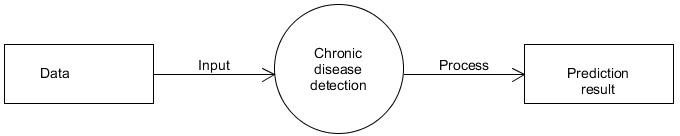
## Introduction

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 project includes use case models, sequence diagrams, and other supporting requirement information.

## Scope

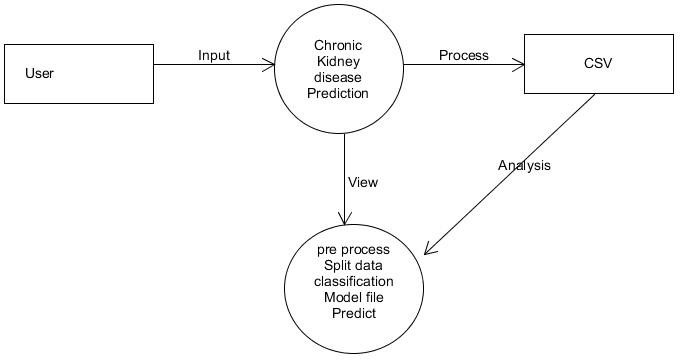
The design Document is for a primary level system, which will work as a basement for building a system that provides a base level of functionality to show feasibility for large-scale production use. The software Design Document, the focus placed on the generation and modification of the documents. The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

## Data Flow Diagram

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

**Figure 4.1: Level 0 DFD**

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



**Figure 4.2: Level 1 DFD**

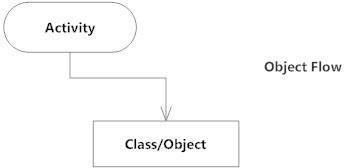
## Activity Diagram

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 measure demonstrating. They can likewise depict the means in an utilization case chart. Exercises demonstrated can be consecutive and simultaneous. In the two cases, an action outline will have a start (an underlying state) and an end (a last state).

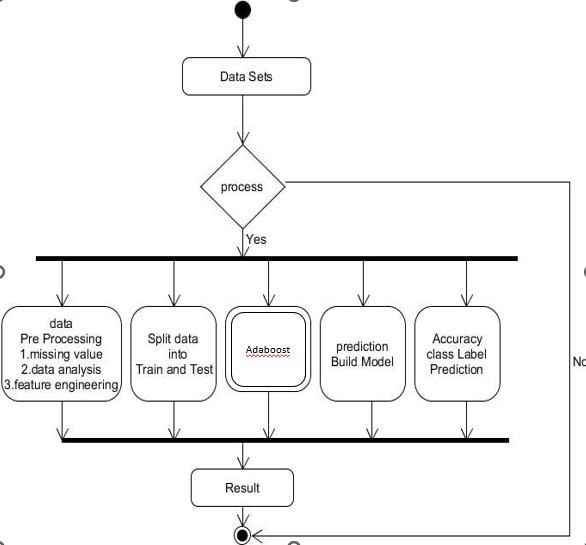








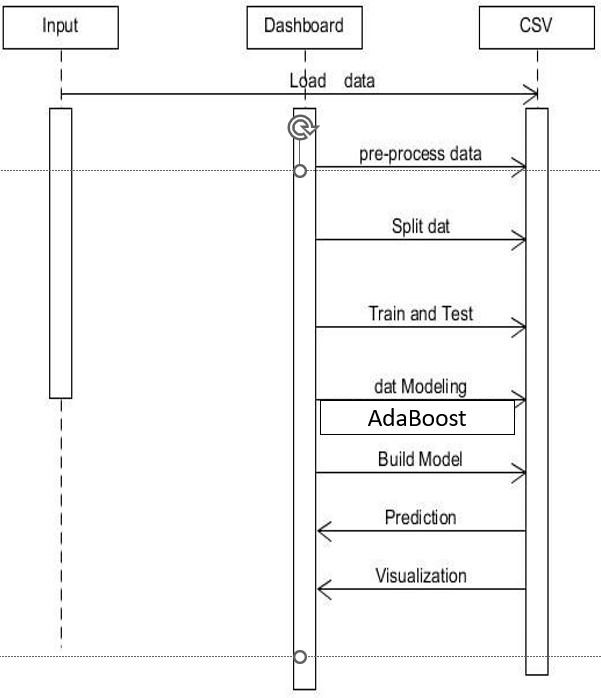
**Figures 4.3: Symbols**



***Figure 4.4 : Activity Diagram***

## 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 envision and 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 another framework.



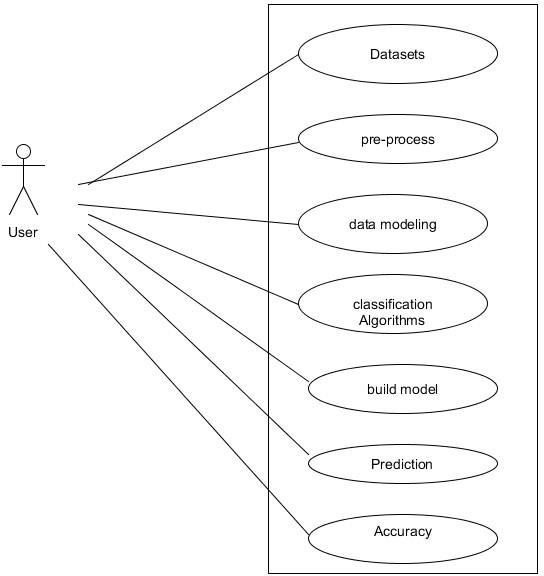


***Figure 4.5 : Sequential Diagram***

## 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 four outlines (action, grouping, cooperation, and Statechart) likewise have a similar reason. We will investigate 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 a framework is investigated to accumulate its functionalities, use cases are readied and entertainers are distinguished.



**Figure 4.5:Use Case Diagram**

# IMPLEMENTATION

## Introduction

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 can be used as a template for creating copies of such module on demand.

This project is implemented using python programming language. Python is [dynamically typed](https://en.wikipedia.org/wiki/Dynamic_programming_language) and [garbage-collected.](https://en.wikipedia.org/wiki/Garbage_collection_(computer_science)) It supports multiple [programming paradigms,](https://en.wikipedia.org/wiki/Programming_paradigms) including [procedural,](https://en.wikipedia.org/wiki/Procedural_programming) object-oriented, and [functional programming.](https://en.wikipedia.org/wiki/Functional_programming)

## 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. In this article, we’ll see basics of Machine Learning, and implementation of a simple machine learning algorithm using python.

Machine learning involves a computer to be trained using a given data set, and use this training to predict 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, the computer 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 is used gives to predict CKD.

### 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.

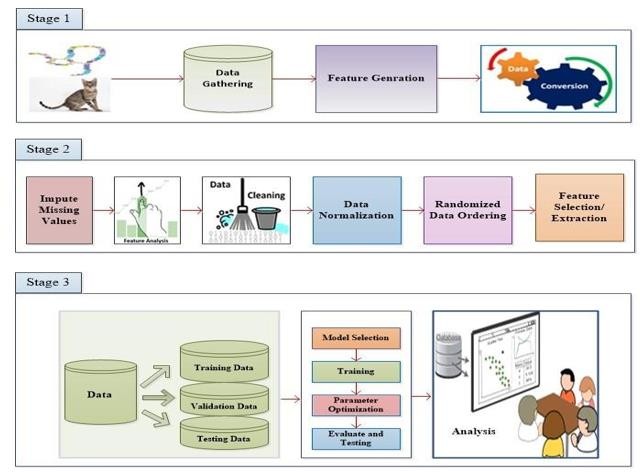
**Training requirements** **:**AI-powered intellectual systems must be trained in a domain, e.g., claims or billing for an insurer. This requires a separate training system, which insurers find hard to provide for training the AI model. Models need to be trained with huge volumes of documents/transactions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

**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. This is important to train predictive models. Generally, insurers struggle to provide relevant data for training AI models.

**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.

**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.

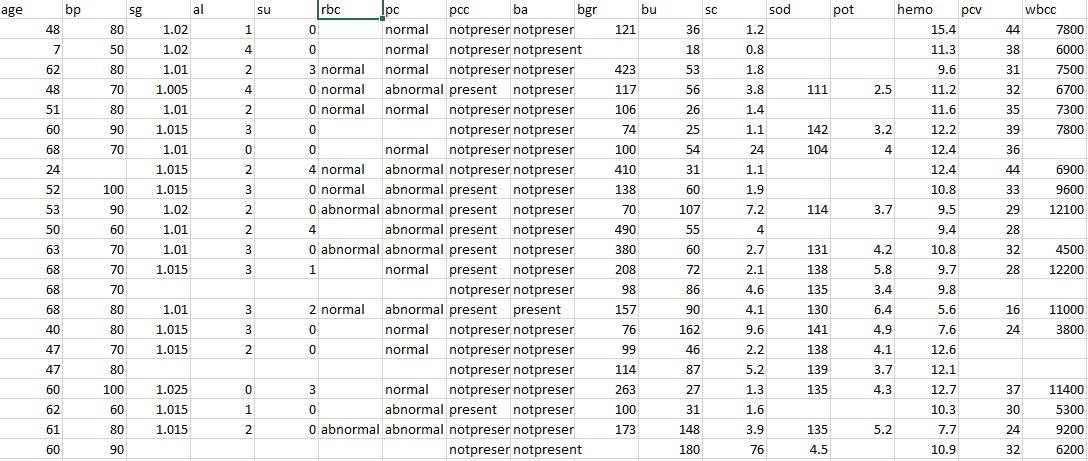
## Architecture



**Figure 5.1:Architecture**

###### **Stage1:**

We collected dataset from Kaggle website that it has 25 features and 1 class label for every chronic kidney disease record, and the features include basic etc age, bp, sugar, serum creatine, sodium, haemoglobin etc.



###### **Stage2:**

**Data Cleaning:**

The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, analysis of data, feature engineering,noisy data etc.

###### **Missing Data:**

This situation arises when some data is missing in the data. It can be handled in various ways.Some of them are:

###### **Ignore the tuples:**

This approach is suitable only when the dataset we have is quite large and multiple values are missing within tuple.

###### **Fill the Missing values:**

There are various ways to do this task. You can choose to fill the missing values manually, by attribute mean or the most probable value

###### **Stage 3:**

The obtained data from stage is taken into consideration then data is trained using the classification algorithm and obtained result is analysed and Showed in the graph using python library.

The obtained data is trained using AdaBoost Learning Algorithm:

## 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.

### 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

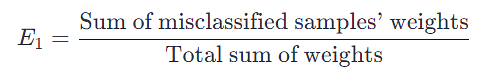
**Step2 – Train weak classifiers**

• Train a weak classifier Mk where k is the current iteration

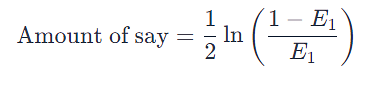
• Calculate the weighted error E1 of the first model:

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

• Calculate the weighted error E1 of the first model:



• Calculate the importance of each model α\_k using formula :



**Step4 – Update data point weight for each data point Wi**

* Correctly classified samples' weights are decreased:



* Misclassified samples' weights are increased:



**Step5 – Normalize the Instance weight**

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

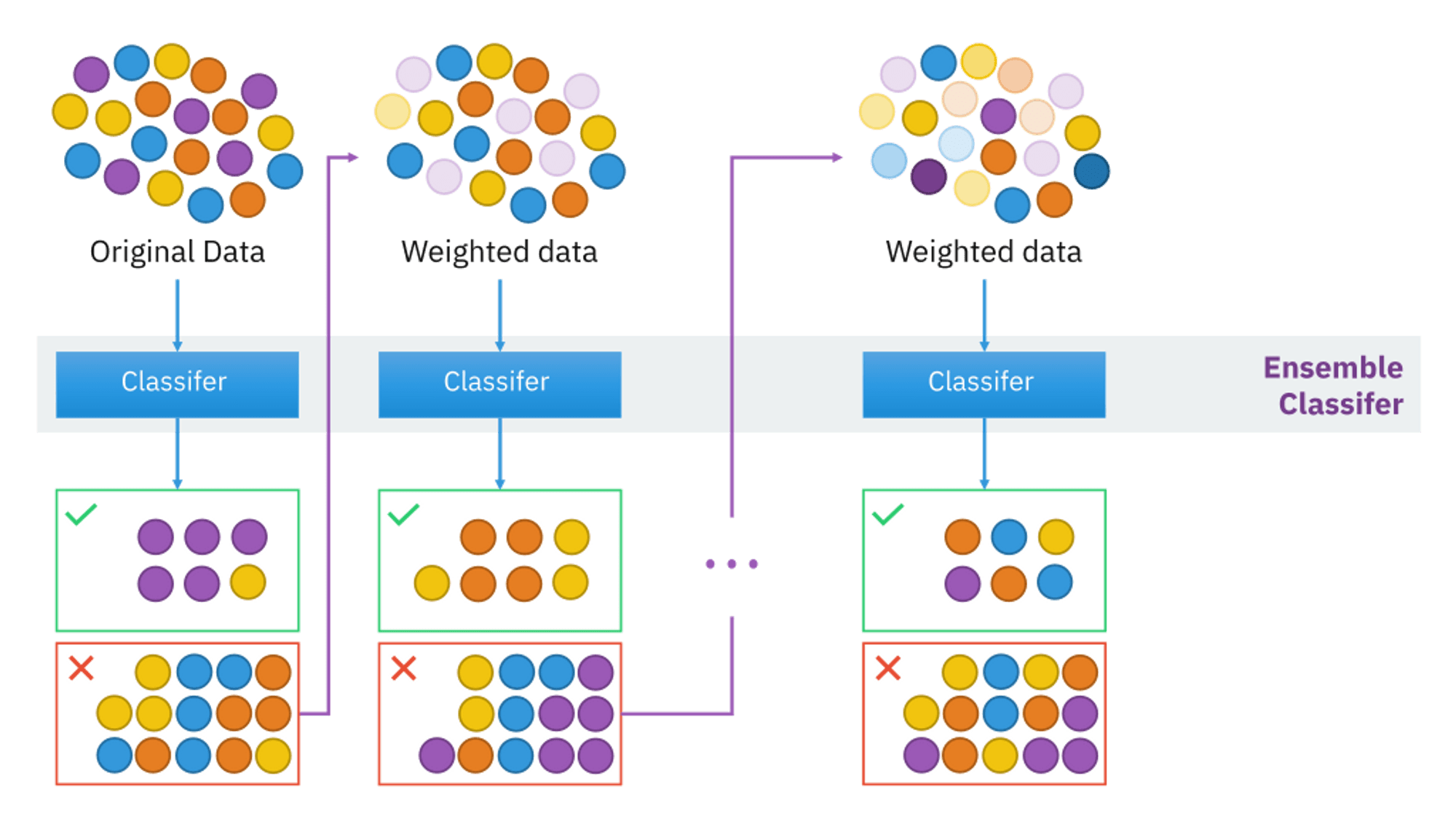
A math equation with numbers and symbols

Description automatically generated

**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 5.2:Working of Ada Boost**

### PseudoCode

Set equal weights for each observation (w=1/n)

for t=1:T:

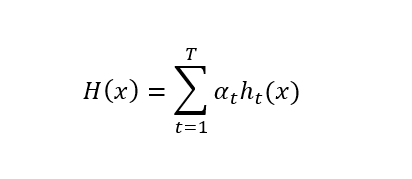
Train a weak classifier

Choose coefficient alpha (importance)

Update weights

End for

Output:



# TESTING

## Introduction

Testing is the way toward running a framework with the expectation of discovering blunders. Testing upgrades 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.

he primary intention is to distinguish blunders and mistake get-prom zones in a framework. 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 by their 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.

## Objectives Of Testing

* Testing in a cycle of executing a program with the expectation of

discovering mistakes.

* A 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 testing makes 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.

## 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 acceptance testing.

### 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 -

* + - Exercise all logical decisions on their true or false side.
    - Execute all loops at their boundaries

### Black Box Testing

It is focused around the practical necessities of the product. It's anything but a choice to white box testing; 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.

* + - * Incorrect or missing capacities
      * Interface blunders
      * Errors in an information structure or outside information base access

### 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. In this progression, every module is discovered to be working acceptably as respects the normal yield from the module.

### Integration Testing

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

### Output Testing

Subsequent to performing approval testing, the following stage is yield trying of the proposed framework, 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.

## Validation

Toward the consummation of the reconciliation testing, the product is totally amassed as bundle interfacing blunders have been revealed and adjusted and a last arrangement of programming tests starts in approval testing. Approval testing can be characterized from multiple points of view, however a straightforward definition is that the approval succeeds

when the product work in a way that is normal by the client. After approval test has been directed as follows:

* The capacity or execution qualities adjust to detail and are acknowledged.
* A deviation from the particular is revealed and a lack list is made.
* Proposed framework viable has been tried by utilizing an approval test and discovered to be working acceptably.

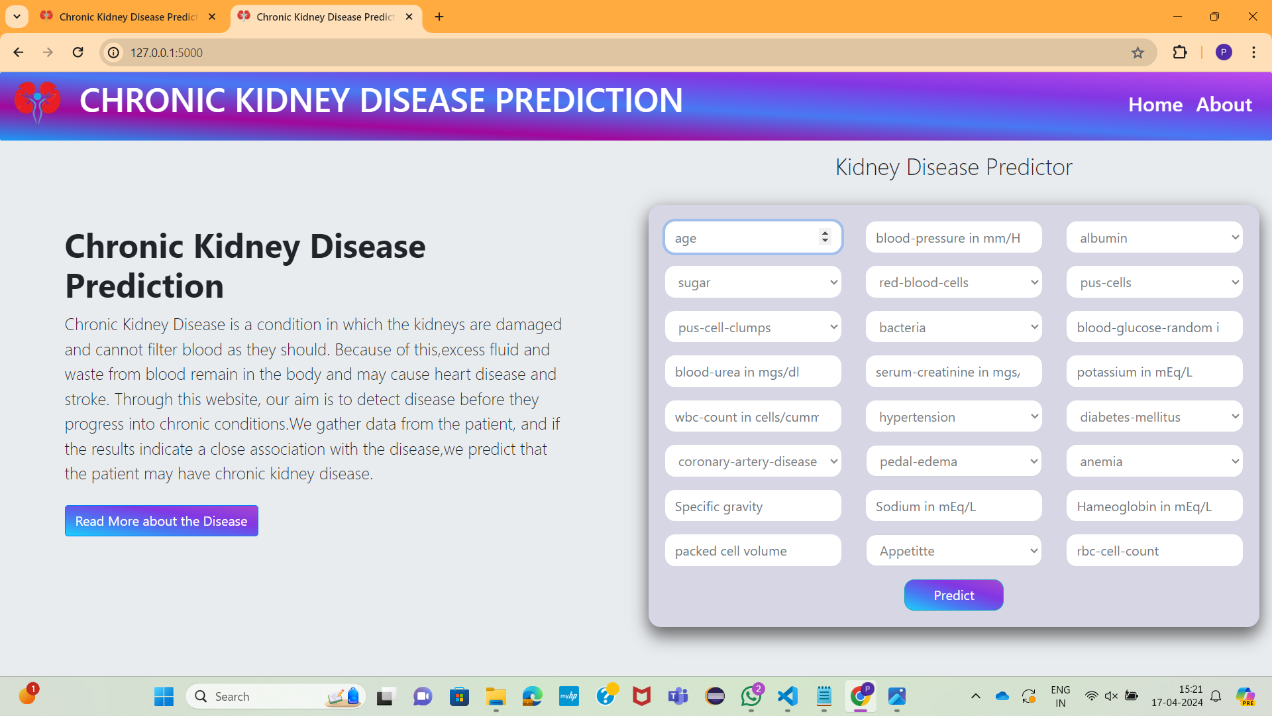
# RESULTS

## Classification Report:

A screenshot of a computer

Description automatically generated

**Figure 7.1:Clasification Report**

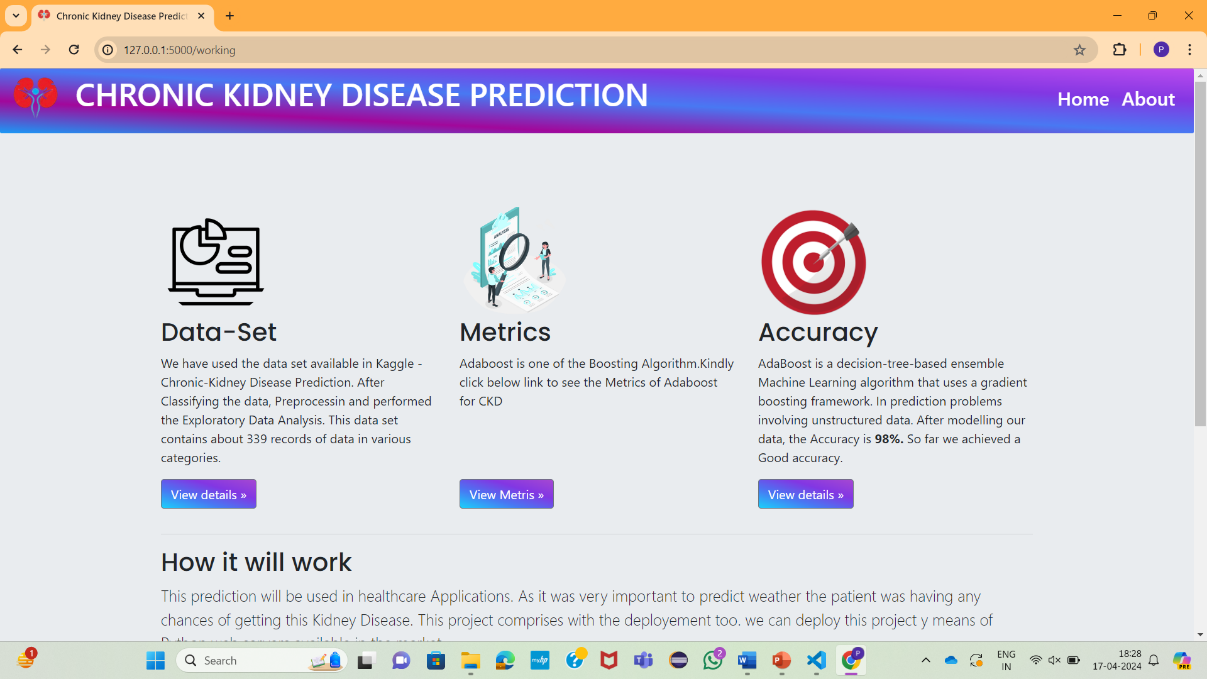
Finally Concluding that Ada Boost Algorithm achieved 98% Accuracy here some of the metrics for the Ada Boost.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated



A screenshot of a computer

Description automatically generated

# CONCLUSION AND SCOPE FOR FUTURE WORK

## Conclusion

The application of data mining techniques for predictive analysis is very important in the health field because it gives us the power to 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%.

## Scope For Future work

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.

# REFERENCES

1. Konstantina Kourou et.al, “Machine learning applications in cancer prognosis and prediction” Computational and structural bitechnology Journal, Elsevier.
2. P.Swathi Baby, T. Panduranga Vital ,”Statistical Analysis and Predicting Kidney Diseases using Machine Learning Algorithms” International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-018, Vol. 4 Issue 07, July-2015,206-210.
3. K.R.Lakshmi1, Y.Nagesh2 and M.VeeraKrishna3, “Performance Comparision of Three Data Mining Techniques for Predicting Kidney Dialysis Survivability”, International Journal of Advances in Engineering & Technology, Mar. 2014, Vol. 7, Issue 1, pp. 242-254.
4. AndrewKusiak, Bradley Dixonb, Shital Shaha, (2005) Predicting survival time for kidney dialysis patients: a data mining approach, Elsevier Publication, Computers in Biology and Medicine 35, page no 311–327.
5. Dr. S. Vijayarani, Mr.S.Dhayanand, “Kidney Disease Prediction Using SVM and ANN Algorithms” IJCBR , ISSN (online): 2229- 6166,Volume 6 Issue 2 March 2015.
6. Prediction of kidney disease stages using data mining algorithms El-Houssainy A. Radya , Ayman S. Anwarb.
7. DATA MINING CLASSIFICATION ALGORITHMS FOR KIDNEY DISEASE PREDICTION -Dr. S. Vijayarani1 , Mr.S.Dhayanand2 .
8. Deep EHR: Chronic Disease Prediction Using Medical Notes Jingshu Liu Zachariah Zhang.



1. Comparative Study of Chronic Kidney Disease Prediction using KNN and SVM Parul Sinha. **[**10] Disease Prediction and Doctor Recommendation SystemDhanashri Gujar1, Rashmi Biyani2, Tejaswini Bramhane3[6]Chronic Diseases Diagnosis using Machine Learning, Shweta Ganiger.
2. A Literature Review on Kidney Disease Prediction using Data Mining Classification Technique Suman Bala1 , Krishan Kuma.

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