# Early Detection of Chronic Kidney Disease Using Machine Learning

## INTRODUCTION

### PROJECT OVERVIEW:

Chronic kidney disease (cdk) has received much attention due to its high mortality rate. Chronic disease have become a concern threatening developing countries, according to the World Health Organization(WHO).CDK is a kidney disorder treatable in its early stages,but it causes kidney failure in its late stages.In 2016, chronic kidney disease caused the death of 753 million people worldwide,where the number of males died was 336 million, while the number of females died was 417 million.It is called "chronic" disease because the kidney disease begings gradually and lasts for a long time, which affets the functioning of the urinary system.The accumulation of waste products in the blood leads to the emergence of other health problems,which are associated with several symptoms such high and low blood pressure, and cardiovascular diseases,nerve damage, and bone problems, which lead to cardiovascular disease. Risk factors for CKD patients include diabetes,blood pressure ,and cardiovascular disease(CVD).CKD patients suffers from side effects,especially in the late stages, which damage the nervous and immune system. In developing countries,patients may reach the late stages, so they must undergo dialysis or kidney transplantation.Medical expects determine kidney disease through glomerular filtration rate(GFR),which describes kidney function. GFR is based on information such as age, blood test, gender, and other factors suffered by the patient. Regarding the GFR value, doctors can classify CKD into five stages.

Early diagnosis and treatment of chronic kidney disease will prevent its progression to kidney failure. The best way to treat chronic kidney disease is to diagnose it in the early stages, but discovering it in its late stages will lead to kidney failure, which requires continuous dialysis or kidney transplantation to maintain a normal life. In the medical diagnosis of chronic kidney disease, two medical tests are used to detect CKD, which are by a blood test to check the glomerular filtrate or by a urine test to check albumin. Due to the increasing number of chronic kidney patients, the scarcity of specialist physicians, and the high costs of diagnosis and treatment, especially in developing countries, there is a need for computer-assisted diagnostics to help physicians and radiologists in supporting their diagnostic decisions.

Artificial intelligence techniques have played a role in the health sector and medical image processing, where machine learning and deep learning techniques have been applied in the processes of disease prediction and disease diagnosis in the early stages.Artificial intelligence approaches have played a basic role in the early diagnosis of ckd.Machine learning algorithms are used for the early diagnosis of ckd.The ANN and SVM algorithms are among the most widely used technolpgies.These technologies have great advantages in diagnosing several fields,including medical diagnosis.The ANN algorithm works like human neurons,which can learn how to operate once properly trained,and its ability to generalize and solve future problems (test data).However,SVM algorithm depends on experience and examples to assign labels to the class.SVM algorithm basically separates the data by a line that achieves the maximum distance between the class data.Many factors affect kidney performance,which induce ckd,like diabetes,blood pressure,heart disease,food and family history.

**PURPOSE**:

Chronic Kidney Disease is very common, but less than 1 in 10 people with CKD ever require dialysis (artificial kidney treatment) or a kidney transplant.Someone with CKD is at increased risk of heart attack or stroke, especially if they smoke or are overweight.People with CKD should have regular checks of their kidney function and blood pressure, and have treatment if their blood pressure is elevated.

LITERATURE SURVEY

EXISTING PROBLEM:

In this the previous prediction of the persistent kidney illness are to tell the accurate values to use some of the algorithms .Early acknowledgement and brief usage of prescribed administration rules are necessary to forestall intensifying kidney work and cardiovascular horribleness in patients.Acknowledgement could be in a roundabout way surveyed by non appearance of ckd.

## 2.3 PROBLEM STATEMENT:

This paper provides a problem statement on the various techniques that have been used to forecast Early Detection Of Chronic Kidney Disease Using Machine Learning. We mainly focused on the researches to provide a prediction algorithm to predict Chronic Kidney Disorders at an early stage.

## 3. IDEATION & PROPOSED SOLUTION

### 3.1 EMPATHY MAP CANVAS:

Data science is capable of collecting the data from the patients such as blood grp age etc. Education around the risk factors is one of the way to improve ckd progression Collecting data in healthcare is not enough the data should be accurate and precise .One of the negative thing is the lack of privacy when comes medical records .The gain through the process is to provide the treatment on time.

### 3.2 IDEATION & BRAINSTORMING:

The advances in sensor networks.communication technologies,data science and statistical processing have rendered ML techniques as important tools in various health-oriented applications.
CDK can be predicted using data pre-processing exploring data analysis.
Kidney disease detection model is proposed to allow early detection of disease through appropriate use of big data for machine learning.Few diadnotic test to check CKD are estimated glomerular filtration rate, Urine test, blood pressure.

### 3.3 PROPOSED SOLUTION:

To provide a prediction algorithm to predict Chronic Kidney Disorders at an early stage. The dataset shows input parameters collected from the CKD patients and the models are trained and validated for the given input parameters.

### 3.4 PROBLEM SOLUTION FIT:

People are required to answer the questions and share the required medical data .Ckd is an asymptomatic disease which is very difficult to diagnose in early stage .Regular testing is recommended for the people with high blood pressure and diabetes.It provides a prediction algorithm to predict the ckd the dataset shows input parameters collected from ckd. To execute this we need to develop a web application by using the data given by the patients.

## 4.REQUIREMENT ANALYSIS

### 4.1 FUNCTIONAL REQUIREMENT:

Functional Requirement defines a function of a software system and how the system must behave when presented with specific inputs or conditions. These may include calculations, datamanipulation and processing and other specific functionality. In this system following are the
functional requirements

1. All the data must be in the same format as a structured data.
2. The data collected will be vectorized and sent across to the classifier.

## 4.2 NON- FUNCTIONAL REQUIREMENTS:

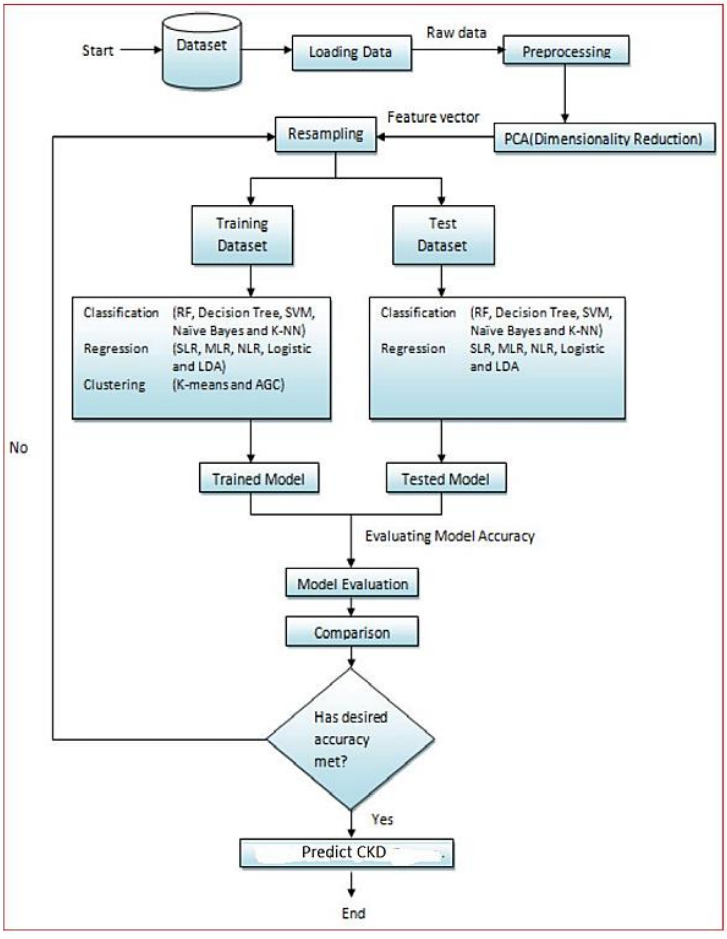
Non functional requirements are the requirements which are not directly concerned with the specific function delivered by the system. They specify the criteria that can be used to judge the operation of a system rather than specific behaviors. They may relate to emergent system
properties such as reliability, response time and store occupancy. Non functional requirements arise through the user needs, because of budget constraints, organizational policies, the need for interoperability with other.

Software and hardware systems or because of external factors such as:-

- Product Requirements
- Organizational Requirements
- User Requirements
- Basic Operational Requirements

## 5.PROJECT DESIGN

## 5.1 DATA FLOW DIAGRAM:



## 5.2 SOLUTION & TECHNICAL ARCHITECTURE:

### SOLUTION ARCHITECTURE:

It is a complex process with many sub processes that bridges the gap between business problem and technology solutions.Decision tree,random forest, support vector machine learning models are constructed to carry out the diagnosis of ckd .The performance of the models are evaluated based on the accuracy of prediction .The models are trained and validated for the given input parameters.

### TECHNICAL ARCHITECTURE:

There are many ways to detect chronic kidney disease using machine learning, but early detection is essential to treatment and management of the disease. One way to detect chronic kidney disease early is to use machine learning algorithms to identify risk factors for the disease. Some risk factors for chronic kidney disease include high blood pressure, diabetes, and a family history of the disease. By identifying these risk factors early, it is possible to prevent or delay the onset of chronic kidney disease.

## 5.3 USER STORIES:

**User Stories**

| User Type | Functional requirement | User story number | User story/task | Acceptance criteria | Priority | Release |
|---|---|---|---|---|---|---|
| Customer (Mobile user, Web user, Care executive, Administrator) | Registration | USN-1 | As a user, I can register for the application by entering my mail, password, and confirming my password | I can access my account/ dashboard | High | Sprint-1 |
| | | USN-2 | As a user, I will receive confirmation email once I have registered for the application | I can receive confirmation email & click confirm | High | Sprint-1 |
| | Dashboard | USN-3 | As a user, I can register for the application through internet | I can register & Access the dashboard with Internet login | Low | Sprint-2 |
| | | USN-4 | As a user, I can register for the application through Gmail | I can confirm the registration in Gmail | Medium | Sprint-1 |
| | Login | USN-5 | As a user, I can log into the application by entering email & password | I can login with my id and password | High | Sprint-1 |

## 6.PROJECT PLANNING & SCHEDULING

### 6.1 SPRINT PLANNING & ESTIMATION:

1. Collect the dataset
2. Clean the dataset

IMPORTING THE LIBRARIES:

1. Pandas
2. NumPy
3. Counter
4. Matplotlib and seaborn
5. Missigno
6. Accuracy score
7. Confusion matrix
8. Train_test_split
9. LabelEncoding
10. Pickle

3. Read the dataset:

You might have your data in .csv files , excel files or .tsv files or something else. But the goal is the same in all cases. If you want to analyze that data using pandas, the first step will be to read it into a data structure that's compatible with pandas.

Let's load a .csv data file into pandas. There is a function for it, called read_csv(). We will need to locate the directory of the CSV file at first (it's more efficient to keep the dataset in the same directory as your program).

4. Understanding data type and summary of features:

How the information is stored in a DataFrame or Python object affects what we can do with it and the outputs of calculations as well. There are two main types of data that are

- Numeric and text data types.
- Numeric data types include integers and floats.
- The text data type is known as Strings in Python, or Objects in Pandas. Strings can contain numbers or characters. For example, a string might be a word, a sentence, or several sentences.

5. Handling the missing values:

Sometimes you may find some data are missing in the dataset. We need to be equipped to handle the problem when we come across them. Obviously, you could remove the entire line of data but what if you are unknowingly removing crucial information.Of course, we would not want to do that. One of the most common ideas to handle the problem is to take a mean of all the values for continuous and for categorical we make use of mode values and replace the missing data.

6.Replacing the missing values:

We can replace the missing values by mean, median or mode by using fillna method.

Check the count of null values after filling all null values using isnull.sum() you should find count as zero at every column.

7.Lable Encoding:

Typically, any structured dataset includes multiple columns with combination of numerical as well as categorical variables. A machine can only understand the numbers. It cannot understand the text. That's essentially the case with Machine Learning algorithms too. We need to convert each text category to numbers in order for the machine to process those using mathematical equations.

8.Splitting the dataset into dependent and independent variable:

To help us with this task, the Scikit library provides a tool, called the Model Selection library. There is a class in the library which 'train_test_split.' Using this we can easily split the dataset into the training and the testing datasets in various proportions.The train-test split is a technique for evaluating the performance of a machine learning algorithm.

**Train Dataset:** Used to fit the machine learning model.

**Test Dataset:** Used to evaluate the fit machine learning model.

10.Model Building;

There are several Machine learning algorithms to be used depending on the data you are going to process such as images, sound, text, and numerical values. The algorithms that you can choose according to the objective that you might have it may be Classification algorithms or Regression algorithms.

Example: 1. Linear Regression.

2. Logistic Regression.
3. Random Forest Regression / Classification.
4. Decision Tree Regression / Classification.

You will need to train the datasets to run smoothly and see an incremental improvement in the prediction rate.

11.Test the model:

Once the model is trained, it's ready to make predictions. We can use the predict method on the model and pass x_test as a parameter to get the output as y_pred.Notice that the prediction output is an array of real numbers corresponding to the input array.

12.Model Evalution:

Finally, we need to check to see how well our model is performing on the test data. There are many evaluation techniques are there. For this, we evaluate the accuracy score produced by the model.

13.Save the model:

Save our model by importing pickle file. Pickle is used for serializing and de-serializing Python object structures, also called marshalling or flattening. Serialization refers to the process of converting an object in memory to a byte stream that can be stored on disk or sent over a network. Later on, this character stream can then be retrieved and de-serialized back to a Python object.

## 7.Coding & solutioning:

### 7.1 Testing:

In this phase all programs (models) are integrated and tested to ensure that the complete system meets the software requirements. The testing is concerned with verification and validation Black System Testing is a level of software testing where a complete and in- tegrated software is
tested. The purpose of this test is to evaluate the systems compliance with the specified requirements. System Testing is the testing of a complete and fully integrated software product. and White Box Testing. System test falls under the black box testing cate- gory of software testing.

## System Testing:

• Usability Testing - Usability Testing mainly focuses on the users ease to use the application,
flexibility in handling controls and ability of the system to meet its objectives.
• Load Testing - Load Testing is necessary to know that a software solution will perform under
real-life loads.
• Regression Testing- Regression Testing involves testing done to make sure none of the changes
made over the course of the development process have caused new bugs.
• Recovery Testing - Recovery testing is done to demonstrate a soft- ware solution is reliable,
trustworthy and can successfully recoup from possible crashes.
• Migration Testing - Migration testing is done to ensure that the software can be moved from
older system infrastructures to current system infrastructures without any issues.

### User acceptance test

The user should be able to input their data into the machine learning algorithm.
The machine learning algorithm should be able to accurately detect early signs of chronic kidney disease.
The user should be able to receive results from the machine learning algorithm that accurately reflect their risk for chronic kidney disease

## Advantages

Some potential advantages of early detection of chronic kidney disease using machine learning include improved patient outcomes and earlier interventions to prevent or delay the progression of the disease. Additionally, machine learning may be able to identify risk factors for chronic kidney disease that are not currently known or that are difficult to detect using traditional methods. This could lead to more targeted and effective treatments for the disease.it can allow for earlier and more effective treatment. If a patient is diagnosed with chronic kidney disease at an early stage, they may be able to receive treatment that can prevent or slow down the progression of the disease. This can improve the patient's prognosis and quality of life.

Some potential advantages of early detection of chronic kidney disease using machine learning include the ability to identify high-risk individuals earlier, potentially leading to earlier intervention and improved outcomes; the ability to more accurately predict which individuals are likely to develop chronic kidney disease; and the ability to monitor individual risk over time.

## Future scope

The early detection of chronic kidney disease (CKD) is a significant challenge for the medical community. While there are a number of risk factors for CKD, such as diabetes and hypertension, the disease can often be asymptomatic in its early stages. This makes early detection difficult, as patients may not present for medical attention until the disease has progressed to a more serious stage.

Machine learning may offer a potential solution for early detection of CKD. By analyzing a patient's medical history and other data, such as demographics and laboratory results, a machine learning model could identify those at risk for CKD and flag them for further testing and monitoring. While more research is needed to develop and validate such a model, the potential benefits of early CKD detection make it a promising area of exploration.

For the purposes of this project we have used five popular algorithms:
Logistic regression and Neural network, Decision Tree, Gaussian NB, KNN. All the algorithms are based on supervised learning. We are determining the best method considering 4 factors namely Specificity, Sensitivity, Log Loss, Accuracy. When plotted on a graph for all the algorithms it was found that Logistic Regression was the best method to use to find Chronic Kidney Disease.

## Accuracy

Accuracy is the number of correctly predicted data points out of all the data points. More formally, it is defined as the number of true positives and true negatives divided by the number of true positives,true negatives, false positives, and false negatives.

## Specificity

Specificity is defined as the proportion of actual negative, which got predicted as the negative(or true
negative)

## Sensitivity

Sensitivity is a measure of the proportion of actual positive cases that got predicted as positive (or true positive).

## CONCLUSION:

In conclusion, the use of data mining methods for data risk analysis is very important in the health sector because it first gives the power to fight diseases and therefore saves people's lives by reversing treatment. In this work, we used a number of learning algorithms to assess patients with chronic renal failure (ckd) and patients with this disorder . Simulation results have shown that SVM classification has proven its performance in predicting the best results in terms of accuracy and minimum execution time. The result of the testing does not achieve 100%. It may be due to calculation and weight. To get a more accurate result, the calculation of the similarity has to modify with the weight. The experts' opinion and the view will determine the most critical features (weight), and it will be used in similarity computation by weighted average. By doing that, the most accurate calculation will be determined. In the future, data collection using the accelerometer sensor will be collected for older people with non-neurodegenerative diseases. An in-depth Artificial Neural Network can be done before getting better performance. Many details can be used to train the learning model. Finally, bring out some of the features that will be most helpful in training the learning model. In the future, the information-driven approach may be used to remove uncertainty as a legal system based on expertise.