iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 Issue: IV Month of publication: April 2021

DOI: https://doi.org/10.22214/ijraset.2021.33816

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue IV Apr 2021- Available at www.ijraset.com

Prediction of CKD (Chronic Kidney Disease) using Machine Learning

Naresh Thoutam[1], Gauri Mandlik[2], Pooja Mane[3], Neha Tiwari[4], Nutan Pandey[5]
[1]Assistant Prof., [2, 3, 4, 5]Student, Sandip Institute of Technology and Research Center, India

Abstract: Machine Learning, Support vector machine, Artificial Neural Network, Decision Tree, Naive In today's era everyone is trying to be conscious about health although due to workload and busy schedule one gives attention to the health when it shows any symptoms of some kind. But CKD is a disease which doesn't shows symptoms at all or in some cases it doesn't show any disease specific symptoms it is hard to predict, detect and prevent such a disease and this could be led to permanently health damage, but machine learning can be hope in this problem it is best in prediction and analysis. Bayes, Data Mining. Chronic Kidney Disease (CKD) implies that the human kidneys are harmed and unable to blood filter in the manner which they should. The disease is designated ‘‘chronic’’ in light of the fact that harm to human kidneys happen gradually over a significant time. This harm can make wastes to build up in your body. Many techniques and models have been developed to diagnose the CKD in early-stage. Among all techniques, Machine Learning (ML) techniques play a significant role in the early forecasting of different kinds ailments. ML techniques have been used for achieving analytical results which is one of the instruments utilize in medical analysis and prediction.

I. INTRODUCTION

Chronic kidney disease (CKD) is a significant public health problem worldwide, especially for low- and medium-income countries. Chronic kidney disease (CKD) means that the kidney does not work as expected and cannot correctly filter blood. About 10% of the population worldwide suffers from (CKD), and millions die each year because they cannot get affordable treatment, with the number increasing in the elderly.

According to the Global Burden Disease 2010 study conducted by the International Society of Nephrology, chronic kidney disease (CKD) has been raised as an important cause of mortality worldwide with the number of deaths increasing by 82.3% in the last two decades [1, 2]. Also, the number of patients reaching end-stage renal disease (ESRD) is increasing, which requires kidney transplantation or dialysis to save patients' lives [1, 3, 4]. CKD, in its early stages, has no symptoms; testing may be the only way to find out if the patient has kidney disease. Early detection of CKD in its initial stages can help the patient get effective treatment and then prohibit the progression to ESRD [1].

It is argued that every year, a person that has one of the CKD risk factors, such as a family history of kidney failure, hypertension, or diabetes, get checked. The sooner they know about having this disease, the sooner they can get treatment. To raise awareness and to encourage those who are most susceptible to the disease to perform the tests periodically, we hope that the disease can be detected with the least possible tests and at low cost. So, the objective of this research is to provide an effective model to predict the CKD by least number of predictors.

II. LITEARTURE SURVEY

Subashini, et al. [3] measured the performance of classification techniques using classifier such as Artificial Neural Network (ANN), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Naïve Bayes, Decision Tree, and Fuzzy. Using Matlab R2016a as analyzing tool, they validated and tested the classification parameters of six methodologies such as Accuracy, Precision, Recall, Specifity, and F-Measure on 400-instances CKD dataset. From the evaluation parameters, they concluded that the fuzzy technique serve the best performance for CKD classification.

P. Sinha, et al. [4] compared two data mining techniques in predicting CKD by using Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) classifier. The experimental analysis performed by Matlab tool showed that SVM classifier outperforms three of four evaluation parameters than KNN classifier. E. Celik, et al. [5] diagnosed and estimated of CKD using J48 and Sequential Minimal Optimization (SMO) algorithm in Weka tool. They proposed to compare Decision Tree to Support Vector Machine (SVM) technique in predicting CKD. The classification stage revealed that Decision Tree has more successful than SVM recognition in correct classification.

K.R. Lakshmi, et al. [6] performed a comparative study in predicting kidney dialysis survivability by using thee data mining techniques such as Artificial Neural Network (ANN), Decision Tree, and Logistic Regression. Data analysis used AVF data of 193 under-hemodialysis patients in Hashimenejad Kidney Center (HKC) of Tehran. Using 10-fold cross validation fold, they concluded that ANN show the high level compare than other two techniques in accuracy and sensitivity parameter. S. Ramya, et al. [7] compared Back-Propagation Neural Network (BPN), Radian Basis Function (RBF), and RandomForest (RF) in diagnosing CKD.

The dataset including 1000 instances and fifteen attributes were obained from medical reports of patients from different laboratories in Combatore. Using R tool, the experimental results revealed that the RBF gains the highest accuracy than other two algorithm.

Sunil D., et al. [8] addressed the use of two data mining techniques in Rapidminer tool for CKD prediction such as Naïve Bayes and Artificial Neural Network (ANN).

The results show that Naïve Bayes outperform ANN in classification accuracy. A. Subas, et al. [9] compared five data mining classifiers such as Artificial Neural Network (ANN), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), C.45 Decision Tree, and RandomForest (RF) Decision Tree in Weka tool. The evaluation parameters show that RF serves the highest accuracy than other three techniques.

S. Vijayarani, et al. [10] used Support Vector Machine (SVM) and Artificial Neural Network (ANN) to classify kidney disease such as Acute Nephritic Syndrome, Chronic Kidney Disease, Acute Renal Failure, and Chronic Glomerulonephritis. From the results of classification accuracy and execution time, SVM can be considered as better classifier compared to SVM.In recent years, few studies have been done on the classification or diagnosis of chronic kidney disease.

In 2013, T. Di Noia et al. [5], presented a software tool that used the artificial neural network ANN to classify patient status, which is likely to lead to end-stage renal disease (ESRD).

The classifiers were trained using the data collected at the University of Bari over a 38-year period, and the evaluation was done based on precision, recall, and F-measure. The presented software tool has been made available as both an Android mobile application and online web application.

Using data from Electronic Health Records (EHR) in 2014, H. S. Chase et al. [6] identified two groups of patients in stage 3: 117 progressor patients (eGFR declined >3 ml/min/1.73m2/year) and 364 non-progressor patients (eGFR declined <1 ml/min/1.73m2) .Where GFR is a glomerular filtration rate that commonly used to detect CKD. Based on initial lab data recorded, the authors used Naïve Bayes and Logistic Regression classifiers to develop a predictive model for progression from stage 3 to stage 4. They compared the metabolic complications between the two groups and found that phosphate values were significantly higher, but bicarbonate, hemoglobin, calcium, and albumin values were significantly lower in progressors compared to nonprogressors, even if initial eGFR values were similar.

Finally, they found that the probability of progression in patients classified as progressors was 81% (73% − 86%) and nonprogressors was 17% (13% − 23%).

Later in 2016, K. A. Padmanaban and G. Parthiban [7] aimed in their work to detect chronic kidney disease for diabetic patients using machine learning methods. In their research, they used 600 clinical records collected from a leading Chennai-based diabetes research center. The authors have tested the dataset using the decision tree and Naïve Bayes methods for classification using the WEKA tool. They concluded that the decision tree algorithm outweighs the Naïve Bayes with an accuracy of 91%.

A. Salekin and J. Stankovic [8] evaluated three classifiers: random forest, K-nearest neighbors, and neural network to detect the CKD. They used a dataset with 400 patients form UCI with 24 attributes. By using the wrapper method, a feature reduction analysis has been performed to find the attributes that detect this disease with high accuracy. By considering: albumin, specific gravity, diabetes mellitus, hemoglobin, and hypertension as features, they can predict the CKD with .98 F1 and 0.11 RMSE.

In the study carried out by W. Gunarathne, K. Perera, and K. Kahandawaarachchi [9], Microsoft Azore has been used to predict the patient status of CKD. By considering 14 attributes out of 25, they compared four different algorithms, which were Multiclass Decision Forest, Multiclass Decision Jungle, Multiclass Decision Regression, and Multiclass Neural Network. After comparison, they found that Multiclass Decision Forest performed the best with 99.1% accuracy.

H. Polat, H. D. Mehr, and A. Cetin [10] in their research used SVM algorithm along with two feature selection methods: filter and wrapper to reduce the dimensionality of the CKD dataset with two different evaluations for each method. For the wrapper approach, the ClassifierSubsetEval with the Greedy Stepwise search engine and WrapperSubsetEval with the Best First search engine were used. For the Filter approach, CfsSubsetEval with the Greedy Stepwise search engine and FilterSubsetEval with the Best First search engine were used. However, the best accuracy was 98.5% with 13 features using FilterSubsetEval with the Best First search engine using the SVM algorithm without mentioning which features were used.

P. Yildirim [1] studied the effect of sampling algorithms in predicting chronic kidney disease. The experiment was done by comparing the effect of the three sampling algorithms: Resample, SMOTE, and Spread Sup Sample on the prediction by multilayer perceptron classification algorithm.

The study showed that sampling algorithms could improve the classification algorithm performance, and the resample method has a higher accuracy among the sampling algorithms. On the other hand, Spread Sub Sample was better in terms of execution time. A. J. Aljaaf et al. [2] examined in their study the ability of four machine learning (ML) models for early prediction of CKD, which were: support vector machine (SVM), classification and regression tree (CART), logistic regression (LR), and multilayer perceptron neural network (MLP).

By using the CKD dataset from UCI and seven features out of 24, they compared the performance of these ML models. The results showed that the MLP model had the highest AUC and sensitivity. It was also noticeable that logistic regression almost had the same performance as MLP but with the advantage of the simplicity of the LR algorithm. Therefore, in our study, we can use the LR algorithm as a start or a benchmark and then use more complex algorithms.

Lastly in 2019, J. Xiao et al. [3] in their study established and compared nine ML models, including LR, Elastic Net, ridge regression lasso regression SVM, RF, XGBoost , knearest neighbor and neural network to predict the progression of CKD. They used available clinical features from 551 CKD follow-up patients. They conclude that linear models have the overall predictive power with an average AUC above 0.87 and precision above 0.8 and 0.8, respectively.

A. Problem Definition

The Chronic Kidney disease is the most important health issues concerning the people as a whole. Chronic diseases lead to morbidity and increase of death rates in India and other low- and middle-income countries. The chronic diseases account to about 60% of all deaths worldwide. 80% of chronic disease deaths worldwide also occur in low- and middle-income countries. In India, probably the number of deaths due to chronic disease found to be 5.21 million in 2008 and seems to be raised to 7.63 million in 2020 approximately 66.7%.

There are approximately 1 million cases of Chronic Kidney Disease (CKD) per year in India. Chronic kidney disease is also called renal failure. It is a dangerous disease of the kidney which produces gradual loss in kidney functionality. CKD is a slow and periodical loss of kidney function over a period of several years. A person will develop permanent kidney failure. If CKD is not detected and cured in early stage then patient can show following Symptoms: Blood Pressure, anemia, weak bones, poor nutrition health and nerve damage, decreased immune response because at advanced stages dangerous levels of fluids, electrolytes, and wastes can build up in your blood and body.

Hence it is essential to detect CKD at its early stage but it is unpredictable as its Symptoms develop slowly and aren't specific to the disease.

B. Goal and Objectives

Reduce chronic kidney disease (CKD) and associated complications, disability, death, and economic costs. CKD, in its early stages, has no symptoms; testing may be the only way to find out if the patient has kidney disease. Early detection of CKD in its initial stages can help the patient get effective treatment and then prohibit the progression to ESRD [1]. It is argued that every year, a person that has one of the CKD risk factors, such as a family history of kidney failure, hypertension, or diabetes, get checked. The sooner they know about having this disease, the sooner they can get treatment. To raise awareness and to encourage those who are most susceptible to the disease to perform the tests periodically, we hope that the disease can be detected with the least possible tests and at low cost. So, the objective of this project is to provide an effective model to predict the CKD by least number of predictors. The objective of this study is CKD prediction using data mining tool. The main task in this study is (1) various decision tree techniques are used for the prediction of the CKD; (2) comparing different decision tree techniques; and (3) finding best decision tree for the CKD prediction.

C. Outcomes

A system which might predict the CKD risk level of a patient with a better accuracy. Model development is based on categorization methods as Decision Tree, ANN, Naive Bayes and SVM algorithms. For Decision Tree, the models give precisions of 85%, for Naive Bayes 77% and 77.3% for Support Vector Machine. Outcomes show a significant accuracy of the methods.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue IV Apr 2021- Available at www.ijraset.com

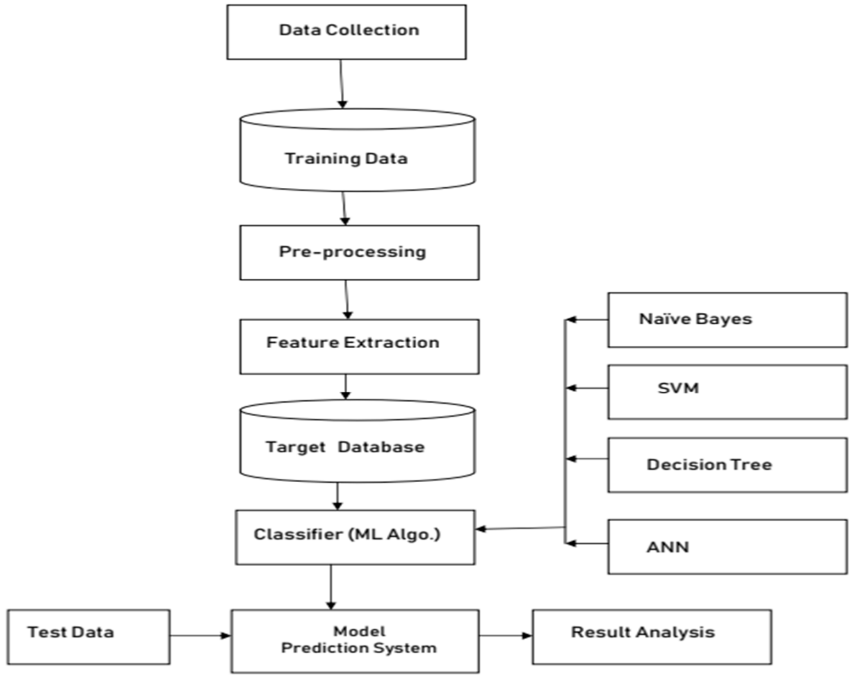


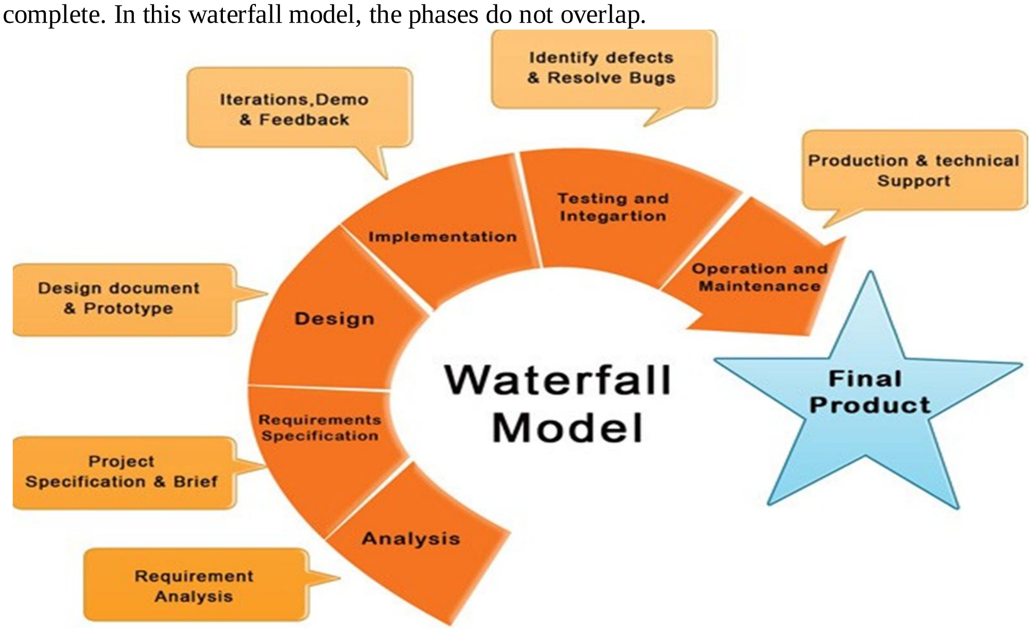
Fig: Proposed System

Proposed System

III. METHODOLOGY

The Waterfall Model was the first Process Model to be introduced. It is also referred to as a linear-sequential life cycle model. It is very simple to understand and use. In a waterfall model, each phase must be completed before the next phase can begin and there is no overlapping in the phases.

The Waterfall model is the earliest SDLC approach that was used for software development. The waterfall Model illustrates the software development process in a linear sequential flow. This means that any phase in the development process begins only if the previous phase is complete. In this waterfall model, the phases do not overlap.



A. *Waterfall Model - Design*

Waterfall approach was first SDLC Model to be used widely in Software Engineering to ensure success of the project. In "The Waterfall" approach, the whole process of software development is divided into separate phases. In this Waterfall model, typically, the outcome of one phase acts as the input for the next phase sequentially.

The sequential phases in Waterfall model are −

1) *Requirement Gathering and Analysis*: All possible requirements of the system to be developed are captured in this phase and documented in a requirement specification document.
2) *System Design*: The requirement specifications from first phase are studied in this phase and the system design is prepared. This system design helps in specifying hardware and system requirements and helps in defining the overall system architecture.
3) *Implementation*: With inputs from the system design, the system is first developed in small programs called units, which are integrated in the next phase. Each unit is developed and tested for its functionality, which is referred to as Unit Testing.
4) *Integration and Testing*: All the units developed in the implementation phase are integrated into a system after testing of each unit. Post integration the entire system is tested for any faults and failures.
5) *Deployment of System*: Once the functional and non-functional testing is done; the product is deployed in the customer environment or released into the market.
6) *Maintenance:* There are some issues which come up in the client environment. To fix those issues, patches are released. Also, to enhance the product some better versions are released. Maintenance is done to deliver these changes in the customer environment.

All these phases are cascaded to each other in which progress is seen as flowing steadily downwards (like a waterfall) through the phases. The next phase is started only after the defined set of goals are achieved for previous phase and it is signed off, so the name "Waterfall Model". In this model, phases do not overlap.

B. *Usage Scenario*

The use case scenario considers the goal of publishing a short story. It breaks down the process of book publishing by describing the actors, the typical workflow in the main success story, and the things that could go wrong, called extensions. When managing project that use UML conventions, there can be temptation to jump straight into the case diagram, with stick guards, ovals, and lots of lines. But if you don’t know your goals and who’s involved, take a step back and write your goals down in prose.

A use case describes how a user uses a system to accomplish a particular goal. A use case diagram consists of the system, the related use cases and actors and relates these to each other to visualize: what is being described? (**system**), who is using the system? (**actors**) and what do the actors want to achieve? (**use cases**), thus, use cases help ensure that the correct system is developed by capturing the requirements from the user's point of view. A use case is a list of actions or event steps typically defining the interactions between a role of an actor and a system to achieve a goal. A use case is a useful technique for identifying, clarifying, and organizing system requirements. A use case is made up of a set of possible sequences of interactions between systems and users that defines the features to be implemented and the resolution of any errors that may be encountered.

While a use case itself might drill into a lot of detail (such as, flow of events and scenarios) about every possibility, a use-case diagram can help provide a higher-level view of the system, providing the simplified and graphical representation of what the system must actually do.

A use case (or set of use cases) has these characteristics:

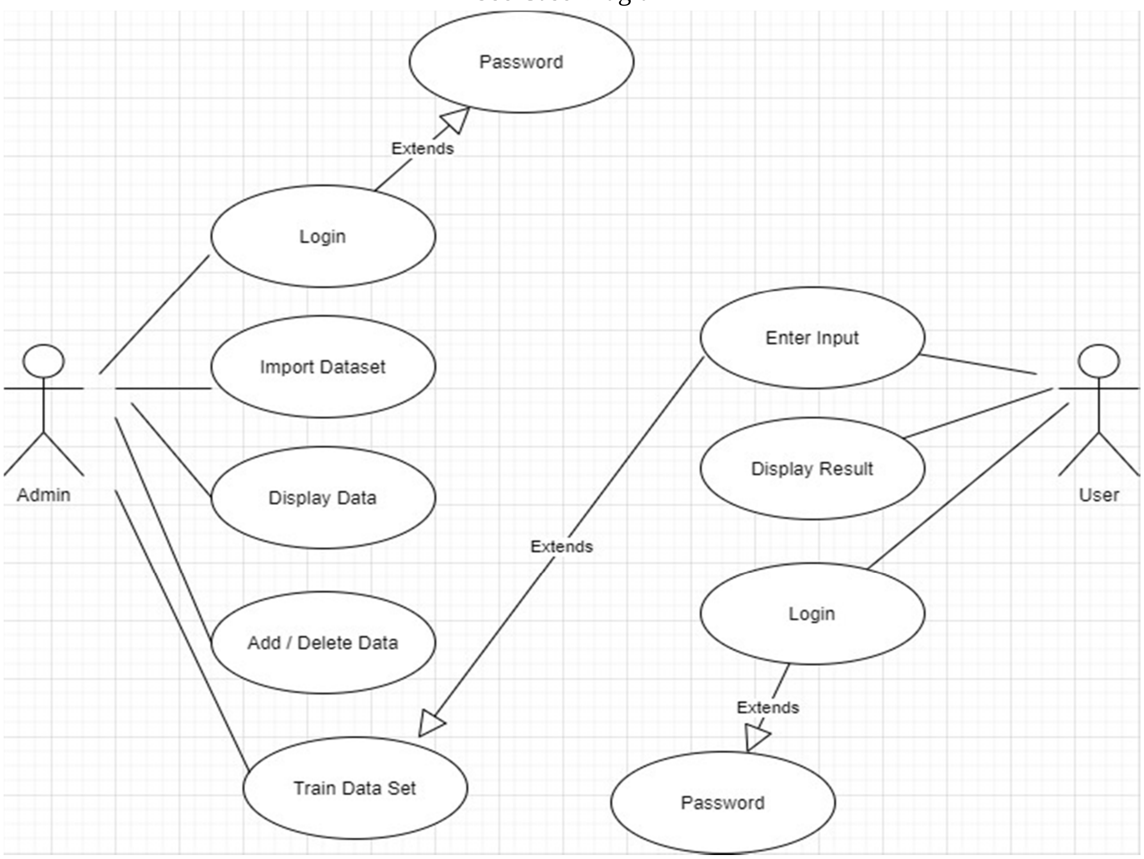
1) Organizes functional requirements
2) Models the goals of system/actor (user) interactions
3) Describes one main flow of events (main scenarios) and possibly other exceptional flows (alternatives), also called paths or user scenarios

a) *Actor:* Actors are usually individuals involved with the system defined according to their roles. The actor can be a human or other external system.
b) *Use Case:* A use case describes how actors uses a system to accomplish a particular goal. Use cases are typically initiated by a user to fulfill goals describing the activities and variants involved in attaining the goal.
c) *Relationship:* The relationships between and among the actors and the use cases.
d) *System Boundary:* The system boundary defines the system of interest in relation to the world around it.
e) *Benefits of Use Case Diagram:* Use cases is a powerful technique for the elicitation and documentation of black-box functional requirements.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue IV Apr 2021- Available at www.ijraset.com

Because, use cases are easy to understand and provide an excellent way for communicating with customers and users as they are written in natural language. Use cases can help manage the complexity of large projects by partitioning the problem into major user features (i.e., use cases) and by specifying applications from the users' perspective. A use case scenario, often represented by a sequence diagram, involves the collaboration of multiple objects and classes, use cases help identify the messages (operations and the information or data required - parameters) that glue the objects and classes together. Use cases provide a good basis to link between the verification of the higher-level models (i.e. interaction between actors and a set of collaborative objects), and subsequently, for the validation of the functional requirements (i.e. blueprint of white-box test).
Use case driven approach provides an traceable links for project tracking in which the key development activities such as the use cases implemented, tested, and delivered fulfilling the goals and objectives from the user point of views.

Use Case Diagram



f) User Profiles

- *Admin:* This user profile is responsible to perform login, import dataset, display data, add or delete data and train the dataset with extending results to user profile.
- *User*: This profile is responsible to perform entering input, display result and authentication.

C. Applications

1) Hospitals.
2) Medical Domain.
3) Society.

D. Hardware Resources Required

1) *Processor:* Pentium IV or higher
2) *RAM:* 256 MB
3) *Space on Hard Disk:* minimum 512MB

E. Software Resources Required

1) Python
2) Django
3) HTML and CSS
4) WampServer

IV. CONCLUSION

At last, we expect that we have reached our results of predicting CKD by applying machine learning algorithms widely. The entire procedure was based on machine learning workflow. We have completed a classification model task with proper practical knowledge to deal with. We have undergone through many steps such as data exploration, data cleaning steps, feature of engineering basics and advanced model feature selections, model selection parameters and hyper parameter tuning using Scikit Learn library from Jupyter notebook. There is a strong possibility that running the project with logistic regression would give much better results

REFERENCES

[1] Siddheshwar Tekale, Pranjal, Sukanya, Ankit, "Prediction of Chronic Kidney Disease Using Machine Learning Algorithm, International Journal of Advanced Research in Computer and Communication Engineering,Vol. 7, Issue 10, October 2018.

[2] Revathy Ramesh, "Chronic Kidney Disease Prediction using machine Learning Models", https://www.researchgate.net/publication/341398109, Article · May 2020 DOI: 10.35940/ijeat.A2213.109119

[3] Bilal Khan, Rashid, Fazal, Ghulam, And Kim, "An Empirical Evaluation of Machine Learning Techniques for Chronic Kidney Disease Prophecy", IEEE Access, Received March 2, 2020, accepted March 14, 2020, date of publication March 18, 2020, date of current version March 30, 2020.Digital Object Identifier 10.1109/ACCESS.2020.2981689

[4] Kerina Blessmore, Noorie, Ronnie, Iyengar, "Risk Level Prediction of Chronic Kidney Disease Using NeuroFuzzy and Hierarchical Clustering Algorithm (s)", International Journal of Multimedia and Ubiquitous Engineering, Vol.12, No.8 (2017), pp.23-36 http://dx.doi.org/10.14257/ijmue.2017.12.8.03

[5] Zixian Wang, Jae Won, Xilin, Yantong, Muning, Anqi, "Machine Learning-Based Prediction System For Chronic Kidney Disease Using Associative Classification Technique", International Journal of Engineering & Technology, 7 (4.36) (2018) 1161-1167.

[6] Reshma, Salma, Ajina, Vishnu, Janisha, "Chronic Kidney Disease Prediction using Machine Learning", International Journal of Engineering Research & Technology (IJERT) http://www.ijert.org ISSN: 2278-0181, IJERTV9IS070092 Vol. 9 Issue 07, July-2020

[7] Marwa Almasoud, Tomas, "Detection of Chronic Kidney Disease using Machine Learning Algorithms with Least Number of Predictors", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 8, 2019

[8] El-Houissainy Rady, Ayman Anwar, "Prediction of kidney disease stages using data mining algorithms", https://doi.org/10.1016/j.imu.2019.100178 Received 9 December 2018; Received in revised form 11 March 2019; Accepted 1 April 2019 Informatics in Medicine Unlocked 15 (2019) 100178 Available online 07 April 2019 2352-9148/ © 2019 Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/BY-NC-ND/4.0/)

[9] I A Prasadana, Hartama, Zarlis, Sianipar, Munandar, Baeha, Alam, "Chronic Kidney Disease Prediction by Using Different Decision Tree Techniques", The International Conference on Computer Science and Applied Mathematic IOP Conf. Series: Journal of Physics: Conf. Series 1255 (2019) 012024 IOP Publishing doi:10.1088/1742-6596/1255/1/012024

[10] Imesh Udara, Damayanthi Herath, "Chronic Kidney Disease Prediction Using Machine Learning Methods", 978-1-7281-9975-7/20/$31.00 ©2020 IEEE Moratuwa Engineering Research Conference (MERCon) 2020 UTC from IEEE Xplore