An Analysis for Predicting Chronic Kidney Disease- Using Machine Learning Approaches

Submitted by

Sadia Haque Piyal

181-35-302

Department of Software Engineering

Daffodil International University

Supervised by

Mr. SAM Matiur Rahman

Associate Professor

Department of Software Engineering

Daffodil International University

This Project report has been submitted in fulfillment of the requirements for the Degree of

Bachelor of Science in Software Engineering.

APPROVAL

This thesis entitled on **"An analysis for predicting Chronic Kidney Disease- using Machine Learning approaches"**, submitted by **Sadia Haque Piyal (ID: 181-35-302)** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and approval as to its style and contents.

BOARD OF EXAMINERS

Chairman

Dr. Imran Mahmud
Associate Professor and Head
Department of Software Engineering
Daffodil International University

External Examiner

Md. Tanvir Quader
Senior Software Engineer
a2i Programme

Internal Examiner 1

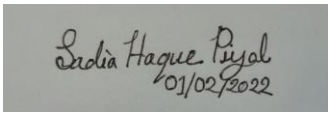
S A M Matiur Rahman
Associate Professor
Department of Software Engineering
Daffodil International University

Internal Examiner 2

Syeda Sumbul Hossain
Lecturer (Senior Scale)
Department of Software Engineering
Daffodil International University

DECLARATION

It hereby declares that this thesis has been done by us under the supervision of **Mr. SAM Matiur Rahman**, Lecturer, Department of Software Engineering, Daffodil International University. It is also declared that neither this thesis nor any part of this has been submitted elsewhere for the award of any degree.



_____

Sadia Haque Piyal

Student ID: 181-35-302

Batch**:** 25th

Department of Software Engineering

Faculty of Science & Information Technology

Daffodil International University

Certified by:

_____

Mr. SAM Matiur Rahman

Associate Professor

Department of Software Engineering

Faculty of Science & Information Technology

Daffodil International University

ACKNOWLEDGEMENT

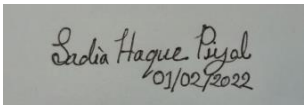
First of all, I am grateful to the Almighty Allah for giving me the ability to complete the final thesis.

I would like to express our gratitude to our supervisor **Mr. SAM Matiur Rahman** for the consistent help of my thesis and research work, through her understanding, inspiration, energy, and knowledge sharing. Her direction helped us to find the solutions of research work and reach my final theory.

I would like to express my extreme sincere gratitude and appreciation to all of our teachers of the **Software Engineering** department for their kind help, generous advice and support during the study.

I am also expressing my gratitude to all of our friend's, senior, junior, who, directly or indirectly, have lent their helping hand in this venture.

Last but not the least, I would like to thank our family for giving birth to us in the first place and supporting me spiritually throughout my life.



Sadia Haque Piyal
181-35-302

Table of Contents

LIST OF FIGURE

LIST OF TABLE

LIST OF NOMENCLATURE

Terms	Nomenclature
CKD	Chronic Kidney disease
AKI	Acute Kidney Injury
XGBoost	Extreme Gradient Boosting
LGBM	Light Gradient Boosted Machine
Ada Boost	Adapting Boosting
KNN	K-Nearest Neighbour
SVM	Support Vector Machine
CHAID	Chi-square Automatic Interaction Detector
ESRD	End-Stage Renal Disease
AKD	Acute Kidney Disease or Disorder
RBC	Red Blood Cell
WBC	White Blood Cell

Abstract

Machine learning is the biggest help now-a-days to develop models of the actual biological systems which are both predictive and informative. Bioscience is the field where we can collect a huge amount of data. By using data mining processes, machine learning has become a growing use in biomedical technology because of the ever-increasing volume and complexity of biological data. Chronic kidney disease or chronic kidney failure is a term when kidneys are damaged and the wastes couldn’t be filtered as kidneys always do. High blood pressure and diabetes are the main causes of CKD. The bad situation occurs when the damage leads to kidney transplant. So, early detection is very much needed in this situation and for predicting CKD I have employed some ML techniques by using a dataset with four hundred clinical data. I have used most of the Machine learning algorithms but four of them(Random forest, XGBoost, Ada Boost, LGBM Classifier) gave me promising results and this model's performance is the best to predict CKD with the given dataset.

Chapter 1
Introduction

1.1 Problem outline:

Kidneys are one of the major organs in the human body. It removes wastes or drugs from our body and also balances our body's fluids. It works to release hormones which controls the pressure of blood and also controls RBC(red blood cells)production. The excess fluid that kidneys remove through urine are highly complex steps. It maintains our body's potassium,acid and salt. The hormones that the kidney produces stimulates RBC production. But if someone got diabetes or high blood pressure it can lead to kidney disease. They are the major cause. If a person has kidney failure in his family history or older aged people have the risk of getting CKD. The symptoms of this disease include feeling low energy or tired more often, concentrating problems,poor appetite, sleeping problems,and urinating more especially at night. Primary kidney disease has these types of symptoms which are called AKI(acute kidney injury) and also called AKD(acute kidney disease or disorder)[11]. Long term AKI or AKD causes CKD(Chronic kidney disease) where kidneys fail to clean the wastes from fluid or blood.[12]

This literature shows the best model description which will predict Chronic kidney disease at an early stage and give the best performance by using five best algorithms with highest accuracy.

1.2 Motivation:

Chronic Kidney Disease(CKD) is a term when kidneys get damaged day by day and to filter the blood kidney faces a lot of problems. People who suffer from this disease have a habit of not drinking proper amounts of water and maintaining healthy food habits. CKD can also happen If a person has type 2 diabetes, high blood pressure or if there is kidney disease or failure in the lineage.CKD and AKI can turn into an end-stage renal disease(ESRD) when renal function has failed miserably where the kidneys are unable to function independently.[13] So, an early identification of CKD can enhance one's life quality significantly. This necessitates the development of a good prediction model to detect CKD at a beginning period. This system will use different types of classifiers through data pre-processing, data transformation process for predicting CKD and also it proposes the best fitted Prediction model for Chronic Kidney Disease.

1.3 Research Objective:

To measure the performance using different types of algorithms to propose the best Prediction framework for the early stage of CKD.

1.4 Research Design:

Here we will see research based on Chronic kidney disease prediction. After giving a description of background study I have attached a dataset description and a table with parameter description. Data pre-processing system shows steps how I pre-processed the data. Then I have given a research methodology where all the algorithms description has been given. And then the result and conclusion has been given. By using different types of machine learning algorithms I have trained the model to predict the CKD with highest accuracy.

Chapter 2

Literature Review

In this paper, Decision tree, SVM and Random Forest algorithm has been used for the prediction of CKD. They showed that among these algorithms Random Forest Classifier model is much better at predicting CKD. Decision Trees accuracy level 94.16%, Support Vector Machines accuracy level 98.33% and Random Forest accuracy level 99.16%. It was based on 70% of the original Training model CKD dataset. Here, they use only three algorithms and the other algorithms have been unused. We need to build a model that helps with more algorithms.[1]

In this study, They used three algorithms (decision tree, random forest algorithm and logistic regression) with the accuracy 98.75% , 97.5% and 98.75%. The model should be upgraded for more algorithms with highest accuracy that can work with more datasets.[2]

In this research, For predicting CKD we need a better model which can predict the accuracy level of the algorithm. Mainly naive bayes and decision trees have been used which are two classifier methods. Naive Bayes methods accuracy was much lower than the accuracy level of the decision tree (99.25%). The performance of the Classification algorithm was Specificity 99.33% and Sensitivity 99.20%. Here, they use only two algorithms and the other algorithms have been unused. We need to build a model that helps with more algorithms.[3]

In this paper, they used three different techniques which are LASSO regression, Wrapper method and feature selection (correlation-based). Here, they use seven algorithms named KNN, Random Forest, linear support vector machine (LSVM), logistic regression, CHAID, C5.0, artificial neural network. LSVM has the highest accuracy level which was 98.86%. Here, The model should be upgraded for more algorithms with highest accuracy that can work with more datasets.[4]

Here, They used two algorithms (KNN and logistic regression) with the accuracy 95.75% and 98.5%. By systematically evaluating prepared tests and indicators, logistic regression generates a variety of alternative trees. Here, They use only two algorithms and the other algorithms have been unused. We need to build a model that helps with more algorithms.[5]

In this paper, GB, XGB, LR, and SVM classifiers have been used for the prediction of CKD. They showed that among these algorithms Random Forest Classifier model is much better at predicting CKD. The accuracy levels are 98.25%, 98.50%, 97.25% and 97.75%.Here, The model should be upgraded for more algorithms with highest accuracy that can work with more datasets.[6]

For predicting CKD we need a better model which can predict the accuracy level of the algorithm. In here, KNN and Gaussian Naive Bayes have been used which are two classifier methods. naive Bayes methods accuracy was much lower(98.30%) than the accuracy level of the KNN (99.14%). Here, they use only two algorithms and the other algorithms have been unused. We need to build a model that helps with more algorithms.[7]

Here, in this paper, They used two algorithms (random forest and extra trees classifier) for the prediction of CKD which have 100% accuracy in the comparison of other models. It has minimal bias for specific attributes. We need to build a model that helps with more algorithms.[8]

Here, they work with the classification algorithm where data has been tested and trained and validated (10-fold cross validation). The GB algorithm gives the best accuracy where specificity was 99.3% and sensitivity was 98.8%. They use only two algorithms and the other algorithms have been unused. We need to build a model that helps with more algorithms.[9]

In this paper, They used four algorithms for the prediction of CKD where the GB classifier has given the best accuracy 99.80%. The model should be upgraded for more algorithms with highest accuracy that can work with more datasets. In ROC and AUC curves this algorithm gives the highest predictable score than other algorithms.[10]

Chapter 3

Research Methodology:

This research will show a CKD prediction model which has two parts such as training and testing data after pre-processing the dataset. While training the data we have used that pre-processed dataset. By using different types of machine learning algorithms we have trained the model to predict the CKD with highest accuracy.

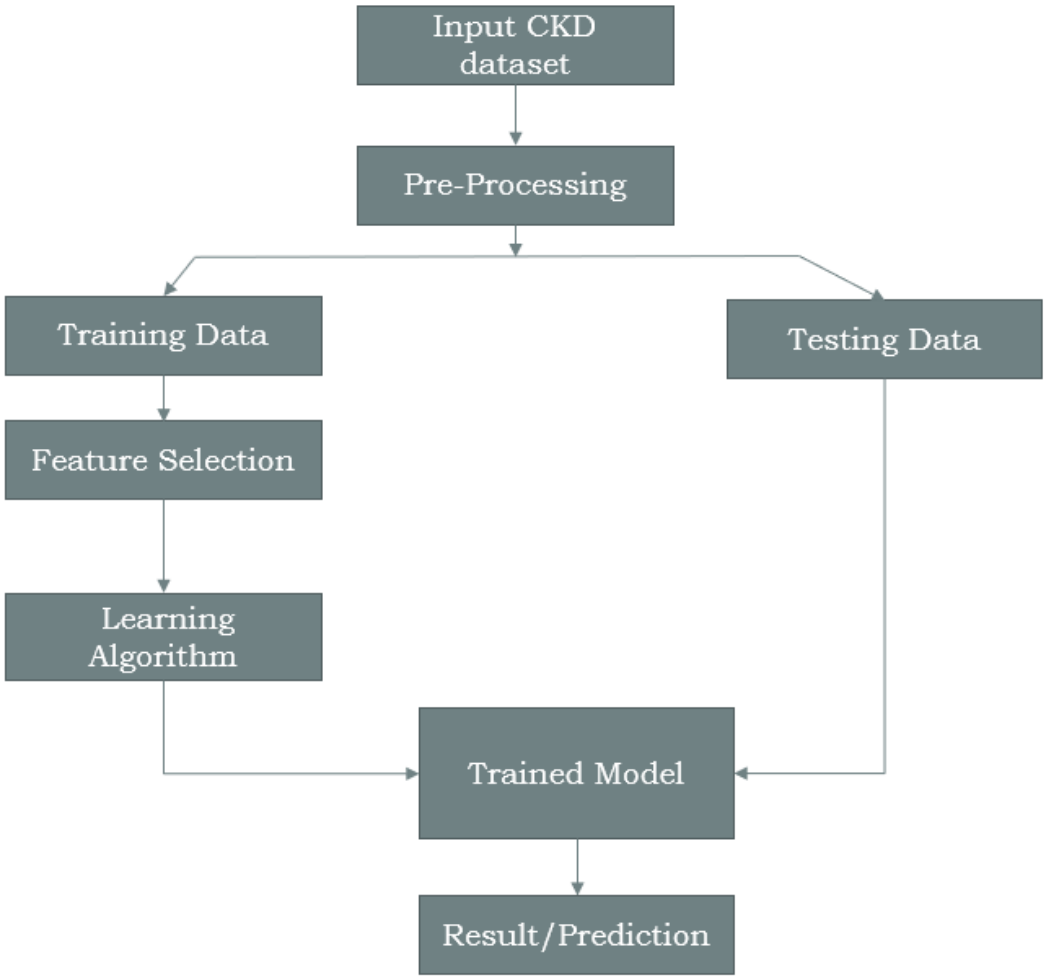


Fig. 1: Workflow of Chronic Kidney Disease prediction.

3.1 Dataset Description:

In this study, I have used a CKD dataset for my prediction model. I have taken this CKD dataset from kaggle[14]. There are 400 data in the dataset and 14 medical predictors where class is only one dependent variable and other parameters are independent variables. The parameter description is given in Table 1.
The diagnostic measurement indicates whether the patient has CKD or not. From the dataset 37.5% of the patients don't have CKD and 62.5% of patients have CKD.

Table 1: Dataset feature description

Predictor parameter	Type	Description
Class	CKD or not CKD	It defines the result of classification.
Bp(Blood Pressure)	Numerical	It shows the patient's blood pressure level where the min value is 50 and the max value is 180.
Sg(Specific Gravity)	Numerical	This test shows the density of patients' urine compared to the water.
Al(Albumin)	Numerical	The normal level is 3.4-5.4 g/dl. If the level gets lower, patients get malnutrition and if they get higher than acute infection, burns happen.[19]
Su(Sugar)	Numerical	It means the sugar level of blood where 140 mg/dl is normal.
Rbc(Red Blood Cell)	Numerical	It counts how many Red Blood Cells are in the patient's blood.
Bu(Blood Urea)	Numerical	It means how much wastes are in the patient's blood.
Sc(Serum Creatinine)	Numerical	It counts the amount of creatinine in the patient's blood.
Sod(Sodium)	Numerical	The normal sodium level is 135-145 mEq/L. If the level falls down then hyponatremia happens.
Pot(Potassium)	Numerical	It indicates the potassium level of the patient's blood where 3.6-5.2 mmol/L is normal and if

		the level is higher than 6 then it will be dangerous.
Hemo(Hemoglobin)	Numerical	It counts how much RBC is in the patient's blood. If the hemoglobin level is low then it's called anemia.[18]
Wbcc(White Blood Cell Count)	Numerical	It counts how many white cells are in a patient's blood. It fights against bacteria, germs, etc, and creates more WBC to fight.[17]
Rbcc(Red Blood Cell Count)	Numerical	Red Blood Cell Counts how much RBC are in patients blood. The normal amount is 4.2-5.6 per microliter of blood for women and 4.7 -6.1 per microliter of blood for men.[16]
Htn(Hypertension)	Numerical	It counts the blood pressure level of the patient's where hypertension means the level is systolic pressure higher than 120 and diastolic pressure lower than 80.[15]

3.2 Data pre-processing:

In this dataset, I have 400 CKD patients data and 14 parameters and the class column is dependent on the other columns where other columns are independent and have a strong relationship with the class column. When I have started an analysis, no missing data has been found. The dataset correlation of all features illustrated in the figure 2. After completing the analysis of correlation and outliers, the experiment showed some major outliers in the dataset. Some major outliers depicted in figure 3. Since, The dataset had some outliers, then the experiment performed interquartile range to remove the outliers.

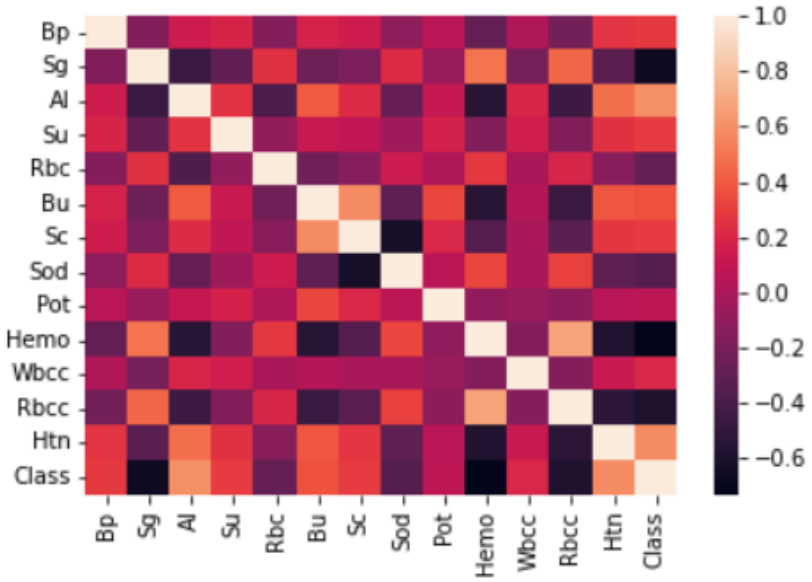


Fig. 2:Dataset correlation of all features.

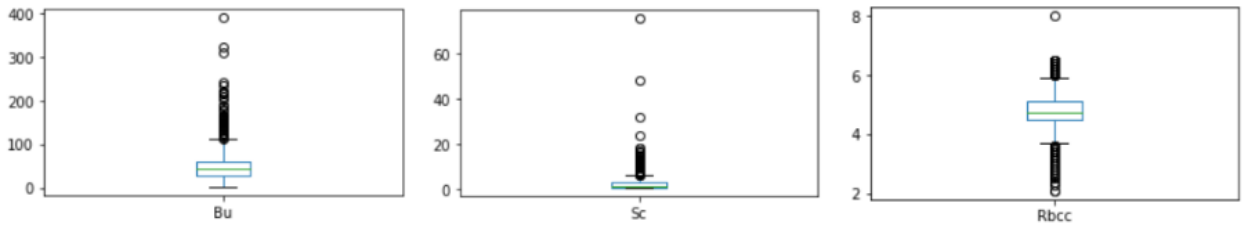


Fig. 3: Some major outliers in the dataset.

3.3 Performance Measure Parameter:

Machine learning model evaluated some important measurement parameters. All measurement parameters generate from confusion metrics. Confusion matrices counted TP, FP,TN and FN from actual data and predicted data. All counted values used to evaluate the model. The following is implication terms,

Table 2: Terms of performance metrics

Actual/Predicted	Actually Positive	Actually Negative
Predicted Positive	True Positive(TP)	False Positive(FP)
Predicted Negative	False Negative(FN)	True Negative(TN)

In this study, there are some important evaluation matrices generated from confusion matrices by their corresponding formula. All formulas describe the important relationship between actual class and predicted class. This study all model evaluated and compare following formula,

Accuracy - The exactness of a classifier is a measure of how effectively it can correctly anticipate situations into the appropriate classification. It's the number of correct predictions divided by the total number of instances in the data collection. It's worth noting that accuracy is highly dependent on the classifier's edge selection, and hence may vary among testing sets. Along these lines, it's not the best method for comparing different classifiers, although it can provide a summary of the classification. As a result, accuracy is frequently calculated using the equation:

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+FP+FN+TN)} \quad \ldots\ldots 1$$

Recall - also known as sensitivity, refers to the rate at which positive predictions are effectively anticipated as positive. This measurement is very important, attractive especially in the medical field of observation correctly analyzed. During this examination, it is more important to correctly identify a dangerous CKD situation than it is to incorrectly discern a primary stage. Hence, Recall is calculate according this formula,

$$\text{Recall} = \frac{TP}{(TP+FN)} \quad \ldots\ldots 2$$

Precision - also known as assurance, is the rate at which both true positive and true negative events are differentiated as obvious positive. This demonstrates how effectively the classifier handles positive observations, but says little about negative ones.

Precision - $\frac{TP}{(TP+FP)}$……3

F1-Score - The weighted harmonic mean of Precision and Recall is the F1-Score. As a result, this score considers both false positive and false negative outcomes.

F1- Score = $\frac{2\times(Recall \times Precision)}{(Recall + Precision)}$ …… 4

3.4 Ensemble Classification Algorithm:

Ensemble methods are ways for enhancing the accuracy of model findings by mixing numerous models rather than utilizing just one. The integrated models considerably improve the accuracy of the outcomes. This is the best way for improving predictability by combining other methods. Ensemble methods are best for classification where they reduce bias and improve accuracy.

1. **Random Forest(RF):**

   Random forest (RF) is a technique for ensemble learning which has been shown to be a very advanced and useful classifier. RF is made up of a number of decision trees, each of which is received training on a different feature vector from the training data. All of the trees in the forest evaluate a new test sample uniquely and produce a classification result. The RF determines the projected class of the test data based on the majority of votes cast across all of the network's trees.

2. **XGBoost:**

   XGBoost is a machine learning method that has recently dominated Kaggle tournaments for structured or tabular data. XGBoost is a high-speed and high-performance implementation of gradient boosted decision trees. Execution Speed and Model Performance are the two reasons to use XGBoost.[21] XGBoost is a distributed gradient boosting toolkit that has been tuned for efficiency, flexibility, and portability. It uses the Gradient Boosting framework to create Machine Learning algorithms. It uses parallel tree boosting to tackle a wide range of data science issues quickly and accurately.

3. AdaBoost:

The AdaBoost algorithm is a Boosting approach used in Machine Learning as an Ensemble Method. The weights are re-allocated to each occurrence, with higher weights applied to improperly identified instances. This is termed Adaptive Boosting.[20] Adaptive Boosting helps to solve classification issues by transforming weak learners or predictors into strong predictors. AdaBoost has a few drawbacks, including the fact that it is based on objective research and is particularly susceptible to homogeneous noise. Weak classifiers might result in low margins and overfitting if they are too weak.

4. LGBM:

It is a tree based gradient boosting algorithm. LGBM stands for Light Gradient Boosting Machine. They are GOSS (Gradient Boosting Decision Tree) or GBDT (Gradient Boosting Decision Tree) (Gradient based one-sided sampling). These are the most extensively used machine learning algorithms for study prediction[27].The GBDT is a boosting algorithm, in which boosting is an all-encompassing collective process that replicates a strong classifier from a collection of weak ones. Its most important characteristics are efficacy, accuracy, and interoperability. It is an extremely effective and accessible advancing tree agenda[27]. With the rise of big data in recent years, GBDT has faced new-fangled issues, particularly in the trade-off between competency and precision.In traditional GBDT implementations, all of the data instances required to obtain the appropriate information are scanned. There will be a proportionality between the number of cases and the number of structures involved as a result of this[27]. GOSS is another algorithm that can be used to forecast a model. Gradient-based One-Side Sampling (GOSS) can be used in conjunction with the LGBM Algorithm. Although "in GBDT there is no instinctual heft for data instances," we note that data instances with varying gradients (the rate of rise or fall along the length of the road with respect to the horizontal is called grade or gradient) play distinct roles in the computation of data exchange. As a result, "high gradients must be used in order to maintain the precision of information gain estimation" when downsampling data instances[28].

Chapter 4

Result and Discussion

In this section, I will make some analytical procedure which will give the best analysis for predicting CKD. This model is put to the test using a variety of measuring experiments. So, after comparing all the classifiers a performance measure will be shown. By learning about many algorithm techniques in various sites and papers I have decided to work on those four ensemble techniques and present the performance and the comparison among those for the prediction of CKD. In figure 3, we can see the accuracy level of the employed model.

Figure 4 shows that LGBM algorithms accuracy level was the best among those algorithms which is 1.0 while XGBoost and Random Forest have accuray 0.99 percent. Meanwhile, AdaBoost algorithms accuracy is 0.98 percent which is much lower than others.

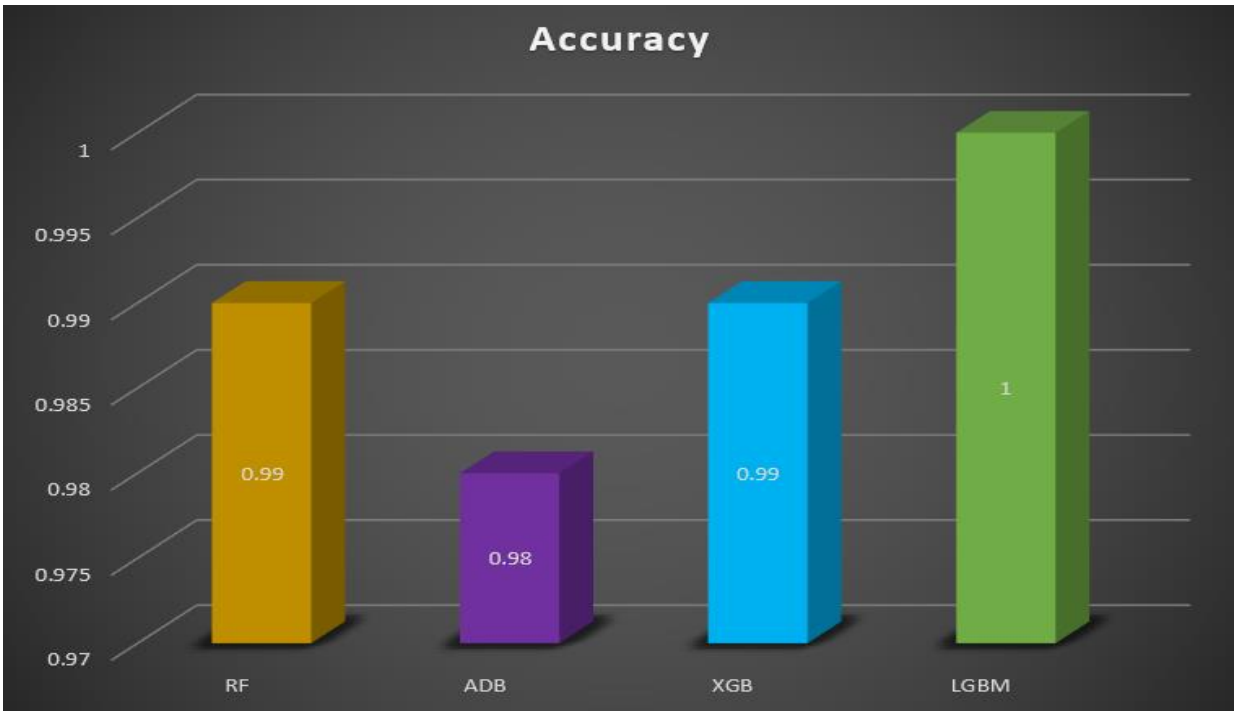


Fig. 4: Accuracy of all employed model (CKD prediction).

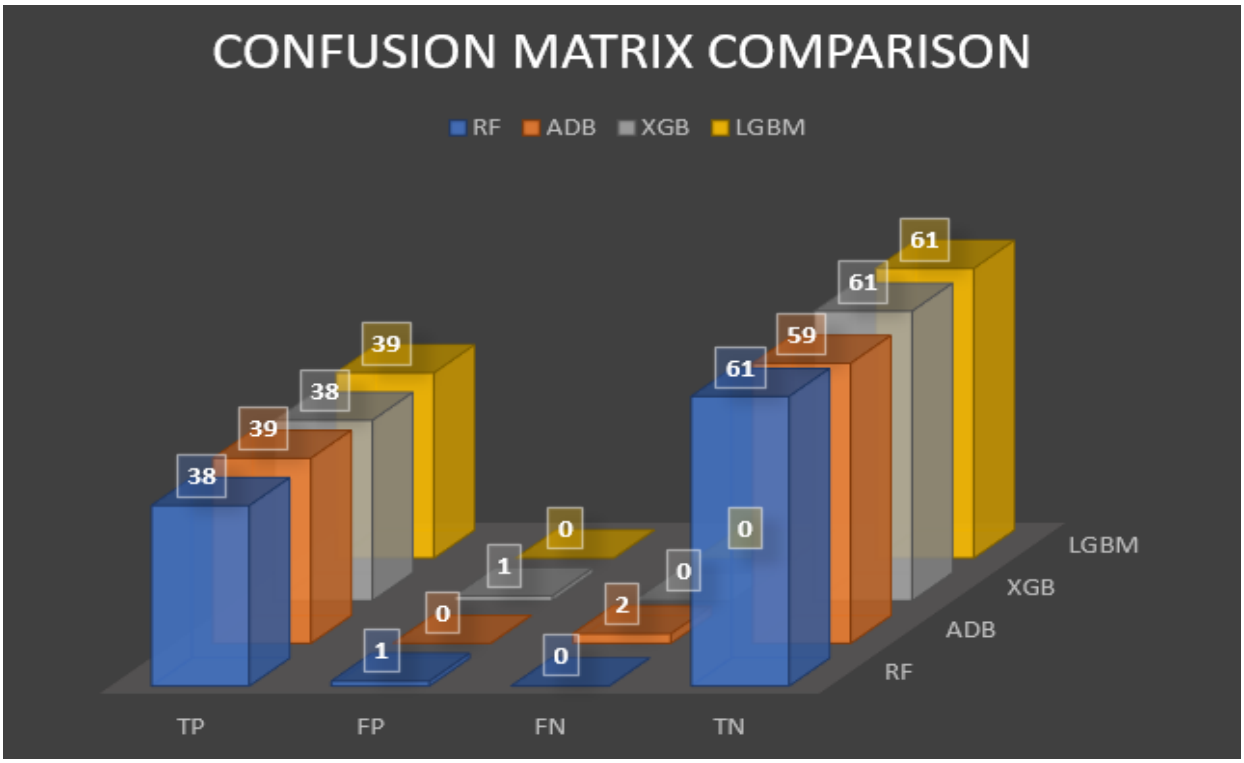


Fig. 5:Confusion matrix analysis for the ensemble method(CKD prediction).

In figure 5 ,we can see that all the four algorithm's confusion matrix analysis where LGBM classifier performance is best and RF and XGB classifiers have medium performance in the confusion matrix. LGBM classifiers have TP(39), FP(0), FN(0) and TN(61) which was best for the prediction among those algorithms.

Fig. 6: performance metrics of four ensemble methods(CKD prediction).

In figure 6, four algorithms performance metrics have been shown where AdaBoost and LGBM algorithms precision level was best both are 1.0. The highest recall was gained by Random Forest, XGBoost and LightGBM methods. In the terms of f1-score, all the classification algorithms have greater performance which is above 0.98 percent.

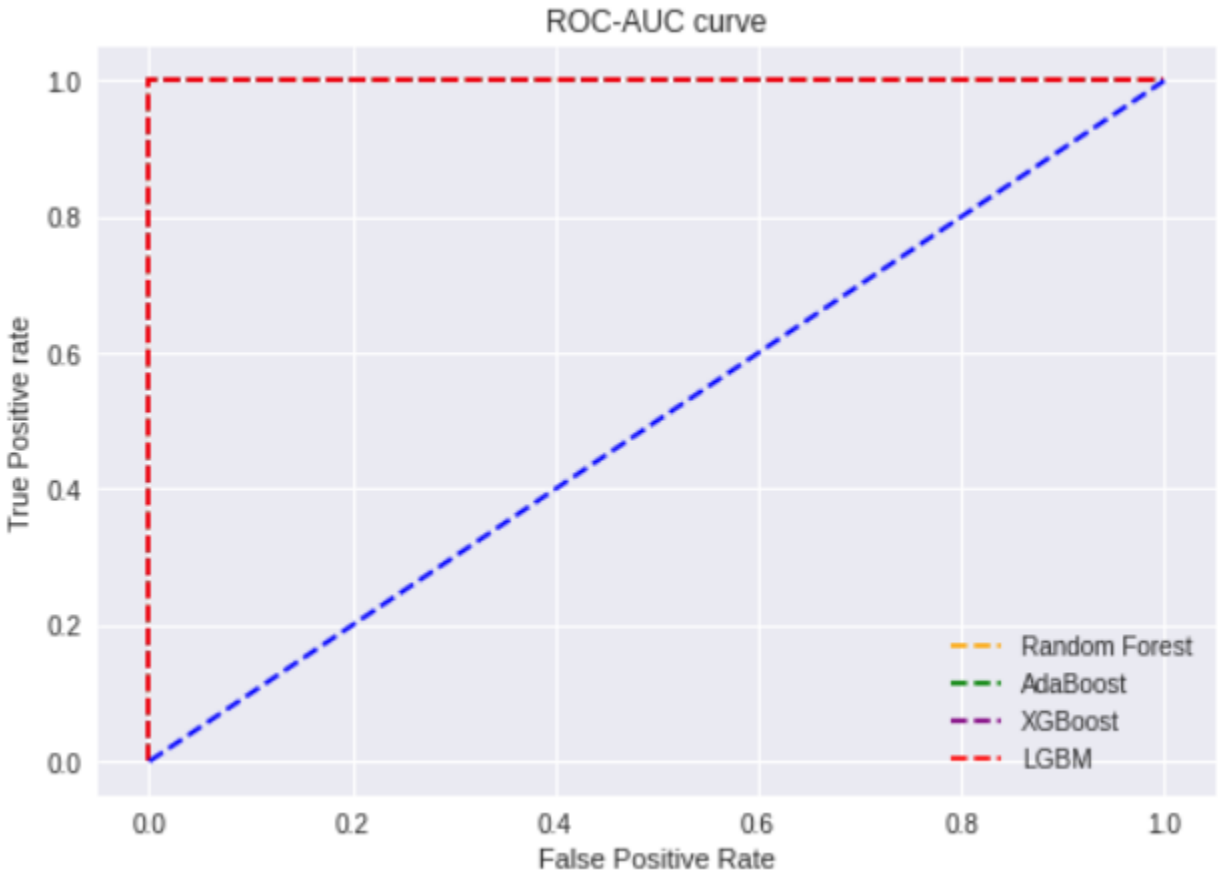


Fig. 7: Receiver Operating Characteristics (ROC) for CKD dataset.

The trade-off between sensitivity (or TPR) and specificity (1 – FPR) is depicted by the ROC curve. Classifiers with curves that are closer to the top-left corner perform better. The test becomes less accurate when the curve approaches the ROC space's 45-degree diagonal. Figure 6 shows the ROC curve for the CKD dataset where we can see that the curve is in the top left corner which means the performance is better.

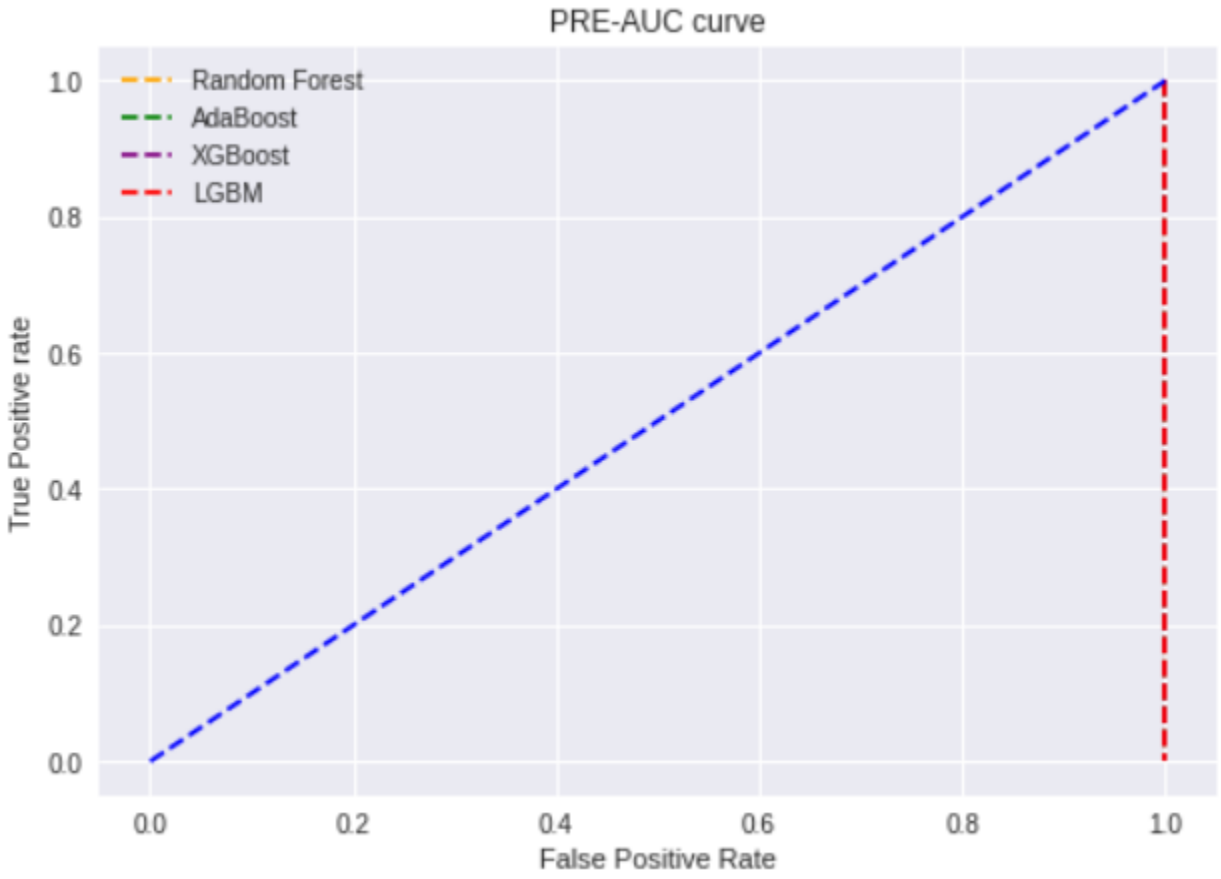


Fig. 8:Precision-recall (PRE)curve for CKD dataset.

For various thresholds, the precision-recall curve depicts the tradeoff among precision and recall. A large area under the curve indicates good recall and precision, with high precision indicating a low false positive rate and high recall indicating a low false negative rate. In figure 7, we can see the precision-recall(PR) curve for CKD dataset. So measuring the performance of four machine learning ensemble classifiers for CKD prediction was very greatly examined. The LGBM classifier shows the best performance above all algorithms.

Chapter 5

Conclusion

CKD nowadays becomes a big threat and an early-stage identification can solve this problem. It will be a big help for saving money and also life. This model also reduces many clinical tests costs. Here, I have used four different types of algorithms like the random forest, LGBM classifier, XGBoost, AdaBoost. Comparing these ensemble technique algorithms we can see that LGBM achieved the highest accuracy which is 1.0 and XGBoost classifier and Random Forest algorithm's accuracy are 0.99 and AdaBoost algorithm's accuracy level is lower than others which is 0.98. In the biomedical sector, machine learning has a great achievement in clinical aid in detecting Chronic Kidney disease. The LGBM developed model is more dependable than above. I think this model can bring a bright future for predicting CKD. In the future, a huge amount of analytical and agent-based data will be collected to develop the model for increasing the performance of prediction while allowing it to distinguish the symptoms of the condition. We recognize that the number and quality of the data in this model will increase and in the future, another model can be developed with more analytical data.

Chapter 6

Reference

[1] Revathy, S., Bharathi, B., Jeyanthi, P., & Ramesh, M. (2019). Chronic kidney disease prediction using machine learning models. . *International Journal of Engineering and Advanced Technology (IJEAT)*, *9*(1).

[2] Samatha, K., Reddy, M. R., Khan, P. F., Chowdary, R. A., & Rao, P. P. Chronic Kidney Disease Prediction using Machine Learning Algorithms.

[3] Scholar, P. G. (2018). Chronic kidney disease prediction using machine learning. *International Journal of Computer Science and Information Security (IJCSIS)*, *16*(4).

[4] Chittora, P., Chaurasia, S., Chakrabarti, P., Kumawat, G., Chakrabarti, T., Leonowicz, Z., ... & Bolshev, V. (2021). Prediction of chronic kidney disease-a machine learning perspective. *IEEE Access*, *9*, 17312-17334.

[5] Abinaya, U., Devi, S. A., Haritha, B., & Raghunathan, T. (2021, May). Noval Approach For Chronic Kidney Disease Using Machine Learning Methodology. In *Journal of Physics: Conference Series* (Vol. 1916, No. 1, p. 012164). IOP Publishing.

[6] Rashed-Al-Mahfuz, M., Haque, A., Azad, A., Alyami, S. A., Quinn, J. M., & Moni, M. A. (2021). Clinically Applicable Machine Learning Approaches to Identify Attributes of Chronic Kidney Disease (CKD) for Use in Low-Cost Diagnostic Screening. *IEEE Journal of Translational Engineering in Health and Medicine*, *9*, 1-11

[7] Roy, M. S., Ghosh, R., Goswami, D., & Karthik, R. (2021, May). Comparative Analysis of Machine Learning Methods to Detect Chronic Kidney Disease. In *Journal of Physics: Conference Series* (Vol. 1911, No. 1, p. 012005). IOP Publishing.[2] Samatha, K., Reddy, M. R., Khan, P. F., Chowdary, R. A., & Rao, P. P. Chronic Kidney Disease Prediction using Machine Learning Algorithms.

[8] Ekanayake, I. U., & Herath, D. (2020, July). Chronic Kidney Disease Prediction Using Machine Learning Methods. In *2020 Moratuwa Engineering Research Conference (MERCon)* (pp. 260-265). IEEE.[4] Chittora, P., Chaurasia, S., Chakrabarti, P., Kumawat, G., Chakrabarti, T., Leonowicz, Z.,... & Bolshev, V. (2021). Prediction of chronic kidney disease-a machine learning perspective. *IEEE Access*, *9*, 17312-17334.

[9] Almasoud, M., & Ward, T. E. (2019). Detection of chronic kidney disease using machine learning algorithms with least number of predictors. *International Journal of Soft Computing and Its Applications*, *10*(8).

[10] Ghosh, P., Shamrat, F. J. M., Shultana, S., Afrin, S., Anjum, A. A., & Khan, A. A. (2020, November). Optimization of Prediction Method of Chronic Kidney Disease Using Machine Learning Algorithm. In *2020 15th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)* (pp. 1-6). IEEE.

[11] Kellum, J. A., Romagnani, P., Ashuntantang, G., Ronco, C., Zarbock, A., & Anders, H. J. (2021). Acute kidney injury. *Nature Reviews Disease Primers*, *7*(1), 1-17.

[12] National Kidney Foundation. Chronic Kidney Disease (CKD) Symptoms and Causes. [Online]. Available: https://www.kidney.org/atoz/content/about-chronic-kidney-disease . Access Date: 11-12-2021.

[13] Ishani, A., Xue, J. L., Himmelfarb, J., Eggers, P. W., Kimmel, P. L., Molitoris, B. A., & Collins, A. J. (2009). Acute kidney injury increases risk of ESRD among elderly. *Journal of the American Society of Nephrology*, *20*(1), 223-228.

[14] Kaggle. Chronic Kidney Disease. [Online]. https://www.kaggle.com/abhia1999/chronic-kidney-disease . Access Date: 12-12-2021

[15] National Institute of Aging. High Blood Pressure and Older Adults. [Online]. https://www.nia.nih.gov/health/high-blood-pressure-and-older-adults . Access Date: 12-12-2021

[16] Cleveland Clinic. High Red Blood Cell Count. [Online]. https://my.clevelandclinic.org/health/symptoms/17810-high-red-blood-cell-count . Access Date: 13-12-2021

[17] Medline Plus. White Blood Count (WBC). [Online]. https://medlineplus.gov/lab-tests/white-blood-count-wbc/ . Access Date: 13-12-2021

[18] Drüeke, T. B., Locatelli, F., Clyne, N., Eckardt, K. U., Macdougall, I. C., Tsakiris, D., ... & Scherhag, A. (2006). Normalization of hemoglobin level in patients with chronic kidney disease and anemia. *New England Journal of Medicine*, *355*(20), 2071-2084.

[19] Zheng, C. M., Ma, W. Y., Wu, C. C., & Lu, K. C. (2012). Glycated albumin in diabetic patients with chronic kidney disease. *Clinica chimica acta*, *413*(19-20), 1555-1561.

[20] Educba. AdaBoost Algorithm. [Online]. https://www.educba.com/adaboost-algorithm/ . Access Date: 13-12-2021

[21] Machine Learning Mastery. A Gentle Introduction to XGBoost for Applied Machine Learning. [Online]. https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/ Access Date: 13-12-2021