Chronic Kidney Disease Prediction Model Based on Machine Learning Technique

Anoop Kumar Kushwaha[1], Anurag Shrivastava[2], Vaibhav Patel[3]

[1]*MTech Scholar,* [2,3]*Assistant Professor*

CSE, Department, NIIRT Bhopal

Abstract - In India and all over the world chronic kidney disease (CKD) is an increasing and serious disease impacting public health. The symptoms of CKD are often appearing too late and many patients inevitably face pain and expensive medical treatments. The ultimate treatment is frequent dialysis or Kidney transplant. Early detection of disease through symptoms can prevent the disease progression by referral to appropriate health care services. Now days, Machine Learning (ML) techniques have been widely used in healthcare sector. ML techniques can help in identifying the potential risk by discovering knowledge from medical reports of patient. Thus helps in preventing the disease progression. Several models for detecting the risk of CKD, proposed in the literature are based on Data Mining (DM) techniques. The use of cloud platforms for data mining tasks is new research area in the field of ML. Several literature and use of platforms confirms a considerable performance improvement in running ML tasks. This research aims at developing a Cloud based Predictive Model to detect the possibilities of CKD and its progression in patients with some health issues like hypertension and diabetes. The proposed model can help physicians to recognize patients at risk and prescribe the treatments and lifestyle changes. The model is trained and tested on the CKD data provided on UCI repository and it is deployed on Microsoft Azure ML platform. The model is built using two algorithms: Two classes Boosted decision tree and two class deep support vector machines. The model built over boosted decision tree algorithm has highest prediction accuracy. The model can also be used to test and predict risk on any unknown data. The results obtained from proposed model are compared with basic benchmark classifiers. Furthermore, the experimental evaluation shows that the proposed work has the potential to get better prediction accuracy to deal with disease risks. The work provides potential and future research directions for predicting the risk of other such diseases.

Keywords— Boosted Decision Tree, Neural Network, Chronic Kidney Disease (CKD), Data Analytics, Health Care, Microsoft Azure, Machine Learning, Prediction Model.

I. INTRODUCTION

The health care sector of the world is the most promising sector nowadays, due to pandemic. Medical Instruments, Medicines, hospital facility, and other required things are lacking everywhere. In this bad situation, it is required to enhance medical technology and its related research. Chronic kidney disease (CKD) poses a serious burden of disease worldwide with substantially increasing number of patients being diagnosed. According to the study [1] published in 2018 Global Burden of Disease study data and methodologies to describe the change in burden of CKD from 1990 to 2016 involving incidence, prevalence, death, and disability-adjusted-life-years (DALYs). Globally, the incidence of CKD increased by 89% to 21,328,972, prevalence increased by 87% to 275,929,799, death due to CKD increased by 98% to 1,186,561, and DALYs increased by 62% to 35,032,384. Figure 1.1 showing the Google search results for CKD [2]. Also, the cost related to CKD care is too high. So early detection and identification of patients with increased risk of developing CKD on the basis of symptoms can improve care by preventive measures to slow disease progression and timely initiation of nephrology care.

CKD is a known common disease, seen by the nephrologists, specialists and practitioner in other fields also.

Enormous complex data is being regularly received by healthcare division about diseases, treatment, patients, medical equipments, hospitals and claims etc. The data requires processing for extraction of knowledge. Data Mining is predominantly helpful in healthcare domain when it is difficult to deal a disease with particular treatment option. DM comprise of efficient techniques and tools to apply on healthcare data for making appropriate decisions towards taking preventive measures and predicting risks of disease.

II. LITERATURE REVIEW

The research paper [1] authors employ experiential analysis of ML techniques for classifying the kidney patient dataset as CKD or NOTCKD. Seven ML techniques together are utilized and assessed using distinctive evaluation measures such as mean absolute error (MAE), root mean squared error (RMSE), relative absolute error (RAE), root relative squared error (RRSE), recall, precision, F-measure and accuracy. The experimental outcomes accomplished of MAE are 0.0419 for NB, 0.035 for LR, 0.265 for MLP, 0.0229 for J48, 0.015 for SVM, 0.0158 for NB Tree and 0.0025 for CHIRP. Moreover, experimental results using accuracy revealed 95.75% for NB, 96.50% for LR, 97.25% for MLP, 97.75% for J48, 98.25% for SVM, 98.75% for NB Tree,

and 99.75% for CHIRP. The overall outcomes show that CHIRP performs well in terms of diminishing error rates and improving accuracy.

In paper [2] prediction is performed using Naive Bayes Classifier and K-Nearest Neigbour algorithm. The data used is collected from the UCI Repository with 400 data sets with 25 attributes. This data has been fed into Classification algorithms. The experimental results show that Naïve Bayes Algorithm gives an accuracy of 96.25%, whereas KNearest Neigbour came up with an accuracy of 100%.

In this paper [3], we conduct the statistical analysis, Machine Learning (ML) and Neural Network application on clinical data set of Uddanam CKD for prevention and early detection of CKD. As per statistical analysis we can prevent the CKD in the Uddanam area. As per ML analysis Naive Bayes model is the best where the process model is constructed within 0.06 seconds and prediction accuracy is 99.9%. In the analysis of NNs, the 9 neurons hidden layer (HL) Artificial Neural Network (ANN) is very accurate than other all models where it performs 100% of accuracy for predicting CKD and it takes the 0.02 seconds process time.

This paper review paper [4] the machine learning (ML), deep learning (DL) and other intelligent cloud based diagnosis models for the three diseases. A detailed introduction to the cloud based healthcare system is also provided. Besides, the reviews of the techniques are made with respect to aim, underlying technique, diagnosed disease and experimental results. At the end of the survey, a detailed comparative study has also been made to examine the possible future work for the readers.

The authors [12] synthesized systematic reviews of risk prediction models for CKD and externally validated few models for a 5-year scope of disease onset. Authors worked on ~234 k patients' data of UK. Seven relevant CKD risk prediction models were identified. All models distinguished well between patients developing CKD or not, with Receiver Operating Characteristic curve (ROC) around 0.90. But, it is concluded that most of the models were poorly calibrated and substantially over-predicting the risk.

The authors [13] predicted CKD using two classification techniques: Naive Bayes and Artificial Neural Network (ANN). The experiment is conducted using Rapidminer tool over dataset containing 400 instances with 25 attributes including class. The dataset from UCI repository [4] is used. The results [27] revealed that Naive Bayes produced more accurate results than ANN.

In study [14] CKD is diagnosed with Adaboost Ensemble Learning (EL) method. For diagnosis Decision tree based classifiers is used. The classifier performance is evaluated using several metrics including area under curve (AUC).The main observation of paper [5] is that Adaboost EL method provides better performance than individual classification. The dataset from UCI repository [4] is used.

III. PROPOSED PREDICATION MODEL

In this research work a cloud based prediction model is proposed, to detect the possibilities of CKD and its progression in patients with some health issues like hypertension and diabetes, is proposed and implemented. The models are trained and tested on the CKD data provided on UCI repository [8] and it is deployed on Microsoft Azure ML platform. The proposed model offered in this work (refer Figure 3.1), actually employs Two class boosted decision tree and Two class deep support vector machine learning algorithm. The model built over Boosted decision tree algorithm has highest prediction accuracy. The model can also be used to test and predict risk on any unknown data. For faster evaluation and lesser overall time cloud platform is used. The dataset requires pre-processing for converting it into a suitable format for getting highly accurate results within smaller time. The pre-processing methods affect a lot in final evaluation results of ML model. It is a good practice to apply such processes on raw data. After applying suitable pre-processing techniques, a method is applied to overcome the missing values. The 'Missing Value Scrubber' is applied to deal with missing values.

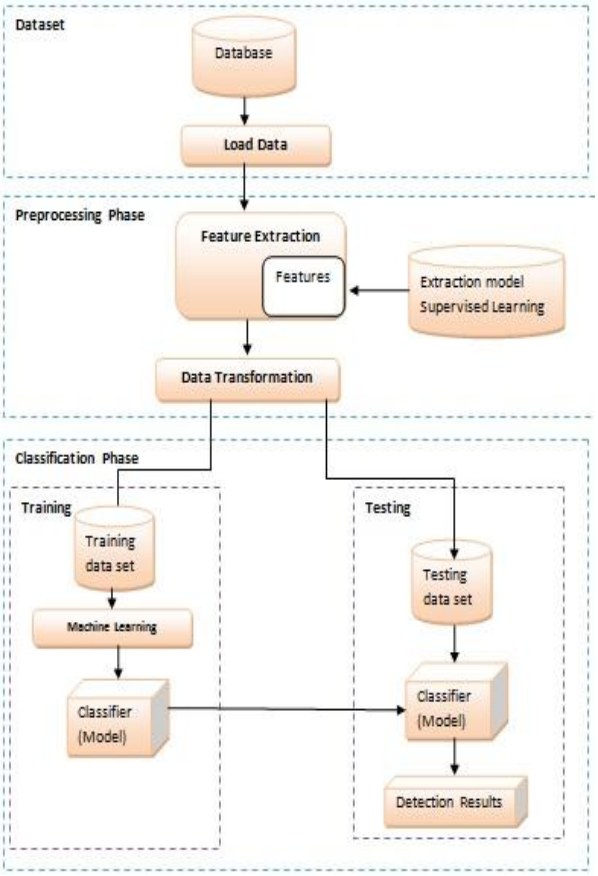


Figure: 3.1

In the next step dataset is split into two subsets known as training and testing set. Generally a small part of dataset is chosen to train the classifier/model. The ratio 40: 60 i.e. train: test is used for this work.

To build the model various ML algorithms are applied and tested iteratively in the next step and best model is determined. The ML methods involving mathematical models and statistical analysis like regression analysis or more complex approaches like Decision Trees and Neural Network algorithm to the data are to be applied to fulfil the purpose of Prediction. The best model based on ML methods is preferred by data scientist to decide many aspects to generate more useful results.

In the proposed Prediction model, the Boosted Decision Tree Algorithm along with other Modules is applied for better Prediction accuracy and faster evaluation. Application of Boosted Decision Tree Algorithm provides better data classification better Prediction accuracy than other models like LR. The Prediction classifier (model) is deployed and tested using test set.

IV. DATASET DESCRIPTION

In various research papers authors used the dataset for experiment purpose. CKD dataset from UCI ML repository [8] is used. The dataset includes 400 instances with 24 attributes and a class attribute. The CKD dataset used in this study is taken from the UCI Machine Learning Repository [8]. The data was donated by Soundarapandian et al. and collected for nearly 2-month period. The dataset comprise of 400 samples represented by 11 numeric and 10 nominal attributes and a class descriptor which is also nominal. Out of 400 samples, 250 samples belong to the CKD group, and the other 150 samples belong to the non-CKD group. Details are given in table 4.1

Table 4.1.

Description	***Details***
Dataset Name	*Chronic Kidney Disease (CKD)*
Source	*UCI Machine Learning Repository*
No. of Instances	*400*
No. Of attributes	*24*
Classes	*{Ckd, Notckd}*
Class Distribution	*Ckd 250*
	Notckd 150
Missing Values	*Yes*

V. EXPERIMENT SETUP

The experiment has done on the Microsoft azure cloud platform. Microsoft provides ML Workspace with ML studio, ML Gallery and ML Web Service Management. The studio is a graphical tool for ML tasks from start to finish. It includes: various data pre-processing modules; a bunch of ML algorithms; An API to access model deployed on Azure. The Studio enables import of datasets, pre-processing methods, ML techniques and more onto its design interface.

Figure: 5.1 Prediction model on Azure ML studio

Before executing chief experimental steps the dataset is uploaded. The main experimental steps that are followed are given in below and represented in Figures 5.1

1. Pre-process the dataset. The step is discretionary if data is already pre-processed. Here missing values in the datasets are handled.

2. Split dataset into training (40%) and testing (60%) set. the split is of 40% for training and 60 % for testing.

3. In train model we need to set the target attribute in column selector so that the datasets can be trained and classification can be done on basis of selected attribute.

4. Apply ML Algorithm to Train the model. This process is repetitive as we have to discover the best ML algorithm for certain tasks. In this work Two Class Boosted Decision Tree and Two classes neural network algorithm is applied on full dataset.

5. Tune model module is required to tune the parameters of algorithms. This is optional for some algorithms/models.

6. Apply "Score Model" module, to Score both the Classifiers built. Test dataset is supplied to this module, so that the model built will classify the data instances. The proposed Model and model based on MLR are scored with standard metrics.

7. Apply "Evaluate Model" module to compare both the models. To speed up the process of evaluation various scripts can be applied. The results visualization can be done in various ways including confusion matrix.

8. Run the experiment. The proposed model has been run properly and performing better in terms of accuracy.

9. The evaluated results are compared and analysis is presented in result analysis. The results show that the proposed model is better in terms of overall accuracy

precision and other parameters also including False Positive rate.

VI. RESULT ANALYSIS

In this research work the evaluation metrics that taken into account are Model's 'Prediction Accuracy' and Precision. The proposed models are achieving good prediction accuracy and precision. Among both the models Boosted decision tree algorithm is performing well as shown in Table 6.1

Table 6.1: Results: Prediction Accuracy and Precision

Models	Prediction Accuracy	Precision
Boosted Decision Tree	*100*	*0.01*
Neural Network	*97.9*	*0.94*

Figure 6.1 show the confusion matrix for boosted decision tree algorithms

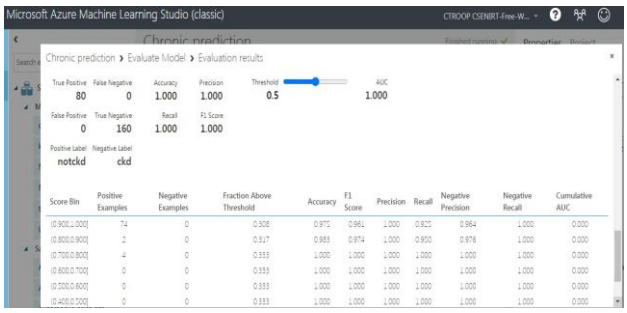


Figure 6.1 Confusion Matrixes for Boosted Decision Tree

Graphical representation of Accuracy and Precision is shown in Figure 6.2

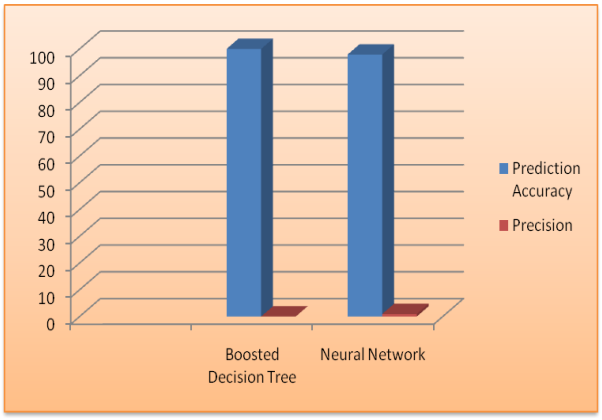
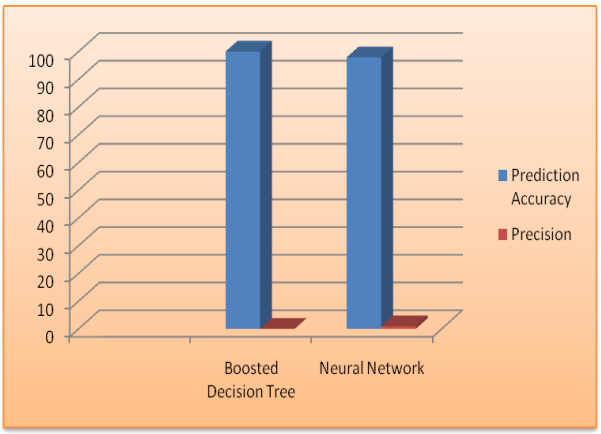


Figure 6.2 : Accuracy and Precision Graph

The confusion matrix actually demonstrates 04 parameters: True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN). Through Confusion Matrix, it is evident the FP rate is also less with the proposed model. False Positive rate signifies the rate of identification of normal patients as suffered from CKD. Figure 6.3 represent the graph between false positive and true negative.



Result 6.3: Graph of False Positive and True Negative

VII. CONCLUSION

The main contributions of this research are: A generalized prediction model for CKD risk detection is proposed. The implementation of proposed prediction model using Azure (Machine Learning) platform is demonstrated. The model is having utility in healthcare domain. Data Mining, CC, Machine Learning (ML), and DM over Cloud are discussed. The importance of Prediction analysis in dreaded disease prediction is studied and requirement of cloud platforms for investigating big datasets is established. The proposed model is independent of computational resource limitations. Few Cloud based ML frameworks for disease data classification is surveyed. Microsoft Azure based ML Environment is studied and explained briefly. The proposed cloud based Prediction model based on 'Boosted Decision Tree' and is compared with a benchmark model 'Neural Network' and evaluated for its effectiveness in terms of class wise predicted Classification Accuracy. Nowadays data comes with three attributes: Volume, Velocity and Variety and such data is termed as Big Data. The number of issues arises while dealing with huge amount of data is scaling and reliable analysis of data. Also, the ML algorithms do not scale well on single system, whereas cloud platforms provide scope to scale the algorithms on big data also. The work proposed here can provide potential approach for addressing multi-class classification problems. The proposed Risk Prediction model can be extended as clinical screening toolkit for early prediction of CKD to check the progression of disease for following proper preventive measures. Several classification and prediction models have been proposed to check CKD onset, but either they are not utilized well or performs over-prediction. This work demonstrates the CKD risk prediction in people. The proposed work will definitely provide significant insight into risk prediction for other diseases also. The model can be further tested for

more parameters and extended for batch prediction by supplying huge dataset.

REFERENCES

[1] Bilal Khan 1, Rashid Naseem 2, Fazal Muhammad 3,Ghulam Abbas 4, (Senior Member, IEEE), and Sunghwan Kim 5 "An Empirical Evaluation of Machine Learning Techniques for Chronic Kidney Disease Prophecy" Received March 2, 2020, accepted March 14, 2020, IEEE 2020

[2] 1Sujata Drall, 2Gurdeep Singh Drall, 3Sugandha Singh, 4Bharat Bhushan Naib "Chronic Kidney Disease Prediction Using Machine Learning: A New Approach" International Journal of Management, Technology And Engineering ISSN NO : 2249-7455 Volume 8, Issue V, MAY/2018

[3] K. B. Anusha, T. PanduRanga Vital, K. Sangeeta "Machine Learning Models and Neural Network Techniques for Predicting Uddanam CKD" International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-2, July 2019

[4] R. Raja1 and B. Ashok2 "An Effective Analysis on Cloud Computing in Healthcare Diagnosis and Prediction Systems" International Journal of Advanced Science and Technology Vol. 29, No. 03, (2020), pp. 12016 - 12029

[5] Jameson K, Jick S, Hagberg KW, Ambegaonkar B, Giles A, O'Donoghue D., "Prevalence and management of chronic kidney disease in primary care patients in the UK". Int J Clin Pract. 2014;68 (9):1110–21.

[6] www.google.co.in/search?q=Chronic+kidney+disease

[7] A. S. Levey, K. Eckardt, U. Tsukamoto, A. Levin, J. Koresh, J. Rossert, D. D. Zeeuw, T. H. Hostetter, N. Lameire and G. Eknoyan, "Definition and classification of chronic kidney disease: A position statement from Kidney Disease: Improving Global Outcomes (KDIGO)," Kidney International, Vol. 67, pp. 2089-2100, 2005.

[8] P. Soundarapandian and L. J. Rubini, http://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease, UCI Machine Learning Repository, Irvine, 2015.

[9] Collins GS, Omar O, Shanyinde M, Yu L-M. A systematic review finds prediction models for chronic kidney disease were poorly reported and often developed using inappropriate methods. J Clin Epidemiol. 2013; 66(3):268–77.

[10] Echouffo-Tcheugui JB, Kengne AP. Risk models to predict chronic kidney disease and its progression: a systematic review. Remuzzi G, editor. PLoS Med. 2012; 9(11):e1001344.

[11] J. K. Han, Micheline, Data mining: concepts and techniques: Morgan Kaufmann, 2001.

[12] I. H. a. F. Witten, Eibe Data Mining: Practical machine learning tools and techniques: Morgan Kaufmann 2005.

[13] C. M. Bishop, Pattern recognition and machine learning: Springer, 2006.

[14] P. Simon, Too Big to Ignore: The Business Case for Big Data: John Wiley & Sons, 2013.

[15] Paolo Fraccaro, Sabine van der Veer, Benjamin Brown, Mattia Prosperi, Donal O'Donoghue, Gary S. Collins, Iain Buchan and Niels Peek, "An external validation of models to predict the onset of chronic kidney disease using population-based electronic health records from Salford, UK", RESEARCH ARTICLE, BMC Medicine (2016) 14:104.

[16] V. Kunwar, K. Chandel, A. S. Sabitha and A. Bansal, "Chronic Kidney Disease analysis using data mining classification techniques," 2016 6th International Conference - Cloud System and Big Data Engineering (Confluence), Noida, 2016, pp. 300-305.

[17] M. D. Başar, P. Sarı, N. Kılıç and A. Akan, "Detection of chronic kidney disease by using Adaboost ensemble learning approach," 2016 24th Signal Processing and Communication Application Conference (SIU), Zonguldak, 2016, pp. 773-776.

[18] Z. Sedighi, H. Ebrahimpour-Komleh and S. J. Mousavirad, "Featue selection effects on kidney desease analysis," 2015 International Congress on Technology, Communication and Knowledge (ICTCK), Mashhad, 2015, pp. 455-459.

[19] Sumit Basu, "Empirical Results on the Generalization Capabilities and Convergence Properties of the Bayes Point Machine", Technical Report, December, 1999