ISSN: 2581-3404 (Online)
International Journal of Innovative Research in Technology and Management, Vol-4, Issue-5, 2020.

Predicting Chronic Kidney Disease Risk Using Recursive Feature Elimination and Machine Learning

Arpit Saxena[1], Amit Ganguly[2], Ajit Kumar Shrivastava[3]

Department of CSE, Sistech-R, Bhopal, (M.P.), India[1,2,3]

ABSTRACT

The Chronic Kidney Disease (CKD) is a global public health issue with a growing incidence, pervasiveness, and high cost. To turn into more focus on real insinuation of CKD and health issues coupled with CKD patients, application of Machine Learning (ML) models becomes necessary. The chief objective of this research work is to evolve a better ML classifier framework for predicting possibility of CKD and its progression in patients with health issues like diabetes and hypertension. The early diagnosis can prevent disease progression and severity through suitable preventive measures and thus reduces treatment cost. In this work, framework for CKD risk prediction is proposed which is based on ranking of features done using Recursive Feature Elimination (RFE) method. The proposed framework employs RFE for eliminating the unrelated features from huge dataset of patients. The elimination of irrelevant features reduces the data to be considered and speeds up the execution of ML algorithms.

The classifier is built using CKD data provided on UCI repository. The model is built using Python based ML libraries. The RFE method increases the performance of classifier based on Logistic Regression (LR). The proposed predictive framework achieved highest prediction accuracy. The framework also generates and alert for patient at risk and can also be tested to predict risk on any new patient.

The framework is implemented using Python ML and data processing and manipulation libraries: Scikit-learn, Numpy and Pandas on Anaconda Navigator IDE. This work presents a positive attempt to predict the risk of a chronic disease. Also, a study is done to know how attribute selection techniques affect the classification accuracy. To evaluate the efficacy of proposed system, three most popular and effective metrics for healthcare data are considered. The metrics are: Accuracy, Precision and F1-Score.

Keywords: Chronic Kidney Disease, Data Mining, Feature Selection, Healthcare, Machine Learning, Predictive Model, Recursive Feature Elimination.

INTRODUCTION

Chronic Kidney Disease (CKD) is a quiet disease condition whose symptoms and signs appear almost too late and are generally nonspecific to reveal an indication for diagnosis or severity of the condition. The primary task of kidney is to filter out waste products developed due to metabolic activities. CKD is a common medical problem in which functioning of kidney is reduced over time. CKD advances many diseases, which in turn, can cause even renal failure. To turn into more focus on real insinuation of CKD and health issues coupled with CKD patients, application of Machine Learning (ML) models becomes necessary.

CKD is a global public health issue with a growing incidence, pervasiveness, and high cost. Around 2.5-11.2% of the population across Asia, Europe, Australia and North America are reported to have CKD [1]. In USA itself disease has affected around 27 million individuals [2]. According to National Kidney Foundation around 59% of all American citizens are at risk of developing CKD in their lifetime [3]. Since last few years, more than 10 million cases every year reported in India. Diabetes mellitus and Hypertension are the major risk factors for CKD prevalence.

Recent studies imply that some of the adverse effects can be delayed or even prevented by early diagnosis and healing [4, 5]. Awareness of CKD in patients is regularly increasing, but still low. More often disease is diagnosed too late when dialysis is needed urgently. Thus caregivers have huge burden of CKD management. Data Mining (DM) methods and technologies aids this situation by discovering hidden relationships and patterns in medical data. For prevention, these methods generally treat the problem as classification problem. Classification is the process of classifying the disease and not disease condition. The predictive models on the other hand identify whether in future a patient face the condition of disease or not.

II RELATED WORK

Salekin and Stankovic [17] built a ML classifier by considering 24 predictive parameters present in famous UCI CKD dataset [7]. They perform feature selection to find out the most relevant attributes for CKD detection according to their predictability. Authors also performed cost-accuracy tradeoff analysis for new CKD detection approach. The detection accuracy achieved is nearly 99% with 12 attributes.

Authors [18] presented a extended survey on the use of feature selection and classification approaches for prediction of chronic diseases. Broad overview of attribute selection methods with their pros and cons is presented. Adaptive and parallel classification systems are also analyzed in this work.

The paper [19] briefly discusses the researches done in this area. An insight on how Health care data can be analyzed with Hadoop/Map Reduce using predictive analytics.

The paper [20] data mining classifiers based prediction for CKD using KNN (K-Nearest Neighbor) and SVM (Support Vector Machine). Authors utilized MATLAB and Hadoop to get accurate results based on the input parameters present in given data set.

Pinar Yildirim [21] considered Back propagation networks for his research. A comparative study of few sampling algorithms was performed based on Multi Layer Perceptron (MLP) for prediction of CKD. This study tells that sampling algorithms can improve the performance of algorithms for classification. Also, learning rate is a vital metric that can significantly performance of MLP.

Authors [22] investigated 4 ML classifiers: SVM, KNN, LR, and decision tree. The performance of these classifiers is evaluated and compared together in order to choose the best model for CKD prediction.

The authors [23] 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 [7] is used. The results [23] revealed that Naive Bayes produced more accurate results than ANN.
The authors [24], employs the fact of dimensionality reduction (feature selection) that improves computation performance of classifiers and produces classified models rapidly. Feature selection makes it popular in DM and ML techniques. In the work, authors employed few such methods followed by ML techniques to classify CKD. It is show that feature selection

ISSN: 2581-3404 (Online)
International Journal of Innovative Research in Technology and Management, Vol-4, Issue-5, 2020.

techniques enables precise classification in least time.

III PROPOSED WORK

The proposed framework is described as five phase model. The framework is depicted in Figure 1. The five phases involves various sub-phases such as pre-processing and application of feature selection technique. The five phases are

- Data Collection Module (DCM)
- Data Processing Module (DPM)
- Model Building Phase (Learning)
- Test (Prediction) Phase
- Evaluation.

Data Collection Module: The outcome of this phase is a file that will pass to next module.

Data Processing Module (DPM) The data received in this module may involve various inconsistencies and noise. These inconsistencies and noise definitely differs classification and prediction accuracies. The major sub-phases of DPM are: Normalization, Transformation and Feature Selection. Feature Selection technique used here is Recursive Feature Elimination.

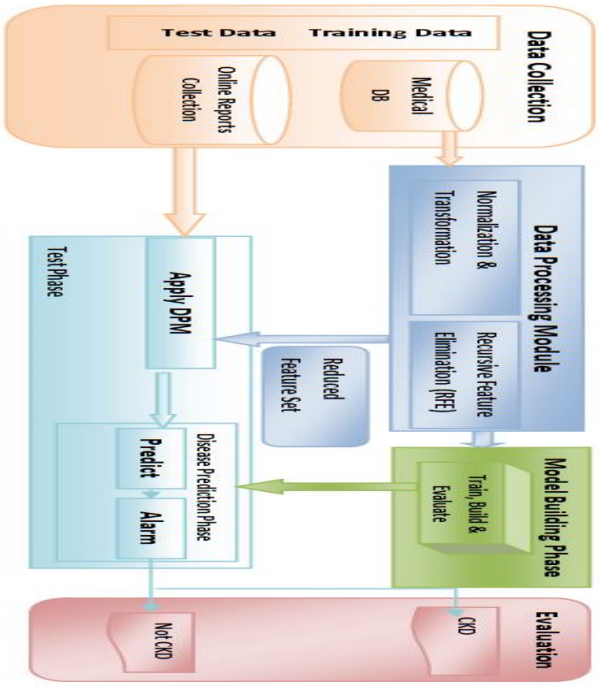

Figure 1: Proposed Framework for CKD Risk Prediction.

Model Building Phase (Learning): The feature selection applied here is RFE. The cycle shown in Figure 2, will continue till a best model is obtained.

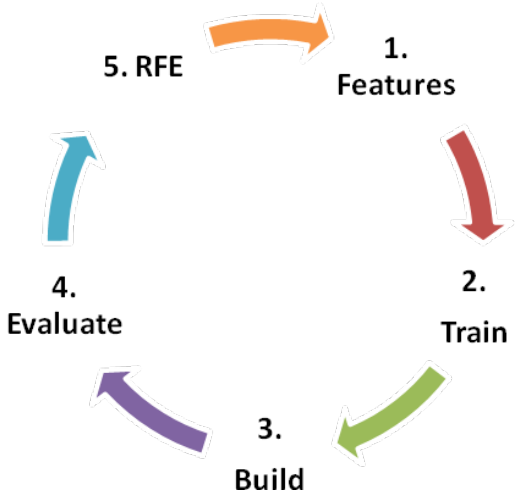

Figure 2: Model Building Cycle.

Test Phase: This phase is having two sub-phases:

- DPM
- Disease Prediction Phase

This phase predicts the class of unseen instance. The phase generates an alarm on the basis of score or probability determined. The alarm indicates that the patient is having CKD Risk or not having risk.

Evaluation: This phase evaluates the proposed model and computes various metrics.

IV RESULT ANALYSIS

The system hardware for experimentation comprise of Intel Core i3 (2.0 GHz.) processor with 8GB RAM and sufficient hard disk space. The system is running (64-bit) operating system: Microsoft Windows – 10. The environment, Tools and Libraries used for experiment are presented in Table 1.

For experimentation purpose CKD dataset donated by Soundarapandian et al. is downloaded from UCI ML repository [7]. The dataset includes 400 instances with 24 attributes and a class attribute. The description of dataset is shown in Table 2. To evaluate the proposed Predictive Classifier for determining CKD against the benchmark classifier

Python ML libraries are used. Simulation is performed with proposed classifier and then with other classifiers.

Table 1: Environments, Tools and Libraries

Environments / Tools / Libraries	Version / Configuration
Anaconda Navigator (GUI)	v_5.3.1; x86_64 bit
Jupyter Notebook [28]	v_5.7.2; 64 bit
Python	3.7.0 and 3.6.6
Scikit-Learn [29]	0.20.2
SciPy [30]	1.1.0
NumPy [31]	1.15.4
Pandas [32]	0.23.4
Seaborn and Matplotlib	0.9.0 and 3.0.2 respectively

Table 2: Dataset Description

Description	Details
Source	UCI-ML Repository [7]
Dataset Name	Chronic Kidney Disease (CKD)
Number of Instances	400
Number of attributes	24
Classes	{Ckd, Notckd}
Missing Values	Yes
Class Distribution	Ckd 250 Notckd 150

Evaluation Parameters:

Prediction Accuracy: Prediction accuracy or rate refers to the percentage of correct predictions among all test data.

Precision: attempts to answer that what proportion of true identifications was actually true?

F1 Score: It is weighted average of Precision and Recall. In case of healthcare data F1 score is useful measure as FP and FN have equal cost.

The parameters are mostly evaluated from confusion matrix.

TP	FP
FN	TN

True Positive (TP): TP means correct predictions (outcomes) of the positive class.

False Positive (FP): FP means incorrect prediction of positive class.

True Negative (TN): TN means correct predictions (outcomes) of the negative class.

False Negative (FN): FN means incorrect predictions (outcomes) of negative class.

The results in tabular format is presented in Table 3

Table 3: Results Comparison

Models	Prediction Accuracy	Precision	F1 Score
LR	97.5	0.93	0.97
SVM	97	0.97	0.85
Proposed	99.5	0.99	1

V CONCLUSIONS

Predictive analytics in healthcare using ML is among new confronts. It helps doctors to predict the cause and disease in early stage of human's life. Based on diagnostics doctors and caregivers can suggest exact treatments and care environment for saving lives. The imbalance, non-availability of real datasets, volume, variety and velocity of medical data sets pose a serious challenge in front of researchers in this field. Also, an extensive knowledge is required to deal in this domain. The major challenge is extraction of right features from medical data and determining disease risk. Apart from accuracy the cost of misclassification is also higher.

The application of SVM, NN and many popular DM algorithms involves a lot of pre-processing, sampling etc. and time. The application of better RFE method and LR gave improved performance and reduces the amount of effort in preprocessing prior to building the model. This work proposed framework for Predicting CKD Risk using RFE feature selection and ML. The approach determined the minimum number of features that are highly correlated with target class on the basis of ranking. The RFE determined the optimal number of features and their ranks. The proposed work is implemented using popular ML libraries: Scikit-learn, Scipy, Numpy and Pandas under Python data science. The proposed work is evaluated and compared with two benchmark models in the literature. The effectiveness is validated against LR and SVM models. The work progresses and achieved the objectives mentioned earlier in this thesis. The important metrics taken for evaluation are: F1-score, Precision and Accuracy. The results have shown improvement up-to 3%.

The results of paper have a set of research proposals. The utilization of efficient feature selection techniques not only reduces the computational efforts but boosts the performance of classification and prediction models. In future, more efforts must be given to the application of such optimized methods to improve the predictive models. The proposed model would be integrated with live medical data to determine the health conditions of patients in real time. Additionally, one can discover the disease stages and symptoms that are often correlated with one another.

REFERENCES

[1] Q.-L. Zhang and D. Rothenbacher, “Prevalence of chronic kidney disease in population-based studies: systematic review,” BMC public health, vol. 8, no. 1, p. 117, 2008.

[2] M. Baumgarten and T. Gehr, “Chronic kidney disease: detection and evaluation,” American family physician, vol. 84, no. 10, p. 1138, 2011.

[3] V. A. Moyer, “Screening for chronic kidney disease: Us preventive services task force recommendation statement,” Annals of internal medicine, vol. 157, no. 8, pp. 567–570, 2012.

[4] G. Remuzzi, P. Ruggenenti, and N. Perico, “Chronic renal diseases: renoprotective benefits of renin–angiotensin system inhibition,” Annals of internal medicine, vol. 136, no. 8, pp. 604–615, 2002.

[5] N. Tazin, S. A. Sabab and M. T. Chowdhury, "Diagnosis of Chronic Kidney Disease using effective classification and feature selection technique," 2016 International Conference on Medical Engineering, Health Informatics and Technology (MediTec), Dhaka, 2016, pp. 1-6.

[6] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Mach. Learn. 46 (2002), pp. 389–422.

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

[8] http://emedicine.medscape.com/article/238798-overview.

[9] Peng, H.C.; Long, F.H.; Ding, C. Feature selection for high-dimensional data: A fast correlation-based filter solution. In Proceedings of the 20th International Conference on Machine Learning, Washington, DC, USA, 21–24 August 2003.

[10] Mohamed, N.S.; Zainudin, S.; Othman, Z.A. Metaheuristic approach for an enhanced mRMR filter method for classification using drug response microarray data. Expert Syst. Appl. 2017, 90, 224–231.

[11] Kohavi, R.; John, G.Wrappers for feature subset selection. Artif. Intell. 1997, 97, 273–324.

[12] Hui, K.H.; Ooi, C.S.; Lim, M.H.; Leong, M.S.; Al-Obaidi, S.M. An improved wrapper-

based feature selection method for machinery fault diagnosis. PLoS ONE 2017, 12, e0189143.

[13] V.Ganganwa, An overview of classification algorithms for imbalanced datasets, International Journal of Emerging Technology and Advanced Engineering, vol.2,issue 4 2012.

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

[15] Han J, Pei J, Kamber M. Data mining: concepts and techniques. Elsevier; 2011.

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

[17] A. Salekin and J. Stankovic, "Detection of Chronic Kidney Disease and Selecting Important Predictive Attributes," 2016 IEEE International Conference on Healthcare Informatics (ICHI), Chicago, IL, 2016, pp. 262-270.

[18] Divya Jain, Vijendra Singh, Feature selection and classification systems for chronic disease prediction: A review, Egyptian Informatics Journal, Volume 19, Issue 3, 2018, Pages 179-189.

[19] A. Batra, U. Batra and V. Singh, "A review to predictive methodology to diagnose chronic kidney disease," 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, 2016, pp. 2760-2763.

[20] G. Kaur and A. Sharma, "Predict chronic kidney disease using data mining algorithms in hadoop," 2017 International Conference on Inventive Computing and Informatics (ICICI), Coimbatore, 2017, pp. 973-979.

[21] P. Yildirim, "Chronic Kidney Disease Prediction on Imbalanced Data by Multilayer Perceptron: Chronic Kidney Disease Prediction," 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC), Turin, 2017, pp. 193-198.

[22] A. Charleonnan, T. Fufaung, T. Niyomwong, W. Chokchueypattanakit, S. Suwannawach and N. Ninchawee, "Predictive analytics for chronic kidney disease using machine learning techniques," 2016 Management and Innovation Technology International Conference (MITicon), Bang-San, 2016, pp. MIT-80-MIT-83.

[23] 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.

[24] 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.

[25] Michael Collins, Robert E. Schapire, “Logistic Regression, AdaBoost and Bregman Distances”, Machine Learning, 48(1/2/3), 2002, Kluwer Academic Publishers, Manufactured in The Netherlands.

[26] Matthew Bihis, Sohini Roychowdhury, “A Generalized Flow for Multi-class and Binary Classification Tasks: An Azure ML Approach”, Available at: https://arxiv.org/ftp/arxiv/papers/1603/1603.08070.pdf

[27] V. Vapnik, The Nature of Statistical Learning Theory, Springer, New York, 1995.

[28] https://jupyter.org/

[29] https://scikit-learn.org/stable/

[30] https://www.scipy.org/about.html

[31] http://www.numpy.org/

[32] https://pandas.pydata.org/

[33] 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.