INTERNATIONAL RESEARCH JOURNAL OF ENGINEERING & APPLIED SCIENCES
ISSN: 2322-0821(O) VOLUME 6 ISSUE 3
ISSN: 2394-9910(P) Jul 2018 - Sep2018
www.irjeas.org

CHRONIC KIDNEY DISEASE RISK PREDICTION BASED ON MACHINE LEARNING TECHNIQUE USING CLOUD PLATFORM

Sourav Ranjan[1]
MTech Scholar
CSE, Department
NIRT Bhopal

Vaibhav Patel[2]
Assistant Professor,
CSE, Department
NIRT Bhopal

Anurag Shrivastava[3]
Associate Professor & Head
CSE, Department, NIRT 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. These models are demonstrated using variety of languages like Python, Java and tools like Weka. 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, Deep Support vector machine, Chronic Kidney Disease (CKD), Data Analytics, Health Care, Microsoft Azure, Machine Learning, Predictive Model, And Predictive Analytics.**

Introduction

Chronic kidney disease (CKD) poses a serious burden of disease worldwide with substantially increasing number of patients being diagnosed. A 2010 study of 2.8 UK adults reported a 5.9 % prevalence of CKD [1]. In India, more than 10 million cases per year (India) are being reported [2], as per reports collected from various recognized hospitals. 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.

Huge 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.

1. **Literature Review**

International Research Journal of Engineering & Applied Sciences, IRJEAS
www.irjeas.org, ISSN(O): 2322-0821, ISSN(P): 2394-9910, Volume 6 Issue 3, Jul 2018-Sep 2018, Page 32-34

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.

The authors [15], 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 shown that feature selection techniques enables precise classification in least time.

ML algorithms play important role in diagnosis of CKD. On the basis of quantitative and qualitative findings, the authors revealed that the Random Forest (RF) classifier achieves the near-optimal performances on the identification of CKD. The RF based model can also be utilized for diagnosis of similar diseases.

2. PROPOSED FRAMEWORK FOR CLASSIFICATION:

In this research work a cloud based predictive model, 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 [4] and it is deployed on Microsoft Azure ML platform. The proposed model offered in this work (refer Figure 4.1), 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. 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 predictive model, the Boosted Decision Tree Algorithm along with other Modules is applied for better predictive accuracy and faster evaluation. Application of Boosted Decision Tree Algorithm provides better data classification better predictive accuracy than other models like LR. The predictive classifier (model) is deployed and tested using test set.

Figure 2: Model built using Azure ML

4. DATASET: For experimentation purpose CKD dataset from UCI ML repository [4] 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 [4]. 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 more discussed in [8].

5. Experimental Results Analysis and Discussion:

In this 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. The Boosted decision tree method is robust and does not over-fit the training data. Also, it scales well as it is available on Cloud platform.

Conclusion & Future Work

A generalized predictive model for CKD risk detection is proposed. The implementation of proposed predictive model using Azure (Machine Learning) platform is demonstrated. The model is having utility in healthcare domain.The proposed cloud based predictive model based on 'Boosted Decision Tree' and is compared with a benchmark model 'Deep SVM' and evaluated for its effectiveness in terms of class wise Predicted Classification Accuracy. 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] 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.

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

[3] 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.

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

[5] 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.

[6] 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.

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

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

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

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

[11] Microsoft Windows Azure. http://www.microsoft.com/windowsazure/.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

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