IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Volume-6, Issue-4 E-ISSN: 2347-2693

Cloud based Predictive Model for Detection of 'Chronic Kidney Disease' Risk

Stuti Nathaniel[1*], Anand Motwani[2], Arpit saxena[3]

[1]Department of computer science, Sagar Institute of Science & Technology Research, Bhopal
[2]Department of computer science, Sagar Institute of Science & Technology Research, Bhopal
[3]Department of computer science, Sagar Institute of Science & Technology Research, Bhopal

Available online at: www.ijcseonline.org

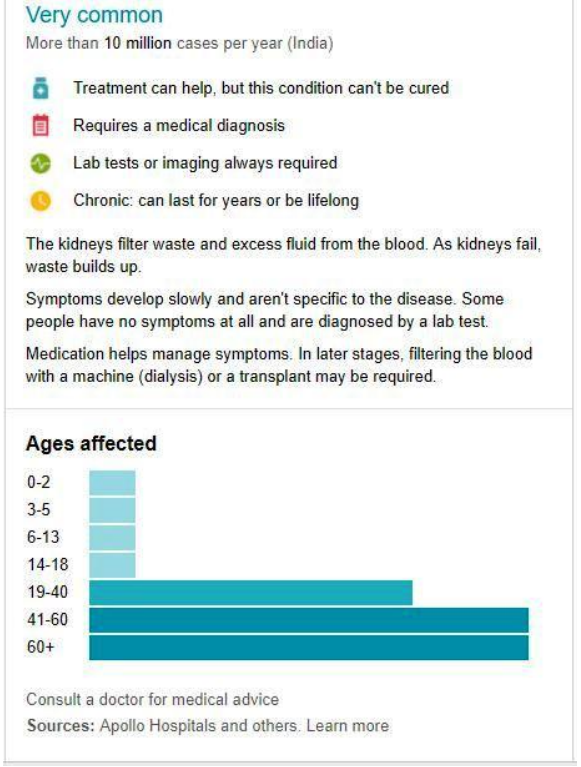
Received: 08/Mar/2018, Revised: 15/Mar/2018, Accepted: 26/Mar/2018, Published: 30/Apr/2018

Abstract- Chronic kidney disease (CKD) is an increasing and serious disease impacting public health worldwide. 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. Machine Learning (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 like classification, clustering and regression etc. These models are demonstrated using variety of languages like Python, Java and tools like Weka and RapidMiner.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.

Keywords: Chronic Kidney Disease (CKD), Health Care, Microsoft Azure, Logistic Regression, Machine Learning, Predictive Model

1. 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 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.



Very common

More than **10 million** cases per year (India)

Treatment can help, but this condition can't be cured

Requires a medical diagnosis

Lab tests or imaging always required

Chronic: can last for years or be lifelong

The kidneys filter waste and excess fluid from the blood. As kidneys fail, waste builds up.

Symptoms develop slowly and aren't specific to the disease. Some people have no symptoms at all and are diagnosed by a lab test.

Medication helps manage symptoms. In later stages, filtering the blood with a machine (dialysis) or a transplant may be required.

Ages affected

Consult a doctor for medical advice

Sources: Apollo Hospitals and others. Learn more

Figure 1: Google Search Results for CKD

Despite various innovations and inventions in health care worldwide, efforts to improve early detection of CKD often remains less futile. Several academicians and researchers in health care division worldwide are working on the detection of CKD problem and trying to develop efficient models to predict and classify the CKD patient, so that the necessary care and preventive measures can be provided to patient. The challenge of detecting CKD risk is handled by DM techniques that are actually based on mathematics and machine learning algorithms. Several models have been developed to predict CKD onset, but most have not been validated outside the setting in which they were developed [3, 4]. We can say that, most of the models only classify the disease as 'CKD' and 'NOT-CKD'. So, there is always a need for predictive tools for early medical diagnosis.

In this research work, we propose and develop 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 models are trained and tested on the CKD data provided on UCI repository [12] and it is deployed on Microsoft Azure ML platform. The model is built with two famous algorithms: Two class Bayes Point and Logistic Regression (LR). Among two, Bayes Point algorithm performed well. The algorithm selection is based on results demonstrated in most of literature that presents the highest accuracy achieved using Naive Bayes (NB) algorithm. Also, in the proposed model Bayes Point Algorithm achieved highest prediction accuracy. The model can also be used to test and predict risk on any unknown data. The proposed model can help physicians to recognize patients' at risk and prescribe the preventive treatments and lifestyle changes.

The rest of this paper is structured as follows. Section 2 discusses ML algorithms used in this work and review of most relevant and recent literature. Sections 3 explain the methodology and proposed model which is build and deployed over Microsoft Azure Cloud. The experimental setup, performance metrics and result discussion is presented in Section 4. Finally, section 5 concludes the research paper.

2. Literature Review

Two Class Bayes Point Machine [5, 6]
It is a Bayesian approach to linear classification. It approximates the optimal Bayesian average of linear classifiers by choosing one "average" classifier known as the Bayes Point. It has several advantages; few of them are listed below:

1. Bayesian classification model are not prone to over-fitting the training data.
2. It does not require data to be normalized.
3. Bayes Point Machine classification models are more robust and easier-to-use.

Logistic Regression [7, 8]
Logistic regression [7, 8] is a well recognized statistical technique that is used for modelling several DM problems. It is a supervised learning method; therefore, we must have to provide a dataset that already contains the class (outcomes) to train the model.

The authors [9] synthesised 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 [10] 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 [12] is used. The results [10] revealed that Naive Bayes produced more accurate results than ANN.

In study [11] 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 [3] is that Adaboost EL method provides better performance than individual classification. The dataset from UCI repository [12] is used.

The authors [13], 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.

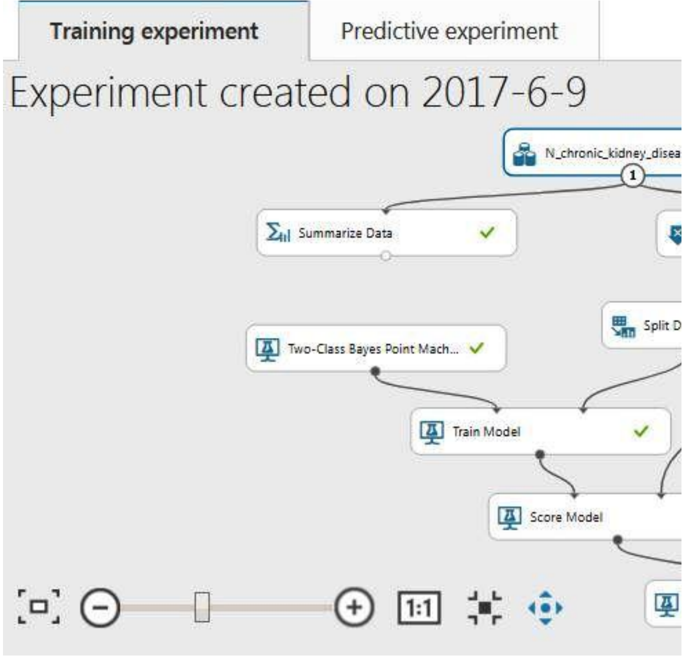
ML algorithms play important role in diagnosis of CKD. On the basis of quantitative and qualitative findings, the authors [14] revealed that the Random Forest (RF) classifier achieves the near-optimal performances on the identification of CKD. The RF based model can also be utilized for diagnosis of similar diseases.

In the work [15] classification models have been built with various classification algorithms along with two attribute evaluator methods to predict and classify the CKD and non CKD patients. These models have applied on dataset available at UCI repository [12].The models have shown better classification performance on attribute reduced dataset than the original dataset. The Models are built using NB, SMO and IBK classifiers achieved classification results 95%, 97.75% and 95.75% respectively.

3. Proposed Work

The proposed Cloud based Predictive model is shown in Figure 2. The model is built and tested over Microsoft Azure ML Workspace [16], the real cloud environment for ML tasks. Finally the model is deployed as web service for testing with unclassified data. The methodological steps that are followed to implement proposed work are mentioned here:

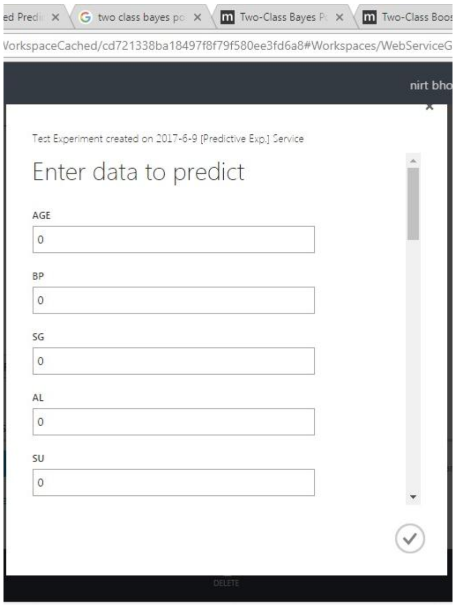
1. Create New Resource within Microsoft Azure Machine Learning Analytics solution.
2. Import/Upload the dataset.
3. Pre-process the dataset. This step involves conversion of attributes to suitable types, i.e. nominal, binary and numeric etc. to apply different algorithms. The pre-process step also involves dealing with missing values. Here "Missing value Scrubber" is used.
4. Randomly split and partition the data into training and test set. Here we used only 40% of data as training set to build the model.
5. Apply Machine Learning Algorithm to Train the model. We applied two algorithms.
6. Now Score the Model (Classifier) with standard metrics.
7. Run the experiment.
8. Evaluate the model using 'Evaluate Model' module.
9. Now build Predictive application.
10. Deploy the project as Web Service.



Experiment created on 2017-6-9

Figure 2: Proposed Predictive Model for Detection of CKD risk

The project is finally viewed as a GUI based application as shown in Figure 3. To test the application, provide the values to appropriate boxes and click on write button to see the results.



Test Experiment created on 2017-6-9 [Predictive Exp.] Service

Enter data to predict

AGE: 0

BP: 0

SG: 0

AL: 0

SU: 0

Figure 3: Snapshot of GUI

4. Experimental Setup and Result Analysis

To implement the proposed work Azure ML workspace [16] is used. Azure ML workspace provides ML studio, a graphical tool that is used to implement DM / KDD process from beginning to end. It includes: a set of data pre-processing modules; a set of ML algorithms; An Azure ML API to access model deployed on Azure. ML Studio allows a user to import datasets, data pre-processing methods, ML algorithms and more onto its design interface.

For experimentation purpose CKD dataset from UCI ML repository [12] is used. The dataset includes 400 instances with 24 attributes and a class attribute. The description of dataset is shown in Table 1.

Table 1: Dataset Description

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*

Two algorithms that are chosen to build the model are discussed in section 2 of this paper.

5. Result Analysis and Discussion

5.1 Evaluation Parameter

Prediction Accuracy (M): Accuracy of a Model M is referred as the percentage of test set instances that are correctly classified by the model M.

Confusion Matrix: Given m classes, $CM_{i, j}$, an entry in a confusion matrix, indicates of tuples in class i that are labelled by the classifier as class j

	C1	**C2**
C1	*True Positive*	*False Negative*
C2	*False Positive*	*True Negative*

Precision: It is also called positive predictive value. It calculates the fraction of positively classified instances that are relevant from among the retrieved instances.

5.2 Results and Analysis

Among both the models Bayes algorithm is performing well as shown in Table 2. The Bayes method is robust and does not over-fit the training data. Also, it scales well as it is available on Cloud platform.

Table 2: Results: Prediction Accuracy and Precision

Models	Prediction Accuracy	Precision
Bayes Point Machine	*97.1*	*0.947*
Logistic Regression	*94.9*	*0.911*

6. Conclusion and Future Work

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. Also, the number of issues arises while dealing with huge amount of data; few of them are scaling and reliable analysis of data. The ML algorithms do not scale well on single system, whereas cloud platforms provide scope to scale the algorithms on big data also. 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] 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] 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.

[4] 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.

[5] https://msdn.microsoft.com/en-us/library/azure/dn905930.aspx

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

[7] https://msdn.microsoft.com/en-us/library/azure/dn905994.aspx

[8] http://www.stat.cmu.edu/~cshalizi/uADA/12/lectures/ch12.pdf

[9] 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.

[10] 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.

[11] 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.

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

[13] 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.

[14] Subasi A., Alickovic E., Kevric J. (2017) Diagnosis of Chronic Kidney Disease by Using Random Forest. In: Badnjevic A. (eds) CMBEBIH 2017. IFMBE Proceedings, vol 62. Springer, Singapore

[15] N. Chetty, K. S. Vaisla and S. D. Sudarsan, "Role of attributes selection in classification of Chronic Kidney Disease patients," 2015 International Conference on Computing, Communication and Security (ICCCS), Pamplemousses, 2015, pp. 1-6.

[16] https://studio.azureml.net/