ISSN- 2394-5125 VOL 7, ISSUE 17, 2020

Chronic Kidney Disease Prediction using Classification techniques

Pawan Agarawal
(pawankosi@gmail.com)
Compocom Institute Of Technology and Management, Jaipur

Sanjay Kumar
(sanjayshrimal@gmail.com)
Compocom Institute Of Information Technology and Management, Jaipur

Abstract
Healthcare industry faces the necessity to manage the growth and process the data into a new actionable insights. Big data and data mining techniques undertakes the prominent challenge by leveraging the improvement towards the value based care of patient. In this paper the usage of Big data analytics and data mining in healthcare, overcomes the challenges of analysing the hidden information and extracting the useful information fromthe massive amount of data such as patients EHR's , which provides an intuition of predicting the chronic kidney disease at the early stage. The aim of the paper is to predict the chronic kidney disease using classification algorithm on structured dataset. There are many classification algorithms namely Logistic Regression, Random Forest, Naïve Bayes, Decision Tree, Support Vector Machine, K-Nearest Neighbours are used in the prediction model to evaluate the performance based on the accuracy, precision and f1-measure on the given dataset. This paper studies the usage of classification algorithm in the prediction modelto help the physicians to make right decision in order to extend the life span.
Keywords: Classification Algorithm, Data mining, Big data, Chronic Kidney disease, Electronic Health Record(EHR's).

I. Introduction

Chronic kidney disease (CKD), is a tenacious abnormality in kidney function for more than 3 months, and affects 8-16% of the population globally. However, patients are unaware of the CKD as this disease tends to slow down for a long period without a symptom or sign and lessthan 5% of patients with early chronic kidney disease report awareness of the disease. ChronicKidney Disease(CKD) is common among the adults who have above 65 of age.

The other symptoms like diabetics, hypertension and heart disease may cause damage to the blood vessels that affects the filtration of blood and wastes are inbuilt inside the body as it leadsto CKD. The patients with severe CKD, may have Edema. It is a swelling that occurs in legs, face, ankles, feet and hands as it cannot remove extra fluid and salt in the body. CKD is diagnosed by a blood and urine test based on the glomerular filtration rate(GFR). The GFR is used to predict the stages of chronic kidney disease, there are commonly five stages and 3 to 5stages are the progression of the CKD. The fifth stage is the end stage as patient need to undergo dialysis or kidney transplant to maintain their health.

It is necessary to handle the task of diagnosing a disease with forethought. Big data and data mining are two different techniques but does the same operation of finding out the hidden patterns from large dataset into a knowledgeable data for the support of decision making by physicians and predicting the disease at the primitive stage to avoid its progression.

II. Literature Survey

A review of techniques in data mining is proven to predict the kidney disease at the primitive stage based on the feature selection and classifying using different algorithms in prediction model using large number of data.

Rucy Kei Chiu et al. [1] The detection of CKD is predicted at the early state by using the neuralnetwork model like BPN, GRNN, MNN and finding out the best algorithm based on accuracy& sensitivity by comparison. which helps the public to known the severity stages of CKD by self- detection in cloud and prevent them by taking appropriate treatment for CKD. As a resultBPN has gained the accuracy of 94.75%.

Naganna Chetty et al. [2] The attribute selection is done using wrappersubsetEval attribute evaluator & Best first search method implemented along with three classification algorithms like Naïve Bayes, SMO & IBK. Based on that NB reduced the dataset to 6 attributes, SMO to12 and IBK to 7 attributes. The accuracy of original dataset using NB, SMO & IBK is 95%, 97.75% & 95.75%. The accuracy of reduced dataset using NB, SMO & IBK is 99%, 98.25% & 100%.

Asif Salekin et al. [3] Algorithms of random forest, neutral network, K-nearest neighbour are used in predicting the CKD of 24 attributes & wrapper method is used to reduce the attributesto 12 for high accuracy detection by applying the lasso regularization, the reduction is further reduced to 10 attributes and priority is ranked by random forest classifier. The accuracy resultedin 99.8% for the reduced attribute.

Uma N et al. [4] In this method they are using one R algorithm for feature subset selection andpredicting the

ISSN- 2394-5125 VOL 7, ISSUE 17, 2020

CKD using Naïve Bayes to classify. Based on this method, 5 attributes are selected and they are calculating the GFR to predict the stages of CKD for preventing at the early stage. As its result the accuracy is 97.5% than the original dataset as it 12.5% greater.

Sai Prasad Potharaju et al. [5] they have used prediction model of rule induction (Trip, oneR, Ridor) and Decision tree algorithm (J48, SimpleCart, ADtree, RandomTree & REPTree) for 400 instances dataset. As a result it produced 150 instances to class CKD & 250 to not CKD, which is imbalanced dataset. To solve this SMOTE of over sampling is used in 1170 instances, were 600 is not CKD & 570 is CKD in 3 samplings. The accuracy is compared based on the parameters like precision, Recall, F-measure, Roc Area with the original data and sampled data.

Patcharaporn Panwong et al. [6] The model collects the data of present disease & congenital disease which leads to CKD and calculate the CKD stage by recalculating the GFR formulae. This model predicts the time interval of stage 3 & stage 5 based on the dataset. The seven algorithms of Decision tree(Random Forest, J48),K-NN(Ibk(1,3,5)),Naïve Bayes, Multilayer perceptron are applied on original dataset, as of result random forest has the high accuracy. The imbalanced dataset shows low accuracy, the SMOTE algorithm is applied. This model evaluates that random forest has high accuracy of 85% compared to original dataset of 60%.

Nusrat Tazin et al. [7] The classification algorithm Decision Tree, NB, SVM & K-NN are compared with the implementation of ranking algorithm of 25, 20, 15 & 10 attributes. The changes in the level of selecting the attribute may affect the accuracy of algorithms used. The Decision Tree has 99% accuracy of 15 attributes after implementing ranking method.

Haya Alasker et al. [8] data mining techniques is used to predict CKD the classification algorithm like BPN, NB, Decision tree, J48, OneR, KNN are used to compare the accuracy level, MAE, RSME & Roc are of 24 attributes and reduced subset features of 8 attributes. Among both the dataset the NB algorithm shows the best result of 99.36% accuracy.

Basma Boukenze et al. [9] The algorithms that are used to predict CKD is to be helpful in decision making to further proceed to treatment. In this five algorithms SVM, NB, MLP, KNN & C4.5 are compared based on accuracy, sensitivity, execution time & specificity. As of results, C4.5 has proved in above all comparison especially by the low error rate & shortest executiontime.

Pinar Yildirim [10] The given dataset is imbalanced as it contains CKD (62.5%) & not CKD (37.5%) of 400 patients. The dataset are balanced using sampling algorithms like Resample, SMOTE & spread sub and the accuracy of all the sampling method is deployed using multilayerperception algorithm by varying learning rate between 0.1 to 0.8. Resample algorithm has thebest accuracy of 0.998.

Made Satria Wibawa et al. [11] The features of the original dataset are reduced to 17 based oncorrelation-based feature selection model and AdaBoost ensemble algorithm is used to detect the CKD along with the base classifiers of NB, K-NN, SVM. The Base classifiers are compared based on the four parameters.. i.e. accuracy, precision, recall & F-measure in CFS model withAdaBoost. The accuracy rate in K-NN is 0.981.

Huseyin Polat et al. [12] The feature selection is used to reduce the computing time & usage of optimized dataset. The method of selection is implemented by two approaches, wrapper & filter approach with two search engine algorithms (Greedy stepwise & Best First search). The wrapper method uses feature subset selection of classifier subset evaluator using Greedstepwise search & Wrapper subset evaluator with Best First search. The Filter method uses correlation feature selection subset evaluator with Greedy Stepwise & Filter Subset evaluator with Best First search. The classification algorithm of SVM is used with feature selection method of wrapper & Filter method. As a result of accuracy rate, the filter approach of Best First search & classifier with SVM algorithm shows 98.56%.

Abdulhamit Subas et al. [13] This model is used to classify the CKD & Non-CKD with the given dataset of 400 samples. The performance of the classification is based on precision, F- measure Accuracy. The machine learning algorithms like SVM, ANN, K-NN, C4.5 Decision Tree & Random Forest are compared to classify the prediction of CKD & Non CKD based onprecision, F-measure & Accuracy. The RF shows 100% in accuracy, precision & F-measure

.

III. Chronic Kidney Disease (CKD)

Kidney disease occurs due the unusual functionality of not removing waste products from the body or not releasing of hormones that regulate blood pressure and imbalance fluids in the body. These functionality are carried out by a blood vessel and it is been affected by a cause of hypertension and diabetics. These problem may occur slowly over a long period and there is no sign of predicting at early stage. However some sign or symptom that occurs at the end stage that can predict the CKD is changes in urination, loss of appetite, swelling in ankles, feet,shortness in breath. The presence of the kidney disease is abide by high levels of albumin in urine and the stages of kidney disease is been predicted by Glomerular Filtration Rate. There are five stages in CKD, the first stage is said to be in normal, if it has the higher level or greaterthan 90ml/min of GFR rate. The second stage is between 60-89ml/min. The third stage is between 30-59ml/min. Severe or the fourth stage is between 15-29ml/min. The last stage has the GFR rate less than 15ml/min, which is the end stage.

Medication are taken to manage the health condition of CKD. When it leads to the end stage, the need of dialysis or kidney transplant is necessary. The kidney transplant may not suit for everyone for the functionality

ISSN- 2394-5125 VOL 7, ISSUE 17, 2020

of kidney in their body. So, the patients can also undergo for an artificial filtration (Dialysis) for two to three times in a week for the circulation of blood usinga machine. These two methods are carried out to increase the life span.

IV. Big Data

Healthcare system generates a large volume of data in storage to process the data, As the traditional methods lacks in their view. To overcome this, big data analytics is capable of storing luminous volume of data, which enhance the potential of digitized EHR's in providingright information and diagnosis to the patients in response, it satisfies the patients in right medication, cost reduction in treatment and preventing from disease. Therefore it improves thequality of life span.

As the use of big data is also viewed in business centric, lot of models are designed based on 5v's- Velocity, Volume, Value, Variety and Veracity. The models designed are continuously changing with a new tools to handle the big data. In healthcare, the models are designed in order to optimize growth level by improving care efficiency in proper treatment, effectiveness based on real-time monitoring of predicting the disease at right time and are alsopersonalized services provided to the patients.

V. Data Mining

Data Mining is used in extracting the knowledge and patterns from huge amount of data. It is the process of computing and analysing data based on different angle, dimension and summarizing them in useful information. The techniques that are used in the health care provides a cost reduction and effective prediction of disease at the right time, which helps in treating the patients. Therefore large number of sample data set are collected and implementedin the prediction model to classify the chronic kidney disease. The prediction model based on the data mining techniques of association, clustering, and classification produces the accuracy
rate of predicting the disease using different algorithms. Performance level of the prediction isbased on the parameters like accuracy, precision and f1-measure.

VI. Comparison Table

Year	Author	Classification Algorithms	Accuracy Rate
2012	Rucy Kei Chiu et al	BPN, GFNN, MNN	BPN-94.75%. MNN– 93.23% GFNN-86.63%
2015	Naganna Chetty et al	Naïve Bayes, SMO & IBK	NB- 99% SMO- 98.25% IBK- 100%
2016	Asif Salekin et al	Random forest, Neural network, K-Nearest Neighbour	RF- 99.8%
2016	Uma N et al	Naïve Bayes	NB- 97.5%
2016	Sai Prasad Potharaju et al	Rule induction (Jrip, oneR, Ridor) and Decision tree algorithm (J48, SimpleCart, ADtree, RandomTree & REPTree)	ADTree– 99.65% Jrip- 99.65% J48- 99.57% SimpleCart- 99.31% REPTree- 99.23% RF- 99.05% Ridor – 98.8% OneR- 94.6%
2016	Patcharaporn Panwong et al	Decision Tree (Random Forest, J48),K-NN(Ibk(1,3,5)),Naïve Bayes, Multilayer perceptron	RF – 86.60% K-NN(Ibk=1)- 86.32% NB- 60.46% MLP- 83.24%

ISSN- 2394-5125 VOL 7, ISSUE 17, 2020

2016	Nusrat Tazin et al	Decision Tree, NB, SVM & K-NN	Decision Tree-99% SVM-97.75% K-NN-95.75%NB-95%
2017	Haya Alasker et al	BPN, NB, Decision tree, J48, oneR,KNN	NB - 99.36% KNN- 99.20% ANN-98.41% J48-98.41% OneR- 97.61% DT- 97.61%
2017	Basma Boukenze et al	SVM, NB, MLP, KNN & C4.5	C4.5-99% MLP-99% SVM-98% K-NN-96% NB-95%
2017	Pinar Yildirim	Resample, SMOTE & Spread Sub Sample	Resample-99.8% SpreadSub Sample-99.7% SMOTE- 99.5%
2017	Made Satria Wibawa et al	NB, K-NN, SVM	K-NN-98.1% NB-98% SVM-97.5%
2017	Huseyin polat et al	SVM	SVM- 98.56%
2017	Abdulhamit Subas et al	ANN, SVM, K-NN, Decision Tree, Random Forest	RF-100% DT-99% SVM-98.5%ANN-98% K-NN-95.75%

VII. Conclusion

In this paper, different data mining techniques are used to classify the Chronic Kidney disease that helps is recognizing the disease based on the feature selection. The feature selection are represented to predict the high performance level of the model based on accuracy, precision, f1-measure on the given dataset. The diagnosis of CKD using the model is estimated with more classifier. Different classification algorithms are applied in the prediction model and compared among them, to choose the best one for the prediction based on the parameters. Therefore thesemodel helps in decision making, giving the right medication and awareness to the patients at the primitive stage.

References

[1] Rucy Kei Chiu, Renee Y. Chen, Shin-An Wang and Sheng-Jen Jian, "Intelligent systemson the cloud for the early detection of Chronic Kidney disease", IEEE, 2012.

[2] Naganna Chetty, Kunwar Singh Vaisla and Sithu D Sudarsan, "Role of Attributes Selectionin Classification of Chronic Kidney disease patients", IEEE, 2015.

[3] Asif Salekin and John Stankovic, "Detection of Chronic Kidney Disease & SelectingImportant Predictive Attributes", IEEE, 2016.

[4] Dr. Uma N Dulhare and Mohammad Ayesha, "Extraction of Action Rules for ChronicKidney Disease using Naïve Bayes Classifier", IEEE, 2016.

[5] Sai Prasad Potharaju and M. Sreedevi, "An Improved prediction of Kidney Disease usingSMOTE", Research gate, 2016.

[6] Patcharaporn Panwong and Natthakan Iam-On, "Predicting Transitional Interval of KidneyDisease stages 3 to 5 using Data Mining method", IEEE, 2016.

ISSN- 2394-5125 VOL 7, ISSUE 17, 2020

[7] Nusrat Tazin, Shahed Anzarus Sabab and Muhammed Tawfiq Chowdhury, "Diagnosis of Chronic Kidney Disease using effective classification and feature selection technique", IEEE,2016.

[8] Haya Alasker, Shatha Alharkan, Wejdan Alharkan , Amal Zaki and Lala Septem Riza, "Detection of Kidney Disease using various Intelligent classifiers", IEEE, 2017.

[9] Basma Boukenze, Abdelkrim Haqiq and Hajar Mousannif, "Predicting Chronic Kidney failure Disease using Data Mining Techniques", Springer, 2017.

[10] Pinar Yildirim, "Chronic Kidney Disease Prediction on Imbalanced data by multilayer perceptron", IEEE, 2017.

[11] Made Satria Wibawa, I Made Dendi Maysanjaya and I Made Agus Wirahadi Putra, "Boosted classifier and Features Selection for Enhancing Chronic Kidney Disease Diagnose",IEEE, 2017.

[12] Huseyin Polat, Homay Danaei Mehr and Aydin Cetin, "Diagnosis of Chronic Kidney Disease Based on Support Vector Machine by Feature Selection Methods", Springer, 2017.

[13] Abdulhamit Subas, Emina Alickovic and Jasmin Kevric, "Diagnosis of Chronic Kidney Disease by using Random Forest", Springer, 2017.