IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 7, July 2020

Early Prediction of Chronic Kidney Disease in Adolescents using Machine Learning

A. Stella[1], Vasanthi Kumari P[2]

Research Scholar, Department of Computer Science, Dayananda Sagar University, Bangalore[1]

Professor, Department of Computer Science, Dayananda Sagar University, Bangalore[2]

Abstract: The rising number of kidney failure in adolescents and young children is of great concern. Pediatric CKD is a dynamic and complex medical and psychosocial disease with unique factors that separate this population from adults. Due to the unique and complex physical, psychological, and family backgrounds, young children may develop damage of kidneys. The long-term mortality for children, adolescents, and young adults with CKD (Chronic Kidney Disease) remains substantially higher than their healthy counterparts. The complex challenges that adolescent and young adult CKD patients face has to be dealt with on a serious note. Adolescents have different CKD etiologies and progress are quite dissimilar to that faced by adults, but have similar multifarious comorbidities. CKD can delay and limit growth. In this paper, various Machine Learning algorithms are used to predict the occurrence of the disease. The benefit of implementing this technique is that the disease can be diagnosed at an early stage based on the various symptoms of the patient and thus can help them to get the diagnosis and treatment on time which will lead to better health and better Quality of Life. Here, the prediction skill of several machine-learning algorithms for early prediction of CKD has been analyzed by usage of predictive analytics, in which the association of data parameters and the target class attributes is done. Predictive analytics enables us to introduce the optimal subset of parameters to feed machine learning to build a set of predictive models.

Keywords: Adolescents, Chronic Kidney Disorder, Machine Learning Algorithms.

I.INTRODUCTION

Premature deaths are claiming more than 2 million lives every year because irreversible treatment for kidney failure is not available. Chronic kidney disease is a major disease that affects the entire world population. For example, in the year 2005, there were approximately 58 million deaths worldwide, with 35 million attributed to chronic disease, according to the World Health Organization [1].The high incidence of death from kidney failure is not only because of lack of awareness and early detection, but also due to shortage of dialysis equipment, the high cost of transplant surgery, and new, stringent government regulations regarding organ donation. 90 per cent of the kidney patients in India cannot afford their treatment [2]. Awareness should start early, so cases can be detected early and dialysis can be prevented. Most researchers have contributed to the prediction of Chronic Kidney Disorder (CKD) in older patients, but, the attention on children and adolescents being affected by CKD has not been considered much.

II. CAUSES OF CKD IN CHILDREN

Chronic kidney disease is complex in both adults and children, but the disease is far from the same between these populations [3].Older people usually have kidney disease caused by high blood pressure or diabetes. But in kids and teens, kidney disease is caused due to various factors:

Infection or Repeated Infections: Urinary infections happen a lot. They usually start in the bladder, but they can travel up and infect the kidneys. The infection can damage gradually and hence the kidneys. Overtime, scars can affect the proper functioning of kidneys.

Structural Problems: The kidneys may not be the normal size or the parts of it may be structured in an unusual way. Cysts can develop in the kidneys, it may be swollen or damaged and this will affect the kidneys.

Glomerulonephritis: Glomerulonephritis is an inflammation of the glomeruli, the kidney's filtering units. It may be caused by: (i) an infection (ii) some drugs or toxic chemicals (iii) an immune system reaction that damages the kidneys (iv) bacterial infection.

Nephrotic Syndrome: Nephrotic syndrome is when a person's glomeruli are damaged.

Data Collection and Preprocessing: CKD in young children ruins their quality of life. In [5], a software tool has been developed that proves the competences of ANN for classification of patients' health status which possibly may lead to End Stage Kidney Disease (ESKD). The classifier is an ensemble of ten ANN networks. The tool has been trained by

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 7, July 2020

using data that had been collected in a period of 38 years at University of Bari. The tool has been improved and made derivable both as a mobile application and as a web application.

The dataset for classifying CKD has been taken from UCI machine learning repository [6]. It has recorded the details of 400 cases, out of which 250 of the cases are patients with CKD and the remaining 150 records are details of normal persons. The classifier attribute indicates whether a patient has a CKD or not. This dataset describes 24 clinical attributes and 1 target attribute, which indicates the classification. The features are divided into three parts clinical history, physical examination and lab tests. According to the properties of the attributes, the target attribute was classified into notckd - "no disease" and ckd -"presence of disease". The extra attributes such as Serum Phosphorus, Serum Calcium and Urinary Calcium excretion, needed for detection and classification of MA disorder should be added to the dataset.

Learning Algorithms: Machine Learning algorithms are used to train and test the data. The following machine learning models can been obtained by using the corresponding predictors on the complete CKD data sets for diagnosing CKD. 1) Regression-based model: LOG 2) Tree-based model: RF 3) Decision plane-based model: SVM 4) Distance-based model: KNN 5) Probability-based model: NB 6) Neural network: FNN Generally, in disease diagnosis, diagnostic samples are distributed in a multidimensional space. This space comprises predictors that are used for data classification (ckd or notckd). Samples of data in the space are clustered in different regions due to their different categories. Therefore, there is a boundary between the two categories, and the distances between samples in the same category are smaller. LOG is based on linear regression, and it obtains the weight of each predictor and a bias. If the sum of the effects of all predictors exceeds a threshold, the category of the sample will be classified as ckd or notckd. When a non-linear relationship exists between dependent and independent variables, it is logical to transform the variables logarithmically. Random Forest are a collaborative learning methodology for problems that require classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class. The mean prediction of the individual trees is determined. RF generates a large number of decision trees by randomly sampling training samples and predictors. Each decision tree is trained to find a boundary that maximizes the difference between ckd and notckd. The final decision is determined by the predictions of all trees in the disease diagnosis.

SVM differentiates between two classes by generating a hyper plane that optimally separates classes after the input data have been transformed mathematically into a high-dimensional space. The SVM has capability to make fine distinctions among classifications, especially when sample sizes are relatively small and a large number of variables are involved. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. It can be used for Regression as well as or Classification. It does not make any assumption on underlying data. SVM learns from stored dataset and during classification rather than instantly from the training set. It performs an action on the dataset. It finds the nearest training samples by calculating the distances between the test sample and the training samples and then determines the diagnostic category by voting. Naïve Bayes classifier is known to outperform even highly sophisticated classification methods. It calculates the conditional probabilities of the sample under the interval by the number of ckd and notckd samples in each different measurement interval. FNN can analyse non-linear relationships in the data sets due to its complex structure, and the sigmoid activation function was used in the hidden layer and the output layer.

III. METHODOLOGY

The combinations of the RF with the rest of the models could be used to establish an integrated model. In [7], LOG, RF, SVM and KNN were run five times on each complete data, and the average time taken are summarized in Table 1. It can be seen that the SVM and KNN take more time than the LOG and RF. In addition, SVM and KNN are also effected by their respective model parameters, so the parameters need to be adjusted before the models are established, which means more manual intervention is needed. For the LOG, there was no additional parameter that need to be adjusted. For the RF, the default parameters of the model were used. Hence, a combination of the LOG and the RF was selected to generate the final integrated model. It can be seen that the SVM and KNN take more time than the LOG and RF. In addition, SVM and KNN are also effected by their respective model parameters, so the parameters need to be adjusted before the models are established, which means more manual intervention is needed. For the LOG, there was no additional parameter that need to be adjusted. For the RF, the default parameters of the model were used. Hence, a combination of the LOG and the RF was selected to generate the final integrated model.

Table 1: The time spent by RF, LOG, SVM and KNN on the complete data.

KNN imputation with K	RF (s)	LOG (s)	SVM (s)	KNN (s)
3	0.382	0.138	16.114	2.796
5	0.376	0.144	15.836	2.788
7	0.386	0.140	16.222	2.864
9	0.396	0.128	16.276	2.822
11	0.394	0.132	16.104	2.766

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 9, Issue 7, July 2020

Procedure for Prediction:

1: Acquire the dataset [6].
2: Preprocess the data [9].
-PCA (Principle Component Analysis): It is used to reduce the dimensionality of a dataset consisting of many variables correlated with each other, either heavily or lightly, while preserving the variation.
3: Integrated Model Selection.[7]
4: Initialize the perceptron and traverse the samples in the new training data set.
5: Return LOG, RF and perception.
6: Input the data into LOG and RF to record the probabilities that the samples are judged as notckd by them.
7: Input the probabilities into the perceptron to obtain the result

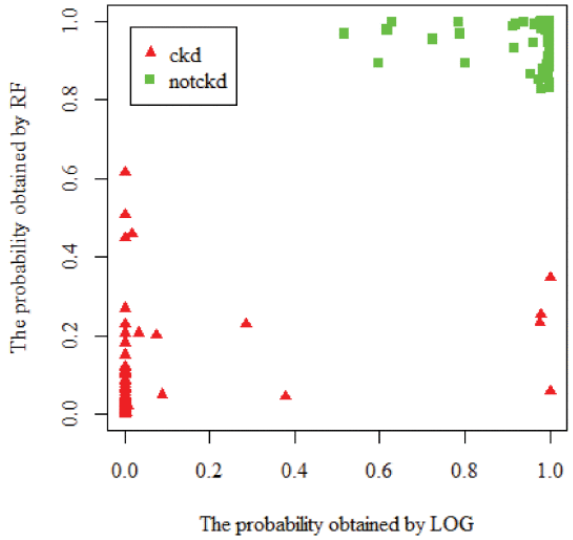

Fig 1: The probability distribution of the samples in the complete CKD data set (at K = 11), the horizontal axis and the vertical axis represent the probabilities that the samples were judged as notckd by the LOG and the RF

Multilayer Perceptron (MLP) Classifier: MLPs are universal function approximators as shown by Cybenko's theorem,[10] so they can be used to create mathematical models by regression analysis. MLPs make good classifier algorithm, because the output variable is categorical. It can be seen from Fig. 1 that the samples have different aggregation regions in the two-dimensional plane. Samples with ckd are concentrated in the lower left part, while the notckd samples are distributed in the top right part. Due to the fact that the results in the two models are different, some samples are located at the top left and lower right, and one of the two models makes the misjudgments. Multi Perceptron can be used to separate samples of two categories accurately by plotting a decision line in the two-dimensional plane of the probability distribution. The Multi perceptron classifier is shown in Fig. 2.The weights can be adjusted for the misjudged cases and can be amended accordingly.

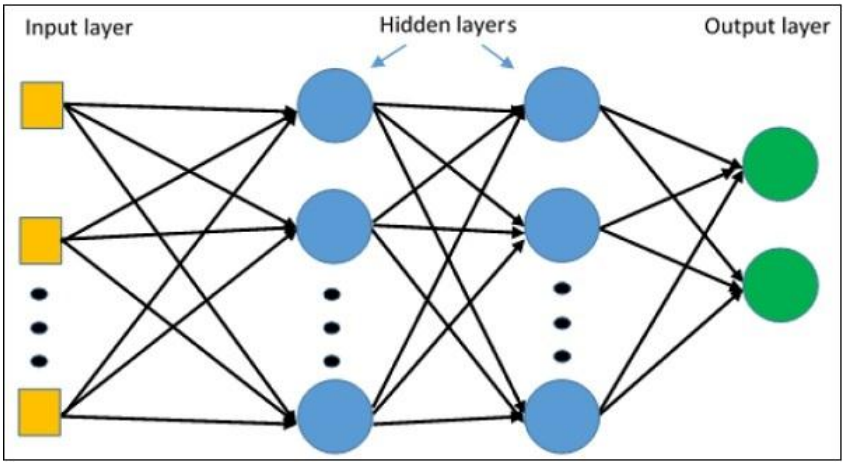

Fig 2: Applying the Multi Perceptron classifier to the new classified data

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 7, July 2020

IV. MODEL EVALUATION AND PRECISION

In this study, the performance is measured using the following metrics.
– True Positive (TP) - Positive instances that will be correctly classified as positive outputs.
– True Negative (TN) - Negative instances that will be correctly classified as negative outputs
– False Positive (FP) - Negative instances that will be wrongly classified as positive outputs
– False Negative (FN) - Positive instances that will be wrongly classified as negative output
– Classification Accuracy - Indicates the ability of classifier algorithm to diagnose of classes of dataset.

Models at different values of K	Actual	Prediction	
		ckd	notckd
Integrated model at K = 3	ckd	250	0
	notckd	0	150
Integrated model at K = 5	ckd	250	0
	notckd	1	149
Integrated model at K = 7	ckd	248	2
	notckd	1	149
Integrated model at K = 9	ckd	249	1
	notckd	0	150
Integrated model at K = 11	ckd	250	0
	notckd	0	150

Table 2: The confusion matrices returned by the integrated models.

The integrated model improves the performance of separate individual models and is superior to almost all the contrast models, with the highest accuracy and F1 score can achieve 100% in Table 2.

V. CONCLUSIONS

KNN imputation could fill in the missing values in the data set for the cases through the misjudgments analysis, LOG and RF were selected as the component models. The LOG achieved an accuracy of around 98.75%, which indicates most samples in the data set are linearly separable. The RF achieved better performance compared with the LOG with the accuracy was around 99.75%. An integrated model combining LOG and RF was established to improve the performance of the component models. From the simulation result, the method of integrating several different classifiers is feasible and effective. This methodology could be extended to more complex situations.

REFERENCES

[1]. Levey AS, Atkins R, Coresh J, et al. Chronic kidney disease as a global public health problem: Approaches and initiatives - a position statement from Kidney Disease Improving Global Outcomes. Kidney Int. Aug 2007; 72(3):247-259.
[2]. Khanna U. The economics of dialysis in India. Indian J Nephrol 2009; 19:1–4.
[3]. Kaspar C.D.W, Bholah R, Bunchman T.E., "A Review of Pediatric Chronic Kidney Disease", Blood Purif, 41: 211– 217, 2016, doi.org/10.1159/000441737.
[4]. K.R Lakshmi, Y. Nagesh, and M. VeeraKrishna, "Performance Comparison of Three Data Mining Techniques for Predicting Kidney Dialysis Survivability", International Journal of Advances in Engineering and Technology, vol.7, no.1, pp. 242-254, March 2014.
[5]. T. Di Noia, V. C. Ostuni, F. Pesce, G. Binetti, D. Naso, F. P. Schena, and E. Di Sciascio. "An end stage kidney disease predictor based on an artificial neural networks ensemble", Expert Systems with Applications, vol. 40, pp. 4438–4445, 2013.
[6]. Soundarapandian P. (2015). UCI Machine Learning Repository [https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease]. Irvine, CA: University of California, School of Information and Computer Science.
[7]. Jiongming Qin, Lin Chen , Yuhua Liu , Chuanjun Liu , Changhao Fengi , Bin Chen, "A Machine Learning Methodology for Diagnosing Chronic Kidney Disease", DOI: 10.1109/ACCESS.2017.
[8]. S. Vanaja and K. R. Kumar, "Analysis of feature selection algorithms on classification: a survey," International Journal of Computer Applications, vol. 96, no. 17, 2014.
[9]. Harri Siirtola,Tanja Saily,Terttu Nevalainen," Interactive Principal Component Analysis", 2017 21st International Conf Information Visualisation.
[10]. G. Cybenko, Elisa Negrini, "Universal Approximation Theorem", https://users.wpi.edu/~msarkis.