International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 4 Apr. 2021, pp: 22-26 www.ijaem.net ISSN: 2395-5252

A Machine Learning Methodology for Diagnosing Chronic Kidney Disease

Seetharaman TJ[1], Giridharan N[2]
[1]Student, K S Rangasamy College Of Technology, Tiruchengode Tamilnadu.
[2]Assistant Professor, CSE Dept, K S Rangasamy College Of Technology, Tiruchengode Tamilnadu.

Submitted: 01-04-2021 Revised: 11-04-2021 Accepted: 14-04-2021

ABSTRACT: Chronic Kidney Disease (CKD) is a global health problem with high morbidity and death rate. Since there are no visible side effects during the beginning phases of CKD, patients regularly neglect to see the illness. Early discovery of CKD empowers patients to get timely treatment to get rid of this infection. Machine learning models can successfully help clinicians accomplish this objective because of their quick and precise acknowledgment execution. In this assessment, we propose an KNN and Logistic regression system for diagnosing CKD. The CKD data set was got from the University of California Irvine (UCI) AI store, which has a tremendous number of missing characteristics. KNN attribution was used to find the missing qualities, which chooses a few complete examples with the most comparative estimations to handle the missing information for each fragmented example. Missing qualities are generally found, all things considered, clinical circumstances since patients may miss a few estimations for different reasons. By implementing KNN and LR for Diagnosing CKD we achieved the Accuracy of 98.6% and 97.38% respectively. Therefore, we theorized that the philosophy could be appropriate to find more clinical information for sickness finding.

Keywords: CKD-Chronic Kidney Disease, KNN-K Nearest Neighbor, UCI- University of California Irvine.

I. INTRODUCTION

Chronic Kidney Disorder (CKD) may be a longstanding disease of kidneys resulting in kidney failure (Inability to get rid of waste & balance fluids. The treatment can help but this condition can't be cured but when identified earlier the seriousness could also be reduced). Normally, kidney filters the waste and excess fluid from the blood because the kidney fails, waste builds up. Early discovery of CKD empowers patients to urge timely treatment to urge obviate this infection. during this assessment, we propose an K Nearest Neighbor and Logistic regression system for diagnosing CKD. Their studies have achieved good leads to the diagnosis of CKD. within the above models, the mean imputation is employed to fill within the missing values and it depends on the diagnostic categories of the samples. actually, patients might miss some measurements for various reasons before diagnosing. additionally, for missing values in categorical variables, data obtained using mean imputation may need an outsized deviation from the particular values. for instance, for variables with only two categories, we set the categories to 0 and 1, but the mean of the variables could be between 0 and 1. it's developed supported feature selection technology, the proposed models reduced the computational cost through feature selection, and therefore the range of accuracy in those models was from 97.75%-98.5%.

CHRONIC KIDNEY DISEASE

Chronic Kidney Disease (CKD) may be a sort of renal disorder during which there's gradual loss of kidney function over a period of months to years. Initially there are generally no symptoms; later, symptoms may include leg swelling, feeling tired, vomiting, loss of appetite, and confusion. Complications include an increased risk of heart disease, high vital sign, bone disease, and anaemia. Causes of chronic renal disorder include diabetes, high blood pressure, glomerulonephritis, and polycystic renal disorder. Risk factors include a case history of chronic renal disorder. Diagnosis is by blood tests to live the estimated glomerular filtration rate (eGFR), and a urine test to live albumin. Ultrasound or kidney biopsy could also be performed to work out the underlying cause. Several severity-based staging systems are in use. Screening at-risk people is recommended. Initial treatments may include medications to lower vital sign, blood glucose, and cholesterol. Angiotensin converting enzyme inhibitors (ACEIs) or angiotensin II receptor antagonists (ARBs) are generally first-line agents for vital sign control, as

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 4 Apr. 2021, pp: 22-26 www.ijaem.net ISSN: 2395-5252

they slow progression of the renal disorder and therefore the risk of heart disease. Loop diuretics could also be wont to control edema and, if needed, to further lower vital sign. NSAIDs should be avoided. Other recommended measures include staying active, and certain dietary changes like a low-sodium diet and therefore the correct quantity of protein. Treatments for anaemia and bone disease may also be required. Severe disease requires hemo-dialysis, peritoneal dialysis, or a kidney transplant for survival. Blood pressure is increased thanks to fluid overload and production of vasoactive hormones created by the kidney via the renin–angiotensin system, increasing the danger of developing hypertension and coronary failure. Urea accumulates, resulting in azotemia and ultimately uremia (symptoms starting from lethargy to pericarditis and encephalopathy). Due to its high systemic concentration, urea is excreted in eccrine sweat at high concentrations and crystallizes on skin because the sweat evaporates ("uremic frost"). Potassium accumulates within the blood (hyperkalemia with a variety of symptoms including malaise and potentially fatal cardiac arrhythmias). Hyperkalaemia usually does not develop until the glomerular filtration rate falls to less than 20–25 ml/min/1.73 m2, at which point the kidneys have decreased ability to excrete potassium. Hyperkalaemia in CKD can be exacerbated by academia (which leads to extracellular shift of potassium) and from lack of insulin. Changes in mineral and bone metabolism which will cause 1) abnormalities of calcium, phosphorus (phosphate), parathormone, or vitamin D metabolism; 2) abnormalities in bone turnover, mineralization, volume, linear growth, or strength (kidney osteodystrophy); and 3) vascular or other soft-tissue calcification. CKD-mineral and bone disorders are related to poor outcomes. Metabolic acidosis may result from decreased capacity to get enough ammonia from the cells of the proximal tubule. Acidemia affects the function of enzymes and increases excitability of cardiac and neuronal membranes by the promotion of hyperkalemia. Anaemia is common and is especially prevalent in those requiring haemodialysis. It is multi-factorial in cause, but includes increased inflammation, reduction in erythropoietin, and hyperuricemia resulting in bone marrow suppression. In later stages, cachexia may develop, resulting in unintentional weight loss, muscle wasting, weakness and anorexia.

KNN IMPUTATION

KNN Imputer by scikit-learn may be a widely used method to impute missing values. It is widely being observed as a replacement for traditional imputation techniques. In today's world, data is being collected from variety of sources and is employed for analysing, generating insights, validating theories, and whatnot. This data collected from different resources may often have some information missing. This may flow from to a drag within the data collection or extraction process that would be a person's error. Dealing with these missing values, thus becomes a crucial step in data pre-processing. The choice of method of imputation is crucial since it can significantly impact one's work. A handful of literature in statistics deals with the source of missing values and ways to beat the difficulty. The best way is to impute these missing observations with an estimated value. In this article, we introduce a guide to impute missing values during a dataset using values of observations for neighboring data points. For this, we use the very popular KNN Imputer by scikit-learn k-Nearest Neighbors Algorithm. Missing values during a dataset are often a hornet's nest for any data scientist. Variables with missing values are often a non-trivial problem as there's no easy answer to affect them. Generally, if the proportion of missing observations in data is little relative to the entire number of observations, we will simply remove those observations. However, this is not the most often case. Deleting the rows containing missing values may cause parting away with useful information or patterns. This happens when the missing values haven't any hidden dependency on the other variable or any characteristic of observations. If a doctor forgets to record the age of each tenth patient entering an ICU, the presence of missing value wouldn't depend upon the characteristic of the patients. n this case, the probability of missing value depends on the characteristics of observable data. In survey data, high-income respondents are less likely to tell the researcher about the amount of properties owned. The missing value for the variable number of properties owned will depend upon the income variable. This happens when the missing values depend upon both characteristics of the info and also on missing values. In this case, determining the mechanism of the generation of missing value is difficult. For example, missing values for a variable like vital sign may partially depend upon the values of vital sign as patients who have low vital sign are less likely to get their blood pressure checked at frequently

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 4 Apr. 2021, pp: 22-26 www.ijaem.net ISSN: 2395-5252

II. EXISTING SYSTEM

The existing system predicts the chronic diseases which are for a particular region and for the particular community. Only particular diseases are predicted by this system. In this System, Big Data & CNN Algorithm is used for Disease risk prediction. For S type data, the system is using Machine Learning algorithm i.e., Decision Tree, Naïve Bayesian. The accuracy of the existing System is up to 94.8%.

The data identified with the undertaking what's more, acquires the qualities of the relating design. This innovation can accomplish exact and practical analyses of sicknesses; subsequently, it very well may be a promising technique for diagnosing CKD.

In the existing work, they streamline machine learning algorithms for the effective prediction of chronic disease outbreak in disease-frequent communities. They experiment with the modified prediction models over real-life hospital data collected from central China. They propose a convolutional neural network-based multimodal disease risk prediction (CNN-MDRP) algorithm using structured and unstructured data from the hospital. It has gotten another sort of clinical instrument with the improvement of data innovation what's more, has an expansive application prospect in view of the fast improvement of electronic wellbeing record. In the clinical field, it has just been utilized to detect human body status break down the significant components of the infection and analyze different sicknesses. For instance, the models worked by machine learning calculations were utilized to analyze coronary illness, diabetes and retinopathy, intense kidney injury, disease what's more, different sicknesses. In these models, calculations in view of relapse, tree, likelihood, choice surface and neural organization were regularly compelling.

III. PROPOSED SYSTEM

Generally, in disease diagnosis, diagnostic samples are distributed in a different space. It comprises predictors that are used for data classification (ckd or not-ckd). Samples of knowledge within the space are clustered in several regions thanks to their different categories. According to the effectiveness of classification, we choose these methods. Logistic Regression (LR) & K-Nearest Neighbour (KNN) are the Algorithm behind this project.

This work investigates how CKD are often diagnosed by using machine learning (ML) techniques. ML algorithms are a drive in detection of abnormalities in several physiological data, and are, with a superb success, employed in several classification tasks. In the present study, variety of various ML classifiers are experimentally validated to a true data set, taken from the UCI Machine Learning Repository, and our findings are compared with the findings reported within the recent literature. The results are quantitatively and qualitatively discussed and our findings reveal that the Logistic regression (LR) classifier achieves the near-optimal performances on the identification of CKD subjects. Hence, we show that ML algorithms serve important function in diagnosis of CKD, with satisfactory robustness, and our findings suggest that LR can also be utilized for the diagnosis of similar diseases. The examinations have accomplished great outcomes in the finding of CKD. In reality, patients may miss a few estimations for different reasons prior to diagnosing. Therefore, we use KNN to fill in the missing qualities and it relies upon the demonstrative classifications of the examples.

IV. SYSTEM MODULES

DATA PROCESSING

Each categorical (nominal) variable was coded to facilitate the processing during a computer. For the values of RBC and PC, normal and abnormal were coded as 1 and 0, respectively. For the values of PCC and BA, present and not present were coded as 1 and 0, respectively. For the values of HTN, DM, CAD, PE and ANE, yes and no were coded as 1 and 0, respectively. For the worth of APPET, good and poor were coded as 1 and 0, respectively. Although the original data description denes three variables sg, AL and SU as categorical types, the values of these three variables are still numeric based, thus these variables were treated as numeric variables. All the categorical variables were transformed into factors. Each sample was given an independent number that ranged from 1 to 400. There is a large number of missing values in the data set, and the number of complete instances is 158. In general, the patients might miss some measurements for various reasons before making a diagnosis. Thus, missing values will appear in the data when the diagnostic categories of samples are unknown, and a corresponding imputation method is needed.

EXTRACTING FEATURE SELECTION

Removing the variables that are neither useful for prediction nor related to response variables and thus preventing these unrelated variables from the models to make an accurate prediction. Here, we use LR to extract the variables that are most meaningful to the prediction. It

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 4 Apr. 2021, pp: 22-26 www.ijaem.net ISSN: 2395-5252

detects the contribution of each variable in the Gini index. The larger the Gini index, the higher the uncertainty in classifying the samples. Therefore, the variables with contribution of 0 are treated as redundant variables. The step of feature extraction was run on complete data set and the combinations are ranked from left to right by the degree the vertical axis represents variables. The horizontal axis represents the degree to which the combination of variables explains the response variable. To distinguish each combination of variables, we used four colors (maroon, skin, grey and blue) to mark the selected variables. The combinations are ranked from left to right by the degree of explanations to the response variable and the right-most combination has the strongest interception to the response variable.

PERFORMANCE INDICATOR

In this study, CKD was set to be positive and not-CKD was set to be negative. The confusion matrix was used to show the specific results and evaluate the performance of the machine learning models. True positive (TP) indicates the CKD samples were correctly diagnosed. False negative (FN) indicates the CKD samples were incorrectly diagnosed. False positive (FP) indicates the not-CKD samples were incorrectly diagnosed. True negative (TN) indicates the not-CKD samples were correctly diagnosed. Accuracy is used to evaluate the performance of the model. They are calculated using the following equations

$$\text{Accuracy} = \frac{TP+TN}{TP+FN+FP+TN}$$

Establishing & Evaluating Individual Models

The following machine learning models have been obtained by using the corresponding subset of features or predictors on the complete CKD data sets for diagnosing CKD.

1) Logistic Regression: LR
2) Distance-based model: KNN

Generally, in disease diagnosis, diagnostic samples are distributed in a multidimensional space. This space comprises predictors that are used for data classification (CKD or not-CKD). 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. According to the effectiveness of classification, we choose the aforementioned methods for disease diagnosis. LR obtains the weight of each predictor. If the sum of the effects of all predictors exceeds a threshold, the category of the sample will be classified as CKD or not-CKD. The final decision is determined by the predictions of all trees in the disease diagnosis. KNN 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. KNN can analyze non-linear relationships.

V. CONCLUSION

The proposed CKD diagnostic methodology is feasible in terms of data imputation and samples diagnosis. After unsupervised imputation of missing values in the data set by using KNN imputation, the integrated model could achieve a satisfactory accuracy. In this assessment, we propose an KNN and Logistic Regression, system for diagnosing CKD Hence, we speculate that applying this methodology to the practical diagnosis of CKD would achieve a desirable effect. In addition, this methodology might be applicable to the clinical data of the other diseases in actual medical diagnosis. However, in the process of establishing the model, due to the limitations of the conditions, the available data samples are relatively small, including only 400 samples.

Therefore, the generalization performance of the model might be limited. In addition, due to there are only two categories (CKD and not-CKD) of data samples in the data set, the model cannot diagnose the severity of CKD. In the future, a large number of more complex and representative data will be collected to train the model to improve the generalization performance while enabling it to detect the severity of the disease. We believe that this model will be more and more perfect by the increase of size and quality of the data.

ACKNOWLEDGEMENT

This work is supported by the Staffs in Department of Computer Science and Engineering in K. S. Rangasamy College of Technology (2020-2021).

REFERENCES

[1]. M. M. Hossain et al., "Mechanical anisotropy assessment in kidney cortex using ARFI peak displacement: Preclinical validation and pilot in vivo clinical results in kidney allografts," IEEE Trans. Ultrason. Ferr., vol. 66, no. 3, pp. 551-562, Mar. 2019.

[2]. E. Hodneland et al., "In vivo detection of chronic kidney disease using tissue deformation fields from dynamic MR imaging," IEEE Trans. BioMed. Eng., vol. 66, no. 6, pp. 1779-1790, Jun. 2019.

[3]. G. R. Vasquez-Morales et al., "Explainable prediction of chronic renal disease in the

colombian population using neural networks and case-based reasoning," IEEE Access, vol. 7, pp. 152900-152910, Oct. 2019.

[4]. Z. Gao et al., "Diagnosis of diabetic retinopathy using deep neural networks," IEEE Access, vol. 7, pp. 3360-3370, Dec. 2018.

[5]. X. Wang et al., "A new effective machine learning framework for sepsis diagnosis," IEEE Access, vol. 6, pp. 48300-48310, Aug. 2018.

[6]. J. Aljaaf et al., "Early prediction of chronic kidney disease using machine learning supported by predictive analytics," in Proc. IEEE Congr. Evolutionary Computation, Jul. 2018.

[7]. W. H. S. D. Gunarathne, K. D. M. Perera and K.A.D.C.P. Kahandawaarachchi, "Performance evaluation on machine learning classification techniques for disease classification and forecasting through data analytics for chronic kidney disease (CKD)," in Proc. IEEE 17th Int. Conf. Bioinformatics and Bioengineering, Oct. 2017, pp. 291-296.

[8]. N. Almansour et al., "Neural network and support vector machine for the prediction of chronic kidney disease: A comparative study," Compute. Biol. Med., vol. 109, pp. 101-111, Jun. 2019.

[9]. M. Alloghani et al., "Applications of machine learning techniques for software engineering learning and early prediction of students' performance," in Proc. Int. Conf. Soft Computing in Data Science, Dec. 2018, pp. 246- 258.

[10]. L. Du et al., "A machine learning based approach to identify protected health information in Chinese clinical text," Int. J. Med. Inform., vol. 116, pp. 24-32, Aug. 2018.

[11]. R. Abbas et al., "Classification of foetal distress and hypoxia using machine learning approaches," in Proc. Int. Conf. Intelligent Computing, Jul. 2018, pp. 767-776.

[12]. M. Mahyoub, M. Randles, T. Baker and P. Yang, "Comparison analysis of machine learning algorithms to rank alzheimer's disease risk factors by importance," in Proc. 11th Int. Conf. Developments in eSystems Engineering, Sep. 2018.

[13]. Q. Zou et al., "Predicting diabetes mellitus with machine learning techniques," Front. Genet., vol. 9, Nov. 2018.

[14]. N. Park et al., "Predicting acute kidney injury in cancer patients using heterogeneous and irregular data," Plos One, vol. 13, no. 7, Jul. 2018.

[15]. Subasi, E. Alickovic, J. Kevric, "Diagnosis of chronic kidney disease by using random forest," in Proc. Int. Conf. Medical and Biological Engineering, Mar. 2017, pp. 589-594.

International Journal of Advances in Engineering and Management
ISSN: 2395-5252

IJAEM

Volume: 03 Issue: 04 DOI: 10.35629/5252

www.ijaem.net
Email id: ijaem.paper@gmail.com