Early Detection Of Chronic Kidney Disease Using Machine Learning

-

ABSTRACT

The huge amounts of data generated by healthcare transactions are too complex and voluminous to be processed and analyzed by traditional methods. Data mining provides the methodology and technology to transform these mounds of data into useful information for decision making. The Healthcare industry is generally “ information rich” , which is not feasible to handle manually. These large amounts of data are very important in the field of data mining to extract useful information and generate relationships amongst the attributes. Kidney disease is a complex task which requires much experience and knowledge. Kidney disease is a silent killer in developed countries and one of the main contributors to disease burden in developing countries. In the health care industry the data mining is mainly used for predicting the diseases from the datasets.

INTRODUCTION

Machine learning and data processing play a vital role in getting more flexible and understandable reports on the idea of varied techniques. Kidneys role act as blood purifiers that remove waste contents while preserving new valuable blood contents like proteins. If the purifiers were damaged, the protein content would be initially leaked, and the substances may seep into urine from the blood. Sometimes the chronic renal disorder is amid high vital sign, which not only is often caused by kidney damage but also further accelerates kidney injury and maybe a significant reason for the adverse effects of chronic renal disorder on other body parts automatically increases the risk of a heart condition and heart-strokes, collection of excess body fluids, anaemia, weakening of bones and deterioration mainly the body will not support for medications. It cannot be detected until the seriousness of the disease is advanced. If detected early, treatment can hamper or refrain kidney function and deny and reduce the opposite effects on new body parts. A biopsy measuring tool called glomerular filtration rate works on the kidneys for removing waste blood contents called creatinine. If the value lies within the range of 60 to 90, it is an early sign of occurring kidney disease; a worth below 60 is typically considered as an abnormal phase.1 Testing urine samples gives the results of protein contents (albumin) within the urine; repeated results of 30

mg or more can signify a drug. Huge vital signs can also point to underlying chronic renal disorder. Distinct machine learning procedures are appropriate for analyzing the data From distinct prospects and reviewing them into useful data.Machine Learning is an application of artificial intelligence(AI) that gives systems the capacity to use analytical strategies to give computers the ability to learn with information and improve from experience without being explicitly configured.

DATA MINING

Data mining is the non-trivial extraction of implicit previously unknown and potentially useful information about data. Data Mining is one of the most vital and motivating area of research with the objective of finding meaningful information from huge data sets. In present era, Data Mining is becoming popular in healthcare field because there is a need of efficient analytical methodology for detecting unknown and valuable information in health data. Medical data mining is used in the knowledge acquisition and analyses the information obtained from research reports, medical reports, flow charts, evidence tables, and transform these mounds of data into useful information for decision making.

LITERATURE REVIEW

This section consists of the reviews of various technical and review articles on data mining techniques applied to predict Kidney Disease.

- DSVGK Kaladhar, Krishna Apparao Rayavarapu and Varahalarao Vadlapudietal . described in their research to understand machine learning techniques to predict kidney stones. They predicted good accuracy with C4.5, Classification tree and Random forest (93%) followed by Support Vector Machines (SVM) (91.98%). Logistic and NN has also shown good accuracy results with zero relative absolute error and 100% correctly classified results. ROC and Calibration curves using Naive Bayes has also been constructed for predicting accuracy of the data. Machine learning approaches provide better results in the treatment of kidney stones.

- J.Van Eyck, J.Ramon, F.Guiza, G.Meyfroidt, M.Bruynooghe, G.Van den Berghe, K.U.Leuven explored data mining techniques for predicting acute

kidney injury after elective cardiac surgery with Gaussian process & machine learning techniques (classification task & regression task).

- K.R.Lakshmi, Y.Nagesh and M.VeeraKrishna presented performance comparison of Artificial Neural Networks, Decision Tree and Logical Regression are used for Kidney dialysis survivability. The data mining techniques were evaluated based on the accuracy measures such as classification accuracy, sensitivity and specificity. They achieved results using 10 fold cross-validations and confusion matrix for each technique. They found ANN shows better results. Hence ANN shows the concrete results with Kidney dialysis of patient records.

- Morteza Khavanin Zadeh, Mohammad Rezapour, and Mohammad Mehdi Sepehri et al [9]l. described in their research by using supervised techniques to predict the early risk of AVF failure in patients. Suman Bala used classification approaches to predict probability of complication in new hemodialysis patients whom have been referred by nephrologists to AVF surgery.

- Abeer Y. Al-Hyari .proposed in their research by using Artificial Neural Network (NN), Decision Tree (DT) and Naïve Bayes (NB) to predict chronic kidney disease. The proposed NN algorithm as well as the other data mining algorithms demonstrated high potential in successful kidney disease.

- Xudong Song, Zhanzhi Qiu, Jianwei Mu introduced data mining decision tree classification method, and proposed a new variable precision rough set decision tree classification algorithm based on weighted limit number explicit region.

DATA MINING TECHNIQUES USED FOR PREDICTIONS

Decision Tree:

The decision tree is a structure that includes root node, branch and leaf node. Each internal node denotes a test on attribute, each branch denotes the outcome of test and each leaf node holds the class label. The topmost node in the tree is the root node. The decision tree approach is more powerful for classification problems. There are two steps in this techniques building a tree &

applying the tree to the dataset. There are many popular decision tree algorithms CART, ID3, C4.5, CHAID, and J48.

Artificial Neural Network (ANN):

It is a collection of neuron – like processing units with weight connections between the units. It maps a set of input data onto a set of appropriate output data. It consists of 3 layers: input layer, hidden layer & output layer. There is connection between each layer & weights are assigned to each connection. The primary function of neurons of input layer is to divide input xi into neurons in hidden layer. Neuron of hidden layer adds input signal xi with weights wji of respective connections from input layer. The output Yj is function of $Yj = f (\Sigma wji xi)$ Where f is a simple threshold function such as sigmoid or hyperbolic tangent function.

Naive Bayes:

Naive Bayes classifier is based on Bayes theorem. This classifier algorithm uses conditional independence, means it assumes that an attribute value on a given class is independent of the values of other attributes. The Bayes theorem is as follows: Let X={x1, x2... xn} be a set of n attributes. In Bayesian, X is considered as evidence and H is some hypothesis means, the data of X belongs to specific class C. We have to determine P (H|X), the probability that the hypothesis H holds given evidence i.e. data sample X. According to Bayes theorem the P (H|X) is expressed as $P (H|X) = P (X| H) P (H) / P (X)$.

K-***Nearest Neighbour:***

The k-nearest neighbour' s algorithm (K-NN) is a method for classifying objects based on closest training data in the feature space. K-NN is a type of instance-based learning. The k-nearest neighbour algorithm is amongst the simplest of all machine learning algorithms. But the accuracy of the k-NN algorithm can be severely degraded by the presence of noisy or irrelevant features, or if the feature scales are not consistent with their importance.

Logistic Regression:

The term regression can be defined as the measuring and analyzing the relation between one or more independent variable and dependent variable. Regression can be defined by two categories; they are linear regression and logistic regression. Logistic regression is a generalized by linear regression. It is mainly used for estimating binary or multi-class dependent variables and the response variable is discrete, it cannot be modelled directly by linear regression i.e. discrete variable changed into continuous value. Logistic regression basically is

used to classify the low dimensional data having non-linear boundaries. It also provides the difference in the percentage of dependent variable and provides the rank of individual variable according to its importance. So, the main motto of Logistic regression is to determine the result of each variable correctly.

Rough Sets: A Rough Set is determined by a lower and upper bound of a set. Every member of the lower bound is a certain member of the set. Every non-member of the upper bound is a certain non-member of the set. The upper bound of a rough set is the union between the lower bound and the so-called boundary region. A member of the boundary region is possibly (but not certainly) a member of the set. Therefore, rough sets may be viewed as with a three-valued membership function (yes, no, perhaps). Rough sets are a mathematical concept dealing with Uncertainty in data. They are usually combined with other methods such as rule induction or clustering methods.

Support Vector Machine (SVM): Support vector machine (SVM) is an algorithm that attempts to find a linear separator (hyper-plane) between the data points of two classes in multidimensional space. SVMs are well suited to dealing with interactions among features and redundant features.

DATA ACQUITION

Patient characteristics included age, gender, education level, marriage status, and insurance status. Medical history comprised history of smoking, history of alcohol consumption, presence of each comorbid condition— diabetes, cardiovascular disease and hypertension. Clinical parameters contained body mass index (BMI), systolic pressure and diastolic pressure. Blood tests consisted of serum creatinine, uric acid, blood urea nitrogen, white blood cell count, hemoglobin, platelets count, aniline aminotransferase (ALT), aspartate aminotransferase (AST), total protein, albumin, alkaline phosphatase (ALP), high-density lipoprotein, low-density lipoprotein, triglycerides, total cholesterol, calcium, phosphorus, potassium, sodium, chloride, and bicarbonate. Te estimated glomerular filtration rate and type of primary kidney disease were also used as predictors. All baseline variables were obtained at the time of subject enrollment.

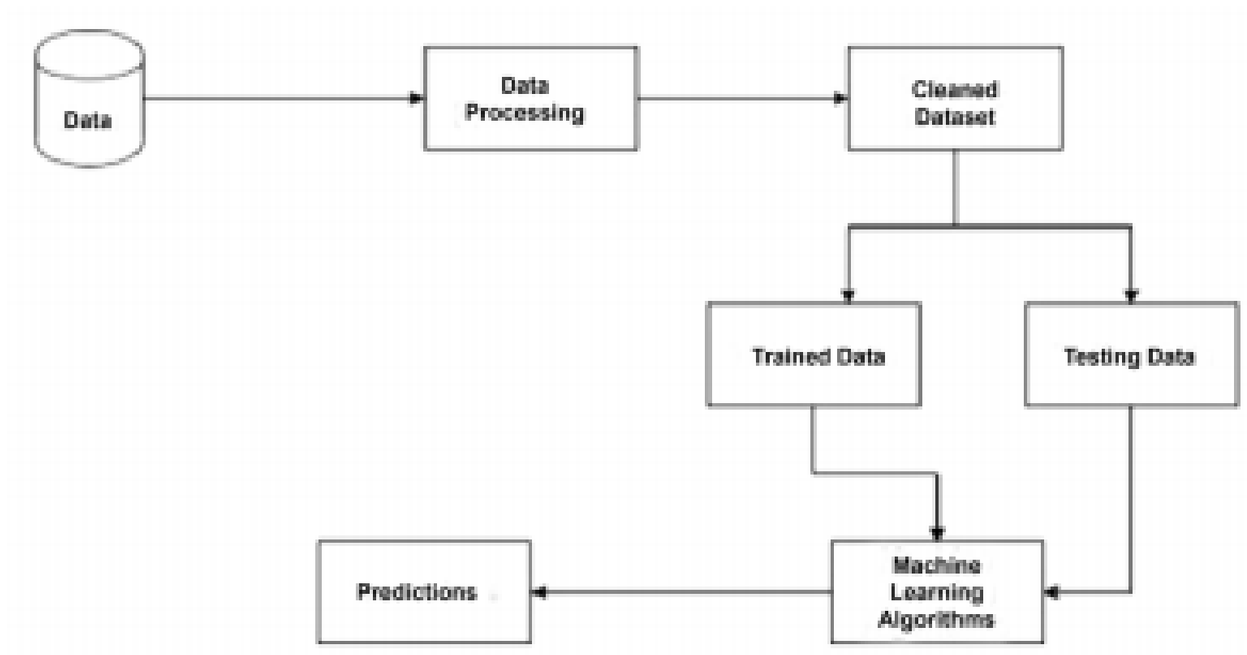
DATA PRE PROCESSING

All categorical variables, such as insurance status, education, and primary disease, were encoded using the one-hot approach. Any variable was removed from model development if the missing values were greater than 50%. Missing data were handled using multiple imputation with five times of repetition,

leading to five slightly different imputed datasets where each of the missing values was randomly sampled from their predictive distribution based on the observed data.

On each imputed set, all models were trained and tested using a fivefold cross validation method. To minimize selection bias, subject assignment to train/test folds was kept consistent across all imputed sets. Data were split in a stratified fashion to ensure the same distribution of the outcome classes (ESKD+ vs. ESKD−) in each subset as the entire set.

GENERAL EXECUTION OF MACHINE LEARNING CLASSIFIER



EXISTING SOLUTION

In the existing system, the previous predictions of the Persistent Kidney Illness are to tell the accurate values to use some of the algorithms in the previous predictions. Early acknowledgement and brief usage of prescribed administration rules are necessary to forestall intensifying kidney work and cardiovascular horribleness in patients with beginning time CKD.4 A significant obstacle in accomplishing these objectives is the thing that might be the absence of acknowledgement by the essential consideration of doctors that their patients have beginning time CKD. Acknowledgement could be in a roundabout way surveyed by the nearness or nonappearance of CKD documentation comprising of words or ideas that convey the nearness of CKD.

PROPOSED SOLUTION

In the suggested system, even more, forecasts must be done. Chronic Kidney Illness is a very harmful health issue that has been spreading out along with expanding because of diversification in lifestyle such as food routines, changes in the ambience, and so on. This project' s main objective is to identify making use of different Classification techniques, and we need to identify the best of the classifiers as We select the dataset of the data containing the previous data related to a database that is used to produce accurate results or predictions that could be better than the existing system we have. So that what we have proposed in this project using the six algorithms, we will find out the accurate result better than the existing or previous.

REFERENCES

1.Radhakrishnan J, Mohan S. KI Reports and World Kidney Day. Kidney international reports. 2017 Mar 1;2(2):125-6.

1. 2. Alaoui SS, Aksasse B, Farhaoui Y. Statistical and Predictive Analytics of Chronic Kidney Disease. In International Conference on Advanced Intelligent Systems for Sustainable Development 2018:27-38. Springer, Cham.

2. 3. Sandeep Reddy Mula, Jaya. CKD Analysis Using Machine Learning Algorithms. International Journal for Research in Engineering Technology. 2018;6:3367-79.

3. 4. Aljaaf AJ, Al-Jumeily D, Haglan HM, Alloghani M, Baker T, Hussain AJ, Mustafina J. Early prediction of chronic kidney disease using machine learning supported by predictive analytics. In2018 IEEE Congress on Evolutionary Computation (CEC) 2018: 1-9. IEEE.

4. Arora M, Sharma EA. Chronic Kidney Disease Detection by Analyzing Medical Datasets in Weka. International Journal of Computer Application. 2016 Jul;6(4):20-6. 6. ClassificationTechniques:Hint:https://www.edureka.co/blog/classification-algorithms/, accessed on June 12, 2020

7. Centres for Disease Control and Prevention. Chronic Kidney Disease Surveillance System website. 2019 Last Accessed: https://nccd.cdc.gov/CKD. Accessed January 7, 2016