ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)
Vol 5, Issue 2, February 2020

Analysis and prediction of Chronic Kidney Disease using Machine Learning Algorithms

[1] Srinitya G, [2] Daniel Madan Raja S

[1][2] Bannari Amman Institute of Technology, Alathukombai, Sathyamangalam, Tamil Nadu, India

***Abstract*—Health is Wealth: Today the world has taken a step forward where each individual is concerned about what he is consuming on a daily basis and analyses the after effects of the food. Every individual is more concerned about his/her everyday food habits and tries to adapt himself to what nature provides him. We are moving towards a technology oriented living where computers in general and data science and analysis in particular plays a major role in every field. A recent survey from World Health Organization (WHO) tells us that the growth of ageing population may increase by 50% in the forth coming decade. Here, in this paper we mainly concentrate on kidney related issues, and try to predict the presence of chronic kidney disease based on certain parameters available from UCI dataset using decision tree based approach.**

***Keywords*---health, chronic kidney disease, machine learning, decision tree**

I. OVER VIEW

WHO survey states that between 2020 and 2050 the number of aged people is supposed to double from 11% to 22% across the globe [1]. Chronic kidney disease is a conditions that causes damage to our kidneys and decrease their ability to function normally and does not keep us healthy thereby affecting our day-to-day routine. When kidney disease worsens, wastes may accumulate to high levels in our blood and makes us feel sick and lazy. Complications such as high blood pressure, anemia, weak bones, and nerve damage may occur leading us to totally get bedridden. Early detection of malfunctioning of kidneys and treatment can help chronic kidney disease from getting worse. When the same condition progresses, the situation may lead to failure of kidneys where they may have to be replaced or the patient may be put to dialysis for his lifetime. Though kidney disease is common among all ages, survey reports reveal that the percentage of people affected by kidney failures generally fall under the above 55 category. This usually makes the person immobilized and deprives him from doing his daily activities.

To detect chronic kidney disease the authors implemented (LDA) Linear Discriminant Analysis and (CSP) common spatial pattern filter[3]. [4] Different classification techniques were applied on patient's record available and the authors proved that adial basis function gives better results

[8] worked on Naïve Bayesian and k-nearest neighbour algorithms and used it to predict the disease. They proved according to their test results that k-nearest neighbour shows more accurate results than naïve Bayesian. [9] the authors are using Datasets to store medical records. They used support vector machine and Bayesian network to predict kidney disease and select the efficient one among them. [11] SVM and KNN classifiers are compared by the authors and based on their accuracy and execution time for CKD prediction they proved KNN classifier is better.

II. PROPOSED MODEL

Figure 1 shows the proposed model for the analysis to be carried out. The data-set with patient data is considered for pre-processing, which includes removing duplicated, filling up of empty locations. This paper focuses on decision tree and C4.5 algorithm the data set was not completely filled because the algorithm works well for discrete values. The next step after pre-processing is to train the dataset and construct a decision tree for each individual data. After training the dataset, this is now used to test the remaining set of values and the result shown indicates whether a patient is affected by Chronic Kidney Disease (CKD) or not (NCKD).

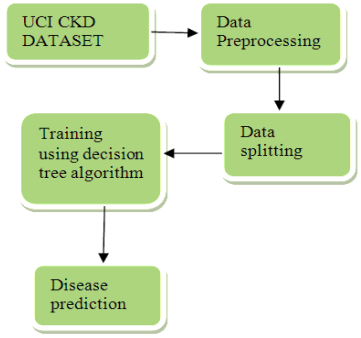

Fig: 1

ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)
Vol 5, Issue 2, February 2020

III. DATA DESCRIPTION

Parameter	Description
age	age (n)
bp	blood pressure (n)
sg	specific gravity
al	albumin
su	sugar
rbc	red blood cells
pc	pus cell
pcc	pus cell clumps
ba	bacteria
bgr	blood glucose random (n)
bu	blood urea (n)
sc	serum creatinine (n)
sod	sodium (n)
pot	potassium (n)
hemo	hemoglobin (n)
pcv	packed cell volume (n)
wc	white blood cell count (n)
rc	red blood cell count (n)
htn	hypertension
dm	diabetes mellitus
cad	coronary artery disease
appet	appetite
pe	pedal edema
ane	anemia
class	class

Table: 1

The dataset for this experiment was taken from the UCI source and it contains patient data with 24 attributes, most of them are clinical and the rest are physiological. Some of the attributes are numerical and some are nominal. The numerical values are indicated in the table by (n) (Table: 1).

IV. IMPLEMENTATION

The implementation begins with the collection of raw data and pre-processing it. The data is then sampled and split into training data and testing data. Training data-set is cleaned up and is trained by using the learning algorithm. The results obtained are optimized. The data is then validated for the correctness of its classification. Lastly, the evaluation of the test data set is performed and classified using the decision tree created for the training data set.

Data from the test data set will be entered and for every input decision tree will be generated by calculating entropy and information gain values as per the rules of c4.5 algorithm explained in the section below. From the root to the leaf node the place at which the incoming node is to be placed will be calculated and it depends upon the homogeneity of the node. Prediction process usually occurs at the leaf node in a decision tree. The results obtained on the classification of CKD and NCKD is listed in Table: 2 and Fig: 2.

V. C4.5 ALGORITHM

C4.5 algorithm builds a decision tree for every input from the training dataset using the concept of information entropy. The training dataset consists of already classified samples.

In each node of the tree the algorithm classifies that particular attribute of the data that effectively splits the samples into subsets that deepen onto one class or the other. The splitting criterion is the normalized information gain (difference in entropy). The attribute that has the highest normalized information gain is chosen to make the decision. This process progresses recursively on the partitioned subsets as well.

This algorithm to start up should have a few base cases.

- All chosen samples in the list belongs to the same class. In this case a leaf node is created stating that the node belongs to the base class.
- The node does not provide any information gain. In this case a node is created higher up in the decision tree meaning to choose that class.
- Instance of previously-unseen class is encountered. In this case also a node is created higher up in the decision tree meaning to choose that class.

Algorithm:

The general algorithm for building decision trees is:

1. Start the process
2. Check each attribute for the above base cases
3. Find the normalized information gain of the attribute
4. Create a decision node that splits on the normalized information gain
5. Repeat on the sub-lists by splitting up on the normalization gain and add those nodes that satisfies the criteria as a child node

VI. RESULTS

classification	% classification TP	% classification TN
CKD	99.25	0.75
NCKD	98.75	1.25

Table: 2

ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)
Vol 5, Issue 2, February 2020

The results obtained clearly indicate that the algorithm proposed in this paper classifies the given data to a convincing extent. With more parameters to be processed and introducing a higher level of precision may show considerable improvement in the results of classification.

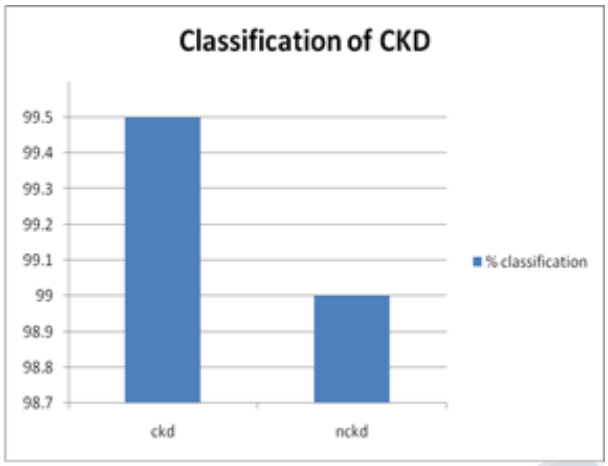

Fig: 2

VII. CONCLUSION

Computer vision especially machine learning works absolutely good in predicting health statistics of humans obtained by clinical diagnosis. Prevention is better than cure, yes but prediction of disease earlier is better to treat people in an effective manner and can save the patient by helping him to get back to his normal routine after a prediction. Many advances in machine algorithms aids us to do this prediction accurately. In this paper, c4.5 learning algorithm is used to predict patients with chronic kidney failure (ckd) disease and patients who do not (nckd) suffer from the disease. The results obtained from applying machine learning algorithms for these types of predictions seems to be convincing and better implementation of computer vision into medical diagnosis help us to do more research of this kind in future.

REFERENCES

[1] https://www.who.int/ageing/10-priorities/en/ accessed on 21 Feb 2019.
[2] https://www.who.int/ageing/en/ accessed on 21 Feb 2019.
[3] Ling Yu, Duke Billie J. et al. Exosomal Gapdh from proximal tubule cells regulate ENaC activity Nov 2016
[4] Verhaar MC, knepper MA et al. Exosomes and the kidney: prospects for diagnosis and therapy of renal diseases. Kidney Int. 80:1138-1145 Aug 2011.
[5] Mirja k, Samoylenko A et al. Exosomes as renal inductive signals in health and disease, and their application as diagnostic markers and therapeutic agents. Front. Cell Dev. Biol. 2015 Oct.
[6] GuillermoGarcia-Garcia, kunitoshi Iseki et al. Chronic kidney disease: global dimension and perspectives. May 2013.
[7] Tharmarajah Thiruvaran et al., IEEE Identifying important attributes for early detection of chronic kidney disease.
[8] Shuo Yang, Ran Wei et al. Semantic inference on clinical documents: Combining Machine learning algorithms with an inference engine for effective clinical diagnosis and treatment (2017)
[9] Meenambal S. et al. Velocity bounded Boolean particle swarm optimization for improved feature selection in liver and kidney disease diagnosis. Expert Syst. Appl. 28-47 (2016).
[10] J Stankovic, Salekin A Detection of chronic kidney disease and selecting important predictive attributes in healthcare informatics (ICHI), IEEE, pp.262-270, oct 2016.
[11] Arora M, Sharma EA. Chronic kidney disease detection by analyzing medical datasets in Weka. Int. Journal. Comput. Appl; vol-6: 20-26, Aug 2016.
[12] D.K. Vawdrey, T.L. Sundelin, K.E. Seamons and C.D. Knustson, "Trust negotiation for authentication and authorization in health care information system," 25th Annual International Conference of IEEE, vol. 2, issue, pp. 1406-1409, 17-21 September 2003.
[13] https://archive.ics.uci.edu/ml/machine-learning-databases/00336/