International Journal for Research in Engineering Application & Management (IJREAM)
ISSN : 2454-9150 Special Issue - NGEPT - 2019

Detection of Pediatric Kidney Disease Based on Data Mining and Machine Learning Using Python

[1]Megha Jangid, [2]Mrs. Jayashree M, [3]Mary Stephina Gilbert, [4]Monika Sharma, [5]Monisha M

[1,3,4,5]Student, [2]Associate Professor, EPCET Bengaluru, India, [1]megha.jangid97@gmail.com, [2]Jayashree.raju2006@gmail.com, [3]stephina.gilbert6@gmail.com, [4]monikasharmar015@gmail.com, [5]m674622@gmail.com

Abstract -As electronic health records are continuously increasing, we add clinical documentation as well as laboratory values and demographics into risk prediction modeling. Data Mining in Healthcare has become a present trend for obtaining accurate results of medical diagnosis. Pediatric Chronic Kidney Disease (CKD) has become an international fitness problem and is a place of concern, it is a situation where kidneys turn out to be damaged and cannot filter toxic wastes within the frame. By using Data Mining Techniques, researchers have the scope to predict the Chronic Kidney Disease at early stage. By keeping this in mind, in our project we plan to develop a prediction model for Chronic Kidney Disease (CKD) progression for children based on the clinical data. We perform feature selection to determine the most relevant attributes for detecting CKD and rank them according to their predictability. When the number of classes are large, and the biases are increased, the Gini-based decision tree method is modified to overcome the known problems, by normalizing the Gini indexes, Instead of using the Gini index for attribute selection, ratios of Gini indexes are used and their splitting values in order to reduce the biases.

Keywords —CKD (chronic kidney disease), Gini index, Data Mining, medical diagnosis, Pediatric.

I. INTRODUCTION

Recent advancements in healthcare are helpful for the specialists for making suitable selections and improving the quality of living of the diseased person. Patients with comparable health problems may be grouped collectively and effective remedy plans can be recommended based totally on patient's records, bodily examination, prognosis and former remedy styles. Data Healthcare in Healthcare using Data Mining techniques is mainly useful in medical discipline where no availability is there for the proof of favoring a selected treatment alternative is located. A Huge volume of complicated information is created constituting patients data, disease record, hospitals bills, medical equipment's, insurance claims, and treatment price and so on. That requires evaluation and processing for extracting useful knowledge from this data. Data mining constitutes of several methodologies and algorithms which is performed on this processed information. With the advancement achieved in machine learning, the incorporation of machine learning in data mining can result in accurate prediction model with accurate results.

The field of clinical disease risk prediction and progression is well developed, with hundreds of models published across many diseases. Given their history predating electronic health records (EHRs), these models have largely been developed with data easily accessible to clinicians. Likewise, current progression risk models for chronic kidney disease (CKD) largely rely on commonly obtained laboratory or vital sign data. CKD affects a large portion of the population is associated with significant morbidity and mortality and is a high-risk clinical condition with frequent adverse events. Despite this, patients with kidney disease frequently go unrecognized, and their care is often suboptimal. Early identification and more accurate prognostication of these patients using better risk prediction models may improve outcomes by facilitating timelier initiation of appropriate therapies, monitoring, and specialty referral.

II. LITERATURE SURVEY

A. Support vector machine-based classification

The SVM has considered as a machine learning algorithm for the classification of two-class problems. The training of support vector machines with the positive and negative types of data is known as one-against-all or one-against-rest. The two-class type SVM is combined for creating a multi-class support vector machine. The classification by this algorithm is represented as in the equation.

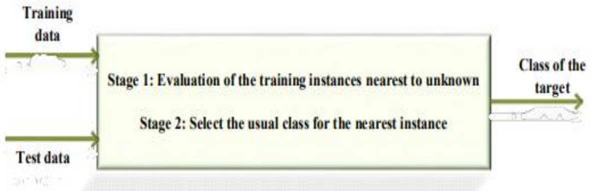
$$F(\theta) = \operatorname{sgn}\left[\left(\sum_{j=1}^{n} \gamma_j x_j \tau(\theta, \theta_j) + b\right)\right]$$

B. K-Nearest Neighbor based classification

The K- Nearest neighbor is the straight forward classifier with greater accuracy. The classification in this algorithm depends upon the similarity measure. The continuity of

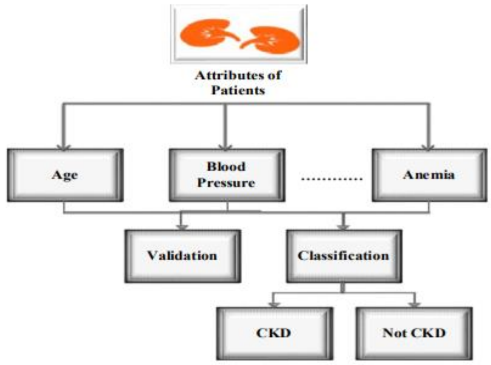
DOI : 10.18231/2454-9150.2019.0573

attributes is very much important in this process. The progression stages in the nearest neighbor algorithm are shown.


Training data
Stage 1: Evaluation of the training instances nearest to unknown
Stage 2: Select the usual class for the nearest instance
Class of the target
Test data

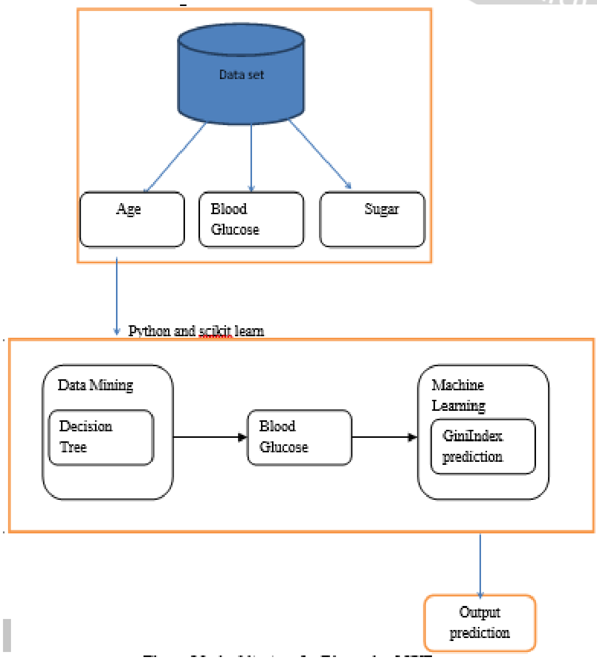
C. *Decision tree-based classification*

This technique is a branching methodology which provides the outputs which are possibly related to the decision. For the generation of proper decision tree approach uses greedy search. Below figure represents the classification of CKD by this technique.



III. Architecture

The clinical data consist of patient's records which are has been considered for the analysis and that dataset is taken from hospitals. The dataset contains certain attributes. Numerical and Nominal values of attributes are considered. Some of the attributes of Chronic Kidney Disease are age, blood pressure, blood glucose level, etc. we analyze and predict the Chronic Kidney Disease and its severity using Chronic Kidney Disease dataset and Data Mining techniques like Classification and Clustering,



Clustering algorithm using decision tree classifier is used where the dataset is fed as input and the corresponding output is fed as input to the prediction algorithm based on gini-Index of scikit machine learning. The clustered output is analyzed by the Classification algorithms and respective results are obtained. Considering all the results obtained, the accuracy is checked to analyze which algorithm is best suitable for prediction of CKD.

IV. Feature Enhancement and Applications

Decision Tree Using Gini index:

Gini index builds decision trees from a set of training data in the same way as id3, using the concept of information entropy. The training data is a set s = s1,s2,... of already classified samples. Each sample si = x1,x2,... is a vector where x1,x2,... represent attributes or features of the sample. The training data is augmented with a vector c = c1,c2,... where c1,c2,... represent the class that each sample belongs to. Gini index uses the fact that each attribute of the data can be used to make a decision that splits the data into smaller subsets.

Gini index examines the normalized information gain (difference in entropy) that results from choosing an attribute for splitting the data. The attribute with the highest normalized information gain is the one used to make the decision. The algorithm then recurs on the smaller sub lists. This algorithm has a few base cases, the most common base case is when all the samples in your list belong to the same class. Once this happens, you simply create a leaf node for your decision tree telling you to choose that class.

It might also happen that none of the features give you any information gain, in this case gini index creates a decision node higher up the tree using the expected value of the class. it also might happen that you've never seen any instances of a class; again, gini index creates a decision node higher up the tree using expected value.

Attribute Selection

Consider a N labeled class pattern partitioned into sets of patterns belonging to classes Ci, i=1,2,3,..l. The population in class Ci is in each pattern has n features and each feature can take two or more values.

For each attribute the Entropy is calculated using the formula:

$$Entropy(C) \equiv H(C) \equiv -\sum_{c} P(C=c)\log_2 P(C=c)$$

Measuring Entropy:

Entropy of class C of set of examples S is

$$H(C) = -p_+ \log_2 p_+ - p_- \log_2 p_-$$

where $p_+ = \frac{p_i}{p_i+n_i}$; p_i positive and n_i negative examples in S

Entropy can be used to measure gain in information about class by branching on attribute a – compute reduction in entropy of class c caused by knowing a value.

Information Gain

$$Gain(C,A) = I(C,A) = H(C) - \sum_{v \in V(A)} P(A=v)H(C|A=v)$$

This measures the information between A and C: the amount of information we learn about C by knowing the value of A.

In traditional Gini index classification, the Gini index measure is used as a heuristic for selecting the attribute that will best partition the training tuples into individual classes. The selected attribute is then used as the testing one at the node of the tree. Let D be a database consisting of |D| = d data tuples. Assume that the class label attribute has n distinct values representing n different classes C1,C2 . . .Cn. gini(D) is defined as

$$gini(D) = 1 - \textstyle\sum_{i=1}^{n} p_i^2$$

where $p_i = \frac{|C_i|}{d}$ is the relative frequency of class C_i in D.

For an attribute A with m distinct values, the database D is partitioned into m subsets D1,D2 . . .Dm. The Gini index of D with respect to the attribute A is defined as

$$gini_A(D) = \textstyle\sum_{i=1}^{m} \frac{|D_i|}{d} \cdot gini(D_i)$$

The reduction in impurity of D with respect to the attribute A is defined as

$$\Delta gini(A) = gini(D) - gini_A(D).$$

In traditional Gini-based classification, the attribute provides the largest reduction in impurity is chosen to split the node. However, it is well known that the method biases multivalued attributes. In addition to having difficulty when the number of classes is large, the method also tends to favour tests that result in equal sized partitions and purity in all partitions.

To overcome these known problems, we normalize the Gini indexes by taking into account information about the splitting status of subsets D1,D2 . . .Dm. The splitting status of D with respect to the attribute A is calculated as

$$split_A(D) = 1 - \sum_{i=1}^{m} \left(\frac{|D_i|}{d}\right)^2.$$

The Gini ratio of D with respect to the attribute A is defined as

$$giniRatio(A) = \Delta gini(A)/split_A(D)$$

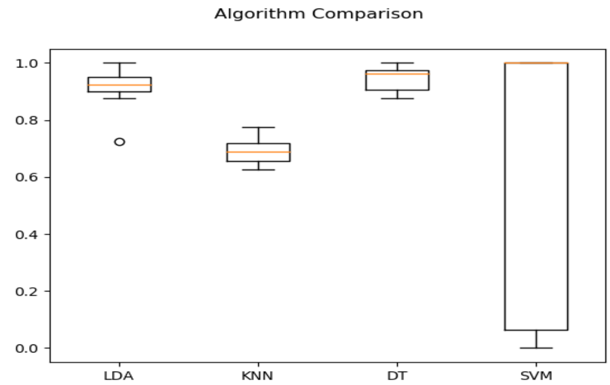
V. Result

A. Figures and Tables

Attribute	Data type
Age	age in years
Blood Pressure	mm/Hg
Specific Gravity	Nominal
Albumin	nominal(1-5)
Sugar	nominal (1-5)
Red Blood Cells	normal,abnormal
Pus Cell	normal,abnormal
Pus Cell clumps	present,notpresent
Bacteria	present,notpresent
Blood Glucose Random	mgs/dl
Blood Urea	mgs/dl
Serum Creatinine	mgs/dl
Sodium	mEq/L
Potassium	mEq/L
Hemoglobin	gms
Packed Cell Volume	nominal
White Blood Cell Count	cells/cumm
Red Blood Cell Count	millions/cmm
Hypertension	yes, no
Diabetes Mellitus	yes, no
Coronary Artery Disease	yes, no
Appetite	good,poor
Pedal Edema	yes,no
Anemia	yes,no

TABLE II

B. Evaluation



VI. Conclusion

The chronic kidney disease can be very well predicted using many classifiers in Data Mining. One can also predict the level of chronic kidney disease using classifiers. As per the observation of different experiments there are some classifiers which gave highest accuracy are Gini Index, entropy, decision tree classifier.

References

[1] Justin Wood, Patrick Tan, Wei Wang, Corey Arnold Department of Computer Science, University of California Los Angeles, CA 90095, USA 2Medical Imaging Informatics Group, University of California, Los Angeles, CA 90095, USA.

[2] R. Subhashini and, M.K. Jeyakumar, Noorul Islam Center for Higher Education, Kanyakumari, Tamil Nadu, India. baskisubha24@gmail.com 2Additional Controller of Examinations, Noorul Islam Center for Higher Education, Kanyakumari, Tamil Nadu, India.

[3] C. H. Jena, C. C. Wang, B. C. Jiangc, Y. H. Chub and M. S. Chen, “Application of classification techniques on development an early-warning systemfor chronic illnesses”.