Proceedings of the Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)
IEEE Xplore Part Number: CFP21OSV-ART; ISBN: 978-1-6654-2642-8

Kidney Disease Diagnosis using Classification Algorithm

S. Pushpalatha PhD
JG College of Computer Applications
Ahmedabad, Gujarat, India
dakshlatha@gmail.com

A. Stella PhD
S. K. Patel Institute of Computer Studies
Gandhinagar, Gujarat, India
rosystella@gmail.com

2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC) | 978-1-6654-2642-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/I-SMAC52330.2021.9640879

***Abstract* — Kidney disease is a larger problem in today's era. Huge population suffers due to kidney problem and mortality rate is increasing largely. Diagnosis of kidney disease and dropping death rate in population is challenging task for health care officials. In this study, we propose the methodology to diagnosis kidney disease using Naive Bayes (NB), Random Forest, Bagging and k-Nearest Neighbor (KNN). Missing value in dataset is handled by Mean imputation technique and transformation is done using LabelEncoder method to transform nominal dataset to numeric dataset. Accuracy percentage and Root Mean Square errors were calculated and compared. From the obtained results Decision Tree algorithm gave an improved accuracy percentage of 100% for diagnosing kidney disease.**

Keywords—Classification, Naïve Bayes, k-nearest neighbor, Random Forest and Accuracy

I. Introduction

Irrespective of any age group the most important disease which gets affected largely is kidney disease. It is considered to be the tenth highest rank of mortality in the global population [8][9]. Major causes for kidney disease are diabetics, blood pressure, improper life style, eating junk foods, lower intake of water, artery disease, etc [10]. Kidney disease can be treated easily with less treatment cost and extended survival rate of patient when if it is identified in early stage. Symptoms for this disease are recognized by patients, but people ignore them which lead to the major risk[8].

Kidney diagnosis in the initial stage is done using blood test [11]. Physician mostly takes decision by comparing the result of previous patient's history. It becomes difficult for physicians to make accurate decision in the diagnosis process with the regular test results. Machine Learning algorithm does a tremendous job to predict and analyse patient's data at earlier stage of disease [3][6].

The flow of this paper is as follows; section 2 describes about description of the algorithms implemented in this paper. Section 3 specifies related work for this finding, in the next section dataset used for performing this study is mentioned. Section 5 specifies the proposed methodology; section 6 describes the algorithmic steps implemented for proposed methodology. Results analysis are done in section 7 and finally conclusion is mentioned in section 8.

II. Machine Learning Algorithms

A. Naïve Bayes Algorithm

Naïve Bayes algorithm, is a probabilistic based classifier which is based on Bayes theorem[6][8]. In Naïve Bayes noisy data and missing dataset are handled effectively. Naïve Bayes classifier predicts independent feature with predicator (Y) to class (X), it derives probabilistically with given tuple of a specific class using (1), P(X) denotes the class attribute present in the dataset.

$$P(X|Y) = \frac{P(Y \mid X)\, P(X)}{P(Y)} \quad (1)$$

B. K-Nearest Neighbor algorithm

K-Nearest Neighbor algorithm follows a non-parametric lazy learning method [7]. It trains all the classifier based on the distance measure using either Euclidian distance measure or Manhattan distance measure represented in (2) and (3). K-Nearest Neighbor classifies only integer dataset for performing classification, so data type conversion is recommended for this algorithm.

Euclidian distance measure: $\sum_{i=n}^{k=1} (x_i - y_i)^2$ (2)

Manhattan distance measure: $\sum_{i=n}^{k=1} |x_i - y_i|$ (3)

C. Bagging

Bagging is a powerful learning method which follows a voting method to classify the dataset using bootstrap method to generate training dataset [12]. Bootstrap aggregation is also done based on the bagging to obtain the subsets of data for training the learners. N different trees are trained independently and average is obtained using (4),

$$f(y) = \frac{1}{N} \sum_{n=1}^{N} f_n(y) \quad (4)$$

where, $f_1(y), f_2(y), \ldots f_n(y)$ are the training set from dataset to build prediction model using N separate training sets, and average them to obtain a low variance learning model.

D. Random Forest

Random forest grows multiple decision trees during training process [4] [12]. Voting mechanism is followed to select the most voted class of the prediction result. This method first performs bootstrap sampling with N sample cases at random. Later, in each node the process first randomly selects n variables and at last it grows into a full tree without pruning. From this we can obtain the higher most predicted result from each single tree. Majority vote can be obtained by taking an average or weighted average of trees.

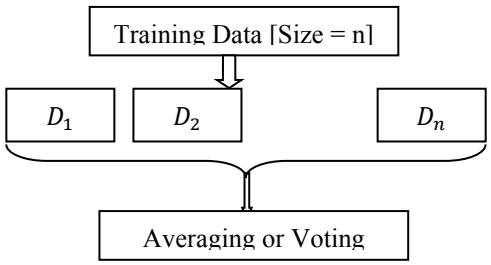

Fig.1 Structure of Random Forest

Random forest uses two parameters named ntree and mtry, ntree specifies total number of trees used and mtry specifies number of features.

E. *Decision Tree Induction*

Top-down recursive tree induction is also known as decision tree induction, it uses an attribute selection measure to identify attribute for every node to build a decision tree. ID3, [Classification and Regression Tree] and ID5 are the three most popular algorithm which follows greedy non-backtracking approaches in decision tree induction. All these three algorithms follow same pattern of divide and conquer approaches. Clinical datasets are partitioned into training and test dataset, training dataset recursively partitions into smaller subset to form a tree.

III. Related Work

El-Houssainy et.al [2] presented a model for predicting stage of kidney disease using data mining algorithms for early detection and characterization. Probabilistic Neural Networks, Multilayer Perceptron, Support Vector Machine and Radial Basis Function algorithms are used in this study. These algorithms are compared among these Probabilistic Neural Networks shows the highest accuracy percentage of 96.7% in comparison with other algorithms.

Zixian Wang et.al. [8] implemented a machine learning based predicting system for chronic kidney disease using Associative Classification Techniques using WEKA. ZeroR, J48, IBK and Naïve Bayes classifiers are used to compare accuracy and error rate of the predicting system. Accuracy rate of 99% were achieved with IBK with Apriori association algorithm.

Marwa Almasoud et.al. [1] implemented the methodology for 'Detection of Chronic Kidney Disease using Machine Learning Algorithms with Least Number of Predictors'. This was conducted for diagnosing chronic kidney disease using logistic regression, support vector machines, random forest, and gradient boosting algorithms. They obtained improved accuracy when implemented with gradient boosting algorithms for detecting kidney disease.

Tekale, S, Shingavi et. al.[5] developed "Prediction of Chronic Kidney Disease Using Machine Learning Algorithm" using decision tree and support vector machine in which support vector machine showed an improved accuracy of 96.75%. In this study they have done pre-processing before implementing the algorithm for prediction.

Sumana De et. al. [4] presented the 'Development of Chronic Kidney Disease Prediction System (CKDPS) Using Machine Learning Technique'. Prediction and comparison of all algorithms was done and concluded random forest algorithm as the best algorithm to predict kidney disease.

From the study of literature review it is noticed that various classification algorithms were used for implementation to predict the diagnosis of kidney disease. In different levels the diagnosis was carried out to achieve the better results. Based on the literature study the proposed methodology in section 5 will be constructed.

IV. Dataset Description

Dataset collected for undergoing this research is taken from secondary data source https://archive.ics.uci.edu/ml/datasets/chronic_kidneydisease, in which many research have done the research. which has 25 attributes [10]. Among these attributes 11 are numeric and remaining 14 attributes are nominal. Class attribute predicts the present or absence of kidney disease. Dataset used in this study is listed in Table – I.

TABLE - I. Dataset Description

Sr. No	DataSet	Description
1	Age	Numeric
2	Blood_Pressure	Numeric
3	Specific_Gravity	Nominal
4	Albumin	Nominal
5	Sugar	Nominal
6	Red_Blood_Cells	Nominal
7	Pus_Cell	Nominal
8	Pus_Cell_Clumps	Nominal
9	Bacteria	Nominal
10	Blood_Glucose_Random	Numeric
11	Blood_Urea	Numeric
12	Serum_Creatinine	Numeric
13	Sodium	Numeric
14	Potassium	Numeric
15	Hemoglobin	Numeric
16	Packed_Cell_Volume	Numeric
17	White_Blood_Cell_Count	Numeric
18	Red_Blood_Cell_Count	Numeric
19	Hypertension	Nominal
20	Diabetes_Mellitus	Nominal
21	Coronary_Artery_Disease	Nominal
22	Sex	Nominal
23	Pedal_Edema	Nominal
24	Anemia	Nominal
25	Class	Nominal

V. Proposed Methodology

Proposed methodology for predicting kidney disease is

represented in Fig. 1. major objective of this research is to predict the presence or absence kidney disease using classification algorithms. Python 2.7.15 software is used to implement the proposed model. Execution steps for proposed methodology are listed below,

1. Dataset for implementing the research is collected from UCI machine learning repository which is listed in Table - I. Collected dataset are stored in a .csv file.
2. Dataset is of raw data which cannot be directly used for implementation. So raw data is preprocessed.
3. Original dataset has 14 nominal attributes and 11 numeric attributes. Nominal values cannot be used for algorithm directly, so transformation of nominal dataset is done. LabelEncoder() method is used to transform nominal dataset into numerical dataset.
4. In the next step the missing values present in each attribute are calculated. As many attributes has missing values Mean Imputation technique is used to impute missing values.
5. Model Processing phase splits dataset into training and test dataset. 80% dataset are considered as training dataset and 20% are considered as test dataset.
6. Naïve Bayes classifier, K-Nearest Neighbor, Bagging, Decision support system and Random forest classifier are applied to perform classification of kidney disease.
7. Comparison of all classification algorithms are done on calculating accuracy and root mean square error.
8. Comparison results shows that Decision Tree classifiers showed an improved accuracy percentage of 100% with minimized error rate.

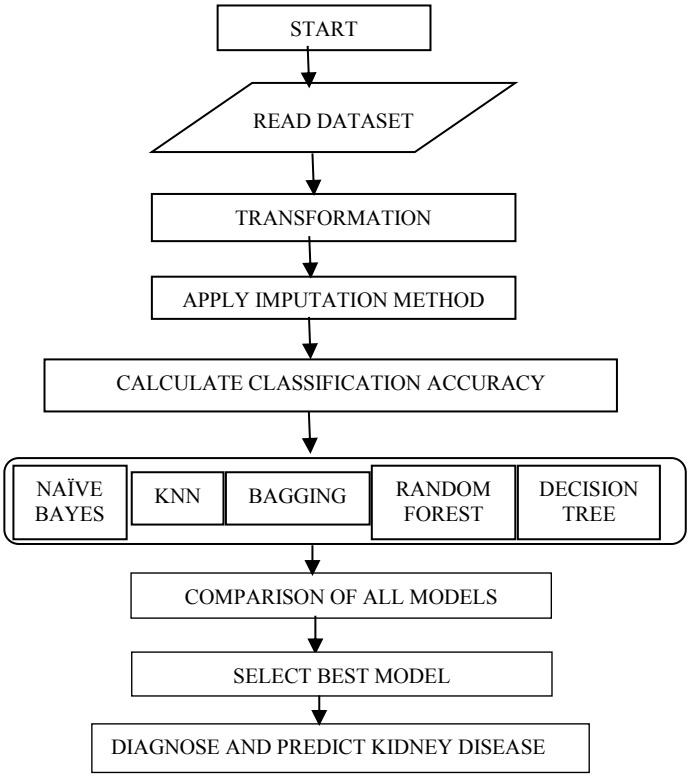

Fig.2 Structure for predicting kidney disease

VI. ALGORITHMIC STEPS

Step – 1: Start

Step – 2: Read the Dataset An

Step – 3: Check dataset attributes

```
If (attributes == 'nominal')
{
        Convert into 'Numerical'
}
```

Step – 4: Calculate missing values

```
If (missing attributes == TRUE)
{
        Apply Mean Imputation technique
}
```

Step – 5: Categorize Target and Response variable

Step – 6: Split training and test dataset as 80% – 20%

Step – 7: Implement classification algorithms for training and test dataset

Step – 8: Calculate classification accuracy, sensitivity, specificity and root mean square error

Step – 9: Select model with higher accuracy and less error used for diagnose and predict kidney disease

Step – 10: Stop

VII. RESULT ANALYSIS

Experimental work for this research is conducted in Python 2.7.15. Dataset collected from UCI machine learning repository are stored in .csv file. Python reads the dataset and checks for datatype of each attributes, later using LabelEncoder() method nominal attributes are converted into numeric attributes.

Using isnull() function it calculates the missing values present In each attributes. As dataset shows large number of missing values, missing values are handled using Mean imputation technique. After applying mean imputation all the attributes are completely fitted with mean values of attributes. Then dataset is divided into training and test dataset for that 80%-20% combination was considered.

Naïve Bayes classifier, K-Nearest Neighbor, Bagging, Decision support system and Random forest classifier are applied for training dataset and test results are obtained and noted. Comparison of test result is done on applied classifiers and results are visualized using bar graph. Error rate for each classifier are noted and compared.

TABLE II. ALGORITHMIC RESULTS

S.No	Algorithm	Accuracy %	RMSE
1	Naive Bayes	97.5	0.31
2	KNN	99.57	0.139
3	Bagging	63.33	1.25
4	Decision Tree	100	0.0
5	Random Forest	99.17	0.12

In Table – II the results obtained from many classifiers are tabulated. Implementation was carried out in Python for getting accurate results. Based on the above tabulated results decision tree shows the performance of 100% accuracy and with 0% root mean square error.

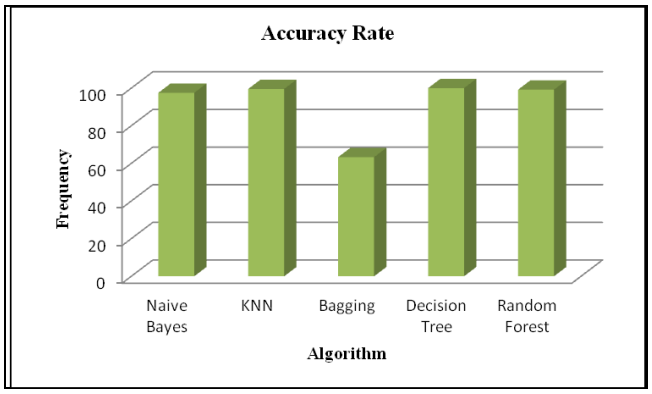

Fig. 3. Comparison results of Accuracy

Comparison results of accuracy obtained by the classification algorithms are visualization in bar graph and shown in Fig. 2. As per the bar graph we can visualize Decision tree algorithm gives the highest accuracy percentage in comparison with other algorithms.

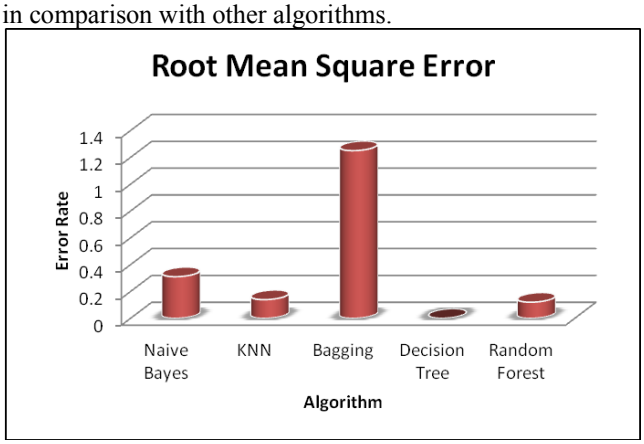

Fig. 4. Comparison results of error rate

Comparison results of Root Mean Square error in Fig. 4 clearly shows the obtained classification algorithm errors which are visualization in bar graph. As per the bar graph Decision tree algorithm gives the least error percentage in comparison with other algorithms.

VIII. CONCLUSION

In this study Naïve Bayes classifier, K-Nearest Neighbor, Bagging, Decision support system and Random forest classifiers are used to predict kidney disease. Implementation work was conducted in Python. Initially, in the pre-processing stage, transformation and missing value evaluation was done and later applied in the classification algorithms. Results were obtained, from the observation which shows, 100% accuracy rate when implemented in Decision tree algorithm and very less root mean square error. Future research should be done using primary dataset and rule based induction to be implemented to calculate the severity of the disease.

REFRENCES

[1] Almasoud, Marwa and Ward, Tomás, "Detection of chronic kidney disease using machine learning algorithms with least number of predictors", International Center for Scientific Research and Studies, 2019

[2] El-Houssainy A Rady, Ayman S Anwar, "Prediction of kidney disease stages using data mining algorithms". Informatics Medicine Unlocked, Elsevier Volume 15, 2019.

[3] Ramya, S, Radha, N, "Diagnosis of chronic kidney disease using machine learning algorithms", International Journal of Innovation Research Computation Communication, Engineering 4, 812–820 (2016)

[4] Sumana De, Baisakhi Chakraborty, "Development of Chronic Kidney Disease Prediction System using Machine Learning Technique", Intelligent Data Communication Technologies and Internet of Things pp 155-164, November 2019.

[5] Tekale, S., Shingavi, P, Wandhekar, S, Chatorikar, A, "Prediction of chronic kidney disease using machine learning algorithm", International Journal of Innovation Research Computation Communication Engineering 7, 92–96 (2018)

[6] Vijayarani S & Dhayanand S, "Data mining classification algorithms for kidney disease prediction", International Journal on Cybernetics and Informatics (IJCI), (2015), pp.13-25

[7] Vinitha, S, Sweetlin, S, Vinusha, H, Sajini, S, "Disease prediction using machine learning over big data", Computer Science Engineering International Journal 8, 1–8 (2018)

[8] Zixian et. al., "Machine Learning Based Prediction System for Chronic Kidney Disease", International Journal of Engineering & Technology, pp 1161-1167, (2018).

[9] https://www.worldkidneyday.org/facts/chronic-kidney-disease/

[10] https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease

[11] https://www.niddk.nih.gov/health-information/communication-programs/nkdep/laboratory-evaluation/glomerular-filtration-rate/estimating

[12] Bashir, U. Qamar, F. H. Khan, "Intelli-Health A medical decision support application using a novel weighted multi-layer classifier ensemble framework," Journal of Biomedical Information, vol. 59, pp. 185–200, 2016.