2020 International Conference for Emerging Technology (INCET)
Belgaum, India. Jun 5-7, 2020

Performance Analysis of Machine Learning Classifier for Predicting Chronic Kidney Disease

Rahul Gupta[1], Nidhi Koli[2], Niharika Mahor[3], N Tejashri[4]
[1,2,3,4] Computer Engineering Department, Delhi Technological University, New Delhi, India
[1]rahulgupta100689@gmail.com, [2]nid1998dec@gmail.com, [3]niharika.mah3@gmail.com, [4]nayakn320@gmail.com

Abstract— **Chronic Kidney Disease (CKD) is a type of chronic disease which means it happens slowly over a period of time and persists for a long time thereafter. It is deadly at its end stage and will only be cured by kidney replacement or regular dialysis which is an artificial filtering mechanism. It is important to identify CKD at the early stage so that necessary treatments can be provided to prevent or cure the disease. The main focus in this paper is on the classification techniques, that is, tree-based decision tree, random forest, and logistic regression has been analyzed. Different measure has been used for comparison between algorithms for the dataset collected from standard UCI repository.**

Keywords— Chronic kidney disease, Prediction, Machine Learning, Decision Tree, Random Forest and Logistic Regression.

I. Introduction

Chronic Kidney Disease (CKD) is a critical health condition worldwide that is a major reason for malicious health outcomes, particularly in countries where income ranges from low-to-middle where millions die regularly due to lack of modest treatment. As per the stages in any chronic disease the fatality is related to the stage it had been without being cured. The high risk factors of CKD are increasing frequency of diabetic patient, hypertension, heart disease, mellitus and family history of kidney failure. If CKD is left undetected and therefore untreated, it can lead to hypertension and in severe cases to kidney failure. WE procured a standard dataset from the UCI machine repository for chronic Kidney Disease. CKD if predicted early and accurately, can benefit patients in many ways. It increases the probability of a successful treatment while also adding years to the person's life. This paper work aims to predict kidney disease by using some of the selected machine learning algorithms and feature selection methods. The objective is to collect the combination of different feature and then have used it as input to the machine learning algorithms. The algorithms have been implemented on the basis of selected features and then we compare their performances.

The paper is organized in sections. Section II discusses research works related to this paper. The Section III gives brief on the method followed in the research as well as describes the concepts used. Section IV presents the results obtained. Section V concludes the research.

II. LITERATURE REVIEW

Use of machine learning algorithms to solve problems in the health sector is not new. Several researchers have tried that. Different techniques and methods have been used for the classification and predict the CKD status of the patient. They have worked on a decision support system to diagnosis and predict the chronic renal failure using the random subspace classification. They used techniques like K-nearest neighbor(KNN) and Naïve Bayes for predicting the presence of CRF[1]. Individual performance and ensemble learners have been used to predict the chronic kidney disease is focused, and they aim for prediction of kidney disease with the use of machine learning algorithm. They use ensemble as well as individual learners with the WEKA tool for the results[2]. In a research paper researchers have worked on a three different prediction algorithms which had been continuous, and categorical the most fitting models with ten predictors and a simplified categorical model with eight predictors.[3] .

Use of five different feature selection techniques to compare the accuracy ,sensitivity and specificity, those are chi-square, gain ratio, information gain, symmetrical uncertainty, relief-f on the two classification algorithm that is multilayer preceptor network(MLPN) and radial-basis function networks(RBFN) with an accuracy of 97% and 98.5% respectively[4]. Researchers have used the random forest algorithm to predict if the candidates have chronic kidney disease or not. He used the original sample to draw and tree bootstrap, then produce the unpruned classification tree choosing best split among all the predictors[5]. P. Yildrim used the multilayer perceptron network and different sampling algorithm to compare predict chronic kidney disease and evaluate the precision, recall, and f-measure for the sampling algorithm[6]. A research was done where they compared the results after applying Support Vector Machine (SVM), Radial Basis Function (RBF), Probabilistic Neural Networks(PNN) and Multilayer perceptron (MLP) algorithms[7]. Researchers did the analysis using data mining classification techniques, that are- Artificial Neural Network(ANN) and Naïve Bayes and compare the performance on the basis of accuracy [8]. In a research they predicted whether a person has kidney diseases using algorithms such SVM and Naïve Bayes. The research was focused on finding which is the best algorithm after they tested their classification accuracy and different execution time performance factors[9] He used the principal component analysis and rule based classifier for the classification of the chronic kidney diseases. The rule-based classifier are – PART, RIDOR and JRIP. Measures used to compare the algorithms are False positive rate, True positive rate, Recall, Precision. [10]

III. METHODOLOGY

In this research the path that will be followed is represented in figure 1. Once we gather the raw data, we preprocess it to reduce any inconsistency like null values. Feature selection such as univariate selection and correlation matrix are applied to find out the features that most

978-1-7281-6221-8/20/$31.00 ©2020 IEEE

appropriately represent the dataset. Using the information, we reduce our attribute size before trying out three classification algorithms. These algorithms are Decision trees, Random Tree and Logistic Regression. Then we use different performance measures to compare the algorithms.

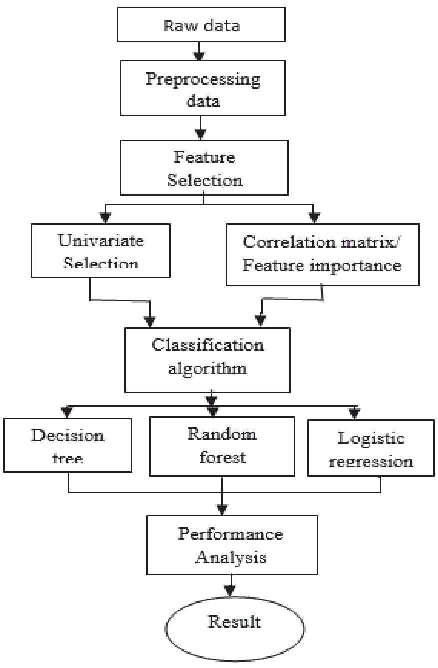

Fig. 1. Proposed method

The Dataset (A) part describe the dataset and its attributes. The Data Mining and preprocessing part (B) describes the preprocessing details. The feature selection (C) part describes the feature selection methods used in our research. The measures of performance evaluation part (D) talks about the parameters on which we will compare the algorithms. The classification algorithms (E) gives details about the algorithms we focused on in our paper.

A. Dataset

We used a dataset from the UCI machine learning repository. We have records of 400 patients in this dataset. The data has 25 attributes as follows:

TABLE I. DATASET

S.NO	ATTRIBUTE	TYPE
1	Age	numerical
2	Blood pressure	numerical
3	Blood glucose random	numerical
4	Blood urea	numerical
5	Serum creatinin	numerical
6	Potassium	numerical
7	Red blood cell count	numerical
8	Packed cell volume	numerical
9	White blood cell count	nominal
10	Haemoglobin	nominal
11	sodium	nominal
12	Sugar	nominal
13	Hypertension	nominal
14	Anemia	nominal
15	Pedal edema	nominal
16	Appetite	nominal
17	Specific Gravity	nominal
18	Bacteria	nominal
19	Coronary artery disease	nominal
20	Albumin	nominal
21	Diabetes mellitus	nominal
22	Red blood cells	Nominal
23	Pus cell clumps	nominal
24	Pus cell	Nominal
25	class	nominal

Numbers of instances in this data set are 400 out of which 62.5% have CKD and 37.5% do not have CKD. There are a total of 25 attributes. The classification need to be done into CKD and NOTCKD. The dataset does contain some missing values that need to be addressed.

B. Data mining and Preprocessing

The dataset we received from the internet source needs to be cleaned as it has null or NA values for various attributes for a given instance. Missing values in the dataset like NA's or blank values are removed by using Pandas library "dropna" and "fillna", which drops either column or rows with missing data and replaces NA's with the mean values of that attribute respectively. The attributes details are as follows:

C. Feature Selection Methods

The performance of every model is heavily impacted with the use of feature selection models. They drastically change the attributes that will be used to train the model as they remove attributes that do not have high impact on the data or negatively impact the data. It is a core concept in machine learning.

Feature selection and Data cleaning are supposed to be the very first thing to do in model designing. In this paper we will apply two feature selection algorithms to obtain important features according to those. Then we would run our classification algorithms on the attributes are deemed important and relevant by both the selection algorithms.

1) Univariate selection: The univariate selection depends on the results of the univariate statistical test. The best features are selected on the test's basis. This selection compares each feature to the target variable, to see whether there is any statistically significant relationship between the feature and the target variable. It is also called analysis of variance (ANOVA). When analysis between one feature and target variable is done, we ignore the other features. That is why it is called 'univariate'. Each feature has its individual test score. Finally, all the test scores are compared, and the features with the top scores will be selected. The formula we have used for the analysis of variance is:

$$F = \frac{\dfrac{\sum \quad n_j(\underline{X_j} - \underline{X})^2}{(k-1)}}{\dfrac{\sum \quad \sum \quad (X - \underline{X_j})^2}{(N-k)}}$$

2) Correlation matrix/ feature importance: On a dataset, the set of correlation values between each pair of its attributes is arranged in the form of a matrix which is called a correlation matrix. Correlation is a statistical term which refers to the estimate of closeness or extent of relation two variables are having with each other, which is a linear relationship. There are many methods of correlation calculation. The most popular method is Pearson Correlation

Coefficient. The formula we have used for computing correlation matrix is:

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

Where,

$$StandardDeviation, \sigma = \sqrt{\frac{\sum \quad (\chi - \mu)^2}{N}}$$

$$Cov(X,Y) = \frac{\sum_{i=0}^{n} \quad (x_i - E(X))(y_i - E(Y))}{N-1}$$

D. Measures of Performance evaluation

To measure the performance of our algorithms we will use tree different evaluation functions. In these functions TP refers to cases that were positive and were predicted positive by the algorithms. TN as negative cases predicted negative. FP are cases predicted positive but were actually negative. FN represents cases that were predicted negative but were actually positive.

1) Accuracy: It is defined as the ratio of correctly predicted observation(true negative + true positive) to the total observations:

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

2) Precision: Precision defines the proportion of predicted positives that are actually positive. It is therefore the ratio of true positives(TP) to the number of cases predicted positive(TP + FP):

$$Precision = \frac{TP}{TP + FP}$$

3) Recall: Recall defines the proportion of actual positives that were predicted positive. It is the ratio of true positives(TP) to the actual positives (TP + FN):

$$Recall = \frac{TP}{TP + FN}$$

E. Classification Algorithms

Classification refers to mapping data of the dataset into predefined groups or classes (predicting the class of unclassified data). Classification is a process related to categorization where we classify a collection of data into different groups. Here we use binary classification that has two classes. So a person would either test positive or negative for CKD. In this research work Decision Tree, Random Forest and Logistic regression are used as classifier to classify whether a person has chronic kidney disease or not.

1) Decision Tree: In decision trees Classification models are built in a tree structure. The dataset is split into smaller subsets and we incrementally establish an associated decision tree. The final outcome of the algorithm is a tree which has decision nodes .The leaf nodes depicts the classification or class. Decision nodes have two or more branches. The uppermost decision node in a tree is called root node. Decision trees can work with both numerical and categorical data.

Algorithm

- With training data D., starts with single node N.
- N becomes leaf if all the data in D belongs to same class. Otherwise attribute 'A' is selection method based on the splitting criterion.
- The instance in 'D' is partitioned accordingly.
- Apply algorithm recursively to each subset in 'D' to all the other subset in 'D' to form decision tree.

2) Random Forest: This algorithm creates a forest of decision trees. It is also a supervised learning classification algorithm. The number of trees in the forest is an indication of the forest's robustness. Also the accuracy of the algorithm is directly proportional to the number of trees in the forest.

Algorithm

- First we randomly select 'p' features from the total 'q' features (where p<<q)
- From the selected' 'p' features, we need to calculate a node, referred as 'd' using the method best split point.
- Using the best split, we need to split the node into daughter nodes.
- Then we repeat the above steps till a number 'l' is reached
- A forest of 'n' number of trees is built by applying the above steps 'n' number of times.

3) Logistic Regression: The core method or the heart of logistic regression is the logistic function, which is also referred to as the sigmoid function .Another type of regression, i.e. linear regression predicts continuous dependent variable. Logistic regression on the other hand predicts categorical dependent variables with the use of a set containing independent variables.

IV. RESULTS

Comparative measures for above algorithms summary:

TABLE II. COMPARISON OF RESULTS FROM DIFFERENT ALGORITHMS.

Algorithm\Measure	Accuracy	Precision	Recall
Decision Tree	98.48	100.0	97.61
Random Forest	94.16	95.12	96.29
Logistic Regression	99.24	98.82	100.0

The final accuracy of the three classification algorithms we used are: {98.48%, 100.0%, 97.61%}

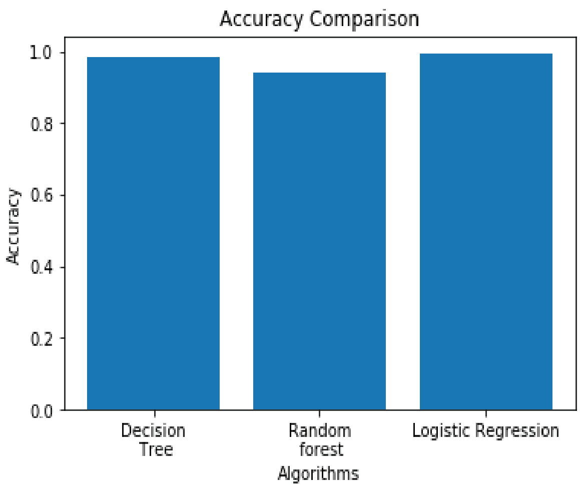

Fig. 2. Accuracy Comparison

The final precision of the three classification algorithms we used are: {100.0 %, 95.12%, 98.82%}

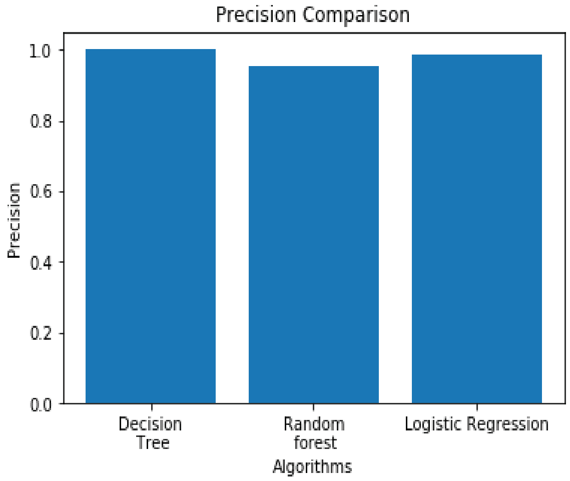

Fig. 3. Precision Comparison

The final recall of the three classification algorithms we used are: {97.61%, 96.29%, 100.0%}

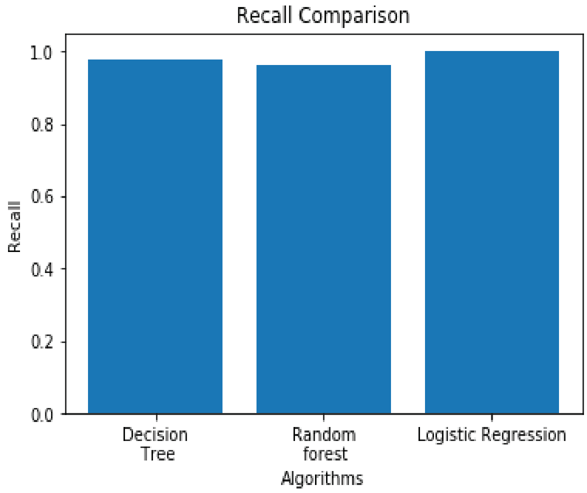

Fig. 4. Recall Comparison

V. Conclusion

We were able to evaluate the performance of different ML algorithms on the Chronic Kidney Disease data set we took from UCI machine learning library[11]. We preprocess the dataset and then used the filter method of feature selection that is univariate selection and correlation matrix along with feature importance to find best features from the dataset. The proposed algorithm that is, Decision tree, Random forest and logistic regression have achieved an accuracy of 98.48, 94.16 and 99.24 respectively. Precision of 100, 95.12 and 98.82 and recall of 97.61, 96.29 and 100. Two feature selecting techniques are combined by leveraging the strength of each the techniques. On comparison we find Logistic Regression with highest accuracy and recall while Decision tree have the highest precision.

References

[1] A. R, G. Sasi, R. Sankar, and O. . Deepa, “Decision Support system for diagnosis and prediction of Chronic Renal Failure using Random Subspace Classification,” 2016, pp. 1287–1292.

[2] D. S. Sisodia and A. Verma, “Prediction Performance of Individual and Ensemble learners for Chronic Kidney Disease,” 2017, pp. 1027–1031.

[3] A. V Kshirsagar et al., “A Simple Algorithm to Predict Incident Kidney Disease,” ARCH Intern Med, vol. 168, no. 22, pp. 2466–2473, 2008.

[4] A. K. Shrivas and S. Kumar Sahu, “Classification of Chronic Kidney Disease using Feature Selection Techniques,” IJCSE, vol. 6, no. 5, pp. 649–653, 2018.

[5] M. Kumar, “Prediction of Chronic Kidney Disease Using Random Forest Machine Learning Algorithm,” Int. J. Comput. Sci. Mob. Comput., vol. 5, no. 2, pp. 24–33, 2016.

[6] P. Yildirim, “Chronic Kidney Disease Prediction on Imbalanced Data by Multilayer Perceptron: Chronic Kidney Disease Prediction,” in Proceedings - International Computer Software and Applications Conference, 2017, vol. 2, pp. 193–198, doi: 10.1109/COMPSAC.2017.84.

[7] E. H. A. Rady and A. S. Anwar, “Prediction of kidney disease stages using data mining algorithms,” Informatics Med. Unlocked, vol. 15, pp. 1–7, Jan. 2019, doi: 10.1016/j.imu.2019.100178.

[8] V. Kunwar, K. Chandel, S. A. sai, and A. Bansal, “Chronic kidney disease analysis using data mining classification techniques,” 2016, pp. 300–305.

[9] V. S and D. S, “Data Mining Classification Algorithms for Kidney Disease Prediction,” Int. J. Cybern. Informatics, vol. 4, no. 4, pp. 13–25, Aug. 2015, doi: 10.5121/ijci.2015.4402.

[10] A. Nway Oo, “Classification of Chronic Kidney Disease (CKD) Using Rule based Classifier and PCA,” Int. J. Adv. Manag. Technol. Eng. Sci., vol. 8, no. 4, pp. 728–733, 2018.

[11] Dua, Dheeru and Graff, and Casey, “UCI Machine Learning Repository,” 2017. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease. [Accessed: 20-Mar-2020].