EasyChair Preprint

№ 3832

Chronic Kidney Disease Prediction by Machine Learning

Abhishek Kumar Pandit, Rohit Prasad Kushwaha and Indra Kumar

July 12, 2020

Chronic Kidney Disease Prediction by Machine Learning

Abhishek Kumar Pandit
Dept. of Computer Science & Engineering
Parul Institute of Engineering. & Technology
Vadodara,India
abhisewa@gmail.com

Rohit Prasad Kushwaha
Dept. of Computer Science & Engineering
Parul Institute of Engineering. & Technology
Vadodara,India
rohitpr9852@gmail.com

Indra Kumar
Dept. of Computer Science & Engineering
Parul Institute of Engineering. & Technology
Vadodara,India
ik512450@gmail.com

Abstract:- The application of machine learning in the field of medical diagnosis is increasing gradually. This can be contributed primarily to the improvement in the classification and recognition systems used in disease diagnosis which is able to provide data that aids medical experts in early detection of fatal diseases and therefore, increase the survival rate of patients significantly. In this paper, we apply different classification algorithms on the dataset available in UCI repository for disease prediction.

Keywords— Machine Learning, Disease Prediction, Chronic Kidney Disease, Decision Tree, Random Forest, KNN, Naive Bayes.

I. Introduction

According to a report by Centers for Disease control and Prevention [1] 38% of US Adults aged 65 years or older are having Chronic Kidney Disease (CKD) in comparison to people aged between 45-64. The healthcare problem of chronic disease is also very important in many other countries. In India the rate of the death by kidney disease is 25.35 which is also a big number in terms of death cases. The early detection of common diseases such as kidney disease and breast cancer could control and reduce the chances of these diseases to be fatal for the patient. With the advancement in machine learning and artificial intelligence, several classifiers and clustering algorithms are being used to achieve this.

Following the methodology used in, this paper presents the use of machine learning algorithms for prediction of chronic kidney disease which is the leading cause of deaths in the US.

The dataset used for building the predictive models in this paper are available and can be downloaded from UCI machine learning library [2]. The data is imported in CSV format and cleaned for use. After data munging and attributes selection, machine learning algorithms including Decision Trees, Random Forest, K-Nearest Neighbours and Naive Bayes, are used for prediction of chronic kidney disease, and a comparison of their accuracy is done for selecting the best model for the disease

dataset. All the analysis and visualization are carried out in python 3.7.6

The paper is presented as follows: Section 2 gives the brief explanation about the machine learning algorithms used. Followed by Section 3 which describes the proposed method for building predictive models. Section 4 explains the experiments and results. Section 5 includes conclusion and future scope of the paper.

II. MACHINE LEARNING ALGORITHMS

A. Decision Tree

Similar to the tree analogy in real life, the Decision tree is a machine learning algorithm, used for both classification and regression analysis [3]. It is a tree-like graph beginning with a single node, and branching into its possible outcomes. Unlike the linear models, a decision tree is a supervised learning, that maps nonlinear relationships as well. The data sample is divided into homogeneous subsets based on the most notable splitter in input attributes. The splitter is identified using various algorithms such as Gini Index, ChiSquare, Information Gain and Reduction in Variance.

B. Random Forest

Random forest is an ensemble of various decision trees, trained with the bagging methodology [4] . Bagging is used for making the model more stable and accurate by approaching averaging model technique. The random forest classifier [5] is basically a collection of decision tree classifiers where each tree is constructed with a number of random vectors and is able to vote for the most favored class for prediction. The injection of randomness in the model prevents it from over fitting and provides better results for classification analysis.

C. K-Nearest Neighbours

In pattern recognition, the K-Nearest Neighbours is a non-parametric method used for classification and regression [6]. In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether K-NN is used for classification or regression:

- In *K-NN classification*, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If $k = 1$, then the object is simply assigned to the class of that single nearest neighbor.
- In *K-NN regression*, the output is the property value for the object. This value is the average of the values of k nearest neighbors.

D. Naive Bayes

In machine learning, naïve Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes theorem with strong (naïve) independence assumptions between the features. They are among the simplest Bayesian network models [7].Naïve Bayes has been studied extensively since the 1960s. It was

introduced (though not under that name) into the text retrieval community in the early 1960s. It also finds application in automatic medical diagnosis.Naïve Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem**.**

III. PROPOSED METHOD

The proposed method for building the predictive model for the disease as follows:

- **Exploration of dataset:** Dataset is explored in the python environment along with a data dictionary of attributes.
- **Data Preprocessing:** Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format. Steps involved in Data Preprocessing:
    1. **Data Cleaning:** The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling missing data, noisy data etc.
    2. **Data Transformation:**This step is taken in order to transform the data in appropriate forms suitable for the mining process.
    3. **Data Reduction:**Since data mining is a technique that is used to handle huge amounts of data. While working with a huge volume of data, analysis became harder in such cases. In order to get rid of this, we use data reduction techniques. It aims to increase the storage efficiency and reduce data storage and analysis costs.
- **Feature Selection:** It is crucial for any predictive modeling and is the process where you automatically or manually select those features which contribute most to your prediction variable or output in which you are interested in.
- **Model Fitting and Testing:** After feature selection, 4 classification algorithms including Decision Trees, Random Forest, K-Nearest Neighbours and Naive Bayes were used with the selected feature and a comparison between their prediction accuracy was done using the Train/Test split method. The test size was set to 0.3, i.e. 70% of the data is used for training and 30% of data is used for testing. Below Fig. 1 shows Steps involved in our method.

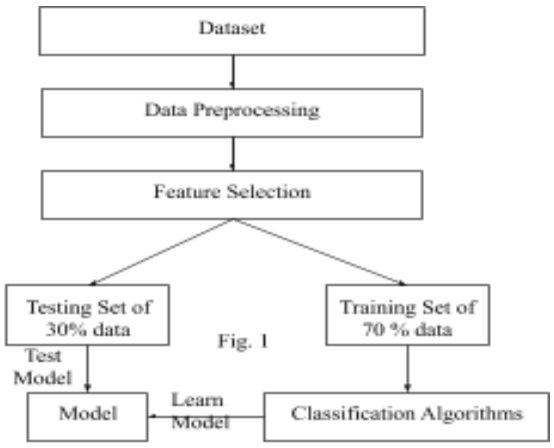

Fig. 1

IV. EXPERIMENTS AND RESULTS

The Chronic kidney Disease(Original) Dataset [2] consists of 25 different attributes with 400 instances. After the Data Preprocessing and Feature Selection, the newly constructed dataset consists of 15 different attributes with 400 instances. The attribute data view of records shown in Table 1.

TABLE 1. ATTRIBUTES IN DATASET

Attributes	Description	Allowed Values
Age	Age in years	Numerical Values
Bp	Blood Pressure	Numerical Values
Sg	Specific Gravity	Nominal Values(1.005, 1.010, 1.015, 1.020, 1.025)
Al	Albumin	Nominal Values(0, 1, 2,3,4,5)
Su	Sugar	Nominal Values(0,1,2,3,4,5)
Rbc	Red Blood Cell	Nominal Values(0,1)
Bu	Blood Urea(mgs/dl)	Numerical Values
Sc	Serum Creatinine(mgs/dl)	Numerical Values
Sod	Sodium(mEq/L)	Numerical Values
Pot	Potassium(mEq/L)	Numerical Values
Hemo	Hemoglobin(gms)	Numerical Values
Wbcc	White Blood Cell Count(cells/cumm)	Numerical Values
Rbcc	Red Blood Cell Count(millions/cmm)	Numerical Values
Htn	Hypertension	Nominal Values(0,1)
Class	Class	Nominal Values(0,1)

During the data preprocessing all the missing values were filled by the mean value of the dataset before modeling. These 16 attributes are selected by performing feature selection on the dataset. Fig. 2 compares the prediction accuracy of the different machine learning algorithms used in this paper for the chronic kidney dataset.

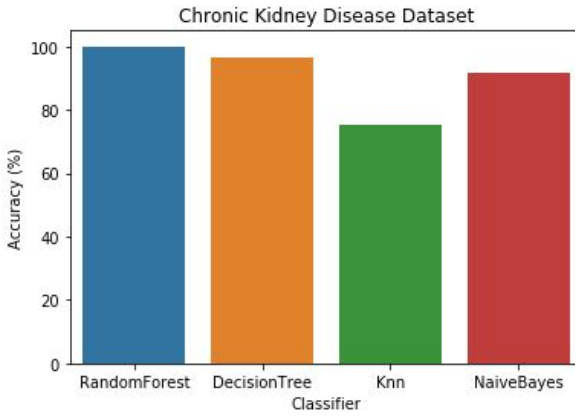

Fig. 2 Comparison of different algorithm

The Classification performances of the classifiers were analysed with respect to standard performance parameters, namely: Accuracy, Recall-Score, Precision-Score and F1-Score. The Formula for calculating these parameters are given below:

$$\text{Precision} = \frac{tp}{tp+fp} * 100$$

$$\text{Recall} = \frac{tp}{tp+fn} * 100$$

$$\text{F1- Score} = \frac{2 * Precision * Recall}{Precision + Recall}$$

Where

tp is the number of true positives,

fp is the number of false positives,

fn is the number of false Negatives.

The Fig. 3 shows the values of Precision, Recall, F1-Score performance metrics besides their training time for all the four classifiers separately for our dataset.

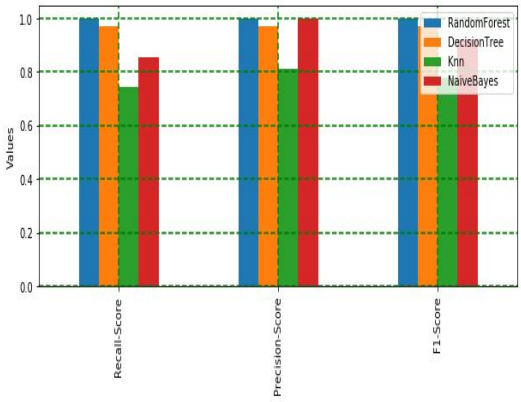

Fig. 3 Performance metrics score for dataset

V. CONCLUSION

The results of this study confirm the application of the machine learning algorithms in prediction and early detection of disease. The prediction accuracy of our proposed method reaches 100% in Chronic Kidney dataset using Random Forest, 96.66% using Decision Tree, 75% using K-Nearest Neighbours, 91.66% using Naive Bayes. The future scope of and improvement of the project involves automation of the steps such as feature selection, data preprocessing. The project can also be used as a training tool for new practitioners after the deployment of the project.

REFERENCES

[1] "Centers for Disease Control and Prevention." [Online]Available:https://www.cdc.gov/kidneydisease/publications-resources/2019-national-facts.html [Accessed: 1-feb-2020].

[2] "UCI Machine Learning Repository" [Online] Available:https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease [Accessed: 24-Sep-2019].

[3] J. R. Quinlan, "Induction of Decision Trees," Mach. Learn., vol. 1, no. 1, pp. 81–106, 1986.

[4]L. Breiman, "Random Forest," pp. 1–33, 2001.

[5]M. Denil, D. Matheson, and N. De Freitas, "Narrowing the Gap: Random Forests In TheDenil, M., Matheson, D., & De Freitas, N. (2014). Narrowing the Gap: Random Forests In Theory and In Practice. Proceedings of The 31st International Conference on Machine Learning, (1998), 665–673. Retrieved from ht," Proc. 31st Int. Conf. Mach. Learn., no. 1998, pp. 665–673, 2014.

[6]Altman, Naomi S. (1992). "An introduction to kernel and nearest-neighbor nonparametric regression" . The American Statistician

[7]McCallum, Andrew. "Graphical Models, Lecture2: Bayesian Network Represention". Retrieved 22 October 2019.