Volume: 4, Issue: 11
Page:101-110
2020

ISSN 2520-4750 (Online) & ISSN 2521-3040 (Print)

International Journal of Science and Business

An Empirical Study of Various Machine Learning Approaches in Prediction of Chronic Kidney Disease

Md. Shafiul Azam, Umme Kulsom, S. M. Hasan Sazzad Iqbal & Md. Toukir Ahmed

Abstract:
In today's era everybody is trying to be conscious about health. Although, due to workload and busy schedule, one gives attention to the health when any major symptoms occur. But Chronic Kidney Disease (CKD) is a disease which doesn't shows symptoms it is hard to predict, detect and prevent such a disease and this can lead to permanently health damage, but some machine learning algorithms can come handy in this aspect for their efficient prediction and analysis. By using data of CKD, patients with 25 attributes and 400 records we are going to use various machine learning techniques like Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree etc. The purposes of our work is to virtuously predicting Chronic Kidney disease and have a comparative analysis among some of the popular machine learning based approaches based on some performance metrics. In our work, it is found that the Random Forest algorithm outperforming other machine learning based approaches we used in the experiment.

IJSB
Accepted 27 October 2020
Published 28 October 2020
DOI: 10.5281/zenodo.4244468

Keywords: CKD, KNN, Machine Learning, Prediction, Performance Metrics.

About Author (s)

Md. Shafiul Azam Assistant Professor, Dept. of Computer Science and Engineering, Pabna University of Science and Technology, Pabna, Bangladesh.
Umme Kulsom, Student, Dept. of Computer Science and Engineering, Pabna University of Science and Technology, Pabna, Bangladesh
S. M. Hasan Sazzad Iqbal, Assistant Professor, Dept. of Computer Science and Engineering, Pabna University of Science and Technology, Pabna, Bangladesh
Md. Toukir Ahmed (Corresponding Author), Lecturer, Dept. of Computer Science and Engineering, Pabna University of Science and Technology, Pabna, Bangladesh.

1. INTRODUCTION

The kidneys are the two organs facing the back of the abdomen. Its function is to clean the blood by removing toxins from the body using the bladder by urinating. When the kidneys are unable to filter out waste products and the body becomes toxic, it causes kidney failure and can lead to death Kidney problems can be classified as serious or chronic kidney disease which is considered serious problems in people 60 years of age and older. The main cause is kidney failure which reduces the rate of gloss filtering. This condition, lasting more than three months, is generally considered to be chronic kidney disease (CKD). The definition and classification of chronic kidney disease (CKD) have changed in the course off time, but current international guidelines define this condition as decreased kidney function estimated by glomerular filtration rate (GFR) of fewer than 60 mL/min per 1·73 m^2, or markers of kidney damage, or both, of at least 3 months time, regardless of the derivative cause (Webster et al., 2007). Diabetes and hypertension are the principal reasons of CKD in all high-income and middle-income countries, and also in many low-income countries. CKD is ranked as the 10th largest cause of word loss (Wang et al., 2018). Chronic kidney disease contains conditions that damage the kidneys and reduce their ability to maintain good health. When kidney disease worsens, waste can build up in our bloodstream and can cause problems such as high blood pressure, anemia, weak bones, poor eating health and nerve damage. Also, kidney disease increases the risk of cardiovascular disease. Chronic kidney disease can be caused by diabetes, high blood pressure, high blood pressure, coronary artery disease, Lupus, Anemia, Bacteria and Albumin in the urine, side effects, Sodium and potassium deficiency in the blood and family history of kidney disease and more. Early detection and treatment can prevent chronic kidney disease. As kidney disease progresses, it can eventually lead to kidney failure, which requires dialysis or kidney transplant to maintain good health (Kumar, 2016).

Kidney disease can affect body's ability to cleanse blood, filter fluid, and help control blood pressure. It can also affect the production of red blood cells and the vitamin D metabolism required for bone health. Usually, people are born with two kidneys. They are on both sides of spine, just above waist. When kidneys are damaged, waste and fluid can build up in the body. This can cause swelling in the ankles, nausea, weakness, poor sleep, and shortness of breath. Without treatment, the damage can be severe and kidneys may eventually stop working. That is very bad, and it can be life threatening. There is also the danger of multiple kidney failure, which can be life-threatening. Awareness is therefore the best way to eliminate the effects of kidney failure. It is a slow-moving disease but is found at the end of stages. At that time, it cannot be cured but can be found at the end of stages. At that time there is no treatment but it can reduce the risk. This will reduce human suffering. The kidneys are the two organs facing the back of the abdomen. Treatment may be possible but may reduce the risk. This problem leads me to use machine learning methods to predict whether people will be affected by kidney failure or not. Most of the medical data is now available in the public database. There are great clinical details. If there is a chance of being identified before the final stage, people can take the necessary steps easily and quickly to reduce the risk. This will reduce human suffering. This severity of the disease was our encouragement to do the research on this topic.

The aim of this thesis is to predict if a person will have a risk of Chronic Kidney Disease. To predict whether the patient will have the symptom at the end-stage affect and that is a challenging task. Thus, there is a need for a mechanism that can accurately predict Kidney Disease. Kidney disease is a slow-motion process in the body. It does not mention its symptoms in the body in the primary stages but in the last stage. From there it is not possible to recover or not remove. It is only possible to reduce risk. It is a Machine learning problem, chosen to be solved using Supervised Learning techniques. The problem of prediction of whether or not a person will have a Kidney belongs to a binary classification domain where the result is how risk the disease carried people.

Data set are collected from UCI Repository. Our algorithm predicts the best accuracy for the given data based on some parameters.

2. REVIEW OF THE LITERATURE

Chronic kidney disease (CKD) is an autoimmune disease. Chronic kidney disease, also called kidney failure, describes a slow loosening of the kidneys. Accurate prognosis for CKD progression over time is needed to reduce costs and mortality rates. There are many researchers working on CKD predictions with the help of various classification algorithms. And those researchers found the expected result of their model. Sinha and Sinha (2015) worked to predict the diagnosis of chronic kidney disease in humans through three classification methods. The classifiers used to support the vector machine and the KNN classifier to predict the disease and the performance of the separator are tested according to accuracy, precision and F-measure. From the analysis they found that of the two divisions SVM and KNN, the KNN division was better than the other. In other paper, authors have learned various machine learning algorithms (Tekale et al., 2018). They analyzed 14 different attributes related to CKD patients and the accuracy of guessing of different machine learning algorithms such as Tree Verification and Vector vector. From the result analysis, it is evident that the decision tree algorithms provide 91.75% accuracy and SVM provides 96.75% accuracy. Vijayarani and Dhayanand (2015) have developed a classification procedure used to classify four types of kidney disease. Comparison of Support Vector Machine (SVM) and Naïve Bayes algorithms for algorithms is made based on the functional characteristics of precision separation and performance time. From the results, it can be concluded that SVM achieves increased editorial performance, producing more accurate results, which is why it is regarded as an excellent separator compared to the Naïve Bayes classifier algorithm. The Naïve Bayes divider separates the data at the highest performance time.

In their study, Tabassum et al. (2017) showed that the classification strategies used to predict the disease whether a patient is affected from CKD or not. Integration techniques are used to combine the same type of person affected under one group. This approach allows physicians to suggest that medications promote and reduce costs. An important goal is to reduce the cost of better treatment. This project consists of one merger algorithm and two partition algorithm. Lakshmi et al. (2014) proposed a three-pronged data mining strategy to predict the survival of dialysis dials. In this study, a variety of data mining techniques (Artificial Neural Networks, Decision tree and Logical Regression) are used to extract information on the interaction between these variables and patient survival. Comparison of the effectiveness of the three data mining techniques is used to extract information. The concepts presented in this study have been relevant and evaluated using data collected from various dialysis environments output information. The concepts presented in this study have been relevant and evaluated using data collected from various dialysis environments. Results are reported. Finally, ANN suggested that kidney dialysis get better results with accuracy and performance. Vijayarani et al. (2015) launched a research project to predict kidney disease using the Support Vector Machine (SVM) and the Artificial Neural Network (ANN). The purpose of this work is to compare the performance of these two algorithms on the basis of their accuracy and performance time. From the test results it is evident that the performance of ANN is better than any other algorithm. Ameta and Jain (2017) suggested diagnostic solutions for the disease by analyzing the data using different classification techniques. This helps to provide quick and appropriate treatment to patients. The biggest challenge is finding the best isolation algorithm process based on the accuracy of type and time to perform fully functional objects. Algorithm with better accuracy and limited time to make it selected as the main algorithm. At this stage, a different amount of precision charging is proved by all the dividing factors. ANN has high class accuracy and the ANN algorithm is considered to be a collection of the best rules for classification strategies.

3. RESEARCH METHODS

3.1 Dataset

The proposed system uses the dataset taken from the UCI Machine Learning Repository named Chronic Kidney Disease has 25 attributes, 11 numeric and 14 nominal. Total 400 instances of the dataset is used for the training to prediction algorithms, out of which 250 has label chronic kidney disease (CKD) and 150 has label non chronic kidney disease (NOTCKD). Some of the attributes in the dataset are age, blood pressure, Red blood cell, class and so on. The dataset is divided into two parts, one for training and another for testing. The ratio of training and testing data is 70% and 30% respectively.

3.1.1 Data acquisition and preprocessing

To satisfy the goal of predicting Kidney Disease using machine learning techniques, it was important to find a data source which had data lying over a wide data range. To load data into the machine learning framework for processing, it was necessary to integrate and transform the data. Data cleaning is a process of converting messy data into tidy data. About 60% of time is spent in cleaning the data. It might seem to come earlier in the process but it is actually an iterative approach. Even after applying model, there might be a need to cleanse the data again. In that case, data set need to be restricted again in the middle of machine learning process. Data cleansing is one of the crucial steps in the process. But in weka tools it is done by internal process. In normal process since missing values can tangibly reduce prediction accuracy, this is a higher priority issue. In terms of machine learning, assumed or approximated values are "more right" for an algorithm than just missing ones. Even if anyone don't know the exact value, methods exist to better "assume" which value is missing or bypass the issue.

3.1.2 Formatting:

The majority of real- world classification problems require supervised learning where the underlying class probabilities and class- conditional probabilities are unknown, and each instance is associated with a class label. In real-world situations, we often have little knowledge about relevant features. A relevant feature is neither irrelevant nor redundant to the target concept; an irrelevant feature is not directly associate with the target concept but affect the learning process, and a redundant feature does not add anything new to the target concept. Data contains many features, but all the features may not be relevant so the feature selection is used so as to eliminate the irrelevant features from the data without much loss of the information. Feature selection is also known as attributes selection or variable selection. In supervised learning, the algorithm works with a set of examples whose labels are known. The labels can be nominal values in the case of the classification task, or numerical values in the case of the regression task.

3.1.3 Sampling:

There may be far more selected data available to work with. More data can result in much longer running times for algorithms and larger computational and memory requirements. We can take a smaller representative sample of the selected data that may be much faster for exploring and prototyping solutions before considering the whole data set.

3.2 Predictive Model Development

The next step that follows in the workflow is choosing a model among the many that researchers and data scientists have created over the years. In the project, naïve Bayes, Decision Tree, Logistic Regression, SVM, KNN, Random forest machine learning algorithms are used.

3.2.1 Model Training:

The specific machine learning task will inform the selection of an appropriate algorithm. Then the data is feed to the model during this phase and we will get a learner. A learner is a ML algorithm that has been trained on some data and adjusted to fit the data as best as possible. Typical step in Machine learning is futurization. Feature vectors are created from training data. These vectors are then used for training the model. Model is generated which is then evaluated in the next phase to get the best model. The 'class' label is very important to be present in the training model for learning purposes. The 'class' or 'label' field tells if a patient has Kidney Disease or not.

3.2.2 Parameter Tuning:

Once the evaluation is over, any further improvement in training can be possible by tuning the parameters. There were a few parameters that implicitly assumed when the training was done. Another parameter included is the learning rate that defines how far the line is shifted during each step, based on the information from the previous training step. These values all play a role in the accuracy of the training model, and how long the training will take. For models that are more complex, initial conditions play a significant role in the determination of the outcome of training. Differences can be seen depending on whether a model starts off training with values initiated to zeroes versus some distribution of values, which then leads to the question of which distribution is to be used. Since there are many considerations at this phrase of training, it's important that we define what makes a model good. These parameters are referred to as hyper parameters. The adjustment or tuning of these parameters depends on the data set, model, and the training process. Once we are done with these parameters and are satisfied we can move on to the last step.

3.2.3 Machine Learning Algorithms:

Here we discuss about Naïve Bayes, Support Vector machine, Logistic Regression, Random Forest, KNN and Decision tree Algorithms and how they work in brief.

Naïve Bayes

It is a classification technique based on Bayes Theorem with an assumption of independence among predictors. In sample terms, a naïve Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability and that is why it is known as 'Naïve'. Naïve Bayes model is easy to build and particularly useful for very large data sets. Among with simplicity, Naïve Bayes is known to outperform even highly sophisticated classification methods (Russell and Norvig, 2002).

Decision Tree

A predictive machine learning model which decides the target value of a new sample based on different attribute values of the available data is decision tree. The different attributes denote by the internal nodes of a decision tree, the branches between the nodes tell us the possible values that these attributes can have in the experimental samples, while the terminal nodes tell us the final value of the dependent variable (Russell and Norvig, 2002).

Support Vector machine (SVM)

The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space (N- the number of features) that distinctly classifies the data points. To separate the two of data points, there are many possible hyperplanes that could be chosen. Our objectives is to find a plane that has the maximum margin, i.e the maximum distance between data points of both classes. Maximizing the margin distance provides some

reinforcement so that future data points can be classified with more confidence (Russell and Norvig, 2002).

Logistic Regression

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression). Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail which is represented by an indicator variable, where the two values are labeled "0" and "1". In the logistic model, the log-odds (the logarithm of the odds) for the value labeled "1" is a linear combination of one or more independent variables ("predictors"); the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value). The corresponding probability of the value labeled "1" can vary between 0 (certainly the value "0") and 1 (certainly the value "1"), hence the labeling; the function that converts log-odds to probability is the logistic function, hence the name (Russell and Norvig, 2002).

Random Forest Classifier

Random forest is a supervised learning algorithm which is used for both classification as well as regression. But however, it is mainly used for classification problems. As we know that a forest is made up of trees and more trees means more robust forest. Similarly, random forest algorithm creates decision trees on data samples and then gets the prediction form each of them and finally selects the best solution by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result (Russell and Norvig, 2002).

K- Nearest Neighbor

In pattern recognition, the K-Nearest Neighbor algorithm (K-NN) is a non-parametric method used for classification and regression (Parul Sinha, 2015). In both cases, the input consists of the K closest training examples in the feature space. K-NN is a type of instance-based learning. In KNN Classification, the output is a class membership. Classification is done by a majority vote of neighbors. If K = 1, then the class is single nearest neighbor. In a common weighting scheme, individual neighbor is assigned to a weight of 1/d if d is the distance to the neighbor. The shortest distance between any two neighbors is always a straight line and the distance is known as Euclidean distance. The limitation of the K-NN algorithm is it's sensitive to the local configuration of the data (Russell and Norvig, 2002).

3.3 Performance Metrics

To assess the result of the study accurately, rather than accuracy alone, some of the other performance metrices were introduced in the result sections too. By observing these metrices, a clear indication of better result was noticed among different folding and splits of the dataset. Performance parameters are the most important factor to compare among classifier methods to get the best classifier. Applied performance metrices includes Accuracy, precision, Recall and F1-Score. These parameters calculated from a confusion matrix which situated in every step of classification (Russell and Norvig, 2002). About confusion matrix and detailed information about these proposed parameters are as follows:

Table 1 Confusion matrix

	Predicted YES	Predicted NO
Actual YES	TP	FN
Actual NO	FP	TN

TP represents the number of correctly classified positive instances.

FP represents the number of misclassified positive instances.

FN represents the number of misclassified negative instances.
TN represents the number of correctly classified negative instances.

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

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

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

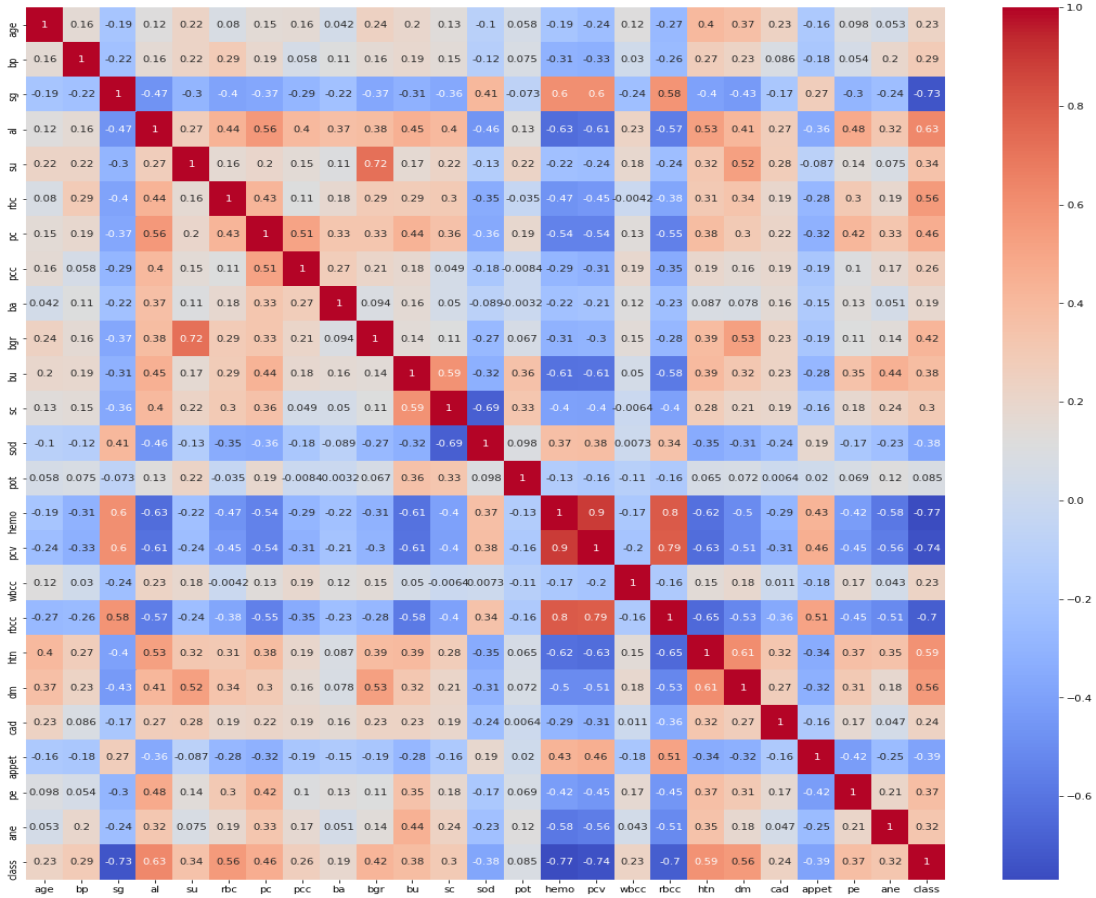
F1 Score = $\frac{2*precision*recall}{precision+recall}$

All the equations are described in detail by Russell and Norvig (2002).

3.4 Experimental Tools

We used python for implementation of different classification algorithm. We used Anaconda for specific computing. Jupyter notebook was used to write and iterate python code for data analysis. Cleaning data set is implemented with python. In jupyter platform the data set is implemented with various classifiers. Like Decision tree, SVM, KNN, Naïve Bayes, Random Forest etc. After loading the dataset, preprocessing steps were done on it as mentioned earlier. Then correlation among features was investigated. Next, we split dataset into training set (70%) and test set (30%). Lastly, we run experiment and analyzed the results.

Figure. 1 Correlation heatmap of the dataset



4. EMPIRICAL RESULTS & DISCUSSION

After implementation of different Machine Learning models, the next step is to find out how the models performed. This is done by running the models on the test dataset which was set aside earlier. The test dataset comprised of 30% of the original data for prediction. Here we will compare the performance of different algorithms and find out the best one for prediction. In this model for predicting CKD, the set of six classification algorithms were used- Logistic Regression (LR), Support Vector Machines (SVM), Naïve Bayes (NB), Decision Tree (DT) and Random Forest (RF) , K nearest neighbor (KNN) were applied on the dataset.

Table 2 Comparison among different classifiers based on different performance metrics. Bold indices best result

Classifiers	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Naive Bayes	96.67	100	94.74	97.3
Random Forest	**100**	**100**	**100**	**100**
Decision Tree	99.17	100	98.68	99.34
SVM	98.33	98.68	98.68	98.68
KNN	74.17	90.91	65.79	76.33
Logistic Regression	91.67	93.42	93.42	93.42

Source: Software output

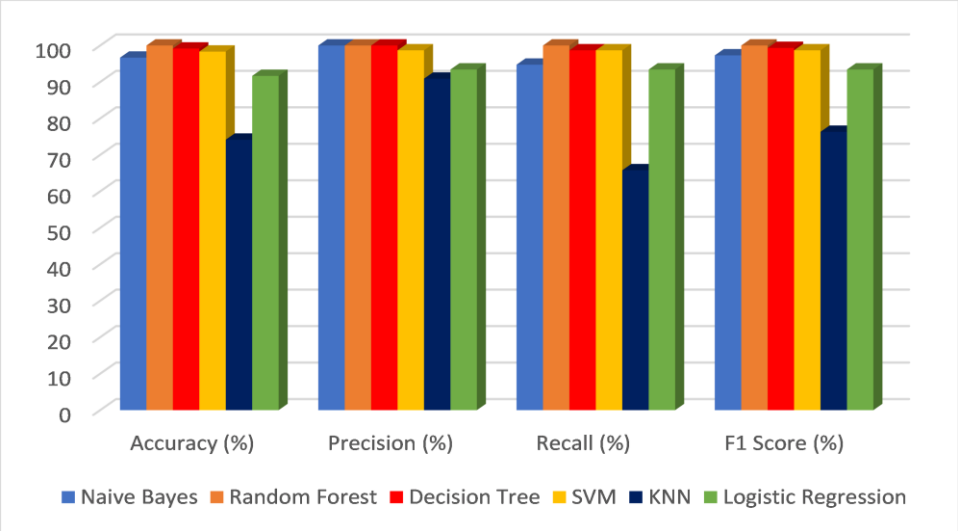


Figure. 2 Bar chart showing the comparison among different classifiers on different performance metrics

From **Table 2** and bar chart in **Figure 2**, we can see accuracy, precision, Recall and F1 score of different classifier algorithms. It is clearly evident that, accuracy of Random forest is 100% and Precision, recall, F1 score are 100, 100, 1.00 respectively. Then the accuracy of Decision tree is 99.17% and Precision, recall and F1 score are 100, 98.68, 0.99 respectively. Support vector machine has 98.33% accuracy. Precision, recall and F1 score of SVM is 98.68, 98.68,0.99 respectively. Naïve Bayes has 96.67% accuracy. Precision, recall and F1 score is 100, 94.74, 0.97 respectively. Logistic regression has 91.67% accuracy. Precision, recall and F1 score is 93.42, 93.42, 0.93 respectively. And K nearest neighbor has 74.17 accuracy and the precision, recall, F1 score has 90.91, 65.79, 0.76 respectively. By comparing all methods, we can say that Random forest classifier gives us the best performance.

5. CONCLUSION & FUTURE DIRECTIONS

The prediction of chronic kidney disease is very important and now-a-days it is the leading cause of death. World's health is badly affected by the chronic disease which is spreading and increasing day by day. The lack or delay in proper treatment can also lead to the death of patients. So, chronic kidney disease is a vital task in medical field. This article presents an approach on various feature selection and classification techniques which can be very helpful for severity analysis for quick chronic kidney disease prediction. This study shows that there is a need to make healthcare professionals aware of reliable feature selection and classification techniques that can be successfully applied on medical databases for the early detection of kidney. Kidney disease prediction system is developed using Naïve bayes, Support vector machine, Decision tree, Random forest, k nearest neighbor and logistic regression classification techniques. Random Forest acts as more effective algorithm than other algorithms implemented in this work to predict Chronic Kidney Disease. However, from the result it is assumed that there may exist overfitting as we saw near perfect result. So, in future, it will be a daunting task to remove overfitting and gather a larger test set and also to run a Principal Component Analysis on the dataset to have a better acclaimed result.

References

Ameta, A., & Jain, K. (2017). Classification of HRS using SVM. *Global Journal of Computer Science and Technology*.

Bala, S., & Kumar, K. (2014). A literature review on kidney disease prediction using data mining classification technique. *International Journal of Computer Science and Mobile Computing*, *3*(7), 960-967.Dalui1, D. (2019). Assessment of Chronic Kidney Disease using clustering techniques. *International Journal of Computer Sciences and Engineering, 7*(18).

Jamgade, A. C., & Zade, S. D. (2019). Disease prediction using machine learning. *International Research Journal of Engineering and Technology*, *6*(5), 6937-6938.Ameta, M. A. (2017). Data Mining Techniques for the Prediction of Kidney Diseases and. *International Journal Of Engineering And Computer Science, 6*(2), 2319-7242.

Haykin, S. (2007). *Neural networks: a comprehensive foundation*. Prentice-Hall, Inc...

Hornik, K., Stinchcombe, M., & White, H. (1990). Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. *Neural networks*, *3*(5), 551-560.

Kumar, M. (2016). Prediction of chronic kidney disease using random forest machine learning algorithm. *International Journal of Computer Science and Mobile Computing*, *5*(2), 24-33.

Lakshmi, K. R., Nagesh, Y., & Krishna, M. V. (2014). Performance comparison of three data mining techniques for predicting kidney dialysis survivability. *International Journal of Advances in Engineering & Technology*, *7*(1), 242.Kumar, M. (2016). Prediction of Chronic Kidney Disease. *International Journal of Computer Science and Mobile Computing, 5*(2), 24-33.

Patil, P. M. (2016). Review on Prediction of Chronic Kidney Disease Using Data Mining Techniques. *International Journal of Computer Science and Mobile Computing*, *5*(5), 135-141.

Russell, S., & Norvig, P. (2002). Artificial intelligence: a modern approach.

Sinha, P., & Sinha, P. (2015). Comparative study of chronic kidney disease prediction using KNN and SVM. *International Journal of Engineering Research and Technology*, *4*(12), 608-12.

Tabassum, S., MBB, G., & Majumdar, J. (2017). Analysis and prediction of chronic kidney disease using data mining techniques. *Int. J. Eng. Res. Comput. Sci. Eng*, *4*(9), 25-32.

Tekale, S., Shingavi, P., Wandhekar, S., & Chatorikar, A. (2018). Prediction of chronic kidney disease using machine learning algorithm. *Int. J. Adv. Res. Comput. Commun. Eng*, *7*(10), 92-96.

Vijayarani, D. S., & Prasannalakshmi, M. R. (2015). Comparative analysis of association rule generation algorithms in data streams. *International Journal on Cybernetics & Informatics (IJCI)*, *4*(1), 15-25.

Vijayarani, S., Dhayanand, S., & Phil, M. (2015). Kidney disease prediction using SVM and ANN algorithms. *International Journal of Computing and Business Research (IJCBR)*, *6*(2), 1-12.

Wang, Z., Chung, J. W., Jiang, X., Cui, Y., Wang, M., & Zheng, A. (2018). Machine Learning-Based Prediction System For Chronic Kidney Disease Using Associative Classification Technique. *International Journal of Engineering & Technology*, *7*(4.36), 1161-1167.

Webster, A. C., Nagler, E. V., Morton, R. L., & Masson, P. (2017). Chronic kidney disease. *The lancet*, *389*(10075), 1238-1252.

Cite this article:

Md. Shafiul Azam, Umme Kulsom, S. M. Hasan Sazzad Iqbal & Md. Toukir Ahmed (2020). An Empirical Study of Various Machine Learning Approaches in Prediction of Chronic Kidney Disease, *International Journal of Science and Business, 4*(11), 101-110. doi: https://doi.org/10.5281/zenodo.4244468

Retrieved from http://ijsab.com/wp-content/uploads/615.pdf

Published by

IJSAB International