Research Inventy: International Journal of Engineering And Science
Issn (e): 2278-4721, Issn (p):2319-6483, PP: 24-28
www.researchinventy.com

Comparing Machine Learning Methods toPredict Chronic Kidney Disease

[1]Subhankar Guha, [2]Sudipta Priyadarshinee

[1]Department of Computer Science and Engineering, Techno Internal Batanagar, India,

[2]Department of Computer Science, Research Scholar, G.M. University, Sambalpur, Odisha, India

Abstract *- Among the most crucial concerns in the healthcare field is the forecasting of chronic kidney disease (CKD). Prediction in the medical area is one of the most fascinating and difficult tasks in daily life. Machine learning processes vast amounts of raw data supplied by the healthcare field into actionable knowledge. Several studies have found that certain characteristics are crucial in improving the accuracy of machine learning methods. This paper employs nine classification models: Naïve Bayes (NB), Bayesian Network, Logistic Regression, K Star, Support Vector Machine (SVM), One R, Projective Adaptive Resonance Theory (PART), Best-First Decision Tree (BF Tree), Optimized Forest to be able to predict (CKD) chronic kidney disease using clinical information. The effectiveness of each classifier is compared to see which is better at predicting chronic kidney disease. for a particular dataset. All experiments are carried out in a simulation environmentusing the WEKA tool.*

Index Terms –CKD; Machine Learning; Prediction

I. Introduction

Chronic kidney disease (CKD) is a condition when the kidneys don't work properly and are no longer cleaning your blood properly. The kidneys' primary role is to filter excess water from the body. and garbage from the blood in order to generate urine; however, if someone has CKD, wastes accumulate in the body. Because of the cumulative damage over time, this condition is considered chronic. It's a widespread ailment that affects people all over the world [1]. Some health problems may occur as a result of CKD. Diabetes, high blood pressure, and heart disease are some of the causes of CKD. CKD is affected by age and gender in addition to these serious illnesses [2]. According to studies, hospitalisation cases of CKD are increasing at a rate of 6.23 percent per year,while the world's death rate is unchanged. [3].

Data mining is appropriate for mining in data when there is a large dataset, however, we can do it with machine learning when the dataset is small. Machine learning has the potential to be useful in a variety of situations such as pattern detection and data analysis [4]. Because there are numerous health datasets available, classifier of machine learning is most suitable to increasing diagnosis prediction accuracy [5]. Algorithms for machine learning are getting more widespread in the field of healthcare as the number of electronic datasets grows fast [6].

II. Related Work

In the diagnosis of CKD, Qin et al. [7] presented data assertion and sample diagnosis. KNN is used to verify data. For diagnosis accuracy, six different classifier methods were used: logistic regression, support vector machine, naive Bayes classifier, random forest feed-forward neural network and KNN. Random forest outperforms these models by 99.75 percent.

Vasquez-Morales et al. [8] used a 40000-instance dataset to create a neural network model for chronic kidney disease forecasting, with a 95% model accuracy.

Because CKD is invasive and expensive, many people reach the end stages without receiving treatment. As a result, early detection of this disease is critical. Amirgaliyev [9] also provided an experimental outcome showed that SVM algorithm achieved 93 percent accuracy.

De Almeida et al. [10] employed Random Forests, Decision Trees, and Support Vector Machines (SVMs) using linear, sigmoid, polynomial, as well as Radial Basis function (RBF) in their study. The author applied the MIMIC-II database to conduct their research. They came to the conclusion that the decision tree and random forest produced the greatest results, with high predictability of 80 percent and 87 percent and so forth.

Almasoud and Ward [11] used a 400-instance CKD dataset with 25 characteristics. In the CKD dataset, they used the filter feature selection approach on features and discovered that albumin, haemoglobin, and specific gravity are feature attributes. They used the dataset to train and verified it using 10-fold cross-validation after feature selection. The gradient boosting technique had the best accuracy of 99.1 percent.

Using clinical data, Sathiya Priya S and Suresh Kumar M [12] used machine learning approaches to predict chronic kidney disease. They employed two algorithms of machine learning: Naive Bayesian (NB) and Decision Tree (DT) method. In comparison to the naive Bayes approach, the Decision tree classifier was determined to be 99.25 percent.

Sujata Drall, Gurdeep Singh Drall, , Bharat Drall, Sugandha Singh and colleagues [13] researched on a dataset about CKD. provided by UCI, which contained 25 characteristics and 400 instances. First, the data was pre- processed, then the missing data was located, updated with 0 and supplied to the dataset. After pre-processing, the authors used an algorithm to find the five the most crucial features, followed by the classification algorithms: Nave Bayes and K-Nearest Neighbour. Since then, KNN was able to obtain the maximum level of accuracy.

III. Proposed Framework

In this research we apply nine classification models: Naïve Bayes (NB), Bayesian Network, Logistic Regression, K Star, Support Vector Machine (SVM), One R, Projective Adaptive Resonance Theory (PART), Best-First Decision Tree (BF Tree), Optimized Forest for the purpose of forecasting chronic kidney disease. Then the classifiers of machine learning are employed on the dataset and estimated the rate of accuracy. Each experiment is subjected to 10-fold cross validation to ensure that the results are free of bias. The primary goal was to find themethod that could best classify the given dataset.

A. Dataset

The suggested approach utilises the dataset on chronic kidney disease from Kaggle, which has 25 features, 11 numeric and 14 nominals. There are 400 instances from the dataset are used to train algorithms for prediction, with 250 labelled chronic kidney disease (CKD) and 150 labelled non chronic kidney disease (NOTCKD). The features in the dataset are age, bp, sg, al, su, rbc, pc, pcc, ba, bgr, bu, sc, sod, pot, hemo, pcv, wc, rc, htn, dm, cad, appet, pe, ane, classification.

B. 10-folds Cross Validation

Cross validation is a method for estimating the effectiveness of a machine learning classifier. It assists researchers in estimating the accuracy of model predictions in practise. There are two kinds of phases in the datasets:testing sets and training sets. Cross validation will be used to compare testing and training sets in order to rule out overfitting and identify how machine learning techniques should produce independent data.

C. Tools and Technique

Weka is useful tool which was utilised to carry out all of the experiments on the classifiers described in this paper. The Weka tool is a gathering of machine learning methods for data mining. It is employed to categorise datasets in an automated fashion using the specified algorithm, for so long as that algorithm is available in the environment.

D. Classification Algorithms

- **Naive Bayes**: The naive bayes algorithm is a well-known classification approach for its simplicity as well as its effectiveness. Bayes law is the only foundation of naive bayes.
  - P(Y/X) = (P(Y/X) *P(X))/P (Y)

The presence of one feature in Naive Bayes has no effect on the presence of other features to put it another way, this theorem presupposes predictor independence [14].

- **Bayesian Network**: Bayesnet is a probabilistic model because it builds models using probability distributions and makes decisions using probability laws. This network is made up of nodes and links in a directed acyclic graph. where each node represents a continuous or discrete variable and each link represents a direct dependency between variables [15].
- **Support Vector Machines (SVM)**: Models of Support Vector Machines (SVM) are finite-dimensional vector spaces in which each dimension represents a "feature" of a certain object. It has been proved to be an excellent method for coping with high-dimensional space issues. This technique is commonly used in document categorization and sentiment analysis because of its computational efficiency on large datasets[16].
- **Logistic Regression:** Logistic regression is a prominent algorithm of machine Learning which belongs to the Supervised Learning method. Logistic regression forecasts the result of a categorical dependent variable. Since then, the outcome must've been discrete or categorical. Yes or No, 0 or 1, true or False, respectively [17].

- **Optimized Forest:** The Forest Optimization (FOA) Algorithm is another method for solving nonlinear problems with optimization, which is inspired by natural processes in forests. Making use of a genetic algorithm to optimise the number of trees in a decision forest in order to find a sub forest with good rate of ensemble accuracy. [18].
- **PART:** Projective Adaptive Resonance Theory is abbreviated as PART. PART is a categorization algorithm using rules. It's a mixture of the C4.5 and RIPPER algorithms. The PART method works well with high-dimensional data. The presence of a hidden layer of neurons in the PART network is a critical element, since it calculates the variances between the output and input neurons and works to reduce the similarity discrepancies [19].
- **BF-Tree:** Best-first decision trees are built in a divide-and-conquer approach. The following is the main notion behind constructing a best-first tree. First, choose a characteristic for the root node and create various branches for it depending on specified criteria [20].
- **K Star:** The test instance's class is defined by the class of associated training instances in K*, which is an instance-based classifier. It distinguishes from many other instance-based learners in that it utilises anentropy-based distance function [21].
- **One R:** OneR is an acronym meaning "One Rule.", is a straightforward but precise classification technique that creates one rule for each data predictor before selecting the rule with the lowest overall error as its "one rule." To develop a rule for each predictor, we establish a frequency table in relation tothe target [22].

IV. Results and Discussion

The experiment is run on the given dataset using a various machine learning approach. In this segment, we evaluate the efficacy of all algorithms in terms of correctly classified instances, incorrectly classified instances, and accuracy. Table 1 displays the outcomes.

Table 1 Classification Accuracy of classifiers

Classifiers	**Correctly classified instances**	**Incorrectly classified instances**	**Accuracy (%)**
Naive Bayes	395	5	98.75
Bayesian Network	399	1	99.75
SVM	398	2	99.5
Logistic Regression	393	7	98.25
Optimized Forest	400	0	100
PART	398	2	99.5
K Star	393	7	98.25
BF-Tree	399	1	99.75
One R	399	1	99.75

From Table 1, it is found that all classifier performed well and Optimized Forest achieves the highest accuracy of100%.

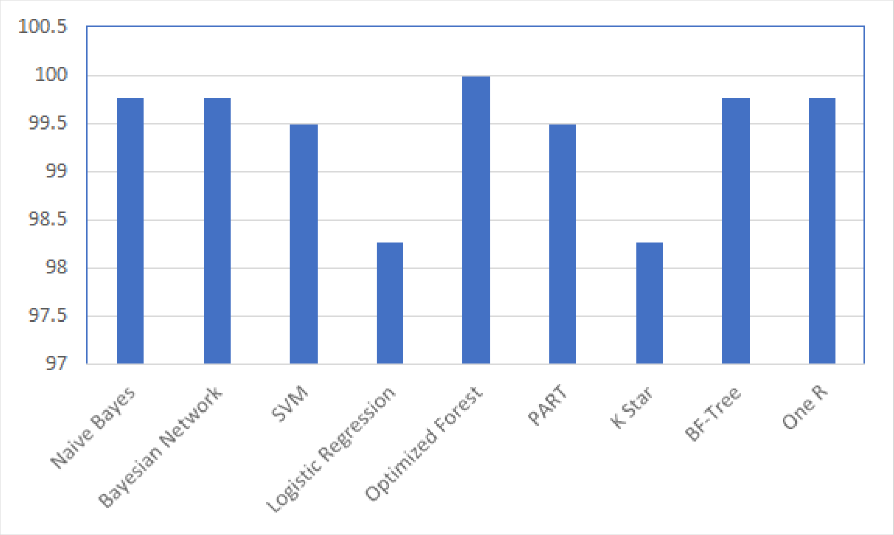

Fig. 1. Comparison of classifier's accuracy

V. Conclusion

(CKD) chronic kidney disease is amongst the most common sickness, and it is vital to have a good diagnosis as soon as possible. Machine learning has proven to be effective in medical therapy. The nine most important machine learning arrangement procedures were studied in this paper for forecasting the chronic kidney disease. The accuracy of the classifiers, we utilised in our paper meets our expectations. From the experiment it is found that Optimized Forest classifier performed well, with an accuracy of 100% in comparison to all other classifiers.

References

[1]. Q.-L. Zhang and D. Rothenbacher, ''Prevalence of chronic kidney disease in population-based studies: Systematic review,'' BMC Public Health, vol. 8, no. 1, p. 117, Dec. 2008.

[2]. W. M. McClellan, D. G. Warnock, S. Judd, P. Muntner, R. Kewalramani, M. Cushman, L. A. McClure, B. B. Newsome, and G. Howard, ''Albuminuria and racial disparities in the risk for ESRD,'' J. Amer. Soc. Nephrol., vol. 22, no. 9, pp. 1721–1728, Aug. 2011.

[3]. W. D. Souza, L. C. D. Abreu, L. G. D. SilvaI, and I. M. P. Bezerra, ''Incidence of chronic kidney disease hospitalisations and mortality in Espírito Santo between 1996 to 2017,'' Wisit Cheungpasitporn, Univ. Mississippi Medical Center, Rochester, MN, USA, Tech. Rep., 2019, doi: 10.1371/journal.pone.0224889.

[4]. T. Xiuyi and G. Yuxia, ''Research on application of machine learning in data mining,'' in Proc. IOP Conf., Mater. Sci. Eng., 2018, doi: 10.1088/1757-899X/392/6/06220.

[5]. A. Dhillon and A. Singh, ''Machine learning in healthcare data analysis: A survey,'' J. Biol. Today's World, vol. 8, no. 2, pp. 1–10, Jan. 2018.

[6]. M. Ghassemi, T. Naumann, P. Schulam, A. L. Beam, I. Y. Chen, and R. Ranganath, ''A review of challenges and opportunities in machine learning for health,'' in Proc. AMIA Joint Summits Transl. Sci., 2020, p. 191.

[7]. J. Qin, L. Chen, Y. Liu, C. Liu, C. Feng, and B. Chen, ''A machine learning methodology for diagnosing chronic kidney disease,'' IEEE Access, vol. 8, pp. 20991–21002, 2020.

[8]. G. R. Vasquez-Morales, S. M. Martinez-Monterrubio, P. Moreno-Ger, and J. A. Recio-Garcia, ''Explainable prediction of chronic renal disease in the colombian population using neural networks and case-based reasoning,'' IEEE Access, vol. 7, pp. 152900–152910, 2019.

[9]. Y. Amirgaliyev, S. Shamiluulu, and A. Serek, ''Analysis of chronic kidney disease dataset by applying machine learning methods,'' in Proc. IEEE 12th Int. Conf. Appl. Inf. Commun. Technol. (AICT), Oct. 2018, pp. 1–4.

[10]. L. Kilvia De Almeida, L. Lessa, A. Peixoto, R. Gomes, and J. Celestino, ''Kidney failure detection using machine learning techniques,'' in Proc. 8th Int. Workshop ADVANCEs ICT Infrastructures Services, 2020, pp. 1–8.

[11]. M. Almasoud and T. E. Ward, ''Detection of chronic kidney disease using machine learning algorithms with least number of predictors,'' Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 8, pp. 89–96, 2019

[12]. Scholar, P. G. "Chronic kidney disease prediction using machine learning." International Journal of Computer Science and Information Security (IJCSIS) 16.4 (2018).

[13]. S. Drall, G. S. Drall, S. Singh, and B. B. Naib, ''chronic kidney disease prediction using machine learning: A new approach,'' Int. J. Manage., Technol. Eng., vol. 8, pp. 278–287, May 2018.

[14]. Rennie, Jason & Shih, Lawrence & Teevan, Jaime & Karger, David. (2003). Tackling the Poor Assumptions of Naive Bayes Text Classifiers. Proceedings of the Twentieth International Conference on Machine Learning. 41.

[15]. https://www.javatpoint.com/bayesian-belief-network-in-artificial-intelligence

[16].] B. Schölkopf, C. Burges, and V. Vapnik, ''Incorporating invariances in support vector learning machines,'' in Proc. Int.

[17]. Conf. Artif. Neural Netw. Berlin, Germany: Springer, 1996, , pp. 47–52

[18]. https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc

[19]. Ghaemi, Manizheh, and Mohammad-Reza Feizi-Derakhshi. "Forest optimization algorithm." Expert Systems with Applications 41.15 (2014): 6676-6687.

[20]. http://infochim.u-strasbg.fr/cgi-bin/weka-3-9-1/doc/weka/classifiers/rules/PART.html
[21]. https://www.sen.uni-konstanz.de/research/research/tools/k-star-algorithm/
[22]. https://www.researchgate.net/publication/33053063_Best-first_Decision_Tree_Learning
[23]. http://rasbt.github.io/mlxtend/user_guide/classifier/OneRClassifier/