International Journal of Recent Research and Review, Vol. XIII, Issue 4, December 2020
ISSN 2277 – 8322

Performance Analysis of Various Machine Learning Classifierson Reduced Chronic Kidney Disease Dataset

Md. Tanjeel Islam Khan[1], Md. Sazzadul Islam Prottasha[2],Tasneem Alam Nasim[3],Abdullah Al Mehedi[4], Md. Appel Mahmud Pranto[5], Nafiz Al Asad[6]

Information and Communication Technology, Bangladesh University of Professionals, Dhaka, Bangladesh
tanjeel27@gmail.com

***Abstract*— Chronic kidney disease (CKD) is considered as a lethal disease all over the world. Chronic kidney disease (CKD) is a condition where kidney shrinks in size and also changes its natural shape. Various machine learning algorithms can be very useful for prediction of CKD. This paper investigates the performance of various machine learning algorithm on chronic kidney disease dataset. Support Vector Machine (SVM), Decision Tree, Naïve Bayes, Random Forest and Logistic Regression are the algorithms considered in this paper. Initially a dataset of 400 instances having 24 attributes is considered. Later feature selection algorithm is used to identify the important attributes and we reduced the uncorrelated attributes and observed the results. Results show that Naïve Bayes achieved the maximum accuracy of 99.1% on reduced chronic kidney disease dataset of 23 attributes. In terms of time complexity decision tree performed better than the other classifiers. It is expected that the application of different machine learning algorithms can help to predict CKD with great accuracy in practice.**

***Keywords*—Chronic Kidney Disease(CKD); Machine Learning, Support Vector Machine(SVM); Decision Tree; Naïve Bayes; Decision Tree, Random Forest, Logistic Regression**

I. INTRODUCTION

A person's kidney health condition can be determined from different biological symptoms and attributes. A healthy kidney's main purpose is to refine the toxic elements from human body. A kidney works more like a filter which filters out all kinds of toxic elements from our blood. The toxic elements are mostly produced in our body. Toxic elements can also be apprehended from the environments. The kidney works as a natural filter and protect the human body from these toxic elements. There many kinds of kidney diseases and the Chronic Kidney Disease (CKD) is one of them. At the moment, CKD is one of the deadliest diseases all around the globe. It is also becoming very deadly in our country because of the new trend of leading an unhealthy lifestyle. CKD is a condition in which a kidney shrinks in size and also changes its natural shape. A diseased kidney cannot perform the job properly which leads to some changes in the patient's body. The changes in the patient's body can be detected with the help of some medical tests. The test results reflect the changes in the body by some biological attributes. Machine learning algorithms have been recognizedas a very accurate method in classifying and predicting diseases over the past few years. The use of machine learning in the medical field is spreading extensively. With the growth of access to different types of medical data has paved the way for modern medical science. The development of medical service and diagnosis of diseases has entered into a new era with the application of machine learning in the medical field.

The main purpose of using machine learning algorithms in the medical field is to help the doctors to predict diseases in a faster and easier way. The popularity of machine learning algorithms in the medical field, especially in disease prediction, has influenced and motivated us to implement Support Vector in CKD prediction and also compare the performance of SVMwith Decision Tree, Naïve Bayes, Random Forest and Logistic Regression.

This paper is divided into several sections. Section II shows previous works using machine learning algorithms to predict diseases. Section III describes the methodology of different algorithms and a brief discussion. Section IV describes the result and analysis and findings of the research. In section V, conclusion of this paper has been described.

II. LITERATURE REVIEW

The use of machine learning algorithms in predicting diseases has become a popular research topic around the globe. Accuracy of prediction depends on different factors on different datasets.

For diagnosing the diabetes mellitus, the authors N.H. Barakat, A.P. Bradley, and M.N.H. Barakat have used the Support Vector algorithm [3]. In this paper, they have used a database consisting of 12 attributes and 4682 instances. It was a diverse dataset including people with different biological characteristics such as waist width, hip width, age, etc. The diagnosis of diabetes was done using a linear SVM kernel. They introduced a cost factor which optimizes the algorithm for better accuracy. The authors found 89% accuracy for diagnosing diabetes using this particular database.

T.K. Wu demonstrates learning disability using ANN and SVM [4]. This paper helps to detect the learning disability in children in an early stage so that proper measures can be taken to counter that. The authors stated that the ANN gives a good performance for most of the cases. ANN classifier correctly identifies and gives a 100% confidence for half of the subjects. SVM gives a quite good performance. The authors found that for some cases SVM gives better performance than the ANN. The authors used genetic algorithm-based feature selection for boosting the accuracy of the algorithm.

The work at [17] used Naive Bayesian classifier, back-propagation learning of neural networks, decision trees and k-nearest neighbors' method to diagnosis heart disease. Back-propagation learning of neural networks gives accuracy of 80% accuracy on a dataset of 327 patients' ECG result.

In [10] authors used K-nearest neighbors, support vector machine with Gaussian kernel, logistic regression and decision tree on predictive analytics for chronic kidney disease. This work used dataset of UCI repository [8] which contains 400 instances with 24 attributes. Support Vector Machine was best approach which gives accuracy of 98.3%.

The work on [11] also used the same dataset to predict chronic kidney disease based on support vector machine by feature selection methods. This work gives accuracy of 98.5% on this algorithm. In [13] Naive Bayes was used on a reduced dataset which gave accuracy of 97.5%.

III. RESEARCH METHODOLOGY

A. *Background*

Support vector machine (SVM) is a supervised machine learning algorithm that gives satisfactory result for analyzing data for classification and regression analysis. Being a supervised machine learning algorithm, we need to input training data set into our machine for SVM. SVM then uses a training data set for building a model by inserting the values (training data set) in its algorithm. SVM uses the model for classification and linear regression when it deals with new data. There are generally two classes and the model assigns the new data that arrives at one of the classes which makes the SVM algorithm a binary linear classifier. A decision tree is a wide part of machine learning that contains both classification and regression [5]. A decision can be visually add explicitly represented in decision tree analysis. As the name suggests, a tree-like model is used to represent a decision. Theattributes or features of a dataset are indicated by the nodes of a decision tree where each node delivers a decision. Leaf node represents the outcome which will be in categorical response or numerical value. Decision tree is written from top to bottom approach.

Naive Bayes Classifier is a classifier algorithm based on the principle of Bayes Theorem [6]. Naive Bayes classifiers work in a simple but efficient way. Feature of datasets in Naive Bayes classifiers are mutually independent. To understand Naive Bayes classifiers knowledge of Bayes theorem is necessary. The Bayes theorem calculates the probability of an event occurring provided that an event has already occurred. Probability Model of Naive Bayes can be written as –

$$\text{Posteriori Probability} = \frac{\text{Conditional Probability} * \text{PrioriProbability}}{\text{evidence}} \quad (1)$$

Mathematical representation can be defined as -

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (2)$$

Though its name suggests regression, logistic regression is used for binary classification [7]. In other

words, logistic regression gives a Booleanoutput. It builds a correlation between on dependable variable and another independent variable by approximating probabilities using a logistic function. Probabilities should be converted into binary value to predict an output. Logistic regression uses a sigmoid function to predict a value. Sigmoid function is an S-shaped curve that takes real values and maps it in 0 and 1. Following equation is used to express sigmoid function -

$$f(x) = \frac{1}{1+e^{-x}} \quad (3)$$

Random Forest is one of the supervised machine learning algorithms which is used for both classification regression [7]. A random forest classifier initially makes decision tree from randomly selected training set. As it is called ensemble method, it chooses the best solution by aggregating and bootstrapping. It overcomes the problem of over fitting from decision tree classifier. It also gives faster and accurate result than decision tree.

B. Dataset

The CKD dataset from the UCI website has been used as both training dataset and testing input for this paper [8]. The dataset includes 400 instances with 25 attributes including age, blood pressure, albumin, sugar, red blood cells, serum creatinine, sodium, potassium, diabetes mellitus, anemia, hemoglobin, blood urea, bacteria etc. The dataset is not from a certain locality, therefore the result wouldn't vary based on the region the people live in. The dataset also doesn't focus on a certain age, meaning that the data are collected from people of random ages. Table I shows a brief description of the dataset and it's attributes. Table II shows the attribute names and their abbreviation of the CKD dataset.

TABLE I

DATASET AND ATTRIBUTES

Dataset Characteristics	Multivariate	**No. of Instances**	400
Attribute Characteristics	Real	**No. of Attributes**	25

Different feature selection algorithms [18] such as Individual Feature selection, Forward and backward propagation has been used to reduce the dimension of CKD dataset. By analyzing the attributes correlation with the class value, we discarded some of the attributes and taken only the important features.

C. Methodology

At first we have preprocessed our dataset and used different feature selection algorithms to identify the important attributes of the dataset. Using the Individual feature selection, we selected 23 attributes, using Forward Feature Selection we selected 21 attributes and using Backward Feature Selection we selected 20 attributes from the CKD dataset. Once the attributes were selected, we applied different machine learning algorithms on the reduced dataset.

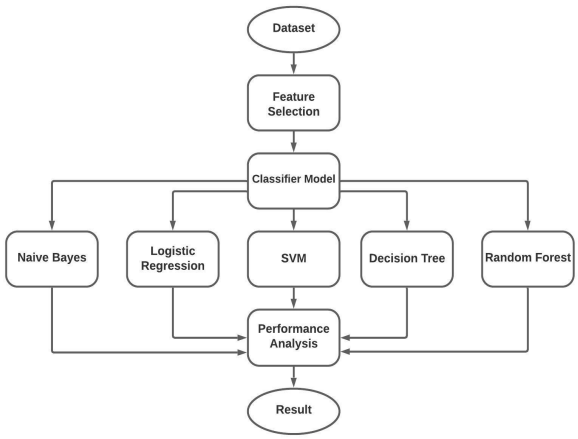

Fig. 1. Methodology for CKD classification

TABLE II

DATASET ATTRIBUTES AND THIR NAMES

ID	**Attributes**	**ID**	**Attributes**
1	Age	13	sod - sodium
2	bp - blood pressure	14	pot - potassium
3	sg - specific gravity	15	hemo - hemoglobin
4	al – albumin	16	pcv - packed cell volume
5	su – sugar	17	wc - white blood cell count
6	rbc - red blood cells	18	rc - red blood cell count
7	pc - pus cell	19	htn - hypertension
8	pcc - pus cell clumps	20	dm - diabetes mellitus
9	ba – bacteria	21	cad - coronary artery disease
10	bgr - blood glucose random	22	appet - appetite
11	bu - blood urea	23	pe - pedal edema
12	sc - serum creatinine	24	ane - anemia
		25	class

10-fold cross validation with 80-20 data split has been used for training and testing of the data

respectively. Then accuracy, precision, recall, f1-score and AUC have been measured using confusion matrix.

IV. RESULT ANALYSIS

The experiment was mainly based on Python programming language. Python3.6 has been used throughout the whole procedure. Jupyter Notebook was used for evaluation of the datasets. Scikit-Learn, Numpy, Pandas and Matplotlib were the main python packages that have been used in this experiment as it provides different types of supervised and unsupervised learning algorithms.It alsointeroperates with Python libraries-SciPy and NumPy. It is used for machine learning and data mining to show live code, equation, visualization and narrative text.

TABLE III
PERFORMANCE ANALYSIS OF VARIOUS CLASSIFIER ON RAW DATASET WITH 25 ATTRIBURES

Classifier	Accuracy	Precision	Recall	F1 score	AUC
Naïve Bayes	98.2%	97.2%	99.7%	98.5%	99.7%
Logistic Regression	98.4%	97.6%	100%	98.5%	99.9%
SVM	97.5%	96.8%	100%	97.3%	99.5%
Decision Tree	92.3%	92.1%	98.4%	94.8%	98.3%
Random Forest	97.9%	96.4%	100%	98.4%	99.5%

In this experiment, confusion matrix has been used to measure the accuracy. A Confusion matrix is a table that is built of true positive, false positive, false negative and true negative values [9]. Accuracy was measured by the total correct value that is true positive and true negative was given from the total value. Precision, recall, f1-score and Area under ROC have also been measured. Precision

TABLE IV
PERFORMANCE ANALYSIS OF VARIOUS CLASSIFIER ON INDIVIDUAL FEATURE SELECTION WITH 23 ATTRIBURES

Classifier	Accuracy	Precision	Recall	F1 score	AUC
Naïve Bayes	99.1%	98.7%	99.7%	99.3%	100%
Logistic Regression	98.1%	97.2%	100%	98.3%	99.3%
SVM	98.9%	97.5%	100%	99.1%	99.9%
Decision Tree	94.9%	93.1%	100%	94.8%	98.1%
Random Forest	98.3%	97.4%	100%	98.7%	99.5%

is whether predicted positive value is predicted true. Recall is when the result actually positive whether it is predicted true. F1- score is measured from a harmonic mean of precision and recall.

TABLE IV
PERFORMANCE ANALYSIS OF VARIOUS CLASSIFIER ON INDIVIDUAL FEATURE SELECTION WITH 23 ATTRIBURES

Classifier	Accuracy	Precision	Recall	F1 score	AUC
Naïve Bayes	97.5%	97 %	99.9%	97.7%	99.5%
Logistic Regression	91.7%	91.1%	98.8%	92.1%	98.5%
SVM	97.7%	97.3%	100%	97.8%	99.7%
Decision Tree	93.9%	93.5%	99.2%	94.6%	98.9%
Random Forest	94.2%	93.3%	99.8%	94.7%	99 %

This experiment evaluated the performance of different classifier with 25 attributes at first and then evaluated performance with feature selection on reduced dataset. Table III shows the shows performance analysis of various classifier on raw data with 25 attributes.Table IV shows performance analysis of various classifiers on feature selection with 23 attributes.Table V shows performance analysis of various classifiers on forward feature selection with 21attributes. Table VI shows performance analysis of various classifiers on backward feature selection with 20 attributes. Table VII shows overall performance of feature selection algorithms with various algorithms' accuracies. Initially raw data was used with 25 attributes.Then we reduced the attributes using individual feature selection, forward feature selection and backward feature selection algorithm.

TABLE VI
PERFORMANCE ANALYSIS OF VARIOUS CLASSIFIER ON BACKWARD FEATURE SELECTION WITH 20 ATTRIBURES

Classifier	Accuracy	Precision	Recall	F1 score	AUC
Naïve Bayes	96.2%	95.8%	99.5%	96.5%	99.2%
Logistic Regression	90.3%	89.8%	98.4%	90.6%	98.1%
SVM	96.8%	96.5%	100%	97.1%	99.5%
Decision Tree	91%	90.3%	98.9%	91.4%	98.3%
Random Forest	92.1%	91.6%	99.2%	92.4%	98.7%

By analyzing the results reported in Table VII we can see that initially the individual feature selection method increases the overall accuracy for each classifier.

TABLE VII
PERFORMANCE ANALYSIS OF VARIOUS CLASSIFIER ALGORITHMS' ACCURACIES

Method	Attributes	Naïve Bayes	Logistic Regression	SVM	Decision Tree	Random Forest
Raw data	24	98.2%	98.4%	97.5%	92.3%	97.9%
Individual feature selection	23	**99.1%**	98.1%	98.9%	94.9%	98.3%
Forward feature selection	21	97.5%	91.7%	97.7%	93.9%	94.2%
Backward feature selection	20	96.2%	90.3%	96.8%	91%	92.1%

However, reducing the attributes further didn't increased the accuracy rather decreased it. Therefore, we can conclude that forward feature selection and backward feature selection algorithm didn't work well for CKD dataset.

In our research we have the following findings about the CKD disease dataset.

1. Naïve Bayes outperformed all other base classifiers in terms of accuracy
2. Decision tree took minimum time to train and test the data, however the accuracy is a bit low.
3. For Individual feature selection with 23 attributes, the classifiers performed best.
4. While the attributes were further reduced, the classification accuracy didn't improve rather it reduced.
5. With reduction of attributes, time complexity also reduced for different classifiers.

V. CONCLUSION

This paper focuses on the prediction of Chronic Kidney Disease on reduced Dataset. The prediction is done on the basis of biological attributes using the naïve Bayes, SVM, decision tree, RandomForestand logistic regression. The algorithms were executed on a CKD dataset which consists of 400 instances and 24 attributes.

The result of the experiment has shown that the Naïve bayes classifier shows better accuracy in predictingCKD than the other machine learning algorithms. For Individual feature selection algorithm with 23 attributes, the accuracy is highest with 99.1%.The other classifiers also gives satisfactory outcome on the CKD dataset. The lowest accuracy is provided by decision tree on 2o attributes during backward feature selection method.

In future, we are planning to work with a large and versatile CKD dataset and implement various feature selection methods on that. Also, we will apply deep learning and other machine learning algorithms on CKD dataset to further increase the accuracy.

VI. REFERENCES

[1] T. J. Hastie, J. H. Friedman, and R. J. Tibshirani, The elements of statistical learning: data mining, inference, and prediction. New York: Springer, pp. 417-438, 2017.

[2] J. Han, M. Kamber, and J. Pei, Data mining: concepts and techniques. Amsterdam: Elsevier, pp 423-425, 2012.

[3] N. Barakat, A. P. Bradley, and M. N. H. Barakat, "Intelligible Support Vector Machines for Diagnosis of Diabetes Mellitus," IEEE Transactions on Information Technology in Biomedicine, vol. 14, no. 4, pp. 1114–1120, 2010.

[4] T.-K. Wu, S.-C. Huang, and Y.-R. Meng, "Evaluation of ANN and SVM classifiers as predictors to the diagnosis of students with learning disabilities," Expert Systems with Applications, vol. 34, no. 3, pp. 1846–1856, 2008.

[5] I. H. Witten and E. Frank, Data mining: practical machine learning tools and techniques with Java implementations. San Francisco, CA: Morgan Kaufmann, pp. 189-199, 2000.

[6] G. Casella, and R.L. Berger, Statistical inference. Belmont, CA: Brooks/COle Cengage Learning, pp 591-592, 2017.

[7] P.-N. Tan, M. Steinbach, A. Karpatne, and V. Kumar, Introduction to data mining. New York, NY: Pearson Education, Inc., 2019.

[8] UCI Machine Learning Repository: Data Set. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease. [Accessed: 30-January-2019].

[9] C. Sammut and G. I. Webb, "Confusion Matrix," in Encyclopedia of machine learning and data mining. New York, NY: Springer, 2017.

[10] A. Charleonnan, T. Fufaung, T. Niyomwong, W. Chokchueypattanakit, S. Suwannawach, and N. Ninchawee, "Predictive analytics for chronic kidney disease using machine learning techniques," 2016 Management and Innovation Technology International Conference (MITicon), 2016.

[11] H. Polat, H.D. Mehr, and A. Cetin, "Diagnosis of Chronic Kidney Disease Based on Support Vector Machine by Feature Selection Methods," Journal of Medical Systems, vol. 41, no. 4, 2017.

[12] C. Nordqvist, "Chronic kidney disease: Symptoms, causes, and treatment," Medical News Today, 13-Dec-2017. [Online]. Available: https://www.medicalnewstoday.com/articles/172179.php. [Accessed: 30-January-2019].

[13] U.N. Dulhare, and M. Ayesha, "Extraction of action rules for chronic kidney disease using Naïve bayes classifier". IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), 2016.

[14] A.S. Levey, J. Coresh,E. and Balk, "National kidney foundation practice guidelines for chronic kidney disease: evaluation, classification and stratification.", ANN Intern Med, pp 137-139, 2003.

[15] S.C Liao, I. and Lee, "Appropriate medical data categorization and specification for data mining classification techniques" Med Inform, Vol 27, no 1, pp 59-67, 2002.

[16] N. Lavrac, E. Keravnou, and B. Zupan, "Intelligent data analysis in medicine," in Encyclopedia of Computer Science and Technology, vol. 42, A. Kent et al., Ed. New York: Marcel Dekker, pp. 113–157, 2000.

[17] M. Kukar, I. Kononenko, C. Grošelj, K. Kralj, and J. Fettich, "Analysing and improving the diagnosis of ischaemic heart disease with machine learning." Artificial Intelligence in Medicine, vol. 16, no. 1, pp. 25-50, 1999

[18] Antonio Arauzo-Azofra, José Luis Aznarte, José M. Benítez, Empirical study of feature selection methods based on individual feature evaluation for classification problems, Expert Systems with Applications,Volume 38, Issue 7, 2011,Pages 8170-8177, ISSN 0957-4174