Data Mining Algorithms for Kidney Disease Stages Prediction

Abdelrahim Koura 1, Hany S. Elnashar2

1 Computer Science Dept., Faculty of Computers and Artificial Intelligent, Beni-Suef University, Egypt,

2 Faculty of Computers and Artificial Intelligent , Beni-Suef University, Egypt hsoliman@fcis.bsu.edu.eg.

Abstract: One of the most common health problems that correlated to serious complications is chronic kidney disease. Early detection and treatment can save it from progression. Machine learning is one tool that used historical data to improve future decision about prediction of chronic kidney disease. The aim of this work is to compare the performance of six different models based on accuracy, sensitivity, precision, recall. In this study, the experiments were conducted on 158 records downloaded from UCI repository. Six algorithms (K-Nearest Neighbor, Naïve Bayes, Support Vector machine, Logistic Regression, Decision Tree, and Random Forest) were implemented on data after preprocessing stage. Evaluation of models resulted in Naïve Bayes and Random Forest accuracy 100%, Sensitivity 100%, Specificity 100%, precision 100 %, Recall 100% respectively. It is concluded that Naïve Bayes and Random Forest are better than other models.

Keywords: Data mining, Kidney Disease(KD), Feed Forward Neural Network; Levenberg-Marquardt; Multi-Layer Perceptron; Particle Swarm Optimization.

1. Introduction

Chronic kidney disease CKD is central problem. CKD disease has an increased incidence and prevalence, poor results and high costs in the world's public health problem [1]. an increasing number of people with kidney failure who are treated with dialysis and transplantation have been recorded during last ten years ref. The major outcomes of chronic kidney disease regardless of cause, include progression to kidney failure, complications of decreased kidney function, and cardiovascular disease (CVD) [2] . Increasing evidence indicates that some of these adverse outcomes can be prevented or delayed by early detection and treatment (2). For many applications, particularly those with complex dimensions, machine learning with classification can effectively be applied [3]. Classification methodology can therefore be used to predict conditions such as cancer and heart disease etc. requiring complicated measurements. This is part of increasing demand for predictive diagnoses and is very interesting. Classification and learning methods have also been proven to be effective in improving the accur acy and recurrence of disease prediction.The research methods in the present laboratory include t he assisting of a vector machine [SVM] and of Random Forest [RF] [3]. This research presents a comparison between more than one machine learning algorithms for prediction of a patient with chronic kidney disease using recorded patient data.

DOI: 10.5281/zenodo.3686791

2. RELATED WORK

Many algorithms based on Machine learning have been used to classify and prediction in the medical analysis and healthcare fields. started by using the Support Vector Machine Algorithm to classify and predict diabetes and pre-diabetes patients, and the results show that SVM is useful to classify patients with common diseases [4]. and have classified Alzheimer's disease by analyze wh ole-brain anatomical magnetic resonance imaging (MRI) for a set of patients, and the results shows that SVM is a promising approach for Alzheimer's disease early detection [5]. for heart disease prediction using the Probabilistic Neural Network Algorithm, Decision tree Algorithm, and Naïve Bayes Algorithm, and PRNN provides the best results compared with other algorithms for heart disease prediction [6]. more have done prediction of HBV-induced liver cirrhosis using the Multilayered Perceptron (MLP) Algorithm and the results shows that the MLP classifier gives satisfactory prediction outputs for liver disease, mostly in HBV-related liver cirrhosis patients [7].

dataset given as input to the network and 30% of the dataset is given as unseen to the network. The implemented network gives the error rate at 0.0773 of MSE and the accuracy as 90 % [10].

3. Materials & Methods

3.1 data Descriptions

This work proposed using some classification techniques to predict the presence of chronic kidney disease in humans. In this study the chronic kidney disease dataset downloaded from UCI Machine Learning Repository which is a collection of databases [8], domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms [9]. The chronic kidney data set collected from the hospital nearly 2 months of period. The original Dataset contains four hundred instances and twenty-five attributes. list of attributes and missing values description Table(1). Before integrating the gathered instances with any classification technique, it is essential to prepare the complete set of data with unique representation.

At first, aggregated data contains missing value, in particular to this work the rows with missing values was removed from dataset. After removing the records with missing values. The final dataset became one hundred and eighty -five instances used for analysis kidney disease. some instance has categorical values for some attribute. The categorical values were replaced with number values for those instances.

This work is performed in Python which is a powerful general-purpose programming language and is becoming an increasingly popular tool in research. Python library allows matrix manipulations, plotting of functions and data, implementation of algorithms. The experimental comparison of K-nearest neighbor (KNN), Support vector machine (SVM), Decision tree (DT), Random forest (RF), Logistic Regression (LR), and Naïve Bayes (NV) are done based on the performance measures of classification accuracy and precision. The dataset used in this analysis consisted of 158 records, one-hundred and fifteen chronic kidney and forty-three was not chronic kidney diseases.

DOI: 10.5281/zenodo.3686791

Precision, Recall and Accuracy percentage and F1score were employed to evaluate the performance of the utilized classification algorithms.

Table(1): list of attributes and missing values description

#	variable	Class	type	missing row
1	Age	Predictor	Numerical	9
2	Blood_pressure	Predictor	Numerical	12
3	Specific_Gravity	Predictor	Categorical	47
4	Albumin	Predictor	Categorical	46
5	Sugar	Predictor	Categorical	49
6	Red_Blood_Cells	Predictor	Categorical	152
7	Pus_Cell	Predictor	Categorical	65
8	Pus_Cell_Clumps	Predictor	Categorical	4
9	Bacteria	Predictor	Categorical	4
10	Blood_Glucose_Random	Predictor	Numerical	44
11	Blood_Urea	Predictor	Numerical	19
12	Serum_Creatinine	Predictor	Numerical	17
13	Sodium	Predictor	Numerical	87
14	Potassium	Predictor	Numerical	88
15	Hemoglobin	Predictor	Numerical	52
16	Packed_Cell_Volume	Predictor	Numerical	70
17	White_Blood_Cell	Predictor	Numerical	105
18	Red_Blood_Cell	Predictor	Numerical	130
19	Hypertension	Predictor	Categorical	2
20	Diabetes_Mellitus	Predictor	Categorical	2
21	Coronary_Artery_Disease	Predictor	Categorical	2
22	Sex	Predictor	Categorical	1
23	Pedal_Edema	Predictor	Categorical	1
24	Anemia	Predictor	Categorical	1
25	Class	Target	Numeric	0

3.2 k nearest neighbor

The k-nearest neighbors (KNN) algorithm is an easy way to implement supervised algorithm that used to solve classification and regression cases [10]. It works in somehow of 8 steps started by Load the data, Initialize K to your neighbors, For each case in the data, Then find the distance between the current node and the desired neighbor from the data, Get the distance and rewrite in an ordered collection, Sort the ordered collection of distances and indices from smallest to largest (in ascending order) by the distances, Pick the first K entries from the sorted collection, and finally Get the labels of the selected K entries, If regression, return the mean of the K labels, If classification, return the mode of the k [11].

3.3 Naive bias

DOI: 10.5281/zenodo.3686791

It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors [12]. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. 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 that this fruit is an apple and that is why it is known as 'Naive'. Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c). Look at the equation below [13]:

3.4 Random Forest

Random forests are one of the most powerful and fully automated machine learning techniques [14]. It does not need any data preparation or modeling expertise, It enables analysts to obtain stunningly effective models, and is also a tool that embodies the power of decision trees, judicious randomization, and ensemble learning to produce amazingly accurate predictive models, insightful classifications of variables, and assumptions of the importance of variables. , For a deeper understanding The basic building block of a random forest is inspired by the decision tree (Classification and Regression Trees) [15], which is one of the machine learning methods for building predictive models of data. Models are obtained by dividing data and building a simple prediction model within each section.

3.5 Support vector

Support Vector Machines are decision planes-based concept [16]. this decision planes define boundaries. A decision plane is one that run as separator between a Groups of objects having classes relations. fig xx show that the items belong either to class GREEN or RED. Where line defines a boundary. on one side of which all items are GREEN otherwise on the left are RED. and Any new items (white item) falling to the right is labeled, as GREEN (or as RED should it at the left of the line). that line L defines decision boundaries.

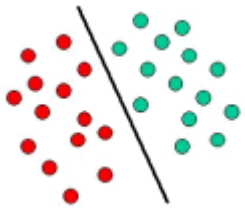

Figure 1: Separation line

This is a classic example of a linear classifier, a classifier separates a set of items to its groups (GREEN or RED) with a line. Most classification tasks are not that simplicity, and more complex structures are required to make an optimal separation, i.e., for classify new objects (test) on the basis of the examples that are available (train). Now it is clear that as in fig xy a full separation of the GREEN and RED objects would require a curve (which is more complex than a last one).

DOI: 10.5281/zenodo.3686791

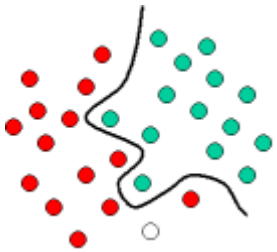

Figure 2: Separation Curve

Classification tasks based on first one to distinguish between items from different class memberships are named as hyper plane classifiers. Support Vector Machines are particularly suited for this type. Support Vector Machine (SVM) is The key classification system, in which hyper- planes are constructed in a multidimensional space separating cases with category marks, carries out classification task [17]. SVM can handle regression and classification tasks as well as various continuous and categorical variables. A dummy variable with cases of either 0 or 1 is generated for categorical variables.

A: {1 0 0}, B: {0 1 0}, C: {0 0 1}

SVM uses an iterative training algorithm to build an optimal hyperplane that minimizes error [18].

3.6 Decision Tree

Using tree representation by mapping each internal node of the tree corresponds to an attribute, and each leaf node corresponds to a class label. To understand Decision Trees algorithm is very easy compared to other classification algorithms [19].

Algorithm of Decision Tree, name all best attributes from dataset in start of the tree root, then make training set into many subsets, each one contains specific data have same value of attribute. And construct leaf nodes of branches of the tree by Redoing last two steps

Class label predictions record need to start by the tree **roots**. Then comparing values of attribute at root with attribute records. From that bases, branch related to that value then jump a new node. Repeating this process until reaching **a leaf node named by** predicted values of the class [20].

3.7 Logistic Regression

Logistic regression is a classification algorithm used to assign observations to a discrete set of classes. Unlike linear regression which outputs continuous number values, logistic regression transforms its output using the logistic sigmoid function to return a probability value which can then be mapped to two or more discrete classes[21].

DOI: 10.5281/zenodo.3686791

4. Result and discussion

In order to test our proposed method, the experiments were conducted for prediction task for chronic kidney disease by separately applying six machine learning algorithms namely: Random Forest (RF), Naïve-Bayes, Decision tree, logistic regression, K-nearest neighbor, and support vector machine. The classification performances of the classifiers were analyzed with respect to the standard performance parameters, namely: Accuracy, Specificity, Sensitivity, Precision, Recall and F1 score as shown in table 2

Table2: Parameters Equation of computation

Sensitivity = $\frac{T}{T+F} * 100$	Specificity = $\frac{T}{T+F} * 100$
Precision = $\frac{T}{T+F}$	Recall = $\frac{T}{T+F}$
Accuracy = Number of correctly classified samples/Total number of samples	F1 score=2*precision*recall/precision +recall

Where:

- TP is Number of true positive classification cases
- FN is Number of false negative classification cases
- TN is Number of true negative classification cases
- FP is Number of false positive classification cases

The sensitivity states the ability of the model to identify positive instances correctly, the specificity shows the ability of the model to identify negative instances correctly and precision indicates the ability of the model to identify positive instances correctly. Precision and recall It is also called positive predictive value. It is defined as the average probability of relevant retrieval. Recall It is defined as the average probability of complete retrieval. The accuracy indicates the percentage of correct classification of both positive class as well as negative class instances. Confusion Matrix It displays the number of correct and incorrect predictions made by the model compared with the actual classifications in the test data. The matrix is represented in the form of n-by-n, where n is the number of classes. The accuracy of each classification algorithms can be calculated from that matrix.F1 that calculates the harmonic mean of the precision and recall because both of them are rates. The overall accuracy and F1 score showed Table 4.

Table 3: The value of Specificity, Sensitivity, Precision, and Recall

	K-NN	SVM	LR	NB	DT	RF
True Positive	35	35	35	35	35	35
True Negative	11	11	11	13	13	13
False Positive	0	0	0	0	0	0
False Negative	2	2	2	0	0	0
Sensitivity	94%	94%	94%	100%	100%	100 %
Specificity	100%	100%	100%	100%	100%	100 %
Precision	1	1	1	1	1	1
Recall	.94	.94	.94	1	.92	1

DOI: 10.5281/zenodo.3686791

Table 4: Accuracy for Each Classifier

Algorithm	Accuracy %	F1 score
K-Nearest Neighbor	95.8	.91
Naïve Bayes	100	1
Support Vector machine	95.8	.91
Logistic Regression	95.8	.91
Decision Tree	97.91	.96
Random Forest	100	1

As per this results figure 3 show graphical comparisons for each algorithm and the given results relative to the other

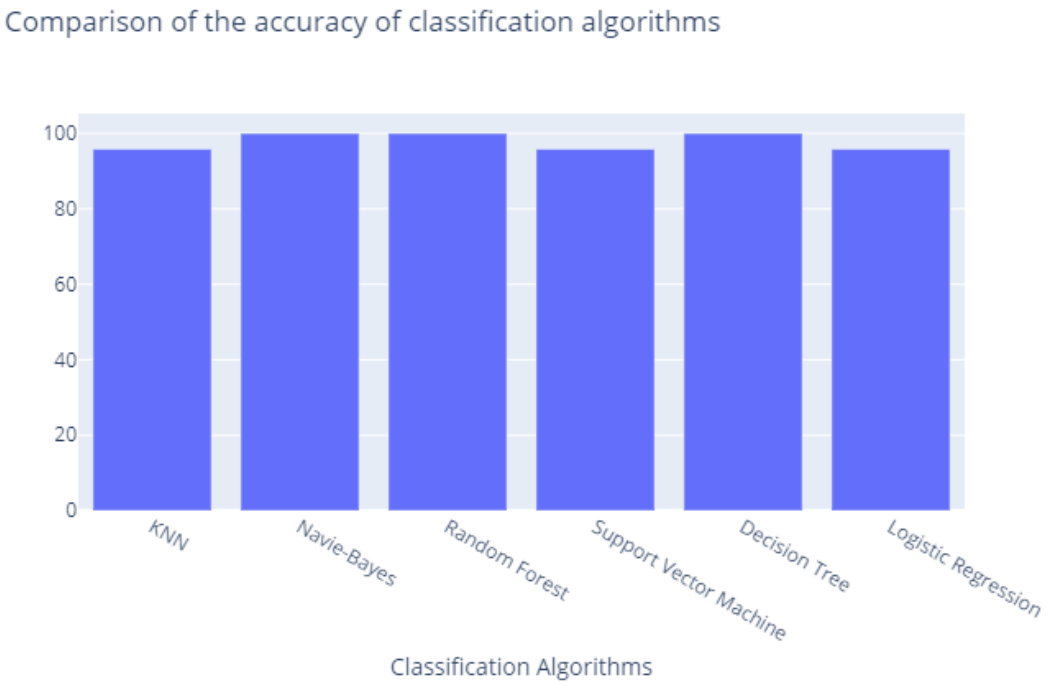

Figure 3: Illustrates the findings of the accuracy of all models

DOI: 10.5281/zenodo.3686791

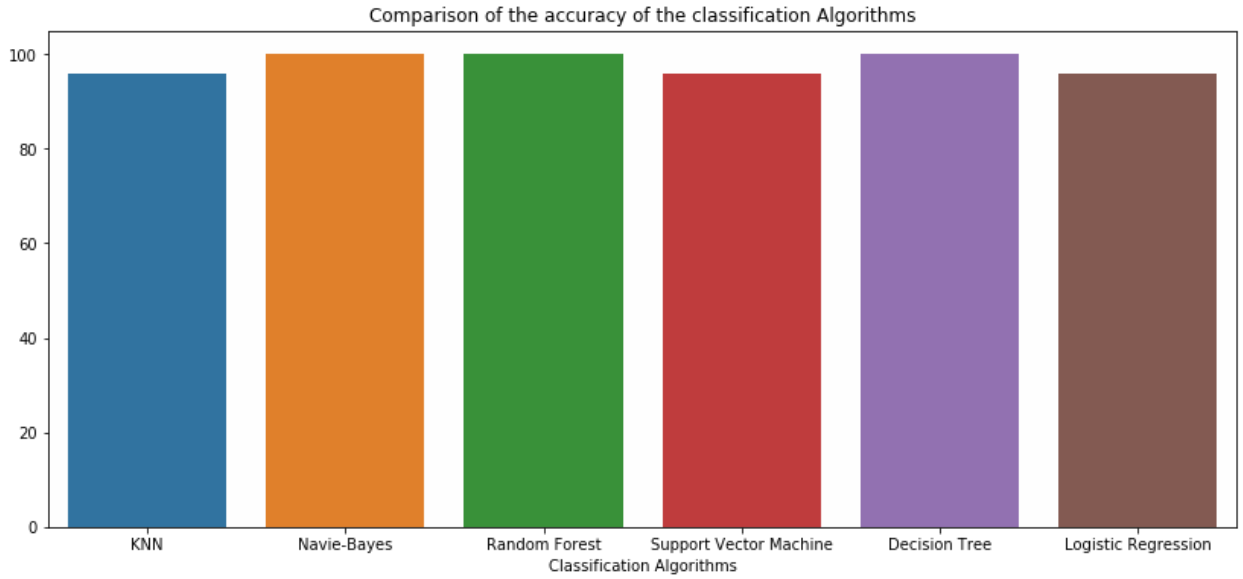

Figure 4: Compairason of accyracy

The algorithms RF, DT, and NB are performed better than other classifiers with sensitivity, specificity and accuracy values 1.00, 1.00 and 1.00 respectively.

5. Conclusion

In this research work classification methods are used to predict chronic kidney disease. six algorithms performance are compared. The experimental results of our proposed method have stated that Random Forest, and Naïve Bayes are produced superior prediction performance in terms of classification accuracy.

References

[1] E. H. A. Rady and A. S. Anwar, "Prediction of kidney disease stages using data mining algorithms," *Informatics in Medicine Unlocked*, vol. 15. Elsevier Ltd, 01-Jan-2019, doi: 10.1016/j.imu.2019.100178.

[2] W. Yu, T. Liu, R. Valdez, M. Gwinn, and M. J. Khoury, "Application of support vector machine modeling for prediction of common diseases: The case of diabetes and pre-diabetes," *BMC Med. Inform. Decis. Mak.*, vol. 10, no. 1, 2010, doi: 10.1186/1472-6947-10-16.

[3] A. S. Levey *et al.*, "National Kidney Foundation Practice Guidelines for Chronic Kidney Disease: Evaluation, Classification, and Stratification," 2003.

[4] A. K. Ahmed, S. Aljahdali, and S. Naimatullah Hussain, "Comparative Prediction Performance with Support Vector Machine and Random Forest Classification Techniques," 2013.

[5] R. A. Nebel *et al.*, "Understanding the impact of sex and gender in Alzheimer's disease: A call to action," *Alzheimer's Dement.*, vol. 14, no. 9, pp. 1171–1183, 2018, doi: 10.1016/j.jalz.2018.04.008.

[6] I. S. F. Dessai, "Intelligent Heart Disease Prediction System Using Probabilistic Neural Network," pp. 38–44, 2013.

[7] Y. Cao, Z. De Hu, X. F. Liu, A. M. Deng, and C. J. Hu, "An MLP classifier for prediction

DOI: 10.5281/zenodo.3686791

of HBV-induced liver cirrhosis using routinely available clinical parameters," *Dis. Markers*, vol. 35, no. 6, pp. 653–660, 2013, doi: 10.1155/2013/127962.

[8] "Center for Machine Learning and Intelligent Systems | University of California, Irvine." [Online]. Available: https://cml.ics.uci.edu/. [Accessed: 29-Dec-2019].

[9] "International Statistical Classification of Diseases and Related Health Problems - World Health Organization - بتكGoogle." [Online]. Available: https://books.google.com.eg/books?hl=ar&lr=&id=Tw5eAtsatiUC&oi=fnd&pg=PA1&ots=o3e2k3qMlF&sig=l9GJ_GapBjCPM0libXs1EMp_jVM&redir_esc=y#v=onepage&q&f=false. [Accessed: 29-Dec-2019].

[10] R. Subhashini and M. K. Jeyakumar, "OF-KNN Technique: An Approach for Chronic Kidney Disease Prediction," *Int. J. Pure Appl. Math.*, vol. 116, no. 24, pp. 331–348, 2017.

[11] "Develop k-Nearest Neighbors in Python From Scratch." [Online]. Available: https://machinelearningmastery.com/tutorial-to-implement-k-nearest-neighbors-in-python-from-scratch/. [Accessed: 29-Dec-2019].

[12] S. H. Khan, S. H. Khan, J. Westin, and M. Dougherty, "DEGREE PROJECT Computer Engineering Programme Reg number Extent Predictive models for chronic renal disease using Decision trees, Naïve Bayes and Case-based methods," 2010.

[13] S. Agarwal, *Data mining: Data mining concepts and techniques*. 2014.

[14] Salford Systems, "Random Forests for Beginners," *Salford Syst.*, p. 71, 2014.

[15] M. Singh, P. K. Gupta, V. Tyagi, J. Flusser, T. I. Ören, and R. Kashyap, Advances in CompuSingh, M., Gupta, P. K., Tyagi, V., Flusser, J., Ören, T. I., & Kashyap, R. (n.d.). Advances in Computing and Data Sciences : Third International Conference, ICACDS 2019, Ghaziabad, India, April 12-13, 2019, Revised Selected Papers, Part. .

[16] R. Nisbet, G. Miner, and K. Yale, "Chapter 9 - Classification," Handb. Stat. Anal. Data Min. Appl., pp. 169–186, 2018, doi: 10.1016/B978-0-12-416632-5.00009-8.

[17] Z. Wang and X. Xue, "Multi-class support vector machine," in Support Vector Machines Applications, vol. 9783319023007, Springer International Publishing, 2013, pp. 23–48.

[18] R. Gholami and N. Fakhari, "Support Vector Machine: Principles, Parameters, and Applications," in Handbook of Neural Computation, Elsevier Inc., 2017, pp. 515–535.

[19] N. Ben Amor, S. Benferhat, and Z. Elouedi, "Qualitative Classification with Possibilistic Decision Trees," in Modern Information Processing, Elsevier, 2006, pp. 159–169.

[20] A. Géron, "Hands-On Machine Learning with Scikit-Learn and TensorFlow."

[21] J. I. E. Hoffman, "Logistic Regression," in Basic Biostatistics for Medical and Biomedical Practitioners, Elsevier, 2019, pp. 581–589.

DOI: 10.5281/zenodo.3686791