Computer Science and Information Technology 7(2): 48-53, 2019 http://www.hrpub.org
DOI: 10.13189/csit.2019.070203

Performance of Datamining Techniques in the Prediction of Chronic Kidney Disease

Kehinde A. Otunaiya*, Garba Muhammad

Department of Computer Science, Kebbi State University of Science and Technology, Nigeria

Abstract Data mining being an experimental science is very important especially in the health sector where we have large volumes of data. Since data mining is an experimental science, getting accurate predictions could be tasking. Getting maximum accuracy of each classifier is necessary. It is therefore important that the appropriate feature selection method should be selected. Feature selection is highly relevant in predictive analysis and should not be overlooked. It helps reduce the execution time and provide a more accurate and reliable result. Therefore, more researches on predictive analysis and how reliable these predictions are needs to be delved into. Application of data mining techniques in the health sector ensures that the right treatment is given to patients. This study was implemented using WEKA. This study is aimed at using 3 classifiers: multilayer perceptron, naive bayes and J48 decision tree in the prediction of chronic kidney disease dataset. The aim of this research is to evaluate the performance of the classifiers used based on the following metrics-accuracy, specificity, sensitivity, error rate and precision. Based on the performance metrics mentioned above, results shows that J48 decision tree gave the best result but naive bayes had the lowest execution time therefore making it the fastest classifier.

Keywords Data Mining Techniques, Chronic Kidney Disease, Naive Bayes, Multilayer Perceptron, J48 Decision Tree

1. Introduction

Chronic kidney disease is a large and growing problem in our world today. Even experts have raised alarm over rising incidences of kidney disease among Nigerians, saying, “each year, 17,000 new cases of kidney failure are diagnosed with only 2,000 having access to life saving dialysis” [1]. Chronic kidney disease includes conditions that damage your kidneys and decrease their ability to keep you healthy [2]. Diabetes, high blood pressure and other disorders might be causes of chronic kidney disease. Early detection of chronic kidney disease and treatment of the disease can most times prevent chronic kidney disease from getting worse. The progress of chronic kidney disease may eventually lead to kidney failure and this requires a kidney transplant or dialysis to sustain the patient’s life [2].

Due to the speedy increase in the size of data in various fields, there is an immense need to understand complex, large and highly informative data [3]. Data mining is the practice of automatically searching large stores of data to discover patterns and trends that go beyond simple analysis. Data mining uses sophisticated mathematical algorithms to segment the data and evaluate the probability of future events. Data mining is also known as Knowledge Discovery in Data (KDD) [4], and it is becoming vital tool in the current decades to turn raw data into useful information. The study of kidney diseases has become one of the most important researches in data mining. There are several data mining methods, which are: Classification, Clustering, Regression and Association Rule Learning [5].

The aim of this research is to analyse and compare the performance of three classifiers in the prediction of chronic kidney disease. Since data mining is an experimental science, getting accurate predictions could be tasking. Getting maximum accuracy of each classifier is necessary, it is therefore important that the appropriate feature selection method is selected. Feature selection is highly relevant in predictive analysis and should not be overlooked. It helps reduce the execution time and provide a more accurate and reliable result [6]. Another problem is how to analyse the classifiers selected, this where the confusion matrix is used. The confusion matrix will be used to analyse the accuracy, precision, classification error, sensitivity and specificity of the selected classifier. Waikato Environment for Knowledge Analysis (WEKA) [7], is used in this study and feature selection is performed to ensure that maximum accuracy is gotten and training time is reduced. The Zero Rule algorithm is the baseline algorithm applied to serve as a reference point to compare other classifiers. The classification algorithms: Naive Bayes, Multilayer Perceptron and Decision Trees are then

implemented and the results are compared based on several performance metrics. The performance metrics used in this study are accuracy, sensitivity, specificity, precision and error rate. The best performing classification algorithm is then identified using the aforementioned metrics.

The remainder of this paper is structured as follows: In Section 2 presents the related works. Data Collection and Synthesis is highlighted in section 3. Section 4 describes the tool used to perform the data mining tasks. Section 5 presents the results and analysis. Finally, section 6 rounds off the paper with the conclusion.

2. Related Works

Mandal [8] presented a paper that provides a review of different soft computing methods in diagnosis and detection of cancer, heart disease & Diabetes disorders acuteness. The survey was carried out for three different types of data of different diseases with cross validation and percentage split for testing new data sets of each. They used WEKA Software tool for J48, ANN and Bayesian Classifier for the implementation in dataset of disease classification but Rough Set theory was implemented using RSES software tools [9]. From their results Rough Set Theory gives maximum accuracy and coverage area but with maximum computational time complexity. On the other hand, Neural and Bayesian Network gave quite satisfactory results. Moreover, the obtained results also suggest that accuracy depends on the quality of normalization of data.

Kidney disease prediction was done using Support Vector Machine (SVM) and Artificial Neural Network (ANN). In their paper, Vijayarani and Dhayanand [8] predominantly focused on, prediction of four types of kidney diseases (Acute Nephritic Syndrome, Chronic Kidney disease, Acute Renal Failure and Chronic Glomerulonephritis. The aim of this paper is to compare the performance of these two algorithms on the basis of its accuracy and execution time. The simulation tool used was Matlab, developed by MathWorks. From the results, it can be concluded that the Artificial Neural Network (ANN) achieves increased classification performance and yields results that are accurate, it was ranked the best classifier when compared with Support Vector Machine (SVM) classifier algorithm.

In their research, Ramya and Radha [10] used machine learning algorithms to diagnose chronic kidney disease. The main objective of the paper is to determine the kidney function failure by applying the classification algorithm on the test result obtained from the patient medical report. The aim of the work is to reduce the diagnosis time and to improve the diagnosis accuracy using classification algorithms. The work deals with classification of different stages in chronic kidney disease according to its severity. The dataset for diagnosis of chronic kidney disease is obtained from medical reports of the patients collected from different laboratories in Coimbatore, Tamil Nadu, India. There are 1000 instances with 15 different attributes related to kidney disease like PID (patients ID), Age, Gender, Weight, Serium-albumin, Serium- sodium, Blood urea nitrogen, Serium creatinine, Serium uric acid, Sodium urine, Urine urea nitrogen, Urine creatinine, Urine uric acid, egfr and Kidney failure. The main contributing attribute to identify the chronic kidney disease is EGFR. Based on the value of the EGFR, the instances were classified as Low, Mild, Moderate, Normal and Severe. The experiment was performed on different algorithms like Back Propagation Neural Network, Radial Basis Function and Random Forest. The Experimental results showed that the Radial Basis function algorithm gives better result than the other classification algorithms and produces 85.3% accuracy.

Lakshmi et al. [11] compared the performance of three data mining techniques for predicting kidney dialysis survivability. In this study, three data mining techniques (Artificial Neural Networks, Decision tree and Logical Regression) were used to elicit knowledge about the interaction between these variables and patient survival. A performance comparison of three data mining techniques is employed for extracting knowledge in the form of classification rules. In this study, the models were evaluated based on the following accuracy measures: classification accuracy, sensitivity and specificity. The results were achieved using 10 fold cross-validation for each model, and are based on the average results obtained from the test dataset (the 10th fold) for each fold. From the results obtained, the ANN model achieved a classification accuracy of 93.852% with a sensitivity of 0.9387 and a specificity of 0.9387. However, it was concluded that the ANN preformed the best of the three models evaluated.

A new chronic kidney disease dataset with three classifiers such as radial basis function network, multilayer perceptron, and logistic regression were proposed in this paper. The obtained result of this experiment shows in terms of prediction accuracy, type I error, type II error, type I error rate, type II error rate, sensitivity, specificity, F-score. Accuracy of the three classifiers are evaluated for the new chronic kidney disease dataset from University of California, Irvine (UCI) repository. Thus, the paper Rubini and Eswaran [12], discussed the result of comparative study of classifiers in medical chronic kidney disease dataset concluding that the multilayer perceptron classifier gave good accuracy over the others.

Basma et. al [13], presented an overview on the evolution of big data in healthcare system, and applied three learning algorithms on a set of medical data. The objective of this research work is to predict kidney disease by using multiple machine learning algorithms, which are Support Vector Machine (SVM), Decision Tree (C4.5), and Bayesian Network (BN), and choose the most efficient one. The results after the implementation of classifiers on Weka showed that NB is the fastest classifier because it

spent the shortest time to build the classification model (0.03s) followed by C4.5. SVM is the slowest one; it took (0.41s) to build its model. Simulation results showed that C4.5 classifier proved its performance in predicting with best results in terms of accuracy and minimum execution time.

3. Data Collection and Synthesis

A. Data and Data Pre-processing

The data obtained for this research is purely real set data and it was obtained from Sir Yahaya Memorial Hospital in Birnin Kebbi, Kebbi State, Nigeria. The Chronic Kidney disease dataset from the UCI Machine Learning Repository was downloaded and studied for a broader knowledge of the factors attributed to the disease. The dataset includes 264 patients with 10 attributes, the attributes are as follows: age, sex, urea, creatinine, sodium, potassium, chlorine, diabetes, hypertension and class. The class is divided into two- YES (for patients with CKD) and NO (for patients without CKD). 158 instances were used for training while 106 instances were for testing.

Data cleaning, data reduction and features selections are important components of data pre-processing. The data collected was transformed to .ARFF file by defining data types. ARFF is an acronym that stands for Attribute-Relation File Format. It is an extension of the CSV file format where a header is used. This header provides metadata about the data types in the columns.

For this conversion, the data was first of all saved with an extension of .CSV from Microsoft Excel and then opened in WEKA using the "ArffViewer" under the "Tools" option to save it with .ARFF extension. This transformation has to be done in order for the data to be used in WEKA. Initially data quality was made sure by checking for missing, incomplete, inconsistent values in the database.

The data after being converted to Arff format was then split into two- the training set and the test set.

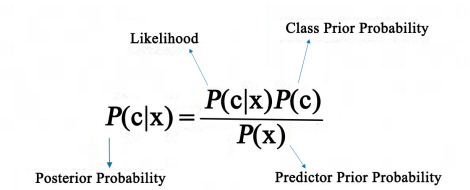
4. The Classifiers Used

The classifiers used are the Multilayer Perceptron, Naïve Bayes and J48 Decision Tree. Let's take a look at how these classifiers work.

- The Multilayer Perceptron is an artificial neural network that has more than one perceptron. It is a feed forward network that is applied on a threshold activation/transfer function. The perceptron is made up of several parts, which is the input layer (this can contain several values), the weights, the weighted sum and the activation function. Now this is how the perceptron works in the simplest way; the input values (y) are multiplied by the weight (w). The multiplied values are added together to give the weighted sum $\sum_w$ which is then applied to the correct activation function (Non-linear activation function are the most used) [14]. The multilayer perceptron consists of several layers, the input is initially fed into the first layer, the output of the first layer is then fed into the second layer and this continues until the output layer is reached. The last layer is often referred to as the output and other layers can be referred to as the hidden layers.

If $W_1.\ Y_1+W_2.Y_2+\ldots+W_n.Y_n > \Theta$ then output is 1 else output is 0

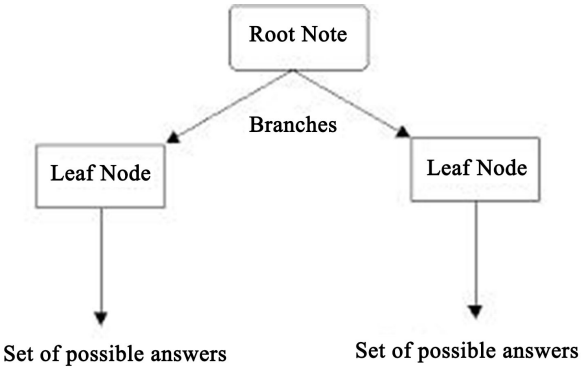
- Naïve Bayes is based on the Bayes theorem and it is useful when dealing with large dataset and it is used to get the base accuracy of the data set. It has speed when predicting the class of the test data. It is a probabilistic classifier that makes classifications using the Maximum A Posteriori decision rule in a Bayesian setting. The Bayes rule is as below:



$$P(c|X)=P(x_1|c)\times P(x_2|c)\times\ldots\times P(x_n|c)\times P(c)$$

Where,

- P (c|x) is the posterior probability of class c given predictor (features).
- P (c) is the probability of class.
- P (x|c) is the likelihood, which is the probability of predictor given class.
- P(x) is the prior probability of predictor [15].
- Using Bayes theorem, the probability of C happening can be found, given that X has occurred. Here, X is the evidence and C is the hypothesis. The assumption made here is that the predictors/features are independent. That is presence of one particular feature of a class does not affect the other. Hence it is called naïve [16].
- J48 Decision tree also known as C4.5 is an extension of the ID3 algorithm. A decision tree builds classification (or regression) models imitating a tree structure. A decision tree breaks down a data set gradually into smaller subsets while at the same time an associated decision tree is incrementally developed. Decision nodes and leaf nodes are the final result of a decision tree. A decision node has two or more branches while the leaf node represents a classification or decision. The root node is the topmost decision node in a tree, which corresponds to the best predictor.



There are several ways of building a decision tree. It's either by splitting or by pruning. Splitting is when datasets are divided into subsets. Pruning is done to address issues of overfitting (this may occur due to the presence of noise and other factors). It trims the original decision tree and improves the generalization capacity of the tree.

5. The Tool Used

The simulation tool used in this research work is Weka 3.6.13. WEKA [7] is an open source tool and has a collection of various machine learning algorithms which can be used to perform data mining tasks. In this tool, the algorithms can be applied directly to a dataset. WEKA contains tools for data pre-processing, classification, regression clustering, association rules, and visualization. The input data file format for WEKA should be in ARFF format.

A. Feature Selection

Feature selection, is a data preprocessing technique that aims to choose a small subset of the relevant features from the original features by removing irrelevant, redundant, or noisy features. Feature selection most times leads to better learning performance in terms of a lower computational cost, higher learning accuracy, and better model interpretability.

There is the unsupervised feature selection and the supervised feature selection. The supervised feature selection is mostly used for classification tasks. The supervised feature selection can effectively select discriminative features to differentiate samples from different classes because of the class labels.

For a supervised feature selection, a subset of features are selected from the features generated from the training data. The data is then processed with the selected features to the learning model. The feature selection phase uses the label information and the characteristics of the data, such as information gain to select important features. The final features selected and the label information, are used to train a classifier, which can be used for prediction [17].

Below were the steps taken to select features in WEKA:

Attributes were selected in WEKA using the "select attributes" option. The attribute evaluator used was "Info Gain Attributes Eval", this evaluates the worth of an attribute by measuring the information gain with respect to the class. The search method used was "Ranker" and this was applied using the full training set.

Feature selection is very important because it improves accuracy, reduces training time and over fitting. Selection of attribute is very essential to narrow down the list of attributes to the most influential to the prediction model. This is done because we want the best accuracy we can get.

Out of the nine attributes (excluding class) urea, creatinine, diabetes, age, hypertension and sex were significant and used for further classification while the insignificant attributes sodium, potassium and chlorine were discarded.

The classifiers applied on the dataset are Multilayer Perceptron, Naive Bayes and J48.

6. Results and Analysis

The 10-fold cross validation test was applied on each classifier to evaluate its performance. The 10-fold cross validation is a technique that divides dataset into ten subsets of equal sizes; one subset is used for testing while the others are used for training. This is done until each subset has been used for testing. 10-fold cross validation is the standard method of evaluation.

- Multilayer Perceptron: We have an accuracy of 85% where 135 instances were correctly classified and 23 instances were incorrectly classified. From the confusion matrix 13 patients were wrongly classified as not having CKD and 10 patients were wrongly classified as having CKD.
- J48: There's an accuracy of 87% where 138 instances were correctly classified and we had 20 incorrectly classified instances. Looking at the confusion matrix, 10 patients were wrongly classified as not having CKD and 10 patients were wrongly classified as having CKD.
- Naive Bayes: In this model, we have 86% accuracy, 136 correctly classified instances and 22 incorrectly classified instances. From the confusion matrix, 4 patients were wrongly classified as not having CKD and 18 patients were wrongly classified as having CKD.

Table 1. Confusion Matrix

CLASSIFIERS	NOT CKD	CKD	
Multilayer Perceptron	64	13	NOT CKD
	10	71	CKD
Naive Bayes	73	4	NOT CKD
	18	63	CKD
J48	67	10	NOT CKD
	10	71	CKD

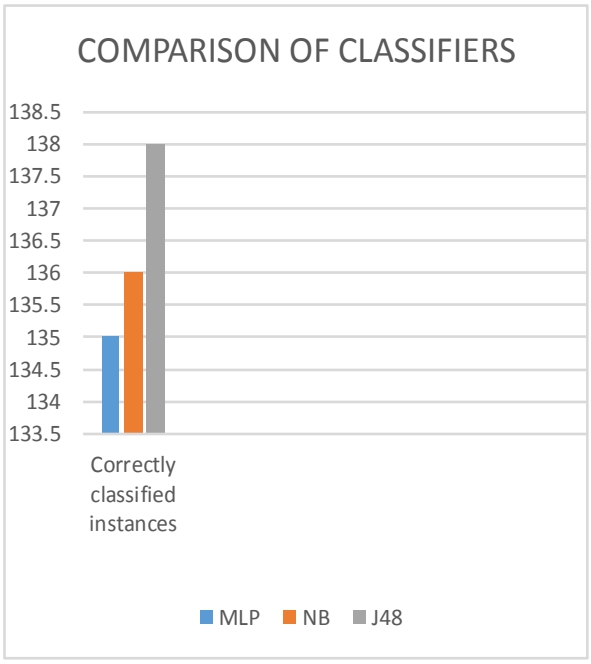


Figure 1. Correctly Classified instances of classifiers.

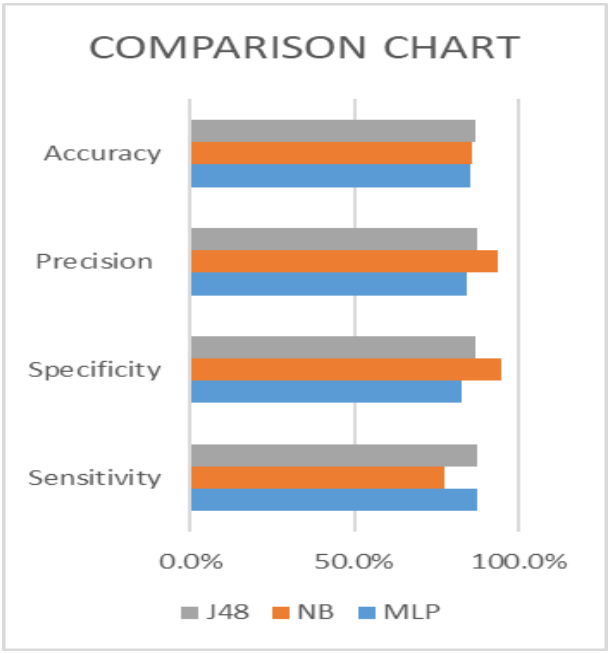


Figure 2. A chart of comparison using various metrics.

From table 2, it is observed that multilayer perceptron has the slowest execution time and naive bayes has the fastest execution time. Multilayer perceptron had the highest incorrectly classified instances (23), therefore having the lowest accuracy of 85.4% unlike J48 which had the lowest incorrectly classified instances (20) making it the best performing classifier with an accuracy of 87.3%. Multilayer perceptron has the highest error rate of 14.5% and J48 has the lowest error rate of 12.6%. The classifier with the highest specificity and precision is naive bayes although it has the worst specificity while the classifier with the lowest specificity and precision is multilayer perceptron it performed better with sensitivity likewise J48. Therefore, J48 is ranked as the best performing classifier but Naive Bayes performed better with its minimum execution time.

Table 2. Comparison Table for the Classifiers Used

PERFORMANCE METRICS	MULTILAYER PERCEPTRON	NAIVE BAYES	J48
Sensitivity	0.877	0.778	0.877
Specificity	0.831	0.948	0.87
Precision	0.845	0.94	0.877
Error rate (%)	14.5	13.9	12.6
Execution Time (secs)	X	S`	0.05
Correctly Classified Instances	135	136	138
Incorrectly classified Instances	23	22	20
Accuracy (%)	85.4	86	87.3

7. Conclusions

Using data mining techniques for predictive analysis especially in the medical field is very important. The classifiers used in this research work are J48 decision tree, Naive Bayes and Multilayer Perceptron and they were evaluated with the following performance metrics: Accuracy, Error rate, Specificity, Sensitivity and Precision. Having applied these classifiers and evaluated them, it is obvious from the results obtained that J48 decision tree had the highest accuracy compared to Naive Bayes and Multilayer Perceptron. Hence, J48 decision tree is highly recommended for similar classification problems.

REFERENCES

[1] *17,000 kidney failure cases diagnosed annually in Nigeria.* [cited 2017 12 July]; Available from: http://www.vanguardngr.com/2016/03/17000-kidney-failure-cases-diagnosed-annually-in-nigeria/.

[2] *National Kidney Foundation: About Chronic Kidney Disease*. [cited 2017 12 July]; Available from: https://www.kidney.org/atoz/content/about-chronic-kidney-disease.

[3] Sharma, N. and R. K. Verma, *Prediction of Kidney Disease by using Data Minining Techniques.* International Journal of Advanced Research in Computer Science and Software Engineering, 2016. 6(9): p. 66-70.

[4] *Oracle: Data Mining Concepts.* [cited 2017 12 July]; Available from: https://docs.oracle.com/cd/B28359_01/datamine.111/b281

29/process.htm - CHDFGCIJ.

[5] Kesavaraj, G.H. and S. S. *A study on classification techniques in data mining. in Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*. 2013 Tiruchengode, India IEEE

[6] Brownlee, J. *An Introduction to Feature Selection.* 2014 [cited 12 August 2018]; Available from: https://machinelearningmastery.com/an-introduction-to-feature-selection/.

[7] Eibe, Fracpete, and Mbatchelor. *Weka: Machine learning software to solve data mining problems* [cited 19 July 2017]; Available from: https://www.cs.waikato.ac.nz/~ml/weka/index.html.

[8] Mandal, S., G. Saha, and R.K. Pal, *A Comparative Study on Disease Classification using Different Soft Computing Techniques.* Transactions on Computer Science Engineering & its Applications (CSEA), 2014. 2(3): p. 45-52.

[9] Jan, G.B. and S.S. Marcin. *RSES and RSESlib- A Collection of Tools for Rough Set Computations*. 2014 [cited 2018 19 December]; Available from: https://www.semanticscholar.org/paper/RSES-and-RSESlib-A-Collection-of-Tools-for-Rough-Bazan--Szczuka.

[10] Ramya, S. and N. Radha, *Diagnosis of Chronic Kidney Disease Using Machine Learning Algorithms*. International Journal of Innovative Research in Computer and Communication Engineering, 2016. 4(1): p. 812-820.

[11] Lakshmi, K.R., Y. Nagesh, and V. M., *Performance Comparison of Three Data Mining Techniques for Predicting Kidney Dialysis Survivability*. International Journal of Advances in Engineering & Technology, 2014. 7(1): p. 242-254.

[12] Rubini, L.J. and P. Eswaran, Generating comparative analysis of early stage prediction of Chronic Kidney Disease. International Journal of Modern Engineering Research (JMER), 2015. 5(7): p. 49-55.

[13] Boukenz, B., H. Mousannif, and A. Haqiq, *Performance of Data Mining Techniques to Predict In Healthcare Case Study: Chronic Kidney Failure Disease*. International Journal of Database Management Systems (IJDMS), 2016. 8(3): p. 1-9.

[14] Sagar, S. *The Fundamentals of Neural Networks.* 2017 [cited 22 March 2018]; Availablefrom:https://towardsdatascience.com/what-the-hell-is-perceptron-626217814f53

[15] Gaurav, C. *All about Naive Bayes.* 2018 [cited 22 March 2018]; Availablefrom:https://towardsdatascience.com/all-about-naive-bayes-8e13cef044cf

[16] Rohith, G. *Naive Bayes Classifier.* 2018 [cited 22 March 2018]; Available from: https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c

[17] Suhang, W, Jiliang, T and Huan, L. *Feature Selection*. Research Gate, 2016. P.2-9.