University of Ibadan
Journal of Science and Logics in ICT Research

Assessment of Selected Data Mining Classification Algorithms for Analysis and Prediction of Certain Diseases

[1]Oguntunde, B. O., [2]Arekete, S. A., [3]Odim, M. O. and [4]Ariyo-Agbaje, G. Y.

Redeemer's University, Ede, Osun State, Nigeria
[1]oguntunden@run.edu.ng [2]areketes@run.edu.ng [3]odimm@run.edu.ng [4]ariyo-agbaje6152@run.edu.ng

Abstract

Medical science generates large volumes of data stored in medical repositories that could be useful for extraction of vital hidden information essential for diseases diagnosis and prognosis. In recent times, the application of data mining to knowledge discovery has shown impressive results in disease analysis and prediction. This study investigates the performance of three data mining classification algorithms, namely decision tree, Naïve Bayes, and k-nearest neighbour in predicting the likelihood of the occurrence of chronic kidney disease, breast cancer, diabetes, and hypothyroid. The datasets which were obtained from the UCI Machine were split into 60% for training and 40% for testing on the one hand and 70% for training and 30% for testing on the other hand. The performance parameters considered include classification accuracy, error rate, execution time, confusion matrix, and area under the curve. Waikato Environment for Knowledge Analysis (WEKA) was used to implement the algorithms. The findings from the analysis showed that decision tree recorded the highest prediction accuracy followed by the Naïve Bayes and k-NN algorithm while k-NN recorded the minimum execution time on the four datasets. However, k-NN also has the largest average percentage error recorded on the datasets. The findings, therefore, suggest that the performance of these classification algorithms could be influenced by the type and size of datasets.

Keywords: *Classification algorithm, Data mining, Decision tree, Naïve Bayes, k-nearest neighbour*

1. INTRODUCTION

Classification is one of the techniques in data mining to allocate objects to one of several predefined groups [1]. Data mining extracts interesting, non-trivial, implicit, previously unknown and potentially useful patterns or knowledge with the help of various techniques in the data gathered from the various sources. Data mining also involves selecting relevant data from the database, pre-processing and cleaning the relevant data, as well as transforming into a suitable form, mining and evaluating the data and afterwards online updating and visualisation. The actual task of data mining is a semi-self-regulating or mechanical investigation of large batches of the dataset for extracting the previously unknown, unusual records and dependencies. The knowledge discovery process involves various selection steps which help in the efficient extraction of the useful data from databases. Furthermore, data mining is one of the essential steps in the KDD process [2].

1.1 Techniques in Data Mining

Several data mining techniques and methods which have been developed and used in data mining research include association, classification, clustering, prediction, and sequential patterns [3]. The focus of this work is the classification technique.

1.1.1 Classification

Classification is one of the fundamental techniques in data mining. Classification techniques are useful to handle a large amount of data; it is used to predict categorical class labels. This model is used to classify newly available data into a class label. Classification is also the process of finding a model that

Oguntunde B. O., Arekete S. A., Odim M. O. and Ariyo-Agbaje G. Y. (2020). Assessment of Selected Data Mining Classification Algorithms for analysis and prediction of certain Diseases. *University of Ibadan Journal of Science and Logics in ICT Research (UIJSLICTR),* Vol. 4 No. 1, pp. 44 - 51

©U IJSLICTR Vol. 4, No. 1, March 2020

describes and distinguishes data classes or concept [4]. Data classification is a two-step method consisting of knowledge step used to make a classification model and a categorisation step used to calculate the class labels for a given data. It serves as descriptive modelling to distinguish between objects of unlike classes. A classification model can also help in predictive modelling to calculate the class label of unidentified records. This process is mainly fitting for describing data sets with dual or diminutive types. It is a systematic approach to construct a classification model from the input data set. It includes Function, Bayesian, Meta-learning, Lazy, Rule-Based, Decision Tree, and Miscellaneous classifiers. Each method utilises a learning algorithm to recognise a model that best fits the liaison between the attribute set and class label of the input data [4].

An essential point of the learning algorithm is to construct the representation with generalisation facility, i.e., the description precisely forecasts the class labels of formerly unidentified instances [3]. Classification techniques like Decision Tree, K-Nearest Neighbour, Support Vector Machines, Naïve Bayesian Classifier, and Neural Networks are considered in this work.

1.1.2 Classification Methods

Three classification techniques are studied in this work, namely: decision tree, Naïve Bayes and *k*-Nearest Neighbour.

a) Decision Trees

Decision tree builds classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The outcome is a tree with decision nodes and leaf nodes. A decision node (e.g., Outlook) has two or more branches (e.g., sunny, overcast, and rainy). Leaf node (e.g., Play) represents a classification or decision. The first decision node in a tree which corresponds to the best predictor is called the root node. Decision trees can handle both categorical and numerical data.

b) Naïve Bayes Classifier

The Naïve Bayes classifier is a probabilistic machine learning model used for classification task based on the Bayes theorem with the independence assumptions between predictors. The Naïve Bayes classifier is built on the motive that the role of a natural class is to estimate the values of features for members of that class. Examples are grouped in categories because they have common values for the features. These classes are often called natural classes [5].

Naïve Bayesian model is easy to build, with no complicated iterative parameter estimation, which makes it particularly useful for massive datasets. Despite its simplicity, the Naïve Bayesian classifier often does surprisingly well and is widely used because it regularly outperforms other classification techniques. The Bayes theorem is expressed in equation (1).

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \qquad (1)$$

Using Bayes theorem, the probability of **A** happening, given that **B** has occurred is expressed. Here, **B** is the evidence, and **A** is the hypothesis. The deduction made here is that the predictors/features are independent. That is, the presence of one particular feature does not affect the other. Hence it is called naïve [6].

c) k-Nearest Neighbour

The *k*-Nearest Neighbour (*k*-NN) technique of classification is one of the most straightforward methods in machine learning used for finding the most similar data points in the training data and making predictions based on their classifications. It is used in recommendation systems, anomaly detection, and semantic searching [7]. The *k*-NN falls under lazy learning, i.e., there is no explicit training phase before classification. Instead, any effort to generalise or abstract the data is completed upon classification. *k*-NN tends to work best on lesser data-sets that do not have many features.

2. Related Works

Pooja and Nasib [8] compared the performances of five classification algorithms, namely: *k*-NN, Neural Networks, Support Vector Machine (SVM), Decision Tree and Bayesian classification on three medical

datasets: heart-statlog, diabetes and hepatitis. Their results showed that SVM offers the most robust method of classification and *k*-NN the least [8]. The work of Nurul and Ahsan analysed the performances of J48 Decision Tree, Neural Network Multilayer Perceptron and Naïve Bayes on heamatological data; their results show that J48 decision tree classifier offers the highest accuracy while Naïve Bayes has the lowest average error rate [9]. Sharma *et al*. [10] performed a comparison of M5P decision tree, K-star Nearest Neighbour, Rule-based Classifier (M5Rule) and Neural Network Multilayer Perceptron on rainfall statistics, admission dataset, tourism dataset, and population dataset. Results obtained show that K-star Nearest Neighbour has the highest accuracy for large datasets, and for small datasets, the performances of all the techniques were comparatively the same [10].

Akter *et. al*., [11] classified hematological data using data mining techniques to predict diseases. The analysis was carried out using random forest tree, neural network and Bayesian network on hematological data. Random forest tree was found to be most efficient, having the highest accuracy and lowest execution time while the neural network has the lowest accuracy. Sakshi *et. al*., [12] applied classification algorithms, namely random forest tree, Naïve Bayes, multilayer perceptron and J48 decision tree on chronic kidney disease dataset. The results obtained shows that multilayer perceptron was found to be more accurate in their studies.

In Oguntunde and Arekete [13], a comparison of Naïve Bayes and k-Nearest Neighbour was made on liver disease and fertility datasets using KNIME. The results showed that k-NN outperformed the Naïve Bayes algorithm in terms of a higher level of interpretability and greater classification accuracy.

The emergence of many new healthcare devices and applications on daily basis, which were however, limited to certain categories of illness had been observed in Ekpo *et. al.*, [14]. The authors stressed the need for more research to evolve techniques for early detection of diseases. In their study, they particularly, explored the significance and available IoT technologies in the e-Health domain.

3. Methodology

Three data mining classifications algorithms, namely: Decision tree, Naïve Bayes and *k*-Nearest Neighbour were employed in the analysis and prediction of the chances of the occurrence of chronic kidney disease, breast cancer, diabetes, and hypothyroid.

3.1 Datasets And Attributes

The datasets were obtained from the UCI Machine Repository. Two different percentage splits of 60% training 40% testing and 70% training 30% testing were examined. The characteristics of the four datasets are presented in Table 1.

Table 1: Datasets characteristics

Features	Dataset			
	Chronic-Kidney disease (CKD)	**Breast Cancer**	**Diabetes**	**Hypothyroid**
DataSet Characteristics	Multivariate	Multivariate	Multivariate	Multivariate
Attribute Characteristics	Real	Categorical	Categorical, Integer	Categorical, Integer
Number of Instances	400 (200 CKD, 150 NOTCKD)	286	768	3772
Number of Attributes	25	10	9	30
Associated Tasks	Classification	Classification	Classification	Classification
Missing Values (?)	Yes (9)	Nil	Nil	Yes (1)

3.2 Evaluation Parameters

The classification techniques selected are evaluated based on the five parameters, namely:

a. Classification Accuracy
b. Execution Time (Speed)
c. Error Rate
d. Confusion Matrix
e. Area Under Curve

Classification Accuracy

Classification accuracy is the ratio of several correct predictions to the number of input samples. The algorithm can correctly predict the class label of new or previously unseen data.

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions\ Made} \quad (2)$$

Execution Time

Execution time is the time taken by the WEKA tool to classify the dataset using a classification algorithm. The mechanism used to measure execution time is implementation-defined. Execution time pertains to the computational cost involved in generating and using the algorithm.

Error Rate

The Error rate is measured in terms of the Mean Absolute Error and Mean Squared Error.

Mean Absolute Error

Mean Absolute Error (MAE) is the average of the difference between the original values and the predicted values. This statistic gives the measure of how far the predictions were from the actual output. However, Mean Absolute Error does not give any idea of the direction of the error, i.e., whether the data are under-predicted or over-predicted.

Mathematically, Mean Absolute Error is given by equation (3)

$$Mean\ Absolute\ Error = \frac{1}{N}\sum_{j=1}^{N}|y_j - \hat{y}_j| \quad (3)$$

Mean Squared Error

Mean Squared Error (MSE) is quite similar to Mean Absolute Error, the only difference being that MSE takes the average of the square of the difference between the original values and the predicted values. With MSE, it is easier to compute the gradient, whereas Mean Absolute Error requires complicated linear programming tools to compute the gradient. As we take a square of the error, the effect of larger errors becomes more pronounced than smaller error. Hence the model can now focus more on the larger errors [15].

$$Mean\ Squared\ Error = \frac{1}{N}\sum_{j=1}^{N}|y_j - \hat{y}_j|^2 \quad (4)$$

Confusion Matrix

Confusion Matrix gives us a matrix as output and describes the complete performance of the technique. There are four important terms:

True Positives: The cases in which the prediction is YES and the actual output is also YES.

True Negatives: The cases in which the prediction is NO and the actual output is NO.

False Positives: The cases in which the prediction is YES, and the actual output is NO.

False Negatives: The cases in which the prediction is NO and the actual output is YES.

Accuracy for the matrix is calculated by taking an average of the values lying across the main diagonal, i.e.

$$Accuracy = \frac{True\ Positives + False\ Negatives}{Total\ Number\ of\ Samples} \quad (5)$$

Area Under Curve (AUC)

Area Under Curve (AUC) is one of the most widely used metrics for evaluation. AUC of a classifier is the probability that the classifier will rank a randomly chosen positive model higher than a randomly chosen negative model. There are two terms in AUC, which are:

True Positive Rate (Sensitivity): True Positive Rate corresponds to the proportion of positive data points that are correctly taken as positive, concerning all positive data points. It is defined as:

$$T_{PR} = \frac{T_P}{F_N + T_P} \quad (6)$$

Where T_{PR} means True Positive Rate, T_P means True Positive and F_N stands for False Negative.

False Positive Rate (Specificity): False Positive Rate corresponds to the proportion of negative data points that are taken as positive, concerning all negative data points.

$$F_{PR} = \frac{F_P}{F_P + T_N} \qquad (7)$$

whereF_{PR} stands for false positive rate, F_P is false positive and T_N denotes true negative.

4. Implementation

WEKA, an open source Java software developed by University of Waikato, New Zealand, that has a group of machine learning algorithms for data mining and data exploration tasks was used for the analysis. The datasets were loaded into WEKA, and series of operations using WEKA's preprocessing filters were performed. Some of the output are presented in figures 1 to 3. Figure 1 shows the output of the training experiment of the Naïve Bayes classifier on the full training set on Diabetes using 70% 30% split. The result showed a correctly classified instances of 95.23% and incorrectly classified instances of 4. 77%. Figure 2 depicts the visualisation of the decision tree of the diabetes sets.

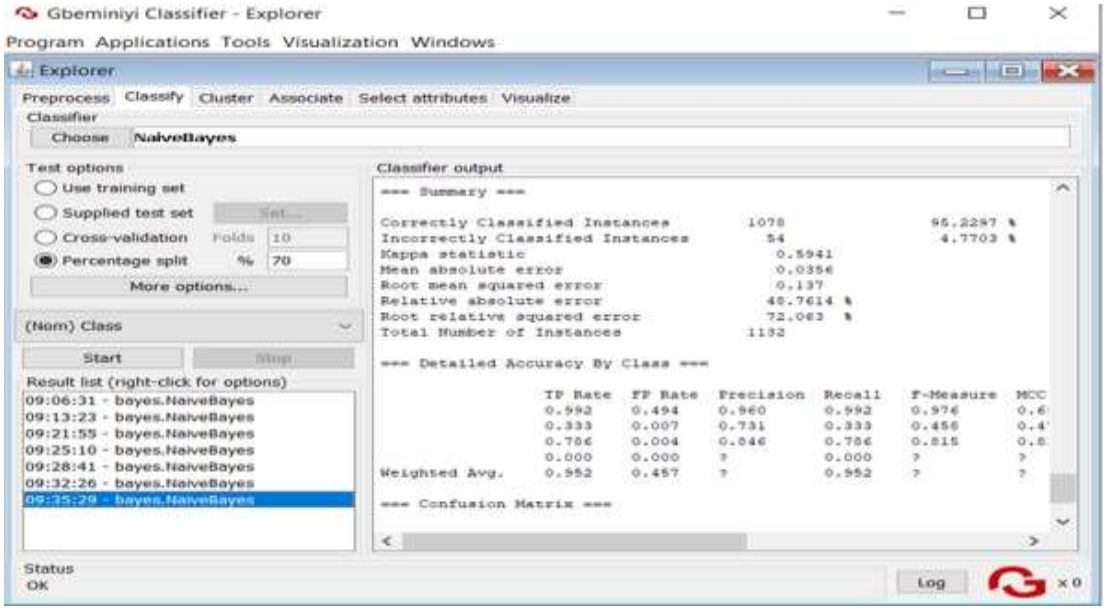

Figure 1: Naïve Bayes classifier on the full training set on Diabetes using 70% 30% split

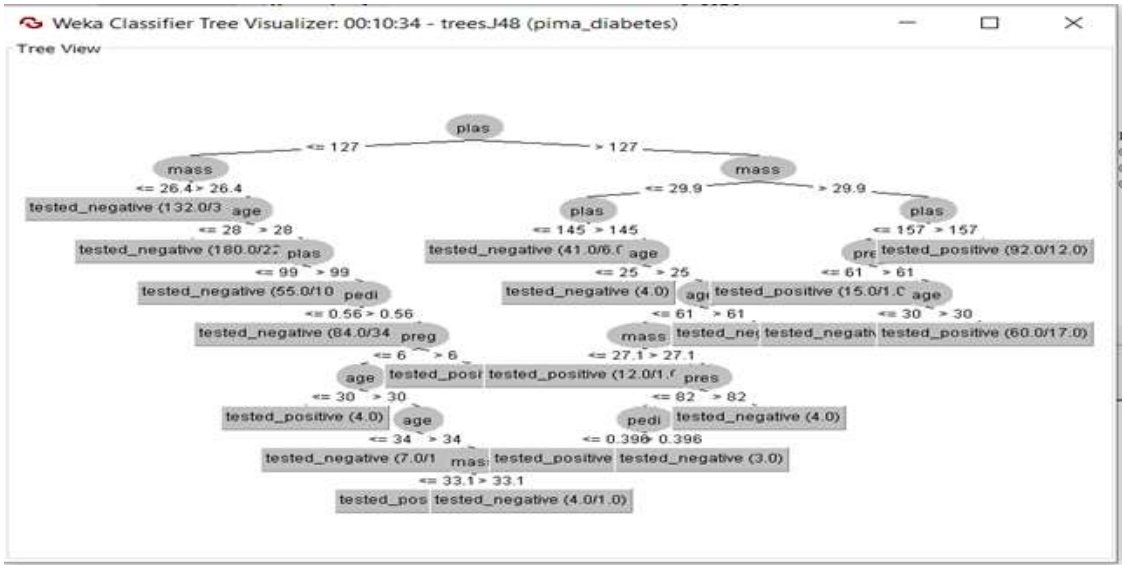

Fig 2: Decision tree of diabetes dataset

The Naïve Bayes classifier on the full training set of chronic kidney disease using 60% 40% is shown in Figure 3. 95.625% and 4.375 instances were correctly and incorrectly classified.

4. Results and Discussion

4.1 Results

The results obtained from the analysis of each data split are presented in Table 2.

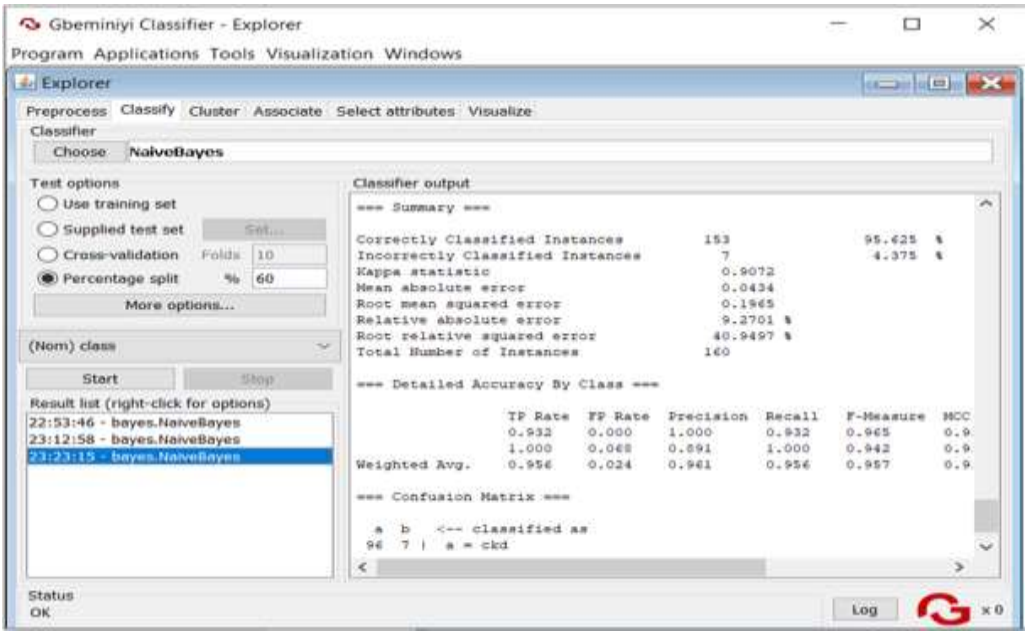

Figure 3: Naïve Bayes classifier training set of chronic kidney disease using 60% 40%

Table 2: Results of the algorithms on the four datasets

	Decision Tree Algorithm		Naive Bayesian Algorithm		K – Nearest Neighbour Algorithm	
Chronic Kidney Disease Dataset (400/25)	**70% 30%**	**60% 40%**	**70% 30%**	**60% 40%**	**70% 30%**	**60% 40%**
Classification Accuracy (%)	97.5	97.5	95.833	95.625	91.667	93.75
Execution Time (seconds)	0.06	0.12	0.01	0.04	0.01	0.01
Error Rate (%)	2.5	2.5	4.1667	4.375	8.333	6.25
Area Under Curve	0.992	0.979	1.000	0.997	0.981	0.988
Breast-Cancer Dataset (286/10)	**70% 30%**	**60% 40%**	**70% 30%**	**60% 40%**	**70% 30%**	**60% 40%**
Classification Accuracy (%)	63.9535	70.1754	67.4419	71.9298	69.7674	75.4386
Execution Time (seconds)	0.03	0.11	0.09	0.02	0.02	0.01
Error Rate (%)	36.0465	29.8246	32.5581	28.0702	30.2326	24.5614
Area Under Curve	0.573	0.572	0.651	0.660	0.650	0.655
Diabetes Dataset (768/9)	**70% 30%**	**60% 40%**	**70% 30%**	**60% 40%**	**70% 30%**	**60% 40%**
Classification Accuracy (%)	76.5217	73.6156	76.9565	75.8956	70.8696	71.0098
Execution Time (seconds)	0.02	0.01	0.02	0.02	0.01	0.01
Error Rate (%)	23.4783	26.3844	23.0435	24.1042	29.1304	28.9902
Area Under Curve	0.743	0.777	0.845	0.834	0.717	0.740

Hypothyroid Dataset (3772/30)	**70% 30%**	**60% 40%**	**70% 30%**	**60% 40%**	**70% 30%**	**60% 40%**
Classification Accuracy (%)	99.2933	99.271	94.9647	95.2286	94.9647	94.3671
Execution Time (seconds)	0.05	0.05	0.02	0.02	0.02	0.01
Error Rate (%)	0.7067	0.729	5.0353	4.7714	5.0353	5.6329
Area Under Curve	0.986	0.991	0.933	0.940	0.974	0.974

4.2 Discussion

The accuracy of the *k*-NN algorithm remained virtually the same. The change is insignificant compared to the other two algorithms. The execution time of the Naïve Bayes algorithm is faster on 70% 30%. *k*-NN execution time remains the same for both data splits. With regards to the error rate, 60% 40% data split has more percentage of recorded errors on decision tree and Naïve Bayesian algorithm than on 70% 30% data split. For the *k*-NN algorithm, the error rate is the same on both data split.

Furthermore, for Chronic Kidney Disease dataset, decision tree accuracy is the same on both data splits (97.5%). The Naïve Bayes algorithm is higher on 70% 30% split (95.83%) than on 60% 40% split in terms of classification accuracy. The accuracy of the *k*-NN algorithm is higher on 60% 40% data split than 70% 30. Moreover, the execution time of decision tree and the Naïve Bayes algorithm is faster on 70% 30% than on 60% 40% data split. *k*-NN execution time remains the same for both split.

Concerning breast cancer dataset, decision tree accuracy was higher on 60% 40% split (70.17%). The Naïve Bayes algorithm is higher on 60% 40% split (71.929%) than on 70% 30% split in terms of classification accuracy. The accuracy of the *k*-NN algorithm is higher on 60% 40% data split than 70% 30. The execution time of Naïve Bayes algorithm and kNN is faster on 60% 40% than on 70% 30% data split. Decision tree execution time is faster on 70% 30% split. With regards to the error rate, 70% 30% data split has more percentage of recorded errors on the three algorithms than on 60% 40% data split.

Moreover, on diabetes dataset, decision tree accuracy and Naïve Bayes algorithms are higher on 70% 30% split (76.52% and 76.95% respectively) as against 73.6165 and 75.8956 on 60%40% split. The accuracy of the *k*-NN algorithm is higher on 60% 40% data split than 70% 30%. The execution time of the Naïve Bayes algorithm and *k*-NN algorithm is the same on both data split. Decision tree execution time is faster on 60% 40% split. With regards to the error rate, 60% 40% data split has more percentage of recorded errors on Naïve Bayes and decision tree while *k*-NN has more percentage of recorded errors on 70% 30% data split.

Nevertheless, for the hypothyroid dataset, decision tree and *k*-NN algorithms accuracies are higher on 70% 30% split (99.29% and 94.68% respectively). The accuracy of the Naïve Bayes algorithm is higher on 60% 40% data split than 70% 30% data split. Also, the execution time of the Naïve Bayes algorithm and decision tree is the same on both data split. *k*-NN execution time is faster on 60% 40% split. With regards to the error rate, 60% 40% data split has more percentage of recorded errors on decision tree and *k*-NN while Naïve Bayes algorithm has more percentage of error on 60% 40% data split.

In a nutshell, the accuracy of decision tree on the average is higher for the four datasets, *k*-NN has more error rate recorded than the other two algorithms and the execution time of the three algorithms varies across the five datasets. The results obtained in this study agrees with that of Nurul & Ahsan [9] in the sense that decision tree has higher accuracy for larger dataset, but others are not and for small dataset performance of the algorithm are comparatively same. Therefore, no particular algorithm is best suited for every situation: the performance of classification algorithms depends on the type and size of datasets. In other words, one algorithm may be more appropriate for one dataset while another algorithm may be more appropriate for another dataset.

5. CONCLUSION

The study examined the performance of three data mining classification algorithms: decision tree, Naïve Bayes and *k*-Nearest Neighbour in analysing and predicting the chances of the occurrence of *chronic kidney, breast cancer, diabetes, and hypothyroid* medical related diseases. The datasets were obtained from the

UCI Machine Learning Repository. Each dataset was split into two, namely: 70% training 30% testing and 60% training 40% testing. The analysis was implemented in the Waikato Environment for Knowledge Analysis v 3.8.3. The three algorithms were compared based on classification accuracy, execution time, error rate, confusion matrix and area under the curve. The decision tree algorithm recorded the most accurate prediction on 70% 30% data split. *k*-NN had the minimum execution time on the datasets. The error rate varied across the two data split. *k*-NN has the largest percentage error on chronic kidney disease dataset, hypothyroid dataset, and diabetes dataset while decision tree has the largest error rate on breast cancer dataset.

Furthermore, the area under the curve varies across the dataset with little or no significant differences. Therefore, no particular algorithm is best suited for all situations, the performance of classification algorithms depends on the type and size of datasets, i.e., one algorithm is more appropriate for one dataset while another algorithm is better on another algorithm.

The datasets used in this project are medium scale datasets; a larger dataset of over 1000 instances is recommended for future works. Furthermore, measurement of other parameters such as interpretability, robustness can be considered.

References

[1] Hency, J. A., & Padmajavalli, R. (2016). Comparative Analysis of Classification Algorithms on Endometrial Cancer Data. *Indian Journal of Science and Technology, 9*(28), 205-211. doi: 10.17485/ijst/2016/v9i28/93846

[2] Saini, S., & Dhankkar, A. (2018). Comparative Analysis of Classification Algorithms Using Weka. *IOSR Journal of Engineering, 8*(10), 29-40.

[3] Diwate, R., & Sahu, A. (2014). Data Mining Techniques in Association Rule: A Review *International Journal Of Computer Science and Information Technologies, 5*(1), 227-229.

[4] Gorade, S. M., Deo, A., & Purohit, P. (2017). Early Identification of Diseases Based on Responsible Attribute Using Data Mining. *International Research Journal of Engineering and Technology (IRJET), 4*(7), 1623-1926.

[5] Poole, D., & Mackworth, A. (2010). Foundations of Computational Agents. Retrieved from Artificial Intelligence. from https://artint.info/html/ArtInt_181.html

[6] Gandhi, R. (2018). Naive Bayes Classifier. *Towards Data Science.* Retrieved May 5, 2019, 2019, from https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c

[7] Soni, D. (2018) Introduction to K-Nearest Neighbours. *Towards Data Science.* Retrieved March 12 2019, 2018, from https://towardsdatascience.com/introduction-to-k-nearest-neighbors-3b534bb11d26

[8] Pooja, M., & Nasib, S. G. (2014). A comparative analysis of classification techniques on medical datasets. *International Journal of Research in Engineering and Technology, 3(6), pp 454 -460., 3*(6), 454-460.

[9] Nurul, A., & Ahsan, H. (2015). Comparison of different classification techniques using WEKA for Haematological Data. *American Journal of Engineering Research (AJER), 4*(3), 55-61.

[10] Sharma, R., Shiv, K., & Rohit, M. (2015). Comparative analysis of classification techniques in data mining using different datasets. *International Journal of Computer Science and Mobile Computing (IJCSMC), 4*(12), 125-134.

[11] Akter, F., Hossin, M. A., Daiyan, G. M., & Hossain, M. M. (2018). Classification of hematological data using data mining technique to predict diseases. *Journal of Computer and Communications, 6*(1), 76-83.

[12] Sakshi, S., Amita, D., & Kamna, S. (2018). Comparative analysis of Classification algorithms using WEKA. *IOST Journal of Engineering (IOSRJEN), 8*(10), 29-40.

[13] Oguntunde, B. O. & Arekete, S. A. (2019). *Naïve Bayes and K-nearest neighbour Data Mining Algorithms: A comparative analysis.* Paper presented at the Pan African Conference on Science, Computing and Telecommunications (PACT) 2019, University of Eswatini Matsapha, Kwaluseni Eswatini.

[14] Ekpo, R. H., Osamor, I. P., Osamor, V. C., Abiodun, T. A., Omoremi, A. O., Odim, M. O. & Oluwasegun Oladipo, O. (2019). Understanding E-health Application Utilizing Internet of Things (IoT) Technologies 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), San Diego, CA, USA, 1465-1469.

[15] Aditya, M. (2018, February 24). *Metrics to Evaluate your Machine Learning Algorithm.* Retrieved from Towards Data Science: https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234