Information Technology and Management Science
ISSN 2255-9094 (online)
2022, vol. 25, pp. 1–7
https://doi.org/10.7250/itms-2022-0001
https://itms-journals.rtu.lv

1862
RIGA TECHNICAL UNIVERSITY

Comparative Study of Chronic Kidney Disease Predictor Performance Given Insufficient Training Dataset

Oluwadamilare Alabi*
Sheda Science & Technology Complex, Federal Capital Territory, Nigeria

***Abstract* – This study compares the performance of Logistic Regression and Classification and Regression Tree model implementations in predicting chronic kidney disease outcomes from predictor variables, given insufficient training data. Imputation of missing data was performed using a technique based on k-nearest neighbours. The dataset was arbitrarily split into 10 % training set and 90 % test set to simulate a dearth of training data. Accuracy was mainly considered for the quantitative performance assessment together with ROC curves, area under the ROC curve values and confusion matrix pairs. Validation of the results was done using a shuffled 5-fold cross-validation procedure. Logistic regression produced an average accuracy of about 99 % compared to about 97 % the decision tree produced.**

***Keywords* – Binary classification, decision tree, logistic regression, machine learning.**

I. Introduction

The use of data to develop decision-making procedures has a long history with diverse approaches aimed at predicting outcomes of medical conditions. One of them being the prompt diagnosis and management of Chronic Kidney Diseases (CKD). A Global Burden of Disease report revealed that the global all-age death rate attributed to CKD increased significantly by 41.5 % between 1990 and 2017 [1], having previously being ranked 19th on the global cause of death index in 2013 [2]. Reducing the mortality rate stemming from kidney disease progression begins with early and accurate diagnosis.

The undesirable outcomes encountered during the treatment of kidney diseases have resulted in the need for improving the detection of such through various methods, which may include engaging the use of selected machine learning approaches. While there are a few machine learning prediction techniques that can provide insight into making informed decisions regarding diagnosis and treatment of diseases, it is often difficult to determine which would be best suitable for the situation at hand. In this particular case, the study examines which is a more suitable algorithm between logistic regression (LR) and classification and regression tree (CART) for predicting the outcome of chronic kidney diseases, given that the amount of available training data has a lesser ratio compared to the test data.

Classification algorithms from the machine-learning knowledgebase are capable of making these predictions by learning non-trivial relationships from a collection of training data, thereby producing a generalization of these relationships that can be used to interpret new separate test data [3]. A typical classification algorithm can examine either the linear or non-linear relationship between a binary outcome and categorical or continuous predictor variables depending on whether the dataset is linearly separable or not. The accuracy of a linear machine on linearly-separable data instances will be unpredictable [4]. In essence, the choice of a suitable algorithm for prediction of data outcomes can be viewed as fundamentally dependent on the kind of dataset in use, whether it be linearly separable or not. This is the basis on which the choice of the classification algorithms used in this study is predicated.

Being a simplistic algorithm, logistic regression is oftentimes the first point of call for classification tasks, less susceptible to overfitting and quick [5]. It is also known to be a linear classifier [6]. A decision tree algorithm, on the other hand, is a tree-like model of decisions made based on the outcomes of tests performed on a set of variables, and can be used on not linearly-separable data. Each branch represents the outcome of the test made and each leaf node obtained represents the final decision taken. A variant of the tree algorithm known as classification and regression tree works by using the Gini index to split the training set into subsets using a feature-threshold value pair, selecting the subset with more decrease in impurity [7], [8]. The two models described above have been employed over the years in the classification of a variety of datasets. Given that both models are being utilized for similar purposes, it is imperative to perform an investigation in order to gain insight into a comparative analysis of logistic regression and decision-tree models, their strengths and their respective weaknesses.

This study will be conducted with the intent to further increase the contributions from computational science to medicine. The findings from this study will provide insight into which algorithm will be a more preferable choice regarding the prediction of chronic kidney disease outcomes especially in developing countries where scarcity of training data abound.

* Corresponding author. E-mail: oi.alabi@shestco.gov.ng

II. Related Work

The comparison of logistic regression and decision tree learning models for binary case classification has been explored in numerous medical domains, for instance, the diagnosis of acute cardiac ischemia [9] in which logistic regression performed better than decision tree using a 60:40 split. In another study, a duplicated chronic kidney disease dataset split in a 70:30 ratio revealed that the decision forest possessed the highest accuracy [10]. A predictive performance comparison of few learning approaches on chronic kidney disease progression was done using an 80:20 proportion of the dataset [11]. As observed above, the amounts of training data have usually been higher than half of the total data sizes. Curiosity arises regarding the prediction accuracy of machine learning algorithms generally whenever the amount of training data is much lower than the amount of test data.

III. Materials and Methods

This section describes the full procedure for this study, which was completed using the Python 3.7.7 library together with machine learning modules in a Spyder 4.1.3 scientific python development environment.

A. Data and Dataset Pre-processing

The dataset used in this study was the UCI machine learning repository's Chronic Kidney Disease dataset [12], where the rows represent observations recorded and the columns represent variables. The original dataset contained a total of 400 rows of observations with one row containing invalid data discarded, leaving a total of 399 valid observations. The dataset contains a range of predictor variables namely 'Sex', 'Age', 'sod', 'Pot', 'Chl', 'Bica', 'Urea', 'Cre', 'UA', 'Alb', 'Classification', each making up a column containing corresponding values for each of the 399 observations. 'Sex' and 'Classification' are nominal variables containing categorical data values. All the variables except 'Classification' were used to determine if a patient had chronic kidney disease or not. Total CKD observations were 199 and 200 were not CKD. The dataset acquired from the repository was observed as having contained eight missing values in the 'age' column which accounted for 2 % of the total number of the column observations, while the 'Sex' and 'Cre' column had only one missing value, about 0.25 % each. The columns 'UA' and 'Alb' contained 355 and 382 missing observations which accounted for 88.9 % and 95.7 %, respectively, and they had to be discarded as they were deemed to be insignificant in predicting the outcome of the disease in patients. The requirement for the learning algorithms to accept the input data is that the dataset has to be totally numerical. Hence, there is the need for the conversion of the alphabetical data values in the columns 'Sex' and 'Classification' to numerical values. To achieve it, the values 'F' and 'KD' were mapped to 1, while values 'M' and 'NKD' were mapped to 0.

Datasets are usually neither complete nor homogeneous; hence, data pre-processing is needed. Missing data may occur for a number of reasons, such as malfunctioning measurement equipment, values that have not passed quality control criteria, human error during data capture, changes in experimental design during data collection and collation of several similar but non-identical datasets [13]. The undesirable implication is that the presence of some missing observations reduces the representativeness of the samples involved, which can result in the distortion of inferences [14]. Also, it is possible for the lost data to generate bias during the estimation of parameters [15]. Therefore, it is exigent to devise means of filling the missing entries in the dataset before use. This can be performed on the dataset by considering a method for imputing the missing data. Imputation methods are beneficial because they have the ability to increase predictive accuracy in the presence of missing-data perturbation [16]. The approach chosen for estimating the values of the missing data in this study incorporates a technique using the k-nearest neighbour predictive algorithm [17], in which the features of N nearest neighbours of the missing feature entry are fetched and either averaged uniformly or weighted by distance to each neighbour to obtain a value. The weighted-distance value was used to fill-in the missing feature entry. The data preparation was performed using Pandas [18].

B. Research Design

The resulting CKD dataset was split into two subsets, namely, the training set and the test set. These subsets were populated with randomly selected observations from the parent set. The training subset was filled randomly with 10 % of the parent set, while the test subset comprised 90 % of the data contained in its parent dataset. This means that the newly created partitions of dataset contained 39 observations in the training set and 360 observations in the test set. Random generator seeds 5, 8 and 10 were incorporated during the random sampling to ensure reproducibility of the procedure. This procedure describes the sampling phase.

To perform the classification tasks in this study, the logistic regression [19] and the classification and regression tree [20] supervised learning models were implemented using the Scikit library [21], a useful machine learning suite that is highly utilized in supervised and unsupervised learning. This logistic regression algorithm utilizes a coordinate descent optimization algorithm implemented with the LIBLINEAR library [22]. The processed dataset which was obtained from the UCI repository was used to train the models. This means that the prediction models were fit to the training data in order to make predictions on the test set. This cycle of training the models and making predictions is performed in three distinct iterations, with the random seeds of the first, second and third iterations set at 5, 8 and 10, respectively, and the results of the performance measurements recorded. Taking into consideration the almost evenly-balanced class distribution of the input dataset, the running speed and classification accuracy have chiefly been chosen in this study to compare the performances of the logistic regression and decision-tree algorithms which were both applied to the same dataset in succession. The running speed of each algorithm compared is measured by computing the time difference in seconds between the moment the model-training phase commences and the moment the prediction phase is immediately completed. Metrics based on confusion matrix pairs, as well as the area under ROC curves were employed in

analysing the performance of classifiers. The area under each ROC curve can reveal the predictability strength of a classifier algorithm. A higher area under curve (AUC) value indicates that the corresponding classifier possesses superior predictive ability [23].

C. Validation

The validity criterion for the dataset is that nominal variables are required to contain categorical values only, while quantitative variables are required to have only numerical values. To leverage the knowledge to be acquired from the conclusion in this study, it is important to verify that neither fictitious nor erroneous results are obtained usually due to overfitting. Shuffled K-fold cross-validation generates a user-defined number of samples (K) of unconnected shuffled train/test data splits, while allowing for a finer control on the proportion of train/test samples in each split pair. In this isolated validation procedure performed on the full dataset, the number of splits was set to 5 and the train/test sample proportion was set to 10:90 for each split, just as the case in the training and prediction procedure. The cross-validation values obtained for the five splits were averaged for each train/test iteration and recorded. This method was selected in this study as a measure of upholding the validity of results emanating from performance evaluation of the learning algorithms. This is because cross-validated estimation of model performance is a more efficient use of data for model validation; however, it can be biased if the number of folds, K, is smaller compared to the number of events [24]. In a study on the effectiveness of K-fold cross-validation, it was shown that the proposed cross-validation method produced a good estimate of traditional methods and was an accurate assessment for linear models, as well as good approximations for non-linear models [25]. $K = 5$ was chosen because it is one of the widely-used and recommended values in literature [26], [27]. Reliability in statistics can be described as the overall consistency of performance of a measure. More consistent results repeatedly translate to higher reliability [28]. Therefore, an approach worthy of consideration is determining the consistency by inspecting the accuracy result values of each learning model over three iterations for any considerable discrepancies.

IV. Results and Discussion

Logistic regression and the classification and regression tree variant of decision tree learning models were implemented on the partitioned chronic kidney disease dataset in an attempt to perform the classification and prediction of the disease.
The results obtained were recorded and are duly presented below. The results are subsequently evaluated in order to determine and compare the performance of both algorithms.

It can be observed from Table I that the decision tree algorithm is always quicker than the logistic regression algorithm. This can be attributed to the classification and regression tree algorithm implicitly performing its feature selection using the Gini impurity criterion, as can be observed in Fig. 2 [29].

TABLE I
Performance Assessment Results of Logistic Regression (LR) & Classification and Regression Tree (CART)

	Running time, s		Area under ROC curve		Accuracy, %		Shuffled 5-fold cross-validation average	
	LR	CART	LR	CART	LR	CART	LR	CART
First iteration	0.010	0.005	0.986	0.983	98.61	98.33	0.9944	0.9800
Second iteration	0.008	0.005	0.997	0.978	99.72	97.77	0.9994	0.9833
Third iteration	0.013	0.005	1.0	0.972	100	97.22	0.9950	0.9861
Average	0.0103	0.005	0.9943	0.9776	99.44	97.77	0.9962	0.9831

Reference [30] showed that computing time depended on the total number of features used, and, consequently, a reduction in the running time due to feature selection was observed when compared with the logistic regression. In addition, the order of time complexity of classification and regression tree can be estimated as $O(m \times nlogn)$ [7] [31], where m is the number of features and n is the number of observations. In contrast, the order of time complexity of logistic regression is given by $O(nd)$ [32], where n is the number of samples and d is the dataset dimension. This translates to shorter execution time for the tree algorithm.

Furthermore, Table I also shows from the values of the average accuracy that the logistic regression algorithm possesses slightly superior accuracy than the decision tree algorithm. Table I also displays the cross-validation averages recorded for each train/test iteration using the shuffled K-fold cross-validation procedure described above, separate from the training and prediction procedure. It can be observed from the cross-validation averages that logistic regression also has slightly higher average values than CART. This can further be corroborated by the figures of the area under curve for both algorithms. Fig. 1, which is collection of ROC curves for the two algorithms after three iterations, further illustrates the superiority of the logistic regression algorithm in classification of chronic kidney disease data. The curves generated reveal that the logistic regression algorithm represented by the solid line possesses a greater area under its curve. This suggests that the logistic regression algorithm has a higher predictive accuracy.

TABLE II
Precision, Recall & F1-Score Results

	Class	Precision		Recall		F1-score	
		LR	DT	LR	DT	LR	DT
First iteration	NKD	0.97	0.97	1.0	1.0	0.99	0.98
	KD	1.00	1.00	0.97	0.97	0.99	0.98
Second iteration	NKD	0.99	0.96	1.00	1.00	1.00	0.98
	KD	1.00	1.00	0.99	0.96	1.00	0.98
Third iteration	NKD	1.00	0.95	1.00	1.00	1.00	0.97
	KD	1.00	1.00	1.00	0.94	1.00	0.97

Table II summarises the results for Precision, Recall and F1-scores of the NKD and KD classes in each iteration, where NKD represents the 'Not Kidney Disease' classification and KD is 'Kidney Disease' classification. Precision is calculated using the ratio of the correctly predicted positives to the total positive predictions. Recall is obtained by calculating ratio of predicted true positives to the sum of true positives and false negatives. A high precision value depicts a low false-positive prediction rate, while a high recall value depicts a low false-negative prediction rate. From Table II, it can be observed that both models produce high precision and recall values for the two classes with logistic regression slightly surpassing the tree model with better precision and recall. This implies that logistic regression is better at not misclassifying a positive sample as negative, as well as returning majority of the total actual positives as predicted positive results. The F1-score is the harmonic mean of precision and recall values. It is an equally weighted metric for balancing precision and recall, requiring both to have high values in order to obtain a high F1-score. Again, logistic regression possesses better scores than the tree model in this regard.

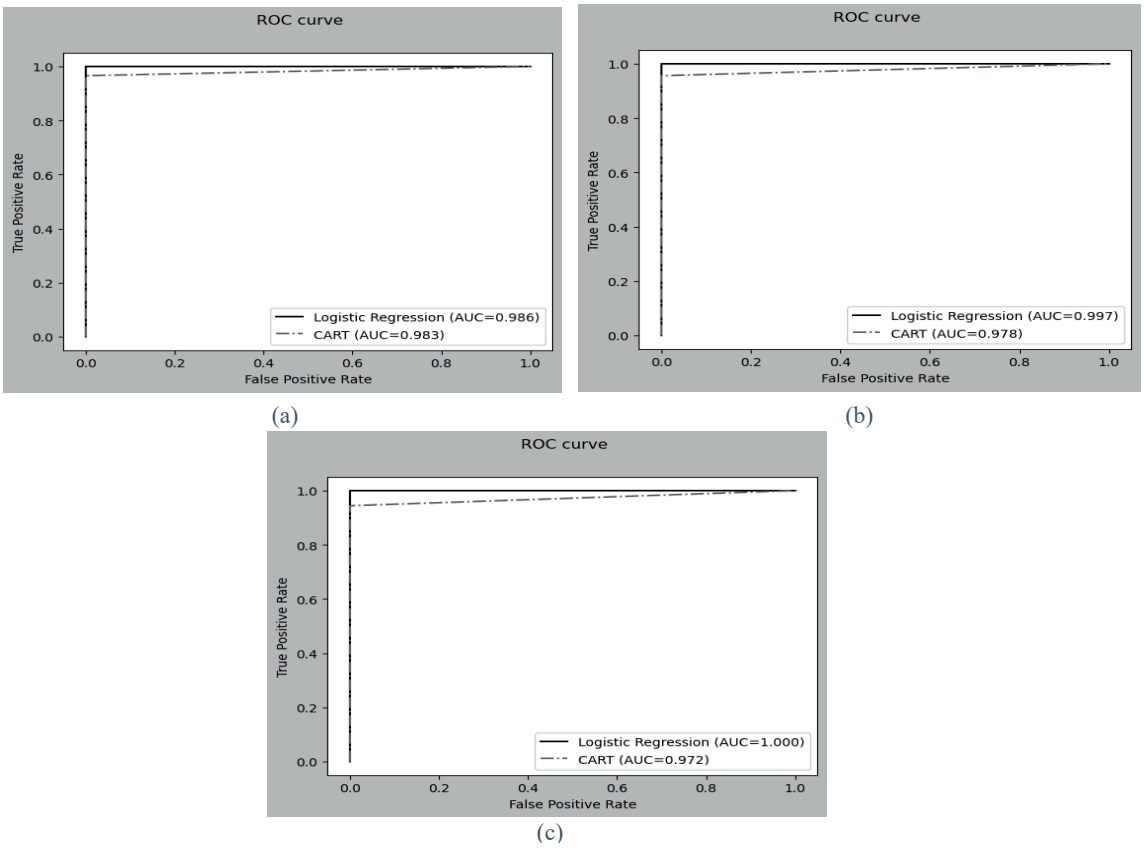

Fig. 1 Receiver operating characteristic curves of the three train/test iterations. (a), (b), (c) represent the iterations at random seed = 5, 8, 10, respectively.

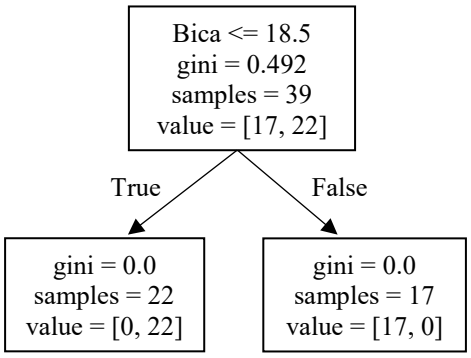

Fig. 2. Graphical representation of the decision tree model-training stage.

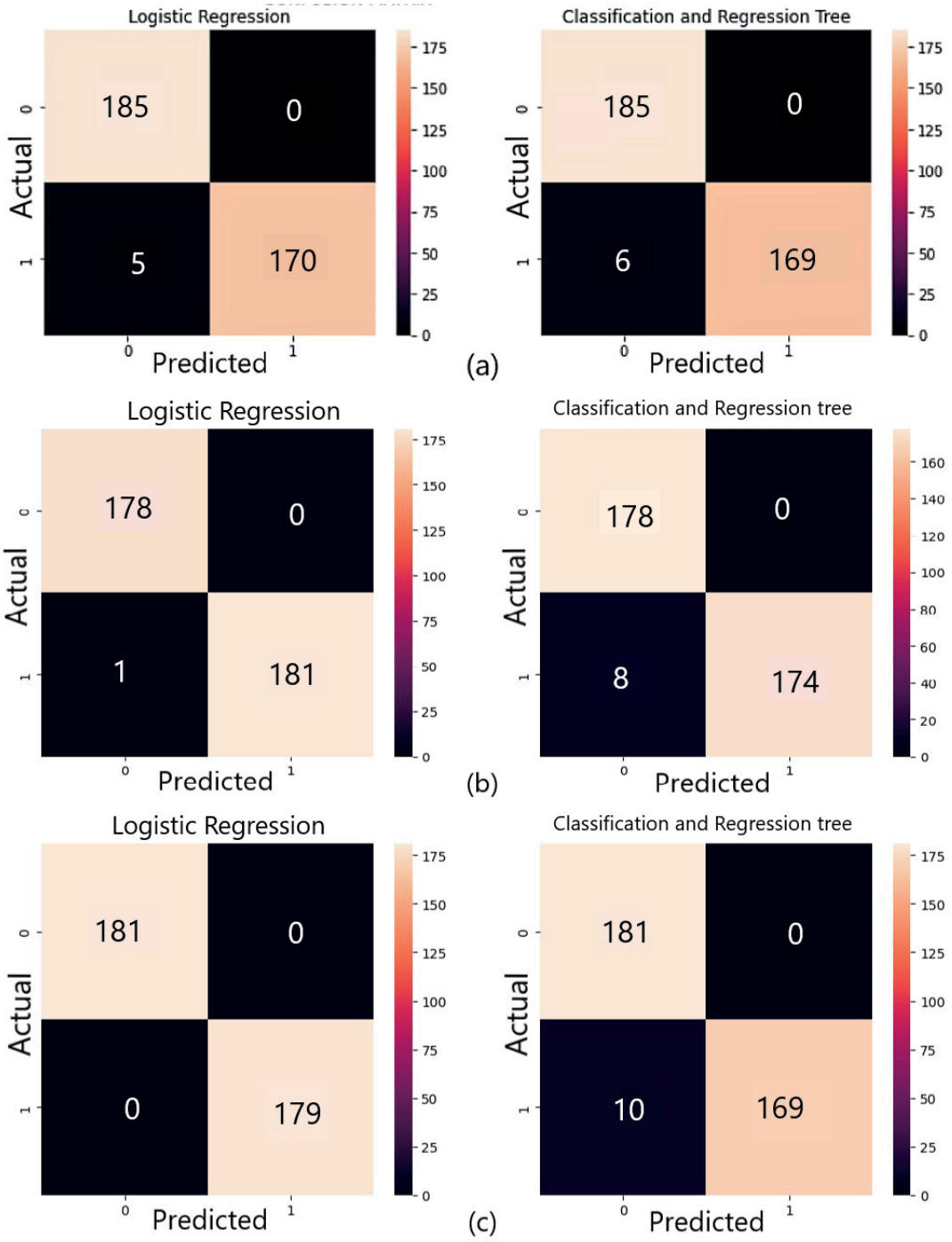

Fig. 3. Confusion matrix pairs from the three iterations of classification tasks using LR & CART. (a), (b), (c) represent the iterations at random seed = 5, 8, 10, respectively. Left-to-right diagonal represents the correctly predicted instances. Right -to-left diagonal represents the number of incorrect predictions.

Fig. 3 illustrates the results of the confusion matrix of each of the algorithms displayed in pairs. This was generated on the test dataset only after prediction in order to assess the classifier performance based on true positives and true negatives, which are the correct classifications, and false positives and false negatives, which are misclassifications. The quality of the output of the two classifiers can be evaluated using the confusion matrix. A closer investigation of the left-to-right diagonal reveals that the logistic regression algorithm has a higher representation of correct predictions. It can also be observed from the right-to-left diagonal that the decision-tree algorithm has a higher occurrence of wrong predictions. The left-to-right diagonal elements represent the number of points for which the predicted result is equal to the true result in the test data, while the right-to-left elements are those that are mislabelled by the classifier.

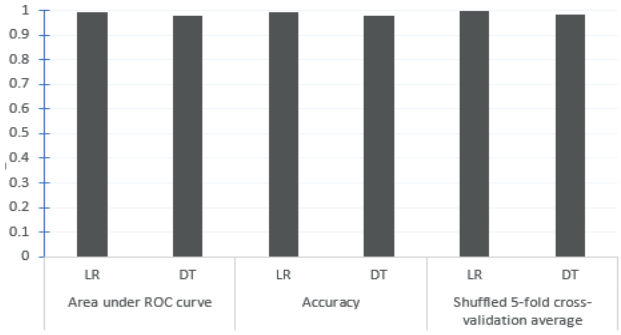

Fig. 4. Visualization of the averages of the major performance metrics of logistic regression and CART classification.

Fig. 4 graphically represents the average values of the three prediction iteration results for the area under ROC values, accuracy values and cross-validation results. Fig. 4 shows that logistic regression has slightly higher averages in all cases.

V. Conclusion

In this study, the author has examined two supervised machine learning classification algorithms for the prediction of kidney disease outcomes. An investigation of the comparison of the performances of logistic regression and CART algorithms has been performed in order to determine more suitable of the two, given inadequate training data. Depending on whether a piece of data is linearly separable or not, the selection of a more suitable algorithm between logistic regression or decision-tree algorithm for classification purposes can be determined. Missing data have been filled with values generated using a k-NN algorithm after discarding variables with missing data greater than 50 %. Logistic regression and CART algorithms have been trained using 10 % training data. Prediction has been performed on the rest of the dataset using the two algorithms as well. The metrics used in assessing performance have been accuracy, area under ROC and confusion matrix. The results reveal that the decision-tree algorithm is faster; however, it is less accurate than the logistic regression algorithm. The validation results obtained through a shuffled 5-fold cross-validation strategy for accuracy bear similarity to the previously obtained average accuracy results. The average accuracy, average area under ROC and average cross-validation values have been higher for the logistic regression model. This implies that logistic regression is a more accurate model for CKD prediction, given scarce training data. Overall, both algorithms are very capable of executing highly accurate classification and prediction tasks.

References

[1] B. Bikbov *et al.*, "Global, regional, and national burden of chronic kidney disease, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017," *The Lancet,* vol. 395, no. 10225, pp. 709–733, Feb. 2020. https://doi.org/10.1016/S0140-6736(20)30045-3

[2] K. J. Jager and S. D. Fraser, "The ascending rank of chronic kidney disease in the global Bburden of disease study," *Nephrology Dialysis Transplantation,* vol. 32, no. 2, pp. 121–128, Apr. 2017. https://doi.org/10.1093/ndt/gfw330

[3] S. R. Mitchell, S. A. Robert, and S. A. Lloyd, "An investigation into the use of machine learning for determining oestrus in cows," *Computers and Electronics in Agriculture,* vol. 15, no. 3, pp. 195–213, Aug. 1996. https://doi.org/10.1016/0168-1699(96)00016-6

[4] C. E. Brodley and P. E. Utgoff, "Multivariate decision trees," *Machine Learning,* vol. 19, no. 1, pp. 45–77, Apr. 1995. https://doi.org/10.1007/BF00994660

[5] V. V. Raghavan, V. N. Gudivada, V. Govindaraju, and C. Rao, Eds., *Cognitive Computing: Theory and Applications*, Volume 35, 1st ed., 2016, p. 183.

[6] G.-X. Yuan, C.-H. Ho, and C.-J. Lin, "Recent advances of large-scale linear classification," *Proceedings of the IEEE,* vol. 100, no. 9, pp. 2584–2603, Apr. 2012. https://doi.org/10.1109/JPROC.2012.2188013

[7] A. Géron, *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*, 2nd ed., CA: Orielly media, 2019, pp. 182–184.

[8] J. Bodine and D. S. Hochbaum, "The max-cut decision tree: Improving on the accuracy and running time of decision trees," *arXiv preprint arXiv:2006.14118,* p. 1, 2020.

[9] W. J. Long, J. L. Griffith, H. P. Selker, and R. B. D'Agostino, "A comparison of logistic regression to decision-tree induction in a medical domain," *Computers and Biomedical Research,* vol. 26, no. 1, pp. 74–97, Feb. 1993. https://doi.org/10.1006/cbmr.1993.1005

[10] W. Gunarathne, K. Perera, and K. Kahandawaarachchi, "Performance evaluation on machine learning classification techniques for disease classification and forecasting through data analytics for chronic kidney disease (CKD)," in *2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE)*, Washington, DC, USA, Oct. 2017, pp. 291–296. https://doi.org/10.1109/BIBE.2017.00-39

[11] J. Xiao, R. Ding, X. Xu, H. Guan, X. Feng, T. Sun, S. Zhu, and Z. Ye, "Comparison and development of machine learning tools in the prediction of chronic kidney disease progression," *Journal of Translational Medicine,* vol. 17, Art no. 119, Apr. 2019. https://doi.org/10.1186/s12967-019-1860-0

[12] D. Dua and C. Graff, *UCI machine learning repository,* University of California, Irvine, School of Information and Computer Sciences, 2017.

[13] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, 1st ed., Morgan Kaufmann, 1999, p. 58.

[14] H. Kang, "The prevention and handling of the missing data," *Korean Journal of Anesthesiology,* vol. 64, no. 5, pp. 402–406, May 2013. https://doi.org/10.4097/kjae.2013.64.5.402

[15] O. F. Ayilara, L. Zhang, T. T. Sajobi, R. Sawatzky, E. Bohm, and L. M. Lix, "Impact of missing data on bias and precision when estimating change in patient-reported outcomes from a clinical registry.," *Health and Quality of Life Outcomes,* vol. 17, Art no. 106, Jun. 2019. https://doi.org/10.1186/s12955-019-1181-2

[16] J. Poulos and R. Valle, "Missing data imputation for supervised learning," *Applied Artificial Intelligence,* vol. 32, no. 2, pp. 186–196, Mar. 2018. https://doi.org/10.1080/08839514.2018.1448143

[17] O. G. Troyanskaya, M. N. Cantor, G. Sherlock, P. O. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, "Missing value estimation methods for DNA microarrays," *Bioinformatics,* vol. 17, no. 6, pp. 520–525, Jun. 2001. https://doi.org/10.1093/bioinformatics/17.6.520

[18] W. McKinney, "Data structures for statistical computing in Python," in *Proceedings of the 9th Python in Science Conference*, Austin, TX, USA, Jul. 2010. https://doi.org/10.25080/Majora-92bf1922-00a

[19] H.-F. Yu, F.-L. Huang and C.-J. Lin, "Dual coordinate descent methods for logistic regression and maximum entropy models," *Machine Learning,* vol. 85, no. 1, pp. 41–75, Oct. 2011. https://doi.org/10.1007/s10994-010-5221-8

[20] L. Breiman, *Classification and Regression Trees*, Chapman and Hall/CRC, 1983.

[21] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research,* vol. 12, no. 85, pp. 2825–2830, 2011. https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf

[22] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *Journal of Machine Learning Research,* vol. 9, no. 9, pp. 1871–1874, 2008.

[23] N. J. Perkins and E. F. Schisterman, "The inconsistency of "optimal" cutpoints obtained using two criteria based on the receiver operating characteristic curve," *American Journal of Epidemiology,* vol. 163, no. 7, pp. 670–675, Apr. 2006. https://doi.org/10.1093/aje/kwj063

[24] R. M. Simon, J. Subramanian, M.-C. Li, and S. Menezes, "Using cross-validation to evaluate predictive accuracy of survival risk classifiers based on high-dimensional data," *Briefings in Bioinformatics,* vol. 12, no. 3, pp. 203–214, May 2011. https://doi.org/10.1093/bib/bbr001

[25] C. Alippi and M. Roveri, "Virtual k-fold cross validation: An effective method for accuracy assessment," in *The 2010 International Joint Conference on Neural Networks (IJCNN)*, Barcelona, Spain, Jul. 2010. https://doi.org/10.1109/IJCNN.2010.5596899

[26] Q. Noirhomme, D. Lesenfants, F. G. Gómez, A. Soddu, J. Schrouff, G. Garraux, A. Luxen, C. Phillips, and S. Laureys, "Biased binomial assessment of cross-validated estimation of classification accuracies illustrated in diagnosis predictions.," *NeuroImage: Clinical,* vol. 4, pp. 687–694, 2014. https://doi.org/10.1016/j.nicl.2014.04.004

[27] S. Lemm, B. Blankertz, T. Dickhaus, and K. R. Müller, "Introduction to machine learning for brain imaging.," *NeuroImage,* vol. 56, no. 2, pp. 387–399, 2011. https://doi.org/10.1016/j.neuroimage.2010.11.004

[28] S. Adelabu, O. Mutanga, and E. Adam, "Testing the reliability and stability of the internal accuracy assessment of random forest for classifying tree defoliation levels using different validation methods," *Geocarto International,* vol. 30, no. 7, pp. 810–821, Feb. 2015. https://doi.org/10.1080/10106049.2014.997303

[29] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research,* vol. 3, no. 7–8, pp. 1157–1182, Mar. 2003. https://psycnet.apa.org/record/2003-10726-001

[30] A. Majeed, "Improving time complexity and accuracy of the machine learning algorithms through selection of highly weighted top k features from complex datasets," *Annals of Data Science,* vol. 6, no. 4, pp. 599–621, May 2019. https://doi.org/10.1007/s40745-019-00217-4

[31] H. M. Sani, C. Lei, and D. Neagu, "Computational complexity analysis of decision tree algorithms," in *International Conference on Innovative Techniques and Applications of Artificial Intelligence*, Cambridge, UK, Nov. 2018. https://doi.org/10.1007/978-3-030-04191-5_17

[32] P. Kumar, "Computational complexity of ML models," *Analytics Vidhya*, 14 December 2019. [Online]. Available: https://medium.com/analytics-vidhya/time-complexity-of-ml-models-4ec39fad2770.

Oluwadamilare Alabi received his Master degree in Information Technology from the National Open University of Nigeria in 2022. Currently, he is a Research Fellow at the Applied Mathematics and Simulation Research Centre of Sheda Science & Technology Complex, Federal Capital Territory, Nigeria. His current research interests lie in the areas of machine learning, deep learning, geoinformatics, and high-performance computing.
E-mail: oi.alabi@shestco.gov.ng
ORCID iD: https://orcid.org/0000-0003-0664-4096