A Comparison of Logistic Regression, Modified Logistic Regression and Naïve Bayes Models for Classifying HIV Viral Load Suppression: The Case of Zombo District in Uganda

Jayer Shepherd, Douglas Candia & Francis Fuller Bbosa

Makerere University

ABSTRACT

Purpose: To enhance the performance of the logistic regression by integrating step-wise procedures and henceforth compare and evaluate its performance with the Logistic regression and Naïve Bayes in classifying HIV viral load suppression (VLS).

Methods: Models for classifying VLS were built using Logistic regression, modified logistic regression and Naïve Bayes classifiers. Accuracy, balanced accuracy and the area under the receiver operating characteristics curve (AUC) were the key performance metrics used to evaluate the generalizability of the various classifiers.

Results: The modified logistic regression model trained on fewer predictor attributes achieved an accuracy of 84.9%, a balanced accuracy of 83.8% and an AUC of 92.6%. The traditional logistic regression model trained on a full set of predictor attributes achieved an accuracy of 84.9%, a balanced accuracy of 83.6% and an AUC of 92.5% whereas the naïve Bayes model achieved an accuracy of 81.6%, a balanced accuracy of 80.5% and AUC of 89.4%.

Keywords: comparison, logistic, modified, naïve bayes, classification, viral load suppression.

Classification: NLM Code: WA 503-610

Language: English

Great Britain Journals Press

LJP Copyright ID: 392883

London Journal of Medical and Health Research

Volume 23 | Issue 13 | Compilation 1.0

A Comparison of Logistic Regression, Modified Logistic Regression and Naïve Bayes Models for Classifying HIV Viral Load Suppression: The Case of Zombo District in Uganda

Jayer Shepherd[α], Douglas Candia[σ] & Francis Fuller Bbosa[ρ]

ABSTRACT

Purpose: To enhance the performance of the logistic regression by integrating step-wise procedures and henceforth compare and evaluate its performance with the Logistic regression and Naïve Bayes in classifying HIV viral load suppression (VLS).

Methods: Models for classifying VLS were built using Logistic regression, modified logistic regression and Naïve Bayes classifiers. Accuracy, balanced accuracy and the area under the receiver operating characteristics curve (AUC) were the key performance metrics used to evaluate the generalizability of the various classifiers.

Results: The modified logistic regression model trained on fewer predictor attributes achieved an accuracy of 84.9%, a balanced accuracy of 83.8% and an AUC of 92.6%. The traditional logistic regression model trained on a full set of predictor attributes achieved an accuracy of 84.9%, a balanced accuracy of 83.6% and an AUC of 92.5% whereas the naïve Bayes model achieved an accuracy of 81.6%, a balanced accuracy of 80.5% and AUC of 89.4%.

Conclusion: The modified logistic regression model outperformed the traditional logistic regression and naïve Bayes models on account of recording higher balanced accuracy and AUC values of 83.8% and 92.6% respectively albeit with fewer predictor attributes. Hence integrating step-wise regression procedures in the traditional logistic regression model can enhance its classification performance leading to better predictions.

Keywords: comparison, logistic, modified, naïve bayes, classification, viral load suppression.

Author α σ ρ: School of Statistics and Planning, Makerere University, Kampala, Uganda. Department of Planning and Applied Statistics School of Statistics and Planning, Makerere University.

I. INTRODUCTION

Logistic regression and Naïve Bayes are among the most used data mining classification techniques [1]. This might be because they have algorithms that are easy to implement [2, 3], their ability to handle both continuous and discrete data [1], application of probability theory in their classification modelling [4] and they produce real-time predictions that can be easily interpreted [5].

The Logistic regression classifier assumes the absence of multicollinearity among the predictors while conducting classifications [6]. However, the performance of the Logistic regression classifier is usually weakened by the presence of multicollinearity among the predictors which may lead to poor classifications [7]. A study by Senaviratna & Cooray [8] reported that the best solution is to understand the cause of multicollinearity and remove the highly correlated variables in the model. However, O'Brien [9], objected to removing the correlated variables from the model because less information would be available potentially leading to suboptimal model performance. Despite this weakness associated with the Logistic regression classifier, several scholars [10, 11, 12] have independently employed the logistic regression classifier while undertaking classification tasks.

Scholars [13, 14] have recommended the use of the Bayesian approach premised on Bayes' theorem as an alternative to the Logistic regression classifier to overcome the problem of multicollinearity because its assumption of mutual independence among the predictors enables each distribution to be independently estimated. Several scholars [15, 16, 17] further stressed that the Naïve Bayes classifier has superior strengths such as being efficient, computationally fast, and does not require a lot of data for training to conduct classifications. Owing to its strength, the Naïve Bayes classifier has outperformed the Logistic Regression classifier in various fields [3, 18, 19] to provide accurate and reliable results. Asharaf et al., [20] recommended the need for more comparative studies of the different data mining techniques to determine their classification ability so that the most optimal model can be chosen. Above all, a number of approaches have been proposed to improve the goodness of fit of the traditional logistic regression classifier in order to overcome its multicollinearity bottlenecks [21, 22]. These include Principal Components Analysis (PCA) [23], Monte Carlo simulation [24] and Variance Inflation factor (VIF) [25], which drops predictors with high VIFs.

Motivated by the performance improvement of enhanced independent classifiers [26] coupled with the fact that stepwise regression procedures are more likely to have lower false classification rates [27]. This study proposes a modified logistic regression classifier which employs the VIF integrated with step-wise regression due to its simplicity [21].

1.1 HIV Viral Load Suppression as a Case Study

HIV viral load suppression (VLS) is the ultimate measure of treatment success for People Living with HIV/AIDS (PLHIV) receiving antiretroviral therapy (ART) [28]. This is in line with the third Sustainable Development Goal (SDG 3) premised on the commitment made by the United Nations (UN) member states to end the AIDS epidemic by 2030 by achieving 95% VLS by 2025 [29, 30]. The consolidated guidelines on the use of ART drugs for treating and preventing HIV infection define all PLHIV receiving ART with HIV viral load (VL) less than 1000 copies/mL as having a suppressed VL [28]. According to the Annual Health Sector Report for the Financial Year 2021/2022, Zombo District achieved VLS of 71% [28]. This falls below the national VLS rate for Uganda of 82% [29] and also below the UNAIDS 95-95-95 target of at least 95% VLS by 2025. Despite several efforts to improve the treatment outcome of PLHIV receiving ART outcome through health education, infrastructural development, bridging the human resource gaps and strengthening the supply chains for essential commodities [31], so little is known about the key factors associated with VLS as well as the performance of various classifiers in determining these factors among PLHIV on ART in Zombo District.

1.2 Classifier Evaluation Metrics

Evaluating the performance of a classifier is paramount as it permits researchers to compare competing models as well as determine the degree to which its results can be generalized to an unseen sample or population from the same distribution from which the existing data were drawn [32]. Several scholars [33, 34, 35] attest that the confusion matrix; which provides a summary of classification outcomes, is the commonest way for evaluating classifier performances.

On the other hand, presenting a confusion matrix by itself in the absence of a suitable summary statistic or metric is insufficient and easily leads to biased interpretations of performance [32]. The most utilized summary statistic emanating from the confusion matrices is accuracy, defined as the number of correct predictions across all classes [36, 37]. However, classification accuracy is a misleading performance metric particularly when the data are not perfectly balanced [36, 38].

The balanced accuracy metric, defined as the arithmetic mean from both the minority and majority classes was suggested to address the above limitation [32]; thus providing more reliable performance evaluations for imbalanced data [39, 40].

1.3 Problem Statement

Many scholars have independently employed logistic regression for classification problems. However, the performance of the Logistic regression classifier is usually weakened by multicollinearity among the predictors which may lead to poor classification results. Naïve Bayes has been suggested as an alternative classifier to overcoming multicollinearity as it assumes mutual independence among the predictors. To this end, limited research has been done to enhance the performance of the logistic regression as well as compare its performance with respect to the naive Bayes classifier.

In order to deal with the existing multicollinearity challenges of the traditional logistic regression classifier, this paper proposes a modified logistic regression classifier which employs a step-wise procedure based on VIFs and hence compare and evaluate its performance with the traditional logistic regression and naïve Bayes classifiers on a similar dataset to determine the most optimal classifier.

II. METHODS

2.1 Data Preprocessing

2.1.1 Data Sources

Data was extracted from Patient forms in one Hospital and nine health facilities of level three (HC IIIs) in Zombo District [41] that are accredited to offer antiretroviral therapy (ART) services for PLHIV who were newly initiated on ART between February 2020 to May 2022. This period conforms with the revised ART guidelines that specify the evaluation of VL for all newly identified PLHIV started on ART after six months of ART treatment [42, 43]. The extracted variables and their descriptions are indicated in Table 1.

Table 1: Description of Variables used in this Study

Sn	Variable Name	Description	Variable Type	Categories
1	HIV Clinic No.	Unique Number is assigned to the HIV patient upon being enrolled on care at the ART clinic in a health facility		
2	Age	Age of the HIV patient in completed years	Continuous	This was transformed into four (4) categories namely; 0 – 9 years, 10 – 19 years, 20 – 49 years, 50 years and above
3	Gender	Gender of the HIV patient	Categorical	M-Male, F-Female
4	Marital	Marital Status of the HIV patient	Categorical	Married, Never Married, Separated, Widow
5	Stage	HIV WHO Clinical Stage of the patient	Categorical	Stage 1, Stage 2, Stage 3, Stage 4
6	regimen	ART regimen	Categorical	DTG-based regimen, LPV-based regimen, NPV-based regimen
7	freq	Daily ART dosage drugs	Categorical	Once per day, Twice per day
8	month_2	HIV patient monthly clinical encounter at the second month	Categorical	Active, Missed Appointment
9	month_3	HIV patient monthly clinical encounter at the third month	Categorical	Active, Missed Appointment, Lost to Follow up
10	month_4	HIV patient monthly clinical encounter at fourth month	Categorical	Active, Missed Appointment, Lost to Follow up
11	month_5	HIV patient monthly clinical encounter at the fifth month	Categorical	Active, Missed Appointment, Lost to Follow up
12	month_6	HIV patient monthly clinical encounter at the sixth month	Categorical	Active, Missed Appointment, Lost to Follow up

Sn	Variable Name	Description	Variable Type	Categories
13	adherence	Adherence to taking ART drugs by the HIV patient during the sixth months period on care	Categorical	Fair, Good, Poor
14	Disclosure	Disclosure of HIV status by the client	Categorical	Yes, No
15	VLS	HIV Viral Load Suppression outcome after 6 months of being on HIV care.	Categorical /Binary variable	0=Suppressed, 1=Non-Suppressed

A total of 1,757 records were extracted, each denoting a newly identified PLHIV started on ART after six months of ART treatment.

2.1.2 Data Cleaning

The researchers examined the data set for missing values, outliers, and addressed discrepancies and observations with missing values were excluded [44].

2.1.3 Data Transformation

Given that the dataset contained both continuous and categorical variables, data discretization was used to create homogenous groups of continuous predictor variables, reducing outliers and minimizing noise formation [45].

2.1.4 Training and Validation Datasets

The researchers split the dataset into training and testing datasets. 70% of the dataset was assigned to the training group for the development of the classifiers. The rest of the dataset (30% of the total cases) was assigned to the validation groups for the assessment of model performance [46].

Analysis

The following classifiers were employed;

Naïve Bayes: It utilizes the Bayes theorem [47] to compute the posterior probability of dependent variable Z given predictor variables $Y = (y_1, y_2, \ldots, y_n)$ the following equation (1).

$$P\left(\frac{Z_n}{y_1, y_2, \ldots, y_n}\right) = \frac{P(Z_n)\prod_{i=1}^{n} P\left(P\left(\frac{y_i}{Z_n}\right)\right)}{\prod_{i=1}^{n} P(y_i)} \quad (1)$$

Where $P(Z_n)$= the probability of Z to be observed

$P(y_i)$= the probability of y to be observed

$P\left(\frac{Z_n}{y_i}\right)$= the posterior probability of class (Z) given predictor (y).

$P\left(\frac{y_i}{Z_n}\right)$= the probability of observing y given Z holds

Since the naïve Bayes assumption is that predictors $(y_1, y_2, \ldots, y_n)$ are conditionally independent of the response variable Z , the posterior probabilities $P\left(\frac{Z=1}{Y}\right)$ and $P\left(\frac{Z=0}{Y}\right)$ are computed for a new sample by dropping the denominator in equation (1) as illustrated in equation (2).

$$x = argmax(P\left(\frac{Z_n}{y_i}\right) = P(Z_n)\prod_{i=1}^{n} \quad (2)$$

Where x is the class of the response variable with the highest probability given a set of variables.

Logistic regression: It uses numerical and or categorical predictors to estimate the likelihood of a dichotomous response variable [5]. The logistic regression model can be expressed as;

$$\log log\left(\frac{P(y_i=1)}{P(y_i=0)}\right) = \alpha + \beta_1 x_1 + \quad (3)$$

Where Y denotes the response variable

x_i denotes the predictor variables

β_i denotes the coefficients of the predictor variables

α denotes the intercept

The probability of p_i is represented by equation (4)

$$p_i = \frac{1}{(1+e^{-x\beta})} \quad \in[0,1] \quad (4)$$

2.1.5 Proposed Modified Logistic Regression

The researchers integrated the backward stepwise regression process [48] into the traditional logistic regression indicated in equation (iii) in order to determine the importance of each predictor variable [49].

The researchers commenced with a full classifier and kept removing predictor attributes with the least significant values (highest P-values>0.05; variables that worsen the model highest), to the trained model, one at a time. For every removal, the trained modified logistic classifier was fit/generalized onto test data until the stopping criteria were met. The criteria to terminate was achieving balanced accuracy metrics similar to or higher than those returned by the traditional logistic regression classifier. The above process was repeated until only variables that generated a parsimonious model were retained in the classifier

2.1.6 Goodness of fit

The 10-fold cross-validation method was employed to validate the accuracy, sensitivity, specificity and balanced accuracy of the classifiers [33] as indicated the equations (5) (6), (7) and (8);

$$Accuracy = \left(\frac{TP+TN}{TP+TN+FP+FN}\right) \quad (5)$$

$$Sensitivity = \left(\frac{TP}{TP+FN}\right) \quad (6)$$

$$Specifity = \left(\frac{TP+TN}{TP+TN+FP+FN}\right) \quad (7)$$

$$Balanced\ Accuracy = \left(\frac{Sensitivity+Specificity}{2}\right) \quad (8)$$

Where in the context of this study, the entries in the confusion matrix were defined as.

1. True positive (TP): is the number of actual "NO" VLS cases classified as "NO".
2. False-positive (FP): is the number of actual "YES" VLS cases classified as "NO"
3. False Negative (FN): is the number of actual "NO" VLS cases classified as "YES".
4. True Negative (TN): is the number of actual "YES" VLS cases classified as "YES".

Software

The data processing and analysis were carried out in R, version 4.1.2 [50], using the R packages "dplyr" version 1.0.7 [51], "caret" version 6.0-90 [52], "pROC" version 1.18.0 [53] and "ROCR" version 1.0-11 [54].

III. RESULTS

In this section, the researchers first present the results from each classifier and then present the comparison results.

3.1 Key Variables for Classification of VLS

Results revealed that fourteen (14) out of 25 variables were key for classifying VLS, namely; "never married", "HIV WHO clinical stage 3",

"HIV WHO clinical stage 4", "daily ART dosage twice", "month 3 lost to follow up", " month 4 lost to follow up", "month 4 missed appointment", "month 5 lost to follow up", "month 5 missed appointment", "month 6 lost to follow up", "month 6 missed appointment", "good ART drug adherence", "poor ART drug adherence" and "disclosure of HIV status by the patient".

Table 2: Key Variables for Classification of VLS for the Logistic vs Modified Logistic Regression Models

Model	Accuracy(%)
Traditional Logistic regression model with all variables (25)	84.9
Modified Logistic regression model with fewer variables (14)	84.9

Table 2 reveals that when the modified logistic regression is trained on the dataset, the number of predictor variables is reduced from 25 to 14. This implies that the modified logistic regression model was able to achieve the accuracy of the traditional logistic regression trained on a full set of variables at the expense of some irrelevant or correlated variables. Hence the resultant variable subset using the modified logistic regression is the most significant set of variables that improves the predictive accuracy of VLS and thus a more robust model for determining VLS.

3.2 Comparison of the Classifiers Performance

The performance of the classifiers was evaluated based on their capacity to classify the instances of the data set into "YES" and "NO" VLS. The researchers utilized 10-fold cross-validation to assess the performance of the three classifiers on previously unlearned data. Computation of the performance metrics indicated in equations 8-11 revealed the results indicated in Fig. 1.

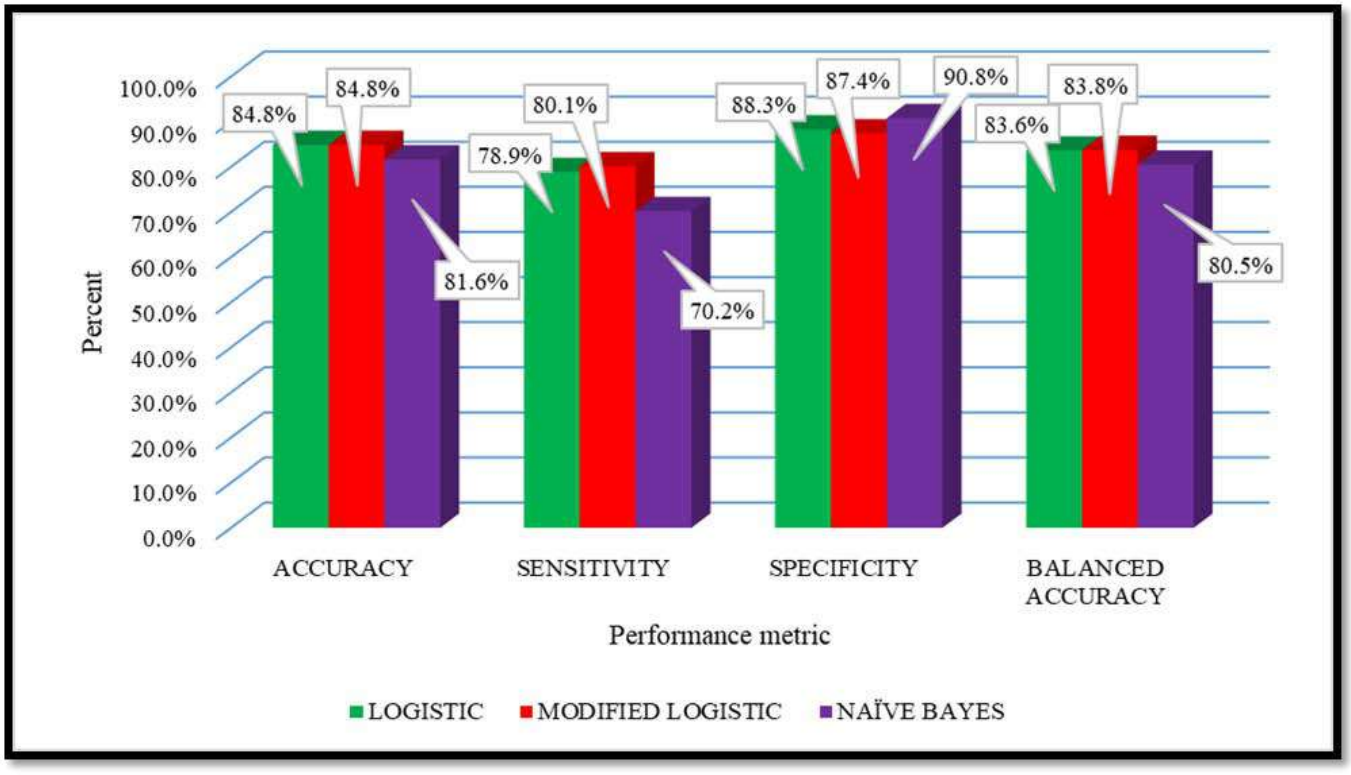

Fig. 1: Comparison of Classifiers' Performance Using 10 Fold Cross-Validation

According to *Fig. 1,* the modified logistic regression model attained the highest performance with respect to the accuracy, sensitivity, and balance accuracy metrics recorded at 84.8%, 80.1%, and 83.8% respectively. This implies that this model correctly classified 84.8%

(accuracy) of PLHIV whose viral load was either suppressed or not suppressed. Additionally, this model also correctly classified 80.1% of PLHIV whose viral load was not-suppressed.

On the other hand, the naïve Bayes classifier registered the highest specificity at 90.8% compared to 88.3% registered by traditional logistic regression and 87.4% obtained by the modified logistic regression classifier. The achieved balanced accuracy results indicate that the proposed modified logistic regression model outperformed the traditional logistic regression and naïve Bayes classifiers by 0.2% and 10.3% respectively.

Comparatively, raw data in Table 1 revealed that the response variable (VLS) comprised uneven proportions of 36% suppressed VL and 64% suppressed VL and therefore the balanced accuracy metric was used as the overall evaluation which balances the precision and recall metrics across each response variable class [55].

3.2.1 Receiver Operating Characteristics (ROC) Curve

The ROC curve *(Fig. 2)* is a graphical illustration of the relationship between the performance of a classifier's sensitivity and specificity [42]. The ROC enabled the researcher to evaluate how well the developed models performed at different thresholds. *Fig. 2* shows that the Modified logistic regression, traditional logistic regression and naive Bayes classifiers' corresponding Area under the Curve (AUC) values were 92.6%, 92.5% and 89.4%. A random model would simply divide the graph in half, giving it an AUC of 50%. For this reason, the classifiers' produced ROC curves supersede a random model, showing that the applied models provide a good measure of separability.

The purple line, which denotes the modified logistic regression, generated a superior cut-off decision level than the other two classifiers since it maximised the true positive rate at the lowest level of false positives (1-specificity).

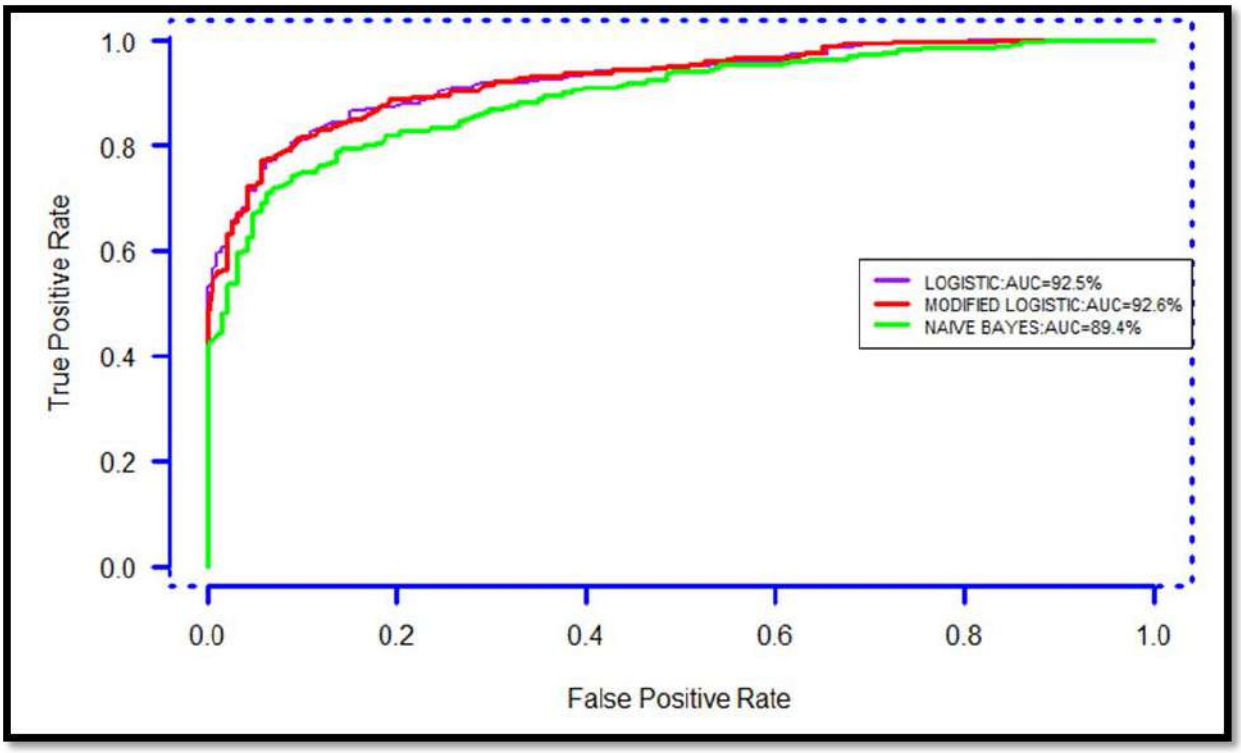

Fig. 2: Comparison of the ROCs for the Classifiers at Various Thresholds

IV. DISCUSSION

The purpose of this study was two fold. Firstly, to enhance the performance of the traditional logistic regression classifier and secondly, to compare and evaluate the performance of the traditional logistic regression, modified logistic regression and Naive Bayes models in classifying VLS. The findings showed that the modified

logistic regression classifier slightly outperformed the traditional logistic regression and naïve Bayes classifiers with regards to accuracy, sensitivity and balanced accuracy whereas the naïve Bayes performed best in terms of specificity.

The proposed modified logistic regression classifier inherits properties of the backward stepwise regression algorithm. This implies that integrating the step wise regression procedures into traditional data mining classifiers can enhance their classification performance as evidenced by the better performance of the modified logistic regression classifier when fitted on previously unknown data samples. This phenomenon is in agreement with those of previous studies [56, 57, 58] that reveal that the performance of the traditional data mining classifiers can be improved by integrating it with other machine learning techniques. In terms of key determinants of VLS, our findings were consistent with those of [59, 60, 61, 62].

Conversely, the study faced a key challenge of available data being limited to data whose variables were regularly gathered from patients and caretakers and recorded in the patient medical records systems for the period under investigation hence the researchers were unable to subject the developed modified model to a higher dimensional dataset in terms of variables and observations from a known population which would return more reliable and robust performance results [63].

V. CONCLUSION

In this study, a modified logistic regression classifier is proposed to further enhance the classification performance of the traditional logistic regression classifier. Furthermore, performance comparisons were made between the modified logistic regression, traditional logistic regression and naïve Bayes classifiers. We found that the modified logistic regression performed slightly better than the traditional logistic regression and naïve Bayes classifiers on account of recording higher balanced accuracy and AUC values of 83.8% and 92.6% respectively albeit with fewer predictor attributes. We attribute this to the fact that the modified logistic regression adapts a step-wise regression procedure which uses a linear combination of the best variables to form a robust classifier, unlike the traditional logistic regression and naïve Bayes. Hence integrating step-wise regression procedures in the traditional logistic regression model can enhance its classification performance leading to better predictions.

ACKNOWLEDGEMENTS

The authors appreciate the efforts of the anonymous reviewers who contributed to improving this research. The financial support from Zombo district Local Government is also deeply appreciated.

Statements and Declarations

Funding: The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Conflicts of interest/Competing interests: The authors declare that they have no conflict of interest

Availability of data and material: The data was sourced from the Patient forms in one Hospital and nine health facilities of level three (HC IIIs) in Zombo District, Uganda and it has been availed/ uploaded as supplementary material.

Code availability: The program scripts/code can be availed by the corresponding author upon request

Authors' contributions: JS was involved in drafting the proposal, data collection, data preprocessing, data analysis, model designing and writing the manuscript. FFB was also involved in data analysis, model designing and writing the manuscript. DC and FB were supervisors of the work. All authors read and approved the final manuscript.

Ethical considerations: The data were accessed with official permission from Zombo District Health Office and personal identification data was de-identified and treated with the utmost confidentiality.

Research involving human participants: Not applicable. No experiment was performed on animal or human subjects.

REFERENCES

1. Kumar Bhowmik, T. (2015). Naive Bayes vs Logistic Regression: Theory, Implementation and Experimental Validation. Inteligencia Artificial, 18 (56), 14–30. https://doi.org/10.4114/ intartif.vol18iss56pp14-30.
2. Dong, Longjun & Wesseloo, Johan & Potvin, Yves & Li, Xibing. (2015). Discrimination of Mine Seismic Events and Blasts Using the Fisher Classifier, Naive Bayesian Classifier and Logistic Regression. Rock Mechanics and Rock Engineering. 49. https://10.1007/s00603-015-0733-y.
3. Samsudin, Nur'Ain & Mohd Foozy, Cik Feresa & Alias, Nabilah & Shamala, Palaniappan & Othman, Nur & Wan Din, Wan Isni Sofiah. (2019). Youtube spam detection framework using naïve Bayes and logistic regression. Indonesian Journal of Electrical Engineering and Computer Science. 14. 1508-1517. https://10.11591/ijeecs.v14.i3.pp1508-1517.
4. D. Seka, D. Seka, B.S. Bonny, B. Bonny, A.N. Yoboué, A. Yoboué, S.R. Sié, S. Sié, & B.A. Adopo-Gourène, B. Adopo-Gourène. (0000). Identification of maize (Zea mays L.) progeny genotypes based on two probabilistic approaches: Logistic regression and naïve Bayes. Artificial intelligence in agriculture, 1, 9-13. https://10.1016/j.aiia.2019.03.001.
5. Prabhat, A., & Khullar, V. (2017). Sentiment classification on big data using Naïve Bayes and logistic regression. *2017 International Conference on Computer Communication and Informatics (ICCCI)*, 1-5.
6. Harris JK. Primer on binary logistic regression. Fam Med Community Health. 2021 Dec; 9 (Suppl 1):e001290. PMCID: PMC8710907.https://10.1136/fmch-2021-001290
7. Khikmah, Lelatul & Wijayanto, Hari & Syafitri, Utami. (2017). Modelling Governance KB with CATPCA to Overcome Multicollinearity in the Logistic Regression. Journal of Physics: Conference Series. 824. 012027. https://10.1088/1742-6596/824/1/ 01-2027.
8. Senaviratna, NAMR & Cooray, T.. (2019). Diagnosing Multicollinearity of Logistic Regression Model. Asian Journal of Probability and Statistics. 1-9. https://10.9734/ajpas/2019/v5i230132.
9. O'Brien, Robert. (2016). Dropping Highly Collinear Variables from a Model: Why it Typically is Not a Good Idea: Dropping Highly Collinear Variables from a Model. Social Science Quarterly. 98. https://10.1111/ ssqu.1227.
10. R. Kumar, S. M. Naik, V. D. Naik, S. Shiralli, Sunil V.G and M. Husain, "Predicting clicks: CTR estimation of advertisements using Logistic Regression classifier," *2015 IEEE International Advance Computing Conference (IACC)*, Banglore, India, 2015, pp. 1134-1138. https://10.1109/IADCC.2015.7154880.
11. Mokhtar, Muhammad & Jusoh, Yusmadi & Admodisastro, Novia & Che Pa, Noraini & Amruddin, Amru. (2019). Fakebuster: Fake News Detection System Using Logistic Regression Technique In Machine Learning. International Journal of Engineering and Advanced Technology. 9. 2407-2410. https://10.35940/ijeat.A2633.109119.
12. M. Al Omari, M. Al-Hajj, N. Hammami and A. Sabra, "Sentiment Classifier: Logistic Regression for Arabic Services' Reviews in Lebanon," *2019 International Conference on Computer and Information Sciences (ICCIS)*, Sakaka, Saudi Arabia, 2019, pp. 1-5, https://10.1109/ICCISci.2019.8716394.
13. Jaya, Mindra & Tantular, Bertho & Andriyana, Yudhie. (2019). A Bayesian approach to multicollinearity problem with an Informative Prior. Journal of Physics: Conference Series. 1265. 012021. https://10.1088/1742-6596/1265/1/012021.
14. Bayman, Emine Ozgur PhD*; Dexter, Franklin MD, PhD, FASA†. Multicollinearity in Logistic Regression Models. Anesthesia & Analgesia 133 (2): p 362-365, August 2021. https://10.1213/ANE.0000000000005593.
15. Ashari, Ahmad & Paryudi, Iman & Tjoa, A Min. (2013). Performance Comparison between Naïve Bayes, Decision Tree and k-Nearest Neighbor in Searching Alternative

Design in an Energy Simulation Tool. International Journal of Advanced Computer Science and Applications. 4. https://10.14569/IJACSA.2013.041105.
16. Cherian, V. A. (2017). Heart Disease Prediction Using Naïve Bayes Algorithm and Laplace Smoothing Technique.
17. Kalcheva, N., Todorova, M & Marinova,g. "NAIVE BAYES CLASSIFIER, DECISION TREE AND ADABOOST ENSEMBLE ALGORITHM – ADVANTAGES AND DISADVANTAGES," in The 6th International Scientific Conference, 2020. https://doi.org/10.31410/ERAZ.2020.153.
18. Hu, Can & Zhang, Chenmeng & Zhang, Zongxi & Xie, Shijun. (2021). Comparative Study on Defects and Faults Detection of Main Transformer Based on Logistic Regression and Naive Bayes Algorithm. Journal of Physics: Conference Series. 1732. 012075. https://10.1088/1742-6596/1732/1/012075.
19. V. Sai Ram Kumar, & Shri Vindhya A. (2022). An Improved Efficiency in Envisioning the Personage Traits over Online Social Media based on Indian Metrics during Pandemic using Novel Naive Bayes Classifier Algorithm Comparing with Logistic Regression Algorithm. Journal of Pharmaceutical Negative Results, 713–722. https://doi.org/10.47750/pnr.2022.13.S04.081.
20. Ashraf, Tahira & Hanif, Asif & Naing, Nyi Nyi & Nadiah, Wan Arfah. (2021). A Comparative Review of Data Mining Techniques for Prediction of Risk Factors of Low Birth Weight. Pakistan Journal of Medical and Health Sciences. 14. 724-727.
21. Chang, M. (2019). On Improving Performance of the Binary Logistic Regression On Improving Performance of the Binary Logistic Regression Classifier. University of Nevada, Las Vegas, Department of Mathematical Sciences. Las Vegas: UNLV Theses, Dissertations. https://dx.doi.org/10.34917/18608608.
22. Siddiqi, N. (2017). Intelligent Credit Scoring: Building and Implementing Better Credit Risk Scorecards (2nd ed.) ISBN: 978-1-119-27915-0
23. Ana M. Aguilera, Manuel Escabias, Mariano J. Valderrama, Using principal components for estimating logistic regression with high-dimensional multicollinear data, Computational Statistics & Data Analysis, Volume 50, Issue 8,2006, Pages 1905-1924, https://doi.org/ 10.1016/j.csda.2005.03.011.
24. Asar, Y. (2017). Some new methods to solve multicollinearity in logistic regression. Communications in Statistics - Simulation and Computation, 46 (4), 2576-2586. https://10.1080/03610918.2015.1053925.
25. Montgomery, D., Peck, E., & Vining, G. (2001). Introduction to Linear regression (3rd ed.). New York: Wiley.
26. Abuassba, A., Zhang, D., Luo, X., Shaheryar, A., & Ali, H. (2017). Improving Classification Performance through an Advanced Ensemble Based Heterogeneous Extreme Learning Machines. Computational Intelligence and Neuroscience. https://doi.org/10.1155/2017/3405463.
27. Ehwerhemuepha, L., & Rakovski, C. (2019, November 7). A comprehensive assessment of automatic logistic regression model selection methods. https://doi.org/10.21203/rs.2.169-60/v1.
28. World Health Organization. (2016). Consolidated guidelines on the use of antiretroviral drugs for treating and preventing HIV infection: recommendations for a public health approach, 2nd ed. World Health Organization. https://apps.who.int/iris/handle/10665/208825.
29. United Nations Programme on HIV/aids. [UNAIDS]. (2021). UNAIDS data 2021. https://www.unaids.org/en/resources/documents/2021/2021_unaids_data.
30. Frescura L, Godfrey-Faussett P, Feizzadeh A. A, El-Sadr W, Syarif O, Ghys PD, et al. (2022) Achieving the 95 95 95 targets for all: A pathway to ending AIDS. PLoS ONE 17 (8): e0272405. https://doi.org/10.1371/journal.pone.0272405.
31. Uganda AIDS Commussion [UAC]. (2020). The National HIV And AIDS Strategic Plan 2020/21–2024/25 (Issue August). https://uac.go.ug/index.php?option=com_content&view=article&id=24:hiv-prevention-1123&catid=8&Itemid=101.

32. Carrillo, H., Brodersen, K.H., Castellanos, J.A. (2014). Probabilistic Performance Evaluation for Multiclass Classification Using the Posterior Balanced Accuracy. In: Armada, M., Sanfeliu, A., Ferre, M. (eds) ROBOT2013: First Iberian Robotics Conference. Advances in Intelligent Systems and Computing, vol 252. Springer, Cham. https://doi.org/10.1007/978-3-319-03413-3_25.
33. Wiharto W, Kusnanto H, Herianto H. Interpretation of Clinical Data Based on C4.5 Algorithm for the Diagnosis of Coronary Heart Disease. Healthc Inform Res. 2016 Jul; 22 (3): 186-95. https://10.4258/hir.2016.22.3.186.
34. Mehdiyev, N., Enke, D., Fettke, P., & Loos, P. (2016). Evaluating Forecasting Methods by Considering Different Accuracy Measures. Procedia Computer Science, 95, 264–271. https://doi.org/10.1016/j.procs.2016.09.332.
35. Wang, Q. (2014). A hybrid sampling SVM approach to imbalanced data classification. Abstract and Applied Analysis, 2014. https://doi.org/10.1155/2014/972786.
36. Brodersen, Kay H. & Mathys, Christoph & Chumbley, Justin & Daunizeau, Jean & Ong, Cheng Soon & Buhmann, Joachim & Stephan, Klaas. (2012). Bayesian Mixed-Effects Inference on Classification Performance in Hierarchical Data Sets. Journal of Machine Learning Research. 13. 3133-3176. https://10.5167/uzh-71594.
37. Bbosa, F. Fuller., Wesonga, Ronald., Nabende, Peter., & Nabukenya, Josephine. (2021). A Modified Decision Tree and Its Application to Assess Variable Importance. 2021 4th International Conference on Data Science and Information Technology, 468–475. https://doi.org/10.1145/3478905.3479245.
38. Bbosa, F.F., Nabukenya, J., Nabende, P. et al.On the goodness of fit of parametric and non-parametric data mining techniques: the case of malaria incidence thresholds in Uganda.Health Technol.11, 9 29–940 (2021). https://doi.org/10.1007/s12553-021-00551-9
39. Kelleher, John; Mac Namee, Brian; D'Arcy, A. (2020). undamentals of Machine Learning for Predictive Data Analytics Algorithms, Worked Examples, and Case Studies. (2nd ed.). Cambridge : MIT Press.
40. Wei, Q., & Dunbrack, R. (2013). The Role of Balanced Training and Testing Data Sets for Binary Classifiers in Bioinformatics. Plos One, 8 (7). https://10.1371/journal.pone.0067863.
41. Ministry of Health [MOH]. (2018). National Health Facility Master List 2018. In Ministry of Health Uganda (Issue November). http://library.health.go.ug/health-infrastructure/health-facility-inventory/national-health-facility-master-facility-list-2018.
42. METS. (2022). UgandaEMR User Manual. https://mets-programme.gitbook.io/ugandaemr-documentation/#ugandaemr-user-manual
43. Ministry of Health [MoH]. (2016). Consolidated guidelines for prevention and treatment of HIV in Uganda. (Issue December). https://www.prepwatch.org/wp-content/uploads/2017/08/consolidated_guidelines_hiv_prevention_uganda.pdf
44. Bhaya, Wesam. (2017). Review of Data Preprocessing Techniques in Data Mining. Journal of Engineering and Applied Sciences. 12. 4102-4107. https://10.3923/jeasci.2017.4102.4107.
45. Chih-Fong Tsai, Yu-Chi Chen, The optimal combination of feature selection and data discretization: An empirical study, Information Sciences, Volume 505, 2019, Pages 282-293, ISSN 0020-0255, https://doi.org/ 10.1016/j.ins.2019.07.091.
46. Li, G., Zhou, X., Liu, J., Chen, Y., Zhang, H., Chen, Y., … Nie, S. (2018). Comparison of three data mining models for prediction of advanced schistosomiasis prognosis in the Hubei province. PLoS Neglected Tropical Diseases, 12(2), 1–19. https://doi.org/10.1371/journal.pntd.0006262.
47. Hosmer, David; Lemeshow, Stanley; Sturdivant, R. (2013). Applie Logistic regression. John Wiley & Sons, Inc. https://doi.org/10.1002/9781118548387.
48. Smith, G. (2018). Step away from stepwise. Journal of Big Data, 5 (32). https://doi.org/10. 1186/s40537-018-0143-6.
49. Hwang, J., & Hu, T. (2014). A stepwise regression algorithm for high-dimensional variable selection. Journal of Statistical Computation and Simulation, 85 (9),

1793–1806. https://doi.org/10.1080/00949655.2014.902460.
50. R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
51. Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2021). dplyr: A Grammar of Data Manipulation. R package version 1.0.7. https://CRAN.R-project.org/package=dplyr.
52. Max Kuhn (2021). caret: Classification and Regression Training. R package version 6.0-90. https://CRAN.R-project.org/package=caret.
53. Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez and Markus Müller (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics, 12, p. 77. DOI: 10.1186/1471-2105-12-77<http://www.biomedcentral.com/1471-2105/12/77/>
54. Sing T, Sander O, Beerenwinkel N, Lengauer T (2005). "ROCR: visualizing classifier performance in R."Bioinformatics_, *21*(20), 7881. <URL: http://rocr.bioinf.mpi-sb.mpg. de>.
55. Mosley, L. (2013). A balanced approach to the multi-class imbalance problem [Iowa State University]. In University Library, IoWA. https://doi.org/10.31274/etd-180810-3375.
56. Wahba, Yasmen & Elsalamouny, Ehab & Eltaweel, Ghada. (2015). Improving the Performance of Multi-class Intrusion Detection Systems using Feature Reduction. International Journal of Computer Science Issues. https://doi.org/10.48550/arXiv.1507.06692.
57. Hoque, N., Singh, M. & Bhattacharyya, D.K. EFS-MI: an ensemble feature selection method for classification. Complex Intell. Syst. 4, 105–118 (2018). https://doi.org/10.1007/s40747-017-0060-x.
58. Gao, Xiang & Wen, Junhao & Zhang, Cheng. (2019). An Improved Random Forest Algorithm for Predicting Employee Turnover. Mathematical Problems in Engineering. 2019. 1-12. https://10.1155/2019/4140707.
59. Maina EK, Mureithi H, Adan AA, Muriuki J, Lwembe RM, Bukusi EA. Incidences and factors associated with viral suppression or rebound among HIV patients on combination antiretroviral therapy from three counties in Kenya. Int J Infect Dis. 2020 Aug;97:151-158. https://10.1016/j.ijid.2020.05.097.
60. Shiferaw, M.B., Endalamaw, D., Hussien, M. et al. Viral suppression rate among children tested for HIV viral load at the Amhara Public Health Institute, Bahir Dar, Ethiopia. BMC Infect Dis 19, 419 (2019). https://doi.org/10.1186/s12879-019-4058-4
61. Nabukeera S, Kagaayi J, Makumbi FE, Mugerwa H, Matovu JKB. Factors associated with virological non-suppression among HIV-positive children receiving antiretroviral therapy at the Joint Clinical Research Centre in Lubowa, Kampala Uganda. PLoS One. 2021 Jan 27; 16 (1): e0246140. https://10.1371/journal.pone.0246140.
62. Opoku, Stephen & Sakyi, Samuel & Ayisi-Boateng, Nana Kwame & Enimil, Anthony & Senu, Ebenezer & Owusu Ansah, Richard & Aning, Bismark & Ojuang, Diana & Wekesa, Doreen & Ahmed, Fatima & Okeke, Chidinma & Sarfo, Ama. (2022). Factors associated with viral suppression and rebound among adult HIV patients on treatment: a retrospective study in Ghana. AIDS Research and Therapy. 19. https://10.1186/s12981-022-00447-2.
63. Yadav, S., & Shukla, S. (2016). Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification. International Conference on Advanced Computing, (6). https://doi.org/10.1109/IACC. 2016.25.