Machine Learning Model for Prediction and Visualization of HIV Index Testing in Northern Tanzania

Happyness Chikusi, Dr Judith Leo, Dr Shubi Kaijage
School of Computational and Communication Science and Technology
Nelson Mandela African Institution of Science and Technology, Arusha, Tanzania

***Abstract*—Human Immunodeficiency Virus Acquired Immunodeficiency Syndrome (HIV AIDS) in Tanzania is still a threatening disease in society. There have been various strategies to increase the number of people to know their HIV status. Among these strategies, HIV index testing has proven to be the best modality for collecting the number of HIV contacts who might be at risk of contracting HIV from an HIV-positive person. However, the current HIV index testing is manual-based, creating many challenges, including errors, time-consuming, and expensive to operate. Therefore, this paper presents the Machine Learning model results to predict and visualise HIV index testing. The development process followed the Agile Software development methodology. The data was collected from Kilimanjaro, Arusha and Manyara regions in Tanzania. A total of 6346 samples and 11 features were collected. Then, the dataset was divided into training sets of 5075 samples and a testing set of 1270 samples (80/20). The datasets were run into Random Forest (RF), XGBoost, and Artificial Neural Networks (ANN) algorithms. The results of the evaluation, by Mean Absolute Errors (MAE), showed that; RF MAE (1.1261), XGBoost MAE (1.2340), and ANN MAE (1.1268.); whereby the RF appeared to have the best result compared to the other two algorithms. Data visualisation shows that 17.4% of males and 82.6 of females had been notified. In addition, the Kilimanjaro region had more cases of people with HIV status from their partners. Overall, this study improved our understanding of the significance of ML in the prediction and visualisation of HIV index testing. The developed model can assist decision-makers in coming out with a suitable intervention strategy towards ending HIV AIDS in our societies. The study recommends that health centres in other regions use this model to simplify their work.**

Keywords*—Index testing; machine learning; random forest; XGBoost; artificial neural network

I. Introduction

Index testing refers to a case-finding strategy that aims to get the exposed contacts of HIV Positive individuals for HIV-testing services. It is also known as partner notification [1]. This person is known as an indexing client. Healthcare workers and counsellors ask index clients to list all their partners, including sexual partners and or injecting drugs partners and their children. The process is voluntary and confidential. In the process, each partner and the children are contacted and informed on the exposure to HIV and offered voluntary testing. The purpose of index testing is to break the chain of HIV transmission. In addition, health workers provide HIV testing to the people who have been exposed to HIV[2]. If the result is positive, they are linked to the treatment, and if the status is negative, they are given prevention services.

There are various HIV testing modalities such as Voluntary HIV counselling and Testing (VCT), Community VCT home-based, mobile, and outreach testing. However, home-based and mobile outreach is costly. Therefore, index case testing was introduced to increase the number of people to know their status, and it has been a promising strategy towards the maximisation of HIV case detection [3]. The current HIV index client testing system does not have an automated system, and the data are collected manually. Therefore, it is challenging to analyse the data and predict HIV index testing. In addition, it requires expertise for data entry and data analysts to do the work. Hence, resulting in additional cost and time-consuming in obtaining the intended results. Not only that but also human errors are unavoidable.

Other researchers applied machine learning in health care specific to HIV AIDS solving various problems like HIV case findings, HIV predictors, Patient-specific current CD_4 count, and prediction of new HIV Index using internet data, however, the methods used were statistical descriptive, estimated index and chi-square.

Therefore, this paper presents the results of the developed Machine Learning model that can help experts make predictions and produce up-to-date data visualisation that is readable and understandable. In addition, the developed Machine-learning model can predict the number of HIV Index testing using partner notification information to identify people who are at risk to contract HIV AIDS. Hence, help decision-makers to come out with a good intervention strategy towards ending HIV AIDS in our societies.

The paper consists of five parts namely: introduction, literature review, material and methodology, results and discussion, conclusion and recommendation.

II. Literature Review

A. Overview of Literature Survey

HIV is an infectious disease that threatens public health globally. According to World Health Organization (WHO), 38% million people are living without HIV globally. 19% do not know their status[4]. Many people living with HIV are

located in middle and low-income countries, with an estimated 68% in sub-Saharan Africa.

WHO Strategy of 2016 to 2021 addresses human rights and equity, with a radical decline in a new HIV infection and reducing death. The global target is to reduce new infection to less than 500,000 by 2020 and end HIV by 2030 as a public threat. The current target is 90 90 90, meaning that 90% know their status, 90% receives quality treatment and care, and the last 90% retain in extended care [5][6].

In the southern part of Africa, various countries have made substantial progress toward the HIV/AIDS Program target of ensuring that 90% of people living with HIV know their status. HIV testing and counselling is the crucial step towards achieving the Joint United Nations Program on HIV/AIDS (UNAIDS) of 90 90 90. However, the target for 2025 is 95% 95% 95%[7].

B. HIV Trends in Tanzania

The HIV status in Tanzania shows that 1.7million people live with HIV, 77,000 new HIV infections, and 27,000 AIDS-related death[8]. The new strategic plan reviewed by the Ministry of Health for 2018 to 2022 is making Index testing Services and Partner Notification services one of the National Strategy for Identification of the People Living with HIV (PLHIV)[9]. Index case testing will support Tanzania to maximise HIV case detection in achieving the first target of 90 (2017- 2022) and the next of 95 for 2025 for males, adolescents, and children.

C. Machine Learning in Health Care

Machine learning(ML) is the use and development of computer systems that can learn and adapt without following explicit instructions use algorithms to analyse and draw inferences from the pattern in data[10]. Machine learning algorithms depend on domain knowledge of the data to create features that make these algorithms work. ML has been used in various domains with data availability, including computer vision, automatic speech recognition, business analysis, natural language processing, and even health care. However, the process demands lots of time and effort for feature selection, and features must extract relevant information from vast and diverse data to produce the best outcome.

Machine learning techniques accurately provide predictions in various applications, such as drug discovery and disease diagnosis, especially with quality data. Machine learning interest is in cancer diagnosis, diabetes, autism subtyping in health care[11]. Also, ML is used to predict cholera disease [12].

Machine learning in HIV/AIDS had applied as follows: Machine learning to identify HIV predictors for screening [13]. Machine learning in the prediction of patient-specific current CD_4 cell count to determine the progression of human immunodeficiency.[14] Prediction of new HIV infection in China by using internet search [15], predicting default from HIV service in Mozambique [16], Another area is improving HIV case findings [17].

Other related works predict HIV index Testing using different methods are Index and target community testing to optimise HIV case findings among men. The process used descriptive statistics, estimated index cascade, and Chi-Square test.[18] Sustained high HIV case finding through Index testing via services register using Microsoft excel.[19] Another study done was about applying machine learning on HIV/AIDS diagnosis and therapy planning[20].

Therefore, this study aims to use machine-learning techniques to predict HIV index testing and visualisation items of Age, Sex, location, and relationship to strengthen the ability to plan, prioritise, and implement the effective intervention.

III. Materials and Methods

A. Materials

The study area selected was the northern part. The Northern party regions include Tanga, Kilimanjaro, Arusha, and Manyara. Kilimanjaro, Arusha, and Manyara had chosen to represent the party. The dataset used in this study was from different health centres and community sites from Arusha, Kilimanjaro, and Manyara. The client information consists of 6346 samples and 11 features, and index-client data consists of 7226 samples with 13 elements.

Python was the programming language used in this study. The reason that led to this programming language took into consideration its ability to offer a variety set of open-source libraries to support machine learning.

B. Methoderations

1) Knowledge discovered from data science: Knowledge Discovered from Data (KDD) is extracting knowledge from various vast quantities of data. In carrying out this study, we selected this approach due to its application in data mining using different algorithms and clearly defined phases. [21]. The study followed an interactive refine at each step (Table I) explains the stages of KDD.

TABLE I. Summary of KDD

NAME	EXPLANATION
Problem Identification	It is the bedrock of all the stages. It involves the study to understand the topic. Simple to have domain knowledge.
Data preprocessing	In this stage, the data are identified and selected. It has data sampling, cleansing, and reduction. This stage is necessary to remove the dirty/noise data and outliers to improve quality.y
Data Mining	This process involved selecting machine learning techniques that will be used to create a model and come out with the desired outcome .e
Pattern Interpretation	Focus on checking the performance of the developed model. It is just a model evaluation process.
Model deployment	This stage is for putting the model to use.

2) Data preprocessing: Data preprocessing is a process that involves various techniques like data selection, data cleaning, data integrations, data reduction, and data transformation. For example, the collected dataset from Kilimanjaro, Arusha, and Manyara had two types of data set the first dataset of client information of 6346 samples with 11 features, and the client index information of 7226 samples with 13 features. The features of client information included client_id, Date, Sex, Age, Residence, contact no, CTC_no, Position, Marital status, HIV knowledge, and the number of HIV indexes. The client index features include client_id, client-name, Contact number, CTC _number, Position, Registration type, Date, Site name, Region, Sex of _index, Age, Type of relationship, and HIV status. Based on the nature of the data and literature review, the preprocessing techniques performed as follows:

The selected dataset was cleaned by ignoring features with no value, and the most occurring feature-filled the missing values; the duplicated value was identified and cleared. The dataset used to make predictions was the client information. Later the two datasets were combined for data visualisation.

Data reduction was made on the following features: client id, date, residence, contact number, and CTC number. The removal was due to the following reasons; the features had no impact on the target (client id, contact no, and CTC number). The features had no values (date and residence). Lastly, the data was transformed into a suitable format for model development. The categorical data were converted into 1 and 0, respectively.

3) Data visualization: Data Visualization refers to the graphical presentation of the analysed data so a user can get insight from it and make decisions [22][23]. Data exploration was done using python.

4) Machine learning algorithms: There are different ways of solving ML problems. ML can be divided into three major parties: Supervised, unsupervised, and reinforcement. Each model may apply algorithms based on the dataset and intended results [24]. Machine learning models are designed to classify things, predict outcomes, find patterns and make informed decisions.

Based on this study, three algorithms were selected for performance comparison to determine the best algorithm for predicting the number of HIV Index testing (based on literature). These algorithms were XGBoost, Random Forest (RF), and Neural network. The study considered all the three ML algorithms to select the best performing. Therefore, these algorithms are explained hereunder.

a) XGBoost: XGBoost is an ensemble algorithm based on gradient boosting that has been explained to be an efficient and reliable machine learning technique in solving challenges[25]. It is an open-source library that works best in speed, performance, and parameter setup[26]. XGBoost is used in classification and regression predictive modelling problems. XGBoost denotes the best algorithm for competition on the Kaggle [27].

b) Random Forest: Random forest is an ensemble learning technique that uses a network of decision trees. Breiman proposed it in 2001. It is used for classification and regression[28][29]. The random forest technique combines various randomised decision trees. It is applied in larger-scale problems. Random sampling enhances the depreciation of the overfitting problem [24]. The randomly generated dataset is used to train the dataset for the ensemble decision tree. Each decision tree will determine output. Fig. 1 below shows how a random forest algorithm is formed.

c) Artificial Neural Network: The artificial neural network, usually called a neural network, is defined as an interconnection of nodes called neurons[30]**.** It works like the human brain works. A collection of neurons created and connected together enables them to send messages to each other. The network is requested to solve a problem, which is performed repeatedly. The more connection is strengthened, the more success is achieved, and the reduced failure.

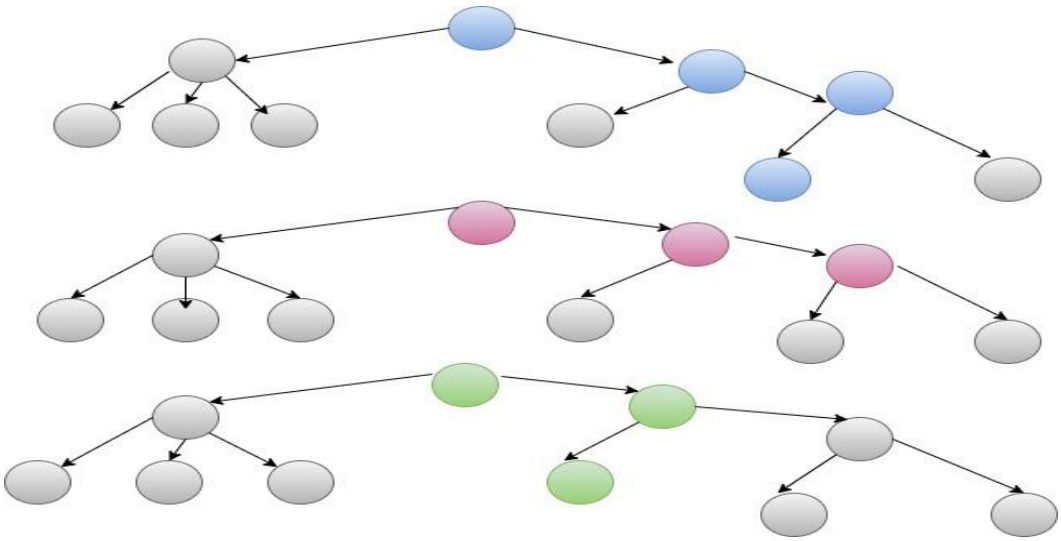

Fig. 1. Random Forests Structure.

The input variables from the data are passed to this neural as a linear connection of various variables. The value multiplied by each characteristic variable is called weight. Now the linear link is applied to nonlinear combinations to provide the ability of nonlinear relationships for neural network modelling. It is used in both classification and regression problems. The artificial neural network is trained by using a random gradient (SGD) and backpropagation algorithm[31]. Fig. 2 shows the structure of an artificial neural network. Each neuro in the input layer represents a column in the input data. Input data is fed to set of neurons and each produces output. Again, each of output is fed to other neuro, which produces another output, which is again fed to the output layer. Error is calculate at this final output layer and again sent back to network for further refine of the output of each neuro. The process is repetitively until the minimal error is obtained.

5) Experimental procedures: The development of models involves major tasks: Acquiring datasets, preprocessing, feature engineering, and model selection. Fig. 3 shows the summary of how the experimental procedures were carried out.

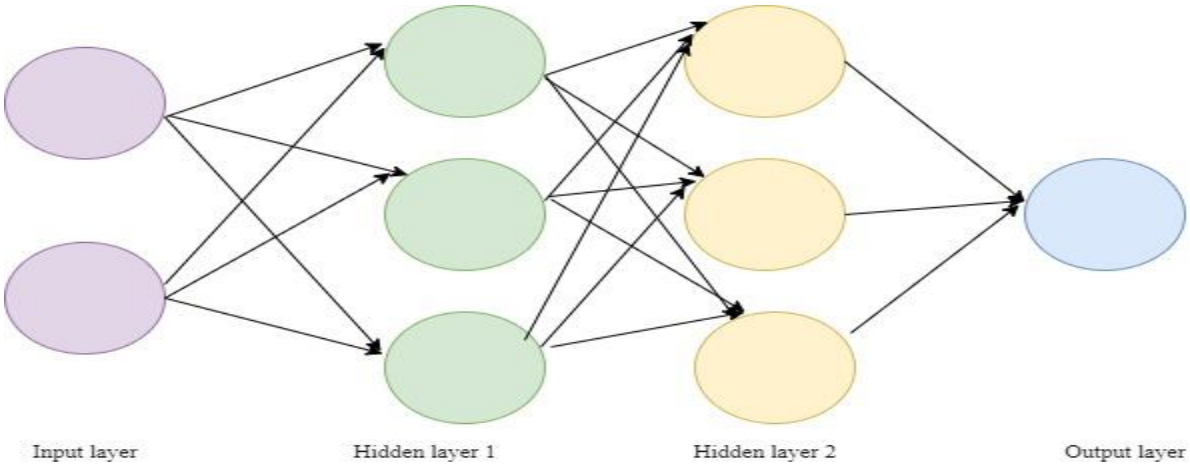

Fig. 2. Neural Network Structure.

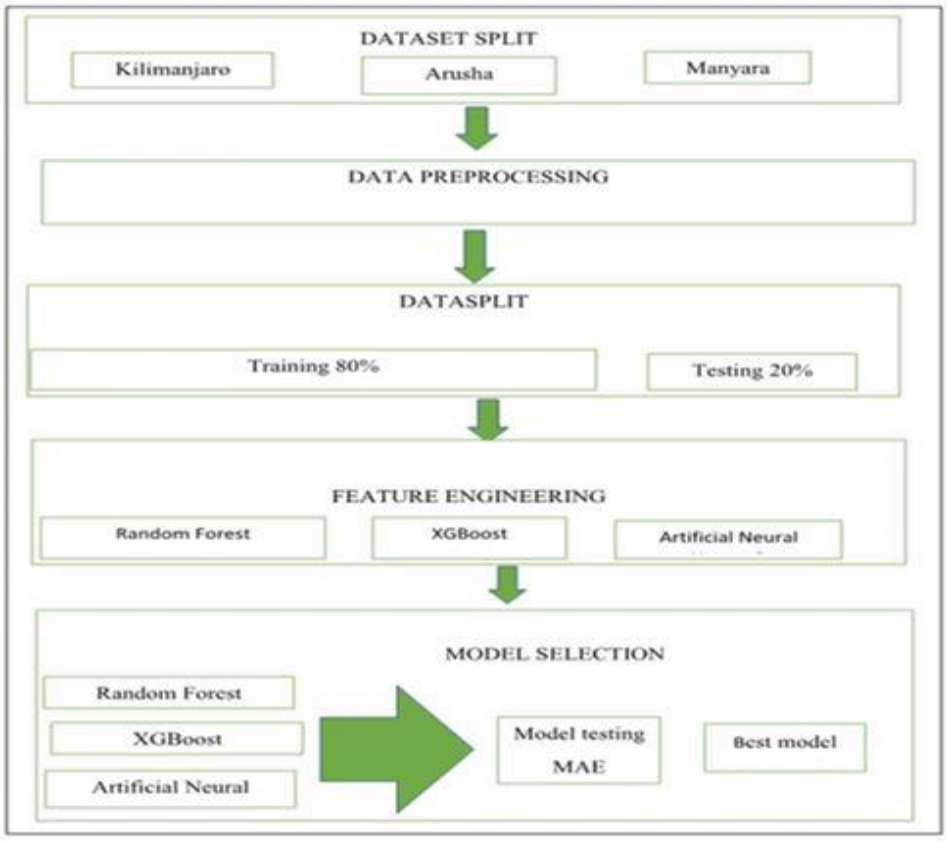

Fig. 3. Experimental Procedures.

6) Evaluation metric: Model evaluation refers to choosing the best-performed model representing data and determining how well the model will work in unseen data. There is a wide variety of evaluation metrics for regression models [32]. According to the literature review, the most used metrics are Mean Squared Error (MSE), RMSE, and MAE. The MSE is calculated as the mean or average squared differences between actual output and predicted target values in a dataset. RMSE is an extension of the mean squared error. MAE score is calculated as the average of the absolute error. In this study, the metric used was MAE due to its simplicity and understandability.

IV. Results and Discussion

A. Results

The subsection explains the results obtained towards developing the HIV index-testing model.

1) Feature engineering: Experiment result from feature engineering showed that people with no knowledge of HIV has a strong coefficient of (0.5). Followed by Marrital_status married (0.175), Age (0.15), Female gender (0.1), Position (influence of someone in the society (0.1), and the rest has a coefficient of less than (0.1). Fig. 4 provides more visualisation of the extracted features using the random forest algorithm. Table II explains in detail the components selected for model development.

2) Data visualization: The section depicts the insight of data from different angles of view. Fig. 5 shows the number of HIV index per client-id. Fig. 6 illustrates the number of HIV indexes by status and site.

Fig. 7 visualise the HIV index versus HIV status and type of relationship. Fig. 8 and Fig. 9 show the number of HIV Index by each region and distributions in term of Age.

3) Model development and evaluation: The result obtained from Model development using three algorithms, as shown in Table III indicates that Random Forest performed well compared to the other two by having the smallest value of MAE: The smaller the value, The desired model.

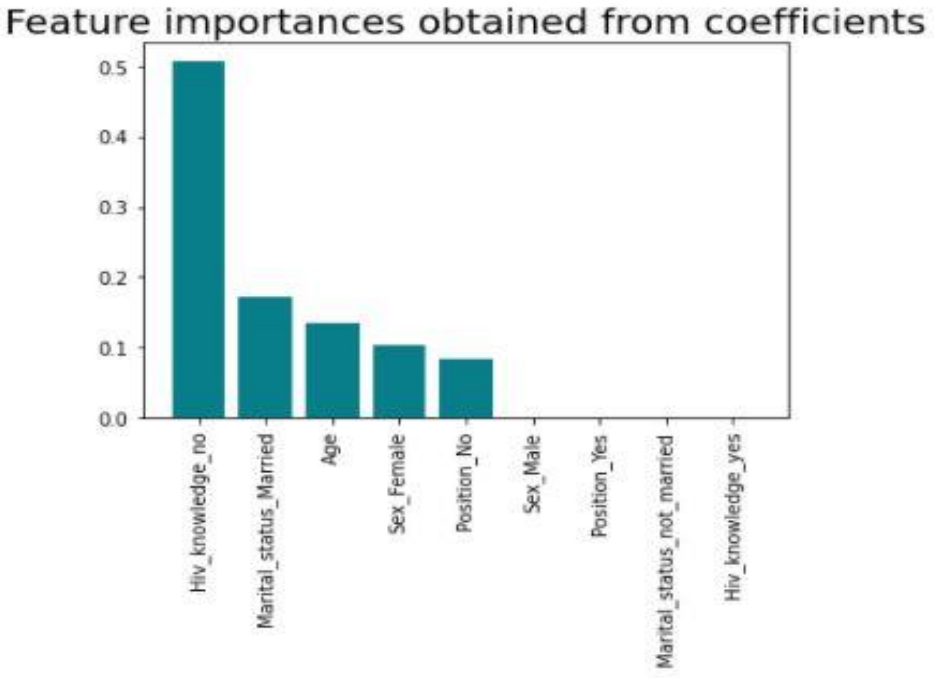

Fig. 4. Feature Engineering.

TABLE II. Selected Feature for Model Development

Variable	Description	Measurement
Age	Age of HI positive client.	number
Sex	Gender of the client (Male/ Female)	1/0
Position	The client is influential in society. Leadership (political, religious, and traditional) values Yes/No	0/1
Marital status	Not married(divorce,widow,widower,never_married)/ married	0/1
HIV_Knowledge	Awareness of client on HIV/AIDS(yes/no)	0/1

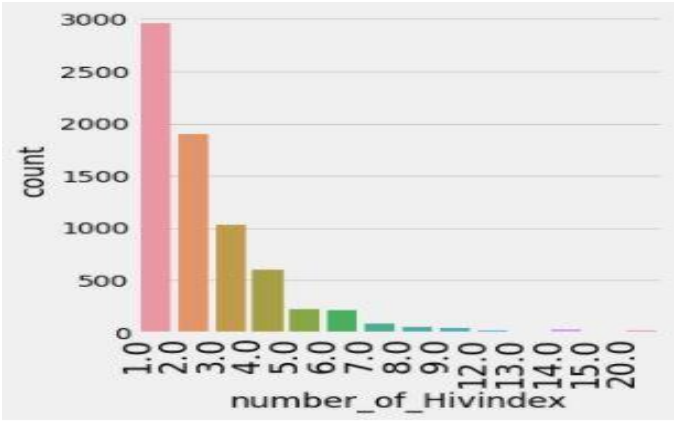

Fig. 5. Number of HIV Index Contacts for each Client_id (Source Google Collab).

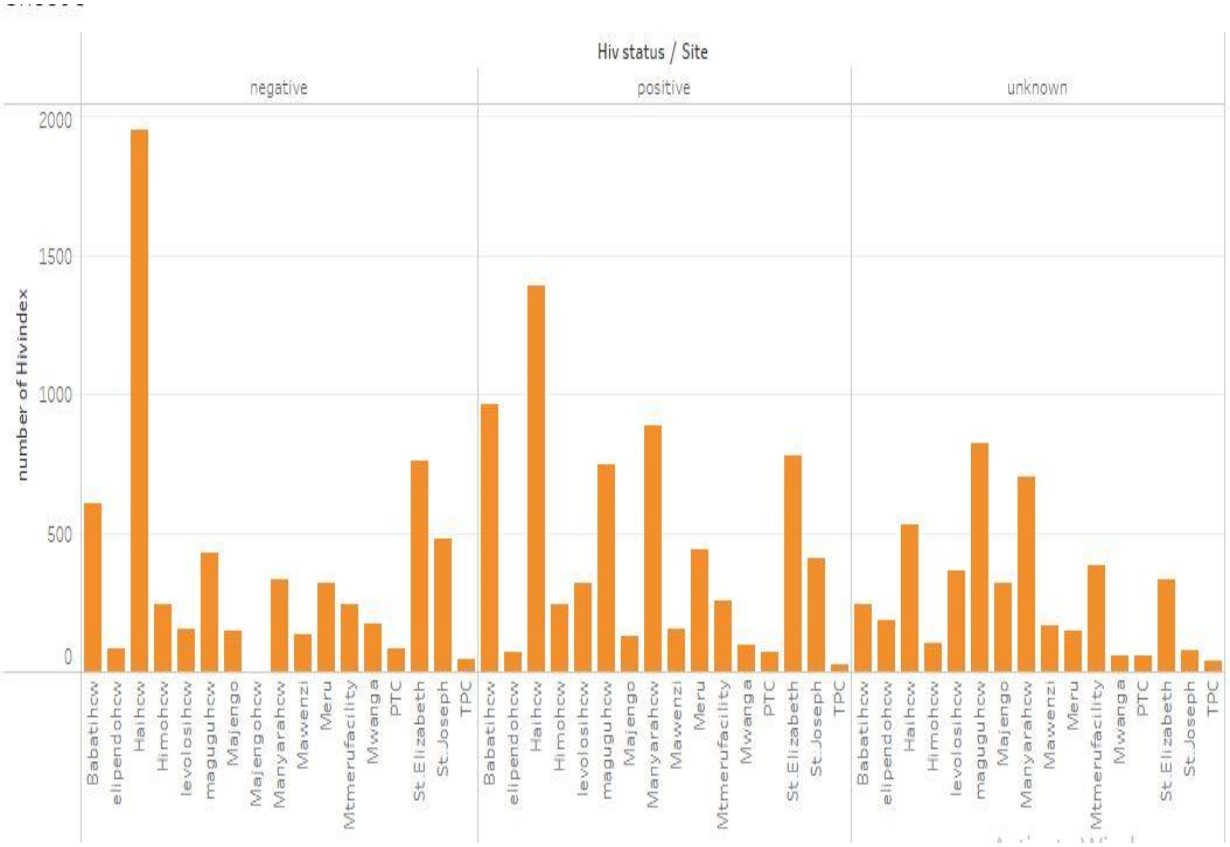

Fig. 6. Number of HIV Index by Status and Site.

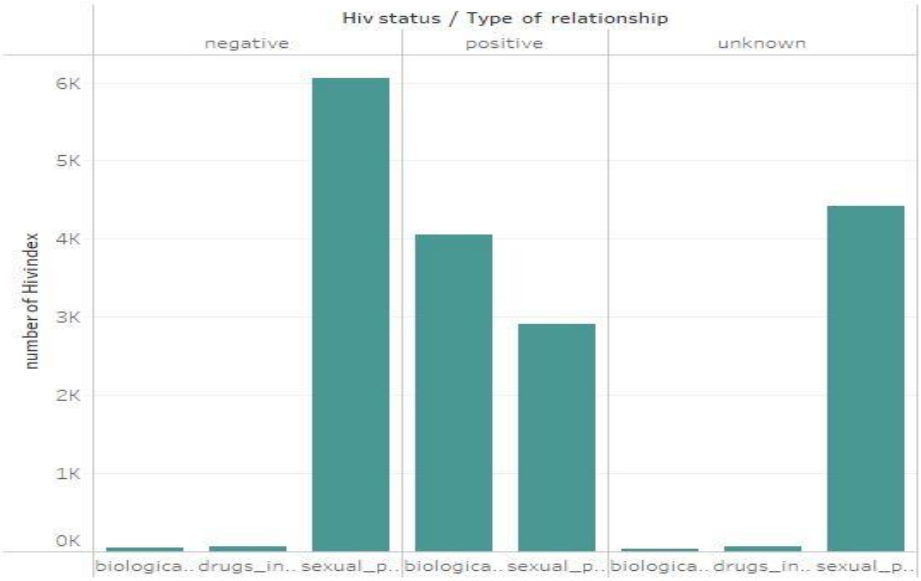

Fig. 7. Number of HIV Index versus HIV Status and Type of Relationship.

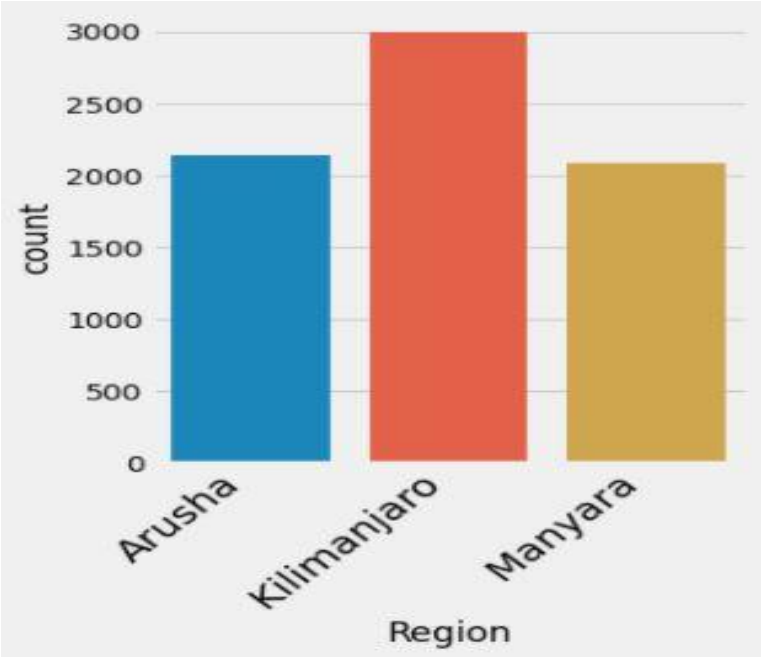

Fig. 8. Diagram shows the Total HIV Indexes in each Region.

number_of_Hivindex distribution across Region by Sex_of_Index

Region ↓	Female	Male	Total number_of_Hivindex
Arusha	81.1%	18.9%	100.0%
Kilimanjaro	83.1%	16.9%	100.0%
Manyara	83.4%	16.6%	100.0%
Grand Total:	**82.6%**	**17.4%**	**100.0%**

Fig. 9. Distribution across the Region.

TABLE III. RESULT OBTAINED DURING MODEL DEVELOPMENT

Serial number(S/N)	Model name	Metric (MAE)
1	Random Forest	1.1261
2	XGBoost	1.2340
3	ANN	1.1268

B. Discussion

The process of understanding the domain knowledge was done thoroughly. Various methods were used to solve the problem to a specified domain. The feature that had a high contribution to the target value was identified. People with no knowledge had led a client to have many client notifications by 35% followed by Position of the client in society 15%, marital status 14%, Age 10%, and Sex 8%.

Data visualisation shows that many clients refer to only one person followed by two and three, while few had up to 12 to 20 people. Kilimanjaro region had high returns compared to the two and a good HIV index specific to the Hai site. The sexual partner notification had a high percentage in information followed by biological children.

The best performance algorithm was a random forest. It had the smallest value of Mean Absolute Error (MAE) of 1.1261. The result remained unchanged after improving the model using the best parameters by GridSearchCV. Lastly, the model was saved ready for deployment.

V. CONCLUSION AND RECOMMENDATION

A. Conclusion

Machine learning is an essential skill in current days. Health care is widely used in many ways, such as decision support, developing medical care guidelines, and applying them in detecting diseases. This paper used machine learning to predict the HIV index and visualisation to help decision-makers develop a suitable intervention strategy to end HIV AIDS as a health threat to society.

However, in achieving the main objective, in addressing the specific goal, the study encountered the following limitations; Missing information in health care data, Lack of enough information in health care such as social-economic and social behaviour information. This information could have an impact on the result. Therefore, the model was developed considering the collected data.

B. Recommendation

Tanzania is one of the sub-Saharan countries with a large rate of people living with HIV. Therefore, client partner notification is vital and can help to yield the target of 95 95 95. However, the study recommends that more researchers and development be required to capture all the required data for better results.

Due to the limitation observed the study recommends that the health care system, especially the unit dealing with HIV/AIDS use the automated system and review the data to be collected for both hospitals and stakeholders to facilitate quality data collection. In addition, HIV knowledge awareness should continuously be given to the community of all ages, and areas.

ACKNOWLEDGMENT

I want to extend my special thanks to the Nelson Mandela African Institution of Science and Technology for granting the opportunity to pursue a master's degree, Center of Excellence in ICT in East Africa (CENIT@EA), for supporting studies and Soft Med company internship took place.

REFERENCES

[1] D.Jerene,*W.Abebe, “Hiv Testing Services Hiv Self-Testing and Partner,” no. December, p. 7, 2017.

[2] U. and C. WH0, PEPFAR, “Partner and Family-Based Index Case Testing,” vol. 1, p. 42, 2015.

[3] M. Katbi et al., “Effect of clients Strategic Index Case Testing on community-based detection of HIV infections (STRICT study).,” Int. J. Infect. Dis. IJID Off. Publ. Int. Soc. Infect. Dis., vol. 74, pp. 54–60, Sep. 2018, doi: 10.1016/j.ijid.2018.06.018.

[4] UNAIDS, “Data 2020,” Program. HIV/AIDS, pp. 1–248, 2020, [Online]. Available: https://www.unaids.org/en/resources/documents/2020/unaids-data%0Ahttp://www.unaids.org/sites/default/files/media_asset/20170720_Data_book_2017_en.pdf.

[5] United Nations Joint Programme on HIV/AIDS (UNAIDS), “To help end the AIDS epidemic,” United Nations, p. 40, 2014, [Online]. Available: http://www.unaids.org/sites/default/files/media_asset/90-90-90_en.pdf.

[6] G. Health, S. Strategies, and W. H. Assembly, “Global Health Sector Strategies 2016-2021 (GHSS) Briefing Note : October 2015,” vol. 2021, no. October 2015, pp. 2016–2021, 2016.

[7] B. Y. Putting, P. At, and T. H. E. Centre, “Prevailing Against Pandemics,” 2020.

[8] National Bureau of Statistics, “Tanzania HIV Impact Survey (THIS) 2016-2017,” Tanzania HIV Impact Surv. 2016-2017, no. December 2017, pp. 2016–2017, 2018, [Online]. Available: https://phia.icap.columbia.edu/wp-content/uploads/2019/06/FINAL_THIS-2016-2017_Final-Report__06.21.19_for-web_TS.pdf.

[9] MOHSS, “National Strategic Framework for HIV and AIDS Response in Namibia 2017/18 to 2021/22,” p. 116, 2017.

[10] SAP Insights, “What Is Machine Learning? | Definition, Types, and Examples | SAP Insights.” 2019, [Online]. Available: https://insights.sap.com/what-is-machine-learning/.

[11] J. A. M. Sidey-Gibbons and C. J. Sidey-Gibbons, “Machine learning in medicine: a practical introduction,” BMC Med. Res. Methodol., vol. 19, no. 1, pp. 1–18, 2019, doi: 10.1186/s12874-019-0681-4.

[12] J. Leo, “A reference machine learning model for prediction of cholera epidemics based-on seasonal weather changes linkages in Tanzania.” NM-AIST, 2020.

[13] C. K. Mutai, P. E. McSharry, I. Ngaruye, and E. Musabanganji, “Use of machine learning techniques to identify HIV predictors for screening in sub-Saharan Africa,” BMC Med. Res. Methodol., vol. 21, no. 1, pp. 1–11, 2021, doi: 10.1186/s12874-021-01346-2.

[14] Y. Singh, N. Narsai, and M. Mars, “Applying machine learning to predict patient-specific current CD 4 cell count to determine the progression of human immunodeficiency virus (HIV) infection,” African J. Biotechnol., vol. 12, no. 23, 2013.

[15] Q. Zhang, Y. Chai, X. Li, S. D. Young, and J. Zhou, “Using internet search data to predict new HIV diagnoses in China: A modelling study,” BMJ Open, vol. 8, no. 10, 2018, doi: 10.1136/bmjopen-2017-018335.

[16] PEPFAR, HRSA, DIMAGI, and ICAP, “Machine Learning for Predicting Default from HIV Services in Mozambique.”

[17] P. Smyrnov, Y. Sereda, A. Lytvyn, and O. Denisiuk, “Improving HIV case-finding with machine learning ML algorithm has performed better or equally well in comparison with rule based algorithm on making decision who should receive additional recruitment coupons due to higher probability of undiagnosed HIV ca,” p. 1223, 2016.

[18] L. K. Mwango et al., "Index and targeted community-based testing to optimize HIV case finding and ART linkage among men in Zambia," J. Int. AIDS Soc., vol. 23, no. S2, pp. 51–61, 2020, doi: 10.1002/jia2.25520.

[19] N. Mahachi et al., "Sustained high HIV case-finding through index testing and partner notification services: experiences from three provinces in Zimbabwe," J. Int. AIDS Soc., vol. 22, no. S3, pp. 23–30, 2019, doi: 10.1002/jia2.25321.

[20] S. Prabhakaran, "Machine Learning Methods for HIV / AIDS Diagnostics and Therapy Planning," PhD thesis, 2014.

[21] M. J. Pazzani, "Knowledge discovery from data?," IEEE Intelligent Systems and Their Applications, vol. 15, no. 2. pp. 10–12, 2000, doi: 10.1109/5254.850821.

[22] G. Chawla, S. Bamal, and R. Khatana, "Big Data Analytics for Data Visualization: Review of Techniques," Int. J. Comput. Appl., vol. 182, no. 21, pp. 37–40, 2018, doi: 10.5120/ijca2018917977.

[23] Z. M. Khalid, "Big Data Analysis for Data Visualization : A Review," no. January, 2021, dDOI 10.5281/zenodo.4462042.

[24] H. H. Rashidi, N. K. Tran, H. Abb, E. V. Betts, L. P. Howell, and R. Green, "Artificial Intelligence and Machine Learning in Pathology : The Present Landscape of Supervised Methods," vol. 6, 2019, doi: 10.1177/2374289519873088.

[25] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A Comparative Analysis of XGBoost," no. November 2019, 2019, doi: 10.1007/s10462-020-09896-5.

[26] C. Bent and G. Mart, "A Comparative Analysis of XGBoost A Comparative Analysis of XGBoost," no. November 2019, 2020.

[27] "XGBoost for Regression.".

[28] W. Lin, Z. Wu, L. Lin, A. Wen, and J. I. N. Li, "An Ensemble Random Forest Algorithm for Insurance Big Data Analysis," vol. 5, 2017.

[29] G. Biau and E. Scornet, "A random forest guided tour," TEST, 2016, doi: 10.1007/s11749-016-0481-7.

[30] M. Vakili and M. Rezaei, "Performance Analysis and Comparison of Machine and Deep Learning Performance Analysis and Comparison of Machine and Deep Learning Algorithms for IoT Data Classification," no. January, pp. 0–13, 2020.

[31] R. Suman, "Understanding Artificial Neural Network With Linear Regression." 2019, [Online]. Available: https://analyticsindiamag.com/ann-with-linear-regression/.

[32] J. Brownlee, "Regression Metrics for Machine Learning," Machine Learning Mastery. 2021, [Online]. Available: https://machinelearningmastery.com/regression-metrics-for-machine-learning/.