JCSDA, Vol. 02, No. 01, 63–82
DOI: 10.69660/jcsda.02012504
ISSN 2959-6912

Using Public Health datasets to predict one's ability to pay for Pre-Exposure prophylaxis (PrEP) services in Uganda

Racheal Nasamula*, Baker Lwasampijja, Louis Henry Kamulegeya, Joan Atuhaire, Umuhoza Natasha, Dhikusoka Flavia, Jonathan Ogwal, Sssenkumba Joseph, Ivan Kagolo, Happy Banonya, Brenda Kabakaari, and John Mark Bwanika

African Center of Applied Digital Health Projects and Research Department, Kampala, Uganda
**corresponding author: rachealnasamula2@gmail.com*

In Uganda, the uptake of pre-exposure prophylaxis (PrEP) as a preventive measure against HIV infection is notably low, despite its proven effectiveness, particularly among high-risk populations (UPHIA, 2020). Although PrEP has historically been available at no cost in government facilities, the recent decrease in HIV medication costs and the shift towards private-sector involvement necessitate a reliable assessment of individuals' ability to pay for PrEP. The growing volume of HIV-related data presents a unique opportunity to leverage artificial intelligence (AI) and machine learning (ML) techniques to identify high-risk sub-populations that are both eligible for and willing to pay for PrEP services. This retrospective study, analyzed three diverse datasets, including, the Uganda Demographic Health Survey, the Uganda Population HIV/AIDS Impact Assessment survey, and a private dataset from the Rocket Health Telemedicine Clinic. The study population included individuals aged 18 years and above that have accessed a private health facility for sexual reproductive health services or products. Statistical methods, including the Chi-square test and Spearman's correlation test, were employed to identify features with a statistical significance to the ability to pay for PrEP. The datasets were aggregated, cleaned and then split into 70% for training and 30% for testing and validation. An ensemble of machine learning classification models was trained using Python and the PyCaret library. The AdaBoost classifier demonstrated superior predictive power, with a recall of 99% and an AUC of 100%, indicating robust prediction capabilities on this dataset. The model achieved a high training score of 99%, suggesting an excellent fit to the training data. Further analysis revealed that factors such as age, gender, employment status, and socioeconomic status were the most influential predictors of the ability to pay for PrEP services. A web application interface was developed using the Streamlit library, allowing individuals and programs to upload data and make predictions about the likelihood of individuals paying for PrEP. The developed tool leverages publicly available data to identify populations capable of paying for PrEP services, fostering a collaborative effort towards achieving better health outcomes and ensuring the sustainability of HIV prevention services.

Keywords: Pre-Exposure Prophylaxis (PrEP), Machine Learning, Predictive Modeling, Artificial Intelligence (AI).

1. Introduction

Infectious diseases like HIV continue to be the leading contributors to mortality rates in Uganda despite the concerted efforts to curb its spread [1]. The Key and priority populations have been identified as the most at-risk sub-populations to this public health scourge and account for the majority of new infections [2].

According to the Uganda Population Health Impact Assessment survey (2020),

the prevalence of HIV was highest in 20-24 years, especially among women (4.2%). In regards to demographic characteristics on where new infections are arising from, the Northern (7.6%) and Central regions, specifically greater Masaka (8.1%), had the highest prevalence rates primarily within urban areas (7.1%) [3]. Pre-exposure prophylaxis (PrEP) is one of the proven biological HIV prevention models and is highly effective for preventing HIV, especially for key populations [4]. And It has been suggested that in the absence of a cure and/or vaccine against HIV, acceptance of prevention interventions like pre-exposure prophylaxis (PrEP) needs to be increased rapidly. Despite the proven efficacy of PrEP, coverage has remained suboptimal within the priority populations; by the end of 2020, there were 58,428 people actively taking PrEP [5] and yet the COP 22 targets were 130,005.

Several factors have been highlighted to contribute to the above trends, including lack of consistent access to PrEP (due to stockouts), the capacity of community-based organizations to roll out PrEP services, a limited number of healthcare providers trained to provide PrEP, community stigma against PrEP use, and personal knowledge and beliefs surrounding PrEP [6]. With this unmet need for PrEP service and the need to achieve self-sustainability, there was a need to diversify the models of delivery and access of PrEP services beyond the public sector. This is because this model is largely donor-dependent and unsustainable regarding funding and leaves out a specific clientele demographic that would prefer convenient, paid-for services in a private setting. The majority of key and priority sub-populations have largely been left out of HIV/AIDS prevention programming as they are unknown regarding finer demographic profiles and characteristics. Therefore, our study used machine learning approaches on open datasets to develop predictive models to better profile and understand this underserved sub-population in HIV/AIDS prevention services.

2. Background

In Uganda, despite the proven efficacy of pre-exposure prophylaxis (PrEP) as a highly effective method of preventing HIV infection, the coverage of PrEP remains suboptimal, particularly among sub-populations who are at increased risk of acquiring HIV infection [7]. Several factors contribute to the low uptake of PrEP, including stockouts of PrEP, a limited number of healthcare providers trained to provide PrEP, community stigma against PrEP use, and personal knowledge and beliefs surrounding PrEP [8]. Furthermore, the delivery of PrEP services has been primarily limited to the public sector, which is largely donor-dependent and unsustainable in funding. This has left out sub-populations who prefer convenient, paid-for services in a private setting [8]. These sub-populations, which are often marginalized and face stigma and discrimination, have largely been left out of HIV/AIDS prevention programming and remain largely unknown regarding their demographic profile and characteristics. The unmet need for PrEP among these categories of people is a significant concern. It highlights the need to better profile and understand this

underserved group and involve the private sector in the sustainable and continuous delivery of PrEP services [9].

PrEP is widely considered as an effective prevention intervention for populations at high risk of acquiring HIV and it has been suggested that in the absence of a cure or vaccine against HIV, acceptance of prevention interventions like pre-exposure prophylaxis (PrEP) needs to be increased rapidly [10]. This study identified the specific characteristics and attributes of sub-population that are eligible for PrEP and willing and able to pay for the service. This provided valuable insights into the needs of these populations and informed the development of targeted and effective PrEP programs that are accessible and inclusive for all who need them.

3. Study objectives

3.1. *Primary objective*

To build a Machine Learning model to predict the ability to pay for Pre-Exposure (PrEP) services in Kampala.

3.2. *Secondary objective*

(1) To profile key population subgroups at higher risk for HIV infections. This is to understand and identify the subgroups at high risk of HIV infections by using existing data and sources. This information has provided a baseline to understand the target population and inform the AI model's design.
(2) To develop and deploy an AI modeling technique for PrEP services market segmentation among (key populations) KPs. This aimed at using machine learning and AI techniques to create a model that will predict the likelihood of key population subgroups to take up PrEP services, given the available information about their demographic, behavioral, and socio-economic characteristics. The model is based on the information collected through data sources such as demographic and health surveys, PrEP uptake data, and other relevant data sets.
(3) To roll out and test the validity and reciprocality of the developed AI model with similar datasets from other settings for PrEP service uptake among KPs. After validation, this aims to implement and evaluate the AI model in real-world settings to determine its effectiveness in determining subpopulations eligible for PrEP and willing to pay for the service.

4. Methodology

4.1. *Study design*

This was a retrospective observational study that used already existing datasets to train machine learning models with the ability to identify the high-risk sub-population eligible for pre-exposure prophylaxis (PrEP) and willing to pay for the services.

This retrospective study analyzed three diverse datasets: the Uganda Demographic Health Survey (UDHS) 2019/2020, the Uganda Population HIV/AIDS Impact Assessment (UPHIA) 2016/2017, and a private dataset from the Rocket Health Telemedicine Clinic collected during 2020–2021. The UDHS dataset, managed by the Uganda Bureau of Statistics (UBOS), provides nationally representative demographic and socioeconomic information. The UPHIA dataset focuses on HIV-related health metrics, identifying trends in prevalence and access to services. The Rocket Health dataset includes private sector health service utilization and payment data, particularly related to reproductive and sexual health services.

All datasets were acquired with appropriate permissions, and personally identi?able information was removed. The datasets were anonymized and harmonized before analysis.

4.2. *Study population*

The study used already existing data sets. These included the Uganda Demographic Health Survey (UDHS), Uganda Population HIV/AIDs Impact Assessment (UPHIA) survey data, and private data from Rocket Health. The datasets provided information on health-seeking patterns, especially for HIV services along with parameters of geographical location, age, sex, and socioeconomic status. UDHS is conducted every four to five years and is housed in the Uganda Bureau of Statistics (UBOS). We used the 2019/2020 data set. The UPHIA dataset is collected every four years and was used to analyze indicators of drivers to new HIV infections, including age categories most affected by regions.

The study did not involve human subjects participants; however, open-source data sets that house relevant sociodemographic,clinical, and other critical variables important in determining an individual's ability to utilize health services were included in the secondary data analysis.

4.3. *Participant selection*

The following eligibility criteria were designed to select participants for whom protocol treatment was considered appropriate.

4.3.1. *Inclusion criteria*

Participants who met the following inclusion criteria were eligible for enrollment in the study:

The datasets that were used in the analysis included; Uganda Demographic Health Survey 2019/2020 (UDHS), Uganda Population HIV/AIDS Impact Assessment 2016/2017 (UPHIA), and a private dataset from Rocket Health Telemedicine Clinic.

4.3.2. *Exclusion criteria*

Any other national dataset other than the three selected datasets was excluded.

4.4. *Study duration*

The study lasted for 18 months, during this period we conducted the collection and curation of the data, development of the machine learning model and the testing and validation of the accuracy of these models.

4.5. *Data analysis*

The data analysis involved merging three datasets, that is, Uganda Demographic Health Survey (UDHS), Uganda Population HIV/AIDS Impact Assessment (UPHIA), and Rocket Health private dataset on common keys and standardizing them to minimize measurement bias. The UDHS dataset, collected by the Uganda Bureau of Statistics (UBOS), provided valuable demographic and socioeconomic information. The UPHIA dataset, which focuses exclusively on HIV clients, served as the benchmark for our modeling goal. The Rocket Health dataset included information on individuals' willingness to pay for maternal health services, complementing the data from UDHS and UPHIA.

To create a comprehensive dataset for our analysis, we systematically cleaned and managed the data, addressing issues such as missing data, outliers, and errors. Descriptive statistics, including mean, median, mode, standard deviation, and interquartile range, were used to summarize the distribution of variables related to the willingness and ability to pay for PrEP services.

Univariate analysis methods, such as frequency tables and histograms, were employed to describe the distribution of each variable and identify any outliers or skewness. Bivariate analysis methods, such as cross-tabulation and scatter plots, were used to examine relationships between the willingness and ability to pay for PrEP services and other variables.

4.6. *Model building*

The PyCaret library was utilised for analysis because it automates complex processes such as data preprocessing, model training, hyperparameter tuning, and evaluation, allowing for rapid experimentation with multiple models. This not only saved time but also ensured consistency with the right insights of all the 15 algorithms. Several machine learning classification algorithms such as AdaBoost Classifier, Gradient Boosting Classifier,Light Gradient Boosting Classifier, Decision Tree Classifier, Extreme Gradient Boosting, Random Forest Classifier, K Neighbours Classifier, Extra Trees Classier, Logistic Regression, Ridge Classifier, Linear Discriminant Analysis, Naïve Bayes, Quadratic Discriminant Analysis, Dummy Classifier, SVM-Linear Kernel. were used to build predictive models for the willingness and ability to pay for PrEP.

services. The performance of the predictive models was evaluated using Recall because of its ability to differentiate true positives. Other metrics such as accuracy, area under the curve (AUC), precision, F1-score, Cohen's Kappa, and Matthews's correlation coefficient (MCC) were used in this study.

Variable importance was determined using methods such as permutation importance, partial dependence plots, or variable importance measures provided by the random forest classifier. The model was validated using cross-validation, bootstrapping, or splitting the data into training and validation sets.

To make the predictive model and its insights accessible to stakeholders, a web application was developed using the Streamlit Python library. The web application included a user-friendly interface that allows users to upload an Excel sheet of relevant demographic and socio-economic characteristics and receive predictions on the individual's ability to pay for PrEP services. Additionally, we created an interactive map of Uganda using QGIS software to display the spatial distribution of individuals who can pay for PrEP services. This map enables stakeholders to identify target areas for interventions effectively.

4.7. *Model Deployment and Application Development*

The trained model was deployed into a production environment, and we opted for a cloud-based deployment. This approach was chosen due to its scalability, flexibility, and ease of access, ensuring the model can handle varying loads and be accessed from multiple locations.

We developed a web application using the Streamlit Python library to facilitate user interaction with the model. Streamlit was chosen for its simplicity and efficiency in creating interactive web applications for data science and machine learning models. The web application included a user-friendly interface that allows users to upload an Excel sheet of relevant demographic and socio-economic characteristics and receive predictions on the individual's ability to pay for PrEP services. This application allows individuals and programmers to upload their data in various formats (e.g., CSV, Excel). Once data is uploaded, the application processes it and generates analysis and predictions in real-time. We hosted the application on a cloud platform ensuring that it can scale with all the data transferred to and from the web and the application is encrypted to protect sensitive information.

In addition, we created an interactive map of Uganda using QGIS software to display the spatial distribution of individuals who can pay for PrEP services. This map enables stakeholders to identify target areas for interventions effectively.

4.8. *Ethical Considerations*

The study protocol, protocol amendments, informed consent documents, and other relevant documents were approved by the Institutional Review Board (IRB). All correspondence with the IRB was retained in the regulatory or trial master file.

Copies of IRB approvals were filed with other study documents. A waiver of informed consent was sought as already existing data records were used. All personal identifying information was anonymized with access restricted to authorized personnel only.

5. Results

5.1. *Demographics*

Table 1 below provides a detailed demographic and socioeconomic variable used in the analysis for this study. A total of 97,330 records were analyzed, with the majority being females 52014 (53%), Males 45055 (46%) and 261 (1%) missing while the age group 25-34 years old contributed the highest number of individuals.

Table 1: Summary of social demographics

No	**Variable**	**Description**	**N% Overall N=97,330**	
1	Gender	Female	52014	(53%)
		Male	45055	(46%)
		Missing	261	(0.26%)
2	Age group	(25-34)years	31184	(32%)
		(35-44)years	21980	(23%)
		(18-24)years	20171	(20%)
		(45-54)years	12437	(13%)
		(55-64)years	6845	(7%)
		(65-74)years	2912	(3%)
		(75+)years	1801	(2%)
3	Level of education	Missing	46330	(47%)
		Completed Secondary	37520	(39%)
		Some Primary	6250	(6%)
		Some Secondary	3334	(3%)
		Completed Primary	2123	(2%)
		No formal education	1773	(2%)
4	Marital status	Married/Cohabiting/Living together	41368	(43%)
		Missing	33479	(34%)
		Never married	13164	(13%)
		Separated	4986	(5%)
		Widowed	4021	(4%)
		Divorced	312	(1%)
5	Religion	Catholic	27084	(27%)
		Pentecostal	23744	(24%)

		Anglican/Protestant	18497	(20%)
		Other	13718	(14%)
		SDA	8583	(9%)
		Moslems	2847	(3%)
		Orthodox	782	(0.9%)
		Other Christians	613	(0.6%)
		Traditional	601	(0.6%)
		Bahai	358	(0.4%)
		Hindu	279	(0.3%)
		None	216	(0.2%)
6	Residence	Urban	50879	(52%)
		Rural	46451	(48%)
7	Region	Central 2	47688	(50%)
		Mid North	12415	(13%)
		East Central	8798	(9%)
		North East	8349	(9%)
		Mid East	6342	(7%)
		Kampala	4254	(4%)
		Central	4029	(4%)
		West Nile	2389	(2%)
		Mid-West	1613	(1%)
		South West	1453	(1%)
8	Occupation	Professional/Technical/ Managerial	43152	(45%)
		Other	27742	(29%)
		Unskilled manual	6386	(7%)
		Refused/ Not provided	5322	(6%)
		Agriculture	4693	(4%)
		Domestic	4454	(4%)
		Sales and service	2554	(2%)
		Clerical	1292	(1.2%)
		Do not know	906	(0.9%)
		Skilled manual	829	(0.9%)
9	Wealth Status	Highest	43841	(45%)
		Middle	27402	(28.2%)
		Lowest	26087	(26.8%)
10	Financial decision	I do	56343	(57%)
		Spouse/Husband	31133	(32%)
		We both do	9604	(10.7%)
		Someone else	113	(0.1%)

		Refused/ Not provided	71	(0.1%)
		Do not know	66	(0.1%)
11	HIV Status	Results not received/refused	80414	(82%)
		Stated HIV negative	13527	(13%)
		Never tested	2680	(4%)
		Stated HIV positive	709	(1%)
12	Sexually active	Ever had sexual intercourse	50225	(51%)
		Missing	46263	(47%)
		Never had sexual intercourse	842	(1%)

5.2. *Model performance*

In this section, a comprehensive evaluation of machine learning models developed using the PyCaret library to identify high-risk sub-populations eligible for PrEP and willing to pay for the services is presented. Our analysis focuses on the deployment and performance of various algorithms in accurately identifying, quantifying, analyzing, and mapping high-risk populations that are eligible for PrEP and have a high ability to pay for the services in Uganda. This section systematically illustrates the effectiveness of the employed models, discussing metrics such as the accuracy, precision, recall, and area under the receiver operating characteristic curve. By dissecting the models' performances in this area of HIV care and treatment services delivery, this section aims to shed light on the potential of machine learning and AI techniques in effectively supporting HIV management and healthcare delivery.

Figure 1 is a structured comparison that outlines the performance of various machine learning classifiers on the prediction of eligibility for PrEP and willingness to pay for the service. Each classifier is evaluated according to several important metrics that gauge their performance as follows. Accuracy: This is the ratio of the number of correct predictions to the total number of input samples. It measures the overall correctness of the model. Higher accuracy means the model is more often correct across all predictions. Area under the curve (AUC): This metric is associated with the receiver operating characteristic curve and reflects a model's ability to distinguish between the classes (positive as eligible for PrEP and willing to pay for the service and negative as not eligibility for PrEP and not willing to pay for the service). A higher AUC value indicates better discrimination capabilities. Recall (or sensitivity): Recall indicates how many of the actual positive cases (eligible for PrEP and willing to pay for the service) the model correctly identified. It is crucial for conditions where missing a positive case can have serious consequences. Precision: This measures the fraction of correct positive predictions out of all positive predictions made. In the context of this study, it reflects how many of the patients the model labeled as eligible for PrEP and willing to pay for the service and were actually eligible for PrEP and willing to pay for the service. F1-score: The F1-score is the harmonic mean of the precision and recall, providing a bal-

ance between the two. It is especially useful when the distribution of classes is not even. Cohen's Kappa: This measures the agreement of the predictive model with the actual data, correcting for chance agreement. It gives a more-robust idea of the model's performance, especially with imbalanced datasets. Matthews's correlation coefficient (MCC): This is a measure of the quality of binary classifications. It takes into account true positives, true negatives, false positives, and false negatives. The MCC is considered a balanced measure, which can be used even if the classes are of very different sizes. Cells highlighted in yellow within Table 2 emphasize notable findings in the machine learning model comparison. These highlights are used to draw attention to results that significantly impact the study's conclusions, such as the highest accuracy rates, the best precision values, or any unexpected patterns in the data that may require further investigation or discussion.

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
ada	Ada Boost Classifier	0.9998	1.0000	0.9999	0.9997	0.9998	0.9997	0.9997	5.6100
gbc	Gradient Boosting Classifier	0.9999	1.0000	0.9999	0.9999	0.9999	0.9998	0.9998	10.7580
lightgbm	Light Gradient Boosting Machine	0.9999	1.0000	0.9999	0.9997	0.9998	0.9997	0.9997	10.5900
dt	Decision Tree Classifier	0.9998	0.9998	0.9998	0.9996	0.9997	0.9995	0.9995	3.6310
xgboost	Extreme Gradient Boosting	0.9999	1.0000	0.9998	0.9998	0.9998	0.9997	0.9997	4.5250
rf	Random Forest Classifier	0.9997	1.0000	0.9997	0.9994	0.9996	0.9993	0.9993	7.3990
knn	K Neighbors Classifier	0.9830	0.9934	0.9688	0.9890	0.9788	0.9647	0.9648	8.9120
et	Extra Trees Classifier	0.9862	0.9997	0.9679	0.9978	0.9826	0.9711	0.9714	8.7600
lr	Logistic Regression	0.9592	0.9812	0.9305	0.9671	0.9484	0.9147	0.9151	9.5290
svm	SVM - Linear Kernel	0.9528	0.9775	0.9239	0.9585	0.9406	0.9014	0.9022	6.1630
ridge	Ridge Classifier	0.9606	0.9780	0.9228	0.9785	0.9498	0.9175	0.9186	3.5740
lda	Linear Discriminant Analysis	0.9606	0.9780	0.9228	0.9785	0.9498	0.9175	0.9186	3.6840
nb	Naive Bayes	0.9659	0.9791	0.9189	0.9963	0.9561	0.9283	0.9303	3.5750
qda	Quadratic Discriminant Analysis	0.8742	0.9484	0.8959	0.8639	0.8671	0.7549	0.7718	3.5840
dummy	Dummy Classifier	0.5963	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000	3.5170

```
AdaBoostClassifier
AdaBoostClassifier(algorithm='SAMME.R', estimator=None, learning_rate=1.0,
                   n_estimators=50, random_state=2)
```

Fig. 1. Willingness to pay prediction: comparison of accuracy, area under the curve (AUC), recall, precision, F1-score, Kappa, and Matthews correlation coefficient (MCC) for different machine learning classifier models. Cells highlighted in yellow emphasize standout performance metrics (i.e., maximum values)

5.3. *Model comparison*

From Figure 1, AdaBoost Classifier (ada), Gradient Boosting Classifier (gbc), Extreme Gradient Boosting (xgboost), and Light Gradient Boosting Machine (light-

gbm) achieved the highest possible scores across almost all metrics (Accuracy, AUC, Recall, Precision, F1, Kappa, MCC), indicating perfect performance and highly reliable on this dataset. However, because AdaBoost classifier has the shortest time in training, and scored 0.9999 emerging the best in recall this indicates that it is marginally better. The AUC measures the model's ability to differentiate between classes (e.g., willing to pay for PrEP and not willing to pay for PrEP). An AUC of 1.000 for AdaBoost classifier, Gradient boosting classifier, Extreme gradient boosting, and Random forest classifier suggests that these models have a strong ability to discriminate between positive and negative cases. Precision assesses how many of the individuals predicted to be eligible and willing to pay for PrEP (positive cases) actually are. AdaBoost and Gradient boosting classifier's precision score of 0.9997 and 0.9999 respectively, suggests that the models are good at ensuring that the positive predictions are likely true. This means using these models might result in fewer false positives. It is critical to understand that machine learning predictions are not perfect tools themselves, but rather, risk stratification aids. Hence, AdaBoost's high performance in recall and AUC metrics indicates that it is a robust model for predicting one's eligibility and willingness to pay for PrEP. The Dummy Classifier serves as a baseline and performs poorly, indicating that the other models are indeed learning and performing well beyond random guessing.

5.4. *The Confusion Matrix*

The confusion matrix for the AdaBoost classifier in identifying individual's eligibility and willingness to pay for PrEP scenario is shown in Figure 2. The confusion matrix provides a snapshot of how well the AdaBoost classifier is performing accurately in identifying those eligible and willing to pay for PrEP as follows. The true positive (11785) reflects the model's ability to identify most of the individuals eligible and willing to pay for PrEP services. The true negative rate (17412) shows that the model is also capable of recognizing individuals who are not eligible and not willing to pay for PrEP services. The false positive (1) points to how the model incorrectly flags one individual as willing to pay yet she/he is not willing to pay for PrEP services. The false negative (1) is a critical metric in clinical settings, as it represents the individual that is willing to pay for PrEP services but incorrectly flagged as not willing to pay by the model. The confusion matrix shows that the AdaBoost classifier is highly effective at predicting the classes accurately, with almost perfect precision and recall. This aligns with the high training and cross-validation scores observed in the learning curve, further confirming the model's robustness and reliability.

All the Receiver Operating Characteristic (ROC) curves for the AdaBoost classifier show an AUC of 1.00, indicating that the model perfectly separates the classes without any misclassifications. The diagonal line (dotted) represents the performance of a random classifier (AUC = 0.5). Since all ROC curves for the AdaBoost classifier lie on the top-left corner (indicating perfect sensitivity and specificity), it

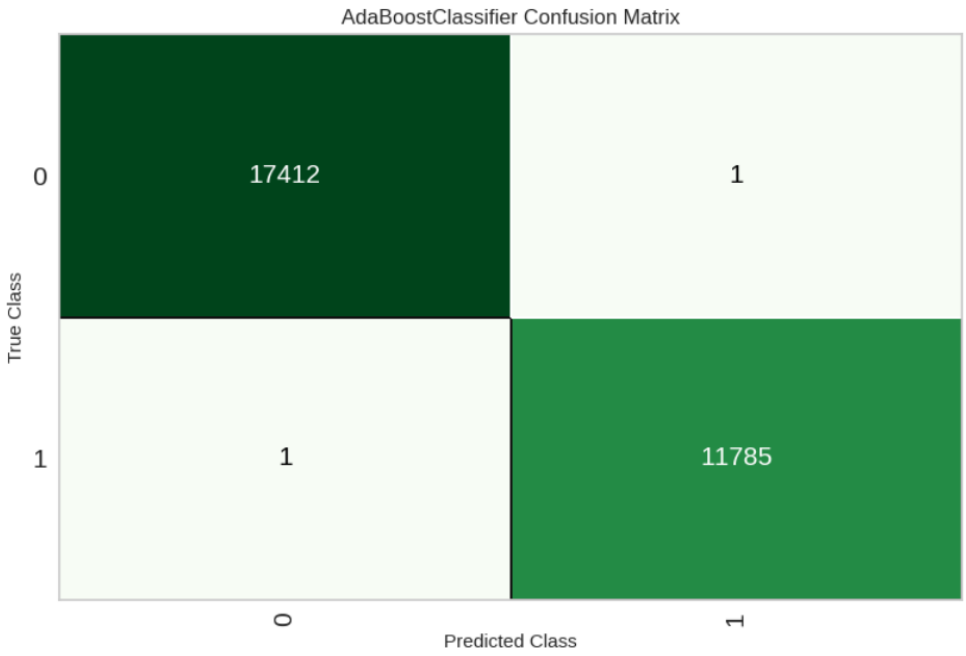

Fig. 2. AdaBoost classifier confusion matrix for eligibility and willingness to pay for PrEP prediction

demonstrates the model's outstanding predictive capabilities. The ROC curves and the corresponding AUC values reiterate the findings from the confusion matrix and the performance table. The AdaBoost classifier exhibits exceptional performance with perfect separation between classes, making it a highly reliable model for this dataset.

Considering the learning curve, the training score remains very high (close to 0.99995) across all training instances, this indicates that the model is able to fit the training data extremely well, almost perfectly. The cross-validation score is slightly lower than the training score but still very high (around 0.99980). There is some fluctuation in the cross-validation score, suggesting some variability in model performance on unseen data. As the number of training instances increase, the training score remains consistently high. The AdaBoost classifier performs exceptionally well with both training and cross-validation scores being very high. The model is able to almost perfectly fit the training data while also generalizing well to unseen data. The small gap between training and cross-validation scores suggests good generalization capabilities, with minimal overfitting. Training Instances: Increasing the number of training instances generally leads to slightly improved cross-validation scores and reduced variability, indicating that the model benefits from more training data. This learning curve indicates a robust model with excellent performance on both training and validation sets, demonstrating its reliability and effectiveness.

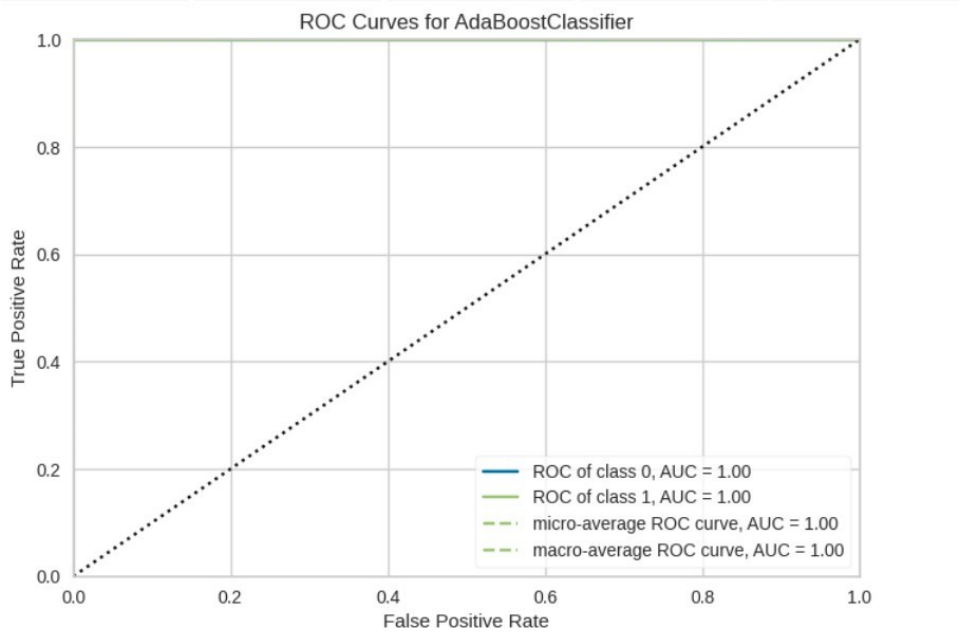

Fig. 3. ROC curves for AdaBoost classifier used for eligibility and willingness to pay for PrEP prediction. (The diagonal line in the graph indicates that lines above it are significant)

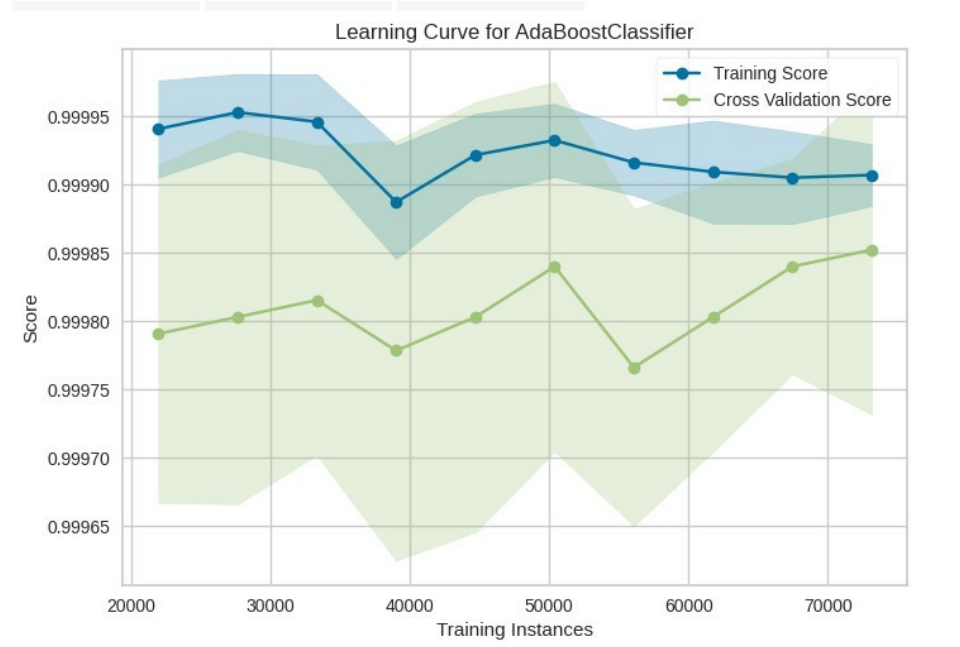

Fig. 4. Learning curves for AdaBoost classifier to predict eligibility and willingness to pay for PrEP services

The analysis of feature importance was crucial in this research as it allowed the researchers to identify the factors(features) that most strongly predict the ability to pay for PrEP services, this may guide future interventions and resource allocation. The feature importance was derived using techniques such as permutation importance, partial dependence plots, and variable importance measures provided by the random forest classifier.

Permutation Importance, evaluates the change in the model's performance (e.g., accuracy, recall) when the values of a given feature are randomly shuffled. A significant decrease in model performance indicates that the feature is important. Partial Dependence Plots (PDP), that shows the relationship between a feature and the predicted outcome while keeping all other features constant. The Partial dependence plots help to visualize how a feature influences the model's prediction, providing insight into its importance. Variable Importance in Random Forests classifiers also calculate feature importance based on how much a feature improves the split quality in the trees. The importance is typically measured using metrics like Gini impurity and information gain, which quantifies the contribution of each feature in reducing uncertainty within the model.

Considering the feature importance plot below, Service location is the most important feature, having the highest variable importance score. This implies that the location where the service is provided significantly influences the model's predictions. This is followed by the Sexually_active feature which indicates that the sexual status of an individual is also a major factor in the model's predictions. These features are followed by other important variables to the model such as, Known_hiv_status, region, age, wealth_quintile, religion, education_level, financial_decision, marital_status, and occupation While variables like age specific groups (45-54 years, 55-64 years), gender, urban have minimal importance to the model as shown in the feature importance plot below.

6. Discussion

This study applied machine learning techniques to predict individuals' eligibility and willingness to pay for PrEP services in Kampala, Uganda. Among the models tested, the AdaBoost classifier outperformed others across all major evaluation metrics, achieving near-perfect scores in accuracy, precision, recall, F1-score, AUC, and MCC. These results highlight the potential of ensemble-based classifiers to support health resource allocation and intervention targeting in public health settings.

One of the most remarkable findings is the extremely high recall (99%) and precision (99.97%) of the AdaBoost classifier. This performance is particularly important in a healthcare context where both false negatives and false positives carry substantial consequences. The model's high performance indicates its reliability in identifying target individuals while minimizing errors in classification [11]. However, an in-depth analysis of feature importance reveals that Service location and sexual activity emerged as the most predictive features, followed by known HIV status,

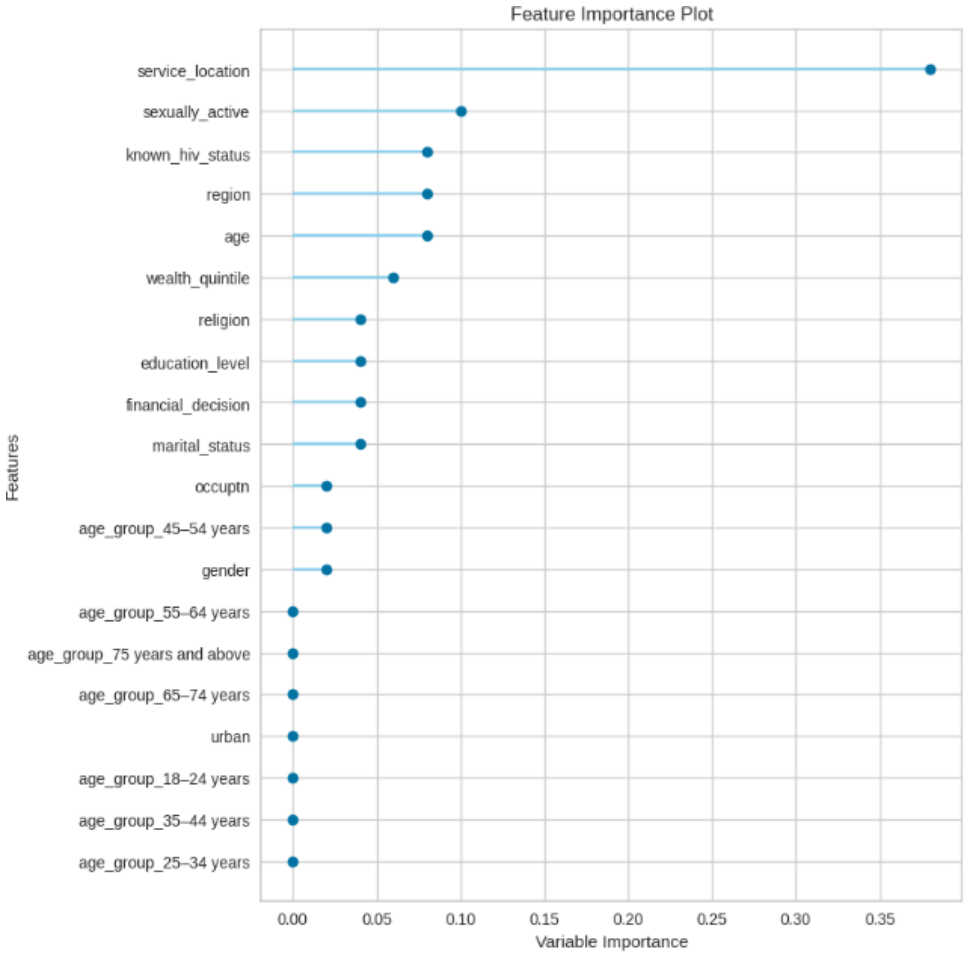

Fig. 5. Feature importance plot for eligibility and willingness to pay for PrEP prediction

region, education, wealth quintile, and religion.

Interestingly, variables that are commonly assumed to be influential, such as age, occupation, and financial decision, had relatively low feature importance scores in the AdaBoost model. This may appear contrary to expectations, especially considering their known relevance in HIV risk and economic behavior, this was also realized in a case study carried out in Netherlands by Adekule [12]. Such findings can be explained by how ensemble models like AdaBoost handle feature redundancy and interaction. Age, occupation, and financial decision-making are often correlated with broader socioeconomic indicators such as wealth quintile and employment status. In such cases, the model tends to prioritize variables that offer the most direct and consistent predictive signals, suppressing others that may be redundant or introduce noise. In the presence of multicollinearity, feature importance becomes a

reflection of statistical contribution rather than contextual significance [13].

Although age is widely recognized as a critical factor in HIV risk and PrEP eligibility, particularly among individuals aged 18–34, who are often at higher sexual activity levels and biological susceptibility, the predictive modeling results revealed that age had a limited impact on predicting individuals' ability to pay for PrEP. This finding may appear surprising but is methodologically consistent with the design and objectives of our study. Specifically, while age correlates with HIV risk, our model was optimized to predict economic capacity rather than medical need.

Features such as employment status, gender, and wealth_quintile more directly reflect financial ability and may absorb the variance that age would otherwise explain. This divergence underscores an important point that epidemiological importance does not always translate into predictive importance in machine learning models. Therefore, while younger individuals remain a key demographic for HIV prevention interventions, targeted strategies for PrEP financing and delivery should prioritize socioeconomic profiling over demographic characteristics alone.

Interestingly, variables such as financial decision-making and occupation, though intuitively relevant to one's ability to pay for PrEP, did not emerge as significant predictors in our model's feature importance analysis (see Fig. 5). This may seem surprising however, it likely reflects the inherent interdependence between these variables and broader socioeconomic indicators such as wealth quintile, education level, and employment status. Machine learning models like AdaBoost are sensitive to multicollinearity and tend to assign greater weight to features that provide the strongest early predictive gains, often de-emphasizing others that are semantically important but statistically redundant. Additionally, the categorical representation of occupation in the dataset may have lacked sufficient granularity to contribute meaningful discriminatory power [14].

6.1. *Model performance*

The predictive model's performance, as illustrated in Figure 1, demonstrates the effectiveness of various machine learning classifiers in predicting eligibility and willingness to pay for PrEP services. The metrics used to evaluate these models include accuracy, area under the curve (AUC), recall, precision, F1-score, Cohen's Kappa, and Matthews correlation coefficient (MCC). These metrics provide a comprehensive assessment of the model's predictive power and reliability.

The AdaBoost classifier achieved an accuracy of 99%, indicating that it correctly predicted the outcome for the vast majority of cases. High accuracy suggests the model's overall correctness, but it is important to consider other metrics to understand its performance in detail.

The AUC score for the AdaBoost classifier is 1.00, which indicates perfect discrimination between those eligible and willing to pay for PrEP services and those who are not. A high AUC value signifies excellent model performance in distinguishing between classes. This is consistent with findings from other studies that

have utilized machine learning models for similar purposes, showing that AUC is a robust metric for evaluating classifier performance [15].

The recall for the AdaBoost classifier is 99%, demonstrating its ability to identify almost all actual positive cases (those eligible and willing to pay for PrEP). High recall is particularly important in healthcare settings where missing a positive case can have serious consequences. Similar studies in predictive modeling have emphasized the importance of high recall in ensuring that vulnerable populations are not overlooked [16].

The precision score for the AdaBoost classifier is 99.97%, meaning that almost all predictions made by the model for those eligible and willing to pay for PrEP are correct. High precision is crucial to minimize false positives, which can lead to unnecessary resource allocation. This aligns with recent work modeling PrEP willingness to pay using ML techniques in sub-Saharan Africa [17]. Studies have shown that precision is vital in health interventions to ensure efficient use of resources [16].

The confusion matrix for the AdaBoost classifier shows true positives (11,785), true negatives (17,412), false positives (1), and false negatives (1), indicating the model's high precision and recall. The ROC curves further corroborate this with an AUC of 1.00, demonstrating the model's perfect separation between the classes without any misclassifications. The feature importance plot reveals that service location and sexual activity are the most significant predictors of the model. Other important variables include known HIV status, region, age, wealth quintile, religion, education level, financial decision, marital status, and occupation. This highlights the multifaceted nature of factors influencing the ability to pay for PrEP services, consistent with other studies that emphasize the importance of considering a wide range of socioeconomic and demographic factors in predictive modelling [18].

The learning curves indicate that the AdaBoost classifier performs exceptionally well with both training and cross-validation scores being very high, demonstrating its reliability and effectiveness. The small gap between training and cross-validation scores suggests minimal overfitting, highlighting the model's robustness.

6.2. *Conclusion*

The study aimed to develop an AI model using the PyCaret library to predict eligibility and willingness to pay for PrEP services in Kampala, Uganda. The AdaBoost classifier emerged as the most effective model, demonstrating superior performance across multiple metrics including accuracy, AUC, recall, precision, F1-score, Cohen's Kappa, and MCC. The model's high performance underscores its potential in identifying high-risk populations eligible for PrEP and willing to pay for the service, thereby facilitating targeted and efficient deployment of PrEP programs.

The findings have significant implications for HIV prevention efforts in Uganda. By accurately identifying individuals who are most likely to benefit from and afford PrEP, healthcare providers can optimize resource allocation and maximize the impact of these interventions. The insights gained from this study can inform the

development of customized financial models and payment structures to improve the affordability and accessibility of PrEP services. As shown in related work applying affordability focused modeling in Uganda [19].

Moreover, the successful application of machine learning techniques in this context demonstrates the broader potential of these methods to support various aspects of healthcare delivery and decision-making. The ability to rapidly analyze large datasets, identify patterns, and make accurate predictions can be leveraged to improve a wide range of public health interventions, from disease surveillance and resource allocation to personalized care and treatment optimization [19].

However, it is important to acknowledge the limitations of the study. While the models exhibit substantial predictive power, further refinement and validation on new datasets from other settings are necessary to enhance their accuracy and generalizability. Additionally, the reliance on existing datasets introduces certain limitations regarding data quality and representativeness. Future research should incorporate longitudinal data and time-series analysis to better capture the dynamic nature of individuals' financial situations and improve the model's predictions over time.

In conclusion, this study highlights the valuable contribution of machine learning in enhancing HIV prevention and treatment efforts in Uganda. The AdaBoost classifier's exceptional performance underscores these techniques' potential to support data-driven decision-making and improve healthcare service delivery. These findings provide a compelling case for integrating AI and machine learning into public health strategies to promote equitable access to essential preventive services and ultimately reduce the burden of HIV/AIDS. Comparable successes have been documented using ML in low-resource health systems [20].

Acknowledgments:

We sincerely acknowledge the support of our consortium partners: Makerere University, the Infectious Diseases Institute (IDI), Makerere AI Lab, and Sunbird AI, whose financial contributions were instrumental to the success of this project.

We also thank the Uganda Bureau of Statistics (UDHS 2019/20), the Uganda Population HIV/AIDS Impact Assessment (UPHIA 2016/2017), and Rocket Health for providing access to their datasets.

We are especially grateful to our colleagues at the African Center of Applied Digital Health Projects Research for their institutional support.

We further recognize the developers of PyCaret, Streamlit, and QGIS, whose tools supported our data analysis and visualization.

The views presented here are solely those of the authors and do not necessarily represent those of the supporting institutions.

References

1. UNAIDS. 2021 UNAIDS Global AIDS Update — Confronting inequalities — Lessons for pandemic responses from 40 years of AIDS. Jt United Nations Program HIV/AIDS [Internet]. 2021;13–7. Available from: https://www.unaids.org/sites/default/files/media_asset/2021-global-aids-update_e n.pdf
2. Doshi RH, Apodaca K, Ogwal M, Bain R, Amene E, Kiyingi H, et al. Estimating the size of key populations in Kampala, Uganda: 3-source capture-recapture study. JMIR Public Heal Surveill. 2019;5(3) 1–9.
3. UPHIA. Uganda Population-based HIV Impact Assessment 2020 Uganda Population-based HIV Impact Assessment. 2020;2020(April).
4. Koss CA, Havlir D V., Ayieko J, Kwarisiima D, Kabami J, Chamie G, et al. HIV incidence after pre-exposure prophylaxis initiation among women and men at elevated HIV risk: A population-based study in rural Kenya and Uganda. PLoS Med [Internet]. 2021;18(**??**):1–22. Available from: http://dx.doi.org/10.1371/journal.pmed.1003492
5. UNAIDS. Annual Progress Report on HIV Prevention 2020. Onusida [Internet]. 2020;(December 2019):3–21. Available from: https://www.unaids.org/sites/default/files/media_asset/10122019_UNAIDS_PCB 45_PPT_Annual-progress-report-HIV-Prevention-2020.pdf
6. Kagaayi J, Batte J, Nakawooya H, Kigozi B, Nakigozi G, Strömdahl S, et al. Uptake and retention on HIV pre-exposure prophylaxis among key and priority populations in South-Central Uganda. J Int AIDS Soc. 2020;23(**??**):1–6.
7. Musoke1 D, Boynton2 P, Butler2 C, Miph Boses Musoke. Health seeking behaviour and challenges in utilising health facilities in Wakiso district, Uganda. Afr Health Sci. 2014;14(4) 128–31.
8. Baxter C, Abdool Karim S. Combination HIV prevention options for young women in Africa. African J AIDS Res. 2016;15(2) 109–21.
9. Amwonya D, Kigosa N, Kizza J. Female education and maternal health care utilization: evidence from Uganda. Reprod Health [Internet]. 2022;19(1) 1–18. Available from: https://doi.org/10.1186/s12978-022-01432-8
10. Lee S, Wickrama KKAS, Lee TK, O'Neal CW. Long-Term Physical Health Consequences of Financial and Marital Stress in Middle-Aged Couples. J Marriage Fam. 2021;83 (4) 1212–26.
11. Pfeiffer J, Li H, Martez M, Gillespie T. The role of religious behavior in health self-management: A community-based participatory research study. Religions. 2018;9(11).
12. Cyr ME, Etchin AG, Guthrie BJ, Benneyan JC. Access to specialty healthcare in urban versus rural US populations: A systematic literature review. BMC Health Serv Res. 2019;19(1) 1–17.
13. Access to and utilisation of health services for the poor in Uganda: a systematic review of available evidence. BMC Public Health. 2023;23(1) 1–13.

14. McMaughan DJ, Oloruntoba O, Smith ML. Socioeconomic Status and Access to Healthcare: Interrelated Drivers for Healthy Aging. Front Public Heal. 2020;8(June):1–9.
15. UNAIDS and MoH. Uganda AIDS Report 2019: Enhancing HIV mainstreaming towards ending AIDS as a public health threat in Uganda by 2030. 2019;(September):80.
16. Rosenberg MS, Gómez-Olivé FX, Rohr JK, Houle BC, Kabudula CW, Wagner RG, et al. Sexual Behaviors and HIV Status: A Population-Based Study among Older Adults in Rural South Africa. J Acquir Immune Defic Syndr. 2017;74(1) e9–17.
17. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit. 1997;30(7) 1145–59.
18. Li S, Zeng Y, Jr WCC, Erfanzadeh M, Mutch M, Zhu Q, et al. Adaptive Boosting (AdaBoost)-based multiwavelength spatial frequency domain imaging and characterization for ex vivo human colorectal tissue assessment. 2020;13(6) 1–18.
19. White T, Algeri S. Estimating the lifetime risk of a false positive screening test result. PLoS One [Internet]. 2023;18(2 February):1–12. Available from: http://dx.doi.org/10.1371/journal.pone.0281153
20. Ransome Y, Bogart LM, Kawachi I, Kaplan A, Mayer KH, Ojikutu B. Area-level HIV risk and socioeconomic factors associated with willingness to use PrEP among Black people in the U.S. South. Ann Epidemiol. 2020;42:33–41.