WELCOME
TO OUR PRESENTATION

Analysis of Demographic and Behavioral Factors in HIV Prediction Using Machine Learning

Our Team

ID	Name	Intake	Section
21222203049	**Nur-A-Kamrul Islam**	41	01
21222203015	**MD. Omar Faruk**	41	01
21222203038	**Tanjila Aktar Shathi**	41	01
21222203036	**Esrat Jahan**	41	01
21222203023	**Rabeya Aktar**	41	01

Supervised By

Ashifur Rahman
Lecturer
Bangladesh University of Business And Technology

Table of contents

01

Introduction

Introduction

Human Immunodeficiency Virus (HIV) remains a major global health challenge, with 39.9 million cases reported in 2023. Despite advancements in antiretroviral therapy, late diagnosis continues to hinder effective control, especially with 20–30% of cases in high-income countries remaining undiagnosed. In Bangladesh, HIV prevalence is concentrated among key populations, reaching 4.1% among people who inject drugs (PWID). Traditional risk assessment methods often overlook complex, non-linear risk factors. This study explores a machine learning–based approach to improve HIV risk prediction using a Bangladesh-specific dataset, emphasizing comprehensive feature engineering, class imbalance mitigation, and explainable AI.

Problem Statement

Despite significant progress in HIV treatment, late diagnosis remains a critical barrier to effective disease management, particularly in regions with concentrated epidemics like Bangladesh. Conventional risk assessment tools lack the ability to capture complex, non-linear interactions between behavioral, demographic, and socioeconomic factors, leading to missed or delayed, identification of high-risk individuals. There is an urgent need for data-driven, interpretable, and region-specific models that can enhance early detection and guide targeted interventions for HIV prevention and control.

Objective

The objective of this study is to develop and evaluate a machine learning–based HIV risk prediction model tailored to the Bangladeshi context. The model aims to:

- Incorporate novel and relevant features through comprehensive feature engineering.
- Address class imbalance using Synthetic Minority Oversampling Technique (SMOTE).
- Ensure model interpretability using SHAP (SHapley Additive exPlanations) values.

This approach seeks to improve early detection and support targeted public health interventions.

02

Methodology

Methodology

Data Preperation

The dataset was cleaned, encoded, and balanced using SMOTE for model training.

Feature Engineering

Age groups, interaction terms, and correlations were engineered to improve the model.

Model Development

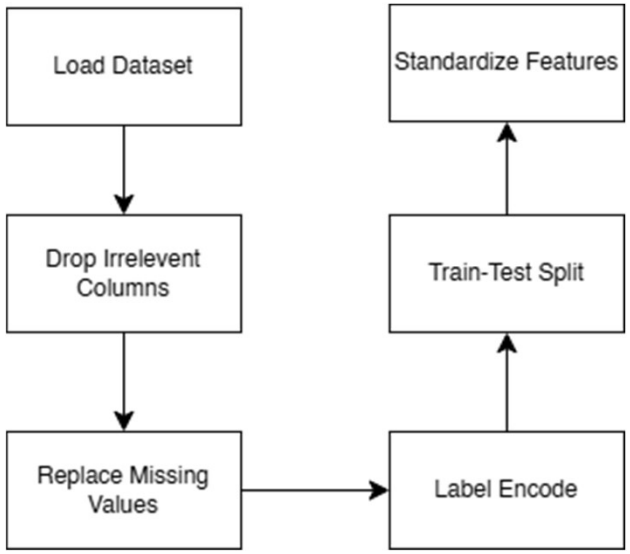
Six classifiers evaluated with five-fold cross-validation using F1-score.

Explainability Analysis

SHAP analysis highlighted age and STD history as key predictors.

Data Preprocessing

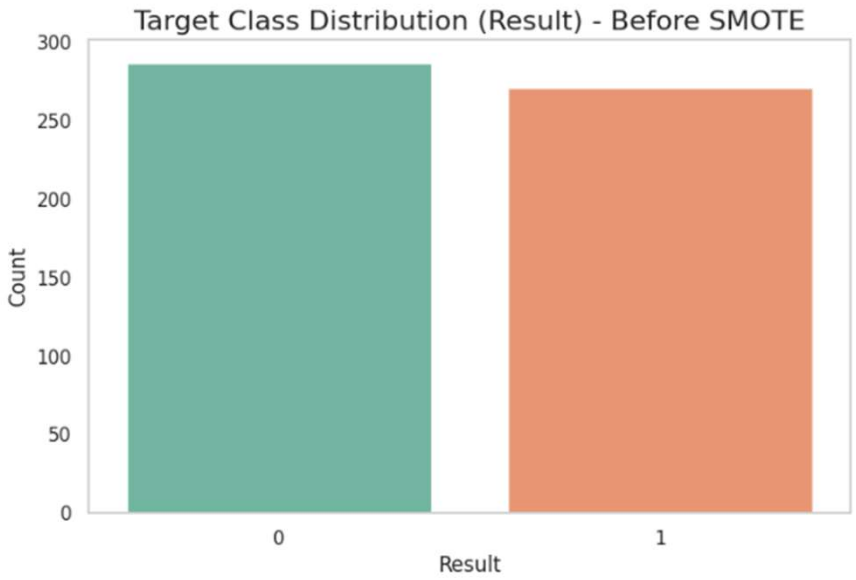
- Data preparation ensured robustness through rigorous preprocessing and critical transformations on 698 samples.



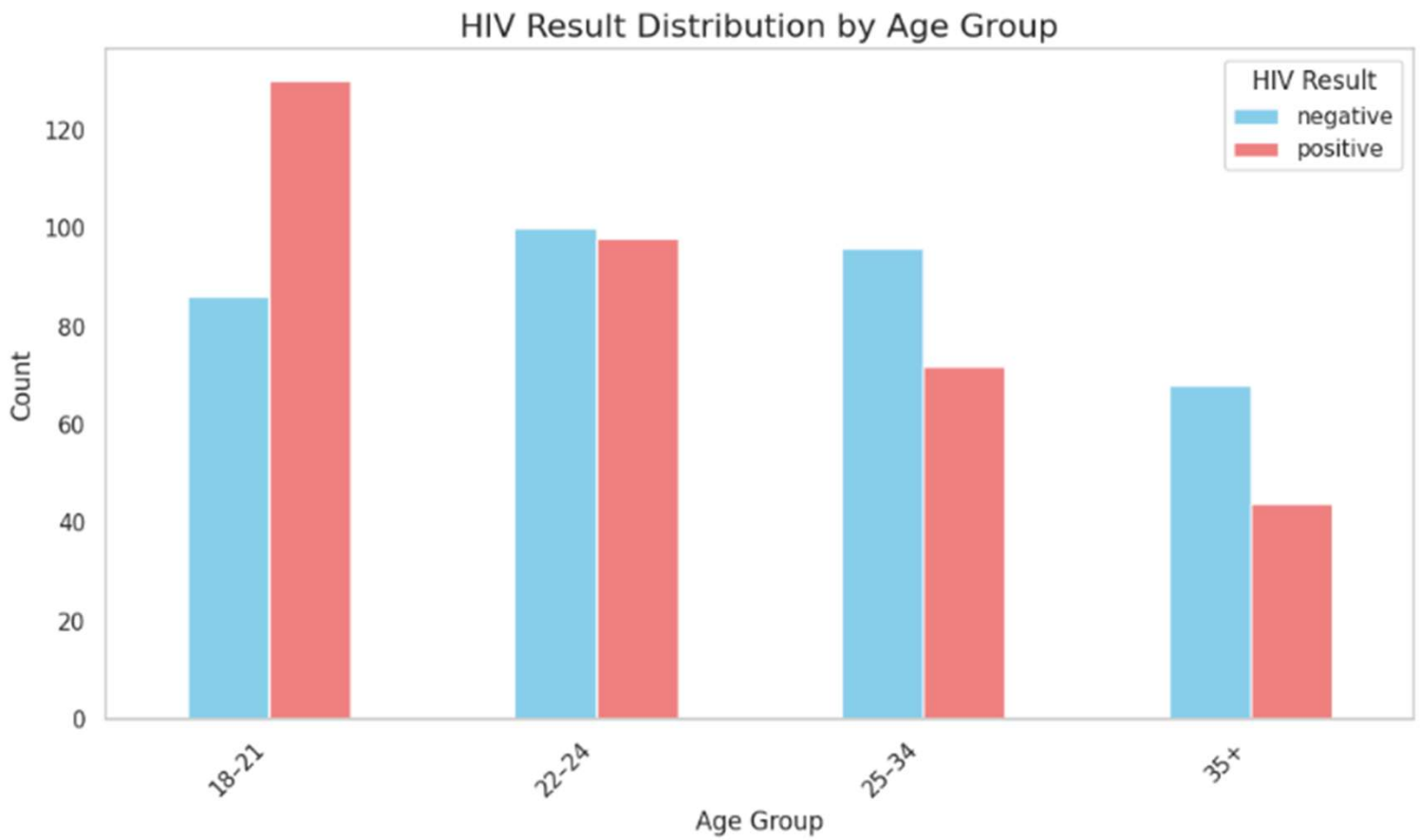
03

Data Visualization

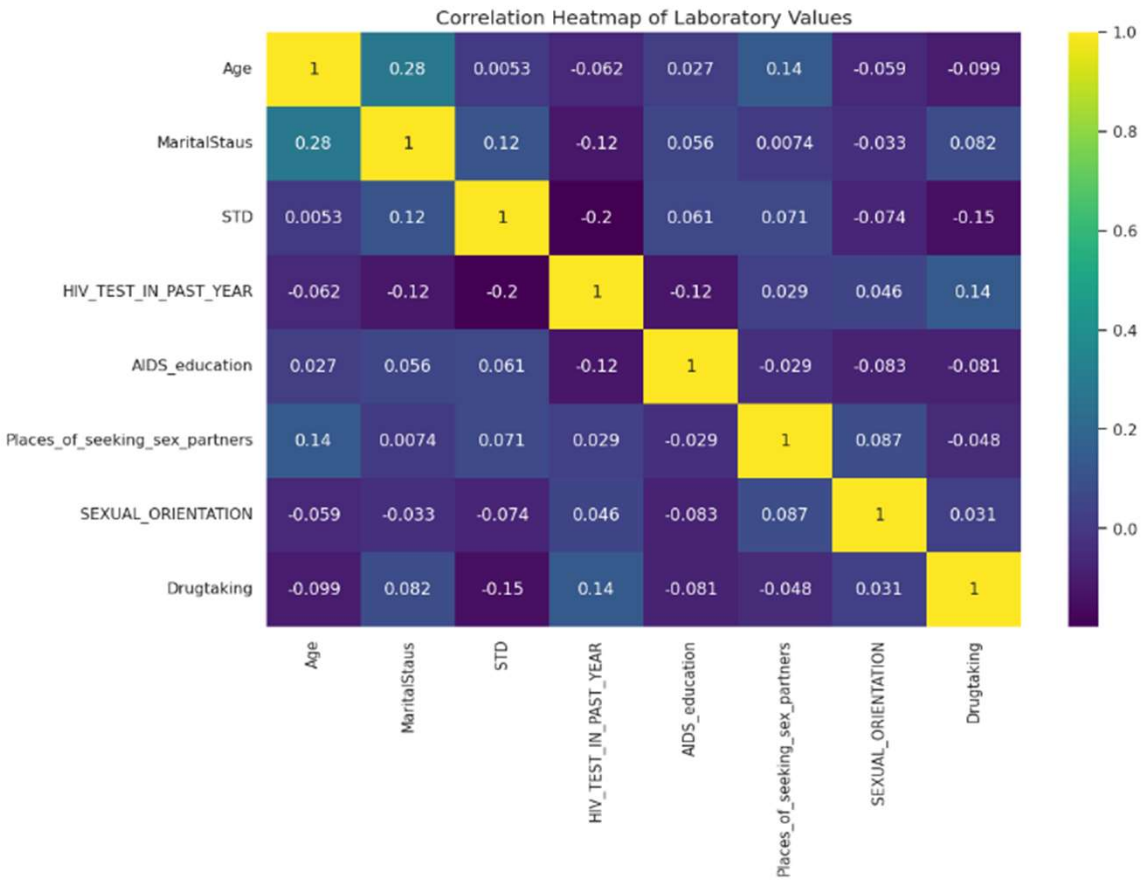
Class Distriubution (Smote)



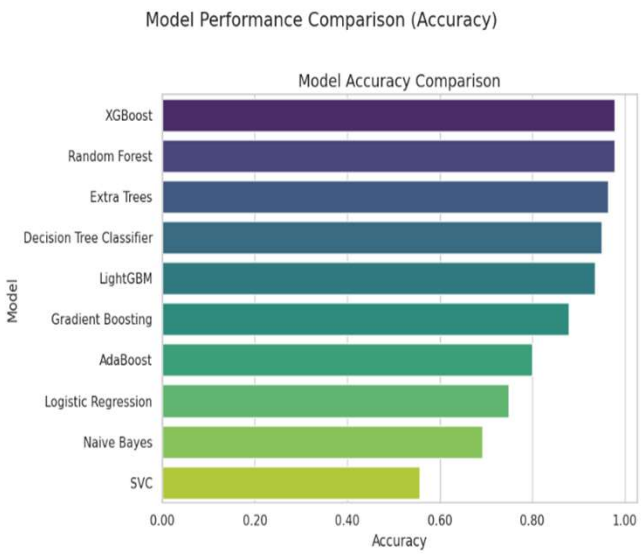
Age Group Discretization

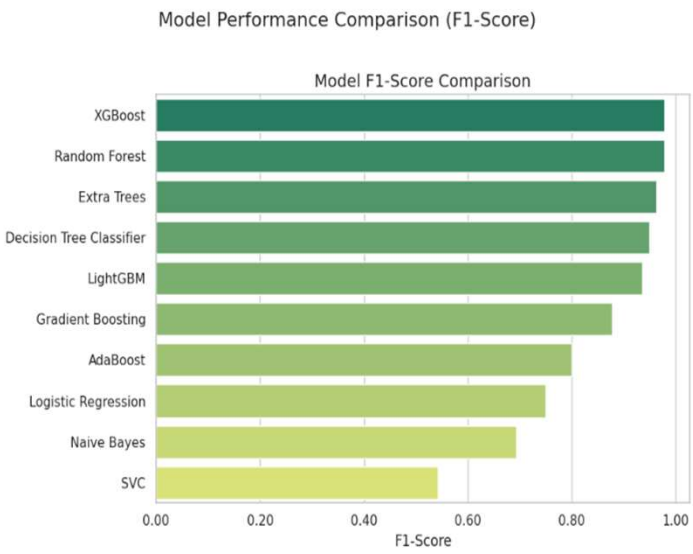


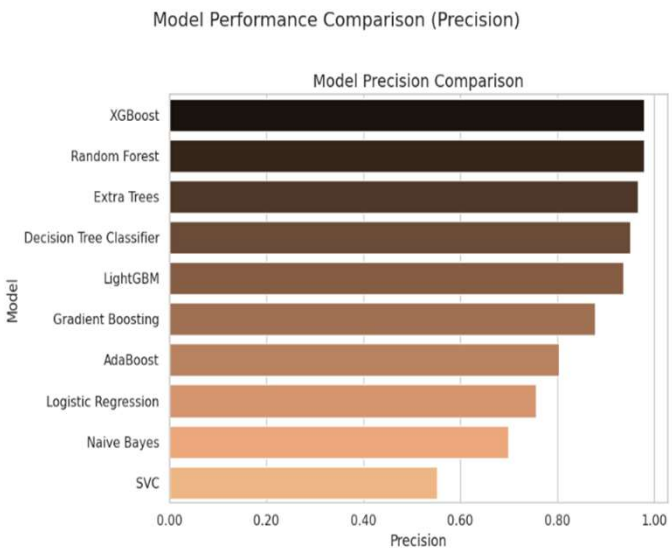
Feature Correlation Heatmap



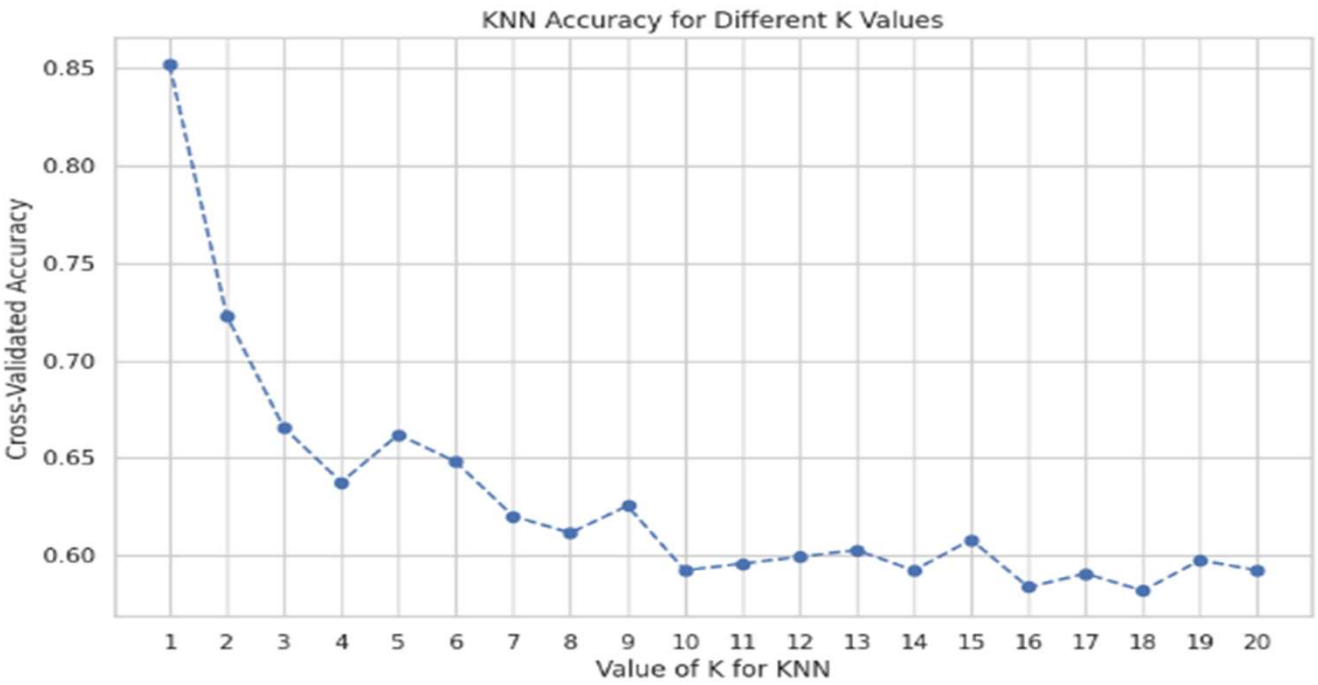
Model Architecture Comparison



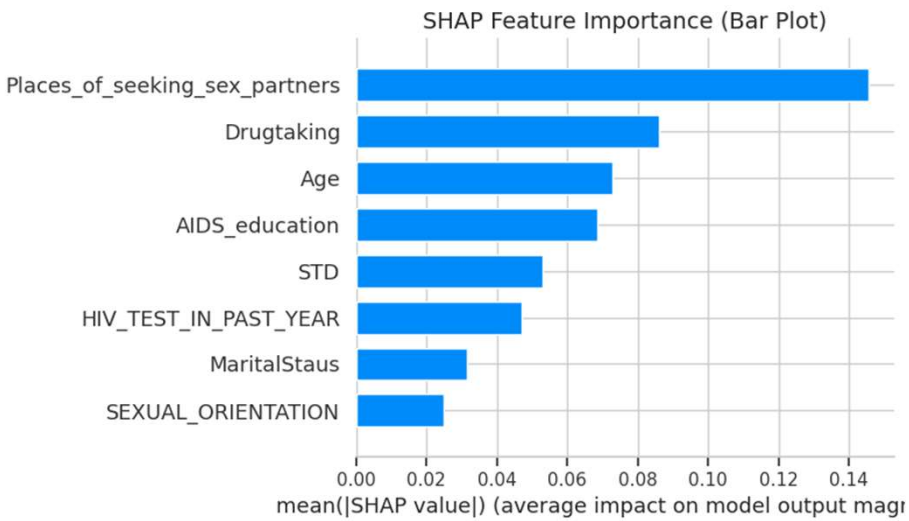


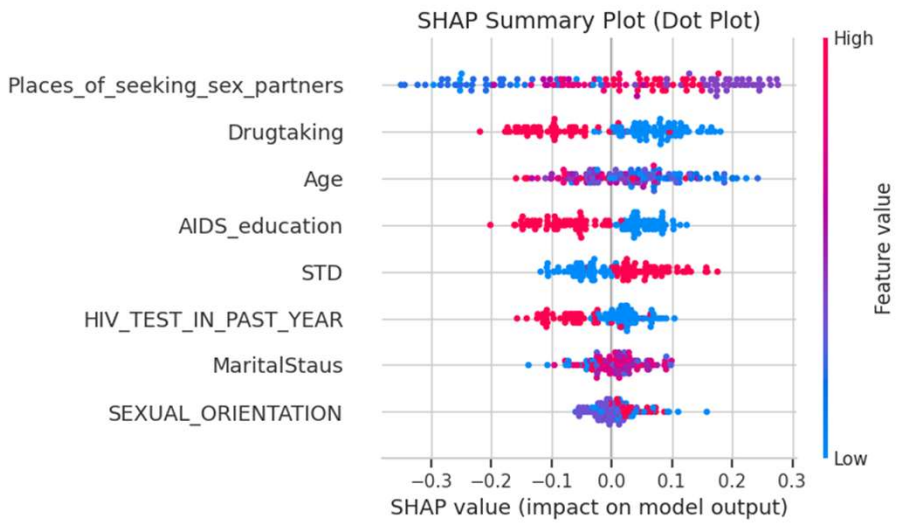


Cross-Validation Schematic

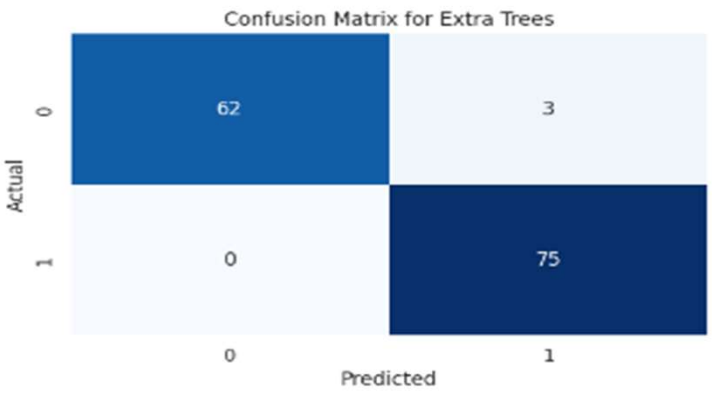


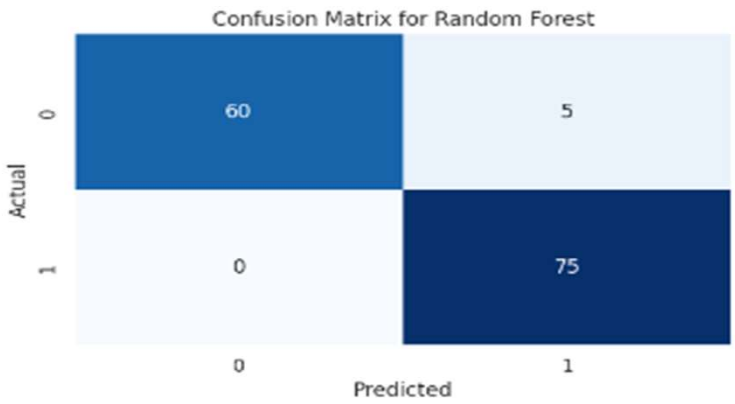
SHAP Summary

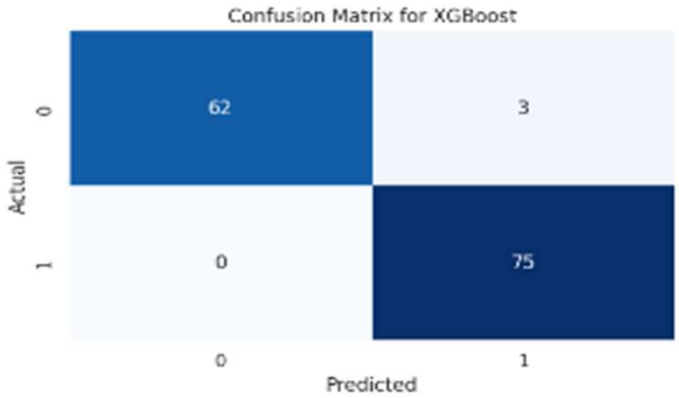


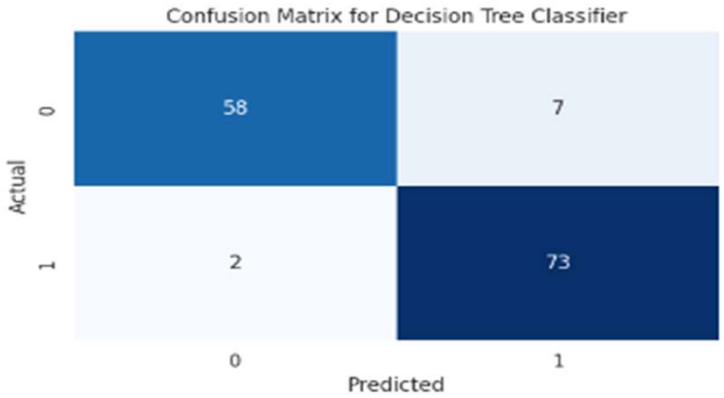


Confusion Matrix

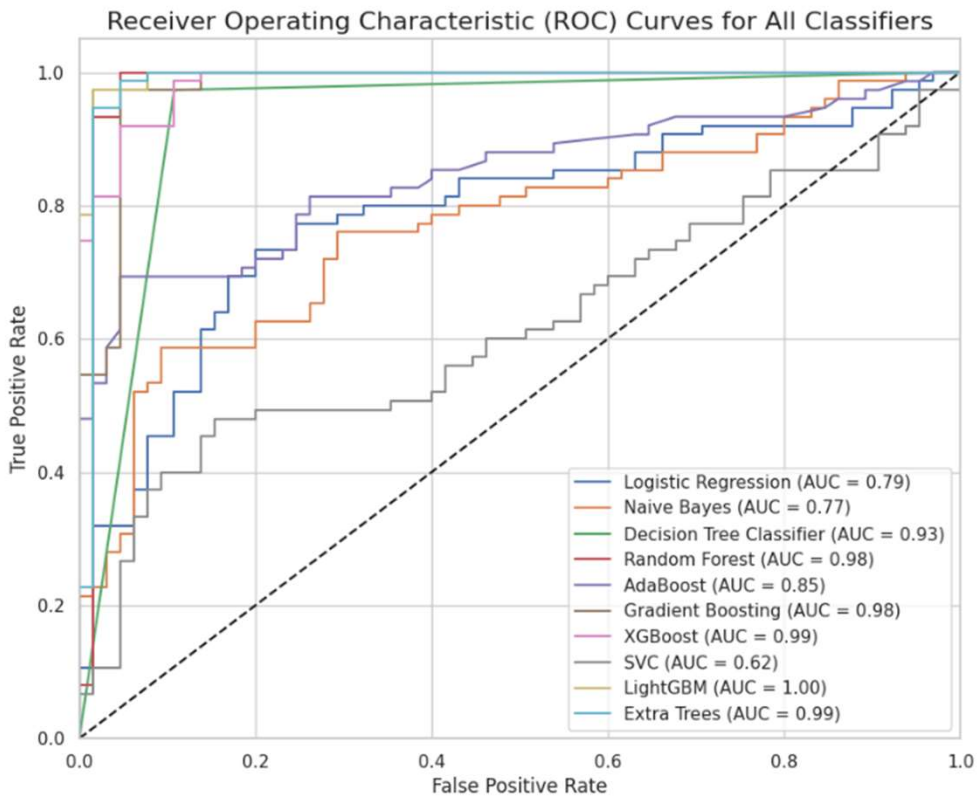








ROC Curves


Receiver Operating Characteristic (ROC) Curves for All Classifiers
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Logistic Regression (AUC = 0.79)
Naive Bayes (AUC = 0.77)
Decision Tree Classifier (AUC = 0.93)
Random Forest (AUC = 0.98)
AdaBoost (AUC = 0.85)
Gradient Boosting (AUC = 0.98)
XGBoost (AUC = 0.99)
SVC (AUC = 0.62)
LightGBM (AUC = 1.00)
Extra Trees (AUC = 0.99)

04
Metrics

Evaluation Metric

- Best Performing Model would be Random Forest. But XGBoost is not far off.

Metric	Random Forest	XGBoost	Decision Tree
Accuracy	97.9%	96.4%	93.6%
Precision	97.9%	96.7%	93.8%
Recall	97.9%	96.4%	93.6%
F1 Score	97.9%	96.5%	93.5%
AUC	0.94+	0.94	0.89

Ensamble Result

What are Ensemble Methods?

- Combine predictions from multiple models to improve overall performance.
- Reduce errors, increase stability, and handle complex data patterns.

Models

1. Random Forest
2. XGBoost
3. Decision Tree

Metric	Value
Accuracy	0.96
Voting Type	Soft

05
Conclusion

Conclusion

- Machine learning models effectively predict HIV status using demographic and behavioral data.
- Random Forest achieved the highest accuracy (97.9%), while ensemble and XGBoost models also performed strongly.
- High precision and recall values indicate reliable identification of both positive and negative cases.
- These results highlight the potential of data-driven approaches to support early HIV detection and targeted interventions.

Thanks!

any questions?