A Data-Driven Model for Predicting Antiretroviral Therapy (ART) Adherence Risk among Kenyan Adolescents

Problem Statement

Adolescents with HIV in Kenya face significant challenges in adhering to their Antiretroviral Therapy (ART) medication. These challenges include lack of awareness of HIV status, limited access to healthcare facilities, socioeconomic issues such as poverty, medication side effects, and parental negligence. Current interventions through education, rural clinic establishment, follow-ups, family support, and side effect management often lack personalized strategies necessary for widespread effectiveness among teenagers. Poor adherence leads to increased risks of HIV progression to AIDS, drug resistance development, HIV transmission, poor health outcomes, reduced quality of life, and potential increases in HIV-related mortality rates among adolescents, placing additional strain on the healthcare system and resources. This study aims to develop a predictive model to identify adolescents at risk of poor ART adherence, track the influence of contributing factors, and enable tailored interventions to improve health outcomes.

Research Objectives

General Objective

To develop and implement a data-driven model that predicts ART adherence patterns among Kenyan adolescents, identifies key influencing factors, and generates insights for personalized intervention strategies.

Specific Objectives

1. To generate synthetic healthcare data that accurately represents ART adherence patterns and associated factors among Kenyan adolescents.

2. To develop a machine learning model for predicting adherence outcomes using demographic, clinical, and social-environmental factors.

3. To implement a feature importance analysis system to identify and rank the impact of various factors on ART adherence prediction.

Data Description

Source of Data

This study will utilize synthetic data generated to simulate real-world ART adherence patterns among Kenyan adolescents. The synthetic data generation will be informed by statistics and relationships documented in multiple authoritative sources:

1. Kenya Population-based HIV Impact Assessment (KENPHIA, 2023): Providing national statistics on HIV prevalence and treatment outcomes among adolescents.

2. Kenya National AIDS & STI Control Programme (NASCOP, 2024): Offering guidelines and aggregated statistics on ART adherence among different demographic groups.

3. Kenya Medical Research Institute (KEMRI, 2023): Reporting adherence rates of approximately 60% among Kenyan adolescents on ART.

Citations Supporting Synthetic Data Approach

Several researchers have successfully employed synthetic data generation for healthcare and specifically HIV research:

Goncalves et al. (2020) developed methodologies for creating and evaluating synthetic patient data for HIV treatment outcomes, demonstrating that synthetic data can effectively mirror statistical properties of real patient data while preserving privacy. Their approach achieved up to 97% similarity with real clinical data patterns while eliminating re-identification risks.

Dankar et al. (2021) conducted a comprehensive comparison of synthetic data generation methods specifically for healthcare applications, finding that certain generative approaches can preserve both statistical utility and privacy protection when original data cannot be shared.

Rankin et al. (2020) specifically used synthetic data for HIV research, demonstrating that machine learning models trained on synthetic HIV patient data performed with comparable accuracy to those trained on real data (within 3-5% performance difference).

Potential Challenges of Using Synthetic Data

While synthetic data offers advantages for sensitive health information, several challenges must be acknowledged:

1. Validation Complexity: As noted by Chen et al. (2021), validating that synthetic data accurately reflects real-world clinical patterns remains challenging without direct comparison to original data.

2. Missing Nuanced Relationships: Xie et al. (2018) highlights that synthetic data may fail to capture subtle relationships between variables that exist in real patient data, potentially omitting important indicators for clinical prediction.

3. Cultural Context Limitations: Haberer et al. (2017) emphasize that HIV adherence behaviors are heavily influenced by cultural and regional factors that may be difficult to accurately simulate in synthetic data generation.

4. Generalizability Concerns: Tucker et al. (2020) caution that models trained on synthetic data may not generalize well to diverse real-world populations, especially when demographic variations are significant.

Data Dimensions and Structure

The synthetic dataset will consist of approximately 10,000 records with the following features categorized by type:

1. Demographic Features:

   - Age (10-19 years, discrete)

   - Gender (Male/Female, categorical)

   - Location (Urban/Rural, categorical)

   - Education Level (Primary/Secondary/Tertiary, categorical)

   - Socioeconomic Status (Low/Medium/High, categorical)

2. Clinical Features:

   - CD4 Count (numerical, cells/mm^3) - Immune system strength indicator

   - Viral Load (numerical, copies/mL) - Measure of HIV virus in blood

   - Treatment Duration (months, numerical) - Time on ART

- Side Effects (None/Mild/Severe, categorical)
- Comorbidities (None/One/Multiple, categorical) - Other health conditions
- Drug Regimen (First-line/Second-line/Third-line, categorical)

3. Social and Environmental Features:

- Family Support (Low/Medium/High, categorical)
- Distance to Clinic (0.0-0.9, continuous) - Distance in normalized units
- Awareness Status (0/1, binary) - Whether adolescent knows their HIV status
- Stigma Experience (Low/Medium/High, categorical)
- Peer Support (Yes/No, binary) - Access to peer support groups

4. Target Variable:

- Adherence Status (0/1, binary) - Where 0 represents good adherence and 1 represents poor adherence

5. Engineered Features (to be created during analysis):

- Age Group (10-13/14-16/17-19, categorical)
- Support-Awareness Interaction (binary)
- Location-Distance Interaction (continuous)
- Viral Suppression Status (Suppressed/Unsuppressed, categorical) - Derived from viral load values
- Treatment Complexity Score (numerical) - Composite score based on regimen type and side effects

Relevance of Attributes to the Study

Each attribute was selected based on its documented influence on ART adherence in the literature, with appropriate citations supporting their inclusion.

Database Implementation with XAMPP

The project will utilize XAMPP as the local development environment for creating and managing the MySQL database. XAMPP provides an integrated platform that includes Apache, MySQL, PHP, and Perl, making it ideal for developing and testing the database components of this project.

The database schema will be implemented with normalized tables for demographic, clinical, and social data, with appropriate primary and foreign key relationships. This relational structure will facilitate queries across different data dimensions and enable efficient data retrieval for analysis.

Tools and ML Resources Required

1. Programming Languages:

   - Python 3.8+ as the primary language

2. Core Libraries and Frameworks:

   - Data Manipulation: Pandas, NumPy

   - Database Connectivity: MySQL Connector for interfacing with XAMPP database

   - Visualization: Matplotlib, Seaborn

   - Statistical Analysis: SciPy

3. Machine Learning Libraries:

   - Scikit-learn for model building and evaluation

   - Imbalanced-learn for handling class imbalance (SMOTE)

   - XGBoost for ensemble models

4. Development Environment:

- XAMPP for local database hosting and management
- Jupyter Notebook for exploratory analysis
- Google Colab for computational resources

Exploratory Data Analysis (EDA)

Planned Data Overview and Distribution Analysis

The initial exploration of the synthetic dataset will examine:

- Distribution of the target variable (Adherence Status) to check for potential class imbalance
- Distribution patterns of key features including demographic, clinical, and social factors

Demographic, Clinical, and Environmental Factor Analysis Plan

Analysis will examine how various factors relate to adherence patterns, including:

- Age, gender, education, and socioeconomic effects on adherence
- Relationships between clinical markers (CD4, viral load) and adherence
- Impact of treatment characteristics (duration, side effects, regimen)
- Effect of social and environmental factors (family support, distance to clinic, stigma)

Feature Correlation Analysis Plan

The study will conduct correlation analysis to examine relationships between features that could inform both model development and potential intervention strategies, focusing on key relationships between clinical, demographic, and social factors.

Feature Importance Analysis Plan

The study will conduct feature importance analysis to identify the most influential factors for adherence prediction, using:

- Logistic regression coefficients for interpretable importance weights

- Feature importance from tree-based models (Random Forest, XGBoost)

- Permutation importance for model-agnostic assessment

Methodology

Data Preparation and Preprocessing

The synthetic dataset will undergo comprehensive preprocessing to ensure optimal model performance:

1. Data Generation:

   - Synthetic data will be generated to simulate real-world characteristics including missing values, inconsistencies, and the natural variability found in clinical datasets

   - Data generation will incorporate appropriate statistical distributions based on published literature

2. Data Cleaning:

   - Handling missing values through appropriate imputation techniques

   - Identifying and addressing outliers using statistical methods

   - Correcting inconsistencies in categorical variables

3. Feature Engineering:

   - Creating age group categories to capture developmental stages

   - Generating interaction features between related variables

   - Deriving clinical status indicators and composite scores

4. Data Transformation:

- Normalizing numerical features to ensure consistent scale
- Encoding categorical variables using appropriate techniques
- Applying transformations to skewed distributions

5. Class Imbalance Handling:

- Implementing SMOTE (Synthetic Minority Over-sampling Technique) to address potential imbalance in adherence status
- Exploring class weighting approaches

Model Development and Evaluation

The project will implement and compare several machine learning algorithms:

1. Logistic Regression:

- Baseline model with interpretable coefficients
- L1 and L2 regularization to prevent overfitting
- Analysis of odds ratios for clinical interpretation

2. Decision Tree Classifier:

- Visualization of decision rules for practical implementation
- Pruning to prevent overfitting
- Extraction of decision paths for high-risk profiles

3. Random Forest Classifier:

- Ensemble approach to improve prediction accuracy
- Built-in feature importance for factor ranking
- Partial dependence plots for understanding feature effects

4. Gradient Boosting Classifier (XGBoost):
   - Advanced ensemble technique for maximizing predictive power
   - Hyperparameter tuning via grid search and cross-validation

Models will be evaluated using appropriate metrics:

- Area Under the ROC Curve (AUC-ROC)
- Precision, Recall, and F1-Score
- Balanced Accuracy
- Confusion Matrix Analysis

Cross-validation (5-fold) will be employed to ensure robust performance assessment, with stratification to maintain class distribution across folds.

Feature Importance Analysis

A central component of this research is analyzing feature importance to identify key factors influencing ART adherence:

1. Model-Based Importance:
   - Extraction of feature coefficients from Logistic Regression
   - Feature importance from tree-based models (Random Forest, XGBoost)

- Permutation importance for model-agnostic assessment

2. Visualization of Feature Effects:
   - Partial Dependence Plots (PDP) to visualize the marginal effect of features
   - Feature interaction visualizations
   - Comparison of importance rankings across different models

The combination of these methods will produce a ranking of factors influencing adherence, providing valuable insights for intervention design and resource allocation.

Expected Outcomes and Applications

Predictive Model Deliverables

The project will produce the following technical deliverables:

1. A machine learning model capable of predicting ART adherence risk
2. A ranking of factors influencing adherence outcomes
3. Basic visualizations showing key feature relationships and importance
4. A database schema for organizing adherence-related data

Practical Applications

The research findings will support several practical applications in clinical settings:

1. Identification of adolescents at risk of poor ART adherence for targeted interventions
2. Evidence-based support strategies based on identified risk factors
3. Training materials and guidelines for healthcare providers
4. Data-driven suggestions for health system improvements

Conclusion

This approach to predicting and understanding ART adherence among Kenyan adolescents integrates data analysis with clinical relevance. By developing a model that predicts risk and explains the factors driving that risk, this research aims to bridge the gap between statistical insight and practical intervention. The emphasis on feature importance analysis provides a link between prediction and action, enabling healthcare providers to develop strategies that address the most influential factors for each individual.