INDONESIAN JOURNAL OF INNOVATION AND APPLIED SCIENCES (IJIAS)
Journal Homepage: https://ojs.literacyinstitute.org/index.php/ijias
ISSN: 2775-4162 (Online)
Research Article

Volume 5	Issue 3	October (2025)	DOI: 10.47540/ijias.v5i3.2275	Page: 277 – 304

Developing a Logistic Regression Machine Learning Model that Predicts Viral Load Outcomes for Children Living with HIV in Gutu District, Zimbabwe

Belinda Ndlovu[1], Fungai Jacqueline Kiwa[2], Martin Muduva[3], Colletor T. Chipfumbu[3], Sheltar Marambi[3]

[1]Informatics and Analytics, National University of Science and Technology, Zimbabwe
[2]Chinhoyi University of Technology, Zimbabwe
[3]Midlands State University, Zimbabwe
Corresponding Author: Belinda Ndlovu; Email: belinda.ndlovu@nust.ac.zw

ARTICLE INFO

Keywords: HIV, Logistic Regression, Machine Learning, Viral Load.

Received : 18 September 2025
Revised : 13 October 2025
Accepted : 30 October 2025

ABSTRACT

HIV remains a major public health issue globally, particularly in poor resource settings such as the Gutu district of Zimbabwe. The study aimed to develop a predictive viral load outcome model for HIV children based on the CRISP-DM research process. Secondary clinical data for children aged 0–17 years in Gutu were retrieved from the Demographic Health Information System (DHIS2). The study identified age, adherence status, gender, and geographical location as correlated with viral load outcomes. A supervised machine learning logistic regression model was trained with data balance and proper management of complexities. Grid search-based hyperparameter tuning was performed to improve model performance further. The evaluation metrics were accuracy, sensitivity, F1 Score, and area under the receiver operating characteristic curve (AUC-ROC). The model's performance resulted in 89% accuracy, with all the metrics showing a strong performance. A confusion matrix was used to visualize the results. The findings add to the knowledge on viral load outcome prediction and HIV care in Zimbabwe. The findings suggest that early diagnosis and targeted interventions can improve viral load outcomes in children in Gutu. For future research, the development of the model will be based on more representative data sets and applied to other settings to determine differences in other regions and understand the dynamics of HIV care in children.

INTRODUCTION

Human Immunodeficiency Virus, or HIV, continues to pose a major global health issue. This is especially true for vulnerable children (Cornell et al., 2019). Viral load (VL) is a primary indicator of the virus's presence. The ideal outcome for HIV-positive children is target not detected, or TND (Davies, 2020). Laboratories routinely monitor viral load levels. Many countries work hard to optimize these levels, particularly in areas with limited resources. Machine learning approaches, like logistic regression, show real promise in handling the complexities of HIV data analysis (Shamount et al., 2022).

Worldwide, about 78 percent of individuals on antiretroviral therapy, or ART, reach viral suppression. That is still short of the World Health Organization's 95 percent goal (Eamsakulrat et al., 2022). Even in wealthier nations, suppression rates for children fall below expectations. For example, just 73 percent of Canadian children sustained suppression over three years (Kakkar et al., 2020). In China, rates varied between 60 and 78 percent (Lao et al., 2023). South Africa saw 73 percent of HIV-positive children achieving suppression (Cornell et al., 2019). Sub-Saharan Africa shoulders much of the burden. Suppression rates ranged from 40 to 77 percent across 12 countries (Hladik et al., 2023). Zimbabwe managed 60 percent suppression,

but children's health outcomes warrant closer attention (Hladik et al., 2023). ART has shifted HIV from a deadly illness to a manageable chronic one (Bouchard et al., 2022). However, the global UNAIDS targets the final 95 percent for suppression (Bouchard et al., 2022).

Sustaining viral suppression is essential for successful HIV management. Current strategies emphasize reaching undetectable levels (Bouchard et al., 2022). Children encounter unique hurdles. These include intricate medication schedules, reliance on caregivers, economic obstacles, and co-occurring infections. Such issues intensify in low-resource areas like Gutu District. Spotting detectable viral load early allows for prompt action. With more clinical data now accessible, predictive models become feasible. Logistic regression stands out for predicting binary outcomes, such as health risks, based on key factors (Nusinovichi et al., 2020). This study develops a logistic regression model from routine data on children in Gutu District. It identifies major predictors of viral load results. Ultimately, it aims to inform better clinical choices.

Despite advancements in HIV therapy, children living with HIV in Zimbabwe's Gutu district continue to face poor viral load (VL) outcomes, resulting in increased vulnerability to opportunistic infections, compromised physical development, and a higher risk of treatment failure due to potential drug resistance (Machekano et al, 2023). With a previously suppressed viral load outcome, a person's viral load count is checked after 1 year, which may be too late, especially in cases of early viral rebounding. The general VL suppression rate for people living with HIV in Gutu is 83% which is way below the targeted 95% of the UNAIDS's targets (Conan et al, 2020). A study on the rural part of the country showed VL suppression rates of 80% for children (Machekano et al, 2023). Due to socioeconomic circumstances, restricted access to medical care as a result of long distances, and poor antiretroviral therapy (ART) adherence, many of these children are at risk of having a detectable, and in worst cases, a high viral load making them prone to weak bodies that can easily be attacked by other diseases (Bouchard, et al, 2022). At the teen stages, some of these children become sexually active, stop medication due to stigma, and hence spread HIV if their viral load is not controlled. Therefore, early detection of their viral load outcomes is essential for prompt treatment and intervention. Machine learning presents a viable method to improve the predictive accuracy of viral load outcomes. No localized models are suited to the unique circumstances of Gutu District's vulnerable children. This research aims to create a localised machine learning model that can precisely forecast the results of viral loads for Gutu children, enabling proactive actions for areas with similar settings in Zimbabwe, hence enhancing the Zimbabwean children population's health.

METHODS

The conceptual framework employed in this study draws on established theories and empirical evidence concerning health behaviors, with a particular focus on HIV management through viral load monitoring. It integrates key insights from the Health Belief Model, Statistical Learning Theory, and Social Detrimental Theory. In doing so, the framework offers a thorough perspective on the factors that shape viral load outcomes in vulnerable populations. Evidence indicates that such an approach can illuminate individual perceptions and broader social influences. Figure 1 demonstrates the interactions among these components and their effects on health behaviors.

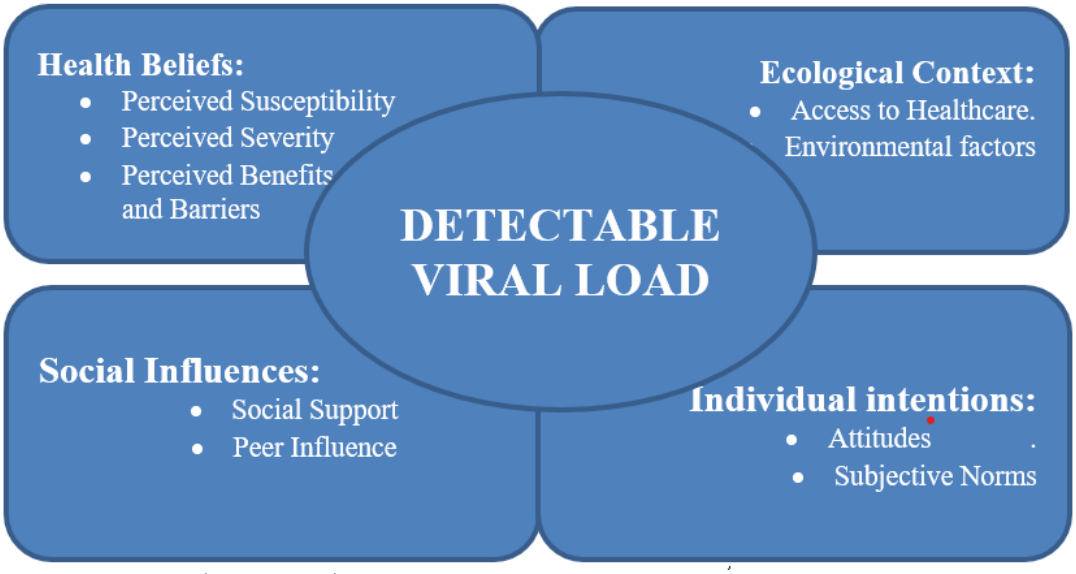


Figure 1. Conceptual Framework

Perceived susceptibility to HIV-related health risks among children appears to heighten motivation for treatment adherence. This effect strengthens when supportive peers and family are involved. On the other hand, social stigma can undermine care-seeking, even with strong positive intentions. These dynamics shape the study's methodology, from data collection to analysis. The research focuses on health beliefs, social and ecological influences, and personal intentions. Ultimately, it seeks to pinpoint intervention opportunities that enhance adherence in vulnerable children.

Health Beliefs

Health beliefs play a very important role in influencing an individual's adherence to HIV treatment, and they take into consideration perceived susceptibility, perceived severity, perceived benefits, and barriers. Perceived susceptibility involves an individual's assessment of their own risk for contracting HIV or facing elevated viral loads stemming from treatment non-adherence. Evidence indicates that greater perceived susceptibility tends to boost motivation for sticking to prescribed regimens. Following that, perceived severity centers on beliefs regarding the gravity of life with HIV, along with the fallout from skipping doses. Such understandings of potential health repercussions shape behaviors promoting better adherence. Finally, perceived benefits and barriers come into play as people balance the upsides of consistent treatment, like improved health outcomes, against hurdles such as medication side effects or logistical challenges in access. Consequently, interventions prove effective when they amplify awareness of benefits while addressing and reducing those barriers.

Social Influences

Social support plays a central role in this framework. Family members, friends, and healthcare providers help promote treatment adherence. This, in turn, leads to suppressed viral loads. Evidence indicates that robust social networks motivate individuals to maintain their regimens consistently. Additionally, peer influence also contributes significantly. Positive relationships with peers can strengthen healthy behaviors in children and adolescents. Consequently, social dynamics emerge as key factors in driving changes in health behaviors.

Ecological Context

This component explores how broader social, economic, and cultural contexts shape health behaviors. Evidence indicates that factors like socioeconomic status, community resources, and cultural beliefs significantly influence individuals' experiences in managing HIV. Access to healthcare plays a pivotal role; i.e., the availability of services, such as counseling and support groups, proves essential for fostering adherence. This highlights the need for systemic interventions aimed at improving healthcare accessibility.

Consequently, environmental factors also matter greatly, with different elements, including caregiver support and household stability, contributing crucially to adherence and achieving undetectable viral loads in children. A child's

surroundings often carry stigma from neighboring families. Empowering those families with accurate knowledge about HIV can make managing the condition far easier for affected children.

Individual Intentions

Attitudes: Individual attitudes toward treatment, shaped by beliefs and experiences, are vital in determining adherence. Positive attitudes toward medication and health can lead to better adherence outcomes. Subjective Norms: The perceptions of what others believe about HIV treatment can influence an individual's intentions to adhere. Norms can either encourage or discourage adherence to behaviours.

This chapter describes the research approach for building a logistic regression model to predict viral load outcomes among children with HIV in Gutu District, Zimbabwe. Predictive analytics served as the core method, emphasizing model development, performance evaluation, and examination of demographic, health, and socioeconomic factors. The approach drew from existing literature on key determinants of viral load suppression. The Cross Industry Standard Process for Data Mining (CRISP-DM) framework was applied to guide the data science effort, offering a structured pathway for predictive modeling.

Research Philosophy

Research philosophies encompass theories about the nature of the world being studied and how knowledge about that reality is produced and justified (Mauthner, 2020). Logistic regression serves as a quantitative method for predicting outcomes. It aligns closely with post-positivism, a paradigm that prioritizes identifying and measuring relationships among variables to explain and forecast phenomena. In this study, the approach addresses the complex factors shaping viral load outcomes in children from Gutu District. Post-positivism contributes to methodological rigor through careful data collection and robust statistical analysis, yielding reliable findings. By adopting this philosophy, the research employs statistical methods to forecast outcomes while navigating inherent complexities. The focus remains on enhancing children's lives, fostering transparency, and leveraging numerical data to uncover patterns, trends, and relationships.

Research Design

Cross Industry Standard Process for Data Mining (CRISP-DM)

The researcher adopted the CRISP-DM methodology as it was suitable for the study. The Cross Industry Standard Process for Data Mining (CRISP-DM) defines a process that provides a framework for carrying out data mining projects (Wirth & Hipp, 2000). The process model is being developed by a consortium of leading data mining users and suppliers, while partly sponsored by the European Commission under the Early Stage Program: Research-Innovation-Training (ESPRIT) program (Project number 24959), mining projects which are independent of both the industry sector and the technology used (Wirth et al., 2000). The CRISP-DM process model intends to make big data mining projects less costly, more reliable, controllable, and faster. Figure 2 shows the six stages of the CRISP-DM methodology.

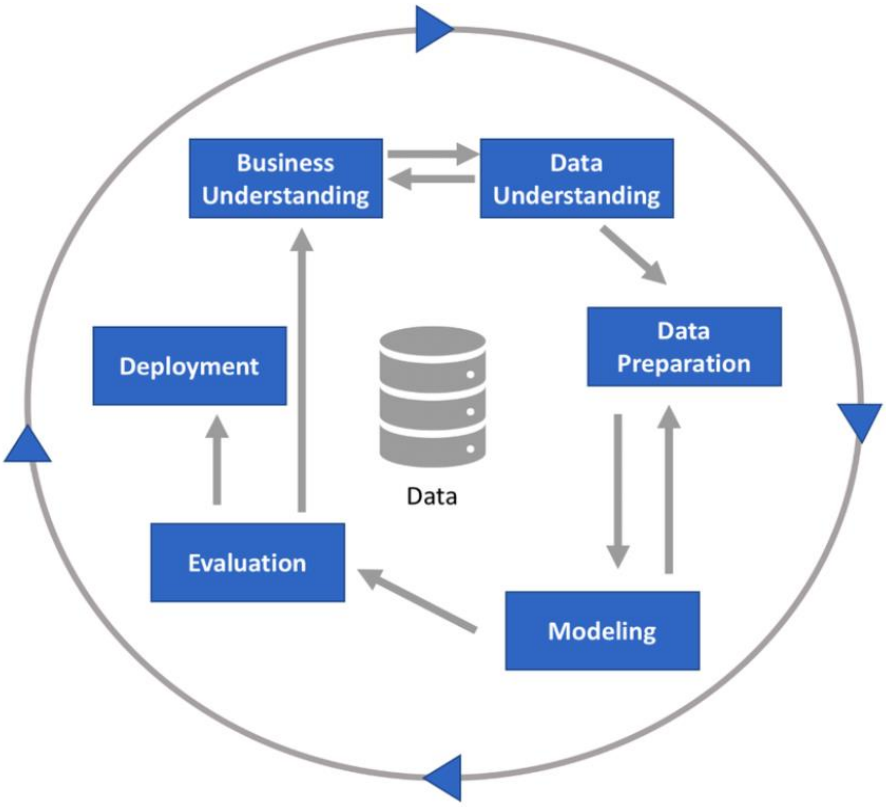


Figure 2. CRISP-DM

Business Understanding

This is the first phase of the Cross-Industry Standard Process for Data Mining (CRISP-DM) architecture, known as Business Understanding. The study has a research framework for developing a predictive logistic regression model for children with HIV in Gutu district. By 2030, the World Health Organization (WHO) and the UN intend to see a 95-95-95 target reached, where the last 95 of viral suppression has been lagging over the years (Bouchard et al., 2022). Determining the factors associated with viral load outcomes might help healthcare workers develop possible intervention strategies that may increase viral suppression rates. The study plan was to develop a logistic regression machine learning model that accurately predicts child viral load outcomes based on various demographic, health, and socioeconomic factors. The model will start by analysing the factors associated with the viral load outcomes, then, based on those factors, build a model that can predict VL outcomes at any given time.

Data understanding

Figure 3 depicts the website from which the data was downloaded.

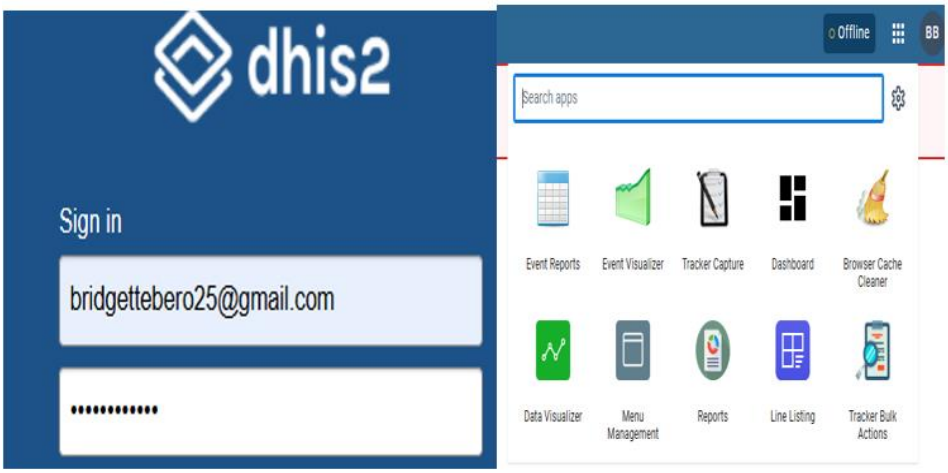


Figure 3. DHIS2 Login

The relevant data was initially downloaded from DHIS2 as a CSV file using the event reports application. The data contained 646 records with variables shown in the table below. The variables are as downloaded from the system. Data was downloaded specifically for children between the ages of 0-17 with a valid viral load result as of the end of 2024. Table 1 shows different attributes that are contained in the data downloaded above and are shown with all their descriptions.

Table 1. Data Set Variable Description

Attribute	Description
Organ	District where the child is monitored as an HIV positive patient.
Organisation	The ward, within the district, where the child lives.
OVC ID	The OVC MIS generates a unique identifier for each child to identify children enrolled in OVC PEPFAR-funded programs uniquely.
Has the beneficiary been linked to ART	Data points that check whether the child is on ART, given that they are HIV positive. (2nd 95 of the UN targets)
ART Number	Another unique identifier generated by the Zimbabwe Ministry of Health and Child Care (MOHCC) to uniquely identify HIV positive clients without exposing their PII.
Sex	Whether the child is male or female
Priority Population	The vulnerability status of that child made them eligible to be tracked and monitored by PEPFAR-funded OVC programs.
HIV Status	The child's HIV status, which a health facility confirms.
Date of birth	The date when the child was born
Age	The age of the child as of 31/12/2024
Health facility	The health center is where the child receives ART medication, ART adherence support, and viral load monitoring.
Date of ART Initiation	When the child was initiated on ART, as informed by the health facility.
ART duration	For how long the child has been on ART since the time of initiation to the end of 2024.
ART status	Classification of the child's adherence, whether it is good or poor, based on Zimbabwe health SOPs.
Adherence Status	Classification of the child's adherence, whether good or poor, is based on Zimbabwe health SOPs.
ZW-Viral load result	The child's actual, valid viral load result as informed by the health facility and confirmed by any Zimbabwe medical laboratory.
VL Status	The status of the viral load result, whether it is still detectable or not detectable.

The data was loaded into Python for analysis. The data was explored using Jupyter Notebook to understand it. Figures 4 and 5 depict snippets of the explanatory data analysis the researcher carried out.

```
import pandas as pd

try:
    df = pd.read_csv('gutu2vlreport311224.csv')
    display(df.head())
    print(df.shape)
except FileNotFoundError:
    print("Error: 'gutu2vlreport311224.csv' not found. Please ensure the file exists in the current directory.")
    df = None # Assign None to df in case of error
except Exception as e:
    print(f"An unexpected error occurred: {e}")
    df = None
```

Figure 4. Loading the data set code

	Organ	Organisation unit name	OVC ID	Has beneficiary been linked to ART?	ART Number	Sex	Priority population 1	HIV Status	Date of Birth	Age	Health Facility for ART	Date of ART initiation	ART duration as @ 2024	A statu
0	Gutu	Gutu ward 21	NAPEtu011117m0	1	08-04-27-2022-A-00012	Male	Children/ adolescents living with HIV	Positive	1/11/2017	7	Gutu - Devure - 100387 - Mission Clinic	1/3/2022	2	I
1	Gutu	Gutu ward 17	AIREtu160516m0	1	08-05-0A-2017-A-00381	Male	Children/ adolescents living with HIV	Positive	16/5/2016	8	Gutu - Chimombe - 100215 - Rural Hospital	20/12/2017	7	I
2	Gutu	Gutu ward 19	AIREtu110307f0	1	08-04-0B-2011-A-00143	Female	Children/ adolescents living with HIV	Positive	11/3/2007	17	Gutu - Chimombe - 100215 - Rural Hospital	22/3/2011	13	I
3	Gutu	Gutu ward 22	AIKAtu280508f0	1	08-04-28-2012-A-00036	Female	Children/ adolescents living with HIV	Positive	28/5/2008	16	Gutu - Magombedze - 100822 - Clinic	19/1/2015	9	Acti
4	Gutu	Gutu ward 36	NEHEtu280411f0	1	08-04-0B-2018-A-00022	Female	Children/ adolescents living with HIV	Positive	28/4/2011	13	Gutu - Chimombe - 100215 - Rural Hospital	26/4/2018	6	I

```
(646, 17)
```

Figure 5. Loading the data set code

Figures 6 and 7 present the code used to visualize the data in the CSV file.

```
# 4. Correlation Analysis (for numerical features)
numerical_features = df.select_dtypes(include=['number'])
correlation_matrix = numerical_features.corr()
plt.figure(figsize=(10, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm')
plt.title('Correlation Matrix of Numerical Features')
plt.show()

#5. Relationship with Target Variable
for col in df.columns:
    if df[col].dtype == 'object' and col != 'VL status':
        plt.figure(figsize=(10,6))
        df.groupby(col)['VL status'].value_counts().unstack().plot(kind='bar', stacked=True)
        plt.title(f'VL status by {col}')
        plt.show()
    elif df[col].dtype in ['int64','float64'] and col != 'VL status':
        plt.figure(figsize=(10,6))
        df.boxplot(column=col, by='VL status')
        plt.title(f'VL Status by {col}')
        plt.show()
```

Figure 7.Visualizing the data set code

```
import matplotlib.pyplot as plt
import seaborn as sns

# 1. Data Overview
display(df.info())
display(df.describe())

# 2. Missing Values
missing_values = df.isnull().sum()
print("Missing Values:\n", missing_values)

# 3. Target Variable Analysis
vl_status_counts = df['VL status'].value_counts()
print("VL Status Counts:\n", vl_status_counts)
plt.figure(figsize=(8, 6))
vl_status_counts.plot(kind='bar')
plt.title('Distribution of VL Status')
plt.xlabel('VL Status')
plt.ylabel('Frequency')
plt.show()
```

Figure 8. Visualizing the data set code

Figure 7 shows the code snippets that are used for visualizing the data to see basic statistics, which include the minimum and maximum values for each attribute, the means, the standard deviations from the means, the data types in each attribute, and the number of records in each attribute. Each attribute was also checked for any missing values. Relationships between each variable and the target variable (Viral load status) were explored and visualised using different suitable visuals to understand the data further. Figure 8 depicts the code snippet for displaying the data's rows and columns.

```
# 6. Data Shape
print(df.shape)
```

Figure 9. Displaying the number of rows and columns

Figure 8 shows the stages that are involved in visualizing the data to see basic statistics, which include the minimum and maximum values for each attribute, the means, the standard deviations from the means, the data types in each attribute, and the number of records in each attribute. Each attribute was also checked for any missing values. Relationships between each variable and the target

variable (Viral load status) were explored and visualised using different suitable visuals to further understand the data.

Data quality is a very important aspect when dealing with data, as it can have social and economic impacts (Wang et al., 2020). The researcher, therefore, also ensured that the data used was of high quality. For this particular study, the data used were secondary data extracted from an information system. There were prior data quality measures that were put in place during the collection and entry of the data. These measures ensured that the data met the data quality dimensions that include completeness, consistency, and reliability.

Data Preparation

At this stage, data was prepared to a state where it was suitable for model building and development. The researcher carried out data wrangling using Jupyter Notebook to develop the final data set for model development. Below are snippets of the data munging steps that were taken:

Figure 9 shows the first step in data preparation, which is data cleaning.

```
# Convert 'Age' to numeric, handling errors
df['Age'] = pd.to_numeric(df['Age'], errors='coerce')

# Check for and handle inconsistencies in sex column
df['Sex'] = df['Sex'].str.lower()  # Convert to lowercase for consistency

# Handling inconsistencies in 'VL status' (if any)
df['VL status'] = df['VL status'].str.strip()

# Checking data types and inconsistencies after cleaning
display(df.info())
display(df.describe(include='all'))
```

Figure 10. Data Cleaning Code

A prior data verification and cleaning was conducted before the actual wrangling of data.

Figure 10 presents the result of the data cleaning above.

```
import matplotlib.pyplot as plt

# Pie chart for 'VL status' distribution
plt.figure(figsize=(6, 6))
df['VL status'].value_counts().plot.pie(autopct='%1.1f%%', startangle=90)
plt.title('Distribution of VL Status')
plt.ylabel('')
plt.show()

# Visualizations for categorical variables
categorical_cols = ['Organ', 'Organisation unit name', 'Sex', 'Priority population 1', 'Art status', 'Adherence status']
for col in categorical_cols:
    plt.figure(figsize=(10, 6))
    df.groupby(col)['VL status'].value_counts().unstack().plot(kind='bar', stacked=True)
    plt.title(f'VL status by {col}')
    plt.xlabel(col)
    plt.ylabel('Frequency')
    plt.xticks(rotation=45, ha='right')
    plt.tight_layout()
    plt.show()

# Visualizations for numerical variables
numerical_cols = ['Age', 'ART duration as @ 2024', 'ZW- Viral load results']
for col in numerical_cols:
    plt.figure(figsize=(8, 6))
    df.boxplot(column=col, by='VL status')
    plt.title(f'VL Status by {col}')
    plt.suptitle('') # Removing the default suptitle
    plt.show()
```

Figure 11. Data visualization after cleaning

Data was also visualized to appreciate the noticeable trends and factors associated with viral load outcomes before model building. Figures 11 and 12 show the next step in data preparation, which is Feature Engineering.

```
import numpy as np
import pandas as pd
from sklearn.preprocessing import StandardScaler

# Identifying categorical columns (excluding the target variable)
categorical_cols = ['Organ', 'Organisation unit name', 'Sex', 'Priority population 1', 'Art status', 'Adherence status']

# Creating dummy variables
df_encoded = pd.get_dummies(df, columns=categorical_cols, drop_first=True)

# Analyzing and transforming continuous features
# Log transformation for 'ZW- Viral load results' if it has a skewed distribution
# Checking for zero values first, add 1 before log if any zeros exist
if (df_encoded['ZW- Viral load results'] == 0).any():
    df_encoded['log_viral_load'] = np.log1p(df_encoded['ZW- Viral load results'])
else:
    df_encoded['log_viral_load'] = np.log(df_encoded['ZW- Viral load results'])
# 2. Standardization for 'Age' and 'ART duration as @ 2024'
scaler = StandardScaler()


# Fitting the scaler on the entire dataset
df_encoded[['Age', 'ART duration as @ 2024']] = scaler.fit_transform(df_encoded[['Age', 'ART duration as @ 2024']])


# Droping original columns after log transformation and standardization
df_encoded = df_encoded.drop(['ZW- Viral load results'], axis = 1)

display(df_encoded.head())
```

Figure 12. Feature Engineering

Sex_male	Art status_IIT	Art status_LTFU	Art status_MA	Adherence status_Poor adherence	log_viral_loac
True	True	False	False	True	0.000000
True	True	False	False	True	0.000000
False	True	False	False	True	3.433987
False	False	False	False	False	3.433987

Figure 13. Feature Engineering

In this stage, the researcher conducted feature engineering to improve model accuracy on unseen data. This was done by creating dummy variables for categorical variables and carrying out log transforms, standardization, and scaling for the appropriate continuous variables. After this stage, the data was ready for model development.

Modelling

The right machine learning model was a crucial choice in creating a viral load prediction model. The study employed a logistic regression supervised machine learning model because the study's outcome variable was binary. Logistic regression is one of the best models, specifically designed to model the relationship between a set of

independent variables and a binary dependent variable (Solutions, 2020). Data was split into training and testing sets as shown below. Figure 13 now shows how the data was split accordingly.

```
from sklearn.model_selection import train_test_split

# Defining features (X) and target (y)
X = df_encoded.drop('VL status', axis=1)
y = df_encoded['VL status']

# Splitting data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42, stratify=y)
```

Figure 14. Data splitting code

The following code was used to develop and train the logistic regression model using the training data set.

```
In [14]: from sklearn.linear_model import LogisticRegression

         # Drop 'Date of Birth', 'Date of ART initiation', and 'Health Facility for ART' columns
         X_train = X_train.drop(['Date of Birth', 'Date of ART initiation', 'Health Facility for ART'], axis=1)
         X_test = X_test.drop(['Date of Birth', 'Date of ART initiation', 'Health Facility for ART'], axis=1)

         # Initialize and train the model
         logreg = LogisticRegression(random_state=42, solver='liblinear', max_iter=1000)
         logreg.fit(X_train, y_train)

Out[14]: LogisticRegression(max_iter=1000, random_state=42, solver='liblinear')
```

Figure 15. Logistic Regression model code

Initially, the model could not be trained due to columns that contained string values, which led to dropping those columns before the actual model training. The regression model from sklearn. linear_model was initialised with specific parameters. The "random_state" parameter is used to control the random number generator, which shuffles the data. This parameter was set to 42, ensuring that the results from the model are reproducible. The "solver" parameter specifies the algorithm to be used for optimisation. It was set to bilinear because it was the most suitable due to the size of the data set, which was small to medium.

The "max_iter" parameter indicates how many iterations the optimization process must run to converge. In this case, it was set to a maximum of 1000 training iterations.

Evaluation

A crucial step after the modelling phase in the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology is the evaluation phase, where the model's performance across training runs is evaluated, which helps to see how it is performing. The testing data set was used for this stage. The following evaluation matrices were employed to evaluate the model:

Accuracy – this metric measures the extent to which the model makes correct predictions, i.e., Accuracy = Number of correct predictions/Total number of predictions. A high accuracy reflects the model's ability to predict most classes well.

Precision = True positives/(True positives + False positives). High precision is an indication of minimised false positives, thus good model performance.

Recall – this metric focuses on the model's ability to identify all the actual positive instances. Recall = True positives/(True positives + False negatives). A high recall minimises false negatives, thus resulting in good model performance.

F1 Score – this metric is a metric that ranges from 0 – 1, which is a combination of precision and recall, balancing the minimisation of false positives and false negatives. F1 Score = 2*[(Precision * Recall)/(Precision + Recall)]. A higher F1-score indicates a better balance between precision and recall.

AUC-ROC (Area under the Receiver Operating Characteristic Curve) – this metric evaluates the model's ability to distinguish between positive and negative classes.

The higher the AUC, the better the model does at distinguishing. If the AUC is 0.5, the model is as good as random guessing. Confusion matrix – a table that visualises the performance of a classification model. It shows the counts of true positives (top left), true negatives (bottom right), false positives (top right), and false negatives

(bottom left). The metric provides a clearer picture of the mistakes the model is making.

Figure 15 are snippets of the code the researcher employed to evaluate the developed model on the testing data set.

```
# Calculating evaluation metrices
try:
    accuracy = accuracy_score(y_test, y_pred)
    precision = precision_score(y_test, y_pred, average='weighted', zero_division=0)
    recall = recall_score(y_test, y_pred, average='weighted', zero_division=0)
    f1 = f1_score(y_test, y_pred, average='weighted', zero_division=0)

    # Calculating AUC-ROC
    y_prob = logreg.predict_proba(X_test)[:, 1]  # Probability of the positive class
    auc_roc = roc_auc_score(y_test, y_prob)

    print(f"Accuracy: {accuracy:.4f}")
    print(f"Precision: {precision:.4f}")
    print(f"Recall: {recall:.4f}")
    print(f"F1-score: {f1:.4f}")
    print(f"AUC-ROC: {auc_roc:.4f}")

except ValueError as e:
    print(f"Error calculating metrics: {e}")

# Printing of the classification report
print("\nClassification Report:\n", classification_report(y_test, y_pred, zero_division=0))

# Generating and visualizing confusion matrix
cm = confusion_matrix(y_test, y_pred)
plt.figure(figsize=(8, 6))
sns.heatmap(cm, annot=True, fmt="d", cmap="Blues",
            xticklabels=logreg.classes_, yticklabels=logreg.classes_)
plt.xlabel("Predicted")
plt.ylabel("Actual")
plt.title("Confusion Matrix")
plt.show()
```

Figure 16. Model Evaluation Code

The outcomes of the evaluation process were assessed to see if the model was performing well. After reviewing, attempts to optimize the model performance were made by hyperparameter tuning using the GridSearchCV, as shown in Figure 16.

```
from sklearn.model_selection import GridSearchCV
from sklearn.metrics import make_scorer, f1_score
from sklearn.preprocessing import LabelEncoder

# Convert the target variable to numerical labels
le = LabelEncoder()
y_train_encoded = le.fit_transform(y_train)

# Defining the parameter grid
param_grid = {
    'C': [0.01, 0.1, 1, 10, 100],
    'penalty': ['l1', 'l2'],
    'solver': ['liblinear']
}

# Defining the scoring metric (using f1_weighted)
scoring = {'f1_weighted': make_scorer(f1_score, average='weighted')}

# Initializing GridSearchCV
grid_search = GridSearchCV(estimator=logreg, param_grid=param_grid, scoring=scoring, refit='f1_weighted', cv=5)

# Fitting GridSearchCV to the training data
grid_search.fit(X_train, y_train_encoded)

# Printing the best hyperparameters and score
print("Best Hyperparameters:", grid_search.best_params_)
print("Best F1-weighted Score:", grid_search.best_score_)

# Getting the best estimator
best_logreg = grid_search.best_estimator_
```

Figure 17. Model Optimization code

The results of this model evaluation were adopted by the researcher, taking note of all limitations to the models' performance.

Ethical Considerations

Ethical considerations were highly important in this research since it involved sensitive medical data for a vulnerable population, which is children living with HIV. The researcher considered the confidentiality principle by seeking approval to use the data to carry out the research. The researcher also downloaded patient-level data, excluding personal identifiable information (PII), to maintain data privacy. The researcher also acknowledged the prior work of other researchers, adhering to the university's criteria. The researcher analysed the data with integrity, avoiding biases and errors as these could promote outcomes that would potentially negatively impact the Zimbabwe Health Sector. A steadfast commitment to ethical considerations was maintained. To project the identity of the individuals, pseudonyms, i.e., P1, P2, P3, P4, and P5, were used in such a way to maintain anonymity.

RESULTS AND DISCUSSION

This chapter presents, analyses, and discusses the results of the study on a viral load prediction model for children living with HIV in Gutu district. Different visuals were used to present the results of the data analysis. The 5 stages of the machine learning pipeline were followed in the presentation and analysis of the data results. These stages include data exploration, data pre-processing, modelling, model evaluation, and model optimisation. The researcher used Python to analyse the data.

Data Exploration results

Explanatory data analysis results showed properties of the data, trends within the data, and possible factors associated with viral load status. Results from the explanatory data analysis that was carried out are discussed. Figure 17 shows the summary of the data contained in our file.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 646 entries, 0 to 645
Data columns (total 17 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   Organ                                 646 non-null    object
 1   Organisation unit name                646 non-null    object
 2   OVC ID                                646 non-null    object
 3   Has beneficiary been linked to ART?   646 non-null    int64
 4   ART Number                            646 non-null    object
 5   Sex                                   646 non-null    object
 6   Priority population 1                 646 non-null    object
 7   HIV Status                            646 non-null    object
 8   Date of Birth                         646 non-null    object
 9   Age                                   646 non-null    int64
 11  Date of ART initiation                646 non-null    object
 12  ART duration as @ 2024                646 non-null    int64
 13  Art status                            646 non-null    object
 14  Adherence status                      646 non-null    object
 15  ZW- Viral load results                646 non-null    int64
 16  VL status                             646 non-null    object
dtypes: int64(4), object(13)
memory usage: 85.9+ KB

None
```

Figure 18. Data structure result

The results from Figure 16 show that the data set, which was being analysed, had 646 records with 17 attributes. The 17 attributes consisted of 4 integer data types and 13 objects. This was an immediate indication that data wrangling would be conducted to have data that could be used for modelling.

Figure 18 outlines the descriptive statistics for the dataset. It summarizes essential measures across the variables examined, including the mean, median, standard deviation, and range. These figures provide a clear snapshot of the data's central tendencies and variability.

	Has beneficiary been linked to ART?	Age	ART duration as @ 2024	ZW- Viral load results
count	646.0	646.000000	646.000000	646.000000
mean	1.0	12.815789	8.578947	215.973684
std	0.0	3.434679	3.492955	449.032035
min	1.0	1.000000	0.000000	0.000000
25%	1.0	11.000000	6.000000	0.000000
50%	1.0	13.000000	9.000000	0.000000
75%	1.0	16.000000	11.000000	66.500000
max	1.0	17.000000	17.000000	6780.000000

	Has beneficiary been linked to ART?	Age	ART duration as @ 2024	ZW- Viral load results
count	646.0	646.000000	646.000000	646.000000
mean	1.0	12.815789	8.578947	215.973684
std	0.0	3.434679	3.492955	449.032035
min	1.0	1.000000	0.000000	0.000000
25%	1.0	11.000000	6.000000	0.000000
50%	1.0	13.000000	9.000000	0.000000
75%	1.0	16.000000	11.000000	66.500000
max	1.0	17.000000	17.000000	6780.000000

Figure 19. descriptive statistics result

Descriptive data analysis was conducted to initially understand the data. The results in Figure 17 showed that the minimum age of the children in the data set was 1 and the maximum age was 17. This was in line with the research, which focused on children living with HIV who are below 18 years of age. This analysis also showed a maximum viral load result of 6780 and a minimum of 0, which was a strong indication that the data had both classes represented in the target variable since a logistic regression model was to be built.

Figure 19 highlights the distribution of missing values, highlighting variables with incomplete data.

```
Missing Values:
 Organ                                0
Organisation unit name                0
OVC ID                                0
Has beneficiary been linked to ART?   0
ART Number                            0
Sex                                   0
Priority population 1                 0
HIV Status                            0
Date of Birth                         0
Age                                   0
Health Facility for ART               0
Date of ART initiation                0
ART duration as @ 2024                0
Art status                            0
Adherence status                      0
ZW- Viral load results                0
VL status                             0
```

Figure 20. Count of missing values result

Figure 19 shows that the data had no missing values; thus, there was no need to employ techniques that replace missing values. This also meant that the study's results were going to give a true reflection of what is in the Gutu district, with no mathematical bias that comes with dealing with missing values.

Figure 20 illustrates the distribution of the target variable, VL status, essential for assessing the prevalence of detectable versus non-detectable viral loads in the study population.

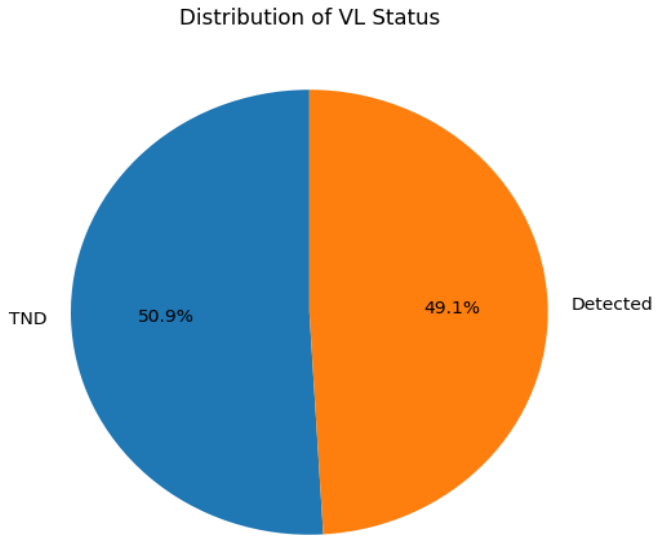


Figure 21. Distribution of VL status result

Figure 20 shows the distribution of the target variable, VL status. The visual showed a big number of children (49.1%) having a detectable viral load result, which is in line with the gap in the literature of most HIV positive people having a detectable viral load result. This result showed that the VL results were not biased more toward one class, which could affect model development, resulting in overfitting. This strongly indicated that a good model was to be developed from the dataset.

Figure 21 illustrates the distribution of viral load status across various organizational units, offering valuable insights into variations in viral load outcomes at different service delivery points, thus pinpointing areas that might need extra support or resources.

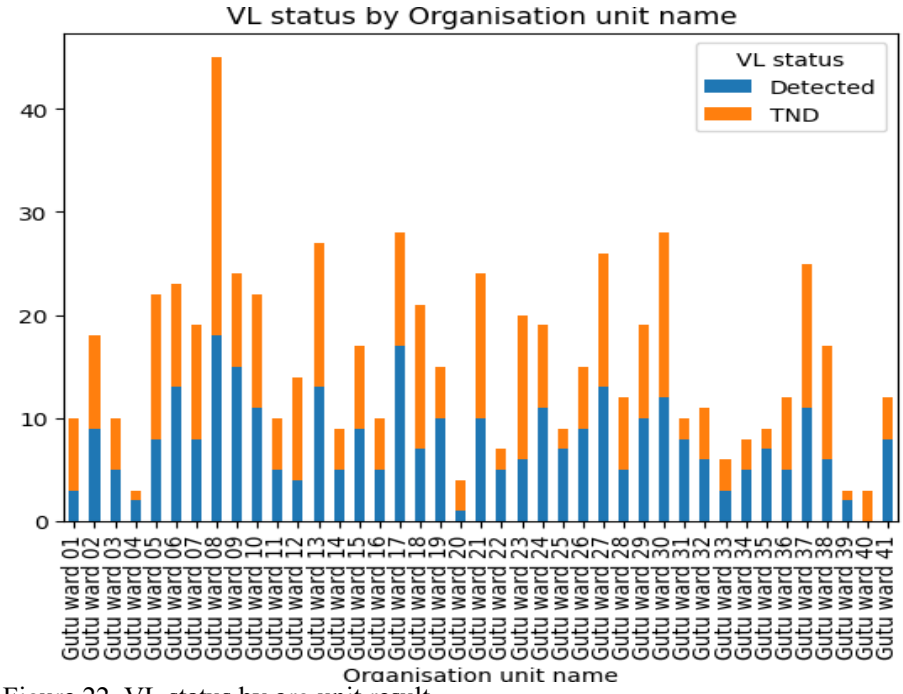


Figure 22. VL status by org unit result

Figure 20 shows that the largest number of children living with HIV lived in ward 8, and the least were from wards 4, 39, and 40. These results were a guide for targeted intervention strategies, as interventions may need to start with the high-volume areas within the district. Figure 22 illustrates the distribution of viral load (VL) status across various health facilities, offering valuable insights into variations in VL outcomes between facilities.

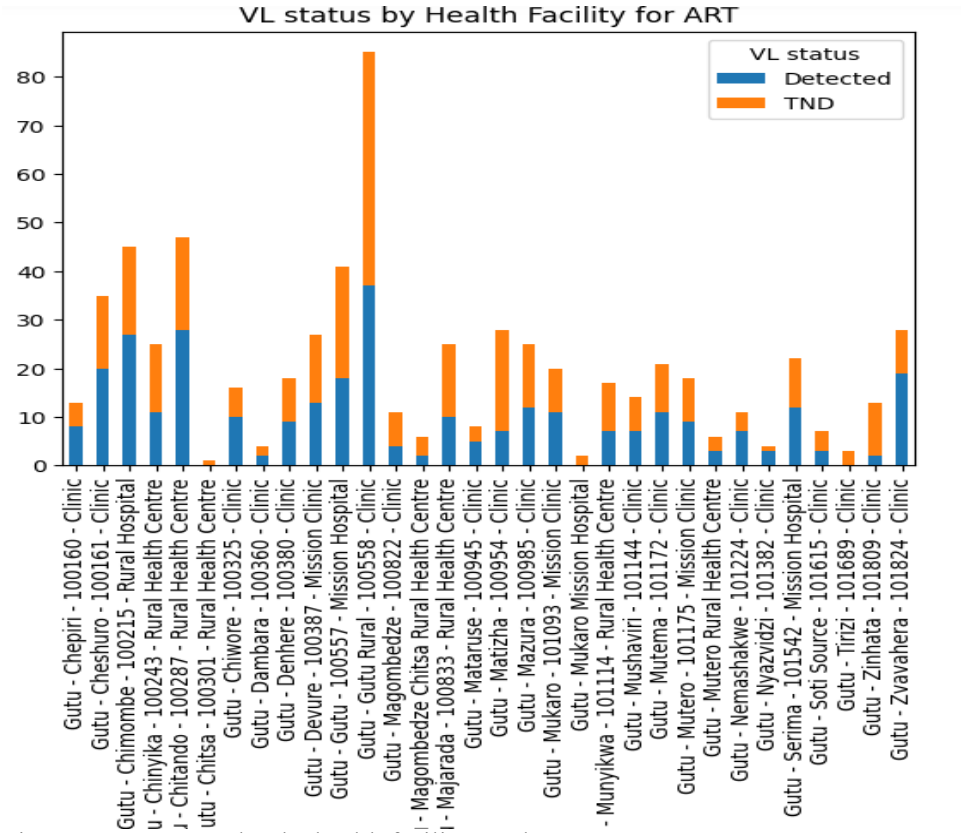


Figure 23. VL status by the health facility results

Results from Figure 19 also show that the largest number of children living with HIV is reported at Gutu Rural Hospital. P1 and P2 also had large numbers of children living with HIV. These high-volume sites also had higher numbers of children with a detected viral load. Intervention strategies were to prioritise these health facilities.

Of note were P3, P4, and P5, whose children all had a TND viral load result as of 2024. This is the ideal viral load that the world is looking forward to for the HIV community. Researchers may have to conduct further research and qualitative studies to understand the factors associated with such good viral load for these facilities for adoption by other facilities within the Gutu district.

Figure 23 shows how viral load status across sex is distributed, illustrating valuable insights into gender-based variations in viral load outcomes.

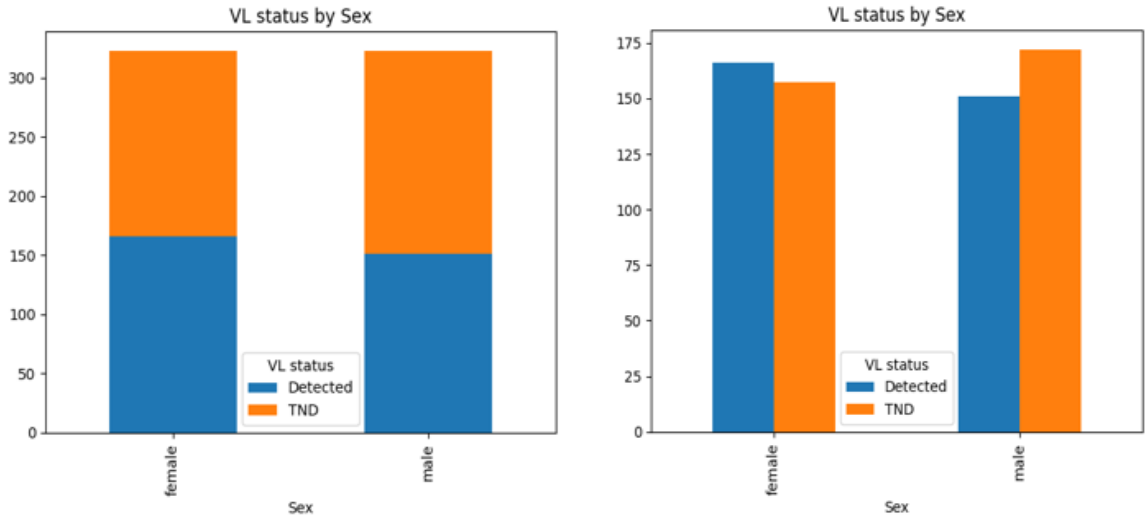
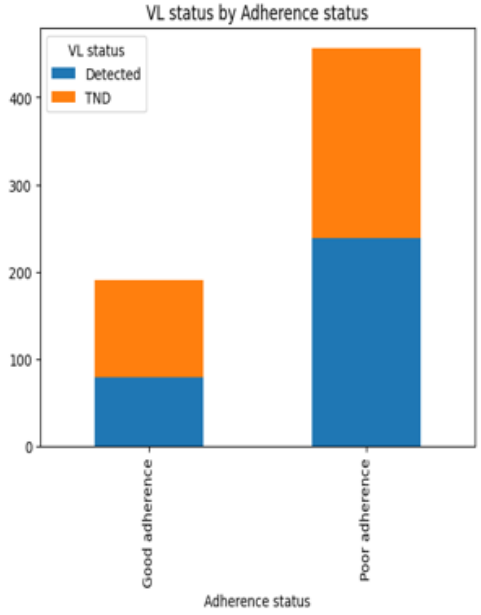


Figure 24. VL status by sex result

Figure 23 generally suggests that the viral load status of children in Gutu district may be slightly affected by the child's gender. The results show that the children with a detected viral load are more likely to be females than males, though the variance is small. Males generally seemed to be doing better in terms of viral load detection, where more males have a "target not detected" status, hence diminishing the chances of spreading HIV to 0%.

Figure 24 illustrates the distribution of viral load (VL) status according to adherence to antiretroviral therapy (ART). This visualization offers valuable insights into how adherence influences VL outcomes. It also highlights potential intervention points to enhance overall treatment success.



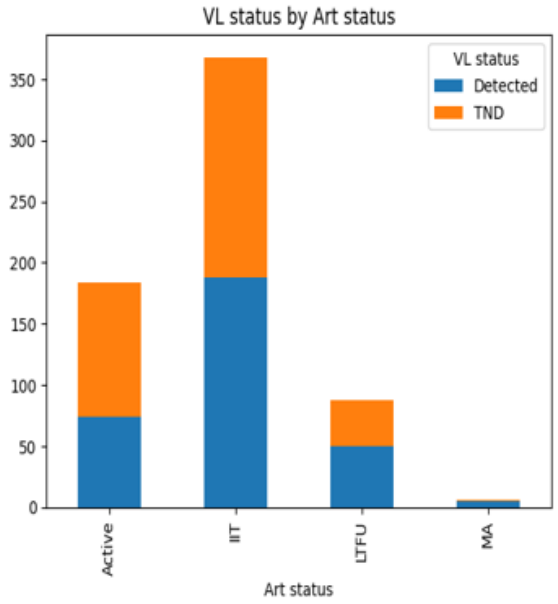


Figure 25. VL status by ART status result

The results suggested that adherence to ART was a strong factor that affected viral load. There were more children with a detected viral load in the class of those who had poor adherence to ART. Children who were active on ART had fewer detected viral load results than children who were interrupted in treatment, lost to follow-up up and those who had missed their appointments. This meant that viral load outcomes were to be improved by improving ART adherence.

Figure 25 shows the distribution of viral load status across different age groups in children. This visualization reveals variations in outcomes, which could guide the development of targeted interventions and clinical approaches tailored to specific ages.

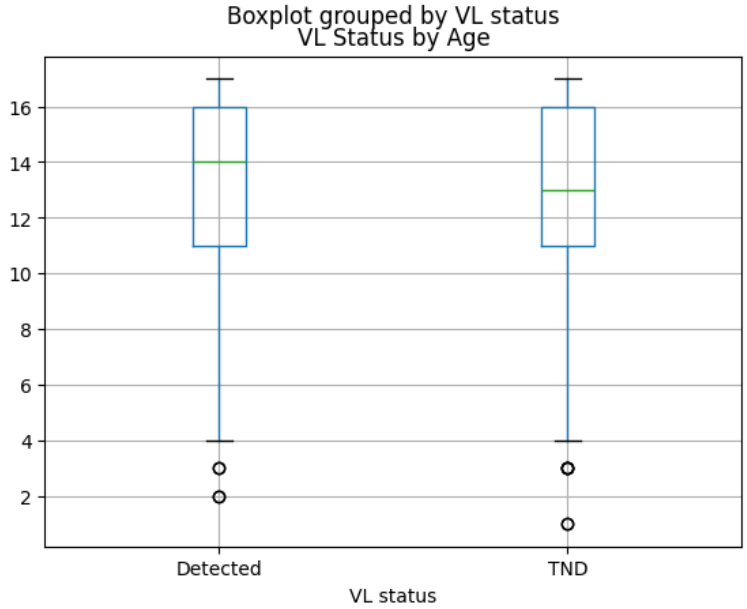


Figure 26. VL status by age result

Figure 25 shows the median age of individuals with "Detected" viral load to be around age 14, slightly higher than the median age of those with "Target Not Detected," around age 13. For "Detected," the box extends roughly from age 11 to age 16. This means the middle 50% of individuals with detectable viral load fall within this age range, while for "TND," the box extends roughly from age 10 to age 16. This means the middle 50% of individuals with target-not-detected viral load fall within this age range. The results suggest a trend towards slightly older individuals having a detectable viral load compared to younger individuals. The older children will be at a transition stage from being a child to becoming an adult. This is the point where they are highly affected by stigma and discrimination, hence affecting their adherence, resulting in detectable viral load. From the results, age was a factor that was slightly associated with children's viral load outcomes.

Figure 26 presents an interpretation of the log odds. This approach sheds light on how shifts in the explanatory variables affect the likelihood of the outcome under study. Studies show that such analyses help clarify these relationships in probabilistic models.

```
Interpretation of Coefficients:
- log_viral_load:
  An increase in 'log_viral_load' is associated with a decreased log-odds of 'VL status' being 'Detected'.
- Has beneficiary been linked to ART?:
  An increase in 'Has beneficiary been linked to ART?' is associated with an increased log-odds of 'VL status' being 'Detec
ted'.
- Adherence status_Poor adherence:
  An increase in 'Adherence status_Poor adherence' is associated with an increased log-odds of 'VL status' being 'Detecte
d'.
- Sex_male:
  An increase in 'Sex_male' is associated with an increased log-odds of 'VL status' being 'Detected'.
- Art status_IIT:
  An increase in 'Art status_IIT' is associated with an increased log-odds of 'VL status' being 'Detected'.
- Organisation unit name_Gutu ward 23:
  An increase in 'Organisation unit name_Gutu ward 23' is associated with an increased log-odds of 'VL status' being 'Detec
ted'.
- Art status_LTFU:
  An increase in 'Art status_LTFU' is associated with an increased log-odds of 'VL status' being 'Detected'.
- Organisation unit name_Gutu ward 30:
  An increase in 'Organisation unit name_Gutu ward 30' is associated with an increased log-odds of 'VL status' being 'Detec
ted'.
- Organisation unit name_Gutu ward 21:
  An increase in 'Organisation unit name_Gutu ward 21' is associated with an increased log-odds of 'VL status' being 'Detec
ted'.
- Organisation unit name_Gutu ward 18:
  An increase in 'Organisation unit name_Gutu ward 18' is associated with an increased log-odds of 'VL status' being 'Detec
ted'.
```

Figure 27. Log Odds interpretation result

Results in Figure 26 shows the log odds coefficient results, which test an event's likelihood. The results showed that Children with poor adherence to their medication have a higher chance of the virus being detected. They also revealed that being male is associated with a higher chance of the virus being detected. Biological, social, or behavioral factors related to being male might influence viral load outcomes in this population. The results also showed that children who have experienced an "Interruption in Treatment" and those who were "Lost to follow-up" (stopped attending clinics or receiving treatment) also have a higher chance of the virus being detected. Stopping and restarting ART can lead to viral rebound, and viral load is likely to increase and become detectable. The results also suggested that there might be differences in healthcare delivery, access to resources, patient populations, or other factors across different wards in Gutu that influence viral load outcomes. These findings could highlight areas needing targeted interventions.

Figure 27 shows results from the chi-squared test analysis that was conducted to interrogate further factors associated with a detectable viral load.

	Feature	Chi2 Statistic	P-value	Cramers V
0	Organ	0.000000	1.000000	NaN
1	Organisation unit name	40.350325	0.454768	0.249924
2	Sex	1.214041	0.270533	0.043351
3	Priority population 1	0.000000	1.000000	NaN
4	HIV Status	0.000000	1.000000	NaN
5	Art status	11.301411	0.010203	0.132267
6	Adherence status	5.628593	0.017670	0.093343

Figure 28. Chi-Squared test result

The results suggested that ART status and adherence status were statistically significant factors associated with viral load status since they had small p-values (0.05), which meant rejecting the null hypothesis and concluding that they affect viral load status. However, other statistical analyses conducted above revealed other factors associated with VL status.

Data Pre-Processing

The data had no missing values and showed consistency from the EDA; thus, no data cleaning was done. A Python code was run to handle missing data and inconsistencies, if any, to ensure the model is being developed using reliable data. Data was then pre-processed in preparation for modelling. Results from this stage are discussed below.

Figure 28 displays the pre-processed data. This cleaned version of the dataset offers a clear perspective, helping to ensure accuracy and reliability for subsequent statistical analysis and predictive modeling.

```
# Identifying categorical columns (excluding the target variable)
categorical_cols = ['Organ', 'Organisation unit name', 'Sex', 'Priority population 1', 'Art status', 'Adherence status']

# Creating dummy variables
df_encoded = pd.get_dummies(df, columns=categorical_cols, drop_first=True)

# Analyzing and transforming continuous features
# Log transformation for 'ZW- Viral load results' if it has a skewed distribution
# Checking for zero values first, add 1 before log if any zeros exist
if (df_encoded['ZW- Viral load results'] == 0).any():
    df_encoded['log_viral_load'] = np.log1p(df_encoded['ZW- Viral load results'])
else:
    df_encoded['log_viral_load'] = np.log(df_encoded['ZW- Viral load results'])
# 2. Standardization for 'Age' and 'ART duration as @ 2024'
scaler = StandardScaler()


# Fitting the scaler on the entire dataset
df_encoded[['Age', 'ART duration as @ 2024']] = scaler.fit_transform(df_encoded[['Age', 'ART duration as @ 2024']])


# Droping original columns after log transformation and standardization
df_encoded = df_encoded.drop(['ZW- Viral load results'], axis = 1)

display(df_encoded.head())
```

ART Number	HIV Status	Date of Birth	Age	Health Facility for ART	Date of ART initiation	ART duration as @ 2024	VL status	...	Organisation unit name_Gutu ward 38	Organisation unit name_Gutu ward 39	Organisation unit name_Gutu ward 40	Organisation unit name_Gutu ward 41	Sex_ma
8-04-27-2022-A-00012	Positive	1/11/2017	-1.694568	Gutu - Devure - 100387 - Mission Clinic	1/3/2022	-1.884950	TND	...	False	False	False	False	Tr
3-05-0A-2017-A-00381	Positive	16/5/2016	-1.403194	Gutu - Chimombe - 100215 - Rural Hospital	20/12/2017	-0.452388	TND	...	False	False	False	False	Tr

Figure 29. Data after pre-processing

As a mathematical model, logistic regression cannot process attributes with non-mathematical data. Some attributes had to be transformed to suit a mathematical state that the model can process. The first section of Figure 27 shows the section where the researcher performed one-hot encoding for categorical variables. The table below, in Figure 27, shows the results after one-hot encoding, where dummy variables were created. Continuous variables such as age and ART duration were also analysed, transformed, and standardized accordingly to achieve the best performing model.

As shown in Figure 29, additional data cleaning steps addressed lingering inconsistencies, outliers, and missing values. This preparation ensured the dataset was ready for reliable analysis.

```
# Drop 'Date of Birth', 'Date of ART initiation', and 'Health Facility for ART' columns
X_train = X_train.drop(['Date of Birth', 'Date of ART initiation', 'Health Facility for ART'], axis=1)
X_test = X_test.drop(['Date of Birth', 'Date of ART initiation', 'Health Facility for ART'], axis=1)
```

Figure 30. Further Data Cleaning

It shows the dropping of columns that contained string values, which could negatively affect the building of the logistic regression model.

Modelling

This is the stage where the model was developed and trained. The data was first split into the training and testing sets. Figure 30 shows the splitting of the data into the training and testing data sets as the first step in data modeling.

```
from sklearn.model_selection import train_test_split

# Separate features (X) and target variable (y)
X = df_selected_features.drop('VL status', axis=1)
y = df_selected_features['VL status']

# Split data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Display shapes of the resulting datasets (optional, for verification)
print("X_train shape:", X_train.shape)
print("X_test shape:", X_test.shape)
print("y_train shape:", y_train.shape)
print("y_test shape:", y_test.shape)
```

```
X_train shape: (516, 8)
X_test shape: (130, 8)
y_train shape: (516,)
y_test shape: (130,)
```

Figure 31. Data splitting result

The results show that the final data set used to develop the actual model contained eight columns. The training data set had 516 rows and eight attributes, while the testing set had 130 rows and eight attributes. Figure 31 shows the development and training of the logistic regression model using the training data set with 1000 training runs.

```
# Initialize and train the model
logreg = LogisticRegression(random_state=42, solver='liblinear', max_iter=1000)
logreg.fit(X_train, y_train)
```

```
LogisticRegression(max_iter=1000, random_state=42, solver='liblinear')
```

Figure 32. Model training result

At this stage, the model was developed and ready to be tested for performance.

Model Evaluation

This is the stage where selected evaluation matrices were used to evaluate the model. The results of the evaluation are shown in Figure 32.

```
Accuracy: 0.8923
Precision: 0.9112
Recall: 0.8923
F1-score: 0.8909
AUC-ROC: 0.8956

Classification Report:
              precision    recall  f1-score   support

    Detected       1.00      0.78      0.88        64
         TND       0.82      1.00      0.90        66

    accuracy                           0.89       130
   macro avg       0.91      0.89      0.89       130
weighted avg       0.91      0.89      0.89       130
```

Figure 33. Model evaluation result

Accuracy - The model's accuracy was 0.8923 (89.23%), which means that the model correctly predicted the viral load status (Detected or TND) for approximately 89.23% of the samples in the test data set.

Precision – The model's precision was 0.9112 (91.12%). This result suggests that when the model predicted a certain viral load status, it was correct about 91.12% of the time.

Recall – Recall was at 0.8923(89.23%), which indicated that the model could identify 89.23% of

all the actual instances of each viral load status in the test data set.

F1-score – The F1 score was 0.8909, which suggested a good balance between precision and recall.

The high accuracy, precision, recall, good F1 score, and AUC-ROC suggested that the model was performing well and is a strong predictor of viral load for children living with HIV in the Gutu district of Zimbabwe. The model could be adopted for further improvement and prediction of viral load outcomes for children living with HIV in Gutu. However, from the analysis and literature, the researcher felt that a better model would have been developed if the data had more attributes. Information such as WHO stage, Drug regimen, orphan status, caregiver type, distance to health facility, etc., may have helped build a robust model. The researcher recommends capturing this data in the system, especially if it has already been collected on hard copies.

AUC-ROC – It was 0.8956, wi

Figure 33 shows a visual of the model's performance in the classification of the children within the testing set of 130 records.

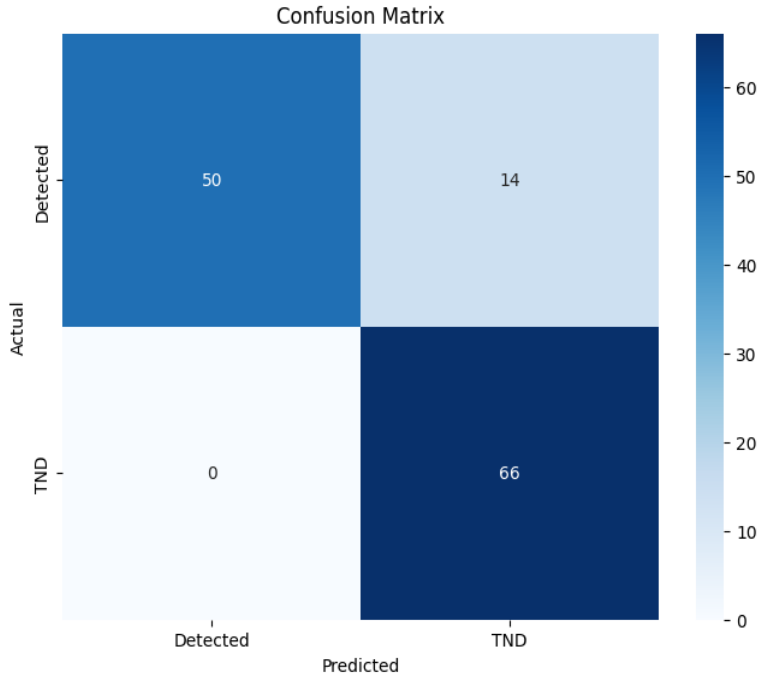


Figure 34. Confusion matrix

The model wrongly classified only 14 records out of the 130. The 50 children detected VL were all predicted as having a detected result. 66 children who actually had their target not detected were predicted to have a TND result.

Model Optimisation

The researcher conducted feature engineering in an attempt to improve the model's performance. Results of the model are shown in Figure 34:

```
Best Hyperparameters: {'C': 0.01, 'penalty': 'l2', 'solver': 'liblinear'}
Best F1-weighted Score: 0.9277254179244666

Optimized Model Evaluation:
Accuracy: 0.8923
Precision: 0.9112
Recall: 0.8923
F1-score: 0.8909
AUC-ROC: 0.9022

Classification Report (Optimized Model):
              precision    recall  f1-score   support

           0       1.00      0.78      0.88        64
           1       0.82      1.00      0.90        66

    accuracy                           0.89       130
   macro avg       0.91      0.89      0.89       130
weighted avg       0.91      0.89      0.89       130
```

Figure 35. Optimised model results

Results from Figure 33 indicated that the optimized model had an improved F1 score and AUC-ROC, which suggested that the optimized model had better performance, though the model's accuracy was the same.

Adherence to Treatment

Treatment adherence plays a vital role in HIV management, particularly for children, where consistent antiretroviral therapy (ART) can lower viral loads and bolster immune function (Bouchard et al, 2022). Nevertheless, adherence faces obstacles like intricate medication schedules, adverse effects, psychological barriers, and limited awareness of treatment's value among children and caregivers (Conan et al, 2020; Goga, 2021; Fuyana et al., 2025). Educational programs that clarify ART's advantages enhance comprehension and adherence, leading to improved viral suppression (Slogrove et al, 2021). Emotional factors, including stress, anxiety, and depression, often undermine caregivers' capacity to oversee treatment (Bhattacharya et al, 2023). Conversely, social support from family and peers bolsters adherence efforts (Gachago et al, 2022). Engaging family in educational initiatives helps children grasp the treatment's purpose, yielding better overall health results (Mavhu et al, 2020).

Stigma and Discrimination

HIV stigma stands as a significant obstacle to accessing medical care and maintaining treatment adherence (Mahamboro et al, 2022). Evidence indicates that it frequently leads to feelings of shame and social isolation among those affected and their families (Gachago et al, 2022). This stigma appears in various forms, including public stigma, self-stigma, and institutional stigma, all of which influence perceptions and hinder access to care (Brahmbhatt, 2020). For children living with HIV, the consequences can be particularly severe, often involving bullying, emotional distress, and difficulties in disclosing their status, which in turn compromise overall health outcomes (Mebrahtu et al., 2023). Awareness campaigns, especially those led by community leaders, have shown promise in dispelling myths, alleviating fears, and promoting more inclusive attitudes toward individuals with HIV (Mahamboro et al, 2021).

Socioeconomic Status

Socioeconomic status significantly influences health behaviours and outcomes, particularly for vulnerable populations like children living with HIV. Individuals from lower socioeconomic backgrounds are believed to frequently face healthcare challenges (Kojima et al, 2021). Among these difficulties are limited access to appropriate transportation, limited educational possibilities, and financial limitations. (Gachago et al, 2022). Children from low-income families may experience higher stress levels and instability affecting their health and well-being (Gachago et al., 2022). Financial limitations can lead to food insecurity, poor housing, and limited access to healthcare services, hindering effective treatment of chronic conditions like HIV (Goga et al, 2021).

Studies indicate that people from lower socioeconomic backgrounds often struggle to get medical care on time as they face issues like living far from clinics and not having reliable transportation, leading to delays in diagnosis and treatment (Slogrove, 2020). Efforts to fix this, such as setting up mobile clinics, help bridge the gaps, ensuring that vulnerable populations receive the care they need regardless of their socioeconomic status (Mavhu et al, 2020).

In some cases, it may be an issue of limited access to quality education where the education system does not have enough resources to support the education of children in that area, thus limiting individuals' understanding of health information and reducing their ability to navigate the healthcare system effectively (Okonji et al, 2021). Educational interventions that empower families to make informed health decisions can be instrumental in improving health behavior.

Cultural Beliefs

Cultural beliefs significantly shape health behaviors and perceptions, affecting how individuals understand and respond to conditions such as HIV (Bhattacharya et al, 2023). In many African contexts, traditional healing practices remain widespread. However, evidence indicates that these practices can sometimes interfere with adherence to prescribed medical therapies (Nichols, 2021). Interestingly, certain families favor these cultural beliefs over conventional medical interventions, resulting in inconsistent treatment patterns. Integrating traditional and modern approaches enhances overall acceptance and compliance with care (Mukumbang et al, 2021).

Cultural perspectives also contribute substantially to HIV-related stigma. Often, the disease is associated with moral shortcomings, which fosters discrimination, reluctance to disclose status, and even virological failure (Poku et al., 2020). Education and awareness initiatives prove essential in mitigating this stigma. In collectivist societies, family dynamics exert a strong influence on health choices. Studies show that engaging families in educational efforts and treatment processes bolsters adherence and fosters a more supportive environment (Cluver et al., 2020).

Peer Influence

Peer pressure is key in shaping health behaviors among children and adolescents in low-resource areas, such as Gutu. Evidence indicates it significantly influences these behaviors (Bhattacharya et al., 2023). Social ties often guide decisions, including adherence to HIV treatment, as young people tend to mirror their peers' actions (Mebrahtu et al., 2023).

Positive influences from peers can boost commitment to health routines. Studies show this might lead to better outcomes (Mavhu et al., 2020). Programs offering peer support and mentorship appear to strengthen adherence and support reaching undetectable viral loads (Mahamboro et al., 2022).

On the other hand, negative pressures like bullying or discrimination could hinder access to care. Research suggests these factors discourage engagement (Pantelic et al., 2022). Fostering safe environments for dialogue and encouraging inclusive attitudes among peers may help counter such effects. Understanding peer dynamics remains crucial. It allows for interventions that build supportive relationships, improve adherence, achieve undetectable viral loads, and enhance overall health for children with HIV.

Empirical Review

Children and adolescents living with HIV (CALHIV) consistently experience sub-optimal viral load outcomes compared to adults (Davies, 2020). They face challenges such as disclosure of HIV status, navigating adherence during major life changes (e.g., school transitions, adolescence), and experiencing stigma and discrimination, which negatively affect their retention in care and achieving and sustaining an undetectable viral load (VL) (Bouchard et al., 2022). Research has generally shown that managing HIV in children presents discrete challenges compared to adults (Cluver et al., 2020). Developmental stages, treatment adherence, and psychosocial environments often affect children more than adults (Bhattacharya et al., 2023). The World Health Organisation says viral load testing is essential for monitoring ART effectiveness and identifying viral suppression success or failure (World Health Organisation, 2021). Achieving the UNAIDS 95-95-95 targets requires 95% of individuals on ART to reach viral suppression, where currently, only about 59% of PLHIV worldwide have achieved viral suppression, with children and adolescents lagging the most (World Health Organisation,

2021). Research has highlighted that although ART availability has increased, viral load suppression rates remain low among children, especially in rural settings (Machekano et al., 2023).

ART adherence is crucial to viral load outcomes (Kim et al., 2022). Many children face difficulties such as complex regimens, medication side effects, and lack of caregiver support, all of which can affect adherence (Nichols, 2021). Stigma related to HIV can discourage families from seeking or continuing treatment (Mahamboro et al., 2022). A child's health can be significantly affected by the emotional effects of living with HIV. Children may experience mental health issues such as anxiety, depression, and loneliness, affecting their willingness to adhere to treatment (Brahmbhatt, 2020). Family dynamics, especially the presence of supportive caregivers, are crucial in influencing children's health behaviours and outcomes (Cluver et al., 2020). Taking note of these psychosocial dynamics is of great significance, i.e., it comes up with interventions that meet medical needs and support the emotional well-being of the children.

Current developments in machine learning have led to the management of chronic diseases such as HIV. Several research studies have shown that machine learning can handle larger volumes of data in such a way as to discover trends and predictions, which is very significant for effective public health intervention (Nash, 2022). Machine learning has emerged as a more effective approach that offers innovative methods to predict patient outcomes (Shamount et al., 2020). Different models, such as decision trees, random forests, and neural networks, have been implemented to evaluate healthcare data, leading to actionable insights (Mukura et al., 2023; Ndlovu et al., 2025). They have proven to help healthcare practitioners in decision-making by discovering hidden patterns and correlations often unseen through traditional statistical techniques (Shamount et al., 2020). Machine learning has been used in HIV care to predict outcomes such as medication adherence, viral load suppression, and disease progression (Almuhaideb et al., 2021; Chiramba et al., 2024). Evidence has shown that ML models can also analyse patient data to predict which individuals may need medical monitoring or are at risk of treatment failure (Almuhaideb et al., 2021). This predictive capability is particularly valuable, especially in resource-limited settings.

Machine learning applications in healthcare offer substantial potential to reduce errors and enhance operational efficiency, as evidenced by recent analyses (Almuhaideb et al., 2021). For instance, predictive models incorporating socio-demographic and behavioral data have shown promise in managing HIV (Makota et al., 2023). Yet, research on treatment outcomes remains limited in Zimbabwe, especially regarding viral suppression in children. Most existing studies prioritize adults (Machekano et al., 2023; Slogrove, 2020). This oversight is particularly troubling, considering the distinct barriers children face, such as adherence to antiretroviral therapy and associated stigma. Consequently, there is an urgent need for localized machine learning models that account for context-specific factors influencing viral load in pediatric populations. Tailored algorithms prove more effective when aligned with regional healthcare dynamics (Almuhaideb et al., 2021). Although the literature underscores machine learning's role in HIV management, significant gaps persist in its practical deployment for forecasting treatment results. Drawing on the example of HIV-infected children in Gutu District, this study seeks to deliver actionable insights for refining predictive models in healthcare.

Gaps in the Literature

Although machine learning has advanced HIV research, significant gaps remain, particularly in localized studies focused on vulnerable populations (Slogrove, 2020). Evidence indicates that most models draw from adult datasets and generalized populations, which diminishes their relevance for children in rural settings like Gutu District (Slogrove, 2020). Zimbabwe's socioeconomic and cultural context plays a major role in shaping health outcomes (Machekano et al., 2023). However, existing studies frequently overlook these elements, thereby restricting the practical applicability of their models. Current research also analyzes clinical, demographic, and psychosocial variables in isolation rather than integrating them (Kakkar et al., 2020). Developing contextualised machine learning models in low-resource environments such as Gutu is crucial, as these can capture the distinct challenges faced by children living with HIV. Such models appear well-suited to detect viral load

failure early, facilitate targeted interventions, and support proactive, real-time monitoring. Ultimately, this early detection can improve disease progression, underscoring the importance of contextualized approaches in enhancing HIV management and health informatics.

CONCLUSION

In conclusion, age, sex, ART status, geographic location, and adherence are features that influence viral load outcomes for children between the ages of 0 – 17 who are living with HIV in the Gutu district of Zimbabwe. Machine learning can be used to predict child viral load outcomes. Logistic regression, a supervised machine learning, can accurately predict these outcomes with high accuracy rates as high as 89%. The logistic regression model developed in this study has a high performance and can be adopted, improved, and integrated into the Zimbabwe health care system if there are enough resources. This will imply earlier identification of risk of a detectable viral load for children, hence prior migratory measures, which will result in an overall high VL suppression rate with VL targets not being detected upon blood testing, thus HIV epidemic control.

CONFLICTS OF INTEREST

The authors declare that there are no conflicts of interest regarding the publication of this paper.

REFERENCES

1. Almuhaideb, A., Dou, D., & Alhajj, R. (2021). *Applications of machine learning in healthcare*.
2. Alyafei, A., & Carr, R. E. (2024). *The Health Belief Model of Behavior Change*.
3. Anuar, H., Shar, S. A., Gafor, H., Mahmood, M. I., & Ghazi, H. F. (2020). Usage of Health Belief Model (HBM) in Health Behavior. *Malaysian Journal of Medicine and Health Sciences*.
4. Bhattacharya, A., Mehrotra, P., & Satyanaravana, V. A. (2023). Barriers to antiretroviral therapy adherence among children and adolescents living with HIV in low-resource settings. *Journal of Pediatric Infectious Diseases Society*, 12(1), 48–55.
5. Bouchard, A., Bourdear, F., Roger, J., Thaillefer, V. T., Sheehan, N. L., Schnitzer, M., & Wang, G. (2022). *Predictive Factors of Detectable Viral Load in HIV-Infected Patients. Aids Research and Human Retrovirus*.
6. Brahmbhatt, H. (2020). Psychological outcomes in HIV-infected youth: The impact of stigma and adherence. *Journal of the International AIDS Society*.
7. Chiramba, N. W., Ndlovu, B., Dube, S., Jacqueline, F., & Muduva, M. (2024). Optimizing Antiretroviral Therapy (ART) Adherence Through Predictive Analytics Using Machine Learning Techniques. *IEOM*, 2021 (July).
8. Fuyana, C., Ndlovu, B., Dube, S., Maguraushe, K., & Malungana, L. (2025). Optimizing HIV Care Through Machine Learning-Assisted Prediction and Personalized Treatment. In V. Bhateja, P. Patel, & J. Tang (Eds.), *Smart Innovation, Systems and Technologies* (Vol. 436, pp. 149–161). Springer Nature
9. Gachago, L., Phiri, S., & Govender, K. (2022). *Socioeconomic factors and HIV treatment outcomes in rural children*. AIDS Research and Human Retroviruses.
10. Gafor, H., Carr, R. E., Shar, S. A., & Anuar, H. (2024). The Health Belief Model of Behavior Change.
11. Goga, A. E., Jackson, D. J., & Lombard, C. (2021). HIV treatment outcomes among children living in resource-limited settings: Challenges and solutions. *Pediatrics and Neonatology*, 1-10.
12. Green, E. C., Murphy, E. M., & Gryboski, K. (2020). The Health Belief Model.
13. Indicator Registry. (n.d.). Retrieved from https://indicatorregistry.unaids.org/: https://indicatorregistry.unaids.org/indicator/people-living-hiv-who-have-suppressed-viral-loads
14. Kakkar, F., Lee, T., Hawkes, M. T., & Brophy, J. (2020). *Challenges to achieving and maintaining viral suppression among children living with HIV*.
15. Kilroy, D. L. (2020). Study of a User-driven Location-based Point-of-Interest Recommendation System.
16. Kim, S. H., Gerver, S. M., & Fidler, S. (2022). The impact of ART adherence on long-term viral suppression. *Current Opinion in HIV and AIDS*, 45-52.

17. Kojima, N., Shrestha, R. K., & Klausner, J. D. (2021). Addressing social determinants of health in HIV prevention and care.
18. Lao, X., Zhang, H., Yan, L., Zhao, H., Zhao, Q., Lu, H., . . . Liang, X. (2023). Thirteen-year viral suppression and immunologic recovery of LPV/r-based regimens in pediatric HIV treatment: a multicenter cohort study in resource-constrained settings of China. *Frontiers in Medicine*, 10.
19. Lopez, M., Antonio, O., & Crossa, A. (2022). Multivariate Statistical Machine Learning Methods for Genomic Prediction. *Multivariate Statistical Machine Learning Methods for Genomic Prediction* (pp. 1–691). Springer International Publishing.
20. Machekano, R., Mapangisana, T., Maposhere, C., Mutetwa, R., Manasa, J., Shamhu, T., Munyati, S. (2023). Clinic-based SAMBA-II vs centralized laboratory viral load assays among HIV-1 infected children, adolescents and young adults in rural Zimbabwe: A randomized controlled trial.
21. Mahamboro, D. B., Fauk, N. K., Ward, P. R., & Merry, M. S. (2022). Stigma and discrimination towards people living with HIV/AIDS: A qualitative study. *International Journal of Environmental Research and Public Health*, 19(6), 3483.
22. Makota, R. B., & Musenge, E. (2023). Predicting HIV infection in the decade (2005–2015) pre-COVID-19 in Zimbabwe: A supervised classification-based machine learning approach. *PLOS Digital Health*, 2(6).
23. Mauthner, N. S. (2020). Research philosophies and why they matter. In *How to Keep Your Doctorate on Track*. Edward Elgar Publishing.
24. Mavhu, W., Willis, N., & Bernays, S. (2020). Enhancing psychosocial support for children living with HIV. *Journal of the International AIDS Society*.
25. Mebrahtu, T., Mohammed, F., & Mensa, G. A. (2023). Psychosocial correlates of ART adherence among adolescents living with HIV. 561-571.
26. Mukumbang, F. C., Ambe, A. N., & Adebiyi, B. O. (2021). Family-centered interventions for managing pediatric HIV.
27. Nagler, T. (2024). Statistical Learning Theory.
28. Nash, D. (2022). Machine learning for HIV care and research: Promises and pitfalls.
29. Ndlovu, B., Maguraushe, K., & Mabikwa, O. (2025). Machine Learning and Explainable AI for Parkinson's Disease Prediction: A Systematic Review. *The Indonesian Journal of Computer Science*, 14(2).
30. Ndlovu, B., Maguraushe, K., & Mabikwa, O. (2025). Machine Learning and Explainable AI for Parkinson's Disease Prediction: A Systematic Review. *The Indonesian Journal of Computer Science*, 14(2).
31. Nichols, J. S. (2021). Adherence challenges among HIV-infected children: Barriers and strategies. *Pediatric Infectious Disease Journal*.
32. Nusinovichi, S., Tham, Y. C., Yan, M. Y., Ting, T. S., Li, J., Sabanayagam, C., & Wong, T. Y. (2020). Logistic regression was as good as machine learning for predicting major chronic diseases. *Journal of Clinical Epidemiology*.
33. Okonji, E. F., Mukumbang, F. C., & Orth, Z. (2021). Socioeconomic disparities in HIV outcomes among African children.
34. Pantelic, M., Sprague, L., & Stangl, A. (2022). The science of stigma reduction.
35. Poku, N. K., Whiteside, A., & Sandkjaer, B. (2020). HIV/AIDS, stigma and globalization.
36. Shamount, F., Zhu, T., & Clifton, D. A. (2020). Machine Learning for Clinical Outcome Prediction.
37. Slogrove, A. L. (2020). Global challenges in pediatric HIV: Unmet needs and opportunities.
38. Slogrove, A. L., Mahy, M., & Abrams, E. J. (2021). Children and adolescents living with HIV: The need for more healthcare innovation. The Lancet Child & Adolescent Health, 612-624.
39. Solutions, S. (2020). What is logistic regression?
40. Topol, E. J. (2020). High-performance medicine: The convergence of human and artificial intelligence.
41. W.C Mukura, N., & Ndlovu, B. (2023). Performance Evaluation of Artificial Intelligence in Decision Support System for Heart Disease Risk Prediction. *Who, 2018*, 83–93.
42. W.C Mukura, N., & Ndlovu, B. (2023). Performance Evaluation of Artificial

Intelligence in Decision Support System for Heart Disease Risk Prediction. *Who, 2018*, 83–93.

43. Wang, R. Y., & String, D. M. (2020). Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems*.
44. Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining. *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, (24959), 29–39.
45. World Health Organisation. (2021). Consolidated guidelines on HIV prevention, testing, treatment, service delivery and monitoring. *Optics InfoBase Conference Papers* (p. 249).