KLE Technological University
Huballi

A Course Project Report on

"HIV Diagnosis"

A Course Project Report Submitted in Partial Fulfillment of the Requirement for the Course of

Exploratory Data Analysis

in

4th Semester of Computer Science and Engineering

by

Abhinandan Onajol -	02FE22BCI003
Varun Gani -	02FE22BCI056
Jaganath Magadum -	02FE22BCI017
Jaganath Malode -	02FE22BCI018

Under the guidance of

Dr. Prema Akkasaligar
Professor,
Department of Computer Science and Engineering,
KLE Technological University's Dr. MSSCET, Belagavi.

KLE Technological University's
Dr. M. S. Sheshgiri College of Engineering and Technology,
Belagavi – 590 008.

June 2024

DECLARATION

We hereby declare that the matter embodied in this report entitled "**HIV Diagnosis**" submitted to KLE Technological University for the course completion of Exploratory Data Analysis (22ECAC210) in the 4th Semester of Computer Science and Engineering is the result of the work done by us in the Department of Computer Science and Engineering(Artificial Intelligence), KLE Dr. M. S. Sheshgiri College of Engineering, Belagavi under the guidance of Dr. Prema Akkasaligar, Professor, Department of Computer Science and Engineering(Artificial Intelligence). We further declare that to the best of our knowledge and belief, the work reported here in doesn't form part of any other project on the basis of which a course or award was conferred on an earlier occasion on this by any other student, also the results of the work are not submitted for the award of any course, degree or diploma within this or in any other University or Institute. We hereby also confirm that all of the experimental work in this report has been done by us.

Belagavi – 590 008
Date :

Abhinandan Onajol (02FE22BCI003)	Varun Gani (02FE22BCI056)
Jaganath Malode (02FE22BCI018)	Jaganath Magadum (02FE22BCI019)

CERTIFICATE

This is to certify that the project entitled "HIV Diagnosis" submitted to KLE Technological University's Dr. MSSCET, Belagavi for the partial fulfillment of the requirement for the course - Exploratory Data Analysis (22ECAC210) by Abhinandan Onajol (02FE22BCI003), Varun Gani (02FE22BCI056), Jaganath Magadum (02FE22BCI019), Jaganath Malode (02FE22BCI018)., students in the Department of Computer Science and Engineering(Artificial Intelligence), KLE Technological University's Dr. MSSCET, Belagavi, is a bonafide record of the work carried out by them under my supervision. The contents of this report, in full or in parts, have not been submitted to any other Institute or University for the award of any other course completion.

Belagavi – 590 008
Date :

Dr. Prema Akkasaligar
(Course Teacher)

Prof. Priyanka Gavade
(Course Coordinator)

Dr. Rajashri Khanai
(Head of the Department)

Abstract

HIV dataset is set of exclusive information is collected to make finding the patterns and analysis of HIV mortality rate. Analyzing this dataset characteristicis crucial for several reasons, such as understandin the trends and patterns of HIV diagnoses. Analyzing linkage to care and viral suppression rates can provide insights into the effecitveness of corrent protocols and highlighting the area of improvement.

HIV(Human Immunodeficiency Virus) it is a critical public health issue globally, to understand the factors influencing HIV diagnoses and outcomes through datadriven insights. in this course project we perform exploratory data analysis to determine what these factors are and how they are correlated with each other.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

- HIV is the one of the most Dangerous disease it cannot be fully curable.

- HIV is primarily transmitted through: Sharing needles or syringes contaminated with HIV-infected blood, Transmission from mother to child during childbirth or breastfeeding (but this risk can be significantly reduced with proper medical interventions)

- HIV disease is not transmitted thorugh water, wind.

1.2 Problem Statement

The main aim is to identify the individuals who are affected by HIV and enabling the HIV prevention and improved health outcomes, and predicting risk of the HIV disease using demographic, behavioral and healthcare data.

1.2.1 Objectives

- Developing Predictive Model.

- Exploring the Geographical distributions of HIV.

- How Demographic Factors such as age, gender, race are associated with HIV diagnoses.

- Analyzing the Trends how HIV diagnoses, AIDS diagnoses and related health outcomes have evolved overtime.

Chapter 2

Knowing the Dataset

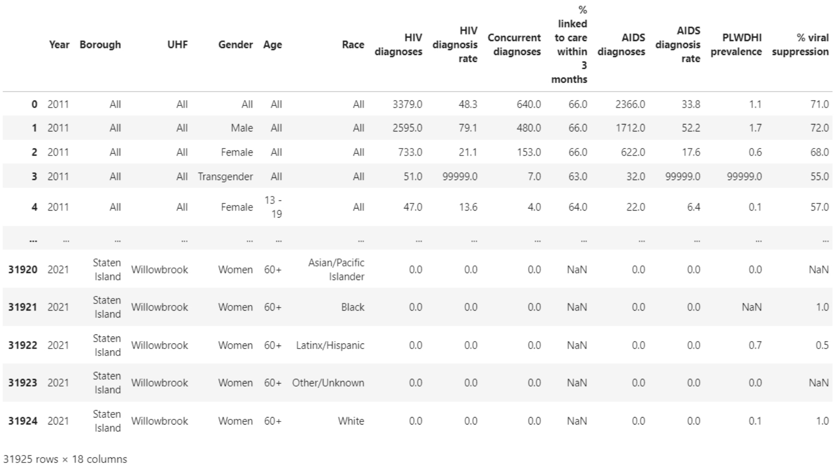
2.1 Dataset

Introduction to the dataset:

- **Source**: https://catalog.data.gov/dataset/dohmh-hiv-aids-annual-report
- **Memory size**:2.6MB
- **Number of instance**s:31926
- **Number of features**:18
- **Snapshot of dataset**

2.2 Features of the Dataset

- **Year**: The year of the data record.
- **Borough**: The borough within New York City where data was collected.
- **UHF**: The United Hospital Fund (UHF) neighborhood within the borough.



	Year	Borough	UHF	Gender	Age	Race	HIV diagnoses	HIV diagnosis rate	Concurrent diagnoses	% linked to care within 3 months	AIDS diagnoses	AIDS diagnosis rate	PLWDHI prevalence	% viral suppression
0	2011	All	All	All	All	All	3379.0	48.3	640.0	66.0	2366.0	33.8	1.1	71.0
1	2011	All	All	Male	All	All	2595.0	79.1	480.0	66.0	1712.0	52.2	1.7	72.0
2	2011	All	All	Female	All	All	733.0	21.1	153.0	66.0	622.0	17.6	0.6	68.0
3	2011	All	All	Transgender	All	All	51.0	99999.0	7.0	63.0	32.0	99999.0	99999.0	55.0
4	2011	All	All	Female	13 - 19	All	47.0	13.6	4.0	64.0	22.0	6.4	0.1	57.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
31920	2021	Staten Island	Willowbrook	Women	60+	Asian/Pacific Islander	0.0	0.0	0.0	NaN	0.0	0.0	0.0	NaN
31921	2021	Staten Island	Willowbrook	Women	60+	Black	0.0	0.0	0.0	NaN	0.0	0.0	NaN	1.0
31922	2021	Staten Island	Willowbrook	Women	60+	Latinx/Hispanic	0.0	0.0	0.0	NaN	0.0	0.0	0.7	0.5
31923	2021	Staten Island	Willowbrook	Women	60+	Other/Unknown	0.0	0.0	0.0	NaN	0.0	0.0	0.0	NaN
31924	2021	Staten Island	Willowbrook	Women	60+	White	0.0	0.0	0.0	NaN	0.0	0.0	0.1	1.0

31925 rows × 18 columns

FIGURE 2.1: HIV Dataset

- **Gender**: Gender of the individuals in the dataset (e.g., Male, Female , Transgender).

- **Age**: Age group of the individuals (e.g., 13 - 19, 20 - 29, etc.).

- **Race**: Race or ethnic group of the individuals.

- **HIV diagnoses**: Number of HIV diagnoses.

- **HIV diagnosis rate**: Rate of HIV diagnoses per 100,000 people.

- **Concurrent diagnoses**: Concurrent diagnoses of HIV and another condition.

- **linked to care within 3 months**: Percentage of HIV-diagnosed individuals linked to care within 3 months.

- **AIDS diagnoses**: Number of AIDS diagnoses.

- **AIDS diagnosis rate**: Rate of AIDS diagnoses per 100,000 people.

- **PLWDHI prevalence**: Prevalence of people living with diagnosed HIV infection.

- **viral suppression**: Percentage of people with viral suppression among those diagnosed with HIV.

- **Deaths**: Number of deaths.

- **Death rate**: Death rate per 100,000 people.

- **HIV-related death rate**: Death rate due to HIV-related causes per 100,000 people.

- **Non-HIV-related death rate**: Death rate due to nonHIV-related causes per100,000people.

Table 2.1: Details of the Features in the HIV Dataset.

Feature Name	**Data Type**	**Distinct Values**	**Missing Values**
Year	Numeric	10	0
Borough	String	5	0
UHF	String	42	0
Gender	String	3	0
HIV diagnoses	String	301	416
HIV diagnosis rate	Numeric	1944	416
Concurrent diagnoses	Numeric	99	116
linked to care within 3 months	Numeric	123	13274
AIDS diagnoses	Numeric	212	337
AIDS diagnosis rate	Numeric	1421	337
PLWDHI prevalence	Numeric	149	3553
viral suppression	Numeric	149	3553
Deaths	Numeric	264	0
Death rate	Numeric	606	1913
HIV-related death rate	Numeric	345	1913
Non-HIV-related death rate	Numeric	446	1913

2.3 Observations

List your observations from the dataset here.

- How are the features? All categorical? Mix?

- It's mix of attributes.
- Are there any missing values? If yes, are they large or small?
- What is the range of data items? How are they distributed?
- Are there any outliers?
- Are any of the features skewed?
- Does any of the features require normalization, scaling?
- Overall what are the characteristics of your dataset?

2.4 Statistical Data Analysis

2.4.1 Minimum and Maximum Values

- **Year**: 2017 to 2021.
- **HIV diagnoses**: 0 to 10.
- **HIV diagnosis rate**: 0 to 73.
- **Concurrent diagnoses**:0 to 2.5 .
- **linked to care within 3 months**: 0.5 to 1.0.
- **AIDS diagnoses**: 0 to 7.5.
- **AIDS diagnosis rate**: 0 to 45.25 .
- **PLWDHI prevalence**: 0 to 3.85.
- **viral suppression**: 0.5 to 1.0.

- **Deaths**: 0 to 7.5.
- **Death rate**:0 to 18.25.
- **HIV-related death rate** 0 to 100.
- **Non-HIV-related death rate** 0 to 6.9999:

2.4.2 Central Tendencies

- **Year**: Mean = 2017.17, Median = 2019.
- **HIV diagnoses**: Mean = 10.93, Median = 0.
- **HIV diagnosis rate**: Mean = 39.137, Median = 0.
- **Concurrent diagnoses**: Mean = 2.066, Median = 0.
- **linked to care within 3 months**: Mean = 8178.288, Median = 1.0.
- **AIDS diagnoses**: Mean = 9.989, Median = 0.0 .
- **AIDS diagnosis rate**: Mean = 33.9494, Median = 0.
- **PLWDHI prevalence**: Mean = 68.204, Median = 0.5.
- **viral suppression**: Mean = 538.04, Median = 0.87.
- **Deaths**: Mean = 14.97, Median = 0.
- **Death rate**: Mean = 7.38, Median = 0.

- **HIV-related death rate**: Mean = 4003.3838, Median = 0.

- **Non-HIV-related death rate**: Mean = 4005.76, Median = 0.

2.4.3 Quartile Scores

- **Year**: Q1 = 2017, Q2 = 2018, Q3 = 2020.

- **HIV diagnoses**: Q1 = 0, Q2 = 1, Q3 = 4.

- **HIV diagnosis rate**: Q1 = 0, Q2 = 4, Q3 = 29.39.

- **Concurrent diagnoses**: Q1 = 0, Q2 = 0, Q3 = 1.

- **linked to care within 3 months**: Q1 = 0, Q2 = 1, Q3 = 67.

- **AIDS diagnoses**: Q1 = 0, Q2 = 0, Q3 = 3.

- **AIDS diagnosis rate**: Q1 = 0, Q2 = 0, Q3 = 18.200.

- **PLWDHI prevalence**: Q1 = 0.20, Q2 = 0.60, Q3 = 1.60.

- **viral suppression**: Q1 = 0.80, Q2 = 0.90, Q3 = 1.

- **Deaths**: Q1 = 0, Q2 = 0, Q3 = 3.

- **Death rate**: Q1 = 0, Q2 = 0, Q3 = 7.90.

- **HIV-related death rate**: Q1 = 0, Q2 = 0, Q3 = 0.

- **Non-HIV-related death rate**: Q1 = 0, Q2 = 0, Q3 =4.20.

Chapter 3

Implement Framework

To perform exploratory data analysis on the HIV dataset, we have followed the following framework. The overall implementation flow is presented.

- **Data loading**: Loading the dataset into our data environment, such as Python, Pandas, R make sure to read the data correctly.

- **Data cleaning**: Handling missing value appropriately, Depending on the attributes present in the dataset. using techniques like impute missing values, remove rows and columns with missing values.

- **Data exploration**: Perform summary statics to gain initial insights into the dataset. This includes the mean, median, mode, standard deviation, maximum and minimum values for numerical features.

- **Data visualization**: Creating visualization to better understand the distribution and relationship between different features.

- **Outlier detection**: Identifying and handling outliers in the dataset makes significant changes in the output.

- **Feature Engineering**: Depending on the analysis goals we have to create the new features or derive new features.

- **Correlation Analysis**: Finding the correlation between different numerical features using correlation matrices or scatter plots. this will helps to understand the relationship between the variables.

- **Skewness and Normalization**: Check the skewness of numerical features. and making data more normally distributed. Normalization or scaling might also be necessary for certain analysis or modeling techniques.

- **Grouping and Aggregation**: Depending on the research questions, might want to group the data by certain attributes.

Chapter 4

Data Pre-processing

- **Handling missing values**:

  Step: Identify and handle misssing values in the dataset. You can either remove rows with missing values or use imputation techniques(e.g., mean, median, mode) to fill in the missing values.

  Result: A dataset with no missing values or with imputed values, making it suitable for further analysis.

- **Dealing with outliers**:

  Step: Detect and handle outliers in numerical features. You can choose to remove outliers or apply transformations(e.g.,log transformations) to mitigate their impact.

  Result: A dataset with outlier-free or transformed numerical features, ensuring more robust analysis.

- **Encoding Categorical Features**:

  Step:Convert categorical features(e.g., level and subject) into numerical format using techniques like one-hot encoding or label algorithms.

  Result: A dataset with numerical representations of categorical features, making it suitable for machine learning algorithms.

- **Scaling Numerical Features**:

Step: Scale numerical features to bring them to a similar maganitude. Common scaling techniques include Max-Min scaling or Standardization(Z-score scaling).

Result: Numerical features with consistent scales, avoiding dominance by feature with longer ranges.

- **Feature Engineering**:

  Step: Create new feature from existing ones or extract meaningful information from features to enhance the dataset.

  Result: A dataset with additional features that might improve the performance machine learning models.

- **Data Transformations**:

  Step: Apply transformations(e.g., logarithmic, square root) to normalise skew numerical features.

  Result: Numerical features with reduced skewness, leading to improved model performance.

- **Handling Date/Time Data**:

  Step: Extract relevant information from date/time columns(e.g., year, month, day) or convert them to a format suitable for analysis.

Chapter 5

Exploratory Data Analysis

5.1 Hypothesis on the Problem Statement

List the hypothesis that you have framed for your problem statements here.

5.2 Analysis

In the each sub-section perform uni-variate, bi-variate or multi-variate analysis of required for each of the hypothesis here.

They should include graphs and interpretation of the results.

Chapter 6

Results and Outcomes

List the outcomes of your course project here.

Conclusions

Paragraph-1: What was your problem statement and why it was important to solve it.

Paragraph-2: How you were able to solve the problem statement.

Paragraph-3: What results you got and their outcomes.

Bibliography