

*AIDS Virus Infection Prediction Model*

In today's connected world, the need for robust measures to compare countries, people, or cities across various dimensions is paramount. Composite indicators serve as indispensable tools in this regard, offering a comprehensive view of complex phenomena that cannot be adequately captured by single indicators alone. This documentation presents a detailed account of my project, aimed at developing a composite indicator using healthcare statistics from the AIDS dataset, to facilitate meaningful comparisons and insights.

The primary source of data for my project is the AIDS Virus Infection Prediction dataset, originally published in 1996. This dataset comprises healthcare statistics and categorical information about patients diagnosed with AIDS, encompassing a diverse array of variables ranging from demographic characteristics to treatment indicators and clinical measures. Prior to analysis, the dataset underwent meticulous preprocessing, including data cleaning, filtering, and handling missing values, to ensure its suitability for composite indicator development.

**Reason for choosing this dataset:**

The choice of the AIDS dataset for our project was driven by its rich and comprehensive healthcare statistics, covering various aspects of AIDS diagnosis, treatment, and outcomes. This dataset offers a nuanced view of HIV/AIDS-related healthcare, including demographic information, clinical measures, and treatment regimens. Its longitudinal nature allows for the exploration of temporal trends and patterns in AIDS-related healthcare over time. Given the global significance of HIV/AIDS, leveraging this dataset enables us to develop a robust composite indicator for comparative analyses across countries, people, or cities, contributing to our understanding of the disease and informing healthcare practices.

Link to the dataset: <https://www.kaggle.com/datasets/aadarshvelu/aids-virus-infection-prediction>

**This dataset contains information about people diagnosed with AIDS and their healthcare history:**

**time:** This shows how long it took for something to happen, like the time until treatment or the time until the study ended.

**trt:** This tells us which treatment the patient received. It's coded as numbers: 0 means they received one type of treatment, 1 means another treatment, and so on.

**age:** Age of the patient when they were first diagnosed with AIDS.

**wtkg:** Weight of the patient when they were diagnosed.

**hemo:** Indicates if the patient has hemophilia or not.

**homo:** Tells us if the patient had homosexual activity.

**drugs:** Shows if the patient has a history of using drugs intravenously.

**karnof:** This is a score that measures how well the patient can do daily activities. Higher scores mean better ability.

**oprior:** Shows if the patient had any antiretroviral therapy before a certain time.

**z30:** Indicates if the patient took a certain type of medication in the 30 days before a certain time.

**preanti:** Tells us the number of days before a certain time the patient started anti-retroviral therapy.

**race:** Indicates the race of the patient.

**gender:** Tells us the gender of the patient.

**str2:** This attribute shows if the patient has taken antiretroviral medication before or if they're new to it.

**strat:** It stratifies or groups patients based on their history of taking antiretroviral medication.

**symptom:** Shows if the patient had any symptoms of AIDS when diagnosed.

**treat:** Indicates the type of treatment the patient received.

**offtrt:** Shows if the patient stopped treatment before a certain time.

**cd40:** This measures a type of white blood cell called CD4 at the beginning of the study.

**cd420:** This measures CD4 levels at a certain time into the study.

**cd80:** This measures another type of white blood cell called CD8 at the beginning of the study.

**cd820:** This measures CD8 levels at a certain time into the study.

**infected:** Finally, this tells us if the patient is infected with AIDS or not.

The problem I wanted to aim for is to address the need for a comprehensive measure to compare healthcare outcomes related to AIDS across different populations or regions.

The AIDS dataset offers a comprehensive repository of healthcare statistics and patient information specific to AIDS diagnosis and treatment. By utilizing this dataset, I could develop a tailored composite indicator that effectively measures and compares healthcare outcomes related to AIDS across various populations or regions. This approach allows for a focused analysis of factors influencing disease progression and treatment effectiveness, ultimately contributing to a deeper understanding of HIV/AIDS epidemiology and informing targeted healthcare interventions.

**Cleaning:**

A screenshot of a computer program

Description automatically generatedInitially looking at the dataset I set out to first check for empty or duplicate values that could cause hinderance further down

upon checking it was found that there were no duplicate values or missing values as this dataset was made with respect to AIDS research. Then I proceeded to produce a descriptive statistical summary of a dataset which would help in understanding non-missing values, the mean, standard deviation, the min and max values in each column.

A screenshot of a computer

Description automatically generated

Based on the colour scheme I used in the above output the following things could be understood, red represents lower values, yellow represents values closer to the middle (median), and white represents higher values. This helped me to identify outliers in the data more easily.

**For example**, in the age column, the minimum value is 12 and the maximum value is 70. The 50th percentile (median) is 34, which means half of the patients in the dataset are younger than 34 and the other half are older than 34. The colour coding might also show that there are more patients in their 30s and 40s (yellow) compared to younger or older age groups (red and white).

Then I got a bar graph and pie chart up to set a clear picture as to how many patients were infected with AIDS or no to get clear idea of the data

A screen shot of a graph

Description automatically generated

**Multivariant Analysis:**

**) Heat Map**

Used ‘heatmap’ for visualization of the correlation matrix. This process is a common

multivariate analysis technique as color-coded representation making it easier to see the

strength of relationships between variables.

A blue and white squares with black and white text

Description automatically generated

**PCA:**

PCA is a critical tool in Multivariate Analysis because it simplifies the complexity in high dimensional data while retaining trends and patterns. This is particularly useful when dealing with multivariate data, allowing easier visualization and better understanding of relationships between variables.

A diagram of a diagram with blue dots

Description automatically generated

The plot 'Residuals vs Fitted' helps to verify if the residuals (errors) have constant variance across all levels of the fitted values. This is crucial because non-constant variance (heteroscedasticity) can lead to inefficient estimates and affect the generalizability of the model. The 'Normal Q-Q' plot is used to assess if the residuals of the model are normally distributed. Normality of residuals is an important assumption in many statistical tests that involve regression models because it affects the validity of the hypothesis tests.

A graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of

Description automatically generated

K-Means clustering using the Elbow Method to determine the optimal number of clusters.

partitioning the data into K distinct clusters and plotting the sum of squared errors (SSE) for

different values of K and looking for a 'knee' in the curve.

A graph with a line

Description automatically generated

**Visualisation:**

Pairplot:

The pairplot shows pairwise relationships between numerical variables (time, age, wtkg, preanti, cd40, cd420, cd80) in the dataset. Each scatterplot in the pairplot represents the relationship between two variables. The diagonal of the pairplot displays histograms of each variable.

**A screenshot of a graph

Description automatically generated**

**Scatter Plots:**

The scatter plot of cd80 vs. cd820 shows how the cd80 varies with the cd820. It helps in understanding the relationship between the cd80 and cd820. The scatter plot of cd820 and cd420 shows how cd820 varies with cd420. It helps in understanding how the condition of the blood cell cd820 varies from cd420 over time and how it could impact the patient.

A diagram of a graph

Description automatically generated with medium confidence

A diagram of a number of blue dots

Description automatically generated

Using catboost classifier I was able to find out the top important features of this dataset that can be used for a prediction model while making one for AIDS infections dataset

|  | **feature\_importance** | **features** |
| --- | --- | --- |
| 0 | 38.948878 | time |
| 1 | 9.192205 | cd420 |
| 2 | 8.730946 | offtrt |
| 3 | 6.128697 | wtkg |
| 4 | 4.814286 | cd40 |
|  |  |  |

A graph with blue squares

Description automatically generated

In conclusion, the development of a composite indicator using the AIDS dataset represents a significant step towards enhancing our understanding of healthcare outcomes related to AIDS. By creating a diverse array of healthcare statistics and patient information, i have constructed a comprehensive measure that facilitates meaningful comparisons across different populations or regions. This project not only provides valuable insights into the factors influencing disease progression and treatment efficacy but also serves as a valuable tool for informing healthcare policies and interventions aimed at improving outcomes for individuals affected by HIV/AIDS. Moving forward, the insights gleaned from this project can guide future research efforts and contribute to the ongoing efforts to combat the global burden of HIV/AIDS.