Web Application for AIDS detection using Machine Learning

Student Name: Sandeep Kumar Daki
Student Number: 21069435

Course: M.Sc. Data Science with Advance research
Supervised by: Michael Kuhn

Chapter 1: Project Plan

Project Title: Web Application for AIDS Detection Using Machine Learning

Research Question

- How do various machine learning algorithms compare in terms of accuracy and efficiency in the detection of AIDS?
- Which features have the highest impact on predicting the infection status?
- How do patient outcomes differ between various demographic groups (e.g., age, gender, weight)?

Aim

This project aims to develop a reliable machine-learning model that can predict AIDS outcomes and compare multiple machine-learning algorithms using this application. The project will use the Streamlit framework to create an interactive web application that can be accessible easily.

Objectives

- To create a machine learning model using a comprehensive dataset on AIDS.
- To train the model with suitable algorithms, such as logistic regression, random forests, and neural networks.
- To validate the accuracy of this model in predicting AIDS based on clinical and demographic indicators.
- To develop a user-friendly web interface that allows users to input their data and receive predictions.
- To incorporate interactive elements in the website for better user engagement and understanding.

Background and Summary

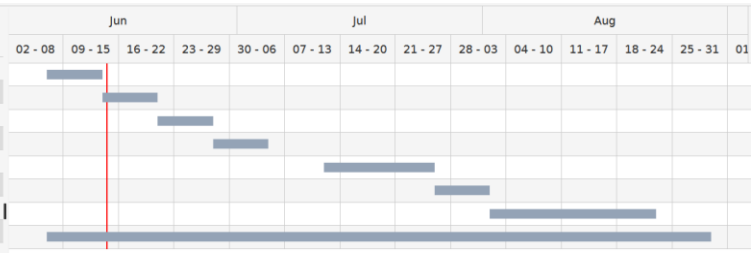
AIDS is a disease that stands for Acquired Immunodeficiency Syndrome, it is widely known as a life-threatening condition caused by a virus known as Human Immunodeficiency Virus (HIV) (*HIV and AIDS*, no date). Currently, there is no cure known cure for HIV, but early detection plays a major part in extending the life of the person affected with AIDS. This is where ML can help in the early diagnosis, or at least help the medical professionals to speed up the process (He *et al.*, 2022).

Initially, an extensive AIDS dataset will be utilized to develop a machine-learning model. Suitable algorithms, such as logistic regression, random forests, and neural networks could be considered for this classification task. After the model has been trained and validated to ensure it predicts accurately, the focus will shift to developing the web interface using Streamlit. Streamlit allows for easy integration of interactive widgets, such as sliders, buttons, and text inputs, which facilitates efficient collection of user inputs. These inputs then act as parameters for the machine-learning model. The Python script that powers the website will incorporate functions to process these inputs through the model and display the predictions on the web page. Streamlit also supports embedding data visualizations, which will present charts and graphs that can help users understand the model's predictions.

Upon completing this project, it is expected that the website will function as a tool that aids in the prediction of AIDS outcomes, leveraging the power of machine learning. This will demonstrate the practical application of data science in solving significant health issues and provide a user-friendly platform for users to interact with the model. The integration of interactive elements and data visualizations will further enhance the understanding and accessibility of predictive analysis, making it a valuable resource for users and researchers alike.

Chapter 2: Task List and Project Time Line

ID	Title	Start Time	End Time
1	Data Collection and Project Plan	06/07/2024	06/14/2024
2	Literature Review	06/14/2024	06/21/2024
3	Data Pre-Processing	06/21/2024	06/28/2024
4	Feature Engineering	06/28/2024	07/05/2024
5	Model Development	07/12/2024	07/26/2024
6	Model Comparison	07/26/2024	08/02/2024
7	Application Development	08/02/2024	08/23/2024
8	Final Report	06/07/2024	08/30/2024



Data Collection and Project Plan: This is a shorter module that is designated to gather the data and plan the project based on the given deadlines.

Literature Review: A relatively longer module to understand the literature present on this project and help to fill the gaps where needed with the help of this research.

Data Pre-processing: This module is responsible for cleaning the data, checking for missing values, cleaning the data of any inconsistencies, etc.,

Feature Engineering: In this module, the data is thoroughly filtered using multiple statistical and mathematical methods to create a final version which is best suitable for model development.

Model Development: This is the most important part of the research because based on the quality of model development the results will be reflected. This module consists of model selection, training, testing, and tuning the model to the given data.

Model Comparison: After the shortlisted models are developed and tested, this module will go in depth of comparing the algorithms based on multiple evaluation metrics.

Application Development: To make the previous steps easier for future use, this module will be responsible for developing an application to automate the process of model testing and comparison.

Final Report: In this final part, the results and the journey of research are documented with a discussion about the process, the results, and the meaning of the results and how they affect the future scope of this project.

Chapter 3: Data Management Plan

Summary of Dataset

The dataset has 4 files each with varying sizes of patient records. One important thing to note is the data does not contain patient identifying information like name, email, etc., Other personal information like age and gender are present. The dataset contains 23 features related to patients along with the target variable. Some of the important types of features available are patient demographic data, treatment type, health indicators, CD4 cell count, and other features. The target variable “infected” indicates if the patient is infected with AIDS or not.

Data collection

The data used in this research report is taken from an open-source website called Kaggle, where all the data is free to use for academics and research purposes.

Source: https://www.kaggle.com/datasets/aadarshvelu/aids-virus-infection-prediction

Document control

In every research, the most important part of report writing/code development is version control. It is important to track the additions and updates made in a document / script. This helps the researcher and the person reading the document to get more information on understanding the different parts of the research (Shepelev, 2021).

In this research project, the version control or document control system used is Git – an Open-Source Version control system and GitHub to host the repository for the research document and code.

Ethical requirements

Since the data is not collected directly from users, no ethical form is needed for this data. The data does not have any personal identification information and it is available to use for research purposes without any restrictions. Hence, the data follows GDPR requirements, conforms to UH ethical policies and it is collected ethically by the primary researchers.

List of References

1. He, J. *et al.* (2022) "Application of machine learning algorithms in predicting HIV infection among men who have sex with men: Model development and validation," *Frontiers in public health*, 10. doi: 10.3389/fpubh.2022.967681.

2. *HIV and AIDS* (no date) *Who.int*. Available at: https://www.who.int/news-room/fact-sheets/detail/hiv-aids (Accessed: June 14, 2024).

3. Shepelev, A. (2021) *Why git is A great documentation management tool*, *Hackernoon.com*. Available at: https://hackernoon.com/why-git-is-a-great-documentation-management-tool-p712339s (Accessed: June 14, 2024).

4. Costales, J. A., Lorico, E. M. and de Los Santos, C. M. (2023) "A comparative sentiment analysis about HIV and AIDS on twitter tweets using supervised machine learning approach," in *2023 5th International Conference on Computer Communication and the Internet (ICCCI)*. IEEE, pp. 27–32. https://ieeexplore.ieee.org/document/10210162

5. Mahto, R. and Sood, K. (2024) "HIV progression and outcome prediction to enhance patient matching for clinical trials," in *2024 IEEE 14th Annual Computing and Communication Workshop and Conference (CCWC)*. IEEE, pp. 0278–0284. https://ieeexplore.ieee.org/document/10427778

6. Mulyadi, W. J. and Qomariyah, N. N. (2023) "Using machine learning to Analyse the effect of antiretroviral therapy (ART) on people with HIV," in *2023 10th International Conference on ICT for Smart Society (ICISS)*. IEEE, pp. 1–5. https://ieeexplore.ieee.org/document/10291717

7. Varshney, N. *et al.* (2023) "Early detection of HIV infection with machine learning from blood test results," in *2023 International Conference on Artificial Intelligence for Innovations in Healthcare Industries (ICAIIHI)*. IEEE, pp. 1–6. https://ieeexplore.ieee.org/document/10489256