**Title:** Comparison of complex and simple machine learning techniques to predict a positive HIV diagnosis

**Group Members:** [Nelson Evbarunegbe](mailto:iyorenelson@gmail.com), [Thomas Potts](mailto:tpotts@umass.edu)

**Justification for a Group Size Exceeding Two:** Group size does not exceed two.

**Background:** Human immunodeficiency virus (HIV) is an ongoing epidemic that has far-reaching impacts across the world, killing tens of millions over the past few decades. There exists a social stigma surrounding HIV’s association with sexuality [1], its lack of a definitive cure,  as well as its adverse effects if it progresses to become acquired immunodeficiency syndrome (AIDS). Though the height of the epidemic seems to have passed, HIV is an incredibly important disease to research using modern tools. In our case, we will be applying machine learning (ML) methods in order to classify HIV outcomes.

**Objective:** Specifically, we will be comparing various machine learning models to determine if predictability can be improved as compared to a basic logistic regression model. This will be a significant contribution to the studies surrounding HIV to test if advanced ML models can perform better than baseline regression.

**Dataset:** The Ethiopian Demographic and Health Survey (EDHS) is a large health survey which includes important attributes regarding HIV [2]. The survey was implemented by the Central Statistical Agency (CSA) and the DHS program with over 80,000 instances and 33 attributes (this includes both discrete and binary variables) with a binary output to predict HIV test status of individuals [2]. The dataset is publicly available on GitHub [3].

Additionally, as a side project in case of more time and not instrumental to our current proposal, we are also interested in the Professional Healthcare Institute of America (PHIA) datasets, but we are unsure if we will have access to the data in a timely manner. PHIA is a HIV survey of sub-Saharan African countries and Haiti, where surveyors would go to people’s houses and ask questions about their health, and provide them with an HIV test [4]. This survey is run by each country’s Ministry of Health and Columbia University, and it is funded by the CDC [4]. We have applied to have access to the data and are waiting for a response.

**Methods:** Our selected advanced machine learning models would be the gradient boosting machine and Gaussian mixture models, while our simple model would be a logistic regressor. Firstly, we would perform feature engineering on the dataset and remove features that may not be important in the prediction of our target variable. After data cleaning, Spearman correlation will be employed to the cleaned dataset, and this will be observed to check the most impacting features. Beyond feature engineering, we build our algorithms via scikit-learn and then evaluate the performance of our models using F1-score and ROC-AUC (classification problem).

The train-validation-test split on the dataset will be 60:20:20 and the target variable will be ‘HIV status’ with a high-dimensional input space of predictor variables (32 features) such as age and sex which can be found in the dataset link attached. The dataset has 83,100 data points.

**Results:** We intend to analyze the effectiveness of the gradient boosting machine and Gaussian mixture models versus our logistic regressor. We expect the advanced machine learning model to outperform the logistic regressor, but we are also interested in the difference in this performance and the effect of the added complexity. Additionally, we are interested in the model evaluation performance of the models based on the HIV dataset.

**Division of Labor (tentative)**:

Nelson Evbarunegbe: Machine learning, model design, evaluation, and optimization

Thomas Potts: Exploratory data analysis, data cleaning, model comparisons, and regression modeling

**References:**

[1] World Health Organization, [Human deficiency syndrome (HIV)](https://www.who.int/news-room/fact-sheets/detail/hiv-aids#:~:text=HIV%20targets%20the%20body%27s%20white,milk%2C%20semen%20and%20vaginal%20fluids.)

[2] D. Mesafint Belete and M. D. Huchaiah, ‘A Deep Learning Approaches for Modeling and Predicting of HIV Test Results Using EDHS Dataset’, Infectious Diseases. IntechOpen, Feb. 08, 2023. doi: 10.5772/intechopen.104224.

[3] EDHS dataset: [EDHS Dataset by danielmesafint](https://github.com/danielmesafint/Datasets)

[4] PHIA Project: [About - PHIA Project](https://phia.icap.columbia.edu/about/)