Contribution ID: **1** Type: **Talk**

An interpretability machine learning approach for identifying factors associated with never taking HIV tests among Tanzanian women

Taking an HIV test can be difficult for a lot of people worldwide. This can generally be associated with several factors, such as anxiety, among others. Using a holistic approach to big survey data, we explore factors underlying a phenomenon of not testing for HIV among Tanzanian women aged between 15 to 49 years. The sample size of the dataset is 14801, with 2685 records of ever tested' and 12116 records ofnever tested'. Several classification techniques were piloted for performance. However, the Logistic Regression and XGBoost (Extra Gradient Boosting) were retained as prediction or classification methods.

Using these two methods, the study examined factors which contribute to the prediction of the target outcome. The XGBoost based methods used are the composite machine learning models of XGBoost and various feature selection techniques. These are XGB-SFFS using sequential forward floating search (SFFS) method, XGB-SFM using Select From Model (SFM) method, and XGB-SHAP using SHapley Additive exPlanations (SHAP) which performs both extraction and ranking of features. The Logistic Regression (LR) based methods using similar extraction methods are LR-SFFS, LR-SFM, and LR-SHAP.

We implemented the models for the purpose of extracting relevant features that mostly influence the target outcome. The models exhibited high competency with almost 100% performance observed for XGBoost based methods, including 10-fold cross-validation metrics. The XGBoost has exhibited a remarkable strength as one of the most efficient and intelligent prediction method with unique properties. We further explored model performance at both default and optimal hyperparameter settings. In terms of feature extraction, XGB-SFFS was computationally more expensive compared to other algorithms. The XGB-SHAP approach emerged as the most effective feature extraction technique. The SHAP technique substantially contributed to interpretability with key insights into the individual impact of features. The top three main factors associated with never testing for HIV among others were factors such as candidates being concerned about test results, the number of entries in pregnancy history and STI awareness. The ensemble frameworks of XGBoost and feature selection methods implemented in this research study present an efficient tool for binary classification using a socio-behavioural dataset. Therefore, one main reason why some people have never tested is the fear of testing positive

Primary authors: Prof. SEITSHIRO, Modisane; Mr NZUZA, Mphiliseni

Presenter: Mr NZUZA, Mphiliseni