

Comparing Decision Trees and Logistic Regression in Predicting HIV among Women in South Africa

Author: Oluwapelumi Oladokun

To facilitate the goal of improving women's reproductive health outcomes in underdeveloped regions, effective HIV risk reduction strategies are needed. One of the strategies is to assess the level at which an individual is at risk of HIV based on factors such as biological, social, and other behavioural factors. In low resource settings where there is a high disproportionate distribution of the number of doctors, staff, facilities and patients, it is important that patients are categorized and prioritized based on their risk to ensure save time and encourage efficient and effective service delivery.

In this study, performances of Logistic regression (LG) and Decision Trees (DT) were compared to predict HIV among women in South Africa because based off current knowledge, no study has used machine learning to predict HIV risk. Data was from the Demographic and Health Surveys (DHS) Program (DHS, 2016). Study participants were 7808 women living in South Africa aged 15 to 49 years. Data was split into training (75%) and testing (25%) datasets. The ROC Curve and statistical evaluation measures (accuracy, true positive, false negative, etc) were used to evaluate the performance of the models.

The decision tree model had the highest accuracy for both training (70.33%) and testing dataset (68.19%). Accuracy for the LG model was 45.62%. The AUCs from the ROC curve reported 0.697 and 0.667 for the DT and LG respectively. This means that on average, a woman will be predicted as HIV negative 69.7% of the time as compared to being HIV positive using the DT model and 66.7% using the LG model.

Machine learning is on the rise and should be used in social science research for developing better theories and making future predictions. Although logistic regression and decision trees have similar purposes, the results conclude that the decision tree algorithm is better in prediction accuracy.