Applying Machine Learning To Classify Disease Status For Selected Notifiable Medical Conditions In South Africa

Introduction

There is a change in disease profiles. Environmental variabilities continue to alter morphological appearances of species necessitating enhancement in diagnostic methods used to detect diseases. The deterministic approaches applied in the current diagnosis methods for Malaria and COVID-19 have presented challenges of low sensitivity and specificity. In this study, we described data structures and disease profiles for Malaria and COVID-19 surveillance data at the National Health Laboratory Services (NHLS), South Africa. We also explored the application of supervised Machine Learning (ML) to classify and predict clinical outcomes for Malaria and COVID-19.

Methods

The COVID-19 surveillance data comprised of 35,202 observations from a unit dataset. The Malaria data was made up of three files; a demographics file, a laboratory results file and a travel-treatment history file of which 40,094 observations were deduced. These datasets were divided into two portions, 75% for model specification and the 25% designated as out-of-sample testing. We compared three supervised ML classifiers: Support Vector Machine (SVM) the K-Nearest Neighbor (KNN) Random Forests (RF) with their variant novelty approaches Isolation Forest (iForest) and One-Class Support Vector Machines (OCSVM) to predict clinical outcomes for Malaria and COVID-19. To account for severe label imbalances, the data with majority class labels was under-sampled to obtain an equal class balance in the target. Novelty detection approaches with iForest and One-Class Support Vector Machines (OCSVM) were also used in classifying and predicting Malaria and COVID-19 clinical outcomes.

Results

Malaria surveillance data was characterized by large proportions of missing data for demographic, syndromic and environmental characteristics. Though complete, compared to Malaria, COVID-19 surveillance data did not follow tidy-data principles. In evaluating classifier predictive power using out-of-sample data with equal representation of clinical outcomes, RF yielded the best predictive power with Area Under Curve (AUC) scores (98%) from Malaria out-of-sample data accounting for distribution weight of clinical outcome. Though not comparable to scores from Malaria data, the RF still scored better than the SVM and KNN classifiers from out-of-sample evaluation over COVID-19 data. Generally, lower classifier performance was observed across all models when subjected to COVID-19 out-of-sample data, where the KNN classifier registered the highest

number of false-positive results. There were significantly higher numbers of False-Negative predictions with the SVM classifiers compared to the RF and KNN. However, the RF performed slightly better in predicting True-Negative observations. By categorizing data with minority clinical outcome representation as outliers, OCSVM predicted more negative observation compared to the iForest.

Conclusions

This study showed the impact of data quality in disease surveillance with respect to predictive modeling for Malaria and COVID-19 medical conditions. The data were characterized by large proportions of incompleteness. Individual demographic characteristics reported and recorded signs and symptoms among other attributes that hold vital information for syndromic disease surveillance were lacking. While supervised ML classifiers performed well with Malaria out-of-sample data, the same methods produced suboptimal results with similar surveillance COVID-19 data. Future studies could explore unsupervised ML approaches on the same surveillance data.