Feature Selection Approach to Improve Malaria Diagnosis Model's Performance for High and Low Endemic Areas of Tanzania

By
Martina Mariki, Neema Mduma and Elizabeth Mkoba
Nelson Mandela African Institution of Science and Technology, Tanzania 2022

Malaria remains a significant cause of death, especially in sub-Saharan Africa, with about 228 million malaria cases worldwide. Parasitological tests, in the form of microscopic and rapid diagnostic tests (RDT), are the recommended and standard tools for diagnosing malaria. However, in areas where parasitological tests for malaria are not readily available, clinical diagnosis is advised. This method is the least expensive and most widely practiced. A clinical diagnosis called presumptive treatment is based on the patient's signs and symptoms and physical findings at the examination. A malaria diagnosis dataset was extracted from patients' files from four (4) identified health facilities in the regions of Kilimanjaro and Morogoro. These regions were selected to represent the country's high endemic areas (Morogoro) and low-endemic areas (Kilimanjaro). The dataset contained 2556 instances and 36 variables. The random forest classifier, a tree-based, was used to select the most important features for malaria prediction. Regional-based features were obtained to facilitate accurate prediction. The feature ranking indicated that fever is universally the most noteworthy feature for predicting malaria, followed by general body malaise, vomiting and headache. However, these features are ranked differently across the regional datasets. Subsequently, six predictive models, using important features selected by the feature selection method, were used to evaluate the performance of the features. The features identified complies with malaria diagnosis and treatment guide lines provided by WHO and Tanzania Mainland. The compliance is observed to produce a prediction model that will fit in the current healthcare provision system.