Benha University,
Faculty of Business, Department of Statistics,
Mathematics and Insurance

Performance Evaluation of Statistical and Machine Learning Models with an Application

PREPARED BY

Ayman S. Hegazy
Teaching Assistant, Department of Statistics, Mathematics, and Insurance,
Faculty of Business, Benha University

SUPERVISED BY

Prof. Mervat Mahdy Ramadan
Professor of Statistics,
Department of Statistics, Mathematics, and Insurance
Faculty of Business, Benha University

Dr. Dina Samir El-telbany
Lecturer of Statistics,
Department of Statistics, Mathematics, and Insurance
Faculty of Business, Benha University

A Thesis Submitted to the Department of Statistics, Mathematics and Insurance, Faculty of Business, Benha University in Partial Fulfillment of the Requirement for the Master of Degree in Applied Statistics

2023

ABSTRACT

Anemia is a prevalent issue in children and is still an important health problem in Ethiopia and other parts of Africa. A child who has anemia does not have enough red blood cells or hemoglobin. Hemoglobin is a type of protein that allows red blood cells to carry oxygen to other cells in the body to survive. So, a prediction of anemia is necessary for planning public health interventions against this disease, and for clinical care of children across the life course. This study proposes statistical and machine learning models to diagnose anemia. Some machine learning techniques also have been used in the stage of data pre-processing to give better results such as whitening transformation with principal component analysis (PCA) has been used in this study to avoid the overfitting problem in addition to *RobustScaler* function to pre-process the data and adjust the outliers. Seven classifiers, including Logistic Regression, Support Vector Classification, k-Nearest Neighbor, Decision Tree, Random Forest, AdaBoost and Gradient Boosting, are implemented in this study. Moreover, the proposed hybrid classifier (LSA) model has been built, which is a combination of the best machine learning classifiers is used in this study: Support Vector Classification, Logistic Regression and AdaBoost by using soft voting strategy. The hyperparameter tuning is performed with cross-validation to find out the optimal value of this hyperparameter to achieve the best accuracy. The performance of the models is evaluated based on the confusion matrix, recall, precision, f1-score, accuracy, and Matthews correlation coefficient (MCC). Although the results show the best accuracy for Support Vector Classification, Logistic Regression, and AdaBoost classifiers reached 99.66%, 99.57% and 99.15% respectively, the proposed LSA model achieved the highest accuracy, which reached to 99.83%, with recall values of 99.80%, precision values of 99.80%, f1-score of 99.80%, and Matthews correlation coefficient values of 99.65% as compared to each an individual classifier.