

1159-00-109

Francis Biney* (fbiney@utep.edu) and **Maria C. Mariani** (mcmariani@utep.edu).

Comparing predictive performance of statistical learning models on medical data.

This work investigates the predictive performance of 10 Machine learning models on three medical data including Breast cancer, Heart disease and Prostate cancer. Furthermore, we use the models to identify risk factors that contribute significantly to these diseases.

The models considered ranged from Logistic regression to Deep-feedforward neural network. The models were grouped according to their similarities and learning style; Linear and non-Linear models.

The model selection and hyperparameter tuning were done using bias-variance tradeoff and cross-validation. The model's performance was improved for each method, by applying early stopping, dropout and removed non-significant variables to avoid overfitting. Different predictive performance measures were used including prediction accuracy, sensitivity and specificity depending of the nature of the response distribution whether balanced or imbalanced to compared the models.

The result show that the non-linear models; SVM-GK, RF and FFNN gave the best predictive performance for the breast cancer data. The linear models; LR-PLS, LR-PC ,LR-Ridge and LR-Enet were preferred for heart disease and mixture of linear and non-linear models; LR-Lasso, LR-Enet, RF and GBM best describes the Prostate cancer predictions. (Received August 02, 2020)