Project 3

Six ML models were compared on many sets of data

Models used: Logistic Regression, Decision Tree Classifier, K Neighbors, SVC, Linear Discriminant Analysis & Gaussian NB

Stroke data with 43400 records, heart disease with 303 records, hepatitis with 305 records and Parkinsons disease with 195 records were used. Accuracy score for training and testing data generated. Best model for each disease identified.

Stroke data was truncated to 3260 and 434 records to study the effect of record size. Comparison plots generated. Larger data size(>10k) is preferred.

Stroke data with 4300 records was also used to study the effect of 10/90, 20/80, 30/70 and 40/60 split of dataset into test/train data. Comparison plots generated.