Data Modeling – Approach

Main objective of the program is

1. To identify diabetes patients who are likely to be readmitted in hospitals, through machine learning techniques, which would benefit the hospitals as they can proactively take measures to reduce the chances of a re-admission.

This dataset has close to 50 feature variables and plotting a co-relation for each of these variables would be impractical.

These variables are a mix of numerical and objective variables and a good majority of the variables are categorical. So I would be implementing One Hot Encoding to enable the predictive power of those variables.

The data has class imbalance. 88% of the patients have not sought readmission in 30 days and only 12% have been readmitted in 30 days. With a greater imbalanced ratio, the decision function favor the class with the larger number of samples, usually referred as the majority class. So the accuracy would be affected.I plan to use undersampling or over sampling from ‘imblearn’ to make only the training data more balanced. The test data would not be touched and would be left as is.

Algorithms would be trained on this balanced trained data and then used to predict the test data.

ML algorithms

I plan to use

Decision Trees,

Random Forest

Logistic Regression

Gradient Boosting Classifier

Adaboosted Classification

and

K Nearest Neighbor (KNN)

And then compare the results to find out which algorithm has the most accurate predictions