**Content**

1. Importing Data

2. Exploratory data analysis

3. Data preprocessing for modeling

4. Model selection by KFold Cross validation

5. Gradient Boosting

1. Hyperparameter Tuning
2. Variable selection by Feature importances
3. Gradient Boosting final model
4. Model evaluation metrics

6. Logistic Regression(2nd best model)

1. Finding optimum cutoff point
2. Comparison b/w cutoff values
3. Model evaluation metrics

**Summary of the Analysis**

Imported raw data in Jupyter then analysed the nature of data by checking type of variables, missing values, correlations b/w variables, and distribution of individual variables.

Next have done EDA with few different variables to understand the persistency of drug on patients (by Gender, Region, etc.)

Prepared data for modeling by encoding different categorical variables and created dummy variables.

To select best model for data: used the Kfold cross validation method, in this applied 'Logistic Regression', 'Decision Tree', 'Random Forest', 'Ada Boost', and 'Gradient Boosting' out of these models **'Gradient Boosting’ and 'Logistic Regression'** had the best scores, so chosen these two models.

Built the **'Gradient Boosting’ model** by tuning their hyperparameters (like max\_depth, max\_features, min\_samples\_leaf, min\_samples\_split, n\_estimators) and selecting top-18 variables by their ‘Feature importances’ scores in model then checked the model performance by ROC, AUC, Classification table, and [Recall and Precision].

Built the **2nd best model ‘Logistic Regression’** by optimum cutoff point and variables used in Gradient Boosting model then checked the model performance by ROC, AUC, Classification table, and [Recall and Precision].