# Medication Compliance Forecasting

I used a boosting ensembler to predict patient’s likelihood of adherence to prescribed regimen.

XGBoost is one of the best ensembler for this problem statement.

Tools - anaconda, jupyter notebook

Python packages - numpy, seaborn, matplotlib, pandas, sk-learn

Steps:

* I performed exploratory data analysis on the given data and found out that there is no missing values in the dataset.
* Age is a continuous variable so I plotted one histogram to check the trend based on this feature. I created five different groups based on the age group like ‘0-15’,’15-30’.... and assigned each patient that group based on their age. New feature name is ‘age\_categories’.
* Patient\_id is not going to be useful for the prediction. It’s there to represent uniqueness. I removed the patient\_id and age features from the training and test dataset.
* Only one categorical feature left in our dataset. XGBoost don’t accept categorical variable so I implemented one hot encoding on this variable. I used onehotencoder from sklearn preprocessing package to perform this operation.
* Now dataset contains five more columns based on one hot encoding and removed that categorical one.
* I implemented feature scaling(standardization) so that every feature is equally important in the final prediction. I used standardScaler from sklearn preprocessing package.
* I split the training data to train and validation in 3:1 ratio so that i can check the f1 on the validation data.I used train\_test\_split from the sklearn model\_selection package.
* Now our data is ready to feed to the XGboost model. I used randomizedSearchCV to find the best hyperparameters suited for our model.
* I used f1\_score ,classification\_report and confusion matrix metrics to check whether my model is a good one. I compared my prediction on the validation data to the actual value and got the f1 score of 0.84.

**F1 Score - 0.84**

* Performed the predictions on test data and created a final result csv which contains each patient adherence with their predicted probability.