# Medication compliance forecasting

## Overview:

The objective was to build a predictive model that will predict a patient's likelihood of adherence to a prescribed regimen.

In-depth analysis along with the code is present here

## Data:

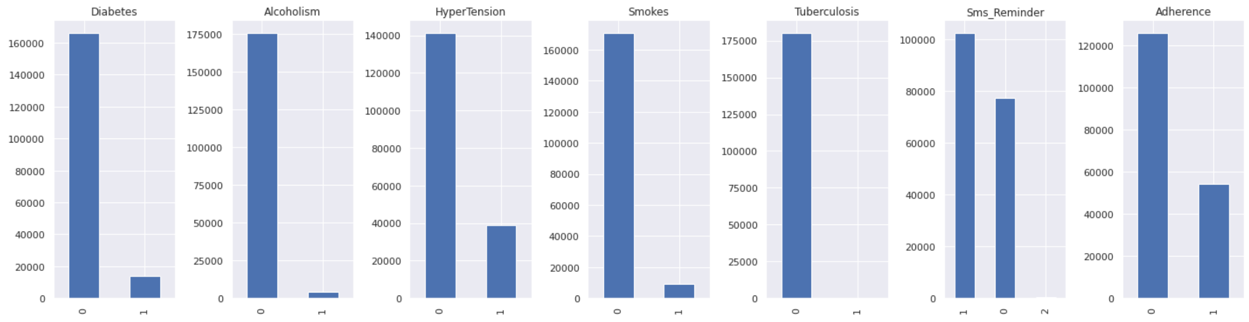
Data consists of id, along with 9 feature columns and 1 target column. The data represents patients and the target variable is adherence.

## EDA:

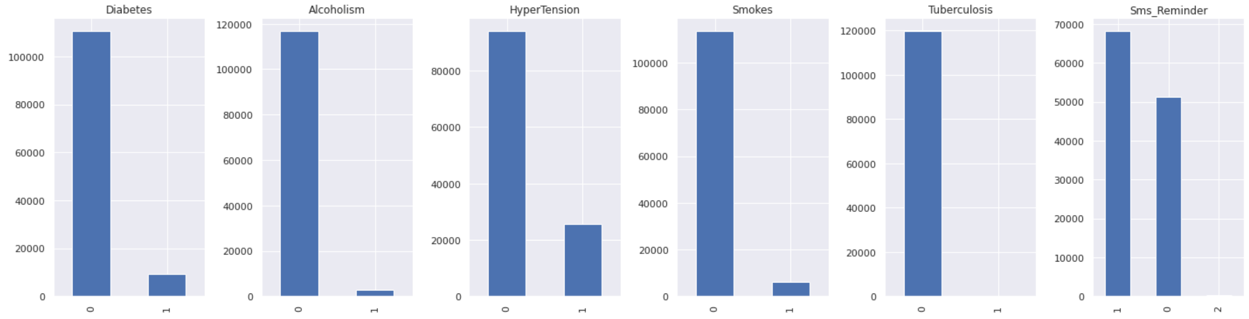
On performing detailed analysis on the training and test datasets, the following points were deduced:

1. Both train and test datasets have a similar distribution of patients with respect to all the features including adherence

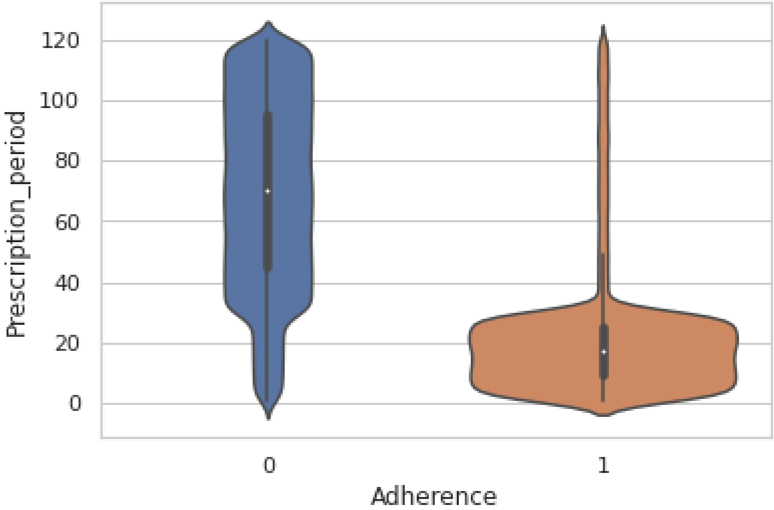
Distribution of features in training data



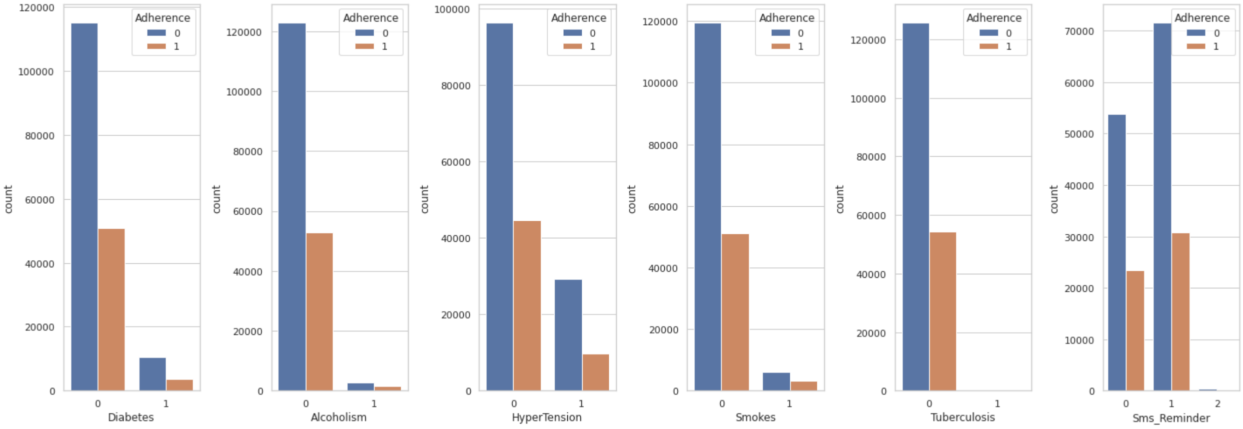
Distribution of features in test data



2. The output feature - adherence was unbalanced in the entire dataset. There are ~30% of patients who adhere to the regimen as compared to ~70% of patients who do not adhere to their regimens
3. If a patient gets a regimen of 30 days or below, there is a higher chance that he/she will adhere to the regimen

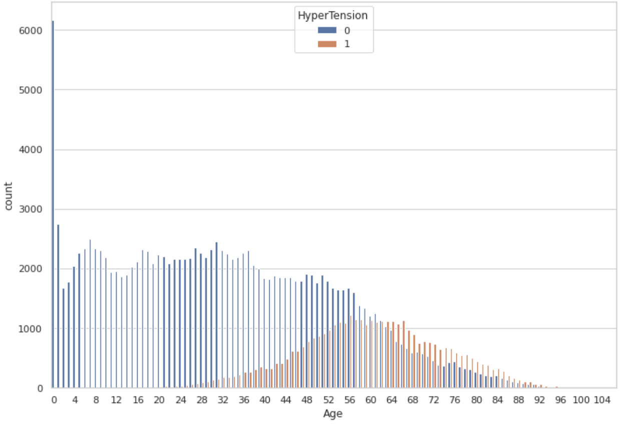


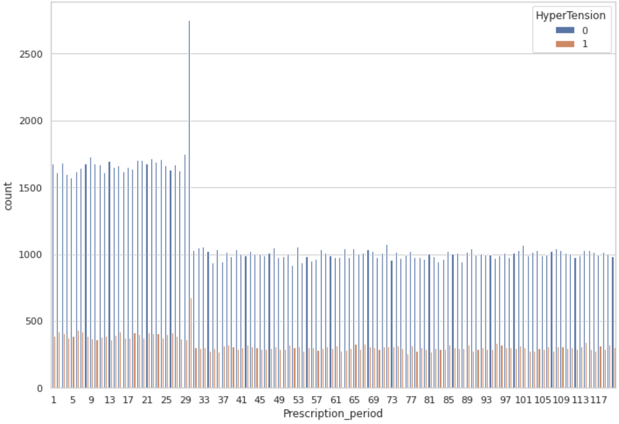
4. SMS reminders do not affect the tendency of a patient to follow a regimen or not follow a regimen. Single SMS reminders are same as double SMS reminders in terms of adherence



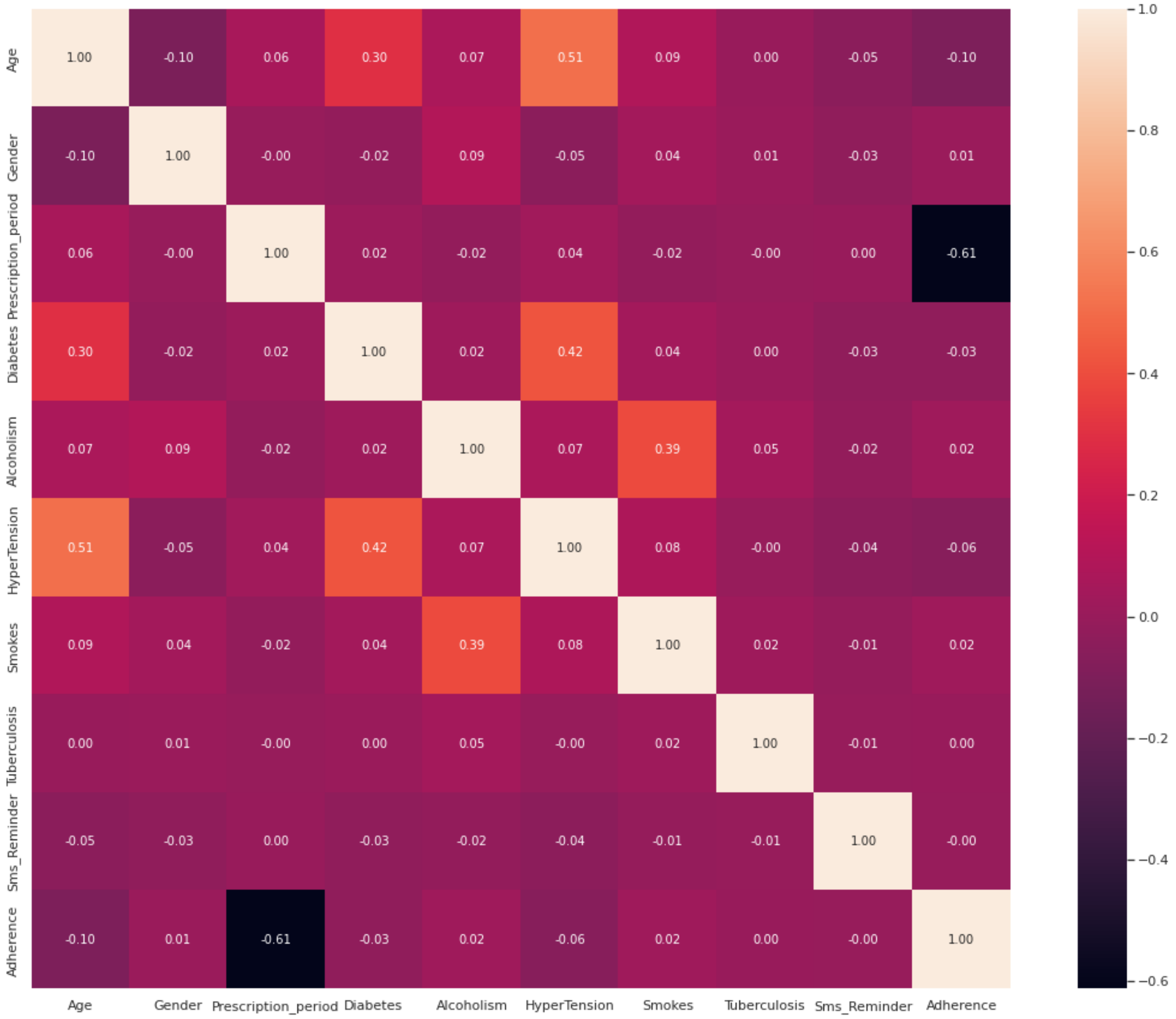
5. There is a higher number of patients who are 0 and 1 year old. Initial analysis suggested that this might be a data anomaly. There is a chance that these are babies. There is also a possibility that while data collection, patients whose age was not known were marked under 0 or 1 category. However, on deep diving into age related datacuts, there was not even a single patient found with age = 0 or 1 and who smokes / drinks / is hypertensive /

diabetic. Hence, these patients were considered to be genuine entries





## Correlation Matrix



| | Age | Gender | Prescription_period | Diabetes | Alcoholism | HyperTension | Smokes | Tuberculosis | Sms_Reminder | Adherence |
|---|---|---|---|---|---|---|---|---|---|---|
| Age | 1.00 | -0.10 | 0.06 | 0.30 | 0.07 | 0.51 | 0.09 | 0.00 | -0.05 | -0.10 |
| Gender | -0.10 | 1.00 | -0.00 | -0.02 | 0.09 | -0.05 | 0.04 | 0.01 | -0.03 | 0.01 |
| Prescription_period | 0.06 | -0.00 | 1.00 | 0.02 | -0.02 | 0.04 | -0.02 | -0.00 | 0.00 | -0.61 |
| Diabetes | 0.30 | -0.02 | 0.02 | 1.00 | 0.02 | 0.42 | 0.04 | 0.00 | -0.03 | -0.03 |
| Alcoholism | 0.07 | 0.09 | -0.02 | 0.02 | 1.00 | 0.07 | 0.39 | 0.05 | -0.02 | 0.02 |
| HyperTension | 0.51 | -0.05 | 0.04 | 0.42 | 0.07 | 1.00 | 0.08 | -0.00 | -0.04 | -0.06 |
| Smokes | 0.09 | 0.04 | -0.02 | 0.04 | 0.39 | 0.08 | 1.00 | 0.02 | -0.01 | 0.02 |
| Tuberculosis | 0.00 | 0.01 | -0.00 | 0.00 | 0.05 | -0.00 | 0.02 | 1.00 | -0.01 | 0.00 |
| Sms_Reminder | -0.05 | -0.03 | 0.00 | -0.03 | -0.02 | -0.04 | -0.01 | -0.01 | 1.00 | -0.00 |
| Adherence | -0.10 | 0.01 | -0.61 | -0.03 | 0.02 | -0.06 | 0.02 | 0.00 | -0.00 | 1.00 |

Following deductions can be made from the correlation matrix:

1. Prescription_period and Adherence are inversely proportional. Smaller the prescription period, the higher the chances of adherence
2. Age is directly proportional to hypertension and diabetes
3. Hypertension and diabetes are also highly correlated
4. Smoking and alcoholism are highly correlated
5. Adherence is not correlated to any of the factors heavily

## Tentative approach:

Based on the EDA and correlation matrix, the following approach was decided:

1. First, train some traditional machine learning models on the unbalanced datasets to get a baseline of accuracy
2. Balance the dataset for the “Adherence” feature and perform feature engineering. Train the above traditional machine learning models again

Reasons for choosing this approach:

1. Since the distribution of train and test datasets are similar, it can be expected that the accuracy of models on training data will be similar to the accuracy of the models on the test dataset
2. Since the adherence counts are not balanced, there is a chance that the model might overfit adherence = no. Hence, balancing the dataset might give the unbiased model

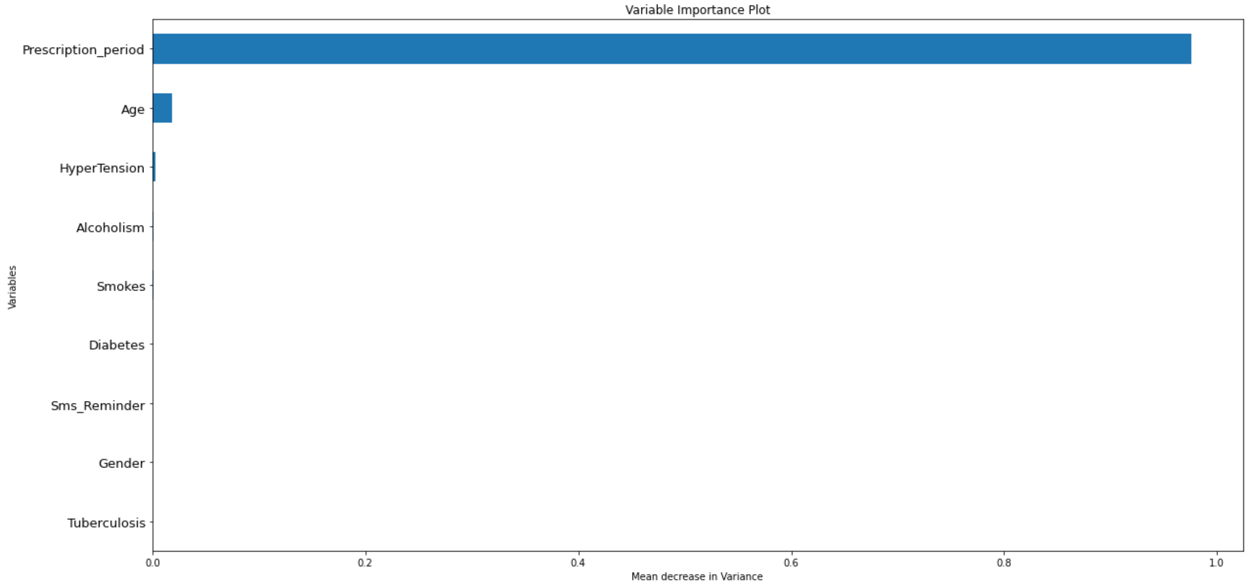
## Traditional ML models on provided dataset

ML models of logistic regression, decision trees and random forest were run on the provided dataset to get baseline scores

## Observations:

1. While training the model on an unbalanced dataset, it was found that prescription period, age and hypertension were the only features that were important in determining the output. All other features were more or less not detrimental in deciding the adherence of

any patient

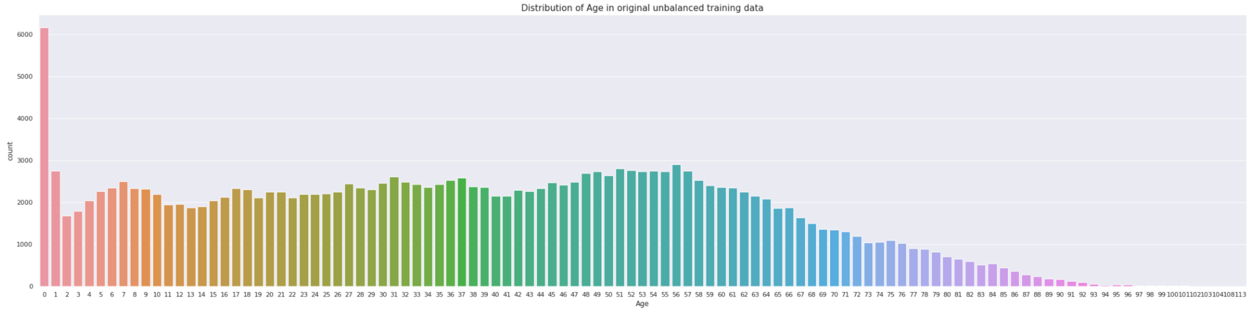


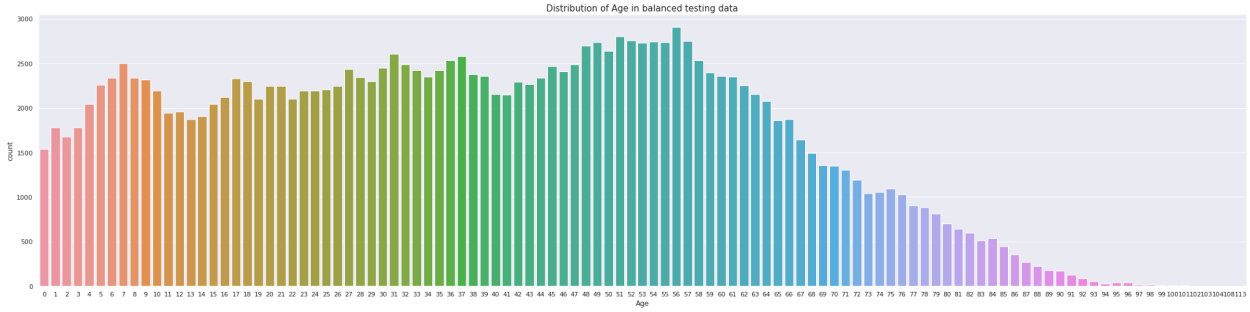
2. All the traditional machine learning models were providing accuracy in the range of 85-89.5% on the training and test dataset
3. Apart from the Regression models present in the jupyter notebook, I also trained decision trees classifiers and Random Forest Classifiers. However, they were also giving accuracy in the same range of 85% - 89.5%. These codes and outputs are present in the jupyter notebooks ‘Extra1’ and ‘Extra2’
4. Cross validation, grid search and weighted class approach did not help in improving output accuracy

## Balancing the dataset

Balancing of the dataset was performed wrt age column only. I have reduced the number of age = 0 and age = 1 to match the distribution of age using undersampling as it is easier to reduce just one age group

Balancing can also be done for the adherence feature as it has 30-70 split. However, I chose not to do that due to the fact that assigning class weights was not affecting the model training.





# Feature engineering

For feature engineering, I created bins for age and prescription period
This was done to bring long tail occurrences into groups

For feature engineering, I decided to perform frequency encoding on age and prescription period and also perform one hot encoding for all the features

# Final model:

For the final model, 7 models were trained:

1. Logistic regression on original dataset with Feature engineering
2. Decision Trees Regressor on original dataset with Feature engineering
3. Random Forest Regressor on original dataset with Feature engineering
4. Logistic regression on age balanced dataset with Feature engineering
5. Decision Trees Regressor on age balanced dataset with Feature engineering
6. Random Forest Regressor on age balanced dataset with Feature engineering
7. Gradient boosting Regressor model was trained on the output of above 6 models to provide the final output

If I had more time, I would have tried the following techniques:

1. Balance the dataset for “adherence” column using either undersampling or oversampling
2. Training more variety of ML models
3. Training a Neural Network
4. Performing creative feature engineering