**Problem description and business understanding**

When patients are prescribed treatment by a health care official, it is expected that the patient adheres to the prescribed treatment regimen. Medication persistence refers to the act of continuing the treatment for the prescribed duration. It may be defined as the duration of time from initiation to discontinuation of therapy [1]. In adequate medication persistence is an age old problem which can have situation-specific alterations in benefit/risk ratios, either because of reduced benefits, increased risks, or both [1].

Though not of interest in this particular case it should be known that Medical compliance is also a problem. Medication compliance (also known as adherence) refers to the degree or extent of conformity to the recommendations about day-to-day treatment by the provider with respect to the timing, dosage, and frequency. It is described as the extent to which a patient acts in accordance with the prescribed interval, and dose of a dosing regimen [1].

Studies have shown that inadequate compliance and non-persistence with prescribed medication regimens result in increased morbidity and mortality from a wide variety of illnesses, as well as increased health-care costs in particular cases [2,3]. Clinical outcomes of treatment depend not only by how well patients take their medications but also by how long they take their medications.

There are many factors which affect a patients persistence such as choice of agent prescribed, comorbidity, and socioeconomic status, despite universal coverage of prescription drug costs [3]. It is important for the sake of improving clinical outcomes to be able to know which patients will less likely to stick with the medication persistence requirements and which ones will not.

Machine learning algorithms are particularly useful for such problems [4]. Our objective is to apply machine learning to predict patient medical persistence with the aim of using this information to identify patients who are least likely to adhere to medical persistence [4]. This information may be useful in helping patients.

Nontuberculous mycobacterial (NTM) lung disease is a general term which refers to a group of disorders characterized by exposure to specific bacterial germs known as mycobacteria [5]. These germs are found in the water and soil and are common throughout the environment as a whole. They usually do not cause illness. In NTM disorders, severity of infection and the disease course can vary greatly from one person to another. The most common symptoms include a persistent cough, fatigue, weight loss, night sweats, and occasionally shortness of breath (dyspnea) and coughing up of blood (hemoptysis) [5].

We will examine a dataset of patients who have NTM. There is a wide range of features which we will use to predict medical persistence in patients. These include but are not limited to patient demographics, provider attributes, clinical factors and disease factors.

[1] J. Cramer, A. Roy, A. Burrell, C. Fairchild, M. Fuldeore, D. Ollendorf and P. Wong, Medication compliance and persistence: Terminology and definitions, *Value in health*, 2, 1(2008).

[2] M. DiMatteo, P. Giordani, H. Lepper and T. Croghan, Patient adherence and medical treatment outcomes: a meta-analysis, *Med Care*, 40, 794–811 (2002).

[3] J. Avorn , J. Monette, A. Lacour A, R. Bohn, M. Monane, H. Mogun and J. lolerier, Persistence of use of lipid-lowering medications: cross-national study, *JAMA*, 279, 1458–62 (1998).

[4] X. Wu,H. Yang, R. Yaun, E. Long and R. Tong, Predictive models of medication non-adherence risks of patients with T2D based on multiple machine learning algorithms, BJM Open Diabetes and Care, 8, 1 (2020).

[5] <https://rarediseases.org/rare-diseases/nontuberculous-mycobacterial-lung-disease/>

**Project lifecycle along with deadline**

### Project description and lifecycle (-5/09/2021)

### Data cleaning (06/09/2021-13/09/2021)

Data understanding

What type of data you have got for analysis

What are the problems in the data ( number of NA values, outliers , skewed etc)

What approaches you are trying to apply on your data set to overcome problems like NA value, outlier etc and why?

### Data cleaning (14/09/2021-19/09/2021)

Data cleansing and transformation done on the data.

Try at least 2 techniques to clean the data ( for NA values : mean/median/mode/Model based approach to handle NA value/WOE and like this try different techniques to identify and handle outliers as well)

Each team member should work on different data cleansing approach.

**Note:**

If one team member is using mean to impute values then other member should experiment on segmented approach or any other model based approach to impute the null values.

### Exploratory data analysis on data (20/09/2021-26/09/2021)

EDA performed on the data

Final Recommendation

### Exploratory data analysis on data (27/09/2021-3/10/2021)

EDA presentation for business users

Last slide of EDA should be dedicated to technical user which should contain recommended models for this data set.

### Model Selection and model building (04/10/2021-11/10/2021)

### Select your base model and then explore 1 model of each family if its classification problem then 1 model for Linear models, 1- Model for Ensemble, 1-Model for boosting and other models if you have time (like stacking)

### Final Project (12/10/2021-15/10/2021)

Power point presentation is **must**.