Data Science Healthcare Project: Assessment of Patient Persistency on Drugs based on Clinical Factors

Table of Contents

Details of Team Members

Group Name: Health+	
Member 1: Sakib Mahmud	**Student Email:** sm1512633@qu.edu.qa **Personal Email:** sakib1263@hotmail.com **Official Email:** sakib.mahmud@qu.edu.qa
	Country: Qatar
	College + Company: Qatar University
	Specialization: Data Science
Member 2: Mohammad Odeh	**Personal Email:** odeh4893@gmail.com
	Country: United Arab Emirates (UAE)
	Specialization: Data Science

Problem Description and Business Understanding

Among the most critical challenges that pharmaceutical companies face in general, the most common one is the task of understanding the "Persistency of a Drug" as per the physician's prescription. To solve this problem "ABC Pharma" company approached an analytics company to automate this process of identification. ABC Pharma provided the company with their recorded data in an Excel file for analysis. The dataset contains four types of predictor parameters (or independent variables) for unique patients (each patient had a unique identifier or ID) such as Patient Demographics, Provider (Doctor/Nurse/Other Medical Staff) attributes, Clinical Factors, and Disease or Treatment Factors. On the other hand, the target or dependent variable for the dataset is the "Persistency Flag" for the patient i.e., whether the patients were persistent on their prescribed medicine(s) or not.

Among the patient demographics parameters, there were patients' age, race, region, etc. Attributes of the physician who prepared the prescription or performed the task of observing the patient might be an important predictor, so it was included. The primary disease for which patients were treated in this case is Nontuberculous Mycobacterial (NTM). Various tests such as DEXA Scans are performed for NTM which produces metrics like T-Score. Clinical Factors like the outcomes of these tests during the Rx and the performance shift during the last 1 to 2 years were also accounted for, along with the Risk Segment of the patient and the possibility of prevalence of multi-risk among the patients. Other treatment factors such as comorbidity of patients for other diseases alongside NTM, Injectable Experience, and concomitancy of various drugs applied on the patient for NTM were also accounted for. All these parameters will be used to produce models using Machine Learning to correctly classify patients based on their "Persistency Flag". Efforts will also need to be given to determine the most influential parameters (or class of parameter) for determining peoples' choice on continuing the medicine.

More than the healthcare perspective of this problem, it has an even more important business perspective. As discussed earlier, one of the challenges for all pharmaceutical companies is to understand the persistency of drugs as per the physician's prescription. The general trend of persistency of pharmaceutical products among a group of patients is downward, as depicted by the "Persistency Curve" in Figure 1. The pharmaceutical companies aim to determine the factors which affect the decline most so that they can address these issues properly and slow down the process (i.e., the smaller slope of the model). Pharmaceutical companies, healthcare organizations, and hospitals in the USA lose billions of dollars per year due to patients not being persistent in their prescribed medicines and/or treatments [1].

Based on the collected data by any pharmaceutical company or an Integrated Delivery Network (IDN), the most important factors behind the lower level of persistence among patients can be determined. The company or organization can address those issues properly to mitigate the process and over time they can resurvey to check for improvements.

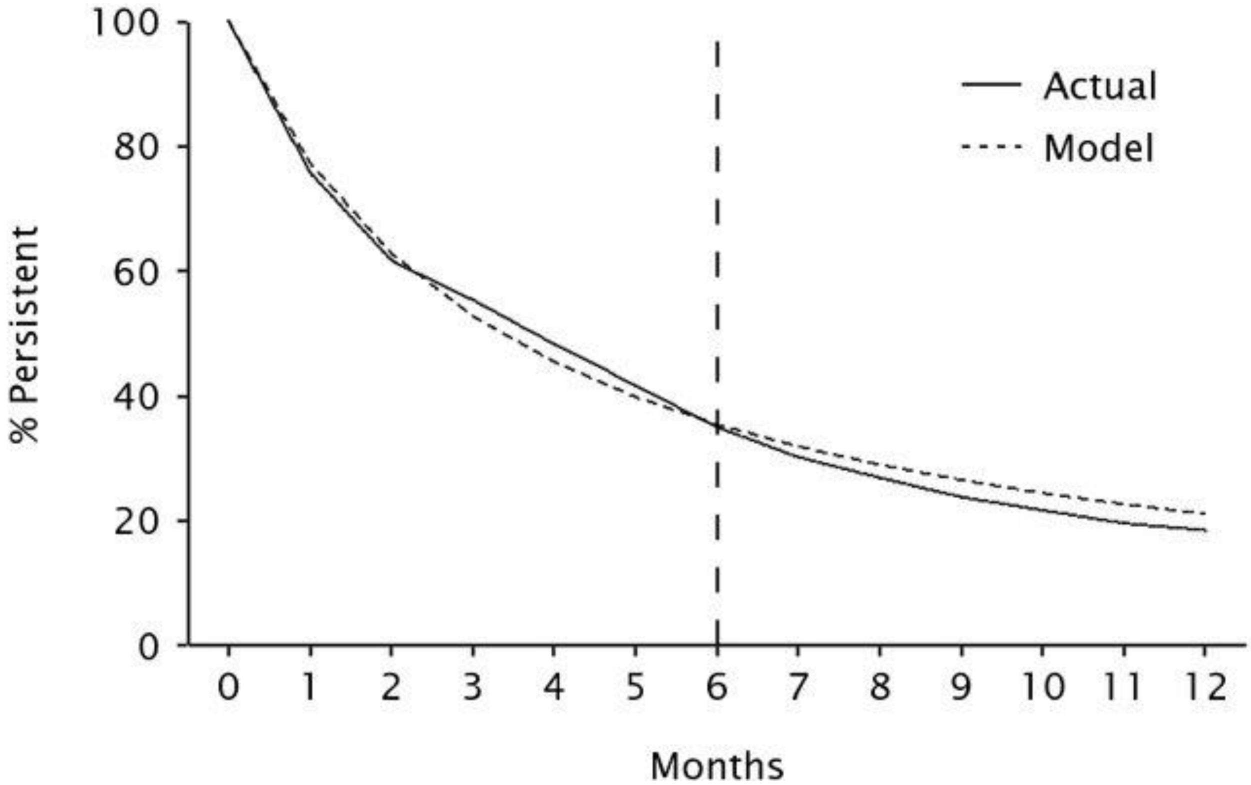

Figure 1: Persistency Curve [2]

From the Persistency Curve in Figure 1, it can be understood that the general trend of the Persistency Curve is downward i.e., patients generally do not adhere to the medicine or the set of medicines prescribed by their doctors. They either stop taking the drug for many reasons or swap to another. Since the overall trend is always downward, the important thing to focus on is how to slow down the rate. The main key behind slowing down this rate is to determine the most influential factor(s) behind so that they can be assessed in time and important business decisions can be taken which could save millions, sometimes even millions for the company or the government. To make people adhere to a certain prescription or guidelines for a long time is not a trivial task and IDNs across the USA were formed primarily due to this.

Dataset Understanding: Descriptive Analysis on the Dataset

The Dataset can be briefly described based on Table 3 below as it was included in the dataset. From this table, it can be understood that there mainly four types of predictor variables for this dataset, the "Target Variable" being the "Persistency Flag". These types are "Demographics", "Provided Attributes", "Clinical Factors" and "Disease/Treatment Factor". All these factors will be analyzed based on inter and intraclass mean and variances.

Bucket	Variable	Variable Description
Unique Row Id	Patient ID	Unique ID of each patient
Target Variable	Persistency Flag	Flag indicating if a patient was persistent or not
Demographics	Age	Age of the patient during their therapy
	Race	Race of the patient from the patient table
	Region	Region of the patient from the patient table
	Ethnicity	Ethnicity of the patient from the patient table
	Gender	Gender of the patient from the patient table
	IDN Indicator	Flag indicating patients mapped to IDN
Provider Attributes	NTM - Physician Specialty	The specialty of the HCP that prescribed the NTM Rx
Clinical Factors	NTM - T-Score	T Score of the patient at the time of the NTM Rx (within 2 years prior from rxdate)
	Change in T Score	Change in T-score before starting with any therapy and after receiving therapy (Worsened, Remained Same, Improved, Unknown)
	NTM - Risk Segment	Risk Segment of the patient at the time of the NTM Rx (within 2 years days prior from rxdate)
	Change in Risk Segment	Change in Risk Segment before starting with any therapy and after receiving therapy (Worsened, Remained Same, Improved, Unknown)
	NTM - Multiple Risk Factors	Flag indicating if a patient falls under multiple risk category (having more than 1 risk) at the time of the NTM Rx (within 365 days prior from rxdate)
	NTM - Dexa Scan Frequency	Number of DEXA scans taken before the first NTM Rx date (within 365 days prior from rxdate)
	NTM - Dexa Scan Recency	Flag indicating the presence of Dexa Scan before the NTM Rx (within 2 years prior from rxdate or between their first Rx and Switched Rx; whichever is smaller and applicable)
	Dexa During Therapy	Flag indicating if the patient had a Dexa Scan during their first continuous therapy
	NTM - Fragility Fracture Recency	Flag indicating if the patient had a recent fragility fracture (within 365 days prior from rxdate)
	Fragility Fracture During Therapy	Flag indicating if the patient had fragility fracture during their first continuous therapy
	NTM - Glucocorticoid Recency	Flag indicating usage of Glucocorticoids (>=7.5mg strength) in the one-year look-back from the first NTM Rx
	Glucocorticoid Usage During Therapy	Flag indicating if the patient had a Glucocorticoid usage during the first continuous therapy
Disease/Treatment Factors	NTM - Injectable Experience	Flag indicating any injectable drug usage in the recent 12 months before the NTM OP Rx
	NTM - Risk Factors	Risk Factors that the patient is falling into. For chronic Risk Factors complete lookback to be applied and for non-chronic Risk Factors, one-year lookback from the date of the first OP Rx
	NTM - Comorbidity	Comorbidities are divided into two main categories - Acute and chronic, based on the ICD codes. For chronic disease, we are taking a complete look back from the first Rx date of NTM therapy and for acute diseases, a period before the NTM OP Rx with one-year lookback has been applied
	NTM - Concomitancy	Concomitant drugs recorded before starting with a therapy(within 365 days before the first rxdate)
	Adherence	Adherence to the therapies

Demographics – Imbalance or Bias in the Dataset

Collected Subject data based on Gender, Race, Age, Region, Ethnicity, and IDN Indicator were assessed in MS Excel using Pivot Table and visualized using Pivot Charts as shown in Figure 2. The extract from the descriptive analysis is that the dataset has some imbalance in terms of demography, from some aspects. For example, the dataset has around 94% female patients compared to only 6% male which shows that the dataset is biased towards females. It is a matter of research that whether the dataset is focused on a certain disease that occurs most to females or females of certain age-class are the ones to reach for medications in large numbers. Nevertheless, due to this class imbalance, any outcome from the dataset will be more representative for female patients.

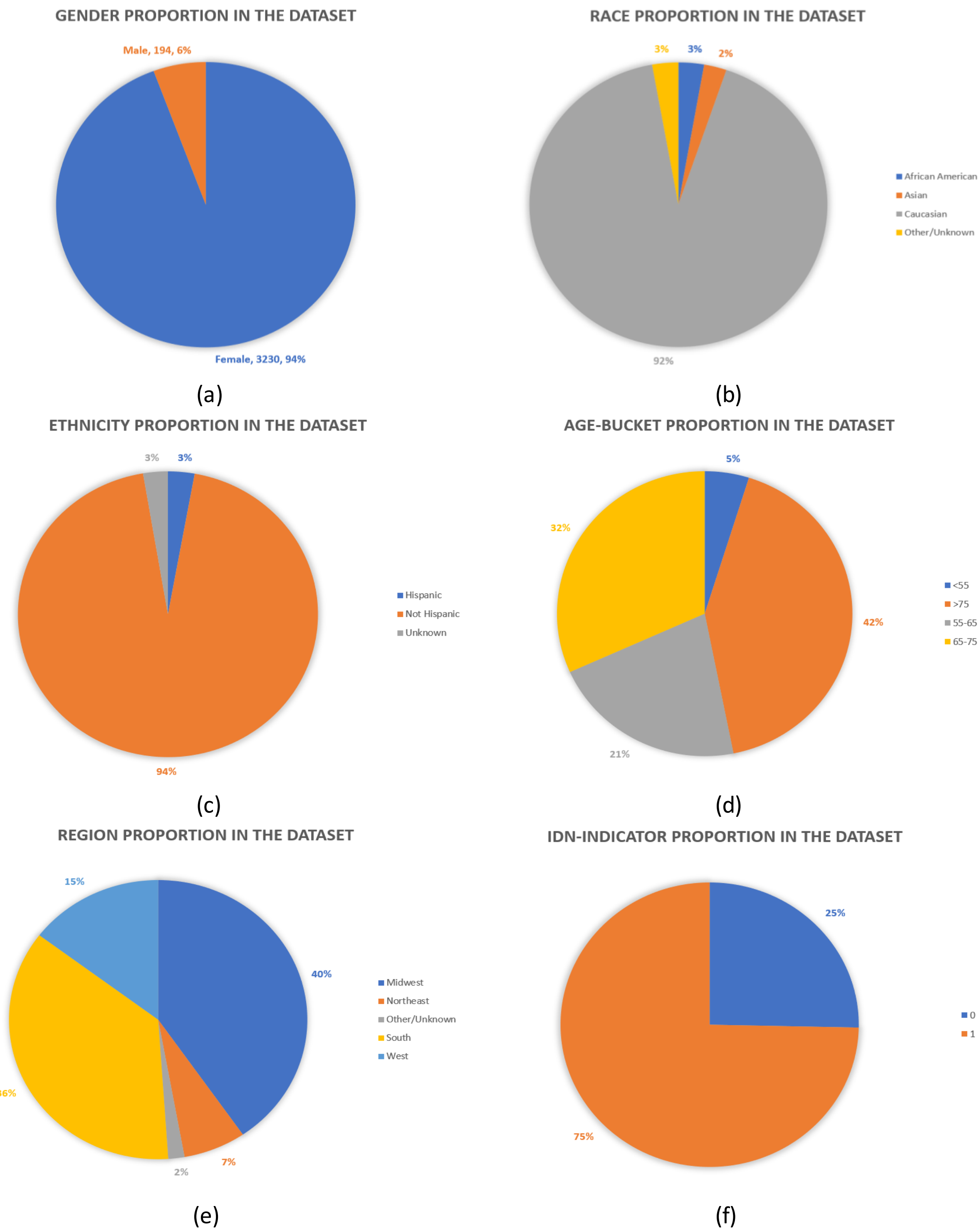

Figure 2: Visualization of the Descriptive Analysis on Demographics of the Subjects

On the other hand, the dataset was dominated by non-Hispanic, Caucasian patients in large proportions. Even though the Region and Age classes were more proportionate, it is clear that most of the patients belonged to higher age classes (only a few lower than 55) and the greatest number of patients came from the Midwest region. One of the most clinically important factors in this analysis was the high proportion of the subjects belonging to the IDN. Around 75% of the subjects belonged to a certain IDN implying the success of forming clusters of health service providers for better patient experience.

Physician Specialty Type and Specialist Flag for the Observing Physician

In this section, the focus was given to the practitioners involved in the cases based on their specialty type and specialist flag (expert on their respective departments).

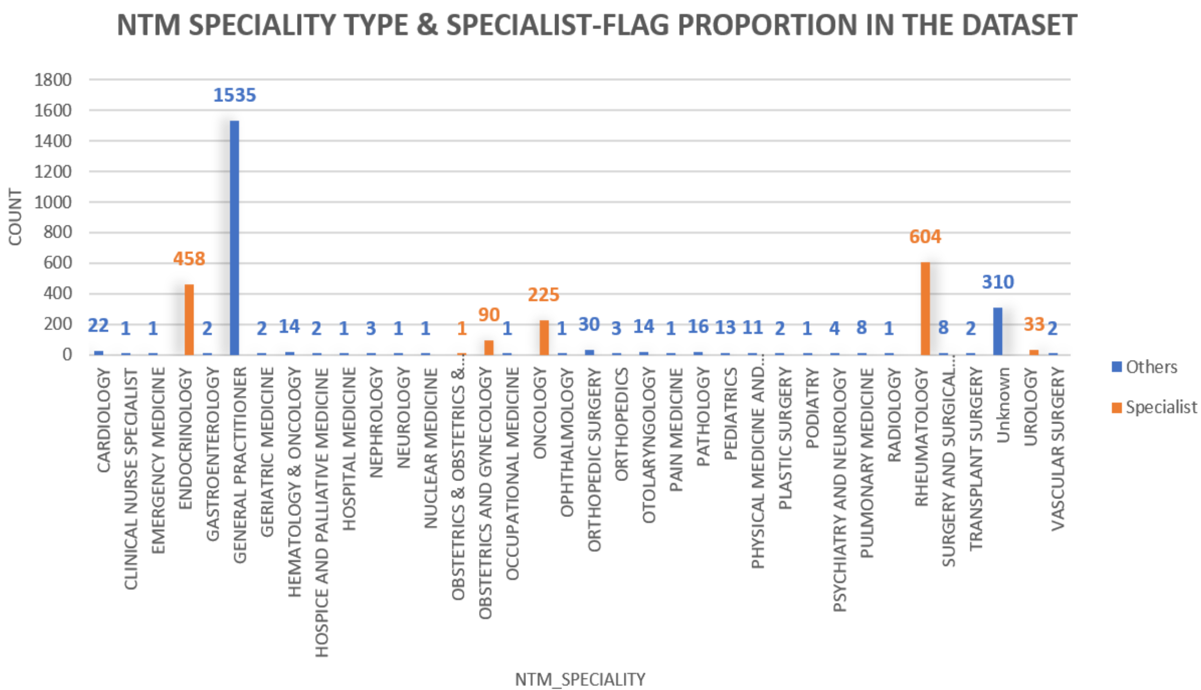

Figure 3: Specialty and Specialist-Flag of the Observing Medical Staff or Physician for the Patient

It can be observed that a large number of physicians who handled the NTM cases were general practitioners, around 45% of the total. But among the other groups, all were not specialists. As the specialist flags indicate, only the physicians belonging to "Endocrinology", "Obstetrics and Gynecology", "Rheumatology" and "Urology" were specialists in those respective fields. It might indicate an important assumption that NTM is more critical for these categories of patients, so specialists had to be involved.

Clinical Factors

The clinical factor can be divided into few major classes such as Comorbidity of Diseases, Concomitancy of Drugs, and Risk Factors among the subjects. Percentage-based column charts are plotted for each sub-category of them and shown in Figures 4-6 respectively. The concomitancy of some drugs is around 35% while other drugs can be as low as 10% of subjects. In this case, the Cholesterol mitigating drug was most commonly used by the subjects alongside the main drug. On the other hand, the comorbidity parameter varies a lot as well-meaning that some diseases are comorbid with NTM than other ones. The comorbidity of Lipoprotein Disorder is around 51%, which is the highest, meaning that these diseases occur commonly together with the main disease.

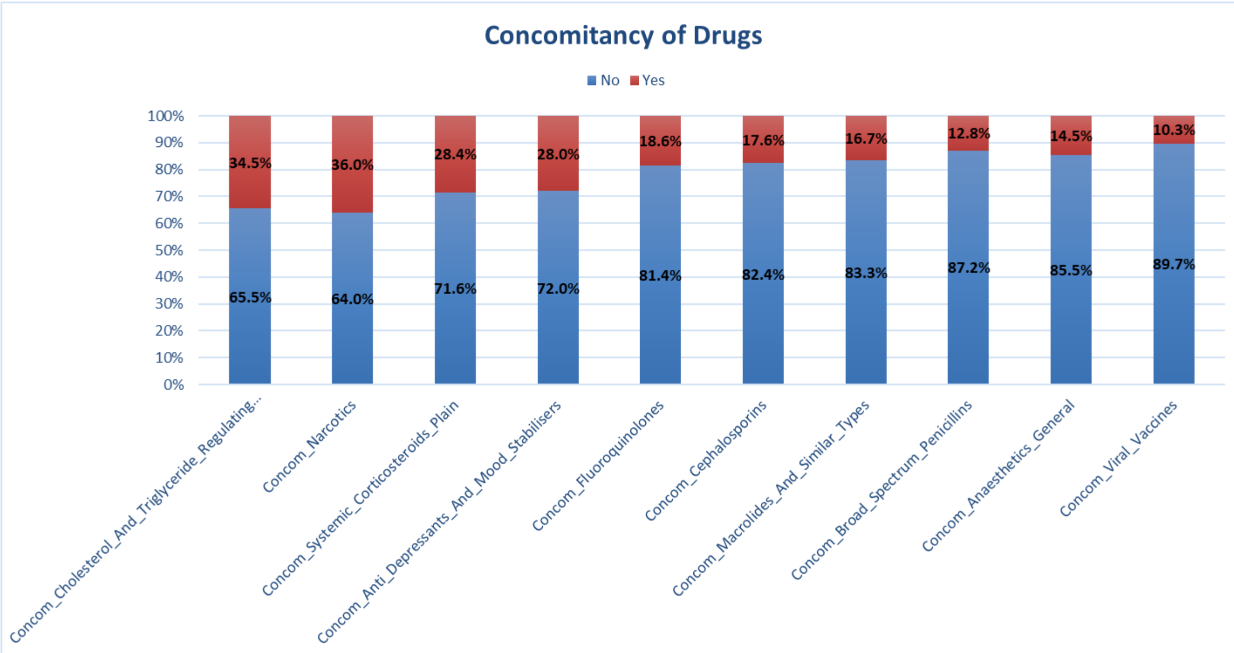

Figure 4: Concomitancy of Various Drugs among Subjects

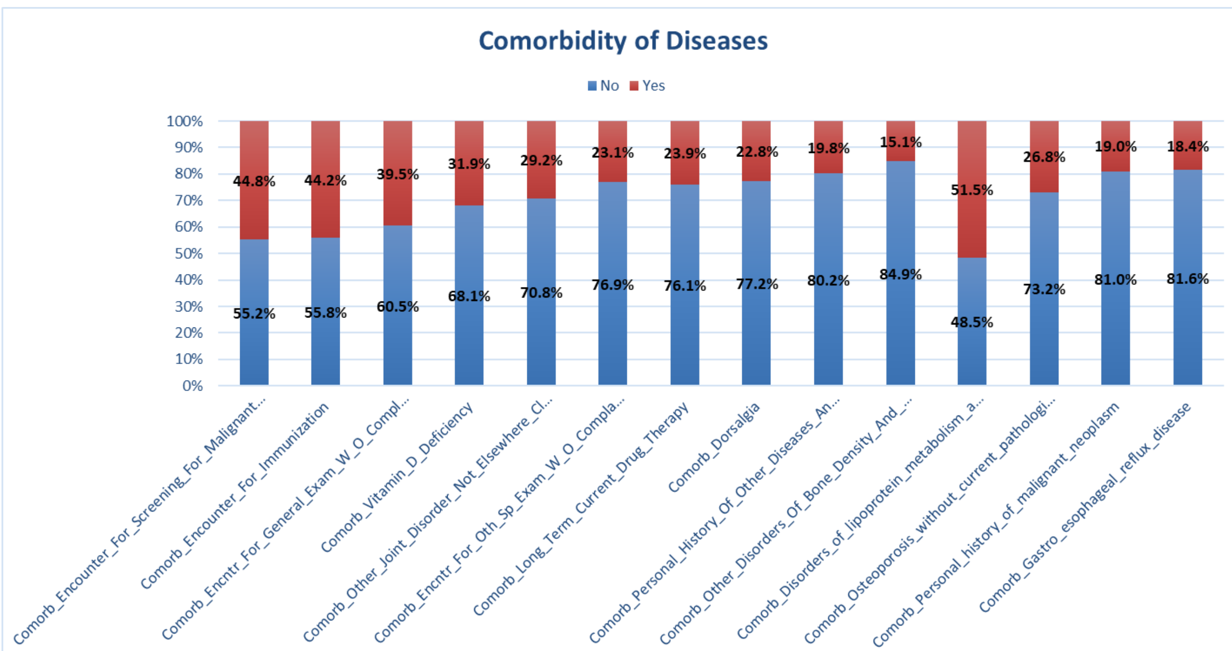

Figure 5: Comorbidity of Various Diseases among Subjects

Various risk factors were observed for the subjects under the study. Among the risk factors, the risk of Vitamin D insufficient was the most acute while the risk of smoking tobacco and the risk of chronic malnutrition were also other influential factors. Nevertheless, the strength of the relations between various clinical factors can be understood from the Machine Learning and Dominance Analysis discussed in the next sections.

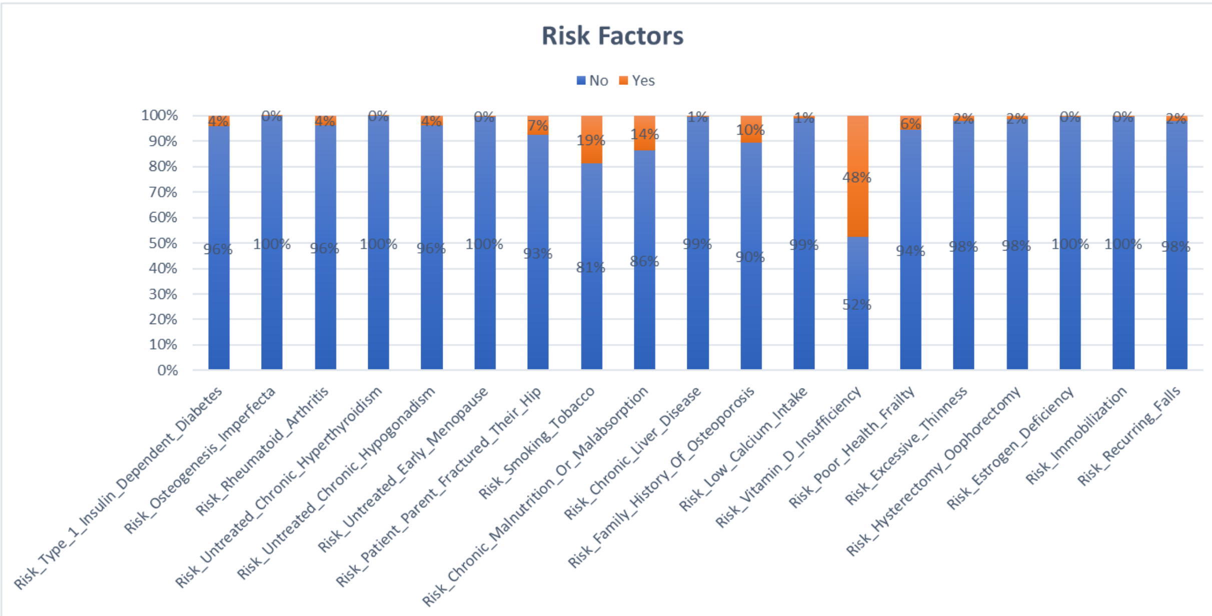

Figure 6: Risk Factors among Subjects

Analysis of Issues in the Dataset

This section briefly discusses the existence of any problem in the dataset such as Null values, Outliers, and Skewness, and the tests performed on the dataset to detect these issues.

Presence of Null Values and Outliers in the Dataset

Using Python PANDAS library's Null value detection commands, the number of null values present in each column and the entire dataset was tested. No Null value could be detected in the dataset proving it to be clean in this aspect. The python code output is provided in Figure 7. Null values are important to be detected and removed or replaced from the dataset since they provide errors while performing Machine Learning and other tasks. Even if they do not show any error during any analysis, output from them is not important or relevant to any analysis. Almost all the variables, independent or dependent, in the dataset were categorical and for this reason, there was no presence of any outlier or abnormal data in the dataset. On the other hand, the dataset did have some imbalance and skewness for some predictors which is discussed in detail in the next section.

```
Total NULL values in the Original DataFrame = 0
Female                     0
Male                       0
African American           0
Asian                      0
Caucasian                  0
                          ..
Risk_Estrogen_Deficiency   0
Risk_Immobilization        0
Risk_Recurring_Falls       0
Count_Of_Risks             0
Persistency_Flag           0
Length: 119, dtype: int64
```

Figure 7: Testing Presence and Number of Null Values in the Dataset

Skewness and Kurtosis in the Dataset

Using Excel's Data Analysis toolbox, various important statistical parameters for the dataset variables were calculated, as shown in Figure 8-10. From Figure 4, the dataset had similar errors for all the variables, no extremity could be noticed while from Figure 5, the demographic parameters were least skewed (closer to being Normal) even though some parameters like Gender were imbalanced. Most skewed were the Risk parameters. Similar observation for the Kurtosis (Figure 9) as it is directly related to Skewness.

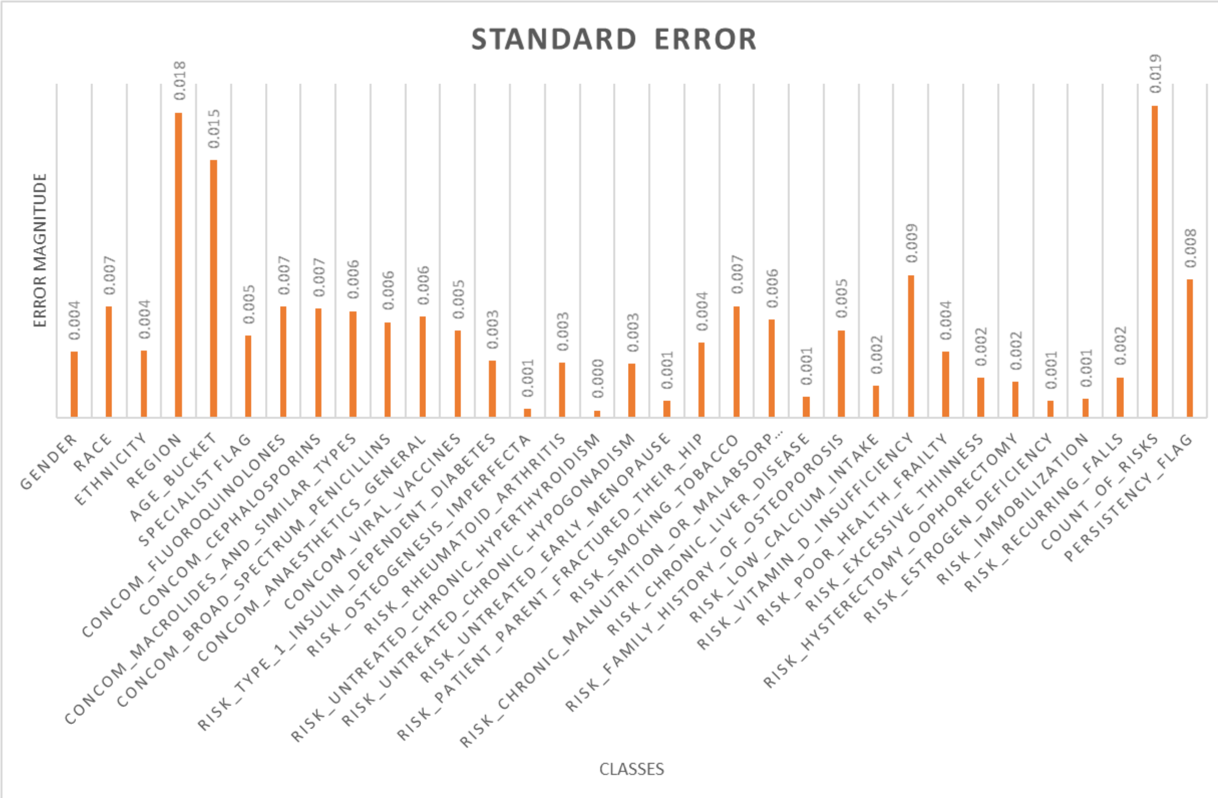

Figure 8: Standard Errors of the Dataset Variables

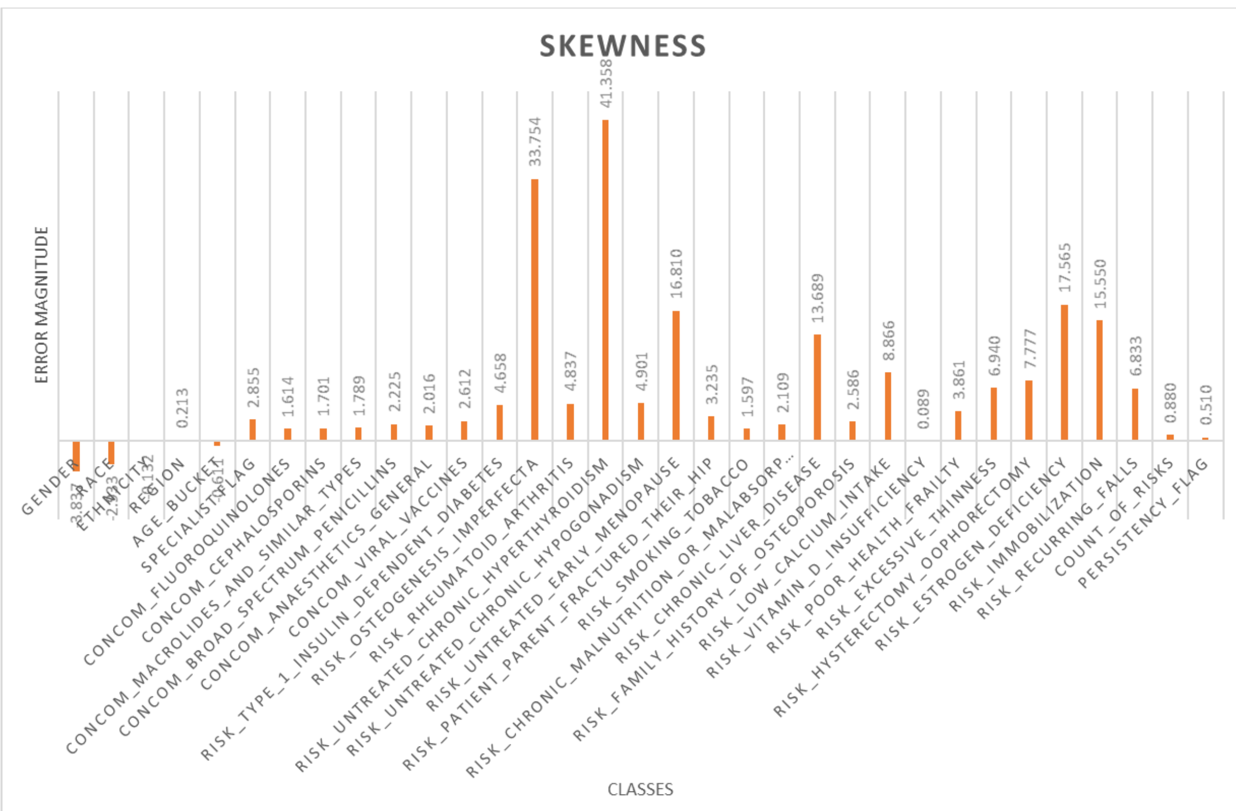

Figure 9: Skewness of the Dataset Variables

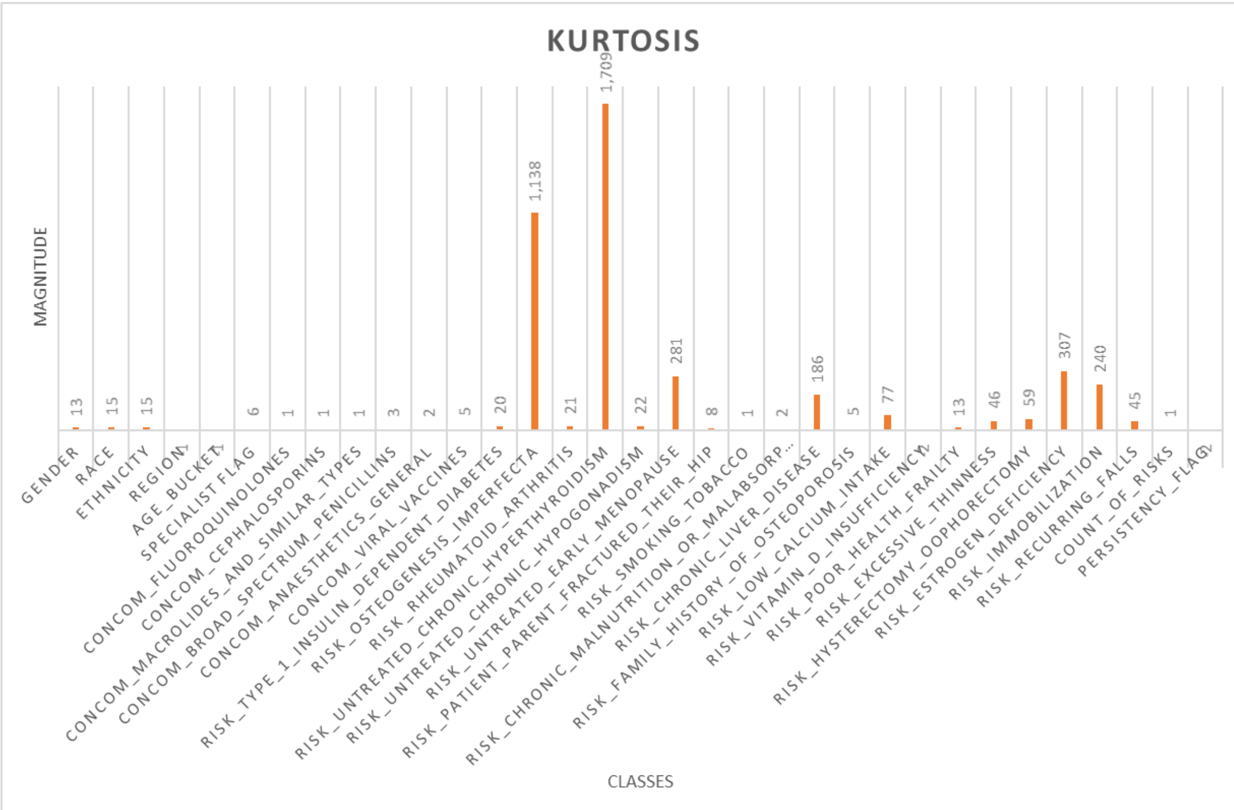

Figure 10: Kurtosis of the Dataset Variables

Approaches to Overcome Problems

As discussed earlier, among 68 parameters, only 2 are numerical, the rest are categorical in the dataset. To convert the data to numerical, we have to apply encoding techniques to it, so it can be translated into numerical labels instead namely Dummy Variables (Example in Figure 11).

Turn your Categorical Column (Ex: "Name")...

Index	Name	8/6/2020
0	Liho Liho	$234.54
1	Chambers	$45.74
2	The Square	$56.22
3	Liho Liho	$32.31

...Into Dummy Indicator Columns

Index	Liho Liho	Chambers	The Square	8/6/2020
0	1	0	0	$234.54
1	0	1	0	$45.74
2	0	0	1	$56.22
3	1	0	0	$32.31

Figure 11: Encoding Techniques to be Applied on the Dataset

GitHub Repo Link

https://github.com/m-odeh/Persistency-of-a-drug

References

[1] https://decisionresourcesgroup.com/solutions/integrated-delivery-networks-and-their-growing-influence-on-regional-healthcare-in-the-us/

[2] https://www.researchgate.net/publication/5055576_How_to_Project_Patient_Persistency