Data Science Project

Healthcare - Persistency of a drug

Group Name: BetterHealth Analytics	
Member Details	Name: Enias Vontas Country: Greece Email: vondas100@gmail.com Specialization: Data Science

Problem Description

One of the challenges for all Pharmaceutical companies is to understand the persistence 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. We have been provided with an Excel file containing some of the company's recorded data. The particular affliction that patients in this dataset were treated for is Nontuberculous Mycobacterial (NTM), which originates from a family of common organisms found in water and soil. This type of infection is rare and can affect people with damaged lungs, or with a weakened immune system. If diagnosed, a patient might need up to two years of treatment and could get infected again in the future.

The dataset contains the target variable 'Persistency_Flag' which indicates whether a patient was persistent with their medication or not. We would like to better understand the factors affecting this variable (our dependent variable). In order to do this, we have been provided with 67 other variables (our independent variables) which can be grouped in four buckets:

- Demographics: with variables such as Age, Race, Gender, etc for each patient.
- Provider Attributes: some information about the provider that wrote the prescription to the patient, with variables such as the Specialty of the Physician, a T-Score which is the result of a scan done to patients of this disease, etc.
- Clinical Factors: certain physiological attributes which could be associated with the disease, with variables such as Usage of Glucocorticoids, Frequency of a Dexa Scan etc.
- Disease/Treatment: Comorbidity factor, divided into two categories – Acute and Chronic and Concomitancy factor, i.e. concomitant drugs recorded prior to starting with the therapy.

All of the above parameters will be considered in our Machine Learning approach in order to better understand the factors affecting a patient's Medication Persistence and to more accurately classify a patient to one of the two categories of our 'Persistency_Flag' variable.

Business Understanding

It can be clear to imagine why a patient not receiving the whole dosage regimen that was prescribed to them can have unwanted results toward the treatment of that patient's illness. Another important unwanted result from this scenario is all the prescribed medication that goes to waste, from manufacturing it, all the way to distributing it to local pharmacies. So it is very important for drug companies and healthcare systems to provide the required medication to patients, but also as important for patients to be consistent with that prescribed medication, otherwise all that drug availability and expenditure would have been for nothing.

Before we talk more about our project from a business understanding, we would like to offer two definitions for a better overall understanding [**1**].

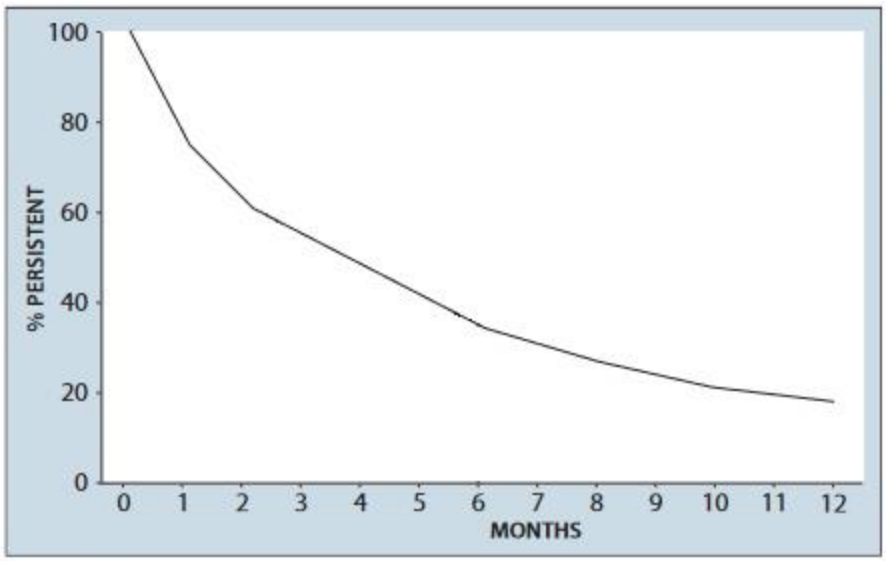
Medication Compliance (Adherence): refers to the degree of extent of conformity to the recommendation about day-to-day treatment by the provider with respect to timing, dosage, and frequency, or the extent to which a patient acts in accordance with the prescribed interval, and dose of a dosing regimen.

Medication persistence: refers to the act of continuing the treatment for the prescribed duration, or the duration of time from initiation to discontinuation of therapy.

Inadequate medication compliance and persistence are age-old problems in the pharmaceutical business. When taken in varying degrees of deviation from the prescribed dosing regimen, medications have situation-specific alterations in benefit-risk ratios, either because of reduced benefits, increased risks, or both. Numerous studies have demonstrated 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 healthcare costs. Factoring in actual compliance and persistence is central to an accurate assessment of effectiveness and cost-effectiveness of therapy.

This source of wasted US healthcare spending every year has the potential to reach even $300 billion, while also affecting pharmaceutical companies [**2**][**3**]. A lot of factors could contribute to a patient stopping, or altering their medication regimen, they could be a physician's time constraint, competing priorities for patients and shortcoming in follow-up initiatives. These factors need to be determined by healthcare providers, as well as pharmaceutical companies in order to address them and control them as much as possible.

It is known that the drug persistence curve has a downward trend and it tends to decrease at a decreasing rate as can be seen in the figure below, where we consider the drug persistence as a percentage, and observe in the duration of a year [4]:



We would like to determine the factors affecting a patient's persistence to their prescribed medication so that companies and doctors could then control for those factors when prescribing medication.

Project Lifecycle and Deadline

The project is due on 15th of August. It has been broken into various sections,which will be completed consecutively, as presented below:

- Problem Understanding
- Business Understanding
- Data Understanding
- Data Cleaning and Feature Engineering
- Model Development
- Model Selection
- Model Evaluation
- Report the accuracy, precision and recall of both the classes of target variable
- Report ROC-AUC
- Deploy the model
- Explain Challenges and Model Selection

As a first section, we will focus on the first two points, that are also underlined. As the project progresses, we will move forward with the other sections, as well as re-evaluate our findings if needed.

Data Understanding

The dataset provided to us contains 3424 patients (each with their own ID), our target variable, the Persistency of the drug, and 67 other variables, which we will use as predictors for our target/dependent variable. The predictor variables can be grouped in four different buckets: 'Demographics', 'Provider Attributes', 'Clinical Factors' and 'Disease/Treatment Factors'. Almost all our predictor variables are categorical and ordinal with the exception of two ('Dexa Scan Frequency' and 'Count of Risks') which are numerical. Our ordinal variables ('Age_Bucket' and T-Scores) are 4 in total and the rest are categorical, either with two categories ('Yes' or 'No') or with more than two categories ('No Change', 'Unkown', 'Worsened', 'Improved').

We provide a table below with all the variables and a brief description for each one:

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 Integrated Deliver Network
Provider Attributes	*NTM - Physician Specialty*	Specialty of the Health Care Personnel 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 Tscore before starting with any therapy and after receiving therapy
	NTM - Risk Segment	Risk Segment of the patient at the time of the NTM Rx (within 2 years prior to rxdate)
	Change in Risk Segment	Change in Risk Segment before starting any therapy and after receiving therapy
	NTM - Multiple Risk Factors	Flag indicating if patient falls under multiple risk category at the time of the NTM Rx (within 365 days prior to rxdate)
	NTM - Dexa Scan Frequency	Number of DEXA scans taken prior to the first NTM Rx(within 365 days prior to 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 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 complete look back from the first Rx date of NTM therapy and for acute diseases, time period before the NTM OP Rx with one year lookback has been applied
	NTM - Concomitancy	Concomitant drugs recorded prior to starting with a therapy(within 365 days prior from first rxdate)
	Adherence	Adherence for the therapies

It is very important to note, that since we have been assigned a Classification Machine Learning problem, we will split our dataset into a 'Train' and a 'Test' set with 80% and 20% of the patients respectively. This is being done in order to 'Train' our model first and then evaluate its performance on the 'Test' set, which we consider as unknown at this point, so as not to 'contaminate' our model building process. There are different schools of thought as to when this split should be done, but **from this point forward** all analysis is being done on the 'Train' set, unless specified otherwise.

As far as any possbible NA or missing values are concerned, we did not find any:

```
Data columns (total 69 columns):
 #   Column                                                  Non-Null Count  Dtype
---  ------                                                  --------------  -----
 0   Ptid                                                    2739 non-null   object
 1   Persistency_Flag                                        2739 non-null   object
 2   Gender                                                  2739 non-null   object
 3   Race                                                    2739 non-null   object
 4   Ethnicity                                               2739 non-null   object
 5   Region                                                  2739 non-null   object
 6   Age_Bucket                                              2739 non-null   object
 7   Ntm_Speciality                                          2739 non-null   object
 8   Ntm_Specialist_Flag                                     2739 non-null   object
 9   Ntm_Speciality_Bucket                                   2739 non-null   object
 10  Gluco_Record_Prior_Ntm                                  2739 non-null   object
 11  Gluco_Record_During_Rx                                  2739 non-null   object
 12  Dexa_Freq_During_Rx                                     2739 non-null   int64
 13  Dexa_During_Rx                                          2739 non-null   object
 14  Frag_Frac_Prior_Ntm                                     2739 non-null   object
 15  Frag_Frac_During_Rx                                     2739 non-null   object
 16  Risk_Segment_Prior_Ntm                                  2739 non-null   object
 17  Tscore_Bucket_Prior_Ntm                                 2739 non-null   object
 18  Risk_Segment_During_Rx                                  2739 non-null   object
 19  Tscore_Bucket_During_Rx                                 2739 non-null   object
 20  Change_T_Score                                          2739 non-null   object
 21  Change_Risk_Segment                                     2739 non-null   object
 22  Adherent_Flag                                           2739 non-null   object
 23  Idn_Indicator                                           2739 non-null   object
 24  Injectable_Experience_During_Rx                         2739 non-null   object
 25  Comorb_Encounter_For_Screening_For_Malignant_Neoplasms  2739 non-null   object
 26  Comorb_Encounter_For_Immunization                       2739 non-null   object
 27  Comorb_Encntr_For_General_Exam_W_O_Complaint,_Susp_Or_Reprtd_Dx  2739 non-null   object
 28  Comorb_Vitamin_D_Deficiency                             2739 non-null   object
 29  Comorb_Other_Joint_Disorder_Not_Elsewhere_Classified    2739 non-null   object
 30  Comorb_Encntr_For_Oth_Sp_Exam_W_O_Complaint_Suspected_Or_Reprtd_Dx  2739 non-null   object
 31  Comorb_Long_Term_Current_Drug_Therapy                   2739 non-null   object
 32  Comorb_Dorsalgia                                        2739 non-null   object
 33  Comorb_Personal_History_Of_Other_Diseases_And_Conditions  2739 non-null   object
 34  Comorb_Other_Disorders_Of_Bone_Density_And_Structure    2739 non-null   object
 35  Comorb_Disorders_of_lipoprotein_metabolism_and_other_lipidemias  2739 non-null   object
 36  Comorb_Osteoporosis_without_current_pathological_fracture  2739 non-null   object
 37  Comorb_Personal_history_of_malignant_neoplasm           2739 non-null   object
 38  Comorb_Gastro_esophageal_reflux_disease                 2739 non-null   object
 39  Concom_Cholesterol_And_Triglyceride_Regulating_Preparations  2739 non-null   object
 40  Concom_Narcotics                                        2739 non-null   object
```

```
 41  Concom_Systemic_Corticosteroids_Plain                2739 non-null   object
 42  Concom_Anti_Depressants_And_Mood_Stabilisers         2739 non-null   object
 43  Concom_Fluoroquinolones                              2739 non-null   object
 44  Concom_Cephalosporins                                2739 non-null   object
 45  Concom_Macrolides_And_Similar_Types                  2739 non-null   object
 46  Concom_Broad_Spectrum_Penicillins                    2739 non-null   object
 47  Concom_Anaesthetics_General                          2739 non-null   object
 48  Concom_Viral_Vaccines                                2739 non-null   object
 49  Risk_Type_1_Insulin_Dependent_Diabetes               2739 non-null   object
 50  Risk_Osteogenesis_Imperfecta                         2739 non-null   object
 51  Risk_Rheumatoid_Arthritis                            2739 non-null   object
 52  Risk_Untreated_Chronic_Hyperthyroidism               2739 non-null   object
 53  Risk_Untreated_Chronic_Hypogonadism                  2739 non-null   object
 54  Risk_Untreated_Early_Menopause                       2739 non-null   object
 55  Risk_Patient_Parent_Fractured_Their_Hip              2739 non-null   object
 56  Risk_Smoking_Tobacco                                 2739 non-null   object
 57  Risk_Chronic_Malnutrition_Or_Malabsorption           2739 non-null   object
 58  Risk_Chronic_Liver_Disease                           2739 non-null   object
 59  Risk_Family_History_Of_Osteoporosis                  2739 non-null   object
 60  Risk_Low_Calcium_Intake                              2739 non-null   object
 61  Risk_Vitamin_D_Insufficiency                         2739 non-null   object
 62  Risk_Poor_Health_Frailty                             2739 non-null   object
 63  Risk_Excessive_Thinness                              2739 non-null   object
 64  Risk_Hysterectomy_Oophorectomy                       2739 non-null   object
 65  Risk_Estrogen_Deficiency                             2739 non-null   object
 66  Risk_Immobilization                                  2739 non-null   object
 67  Risk_Recurring_Falls                                 2739 non-null   object
 68  Count_Of_Risks                                       2739 non-null   int64
dtypes: int64(2), object(67)
memory usage: 1.5+ MB
```

We can see that for each variable we have 2739 non-null values, meaning that there are no null data points in our dataset. But there are 5 variables: Ntm_Speciality, T score During Rx, Change in T score, Risk Segment During Rx and Change in Risk Segment which have an 'Unknown' category, meaning that no value was observed. In the table below we have the number of 'Unknown' counts in each of them:

Variable	'Unknown' category counts
Ntm_Speciality	258/2739 = 9.42%
Risk_Segment_During_Rx	1223/2739 = 44.65%
Change_Risk_Segment	1802/2739 = 65.8%
Tscore_Bucket_During_Rx	1223/2739 = 44.65%
Change_T_Score	1223/2739 = 44.65%

The variable 'Change in Risk Segment' has over 65% of its observations marked as 'Unknown', so we decide to drop this column altogether. As for the other variables, we will try to Impute the missing values, since they are less than 60% in each column. The Imputation method decided for each variable will be performed on Training and Test sets so as not to 'leak' information in our training set, which we will use for model building, from our test set, which we will use for classification of the Persistency Flag variable.

Imputation

Rubin [**5**] classified missing data problems into three categories. In his theory every data point has some likelihood of being missing.

1. If the probability of being missing is the same for all cases, then the data are said to be missing completely at random (**MCAR**), meaning that if a certain value is missing, it has nothing to do with hypothetical value and with the values of other variables.
2. If the probability of being missing is the same only within groups defined by the *observed* data, then the data are missing at random (**MAR**), meaning that the propensity for a data point to be missing is not related to the missing data, but it is related to some of the observed data.
3. If the probability of being missing varies for reasons that are unknown to us, then the data are said to be missing not at random (**MNAR**), meaning that missing value depends on a hypothetical value, or on some other variable's value. Usual strategy for this case is to gather more data, which in our case we cannot do at the moment.

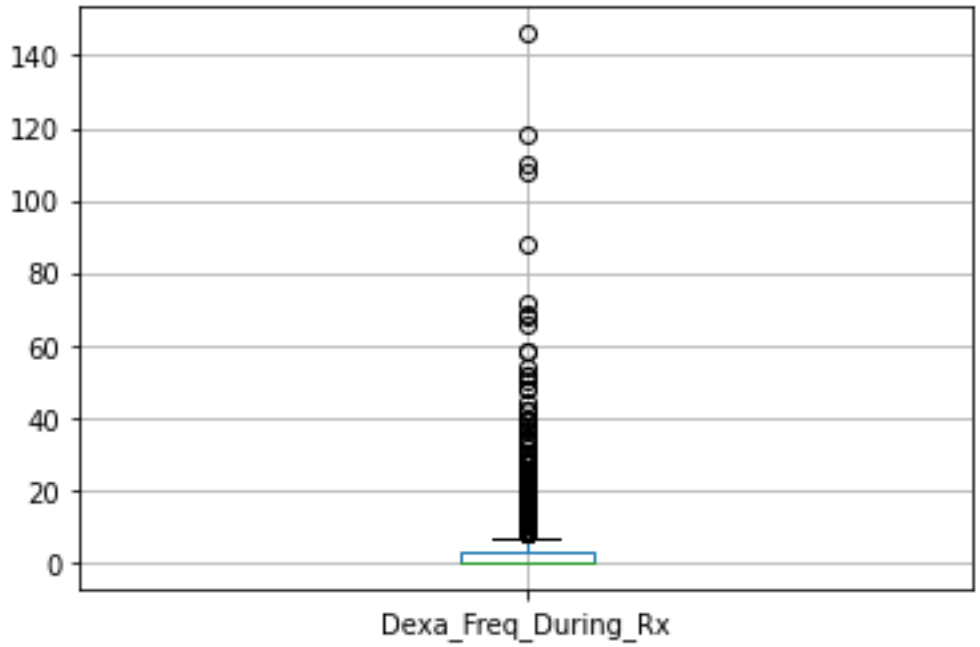
We will consider our variables' missing values to be unrelated to the missing data, but could be related to the observed data (MAR), so we will impute them. For all these variables, we will impute the missing values with the KNNImputer method, with 'number of neighbors' = 1. This method takes into account the whole dataset and each missing feature is imputed using values from its nearest neighbor, where distance is calculated via a eucledian metric that supports missing values (**nan_euclidean_distances**[**6**]). We present below as an example two of the variables in our training set, before and after imputation:

Variable	Before Imputation	After Imputation
Risk_Segment_During_Rx	High Risk: 763 Unknown: 1421 Low Risk: 753 Total: 2937	High Risk: 1292 Low Risk: 1645 Total: 2937
Tscore_Bucket_During_Rx	<=-2.5: 805 >-2.5: 711 Unknown : 1421 Total: 2937	<=-2.5: 1491 >-2.5: 1446 Total: 2937

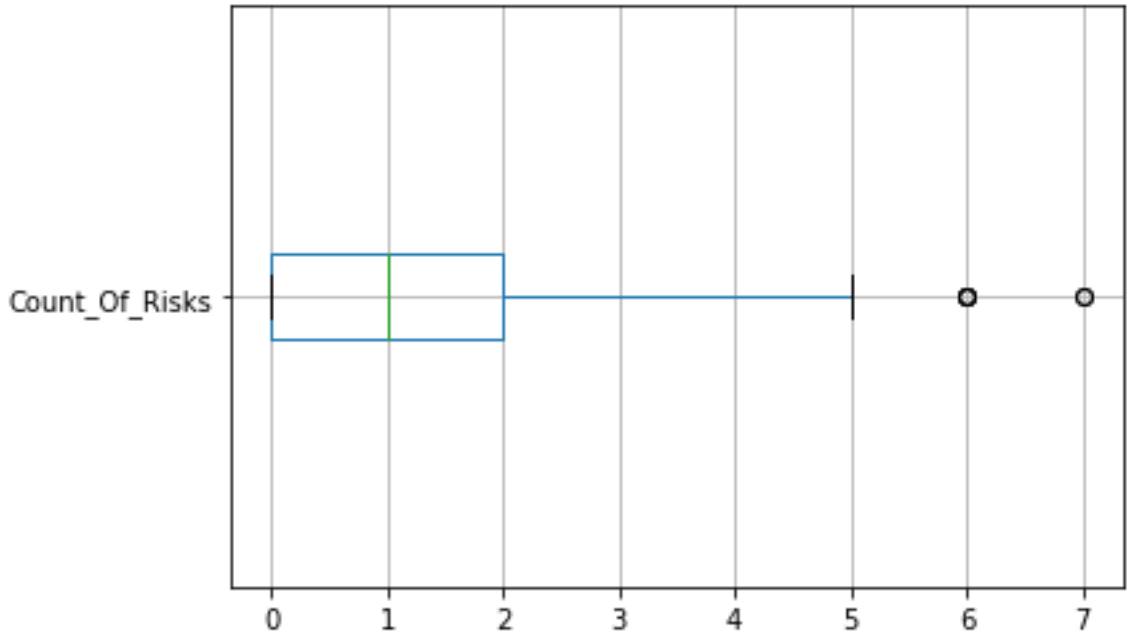
There are no more missing values, 'Unknown' label, and the values imputed do not seem to have affected the distribution of the labels by much.

Outliers

We present below the boxplot for the Frequency of Dexa Scans the patients did, where we have the number of scans taken during the patient's first continuous therapy. An NTM infection therapy can take many months. A patient is considered cured when samples taken from them show no sign of NTM infection for at least 12 months [**7**]. A big percentage of the patients did not have a Dexa Scan (2488, or 72.66%), as can be seen by the distribution plotted in the boxplot, which has a mean of 3 scans and 75% of our observation (3rd Quatrile) are up to 3 scans. For the rest of the 25% of our observations, we can observe some values that can be considered quite high, some even greater than 100, meaning that in a 36 month therapy period the patient had a Dexa scan every 7.5 days. For patients with osteoporosis, one scan every year is recommended [**8**] so there are numbers presented here that mgiht seem out of the ordinary. Unfortunately, we do not have further information about these patients, whether they actually did that many scans or these values might be typos, or miscalculations. And so, we cannot remove these values, just on the basis that they are inconvenient.



A different picture can be seen in our 'Count of Risks' variable, as presented in the figure below. A median value of 1, with 75% of the patients presenting from 0 up to 5 possible Risk factors, and 2 patients having 7, and 6 patients having 6 Risk factors. We also have the Z-scores of this variable, where we can see that the 2 patients with 7 Risk factors are 5.26 Standard Deviations away from the variable's mean, while the other 6 patients with 6 Risk factors are 4.34 Stadard Deviations away from the variable's mean. While these are values that are very far away from the mean, they cannot be easily considered as outliers, and so we do not remove them from our analysis.



	0
460	5.257211
1363	5.257211
2133	4.345216
431	4.345216
2016	4.345216
...	...
1444	-1.126759
1443	-1.126759
1440	-1.126759
1435	-1.126759
2738	-1.126759

[2739 rows x 1 columns]

References

1. https://www.sciencedirect.com/science/article/pii/S1098301510604950
2. https://curanthealth.com/top-barriers-to-patient-persistence/
3. https://www.pharmexec.com/view/top-barriers-patient-persistence
4. https://faculty.wharton.upenn.edu/wp-content/uploads/2012/04/FS8.Lee-Fader-Hardie.pdf
5. 1976. “Inference and Missing Data.” *Biometrika* 63 (3): 581–90
6. https://scikit-learn.org/stable/modules/impute.html#knnimpute
7. https://www.lung.org/lung-health-diseases/lung-disease-lookup/nontuberculous-mycobacteria/diagnosing-and-treating-ntm
8. https://www.nof.org/patients/diagnosis-information/bone-density-examtesting/

Github Repo Link:

https://github.com/EniasVontas/Assignments/tree/main/Week8