Data Glacier

Your Deep Learning Partner

Healthcare

Persistency of a drug

Exploratory Data Analysis

Jan 16th, 2023.

Submitted by Nahari Terena
LISUM15 – Data Science Specialization

Outline

- Business Understanding
- Approach
- EDA
- Summary
- Recommendations

Business Understanding

- ABC Pharma contacted us to analyze the patients' data to have a better understanding of the factors that significantly impact the persistence of their drug. The aim is to know if a patient, based on private information, will follow the prescription and continue taking the medication for all the treatment time or not.
    - We aim to develop a web app to predict if a patient will get or not a drug schedule.

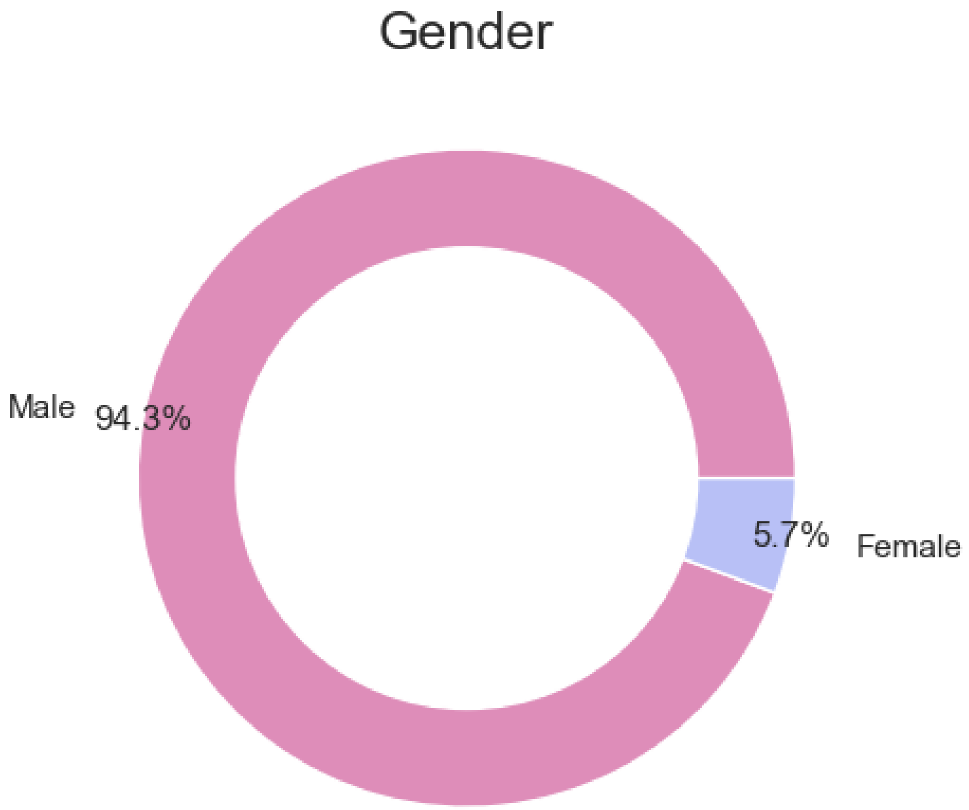
Approach

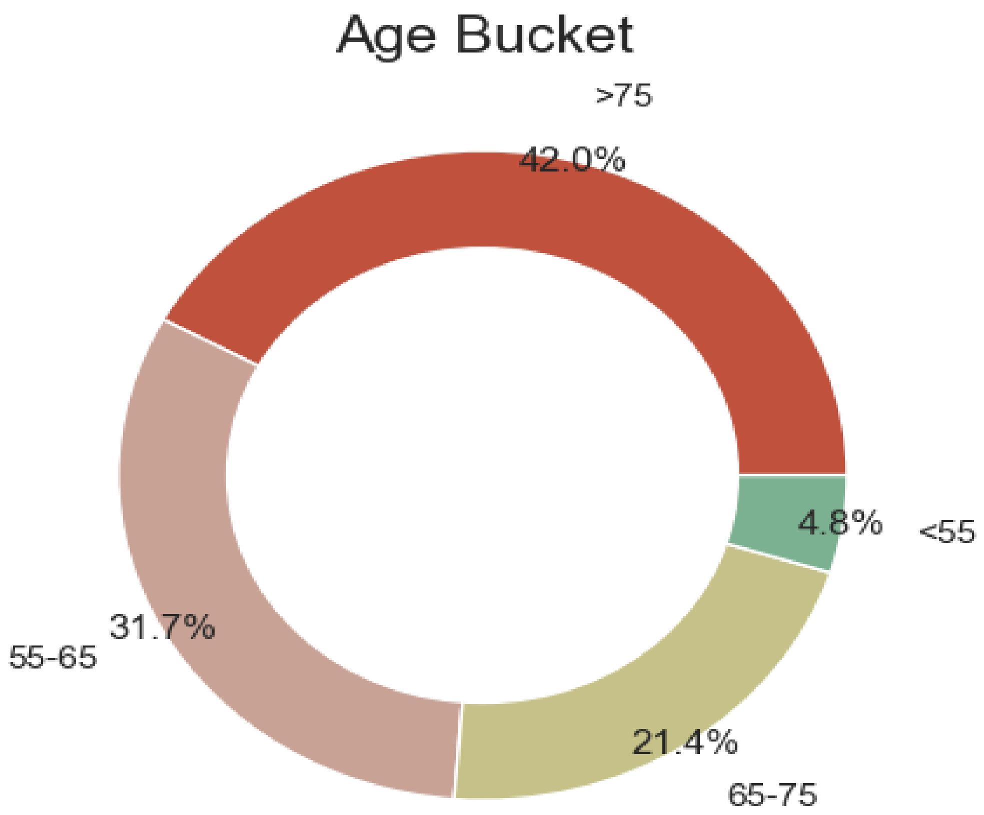
The avaliable data is the Healthcare dataset, which contains 3,424 patients and 69 features.

For each patient, we can check information about demographics, clinical records, other diseases such as comorbidity, risk factors and their physician's speciality.

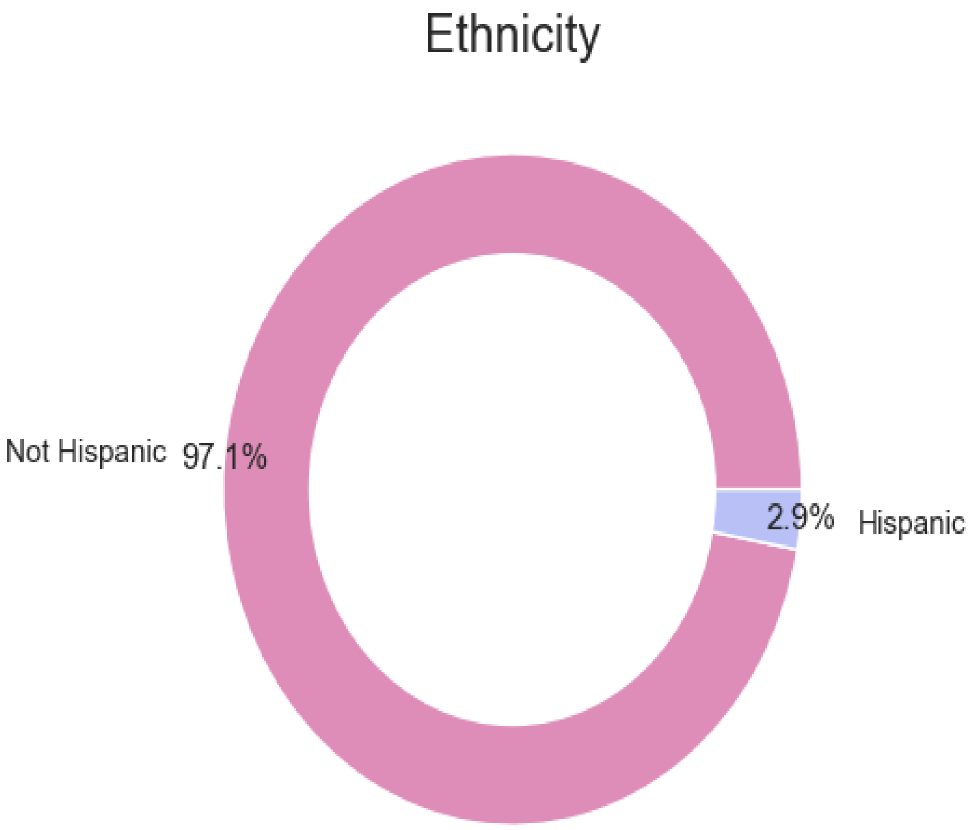
The exploratory data analysis (EDA) take into account the whole dataset so we can have insights from the analysis.

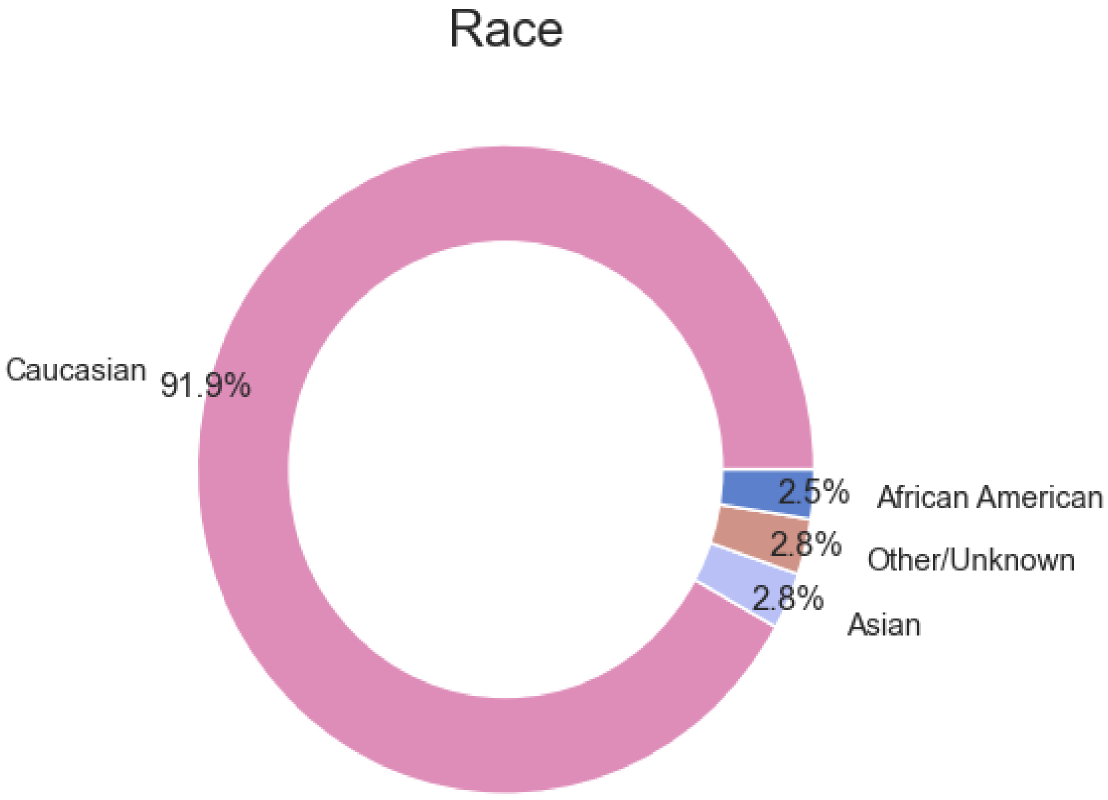
Demographics



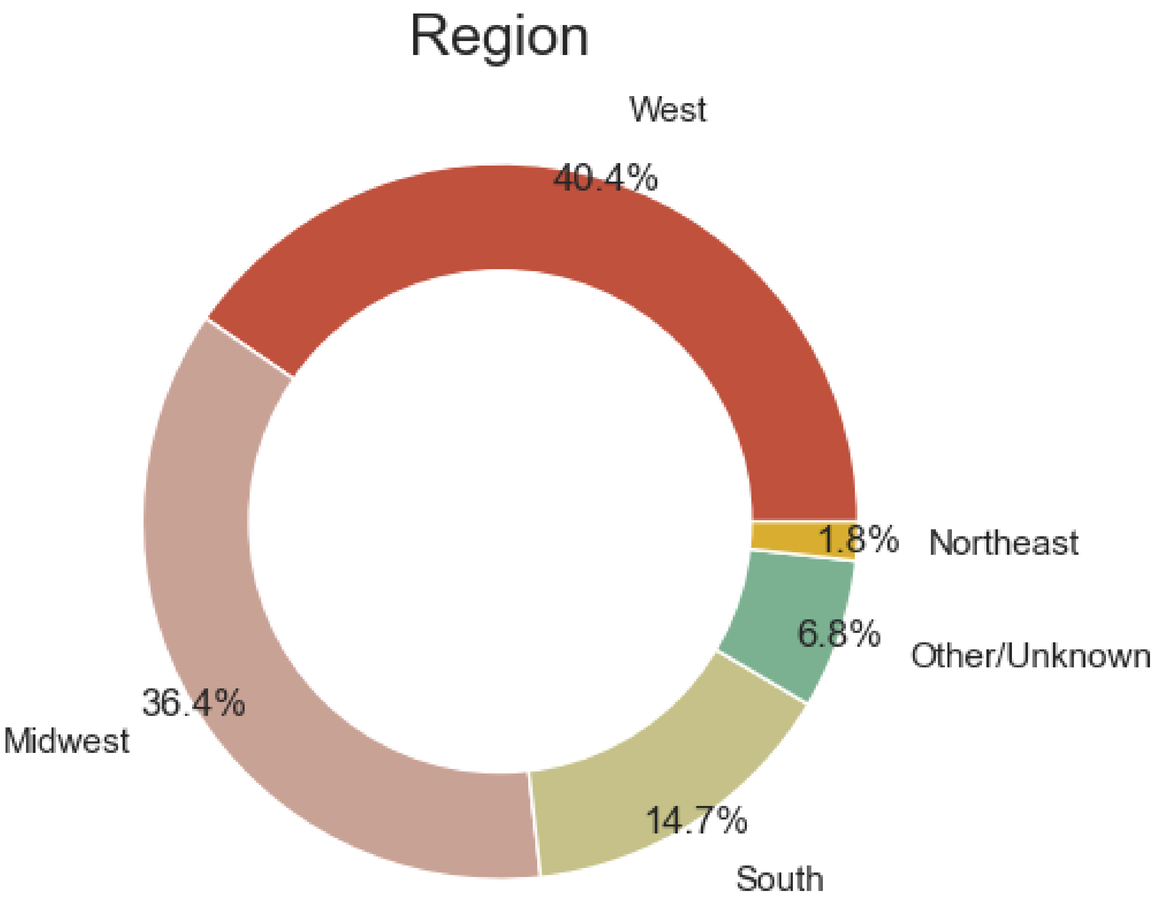


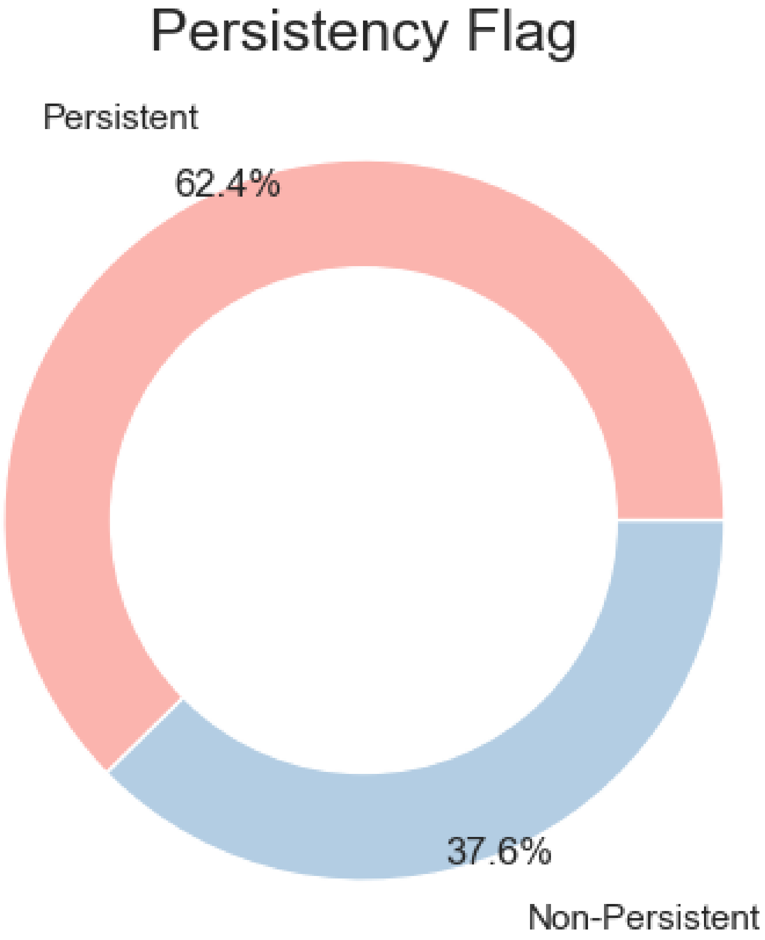
Demographics



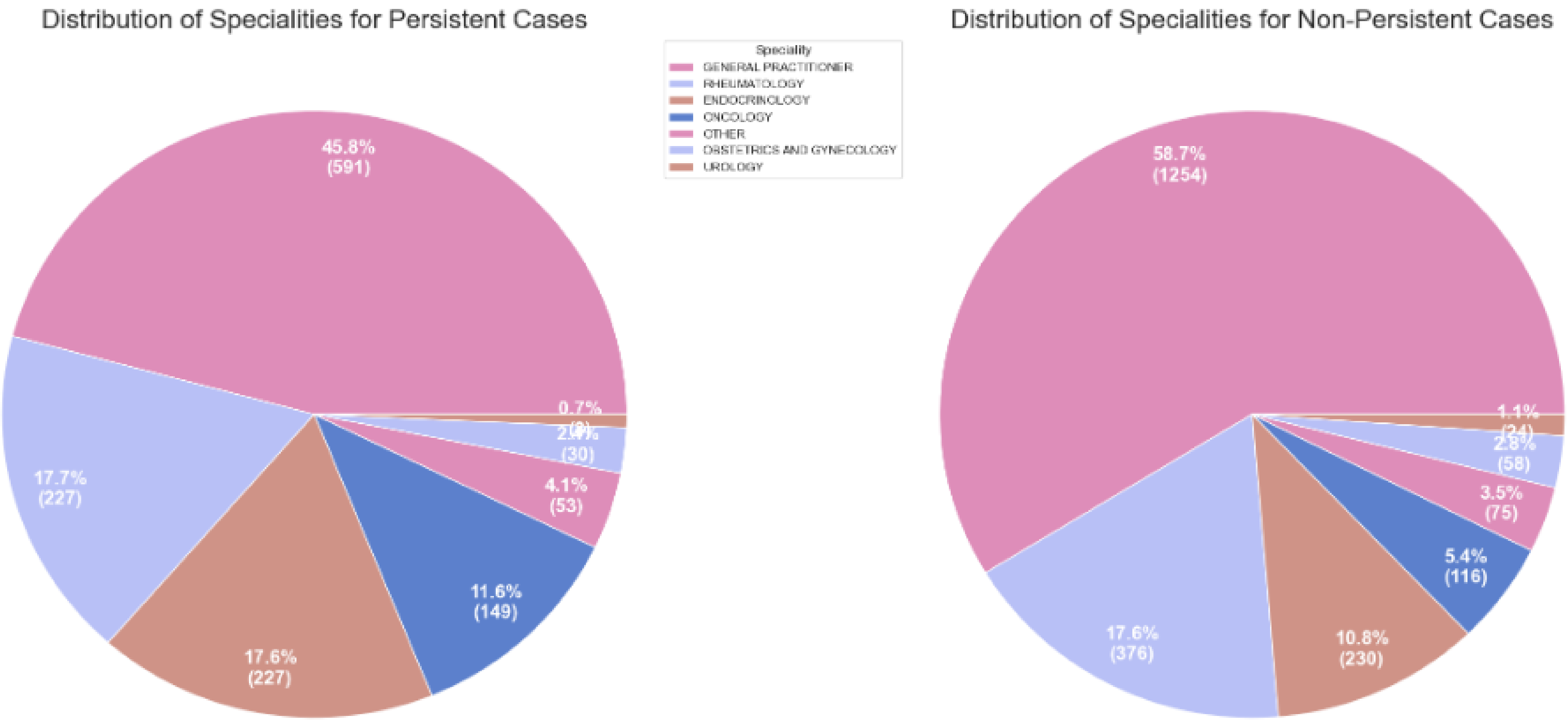


Demographics





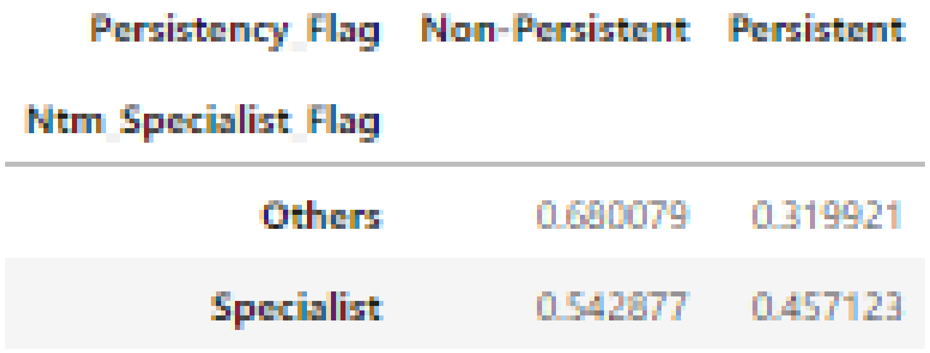
Clinical Factors



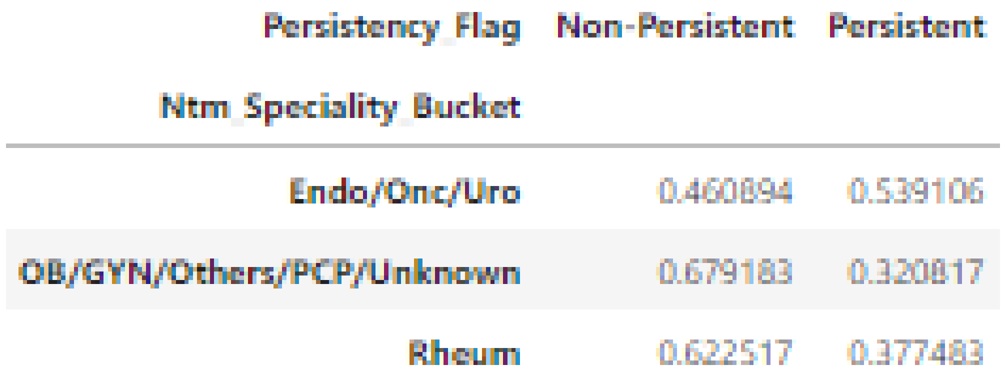
Distribution of Specialities for Persistent Cases

Distribution of Specialities for Non-Persistent Cases

Clinical Factors



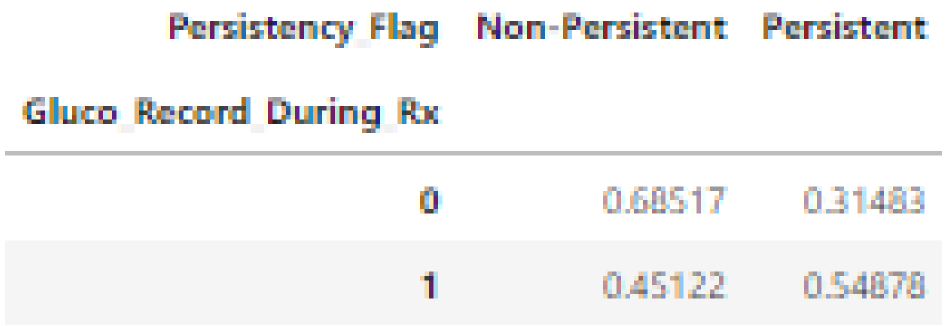
Persistency_Flag	Non-Persistent	Persistent
Ntm_Specialist_Flag		
Others	0.680079	0.319921
Specialist	0.542877	0.457123



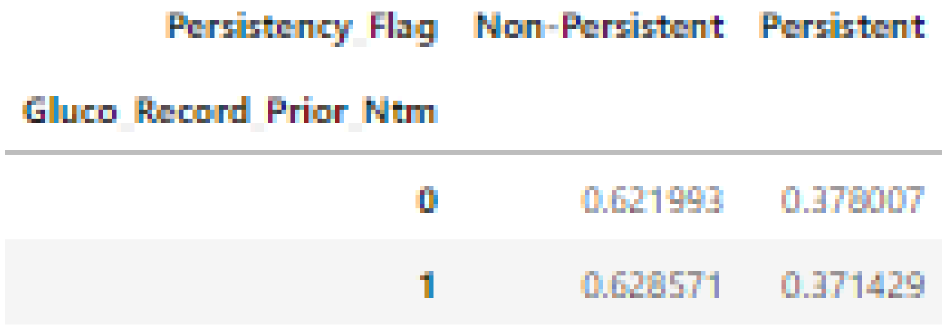
Persistency_Flag	Non-Persistent	Persistent
Ntm_Speciality_Bucket		
Endo/Onc/Uro	0.460894	0.539106
OB/GYN/Others/PCP/Unknown	0.679183	0.320817
Rheum	0.622517	0.377483

It seems that Rheum flag in Ntm_Speciality have some useful information.

Clinical Factors



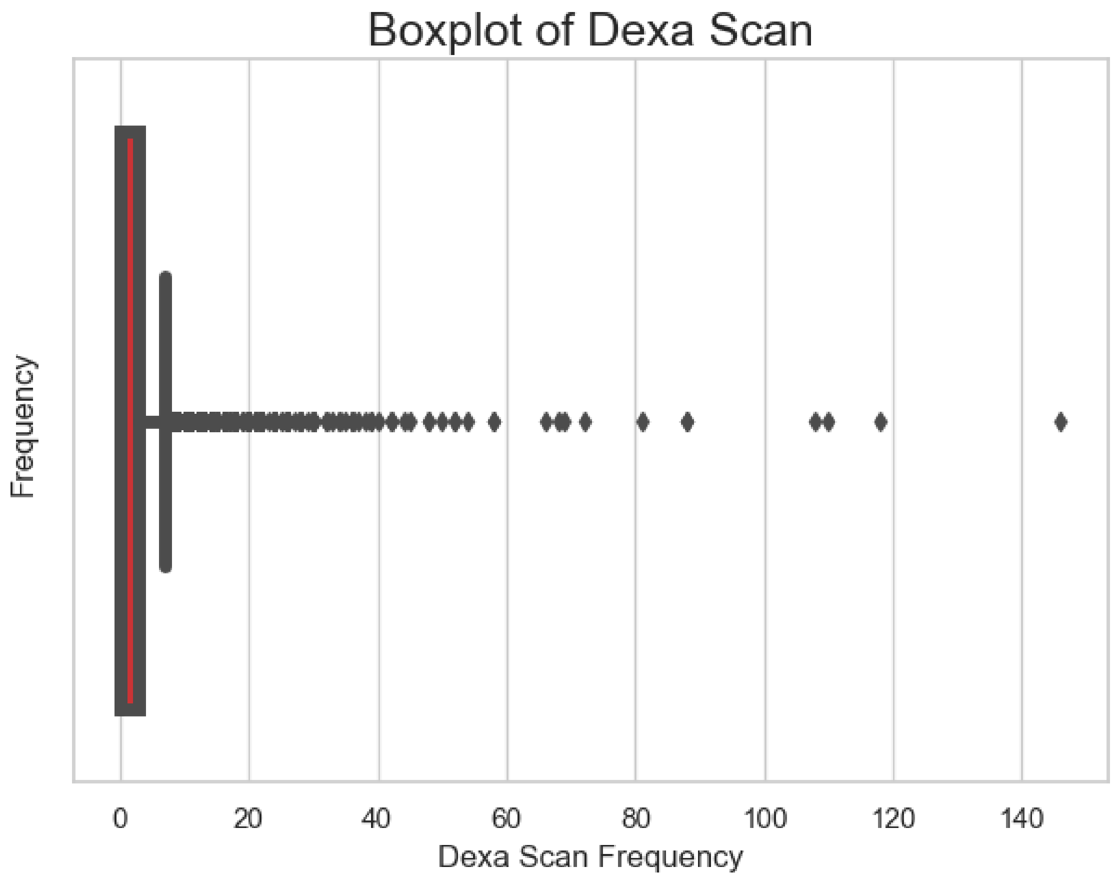
Persistency_Flag	Non-Persistent	Persistent
Gluco_Record_During_Rx		
0	0.68517	0.31483
1	0.45122	0.54878



Persistency_Flag	Non-Persistent	Persistent
Gluco_Record_Prior_Ntm		
0	0.621993	0.378007
1	0.628571	0.371429

It seems that Gluco_Record_During_Rx seems to be more useful than Gluco_Record_Prior_Ntm to predict the target.

Clinical Factors



We decided to maintain the outliers so we may probably identify these patients as a group itself.

Disease/Treatment Factor

There are some significant differences between genders:

- Women seem to be more affected by vitamin D deficiencies.
- More than twice as many women as men have passed as screening for malignant neoplasms.
- Four times as many men as women suffer from Hypogonadism (untreated).

There are some risks and other factors that seem to be significantly higher in South and West regions. It might be interesting to find out about socioeconomic factors aside.

There seems to be some remarkable differences between Asian and other races. They are probably due to cultural factors and other behaviors.

Patients older than 65 years old are affect by the risks in higher proportion.

Disease/Treatment Factor

About 99% holds at least one risk, comorbid or concomitant factor.

It’s easy to see that most of the patients already hold comorbidity factors, while holding risk factors is less common.

The main comorbidity factor is related to lipoproteins and metabolism (cholesterol). On the other hand, the main risk factor is deficiency in vitamin D. More than one third has been found to have taken narcotics.

Summary

- **Demographics** variables seems to be relevant because it brings the context in which the patient is inserted. There are some notable differences between gender and risks factors. Also, when regarding being **Asian** or another race.

- Most of the patients already hold at least one comorbidity factor, on the other hand, holding risk factor is less common.

- Information about **Dexa Scan** and **Gluco Record** have a useful information for the classification.

- The distributions of frequency for the target variable by speciality are pretty similar. Thus, we may rule out the possibility that one of the factors is who prescribed the drug in the first place.

Recommendations

Regarding the prediction model, we may consider interpretable model such as decision tree to support the predictions of more complex models.

- Support Vector Machine algorithm to classify the persistence of patients. A linear kernel has been used, obtaining an accuracy of 83.5% over testing data (25% out of the whole dataset)

- Random Forest algorithm to classification. The algorithm has 1000 estimators, max depth of 10, obtaining an accuracy of 81% and AUC of 89% over testing data.

- Logistic Regression algorithm for binary classification. Using GridSearchCV for optimization, the LR model uses 204 columns, after one-hot-encoding, to train. The F1 scor obtained is 82%.

Thank You

Data Glacier

Your Deep Learning Partner

Github Repo https://github.com/naharift/DataGlacier/tree/main/Week11