Drug Persistency Project

Virtual Internship: Week 11 Presentation on EDA on the Dataset

Group Name: Attack on Data
Group Members:
Armin Khayati (United Arab Emirates)
Ezzuldin Zaky (United Arab Emirates)
Orcun Sami Tandogan (Turkey)
Date: 10-May-2022

Attack on Data Group Members Information

Group Name: **Attack on Data**

Team Members:

Name	Email	Country	College/Company	Specialization
Armin Khayati	Northatlas@gmail.com	UAE	-	Data Science
Ezzuldin Zaky	Ezzulding.zaky@gmail.com	UAE	American University of Sharjah	Data Science
Orcun Sami Tandogan	tandogan.orcun@metu.edu.tr	Turkey	Middle East Technical University	Data Science

Background – Drug Persistency case study

- One of the challenges for all pharmaceutical companies is to understand the persistency of drug as per the physician prescription. To solve this problem ABC pharma company approached an analytics company to automate this process of identification.
- Objective : Find important features and prepare them by Feature Engineering and Feature Selection techniques for training with machine learning algorithms.
- The analysis has been divided into several parts:
  - Data Understanding
  - Data Cleaning
  - Data insights and visualization
  - Recommendations

Data Exploration

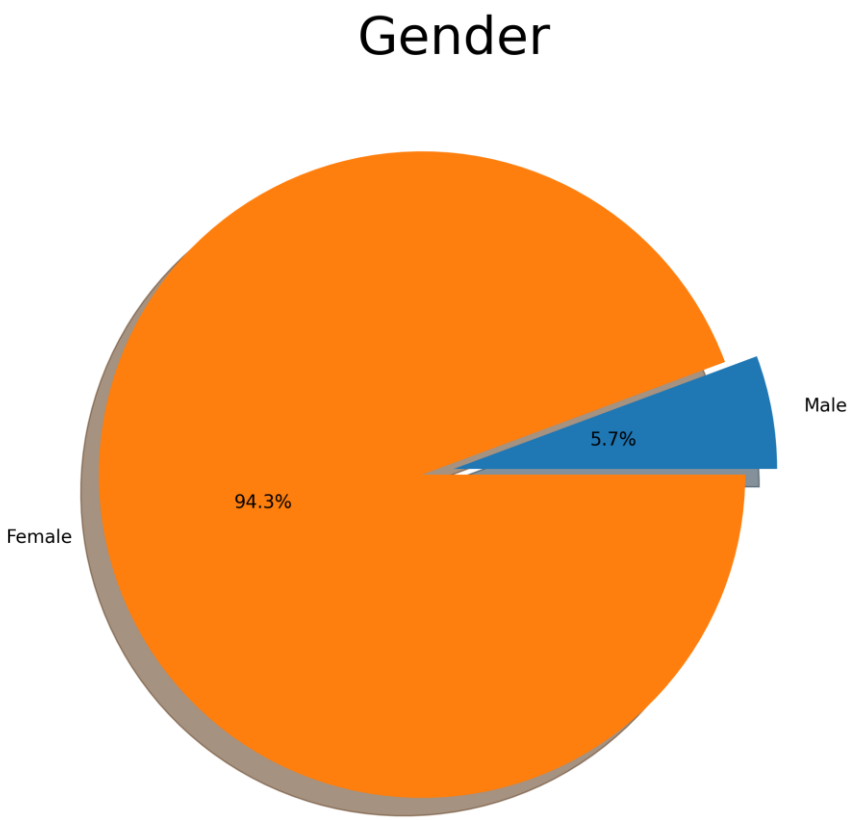
- 68 Features, two numerical and 66 categorical, including :
  - General features such as (Patient general info)
  - Diseases/Drugs Factors
  - Clinical Factors

- Total number of patients : 3424

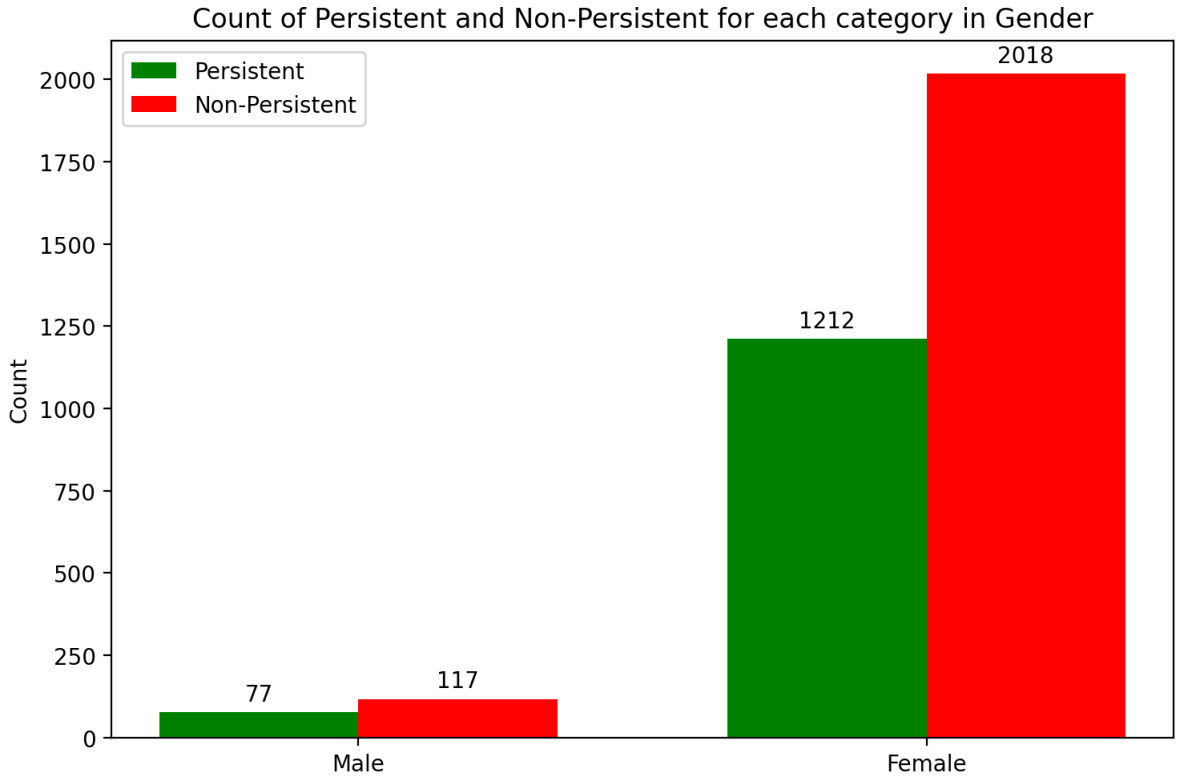
Assumptions:

- Dataset is imbalanced.
- It follows a normal distribution.
- Patients' data were gathered accurately without any errors in testing or examination.

Patients General Info Analysis

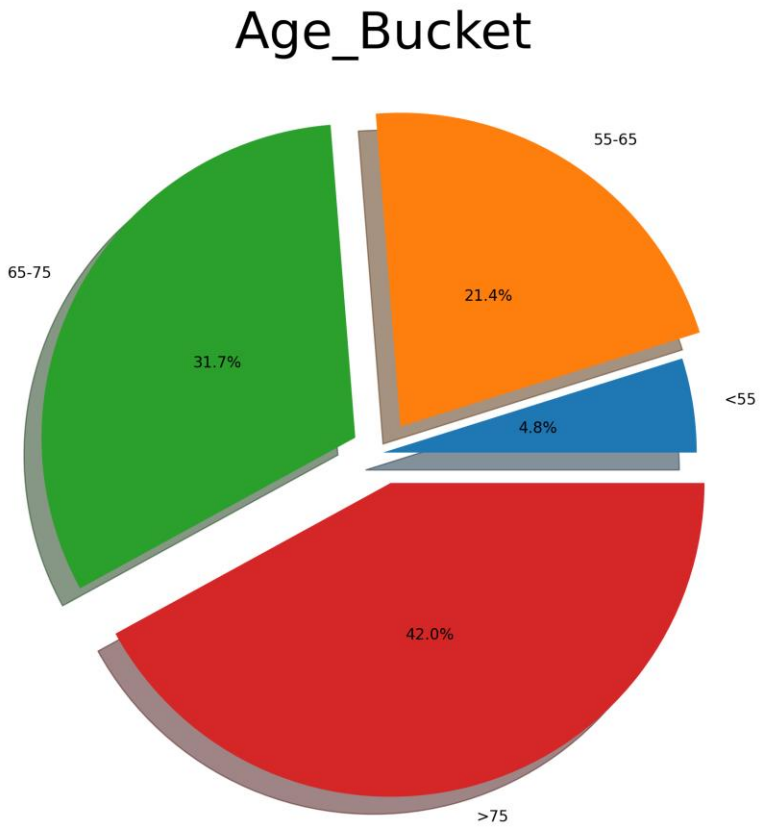


Gender Ratio

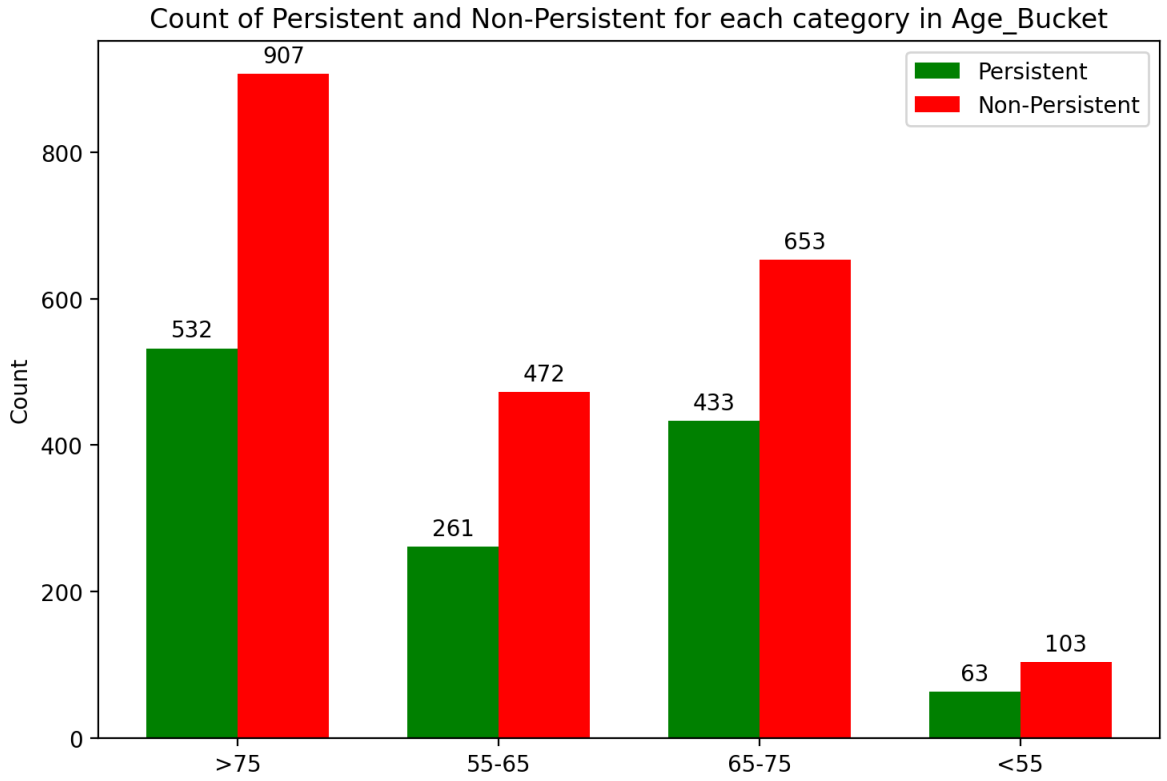


Gender Ratio vs. Persistency Flag

Patients General Info Analysis

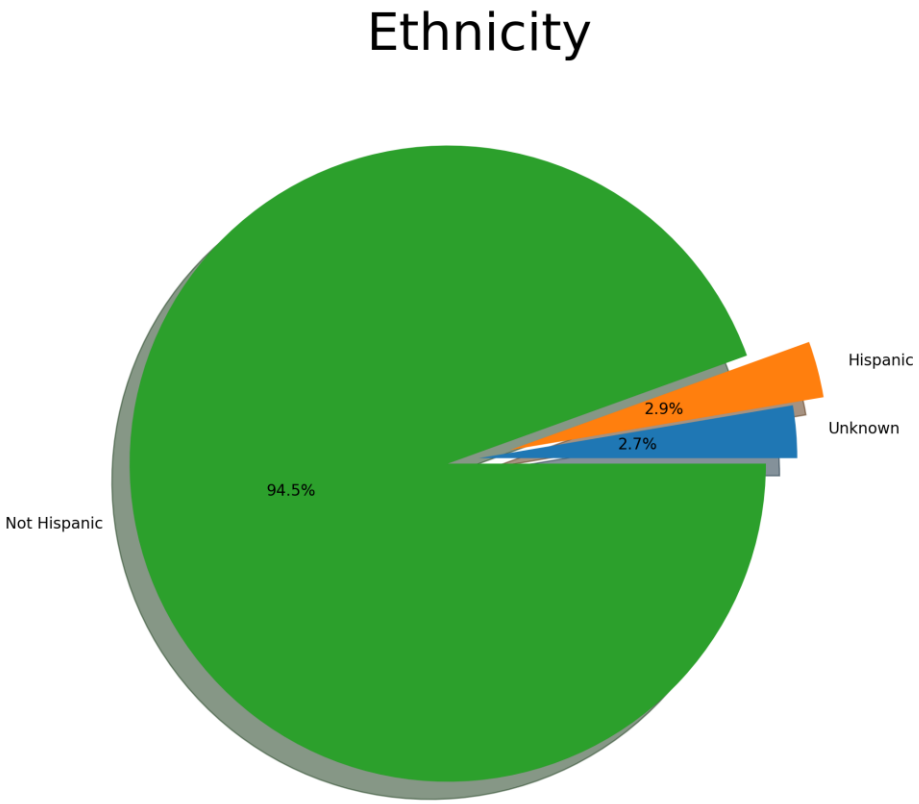


Age Ratio

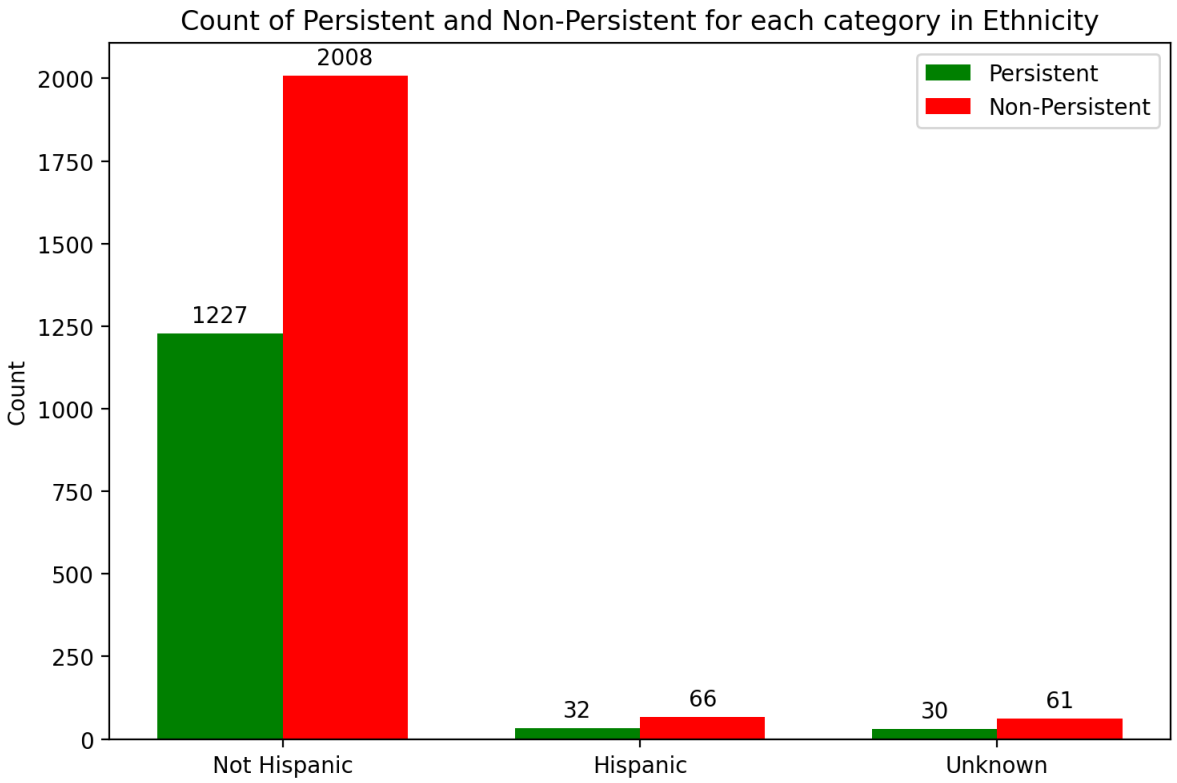


Age Bucket vs. Persistency Flag

Patients General Info Analysis

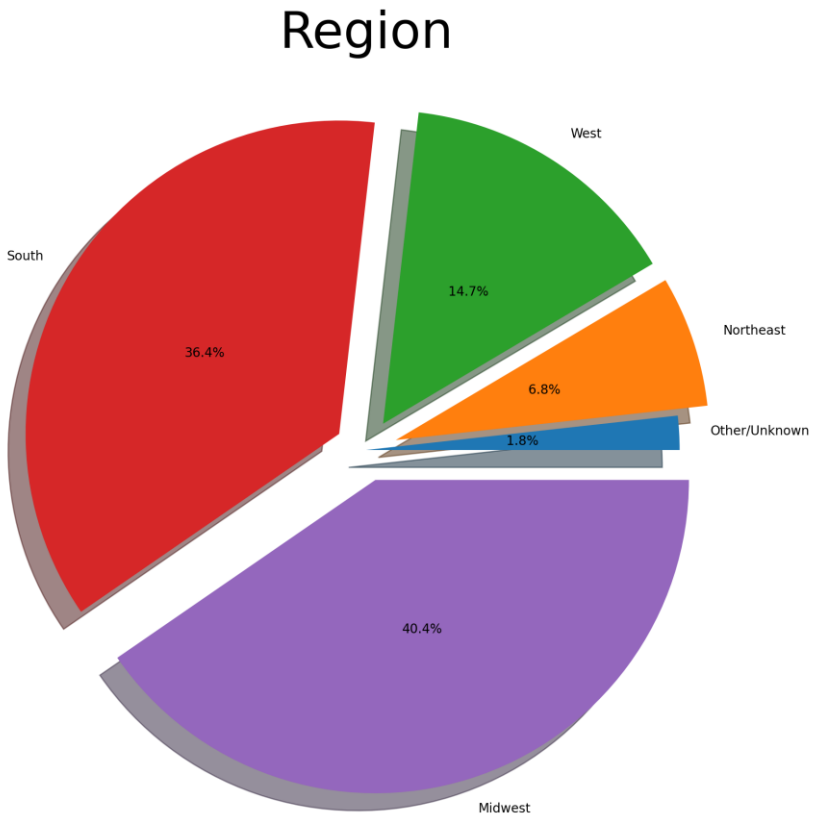


Ethnicity Ratio

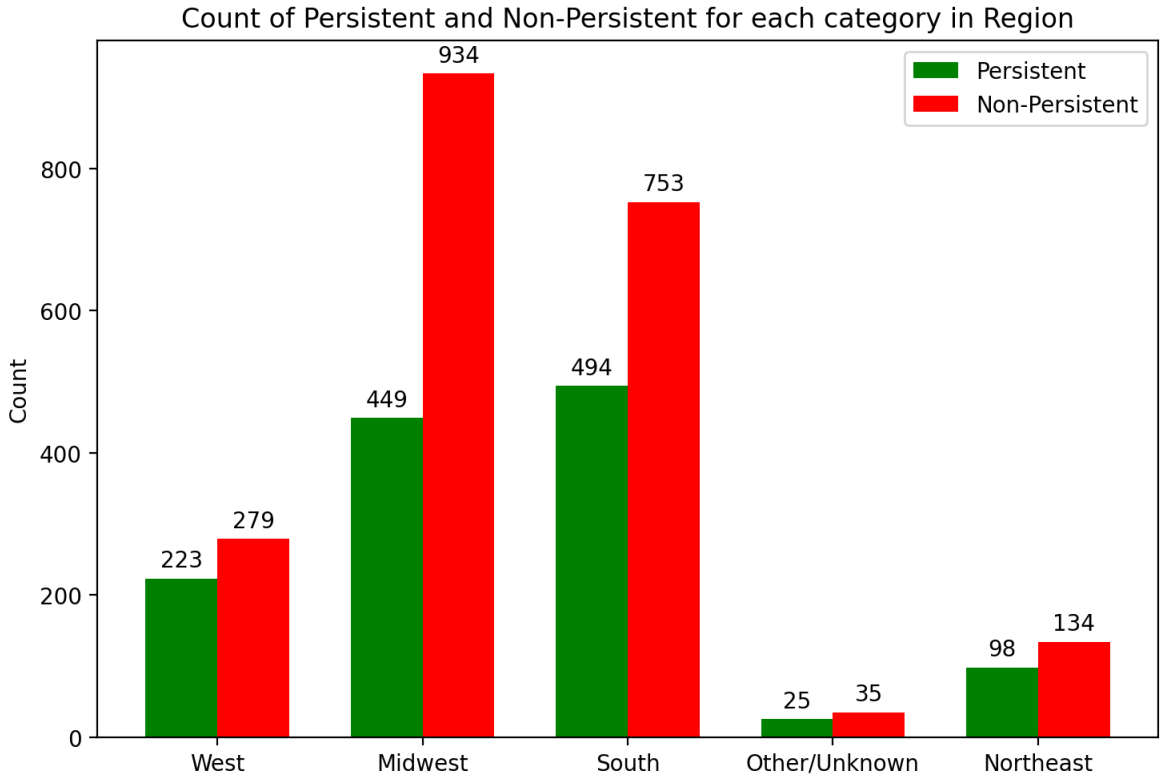


Ethnicity vs. Persistency Flag

Patients General Info Analysis

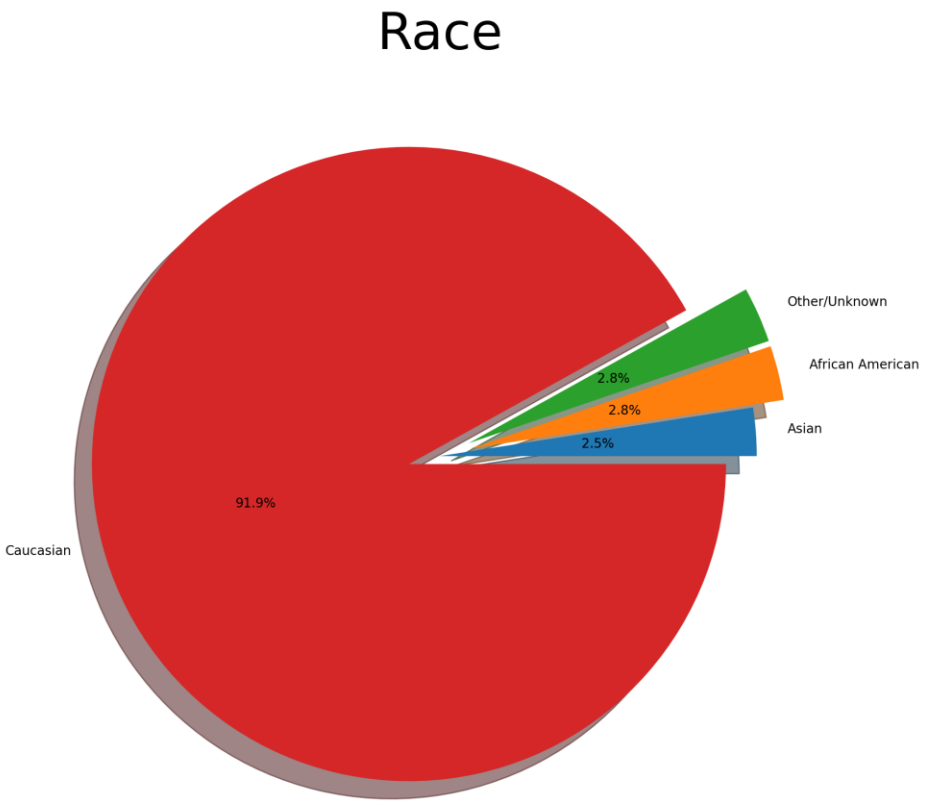


Region Ratio

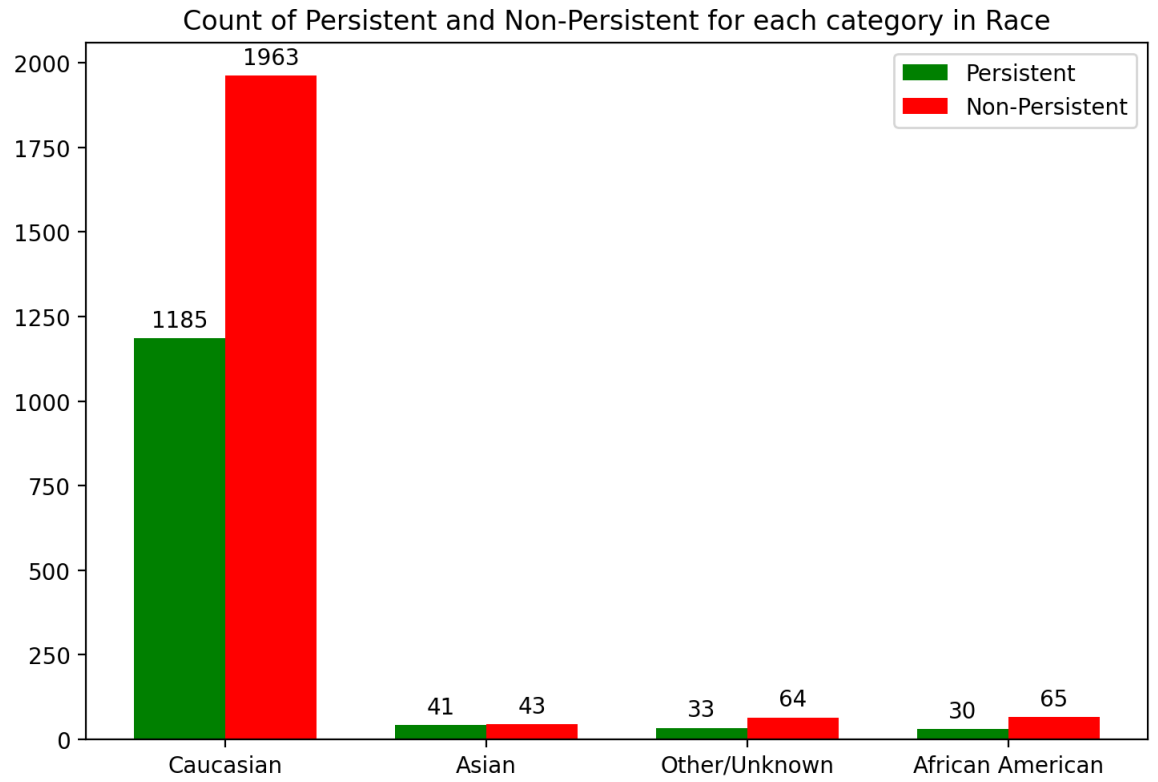


Region vs. Persistency Flag

Patients General Info Analysis

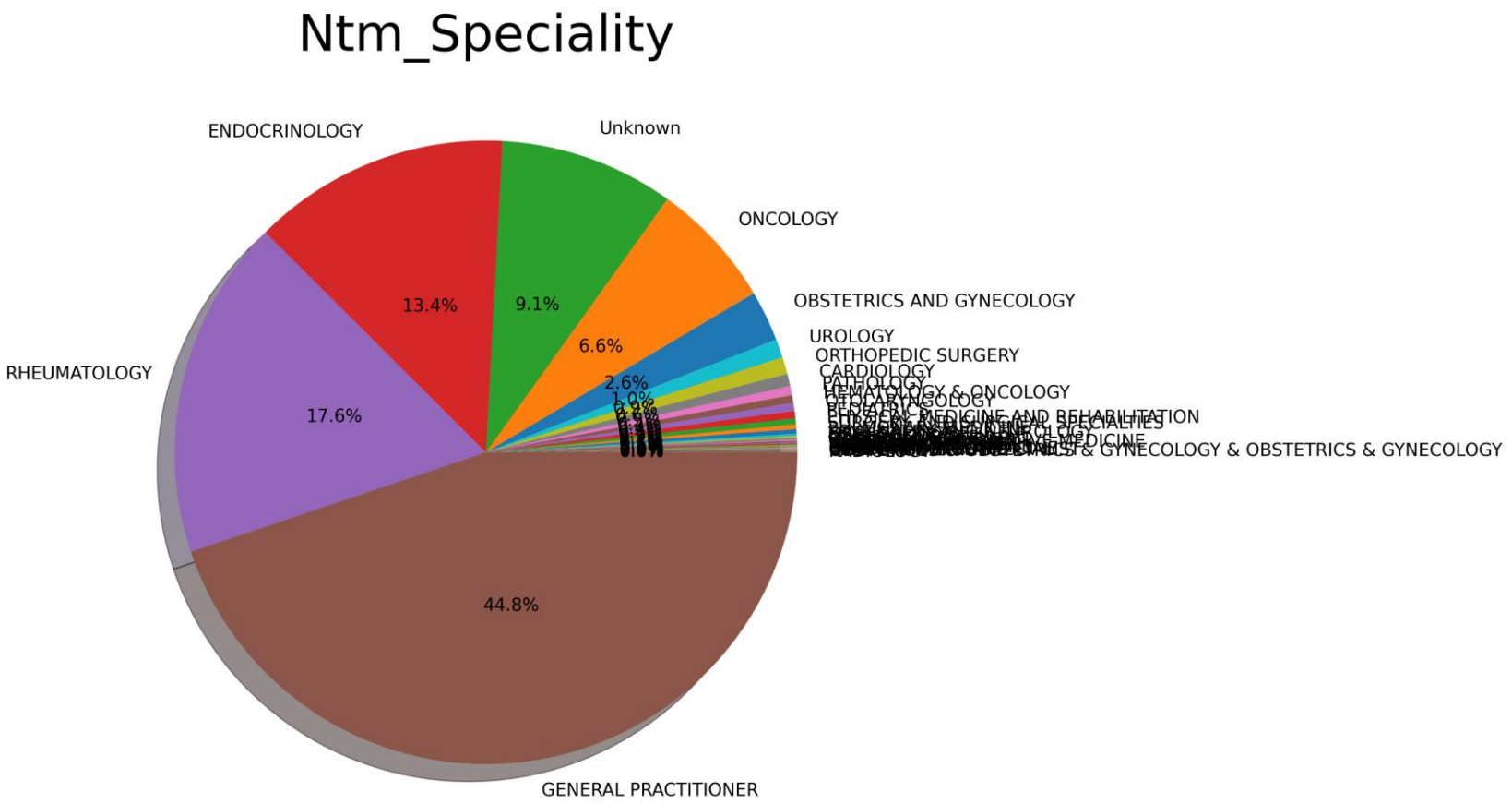


Race Ratio



Race vs. Persistency Flag

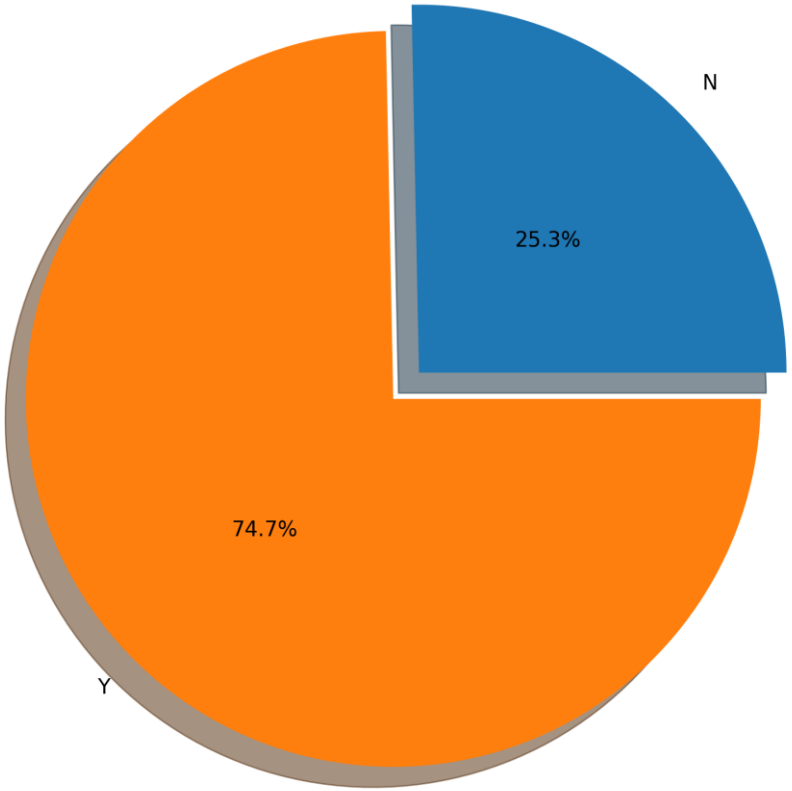
Patients General Info Analysis



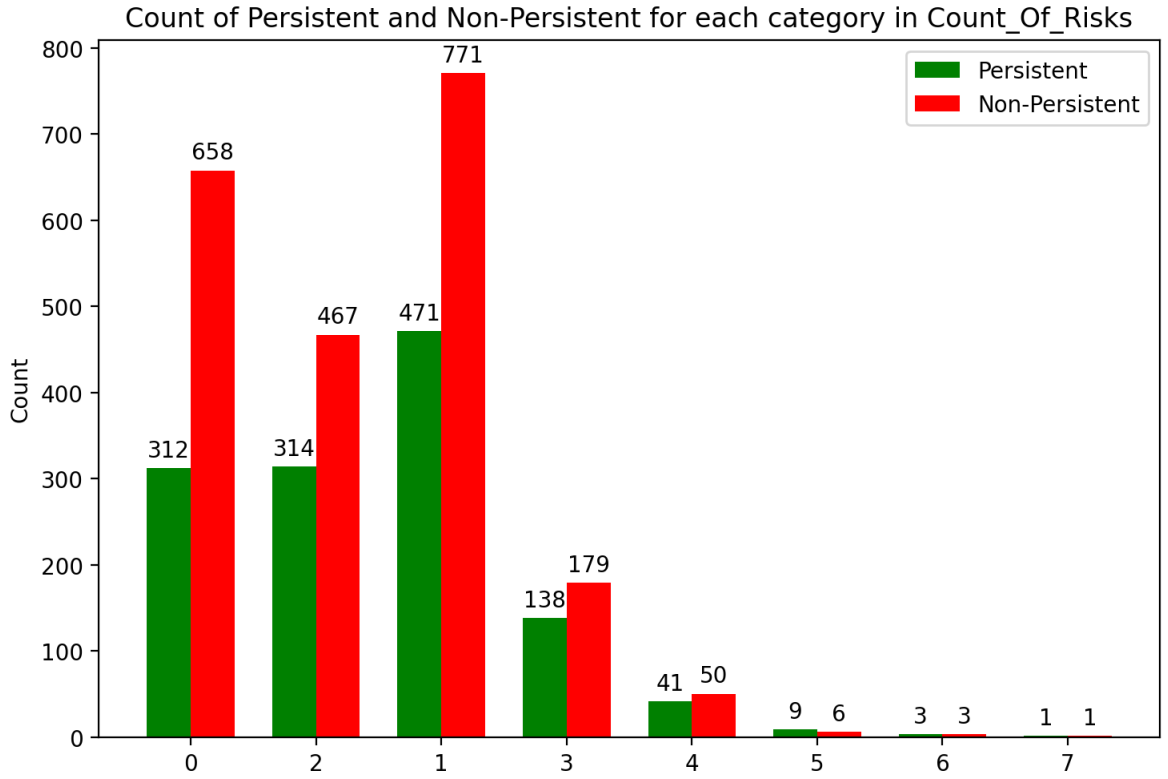
NTM_Speciality Ratio

Patients General Info Analysis

Idn_Indicator

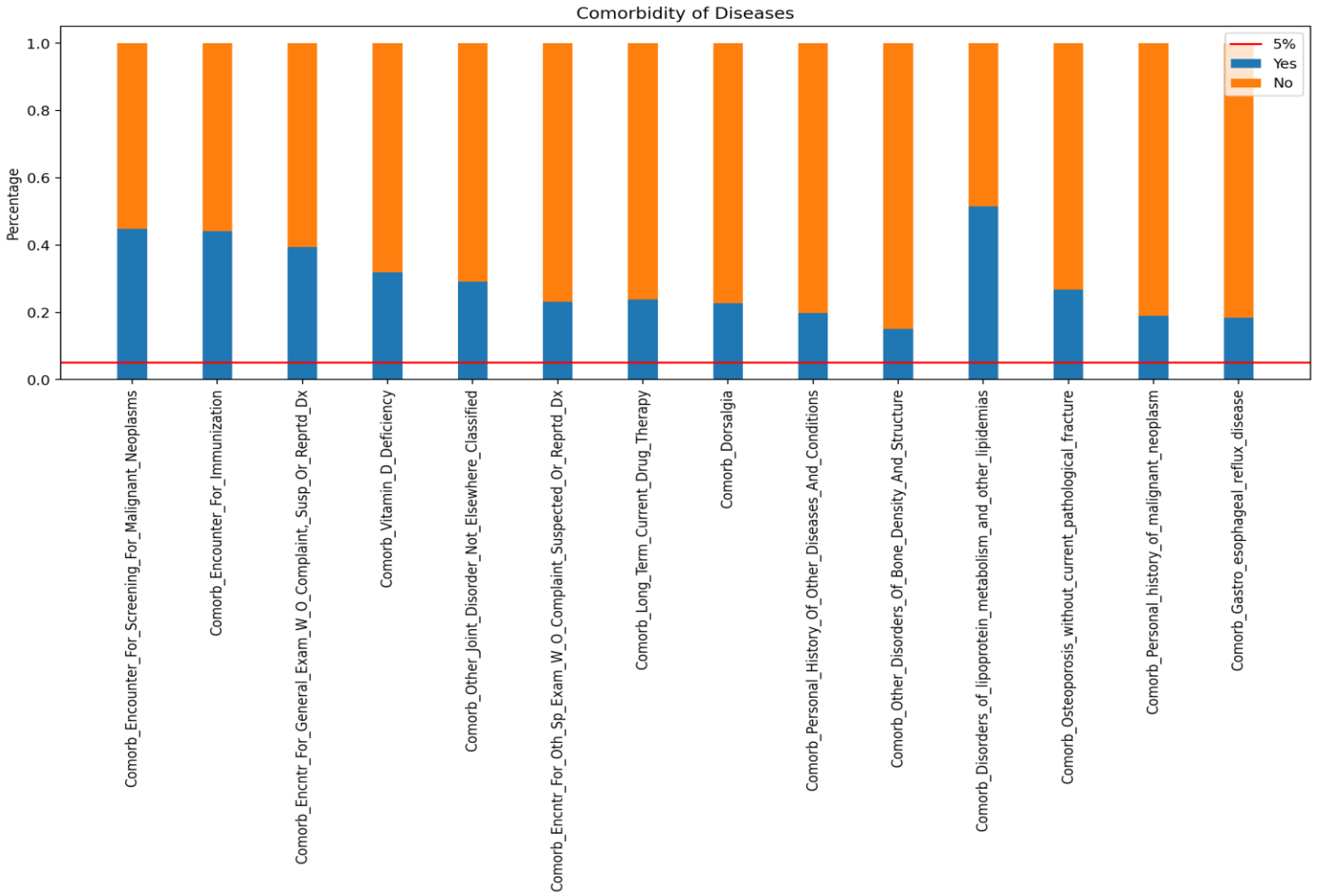


IDN Indicator Ratio



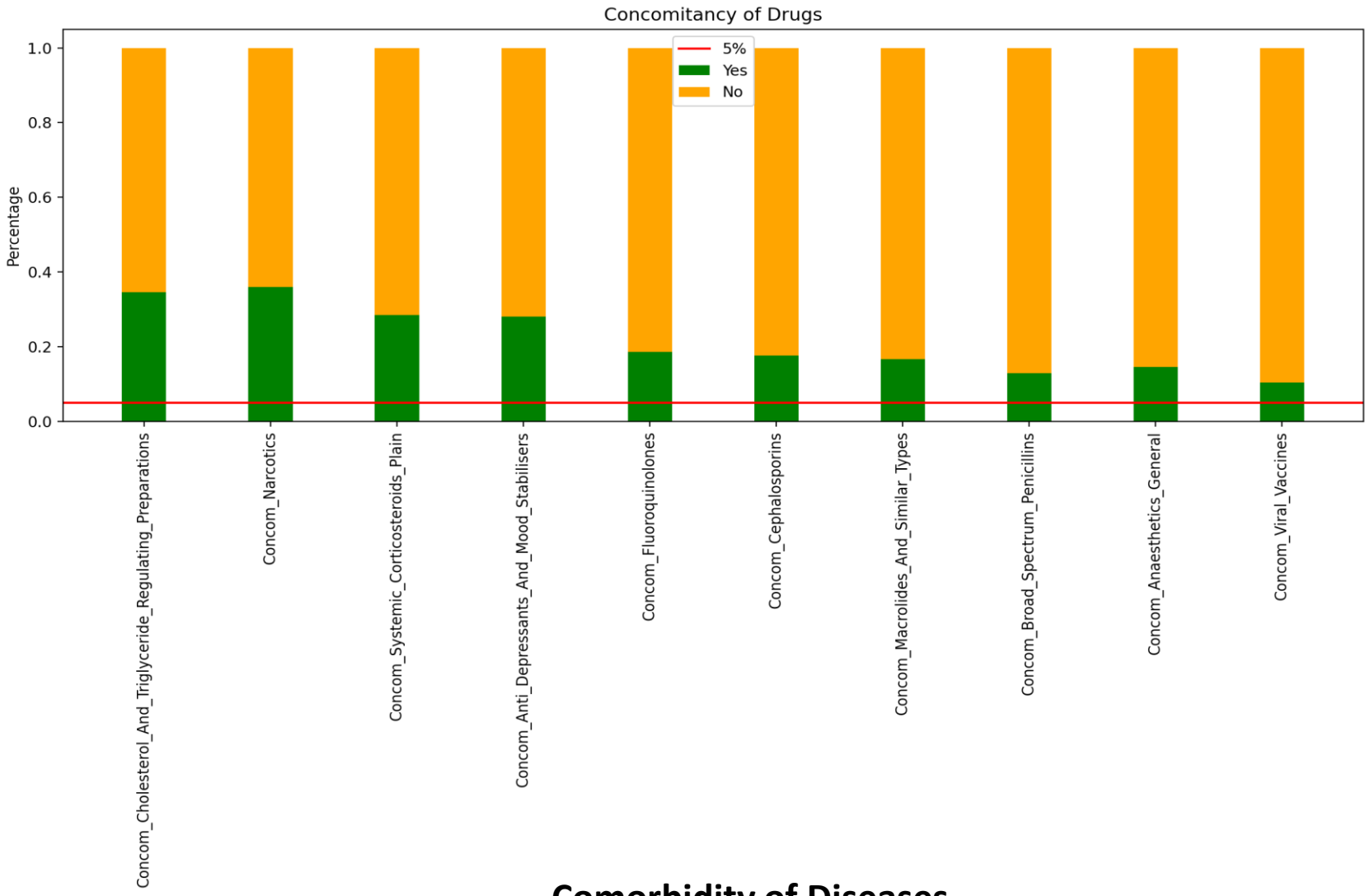
Count of Risk vs . Persistency Flag

Concomitancy of Drugs



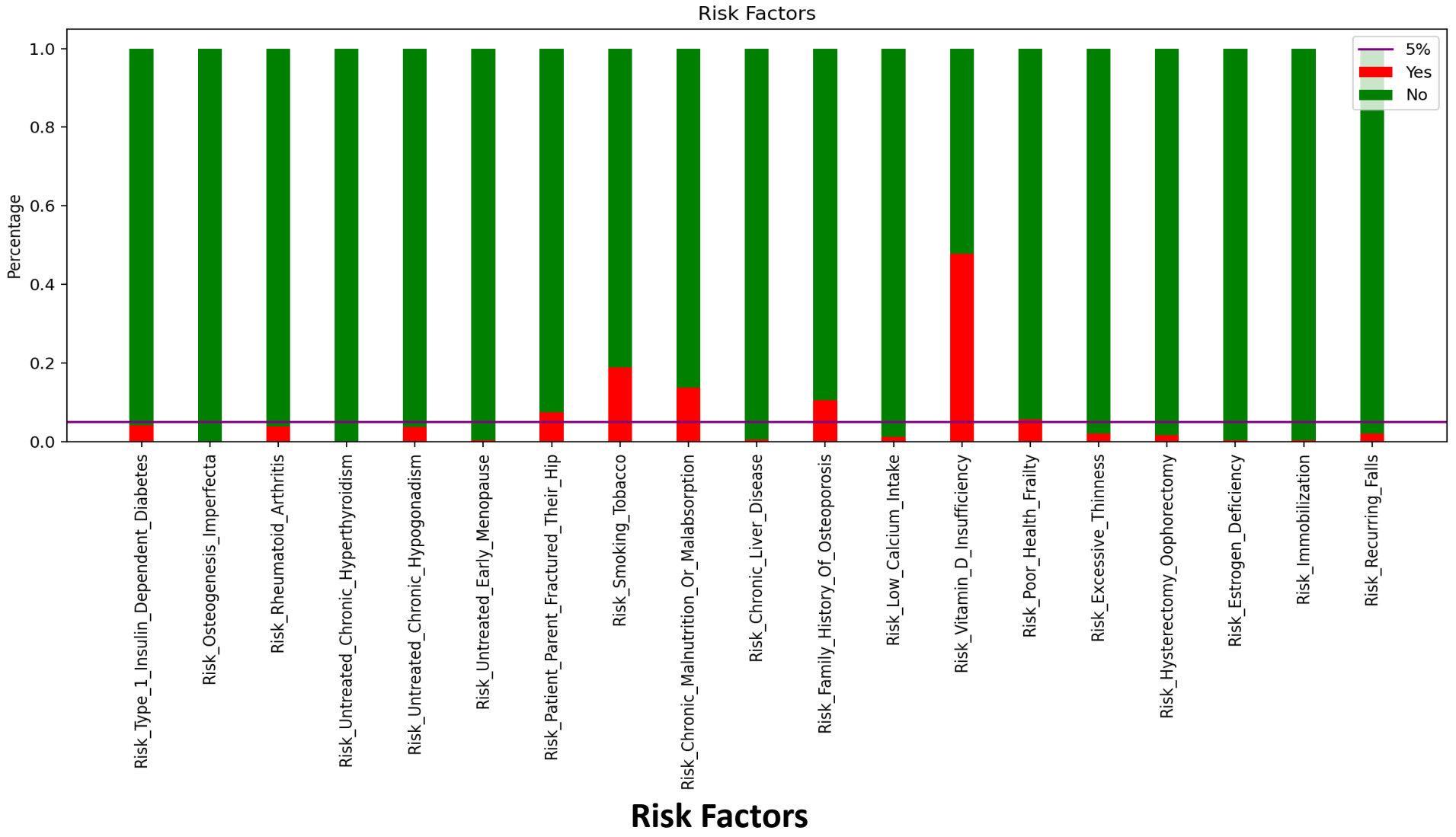
Concomitancy of Drugs

Comorbidity of Diseases

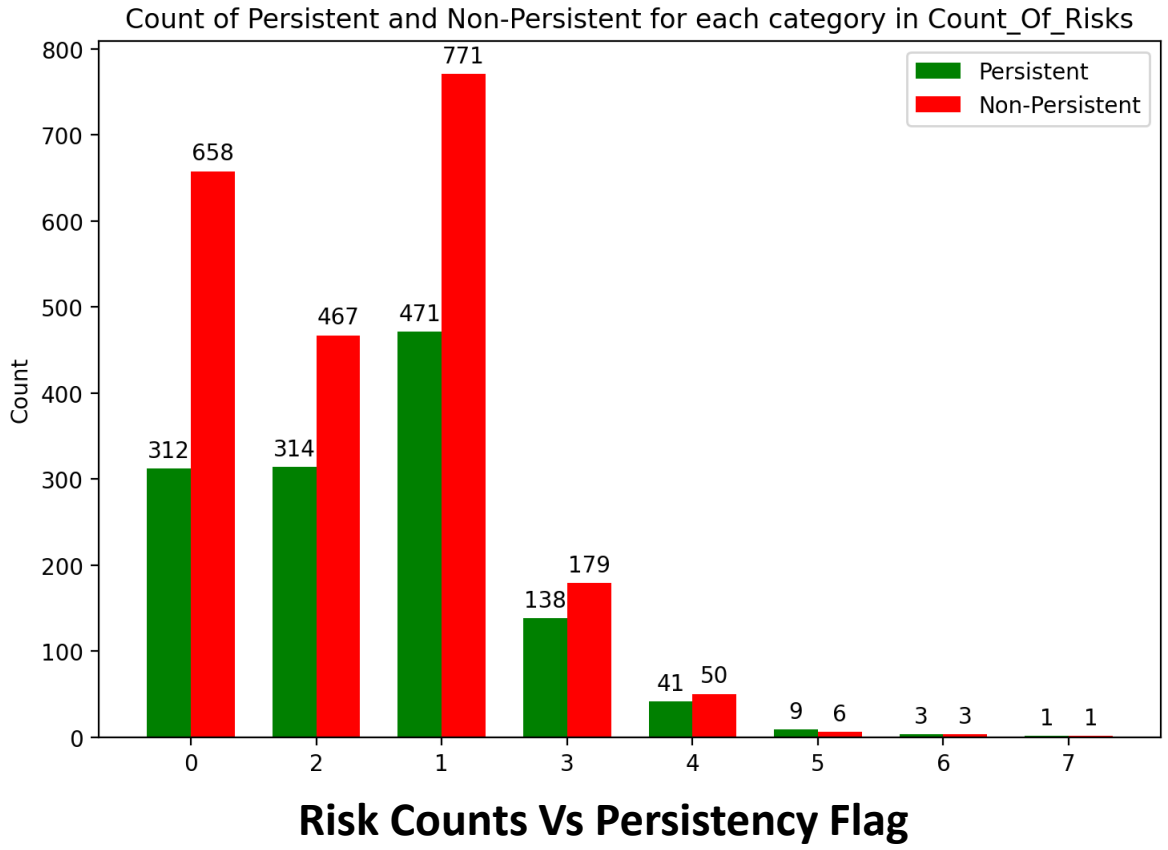


Comorbidity of Diseases

Risk Factors



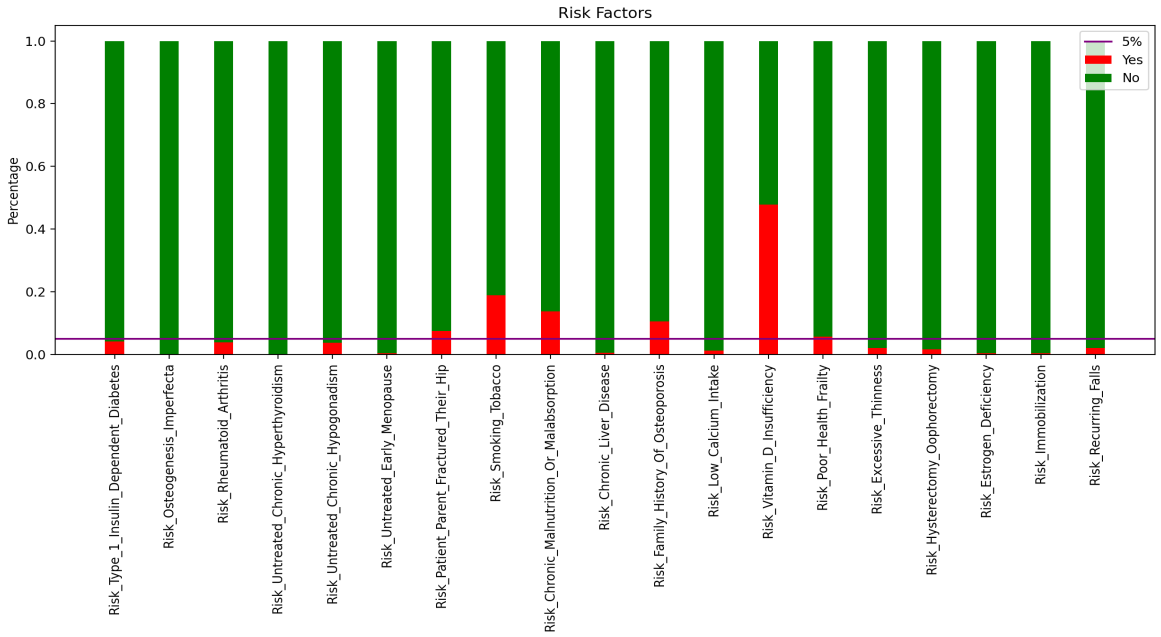
Risk Factor Analysis



Risk Counts Vs Persistency Flag

- Number of non-persistency is higher in lower counts of risks.
- Patients with zero count of risk have the highest Non-Persistent Ration
- Low risk patients were found to be less persistent than the high-risk ones.

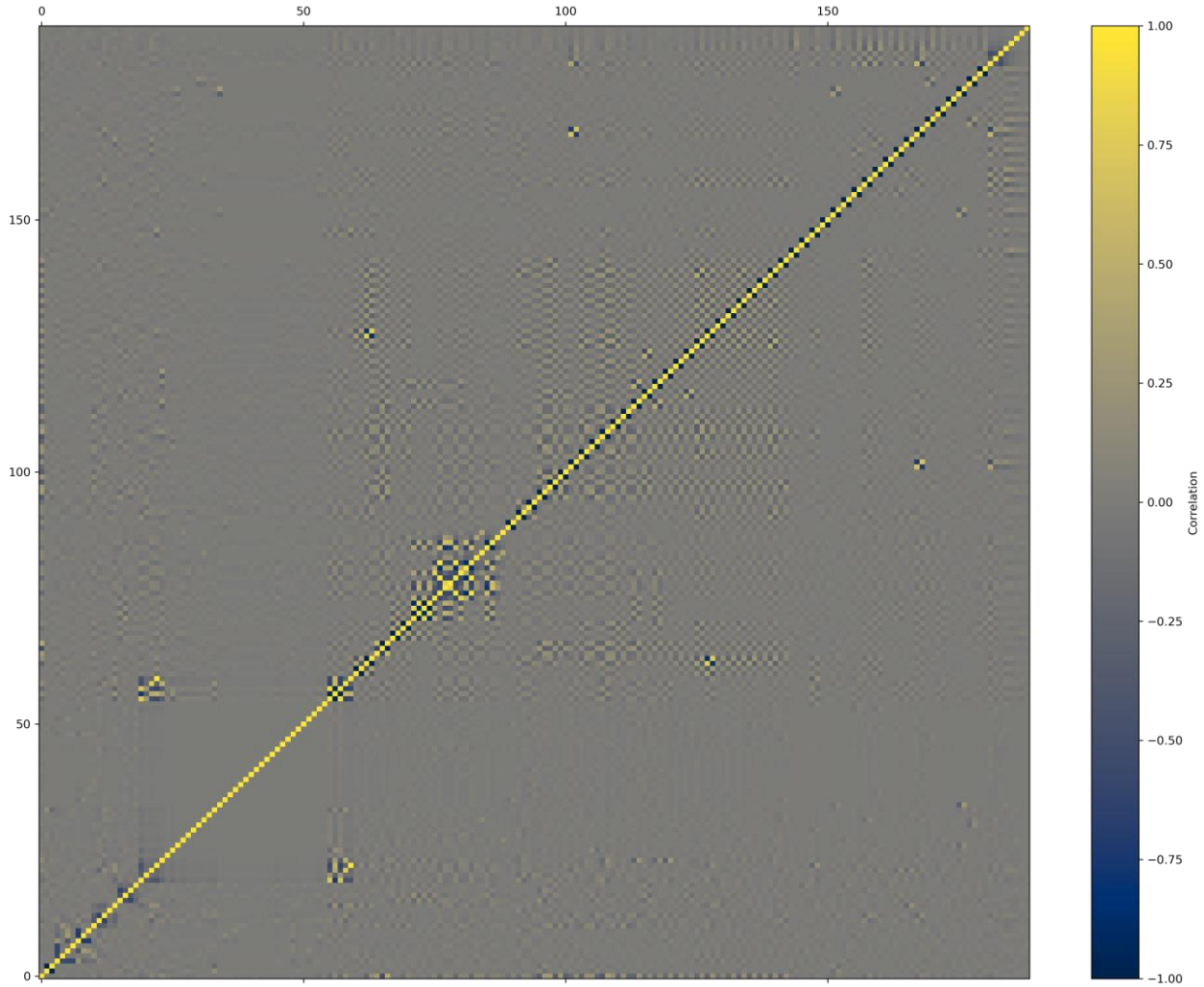
Risk Factors



Risk Factors

- Among the risk factors most of them have less than 5% chance to endanger treatment.
- The risk factor with highest chance is Vitamin D Insufficiency and others above 5% are:
  - Poor Health Frailty
  - Family History Of Osteoporosis
  - Chronic Malnutrition Or Malabsorption
  - Smoking Tobacco
  - Patient Parent Fractured Their Hip
- Rest of the factors have less than 5% risk to endanger treatment.

Features Correlation



Correlation Matrix

- We will be Removing variables with more than 98% correlation.

Recommendations

From the Exploratory Data Analysis (EDA) done on the dataset, we will recommend these instructions:

1. Handling Unknown values for Race, Region, and Ethnicity Variables
    - Using mode as an imputer as an imputer on Race and Ethnicity variables.
    - For Region variable, because most of the people with Unknown Region have Not Hispanic Ethnicity, and Most of people with Not Hispanic Ethnicity, have Midwest Region, we will replace Unknown Regions with Midwest.
2. Handling Rare Labels: Finding categories less than 5 percent in each variable, then merging those categories into one or drop them if the variable only has 2 categories (e.g., Y/N) and cardinality of one them is less than 5 percent.
3. Grouping integer values of Count_Of_Risks variable into two bins: Bin 1 is [0,1,2,3] and Bin 2 is [4,5,6,7].
4. **One hot encoding** all the variables after doing above tasks
5. Removing variables with more than **98% correlation**.

Recommendations

Then Based on previous slide, we recommend these machine learning techniques:

- Gradient Boosting
- Random Forest
- Logistic Regression
- SVC with linear kernel

Thank You