Data Glacier

Your Deep Learning Partner

Week 13: Data Science Healthcare Final Project Report

Name: Serhat Uğur

E-mail: ugur.serhat@outlook.com

Country: Turkey

University: Anadolu University

Specialization: Data Science

GitHub Repo Link: https://github.com/serhatugur/data-science-internship

Table of Contents

Problem Description

One of the challenges for all pharmaceutical companies is to understand the persistence of drugs as per the physician's prescription. To solve this problem, ABC Pharma Company approached an analytics company to automate the identification process.

Business Understanding

“Persistency_Flag” is the main target variable. This variable contains whether the patient is persistent or not. My main objective is to automate the process to find out if the patient will be persistent or not.

Data Understanding

Columns:

1- Unique Row Id

- **Patient ID:** Unique ID of each patient.

2- Target Variable

- **Persistency_Flag:** Flag indicating if a patient was persistent or not.

3- 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.

5- 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 (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 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 prior to 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 Facture 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

6- Disease/Treatment Factor

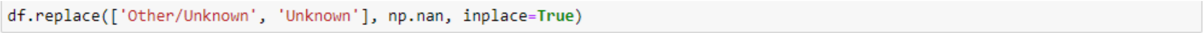
- **NTM - Injectable Experience:** Flag indicating any injectable drug usage in the recent 12 months before the NTM OP Rx
- **NTM - Risk Factors:** Risk Factors that the patient is falling into. For chronic Risk Factors complete lookback to be applied and for non-chronic Risk Factors, one year lookback 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

The original dataset contains 3424 rows and 69 columns. Most of the columns have an object datatype.

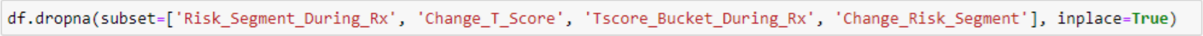
Data Cleaning

1- Cleaning Unknown Values(object)

There are so many unknown values in the columns of Change_Risk_Segment, Change_T_Score, Tscore_Bucket_During_Rx and Risk_Segment_During_Rx. So I will eliminate these unknown values.



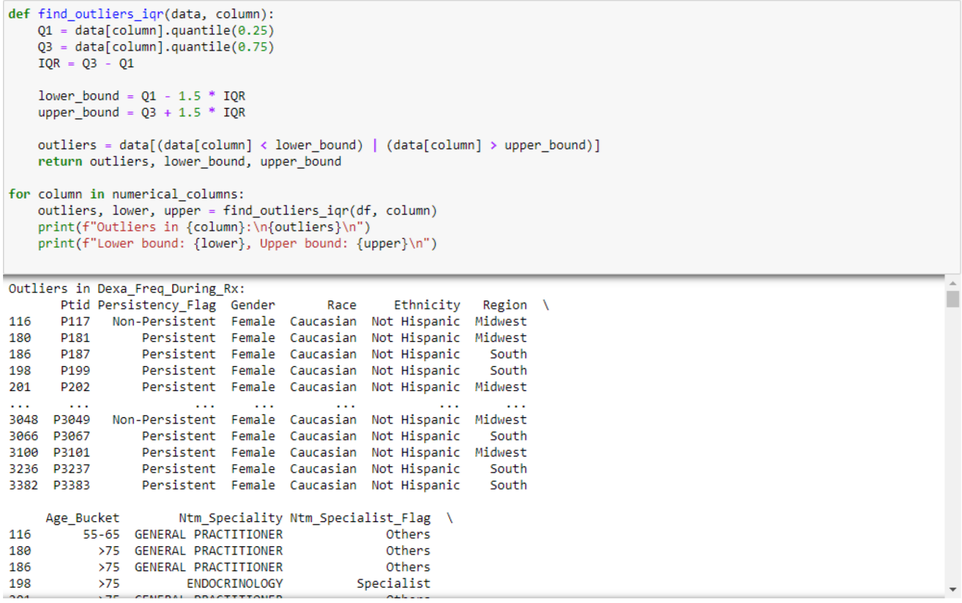
```
df.replace(['Other/Unknown', 'Unknown'], np.nan, inplace=True)
```



```
df.dropna(subset=['Risk_Segment_During_Rx', 'Change_T_Score', 'Tscore_Bucket_During_Rx', 'Change_Risk_Segment'], inplace=True)
```

I converted Other/Unknown and Unknown values to NaN and then dropped the NaN values.

2- Detecting Outliers

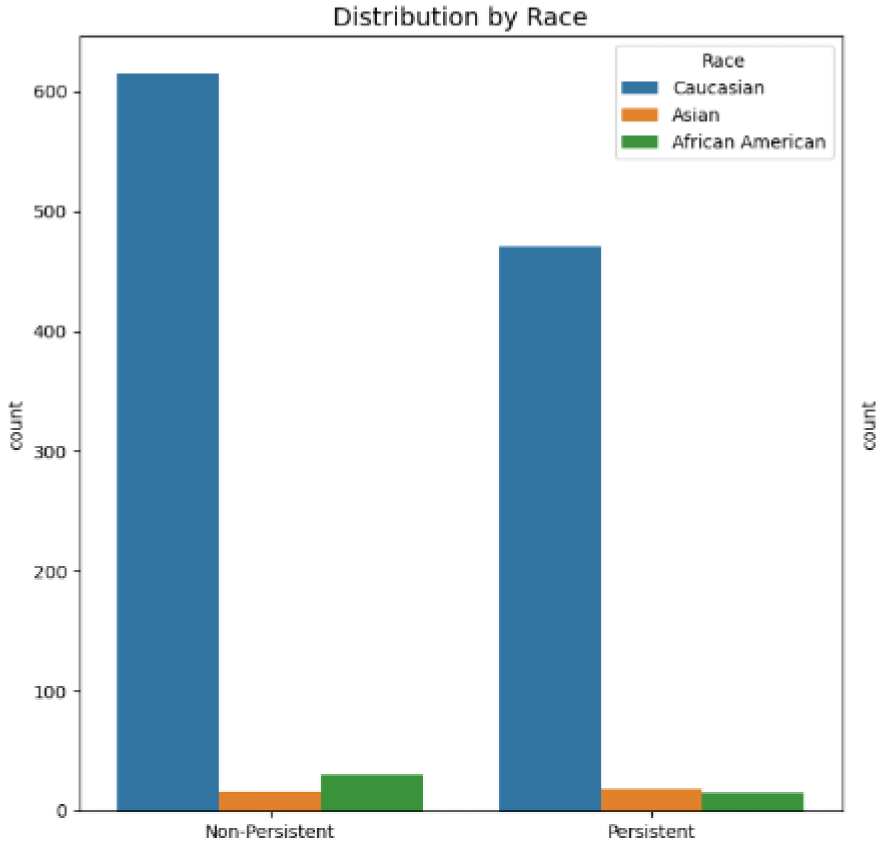


```
def find_outliers_iqr(data, column):
    Q1 = data[column].quantile(0.25)
    Q3 = data[column].quantile(0.75)
    IQR = Q3 - Q1

    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR

    outliers = data[(data[column] < lower_bound) | (data[column] > upper_bound)]
    return outliers, lower_bound, upper_bound

for column in numerical_columns:
    outliers, lower, upper = find_outliers_iqr(df, column)
    print(f"Outliers in {column}:\n{outliers}\n")
    print(f"Lower bound: {lower}, Upper bound: {upper}\n")
```

```
Outliers in Dexa_Freq_During_Rx:
       Ptid Persistency_Flag  Gender       Race     Ethnicity   Region  \
116    P117   Non-Persistent  Female  Caucasian  Not Hispanic  Midwest
180    P181       Persistent  Female  Caucasian  Not Hispanic  Midwest
186    P187       Persistent  Female  Caucasian  Not Hispanic    South
198    P199       Persistent  Female  Caucasian  Not Hispanic    South
201    P202       Persistent  Female  Caucasian  Not Hispanic  Midwest
...     ...              ...     ...        ...           ...      ...
3048  P3049   Non-Persistent  Female  Caucasian  Not Hispanic  Midwest
3066  P3067       Persistent  Female  Caucasian  Not Hispanic    South
3100  P3101       Persistent  Female  Caucasian  Not Hispanic  Midwest
3236  P3237       Persistent  Female  Caucasian  Not Hispanic    South
3382  P3383       Persistent  Female  Caucasian  Not Hispanic    South

     Age_Bucket        Ntm_Speciality Ntm_Specialist_Flag  \
116       55-65  GENERAL PRACTITIONER              Others
180         >75  GENERAL PRACTITIONER              Others
186         >75  GENERAL PRACTITIONER              Others
198         >75         ENDOCRINOLOGY          Specialist
```

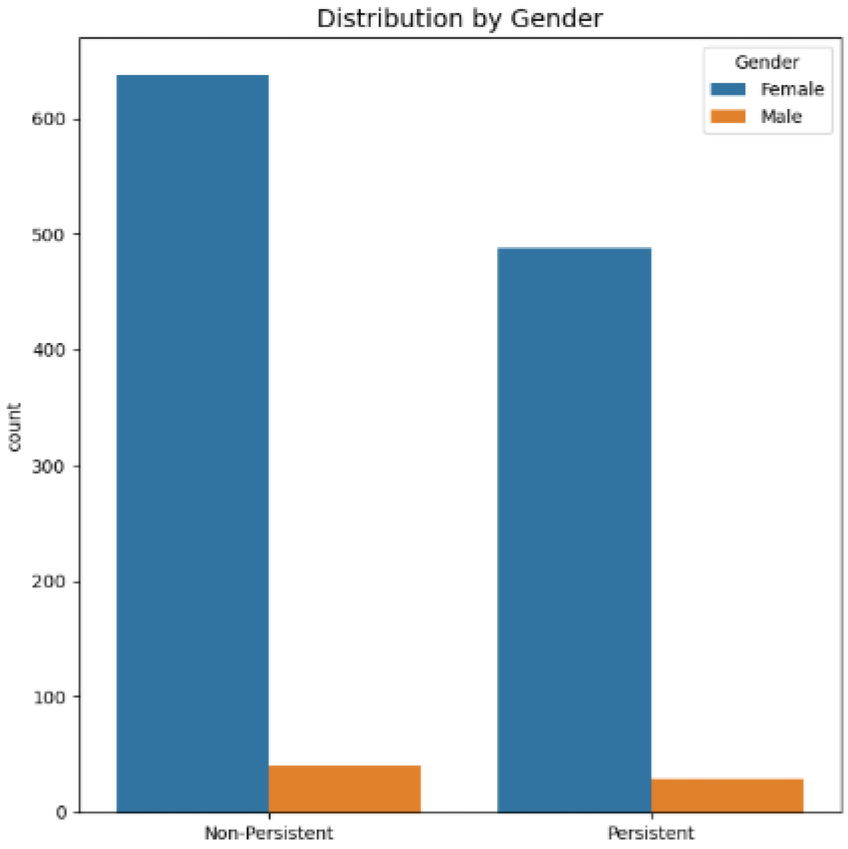
Exploratory Data Analysis

1) Distribution by Race



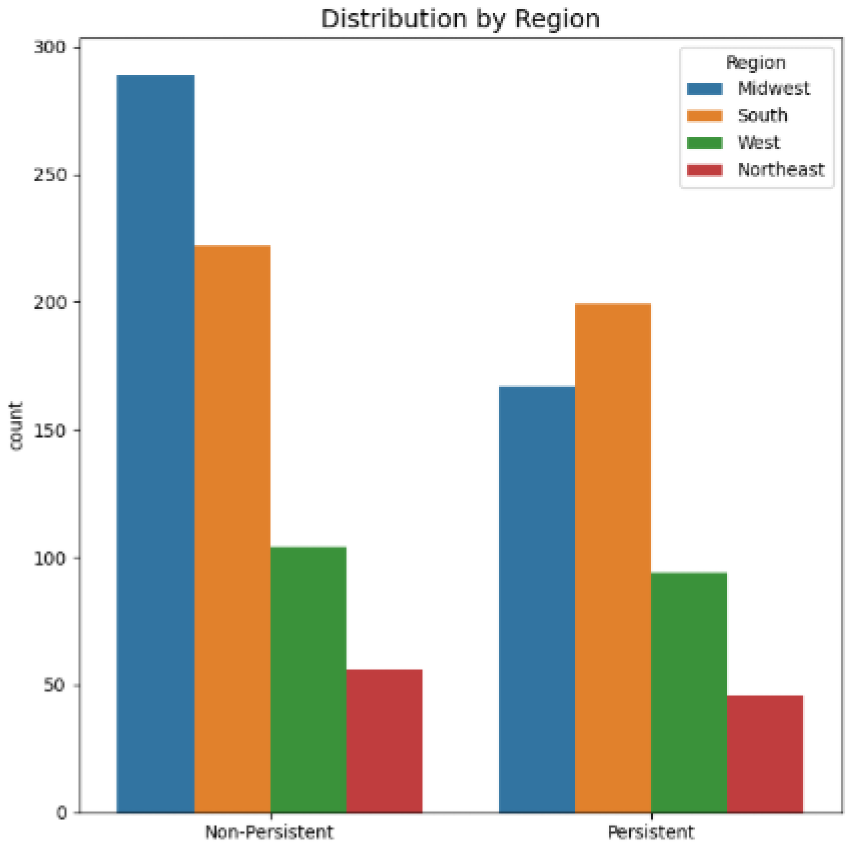
Caucasian is the most common race in the study.

2) Distribution by Gender



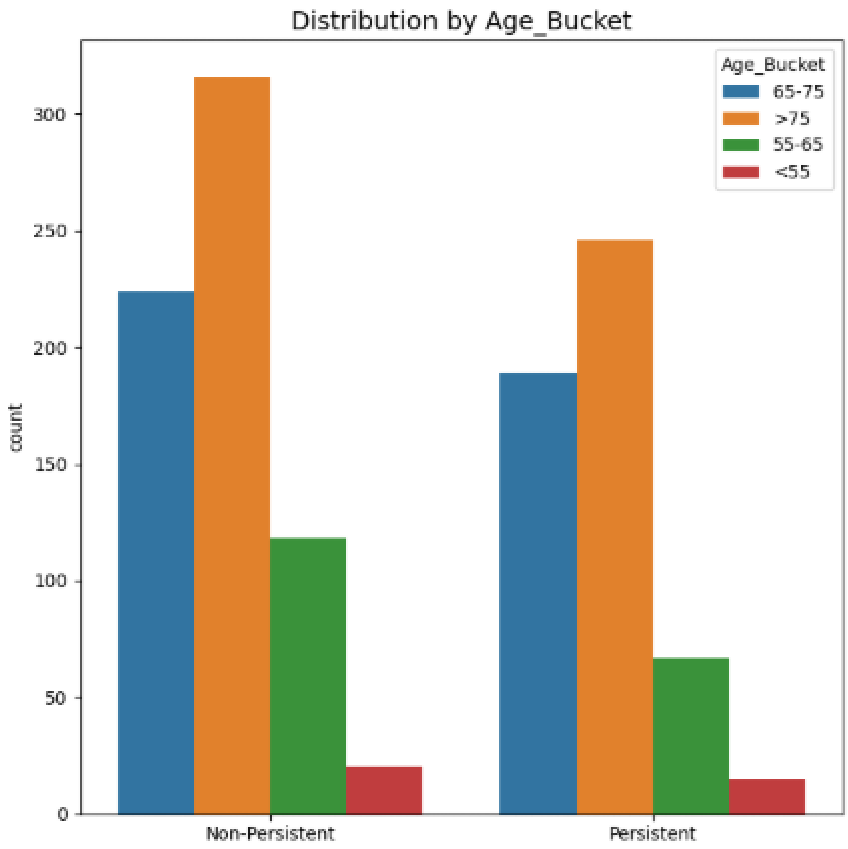
In both groups, females number much more than the number of men.

3) Distribution by Region



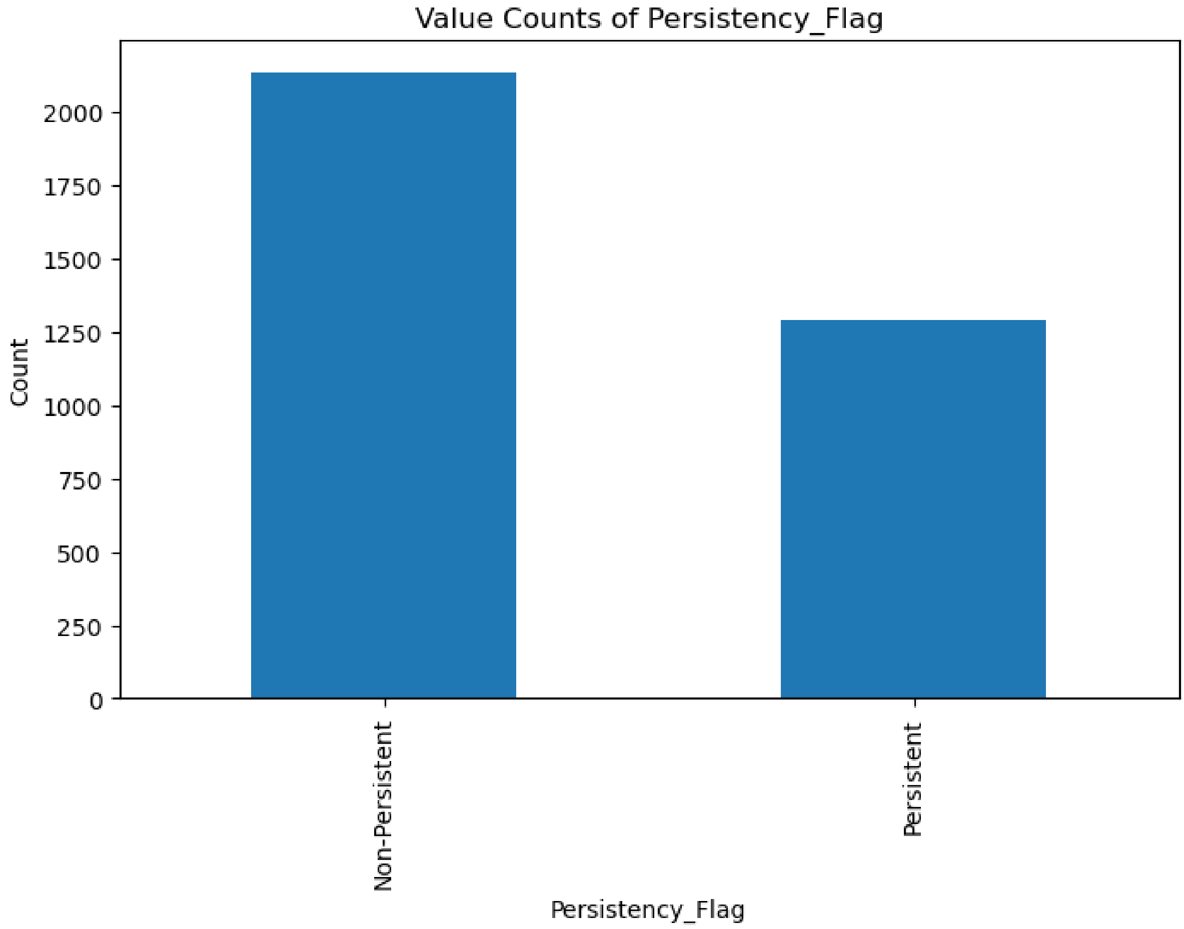
In both groups, the Midwest and South regions have more patients than the other regions.

4) Distribution by Age



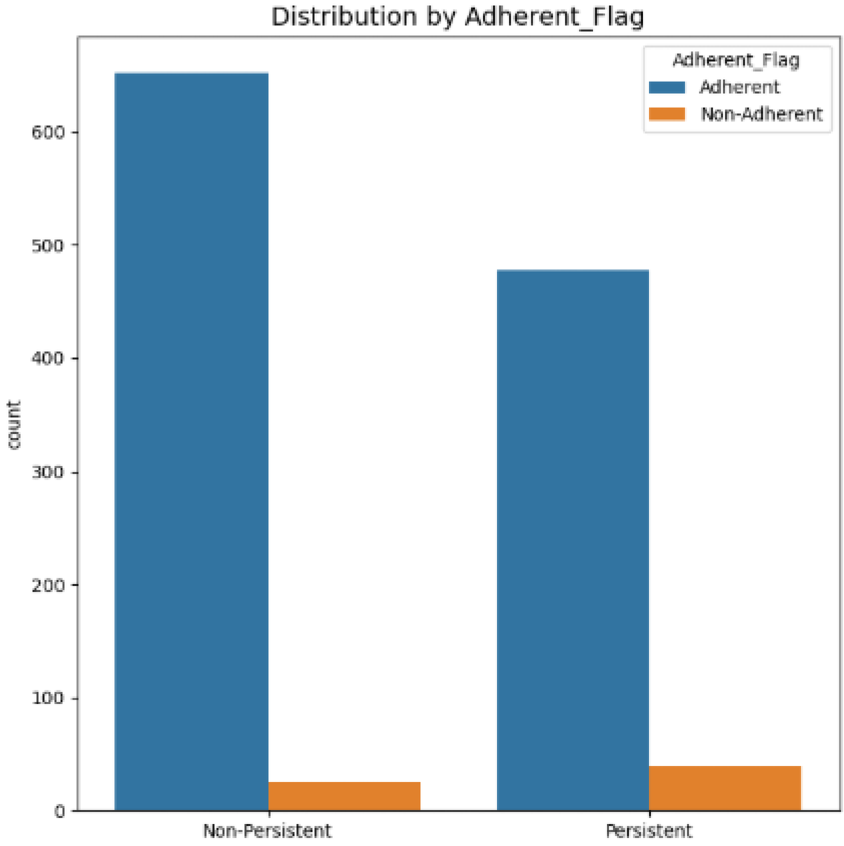
In both groups, 65+ age patients count higher than the others.

5) Non-persistent vs. Persistent



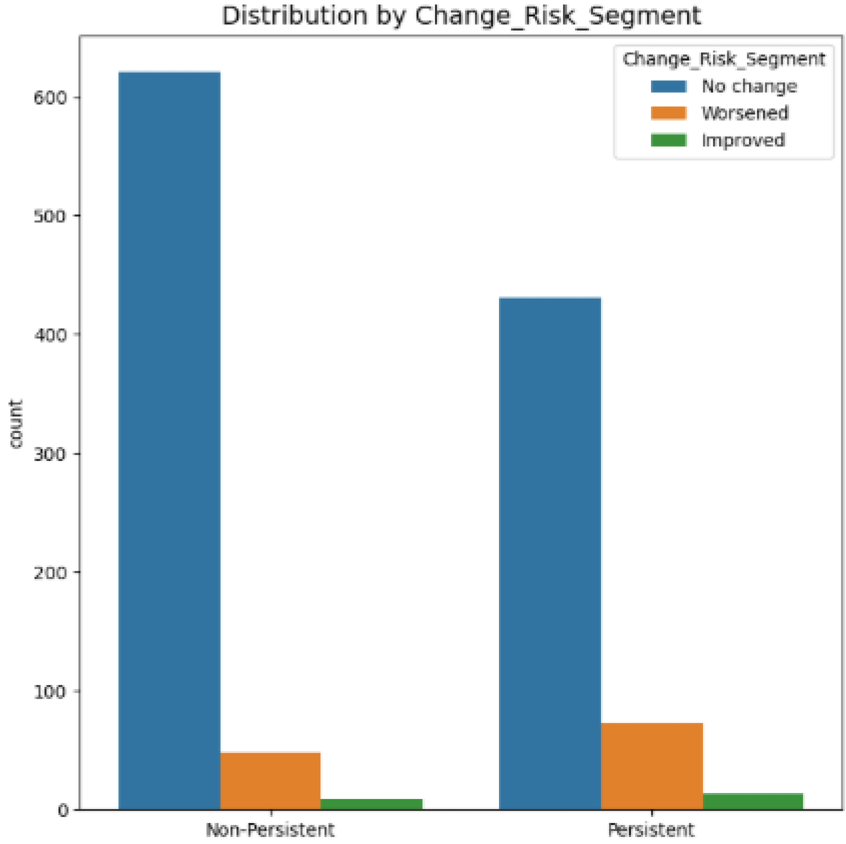
There are more non-persistent patients than the persistent group.

6) Distribution by Adherent_Flag



Most of the patients are following their medical advice.

7) Distribution by Change_Risk_Segment



Mostly, there are no big risk changes for both groups. However, persistent groups' risks tend to be higher compared to non-persistent patients.

Final Recommendation

The main goal of this analysis is to predict the persistence of patients as accurately as possible and provide actionable recommendations for improvement. The following steps outline the approach taken and the proposed next steps:

1-Model Development and Validation

This project will implement several machine learning models, such as logistic regression, random forest, and gradient boosting, to find reliable predictions.

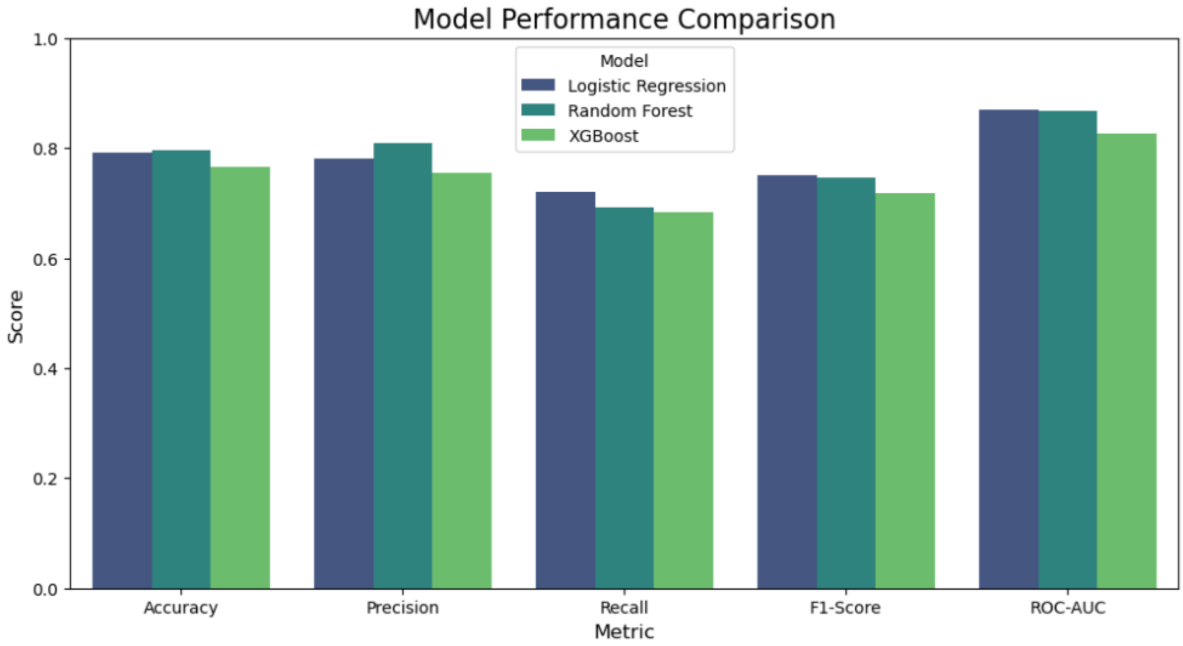
2-Model Evaluation

Key performance metrics such as accuracy, precision, recall, F1-score, and AUC-ROC will be used for model comparisons. The most effective model will be chosen by balancing predictive power and interpretability.

Model Selection and Building

For this project, I used Logistic Regression, Random Forest and XGBoost.

Model Performance Comparison



After doing cross-validation, we can choose logistic regression for our project.