Week 13: Data Science Healthcare Final Presentation

- **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

Agenda

- Problem Description
- Business Understanding
- Data Understanding
- Exploratory Data Analysis
- Proposed Modeling Technique
- Model Selection and Building

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

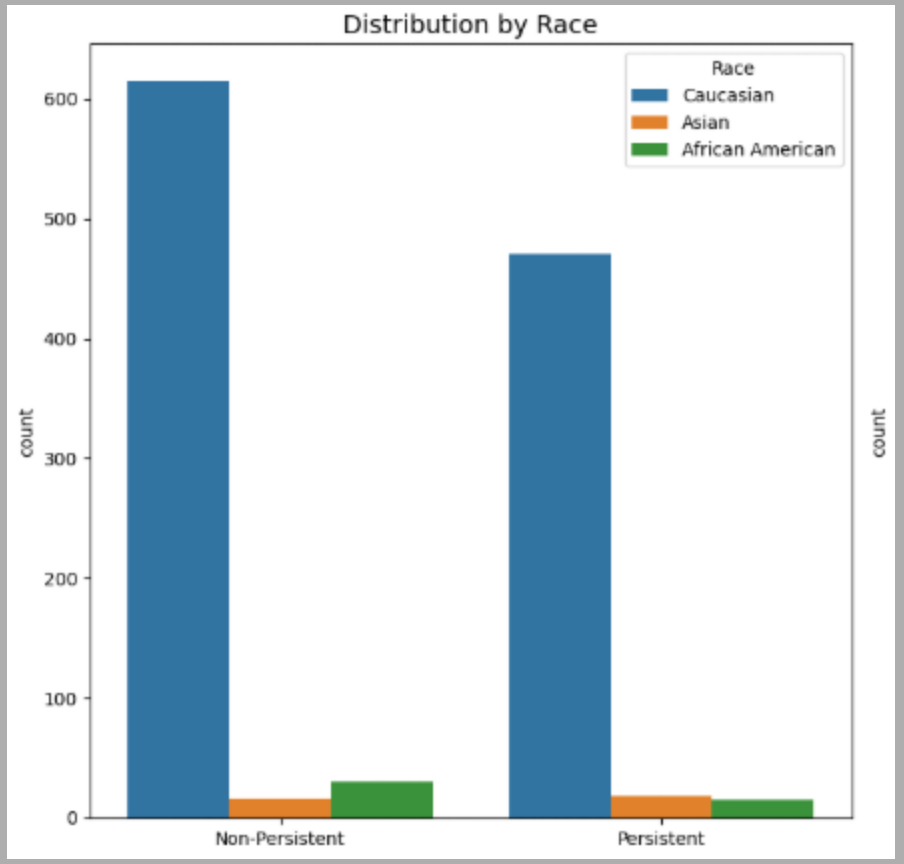
- The original dataset contains 3424 rows and 69 columns. Most of the columns have an object datatype.

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	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 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 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 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

EXPLORATORY DATA ANALYSIS

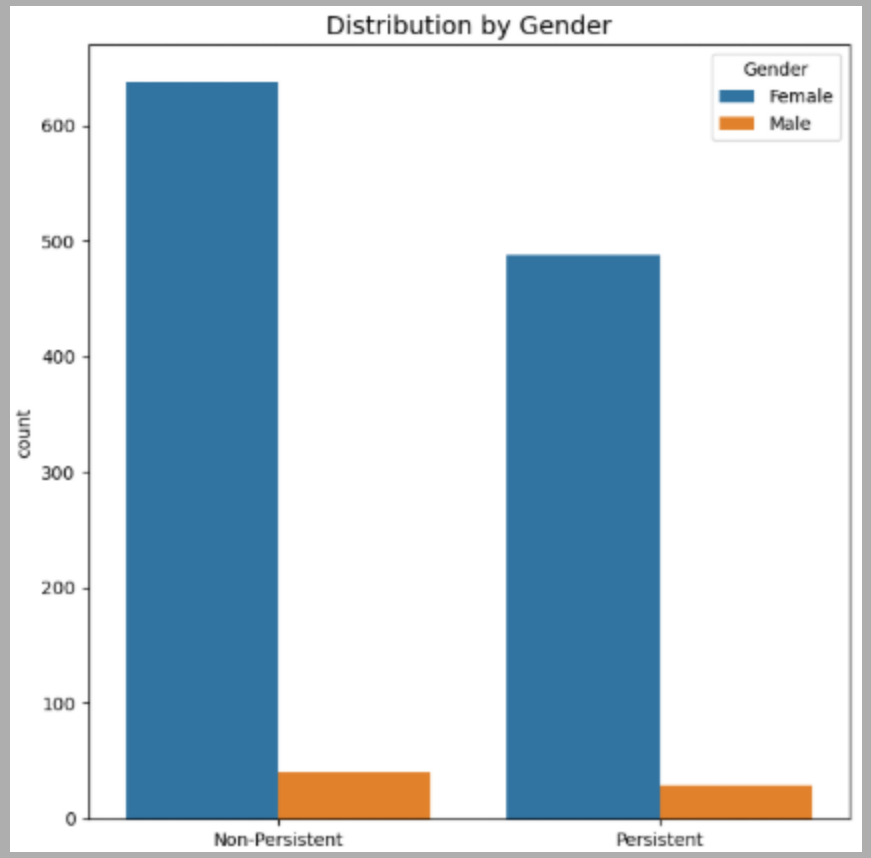
Distribution by Race

Caucasian is the most common race in the study.



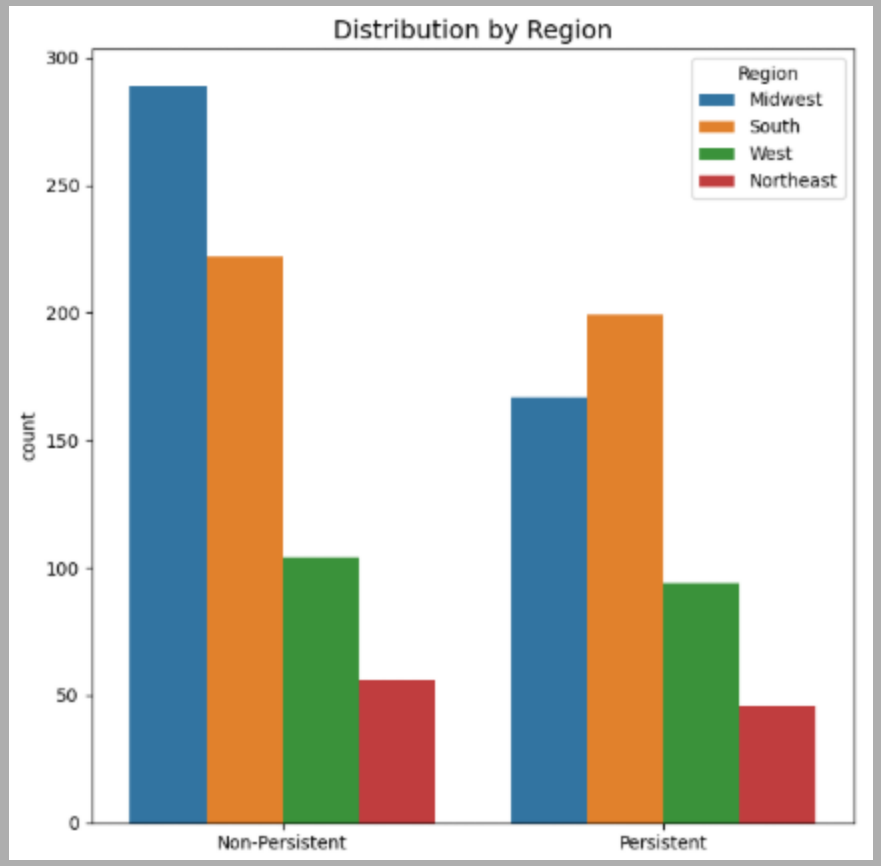
Distribution by Gender

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



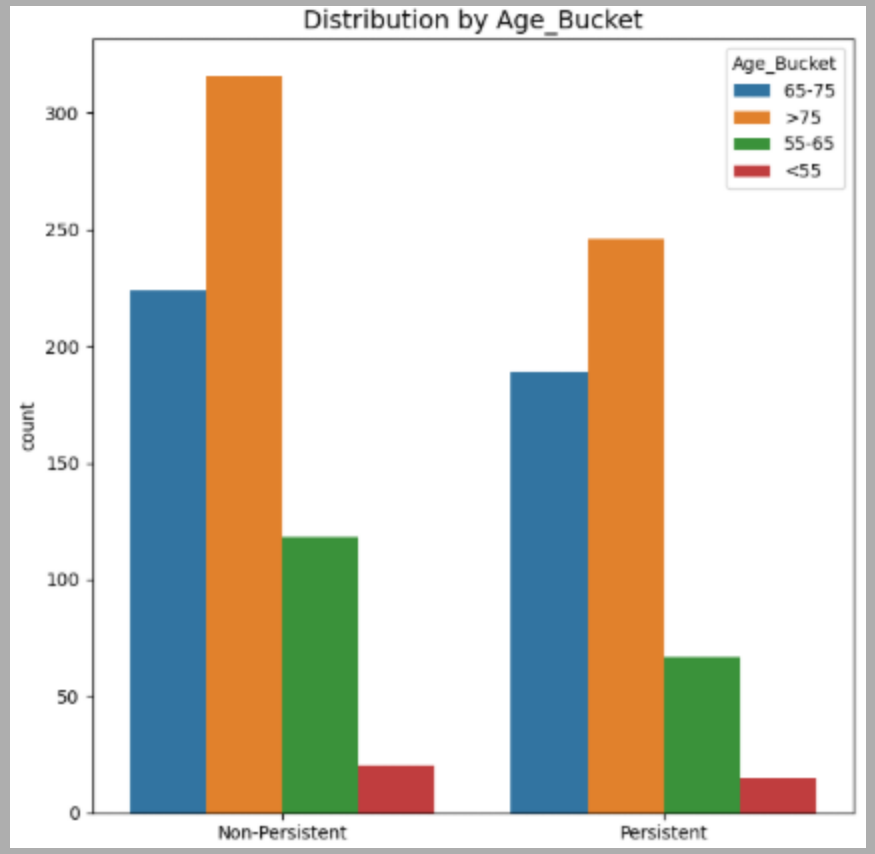
Distribution by Region

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



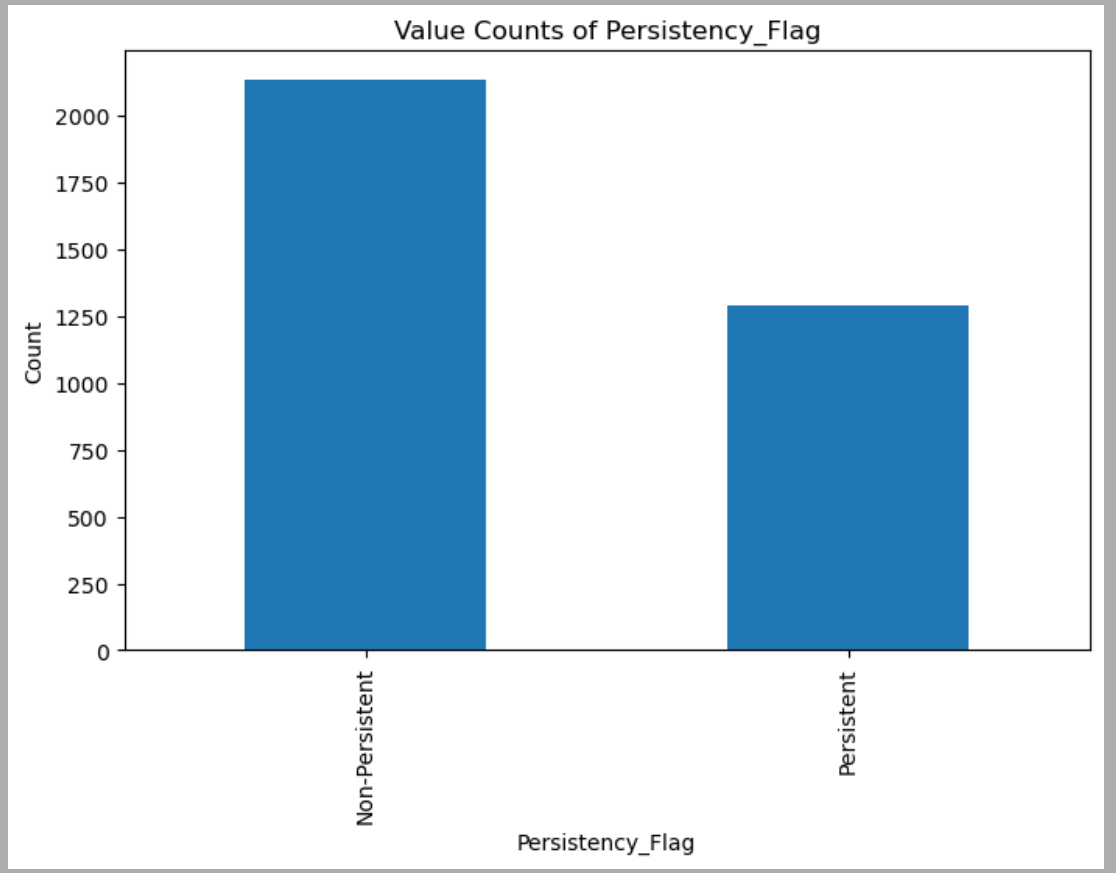
Distribution by Age

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



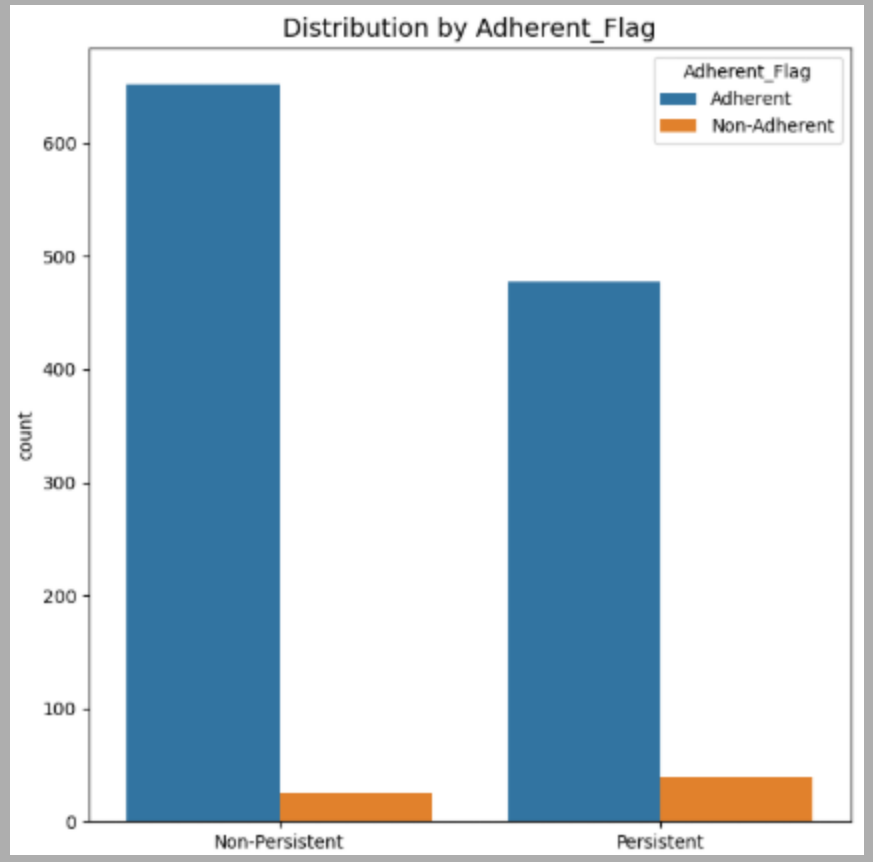
Non-persistent vs. Persistent

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



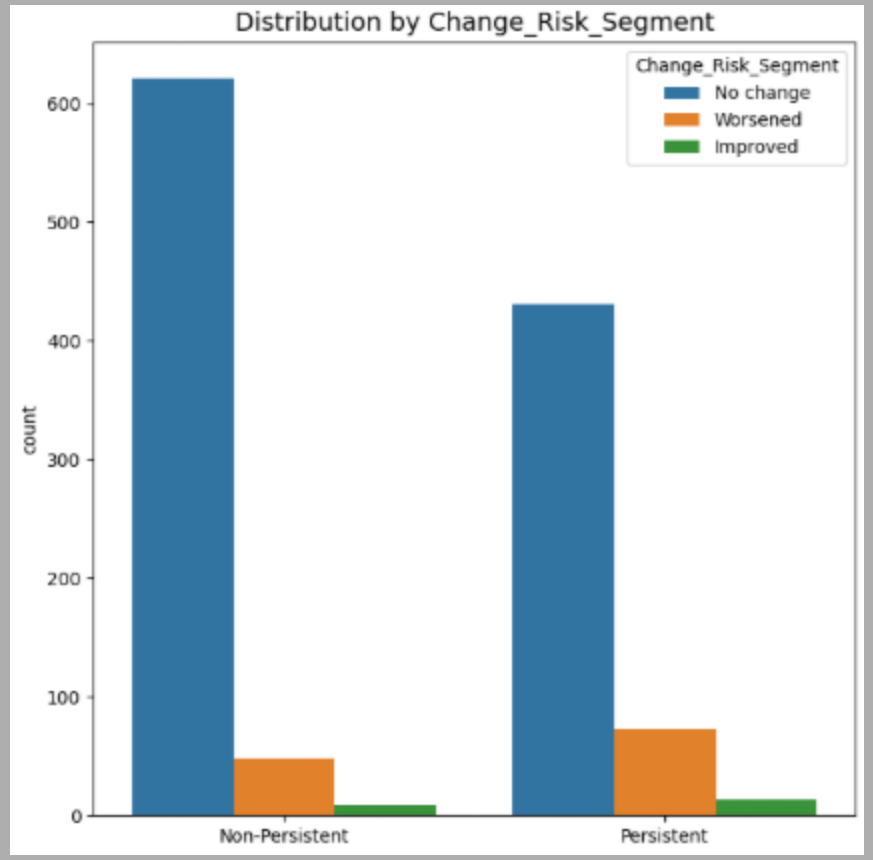
Distribution by Adherent_Flag

Most of the patients are following their medical advice.



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.

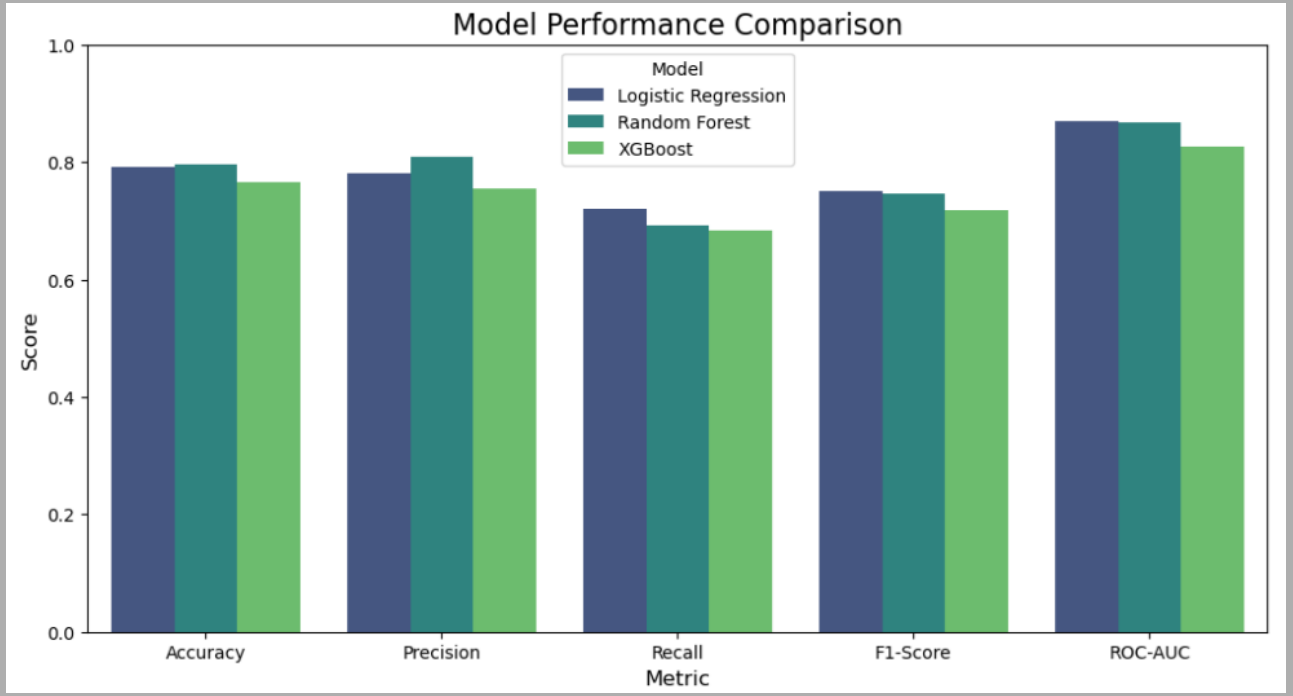


Proposed Modeling Technique

- 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:
- **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.
- **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. However, after doing cross-validation, we can choose logistic regression for our project.



Thank You for Your Time.