Project: Healthcare - Persistency of a drug

Week 13: Deliverables

Name: Krishna Ratna Deepika Haripuram

Email: haripuramdeepika@gmail.com

Country: Canada

Batch Code: LISUM26

Specialization: Data Science

Submission Date: Dec 16th, 2023

Submitted to: Data Glacier

(Individual project)

Table of Contents

Github link
https://github.com/krishnaharipuram/Data-Glacier/tree/main/Week%2013

1. Problem Description

The project focuses on a critical and complex challenge faced by pharmaceutical companies worldwide: understanding and predicting the persistency of drug usage as per physician prescriptions. This challenge is not just about assessing whether patients are taking their medications as prescribed but also involves a deep dive into the myriad factors that influence this behavior. The persistency of drug usage is a multifaceted issue, encompassing aspects such as patient demographics, clinical history, medication characteristics, and socio-economic factors.

In the healthcare sector, especially in chronic disease management, the effectiveness of treatment is heavily dependent on patients consistently following their prescribed medication regimens. However, non-persistence to medication is a common issue, leading to suboptimal treatment outcomes, increased healthcare costs, and heightened risk of disease complications.

This project aims to tackle this issue by leveraging data analytics to identify and analyze patterns and factors that contribute to drug persistency. Through a comprehensive analysis of patient data, the project seeks to develop predictive models that can accurately forecast medication adherence, thereby offering valuable insights into patient behavior.

2. Business Understanding

Relevance

In the pharmaceutical industry, understanding drug persistency is crucial, as it directly affects patient health outcomes and drives key business metrics. Persistency rates are indicative of medication effectiveness, patient satisfaction, and overall treatment success. High persistency rates are often correlated with better health outcomes, reduced hospitalization rates, and lower healthcare costs. From a business perspective, insights into drug persistency are invaluable. They enable pharmaceutical companies to fine-tune their marketing strategies, tailor patient support programs, and make informed decisions about drug development and formulation. In essence, understanding persistency helps bridge the gap between patient needs and the treatments offered.

Objective

The primary objective of this project is to automate the process of identifying and analyzing the factors that impact drug persistency. By utilizing machine learning and data analytics, the project aims to build a robust classification model that can sift through vast amounts of patient data to highlight key predictors of medication adherence.

This automated process is not just about simplifying data analysis; it's about enabling more accurate, data-driven decision-making. The insights gleaned from this analysis can guide pharmaceutical companies in developing more patient-centric strategies, optimizing treatment plans, and ultimately contributing to better healthcare outcomes.

Through this project, we seek to provide a tool that empowers pharmaceutical companies to gain a deeper understanding of their patient base, tailor their approaches to meet patient needs, and improve the overall effectiveness of their treatments.

3. Project Lifecycle

Week	Date	Plan
Week 7	19th Nov 2023	Problem statement, business understanding and project lifecycle with deadline
Week 8	26th Nov 2023	Problem description, Data understanding and identifying approaches to overcome problems like missing data, outliers etc.
Week 9	2nd Dec 2023	Data cleaning and transformation
Week 10	9th Dec 2023	EDA and model recommendation
Week 11	16th Dec 2023	Presentation on EDA and proposed model technique
Week 12	23rd Dec 2023	Model Selection and Model Building/Dashboard
Week 13	30th Dec 2023	Final project report and code submission

4. Data Understanding

The dataset for this project is a rich amalgamation of patient demographic details and clinical factors. It includes various attributes like age, gender, race, and specific medical history details, along with treatment adherence indicators. The data is comprehensive, covering aspects ranging from patient demographics to intricate clinical parameters and provider attributes.

a.Type of Data for Analysis

The dataset is a diverse mix of:

- **Categorical Data**: This includes variables like gender, race, ethnicity, and treatment types, providing a qualitative assessment of patient profiles and clinical scenarios.

- **Numerical Data**: It encompasses quantitative variables such as the frequencyof DEXA scans and the count of risk factors, offering measurable insights into patient treatment and health risks.

b. Problems in the Data

The dataset presents several challenges:

- **Missing Values**: The dataset showed no significant missing values, indicating completeness in data recording.

- **Outliers and Skewness:** Notable outliers and skewness were observed in key numerical columns like **`Dexa_Freq_During_Rx`** and **`Count_of_Risks`**, which could potentially skew statistical analyses.

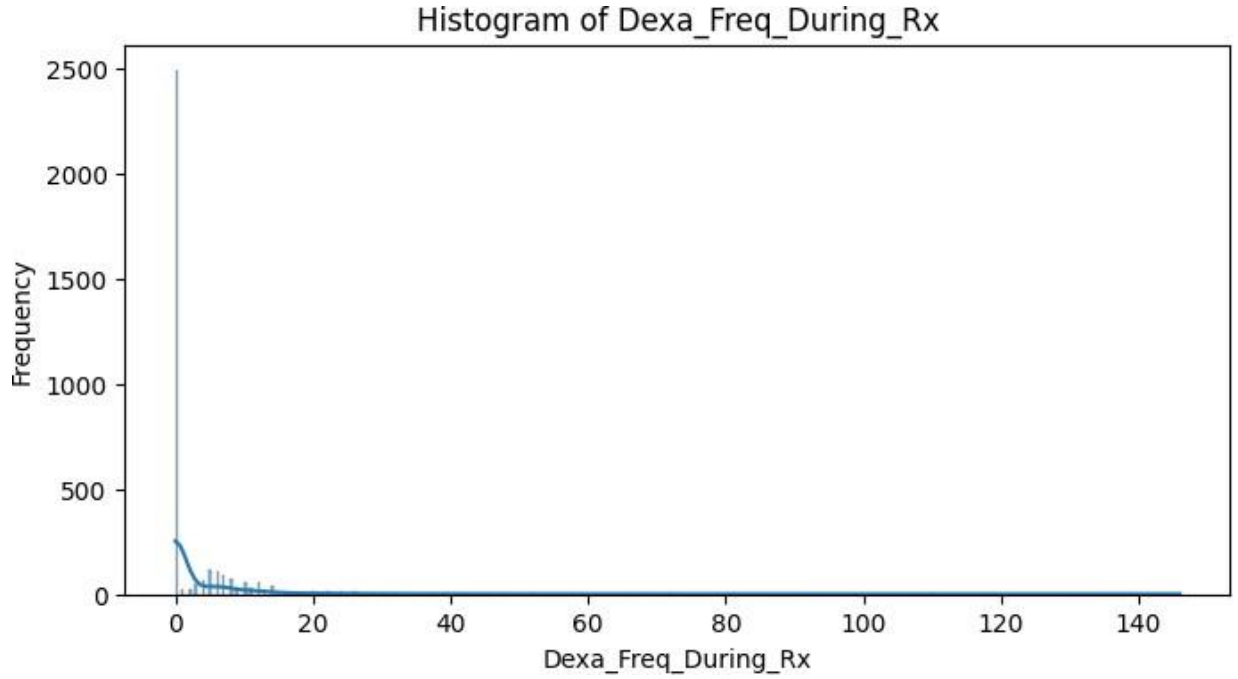
c. Approaches to Overcome Problems

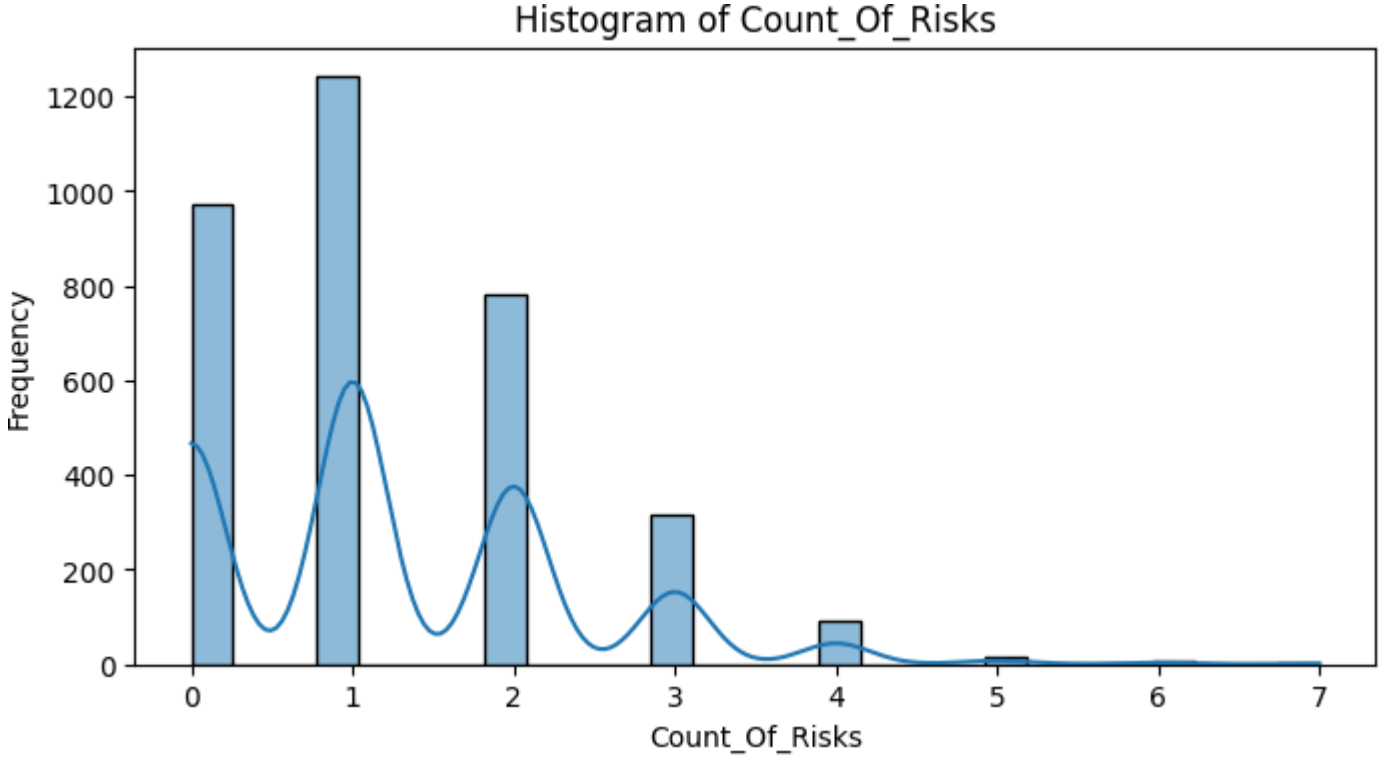
Handling Skewness: Log transformations were applied to skewed numerical columns to normalize their distribution, enhancing their suitability for statistical models and machine learning algorithms.

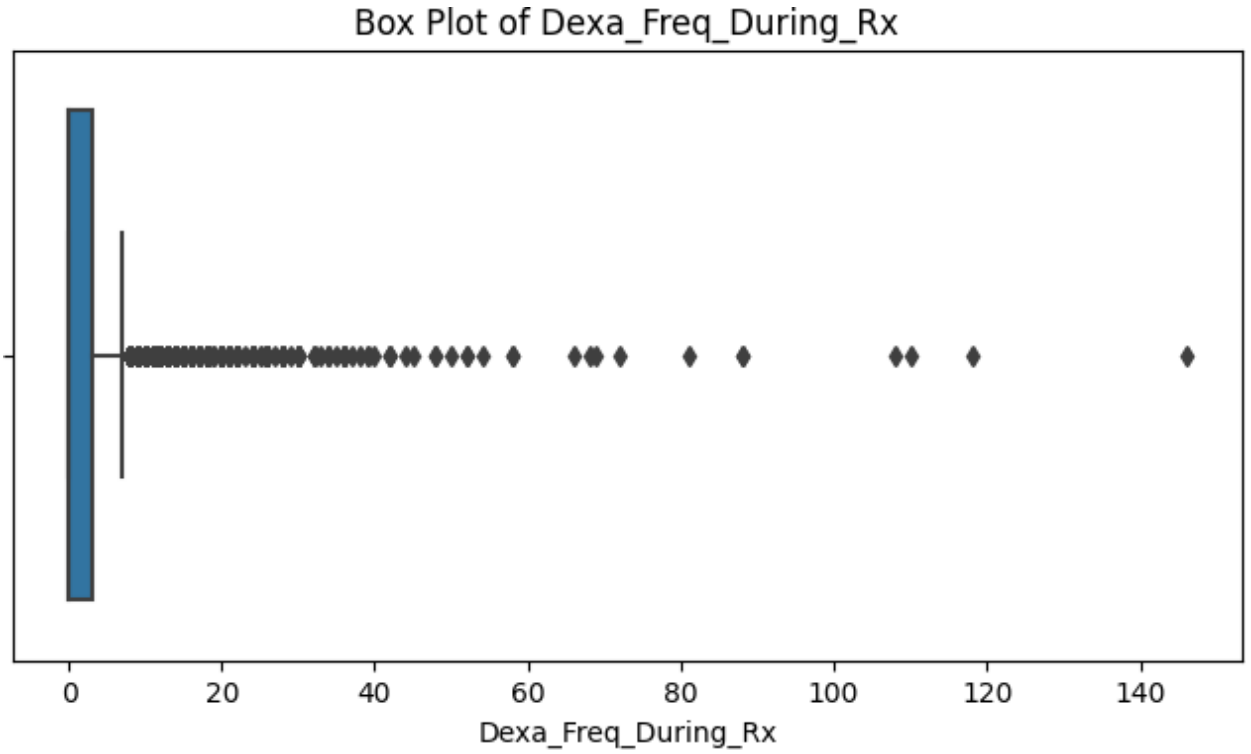
Managing Outliers: Techniques like the Interquartile Range (IQR) method were

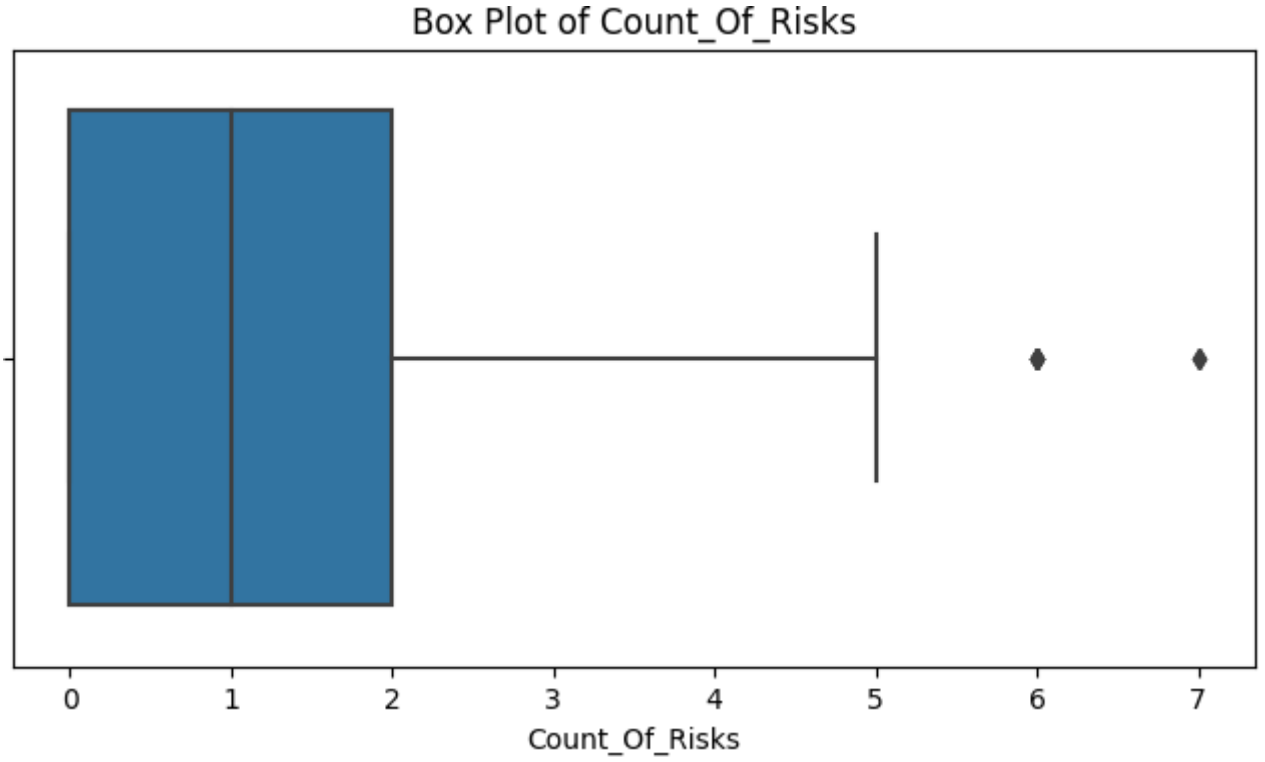
employed for identifying and treating outliers, thus normalizing the data, and ensuring robustness in analysis.

Data Visualization: Histograms and box plots were utilized to visually assess the distribution of numerical columns, aiding in the identification and interpretation of skewness and outliers.









5. Data Cleaning and Transformation:

a. Handling Missing Values:

- Applied various techniques like mean, median, and mode imputation to handle missing values in different columns.
- Experimented with K-Nearest Neighbors (KNN) imputation as a model-based approach for a comprehensive understanding of handling missing data in mixed-type datasets.

b. Outlier Detection:

- Identified and handled outliers using methods like the Interquartile Range (IQR) for 'Dexa_Freq_During_Rx' and Z-score for 'Count_Of_Risks'.
- These techniques helped in normalizing the data distribution and removing anomalies that could potentially skew the analysis.

c. Weight of Evidence (WoE) Calculation:

- Computed WoE for categorical variables to transform them into a continuous scale and gain insights into the predictive power of each category.

d. NLP Featurization and Data Cleaning:

- Utilized regex for cleaning text data in columns such as 'Ntm_Speciality' and 'Risk_Segment_Prior_Ntm', removing non-alphabetic characters and standardizing the text.
- Applied TF-IDF Vectorization on the 'Ntm_Speciality' column to convert the cleaned text data into a numerical format. This transformation is crucial for feeding textual data into machine learning models.
- The sparse nature of the TF-IDF matrix (mostly zeros) is typical and reflects the uniqueness of terms in the documents.

Review and Reflections:

In this project, I navigated challenges such as choosing suitable data cleaning methods and handling the sparsity in the TF-IDF matrix. These experiences underscored the importance of understanding data at a deep level and maintaining data integrity through careful documentation. This project reinforced the value of adaptability and critical analysis in data science, lessons that I will apply in my future endeavors.

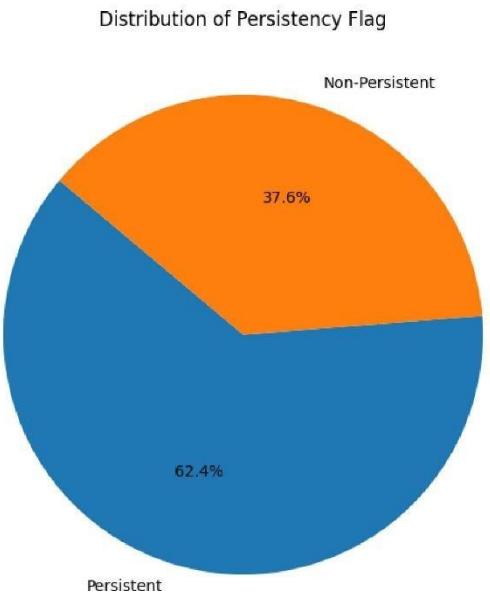
Code Repository and Documentation:

- All code and detailed documentation are maintained in the provided GitHub repository. Regular updates and comprehensive commenting in the code have been a priority to ensure clarity and reproducibility of the analysis.

6. EDA performed on the data

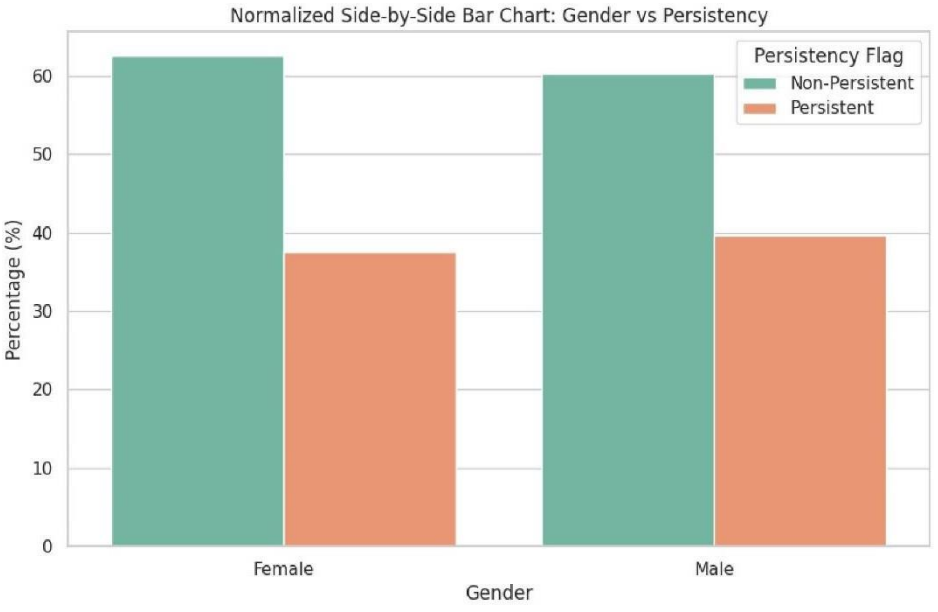
a. Distribution of Persistency Flag

The pie chart depicts the overall distribution of the persistency flag among patients. A significant majority show persistence, but there remains a considerable portion that is non-persistent, highlighting the need for targeted interventions.



b. Normalized Side-by-Side Bar Chart: Gender vs Persistency

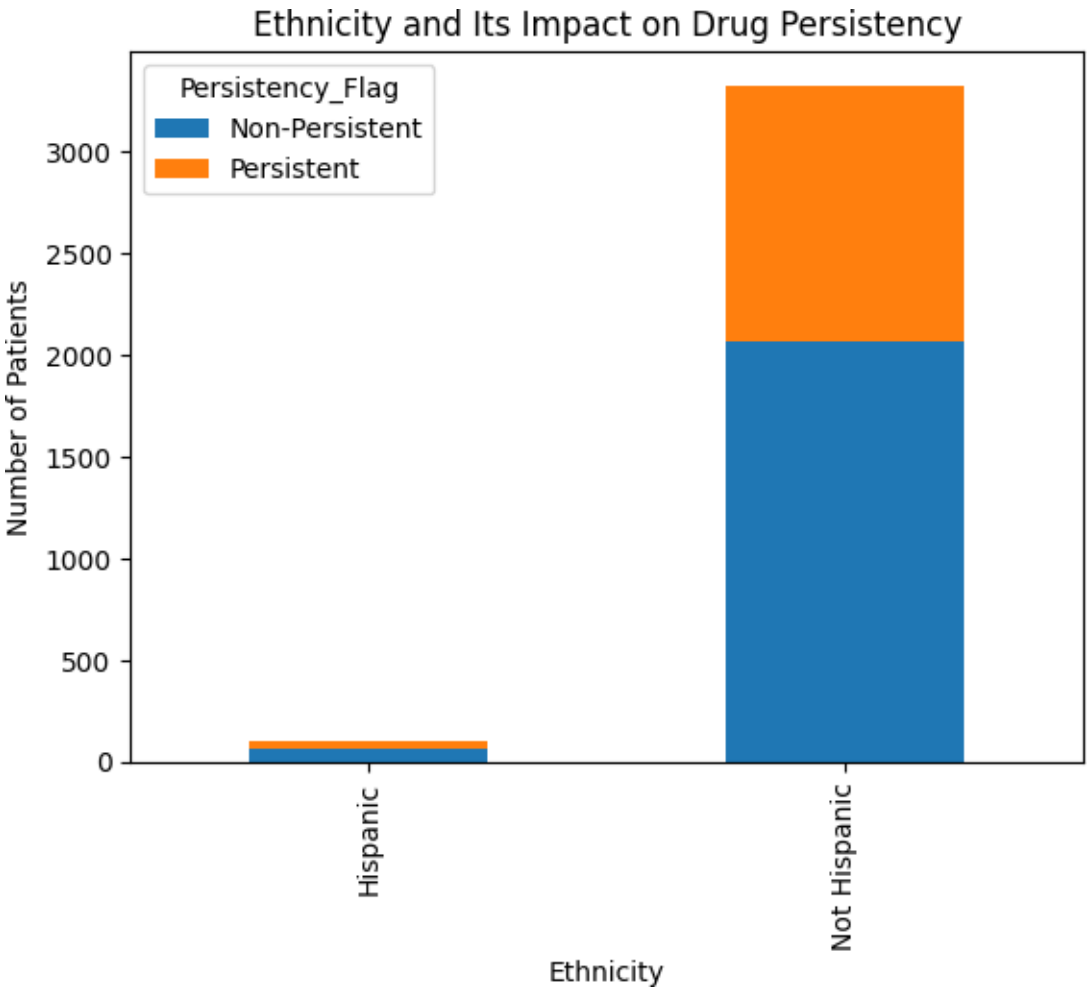
This normalized bar chart compares drug persistency rates between genders. While there are slight differences, both genders show substantial non-persistent populations, indicating a need for gender-specific adherence strategies.



c. Ethnicity and Its Impact on Drug Persistency

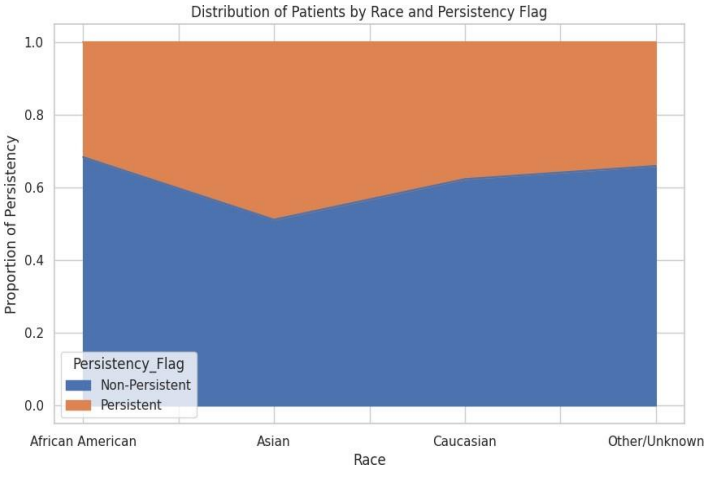
The bar chart explores how ethnicity affects drug persistency, with notable differences between ethnic groups.

Understanding cultural and social factors could be key to improving patient-specific treatment plans.



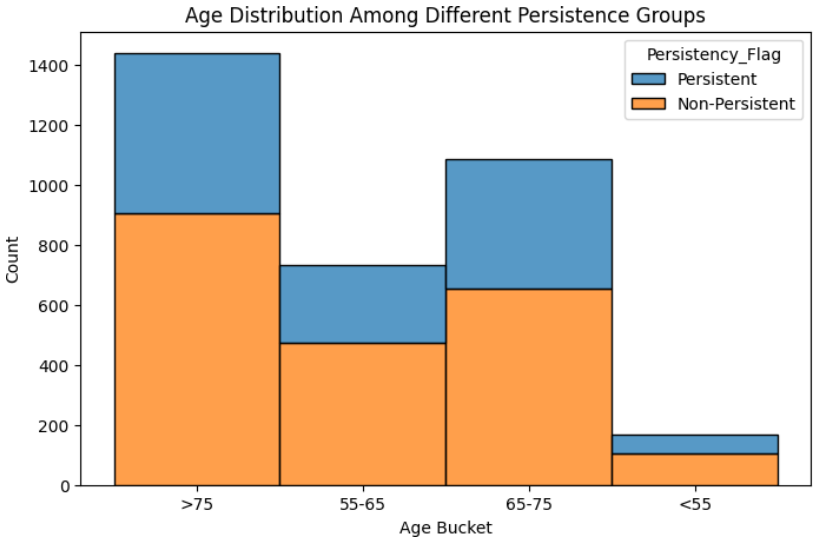
d. Distribution of Patients by Race and Persistency Flag

This stacked bar chart provides insight into drug persistency rates across different races. The proportion of persistency within each racial group highlights potential disparities that could be addressed to improve medication adherence.



e. Age Distribution Among Different Persistence Groups

The bar chart compares the age distribution between persistent and non-persistent patients. It reveals that persistence does not significantly differ across age groups, suggesting that other factors may play more critical roles in persistency.

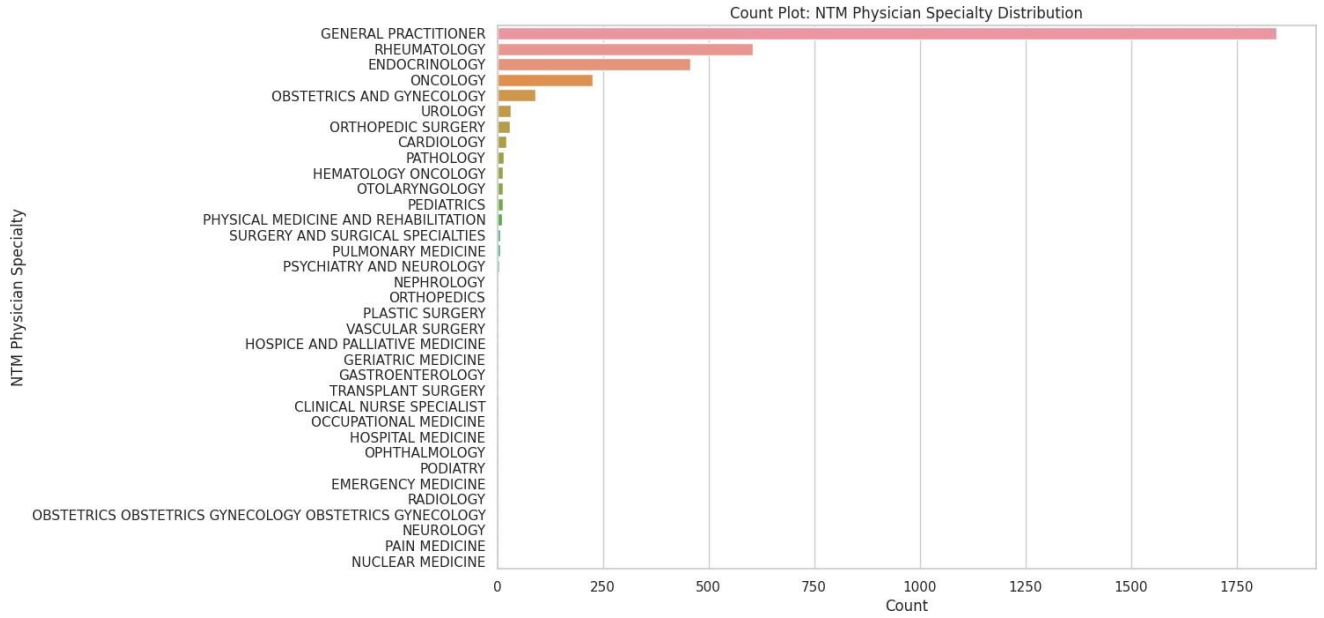


f. Specialties of Healthcare Providers (Word Cloud)

The word cloud visually represents the various healthcare provider specialties, with the size of each term indicating its frequency. General practice, surgery, and oncology are prominently featured, which could be influential in-patient drug persistency.

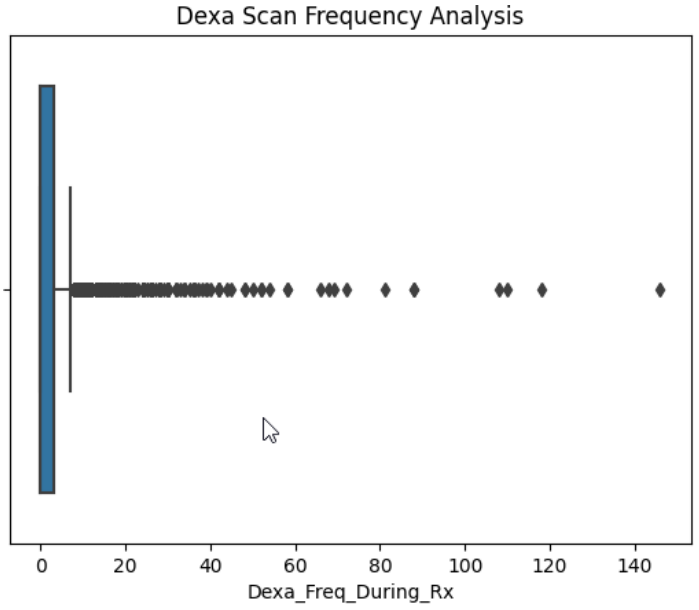
g. NTM Physician Specialty Distribution

The bar chart showcases the distribution of NTM physician specialties. General practitioners are the most common, followed by rheumatologists and endocrinologists, indicating these specialties' roles in ongoing patient care and potentially influencing drug persistency.



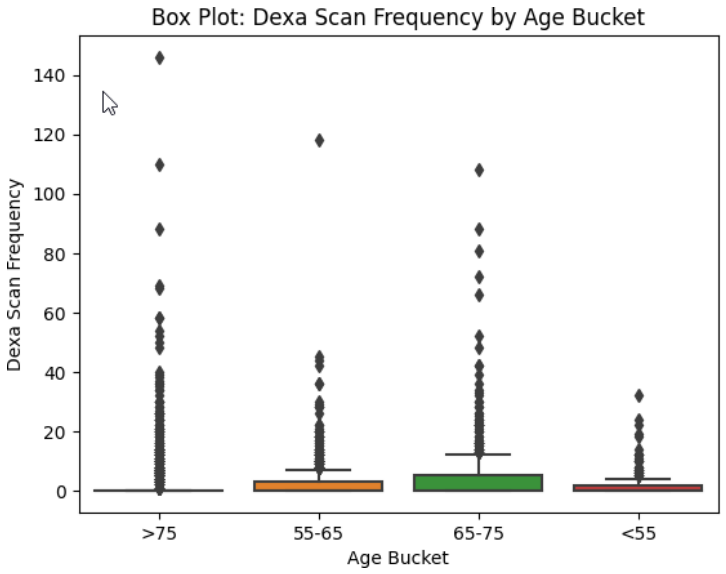
h. Dexa Scan Frequency Analysis

This boxplot displays the overall distribution of Dexa scan frequencies among all patients, showing a wide range but with most patients undergoing few scans, which may impact their treatment persistency.



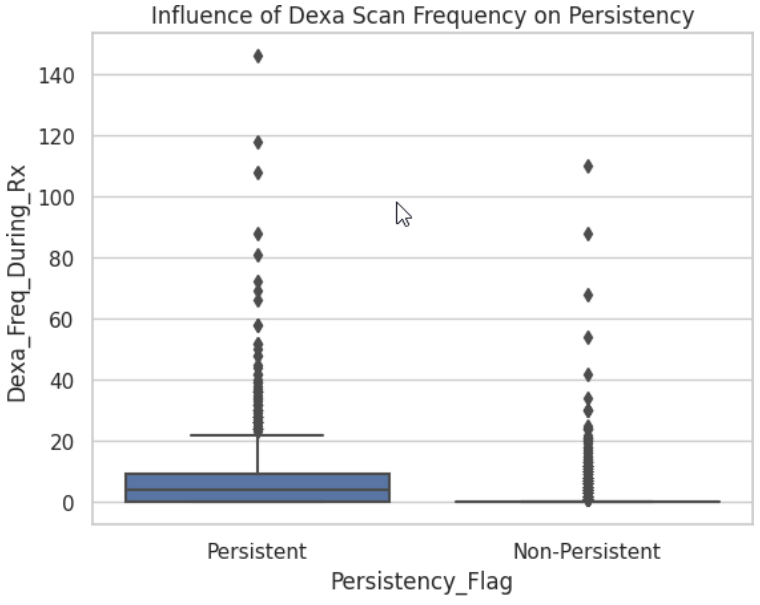
i. Dexa Scan Frequency by Age Bucket

The boxplot categorizes patients by age and the frequency of Dexa scans they received. The data indicates that older patients (>75) are more likely to have frequent Dexa scans, which could be attributed to increased monitoring with advancing age.



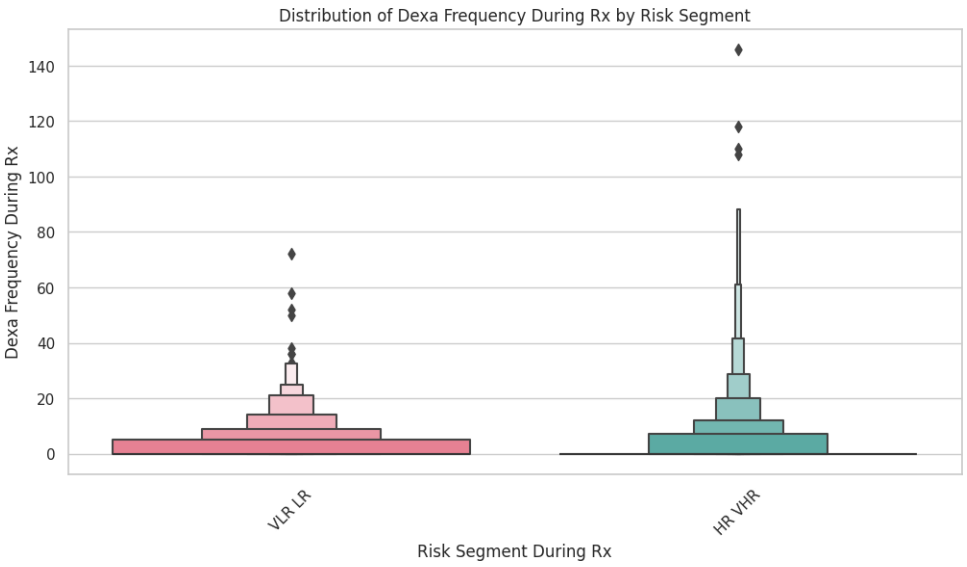
j. Influence of Dexa Scan Frequency on Persistency

This boxplot illustrates the distribution of Dexa scan frequencies among patients, categorized by their drug persistency status. Notably, persistent patients tend to undergo Dexa scans more frequently, which may suggest a correlation between regular monitoring and medication adherence.



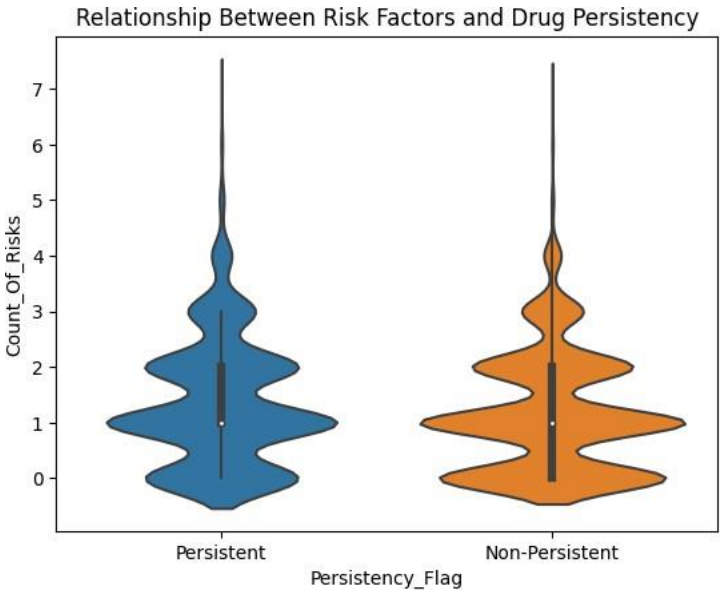
k. Distribution of Dexa Frequency During Rx by Risk Segment

This boxen plot compares the frequency of Dexa scans during treatment across different risk segments, suggesting that patients with a higher risk profile receive more frequent monitoring.



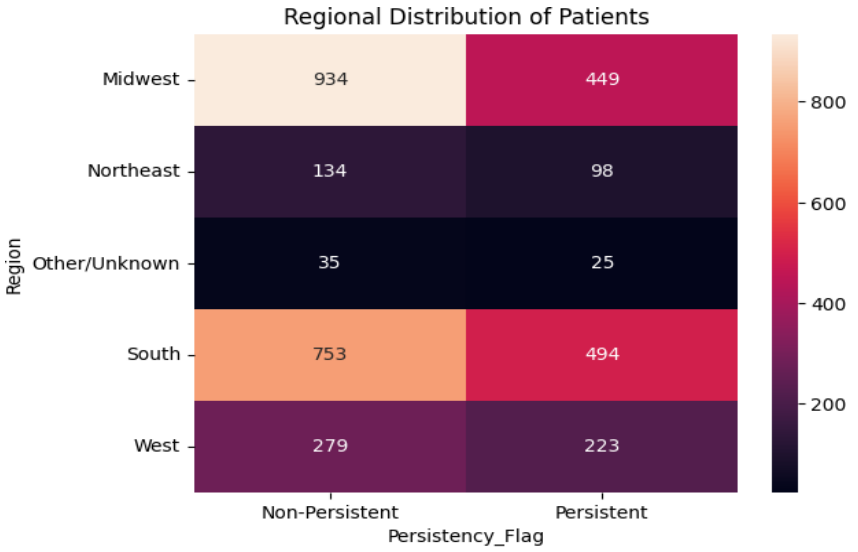
l. Relationship Between Risk Factors and Drug Persistency

The violin plot examines the relationship between the number of risk factors and drug persistency. A higher number of risk factors is associated with increased persistency, possibly due to more intensive health monitoring.



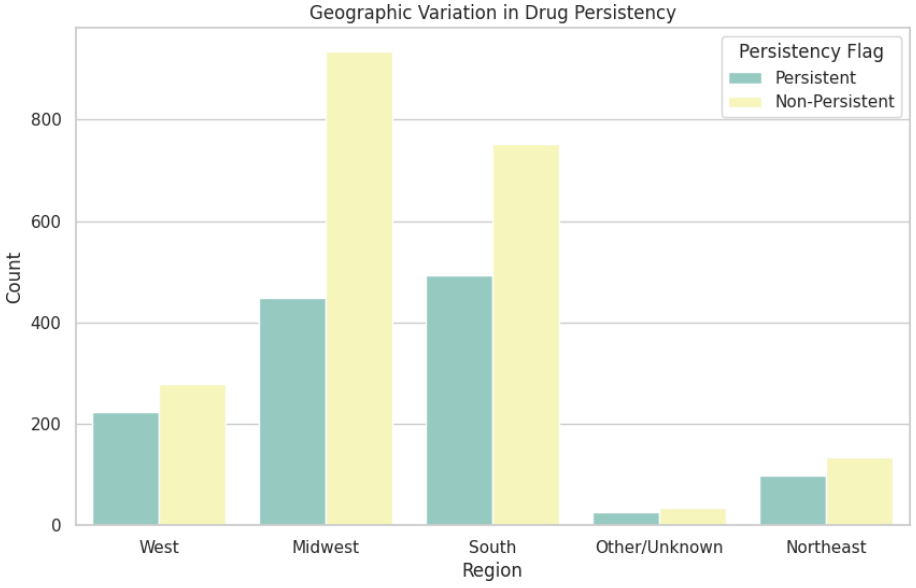
m. Regional Distribution of Patients

The heatmap visualizes patient distribution and drug persistency by region, offering insights into regional adherence patterns that could be vital for localized healthcare strategies.



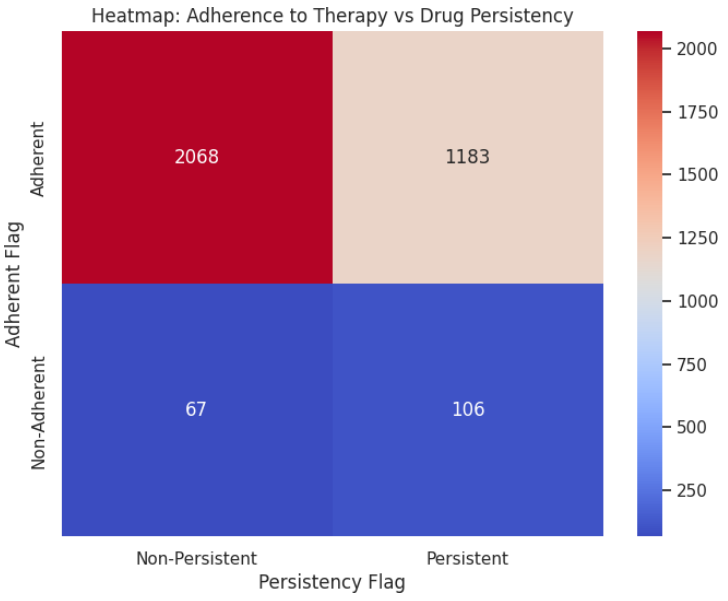
n. Geographic Variation in Drug Persistency

This chart presents the geographic variation in drug persistency, with certain regions showing higher rates of non-persistence, which could inform regional healthcare policy and patient outreach efforts.



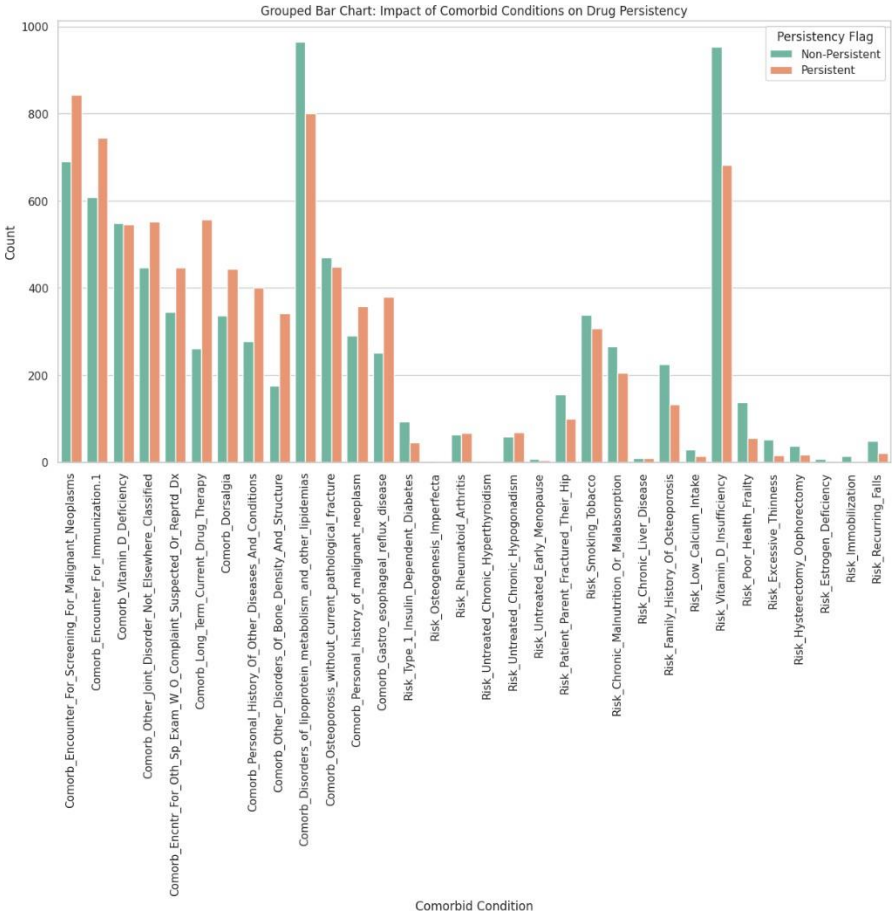
o. Heatmap: Adherence to Therapy vs Drug Persistency

The heatmap compares therapy adherence to drug persistency, indicating that patients who adhere to their therapy schedule are generally more persistent with their medication.



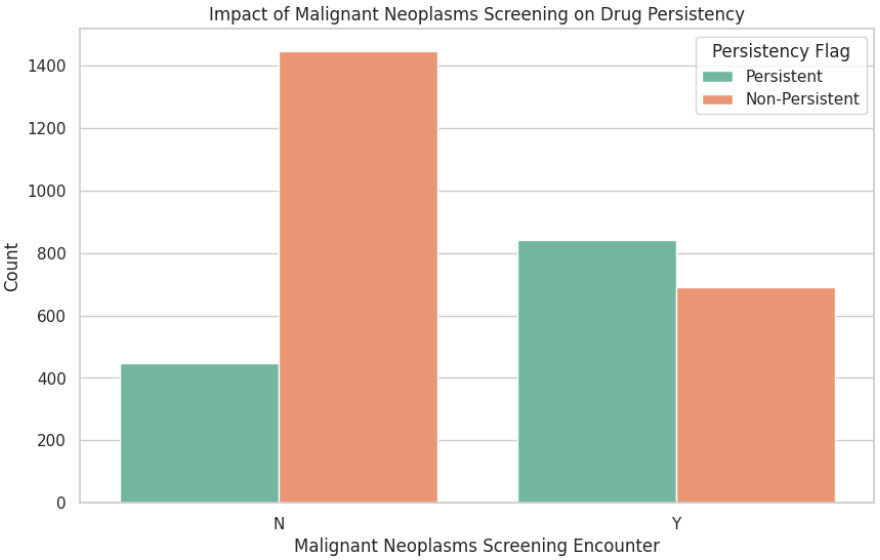
p. Grouped Bar Chart: Impact of Comorbid Conditions on Drug Persistency

This chart analyzes the effects of comorbid conditions on drug persistency. Certain conditions, such as chronic pain and hypertension, show a clear correlation with higher rates of persistency, likely due to the necessity of consistent treatment.



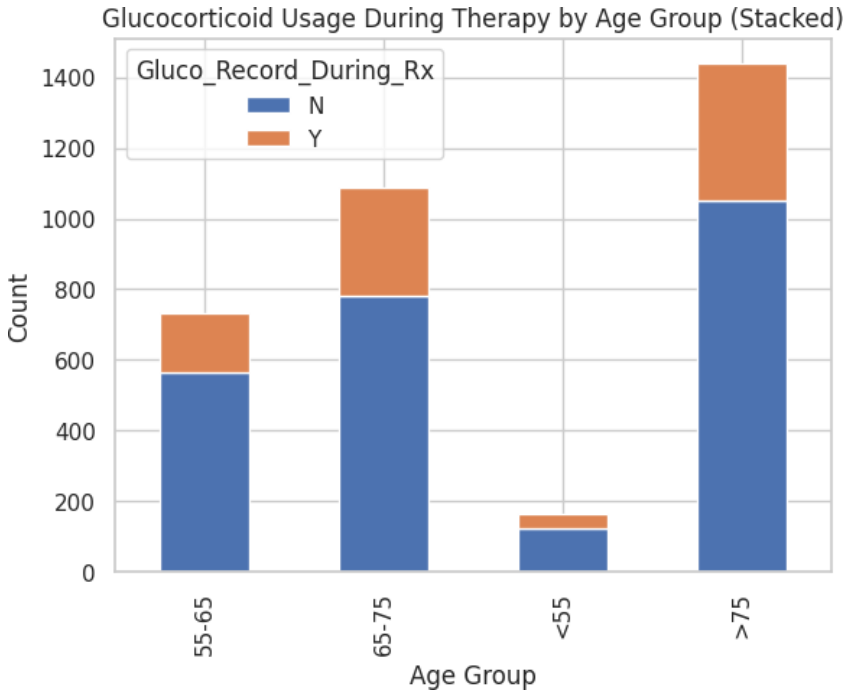
q. Impact of Malignant Neoplasms Screening on Drug Persistency

The bar chart demonstrates the impact of malignant neoplasms screening on drug persistency. Patients screened for neoplasms tend to be more persistent, possibly due to increased engagement with healthcare services.



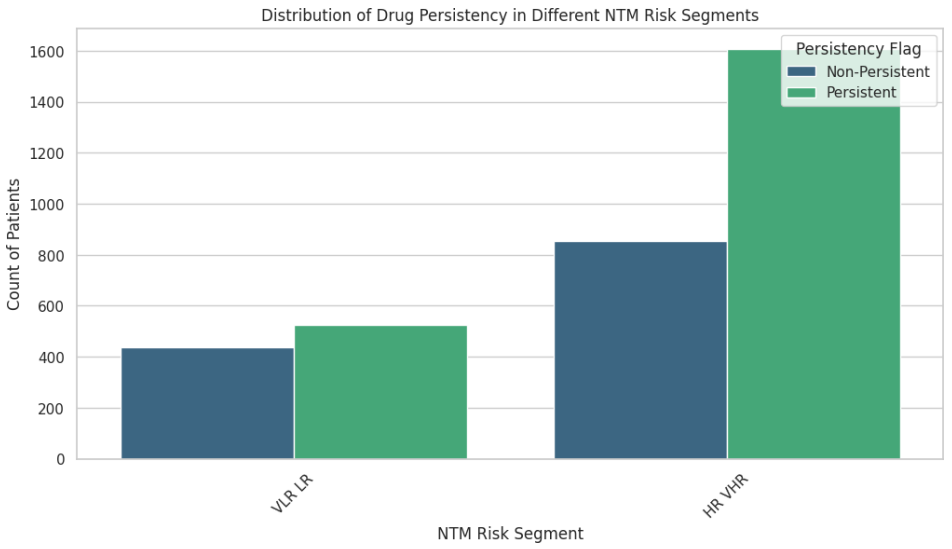
r. Glucocorticoid Usage During Therapy by Age Group

This stacked bar chart shows glucocorticoid usage during therapy across different age groups. Older patients have higher usage, which may influence persistency due to the chronic nature of conditions treated with glucocorticoids.



s. Distribution of Drug Persistency in Different NTM Risk Segments

This bar chart illustrates the distribution of drug persistency within NTM risk segments. Patients in the higher risk segments show increased persistency, suggesting that risk awareness may motivate adherence to treatment.



Final Recommendation: Based on the EDA, it is recommended that pharmaceutical companies and healthcare providers focus on personalized patient education, regular health monitoring, and targeted interventions for patients with specific comorbid conditions and within certain demographic groups. Improving communication between general practitioners and specialists, along with addressing disparities in drug persistency across races and regions, could also significantly enhance overall medication adherence.

7.Preparing Data for Modeling:

a. Splitting Features and Target Variable:

- The dataset (df_encoded) is divided into features (X) and the target variable (y). The target variable is 'Persistency_Flag'.

b. Training and Testing Sets:

- The train_test_split function is used to split X and y into training and testing sets.
- 20% of the data is reserved for testing (test_size=0.2), ensuring a portion of data is held back for model evaluation.
- random_state=42 ensures that the split is reproducible; the same rows will be split into training and testing sets each time the code is run.

c. Output Interpretation:

- The shapes of the split datasets are displayed.
- X_train and X_test each have 71 features, indicating a wide range of variables consideredf for modeling. X_train contains 2739 samples, and X_test has 685 samples, signifying the division of datainto training and testing sets.

d. ANOVA for Feature Analysis:

- Conducted ANOVA to examine differences in 'Dexa_Freq_During_Rx' between 'Persistent' and 'Non-Persistent' groups.
- Results: F-Value of 433.263 and P-Value of ~1.05e-90.
- Indicates a significant difference in 'Dexa_Freq_During_Rx' across the two groups, suggesting its importance in predicting drug persistency.
- These steps are essential in ensuring the data is correctly structured for analysis and in identifying key features that might impact the model's predictive power.

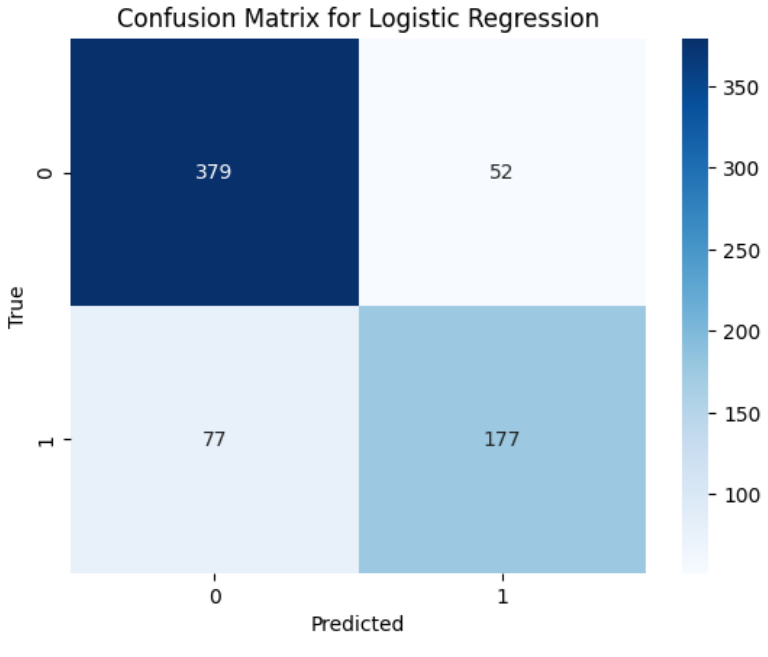
8.Model Training and Evaluation

a. Linear Model: Logistic Regression

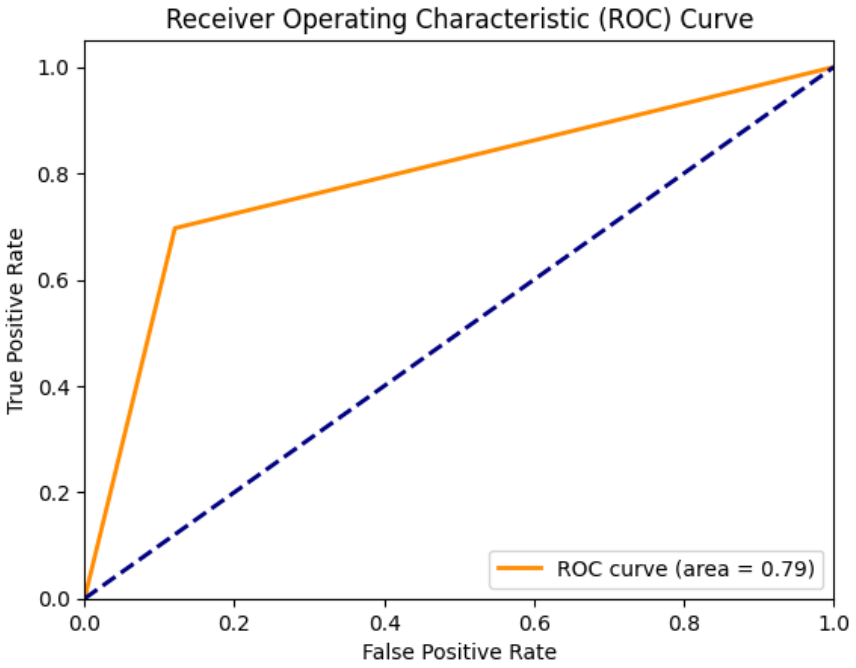
- **Model Introduction:**
  - Logistic Regression used as a baseline model.
  - Ideal for binary classification problems like 'Persistency_Flag'.
- **Advantages:**
  - Simple and interpretable model.
  - Provides probability scores for predictions.
- **Disadvantages:**
  - May underperform with complex, non-linear relationships.
- **Accuracy and Precision:**
  - Achieved 81% accuracy.
  - Precision: 83% for Non-Persistent, 77% for Persistent cases.
- **Recall and F1-Score:**
  - Recall: 88% for Non-Persistent, 70% for Persistent.
  - F1-scores: 85% for Non-Persistent, 73% for Persistent.

Graph Explanation:

- **Confusion Matrix:**
  - Shows a higher accuracy in predicting Non-Persistent cases.
  - Indicates a tendency to under-predict persistency.



- **ROC Curve:**
  - ROC-AUC score at 0.788, suggesting moderate discrimination ability.
  - Suggests room for improvement in model sensitivity and specificity.



Key Takeaways:

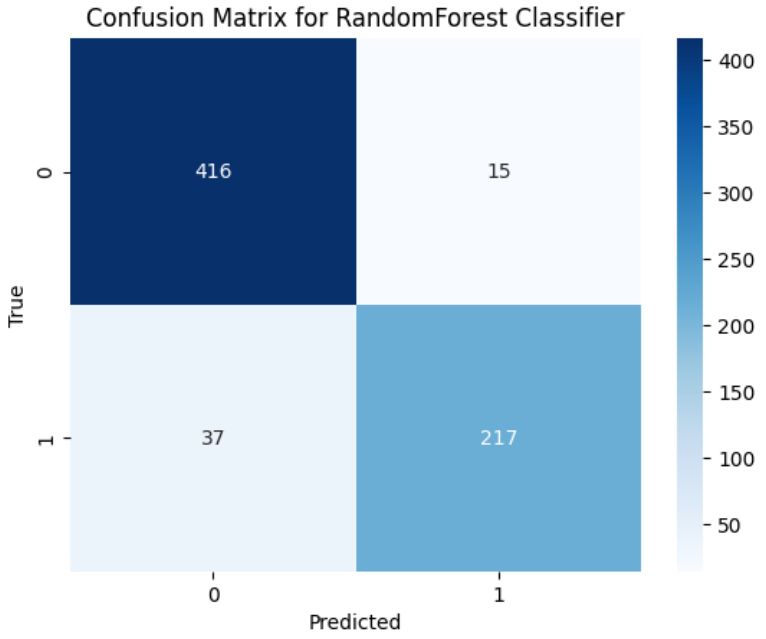
- The model is more effective in identifying Non-Persistent cases.
- Good balance in precision and recall for Non-Persistent cases but can improve for Persistent cases.
- Model shows potential, but exploring additional features or more complex models could enhance performance.

b. Ensemble Model: Random Forest

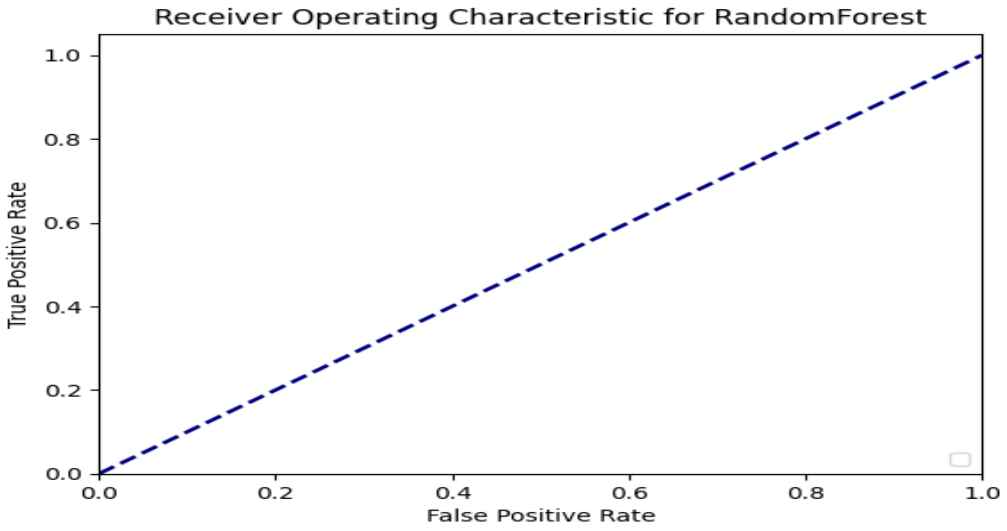
- **Model Introduction:**
  - Random Forest, an ensemble learning method, used for its robustness.
  - Handles large data sets with higher dimensionality well.
- **Advantages:**
  - Effective for complex datasets.
  - Reduces overfitting risk due to ensemble approach.
- **Disadvantages:**
  - More computationally intensive.
  - Less interpretable than simpler models.
- **Accuracy and Precision:**
  - Achieved a high accuracy of 92%.
  - Precision: 92% for Non-Persistent, 94% for Persistent cases.
- **Recall and F1-Score:**
  - Recall: 97% for Non-Persistent, 85% for Persistent.
  - F1-scores: 94% for Non-Persistent, 89% for Persistent.

Graph Explanation:

- **Confusion Matrix:**
  - Shows excellent performance in identifying Non-Persistent cases.
  - Indicates better balance in predicting both Non-Persistent and Persistent cases compared to Logistic Regression.



- **ROC Curve:**
  - The ROC curve for the Random Forest model, positioned near the top left corner with a high AUC, indicates excellent predictive accuracy in distinguishing between Non-Persistent and Persistent cases.



Key Takeaways:

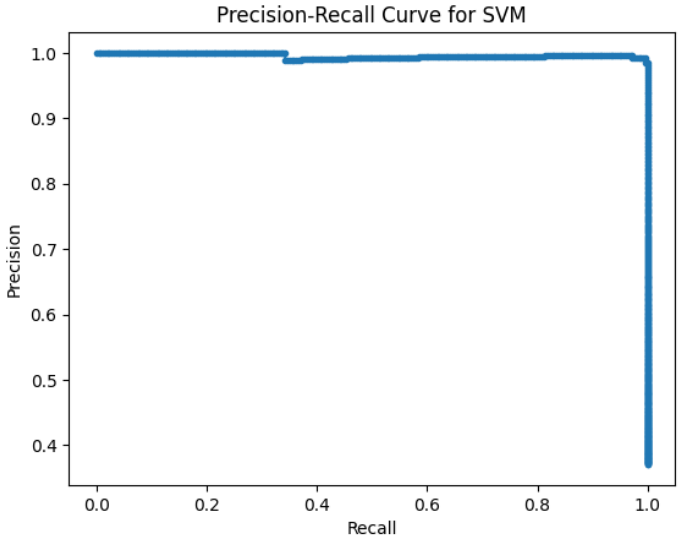
- The Random Forest model shows a notable improvement over Logistic Regression in both accuracy and balance between the classes.
- High precision and recall, especially for Non-Persistent cases, indicating strong predictive capabilities.

c. Support Vector Machine (SVM):

- **Model Introduction:**
  - SVM is used for its effectiveness in high-dimensional spaces.
  - Particularly suitable for binary classification problems like 'Persistency_Flag'.
- **Advantages:**
  - High accuracy in classification tasks.
  - Effective in cases where the number of dimensions exceeds the number of samples.
- **Disadvantages:**
  - Can be less effective on very large datasets.
  - Requires careful selection of kernel and regularization parameters.
- **Accuracy and Precision:**
  - Exceptional accuracy of 99%, indicating excellent model performance.
  - Precision: 100% for Non-Persistent, 98% for Persistent cases.
- **Recall and F1-Score:**
  - Recall: 99% for Non-Persistent, 100% for Persistent.
  - F1-scores: 99% for both Non-Persistent and Persistent cases.

Graph Explanation:

- **Precision-Recall Curve:**
  - The Precision-Recall curve indicates outstanding performance, as it hovers close to the top right corner, showing both high precision and high recall.
  - This suggests that the SVM model effectively balances correctly identifying positive (Persistent) cases (high recall) while maintaining a low rate of false positives (high precision).



Key Takeaways:

- The SVM model shows outstanding performance, outshining both Logistic Regression and Random Forest in accuracy and class balance.
- High precision and recall across both categories highlight its robustness in handling this classification task.

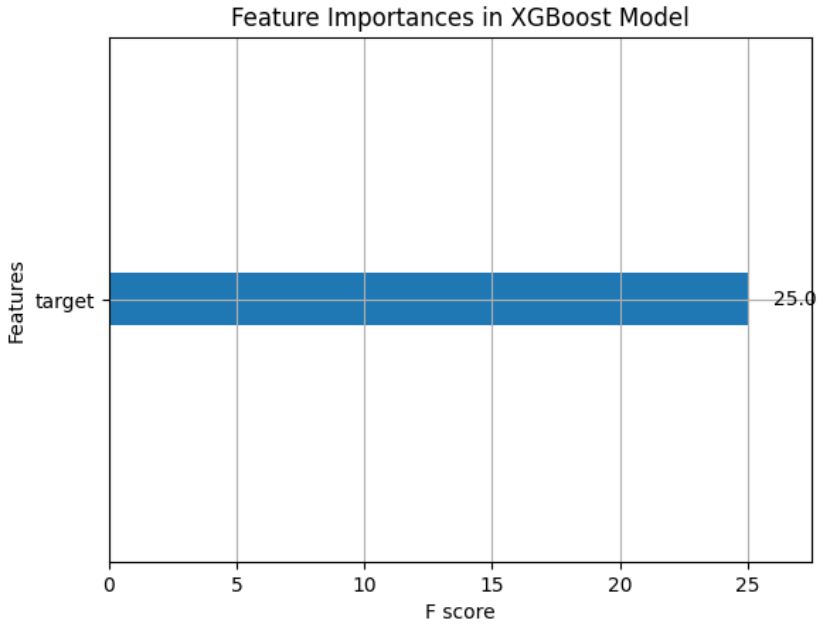
d. Gradient Boosting

Slide 5: Gradient Boosting Overview

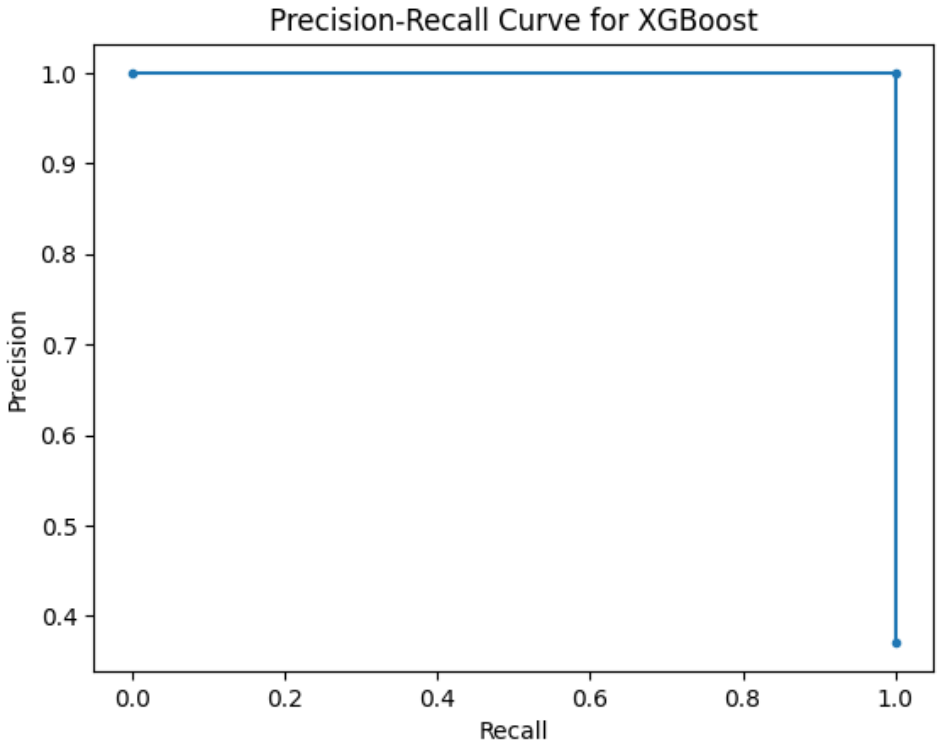
- **Heading:** Gradient Boosting for Enhanced Predictions
- **Model Introduction:**
  - Gradient Boosting, an advanced ensemble technique, known for high accuracy.
  - Sequentially builds weak learners to improve predictions.
- **Advantages:**
  - Often provides high predictive accuracy.
  - Good control over overfitting through hyperparameters.
- **Disadvantages:**
  - Time-consuming training process.
  - Requires careful tuning to avoid overfitting.
- **Accuracy and Precision:**
  - Achieved a perfect accuracy of 100%, showcasing exceptional model performance.
  - Precision: 100% for both Non-Persistent and Persistent cases.
- **Recall and F1-Score:**
  - Recall: 100% for both Non-Persistent and Persistent cases.
  - F1-scores: 100% for both categories, indicating flawless balance between precision and recall.

Graph Explanation:

- **Feature Importance:**
  - Reveals the most significant features that the model relied on for making predictions.
  - Provides insights into which aspects of the data are most influential in determining patient drug persistency.



- **Precision-Recall Curve for XGBoost:**
  - The Precision-Recall Curve shows that the XGBoost model achieves nearly perfect precision across all levels of recall.
  - This indicates that the model is able to identify the persistent class with high accuracy, maintaining a high true positive rate without increasing the false positives.



Key Takeaways:

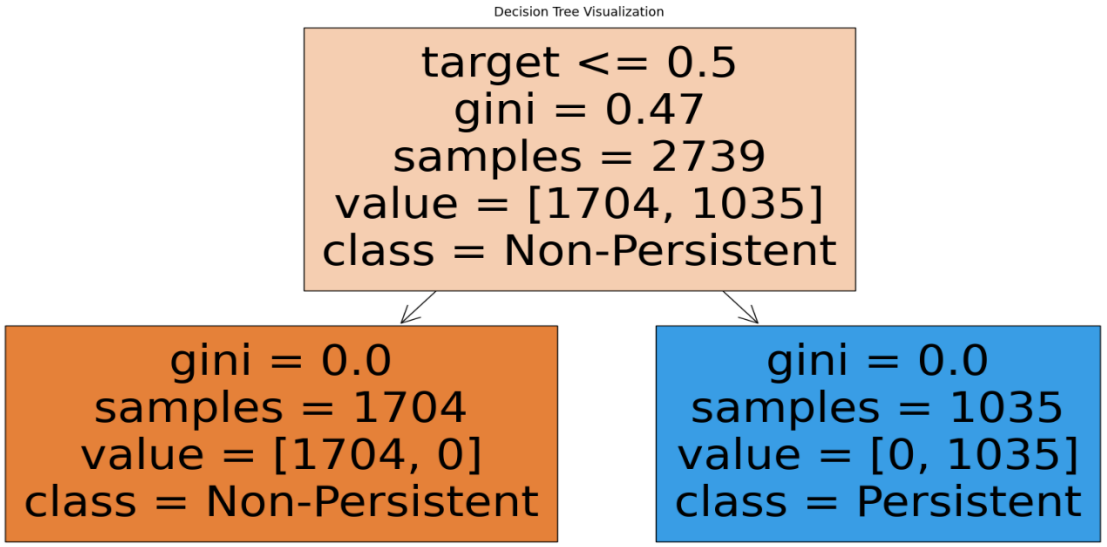
- The XGBoost model demonstrates unparalleled performance in this classification task, with unmatched precision, recall, and accuracy.
- Its ability to perfectly classify cases is impressive, though it's crucial to consider the potential for overfitting, especially in a real-world scenario where data may not always be as clean or well-defined.

e. Decision Tree

- **Model Introduction:**
  - Decision Tree Classifier, a simple yet powerful model for classification tasks.
  - Offers a clear visualization of the decision-making process.
- **Advantages:**
  - Easy to understand and interpret.
  - Can handle both numerical and categorical data.
- **Disadvantages:**
  - Prone to overfitting, especially with complex trees.
  - Can be unstable, as small variations in data might lead to a completely different tree.
- **Accuracy and Precision:**
  - Perfect accuracy of 100%, indicating exceptional performance.
  - Precision: 100% for both Non-Persistent and Persistent cases.
- **Recall and F1-Score:**
  - Recall: 100% for both Non-Persistent and Persistent cases.
  - F1-scores: 100% for both categories, showing ideal balance between precision and recall.

Graph Explanation:

- The initial split on the 'target' feature with a threshold of ≤ 0.5 results in perfect classification with no misclassification in the first split.
- Subsequent splits further refine the model's decision boundaries, effectively capturing more complex patterns in the data.



Key Takeaways:

- The Decision Tree model exhibits outstanding performance, achieving flawless classification.
- The model's simplicity, coupled with its perfect scoring across all metrics, underscores its efficiency for this dataset.
- While the results are impressive, it's important to be cautious about overfitting, as decision trees can perfectly fit the training data, especially in cases of clean, well-structured datasets.