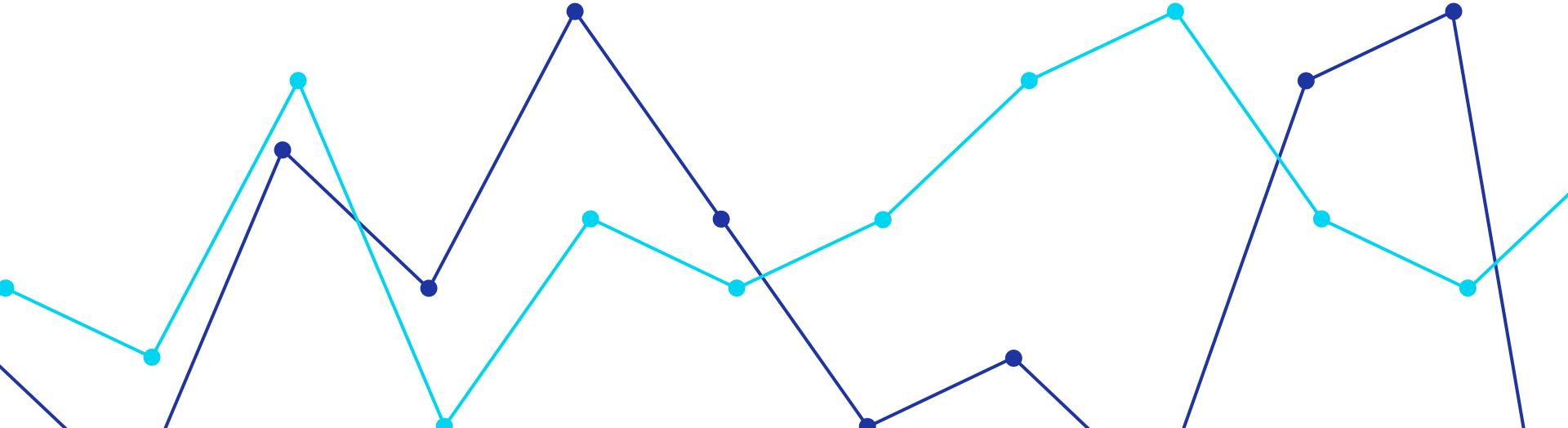
Healthcare - Persistency of a Drug

EDA and Model Recommendation

Jonathan Allen | Matthew Iversen



Problem Description

This project revolves around understanding the persistence of a drug based on physician prescriptions for a patient. The objective is to automate predictions of a persistency flag using a classification machine learning model. After this analysis, we will identify the most important factors in predicting persistence. Before we can make a model to accomplish this goal, we will first delve into the data to find patterns and problems that may need to be addressed.

Steps of Data Cleaning

Define and Display Data

Began with targeting, loading, and displaying our given dataset within our chosen software for conducting the given project.

Determine Shape of Dataset

Finding the number of samples and features. Secondly, deriving the shape of the data. Lastly, reviewing and highlighting about our restrictions and compatibilities when instructing further certain code or algorithms.

Determine Data-Types

Used in order to acknowledge differences in data such as numerical and categorical. As a result, we can guide to visualization and analysis techniques, ensuring the selection of suitable methods for meaningful insights.

Determine File Summary

Executed to give insights to the number of rows and columns within the dataset. Also, the file size, for memory usage insight.

Decipher Unique Variables

At first, gaining an initial understanding of the dataset. Aided in handling missing values, outliers, or inconsistent data. Allowed identification of any unexpected or irregular values that would need attention and ultimately cleaning.

Determine Missing Values

Assessed the completeness of our dataset and identified potential issues with data collection, entry, or storage.

Steps of Data Cleaning

Spell Check

This function identified pairs of string values within each object-type column that have a similarity ratio above the specified threshold, indicating potential typos. Only columns with string data types were analyzed, and any specified columns were excluded.

Outliers

To ensure that our summary statistics and metrics are accurate if needed. Keeping disportionality influencers under supervision and thus control. Detecting and managing outliers also helps ensure that the underlying assumptions are met.

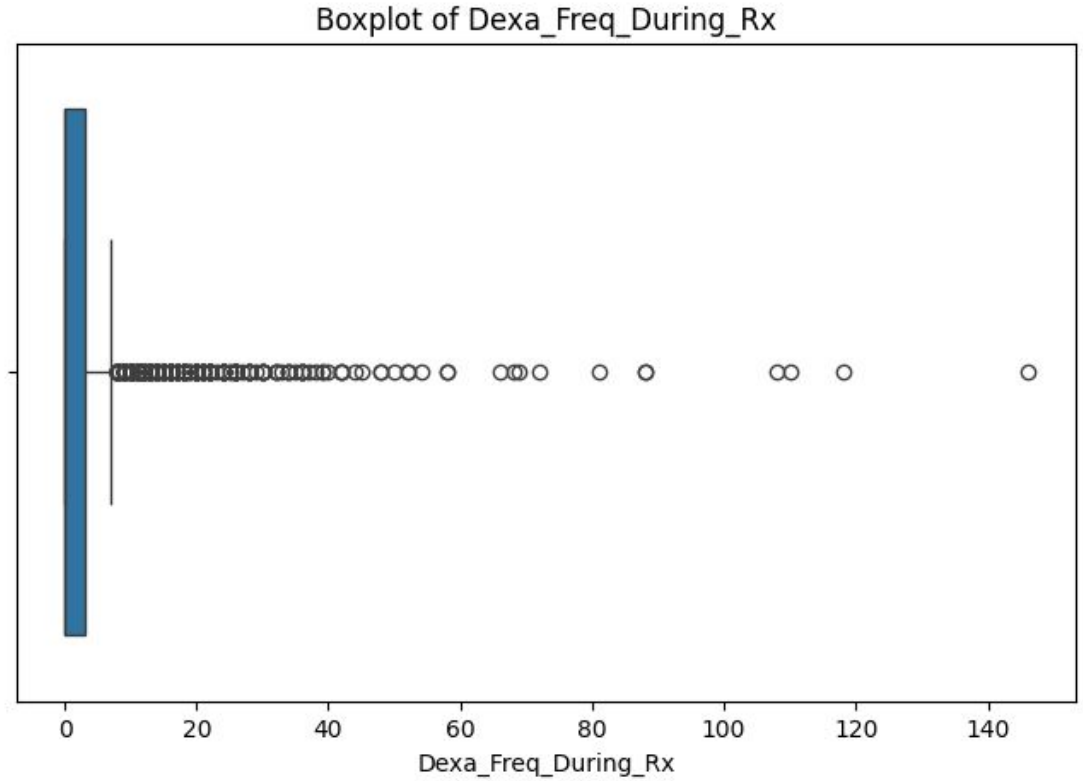
Checking for Duplicate Rows and Columns

Removes the duplicate columns and prints the removed columns, if any.

Validated Number and Names of Columns

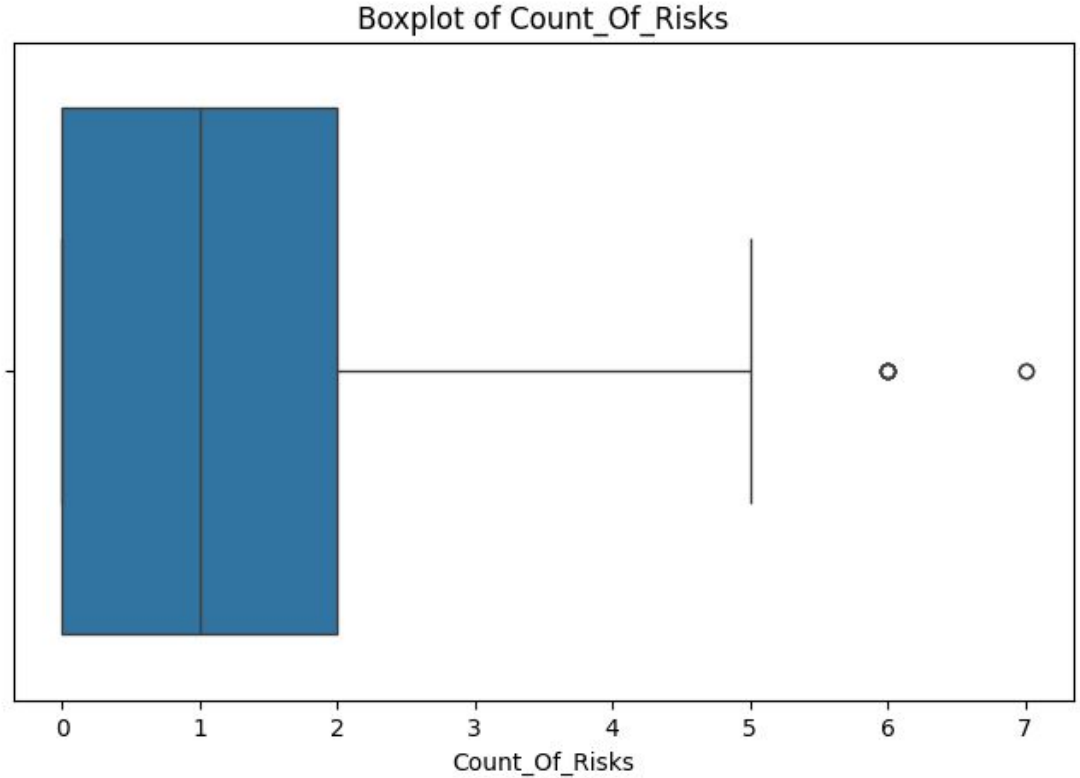
Validated whether or not the number of columns and their header names matched to what their expected to be within the dataframe.

Looking for Outliers in Numerical Columns



Pulling from the csv's data descriptions directly, Dexa_Freq_During_Rx is defined as "Number of DEXA scans taken prior to the first NTM Rx date (within 365 days prior from rxdate)". Numbers over 100 are highly suspicious as getting this many scans within a year period seems unusual. This would call for a talk with a stakeholder with more experience in the area as it appears to have outliers.

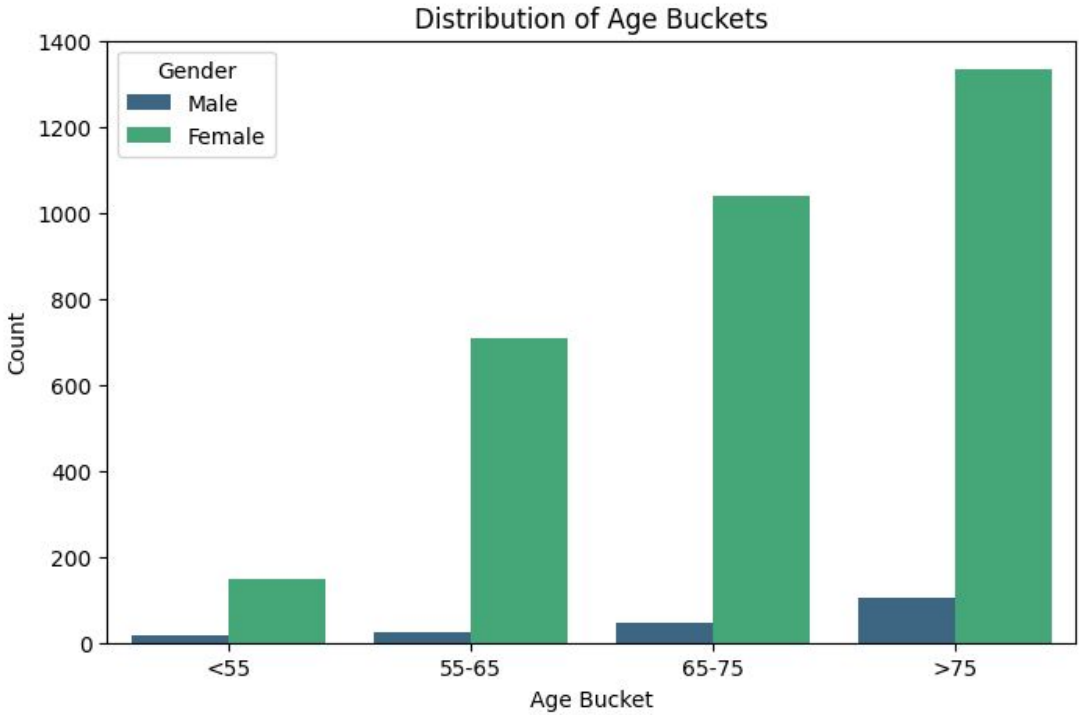
Looking for Outliers in Numerical Columns



While their are statistical outliers, 7 is not unthinkable for the amount of health risks a human can have. These two data points will be left in as they do not appear to be outliers.

Exploratory Data Analysis

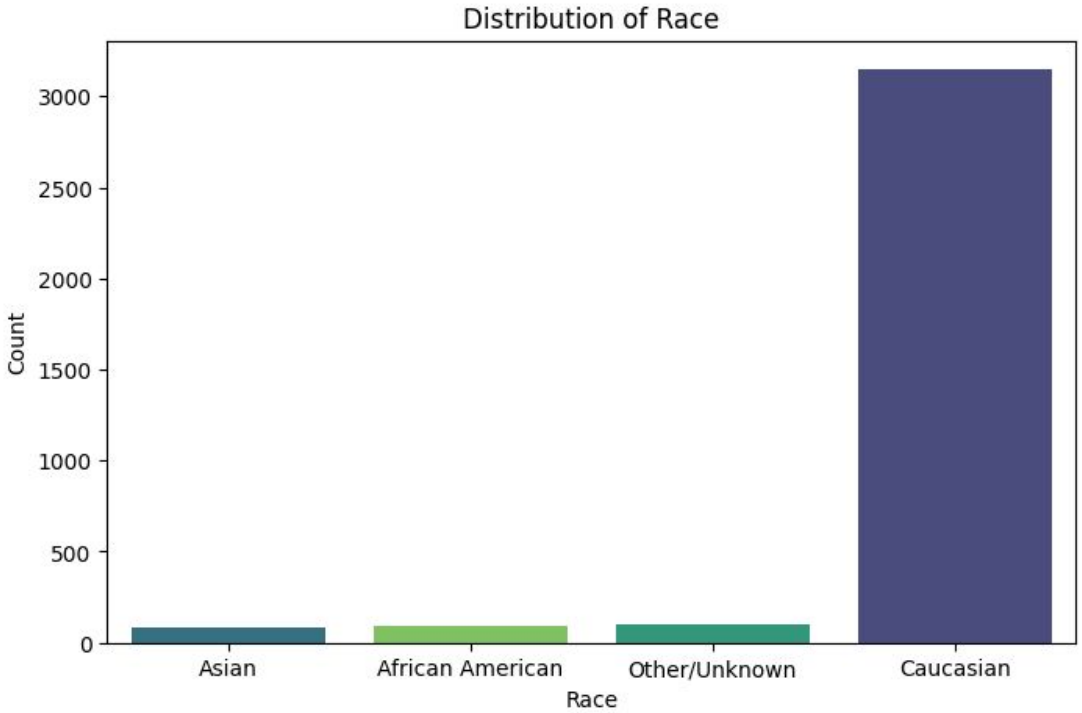
Exploring Patient Demographics



For age, it is important to note that this dataset is largely made up of 65+ patients and will not represent younger patients as well in future analysis and modeling. This will create a bias for older patients and will require future adjustments. Additionally, it will be worthwhile to see if different age buckets require different approaches.

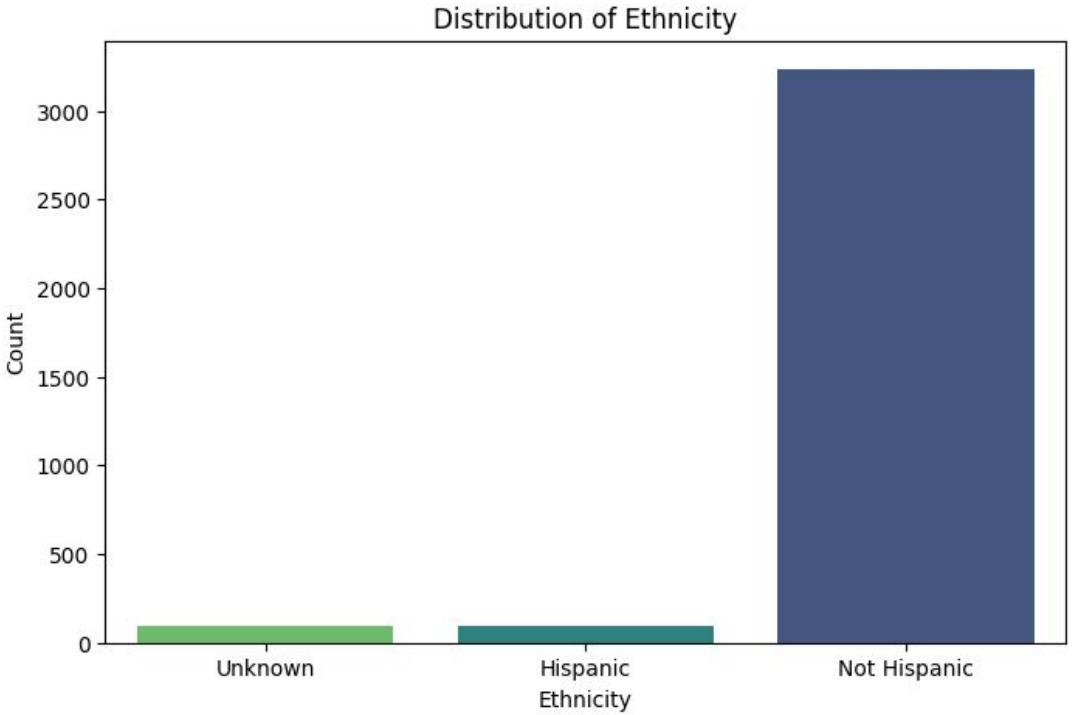
For gender, it is important to note that data is 94% females. This will also create a heavy bias toward female analysis and predictions if left unchecked.

Exploring Patient Demographics



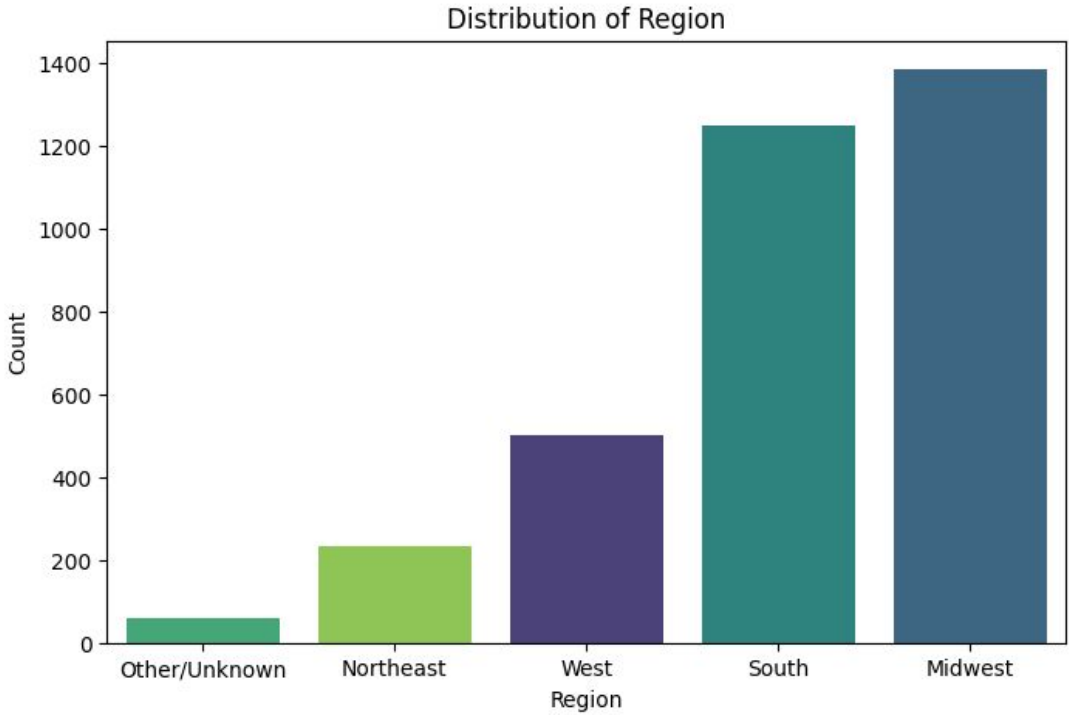
Race is highly biased toward caucasian patients. This can lead to future analysis and models only working well for caucasian patients. Oversampling other races or getting more data is suggested.

Exploring Patient Demographics



As is most demographics in this dataset, there is a heavy bias toward a large group, which in this case is non-hispanic patients. This data would benefit from a more balanced split in addition to more categories split from the non-hispanic bin.

Exploring Patient Demographics

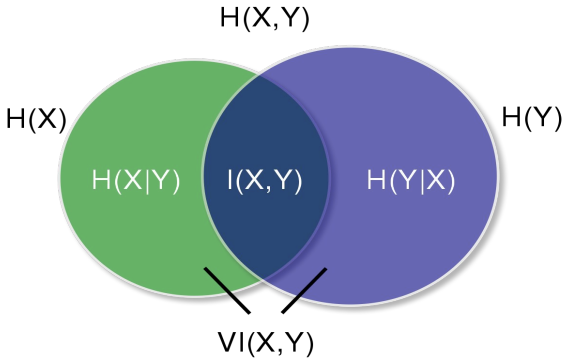


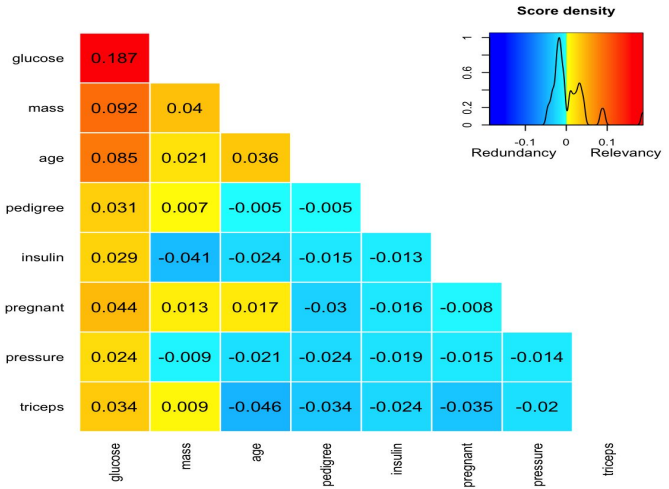
Region's skew is not as significant as the other categories, but does favor the south and midwest. The northeast and west region have enough data that oversampling would work, but the other/unknown region could pose a problem for accuracy.

Mutual Information

Given a large clinical database of longitudinal patient information including many covariates, it is computationally prohibitive to consider all types of interdependence between patient variables of interest.

Mutual information is a measure of the amount of information that knowing the value of one variable provides about another variable. In the context of a healthcare dataset, you can use mutual information to quantify the dependence between different features (variables) in the dataset. The scikit-learn library in Python provides a convenient function for calculating mutual information.





Mutual Information

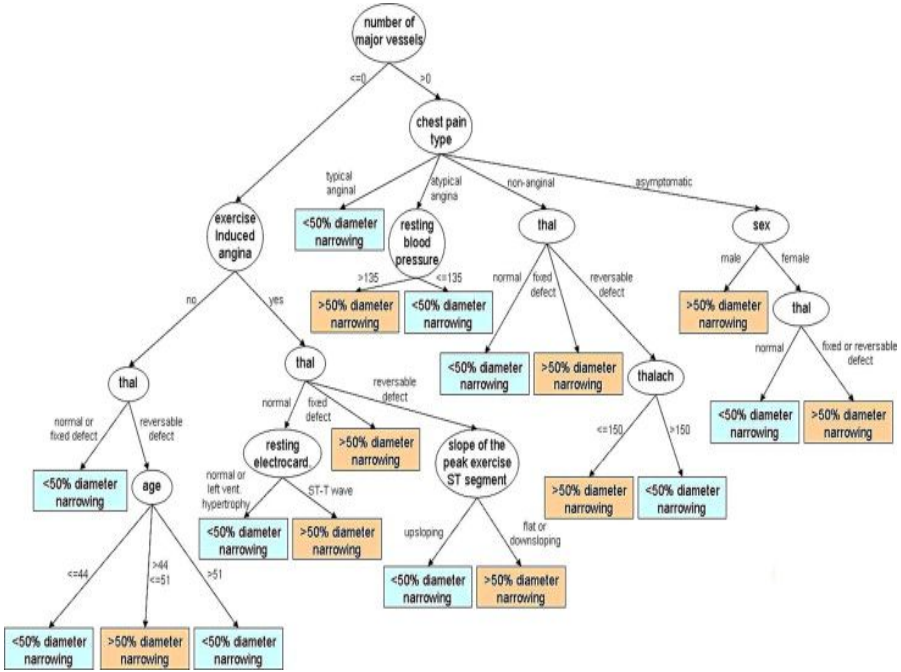
Utilizing the mutual information method(s) on a dataframe comprising diverse medical techniques involves a sophisticated analysis aimed at uncovering the inherent relationships and dependencies among the various variables.

In our context, mutual information is a metric which quantified the amount of information shared between variables, making it particularly useful in the context of medical data where intricate interconnections exist.

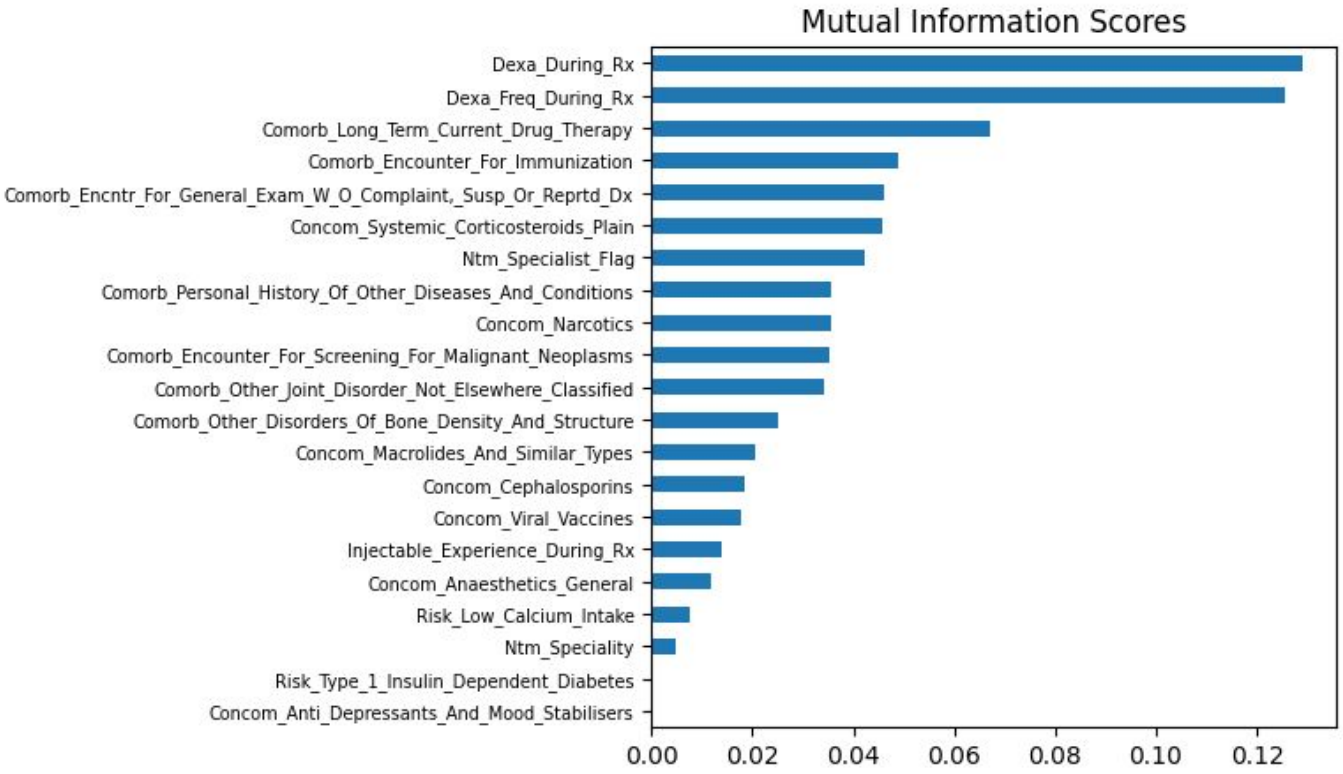
By applying to our data (populated with different medical techniques), we were able to gain insights into the extent to which these techniques contribute unique and shared information.

As a result, identifying key associations and dependencies, facilitating a comprehensive understanding of how different medical procedures interact.

When applied in a real healthcare sector scenario, such a method could potentially influence important decision-making processes. The analysis can help uncover patterns, dependencies, and potential synergies among provided variables, ultimately contributing to more informed and effective medical practices.



Mutual Information



These are the top features we will focus on when moving on to the modeling phase. The MI scores show how much a column accounts for the variability in the target feature, which is Persistency_Flag in this case.

Data Limitations

- This dataset is primarily 65+ caucasian (non-hispanic) females from the South and Midwest. Because of this, future models and statistical analysis based on this data will favor this demographic and will likely not be accurate with the general population.
- Dexa_Freq_During_Rx is a column that was initially shown to have many outliers. As this is the second most correlated column to Persistency_Flag, the accuracy of this column would need to be confirmed with stakeholders before finalizing conclusions.
- There are 166 male, 276 non-caucasian, 189 non-hispanic, and 60 unknown location patients. Relative to the 3424 patients, these are very small categories and could cause inaccurate results when oversampling.

Final Recommendations

- It is recommended to get a more inclusive sample so that the model can have higher confidence for general use. This could come from improved sampling methods or just more samples overall.
- Dexa_During_Rx and Dexa_Freq_During_Rx both have high correlation to the Persistency_Flag and should be analyzed further once their reliability is confirmed.
- While computationally expensive, creating new features that attempt to find correlations with combinations of columns could help improve model accuracy.

Model Recommendations

- As no more data will be provided for this project, oversampling the underrepresented demographics is recommended for making a single-general model.
- A stratified model, stratified sampling, or cross-validation could all help with the under representation of certain demographics.
- If more data and diverse data were provided, building multiple models for different demographic groups could yield highly-tailored recommendations.