Data Glacier

Your Deep Learning Partner

Drug Persistency Project

Virtual Internship: Final Presentation

Group Name: Health+

Group Members:

Mohammad Odeh (United Arab Emirates)

Sakib Mahmud (Qatar)

Date: 11-May-2021

Background – Drug Persistency case study

- One of the challenge 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 : Gather insights on the factors that are impacting the persistency, build a classification for the given dataset.

The analysis has been divided into three parts:

- Data Understanding
- Data insights and visualization
- Recommendations

Data Exploration

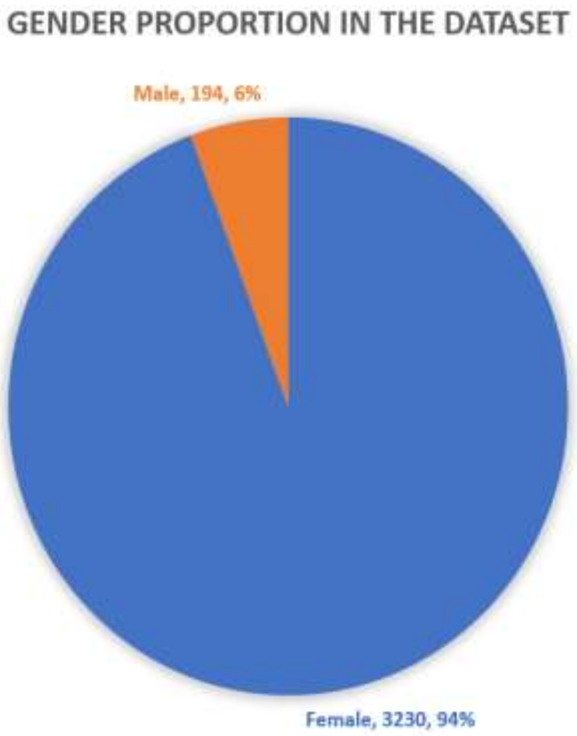
- 68 Features, including :
  - General features such as (Demographics, Provider Attributes)
  - Diseases/Drugs Factors
  - Clinical Factors

- Total number of patients : 3424

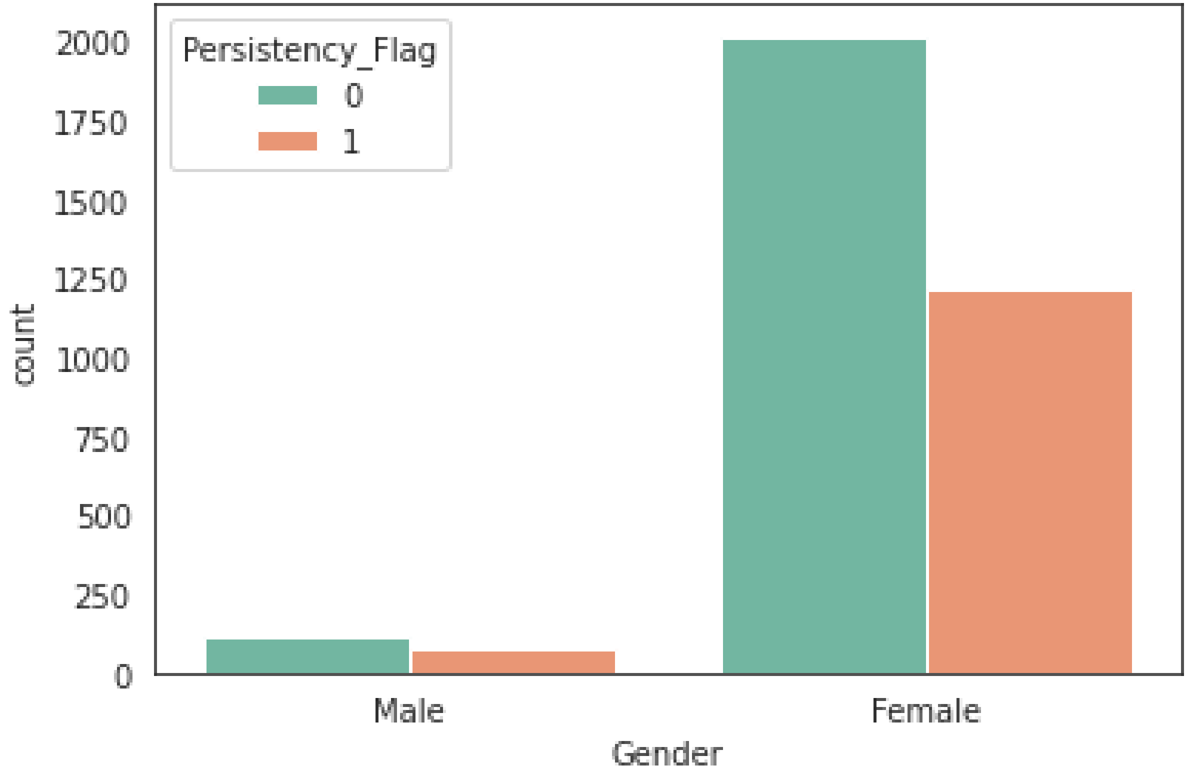
Assumptions:

- The data follows Normal Distribution.
- Patients' history data were recorded accurately without any errors in testing or examination.

Demographics Analysis

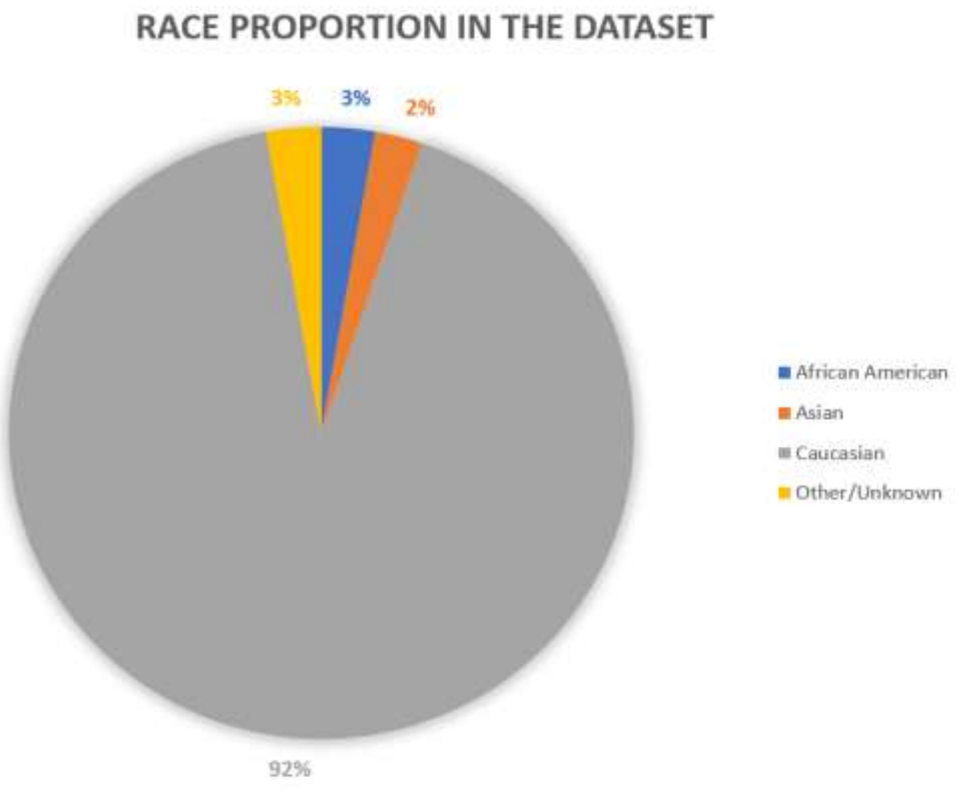


Gender Proportion

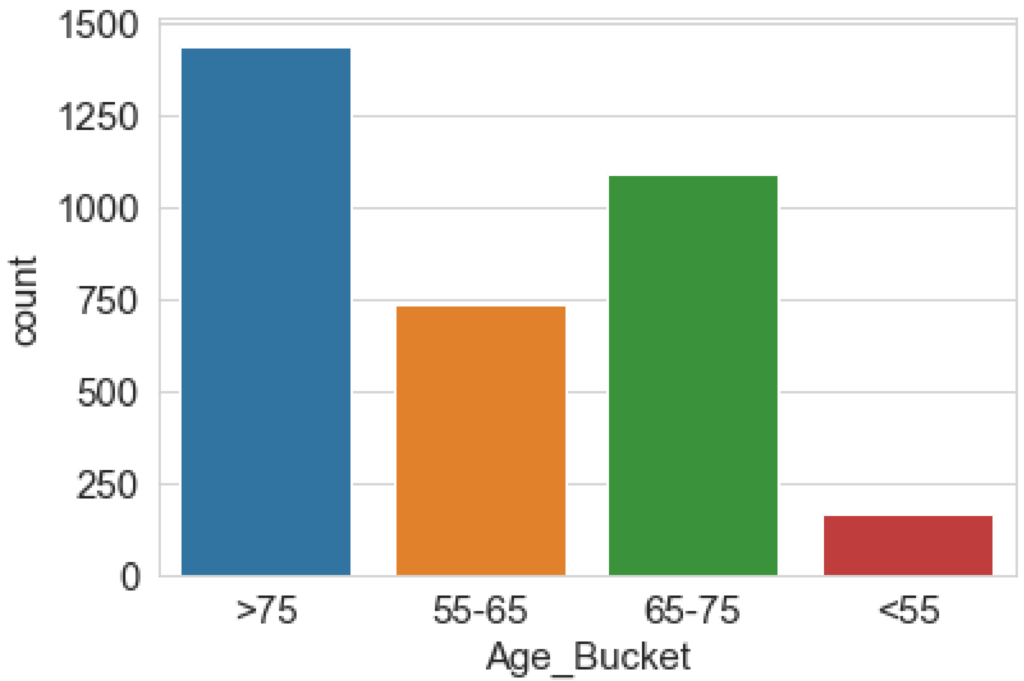


Gender Proportion vs. Persistency Flag

Demographics Analysis

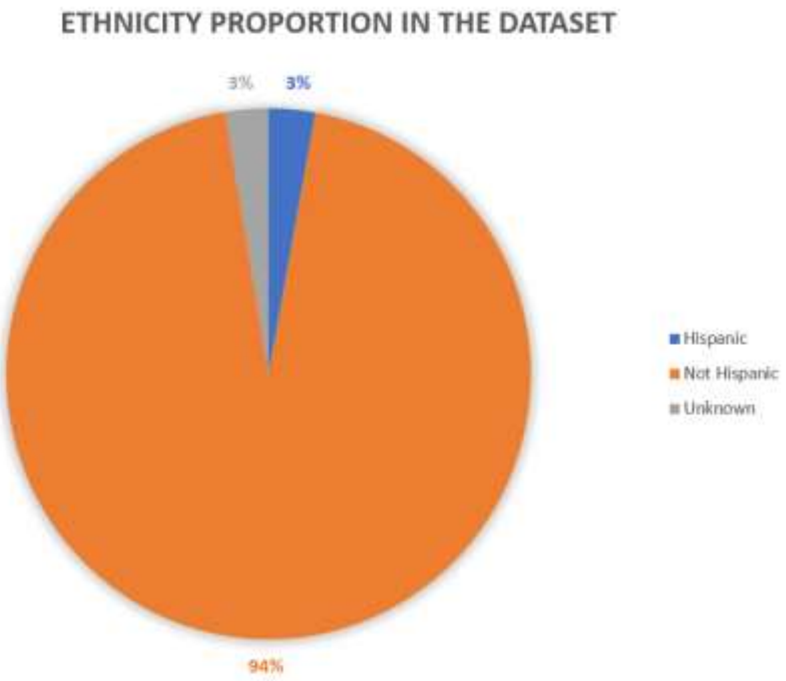


Age Proportion

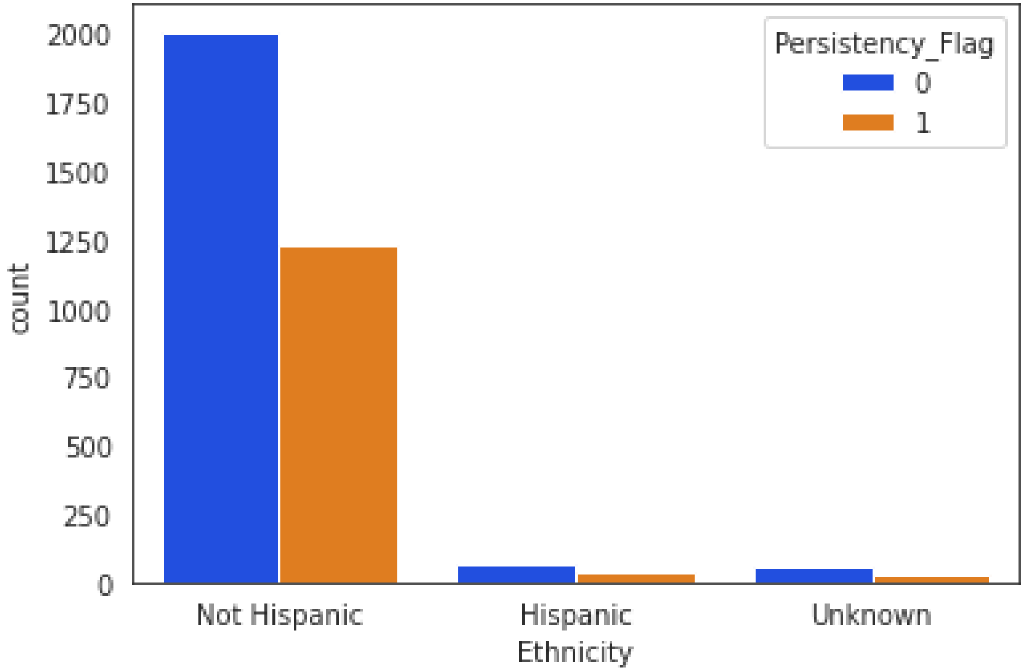


Age Bucket vs. Persistency Flag

Demographics Analysis

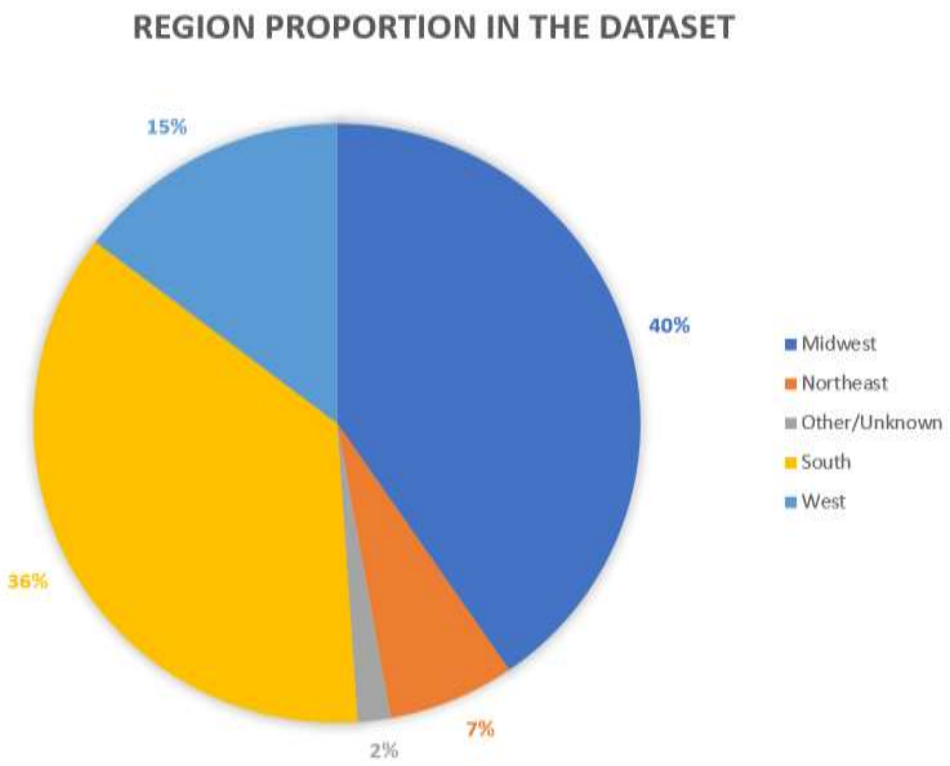


Ethnicity Proportion

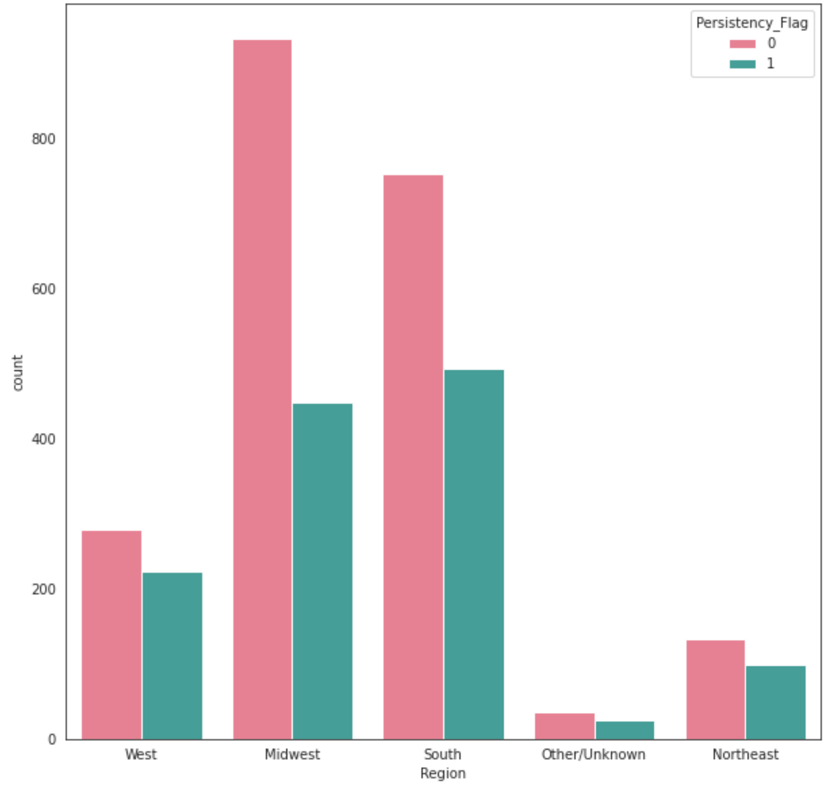


Ethnicity vs. Persistency Flag

Demographics Analysis

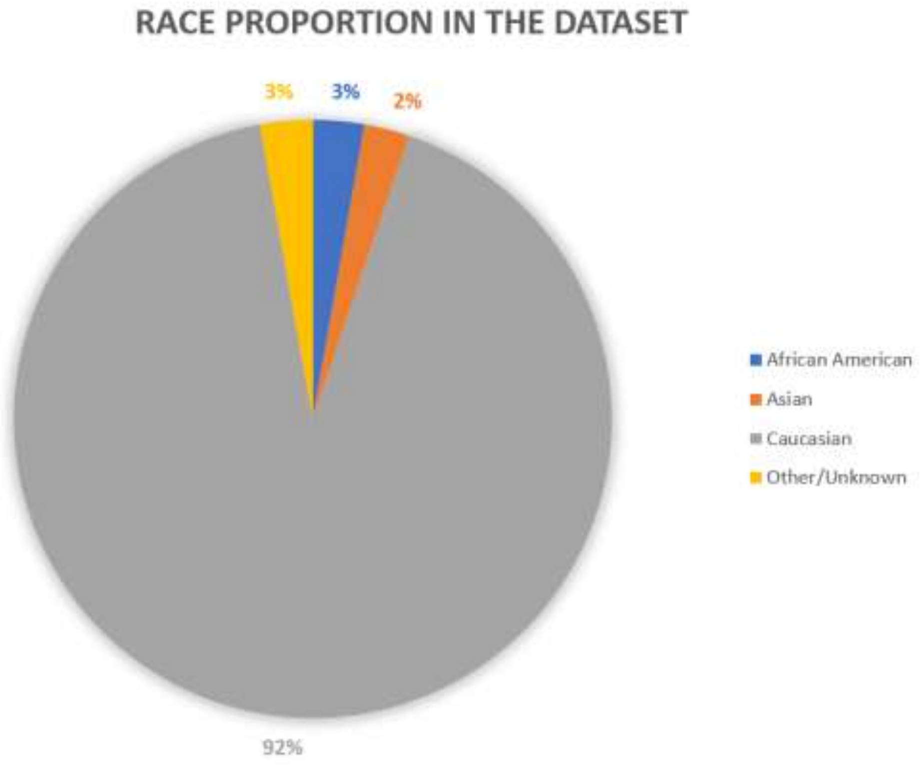


Region Proportion

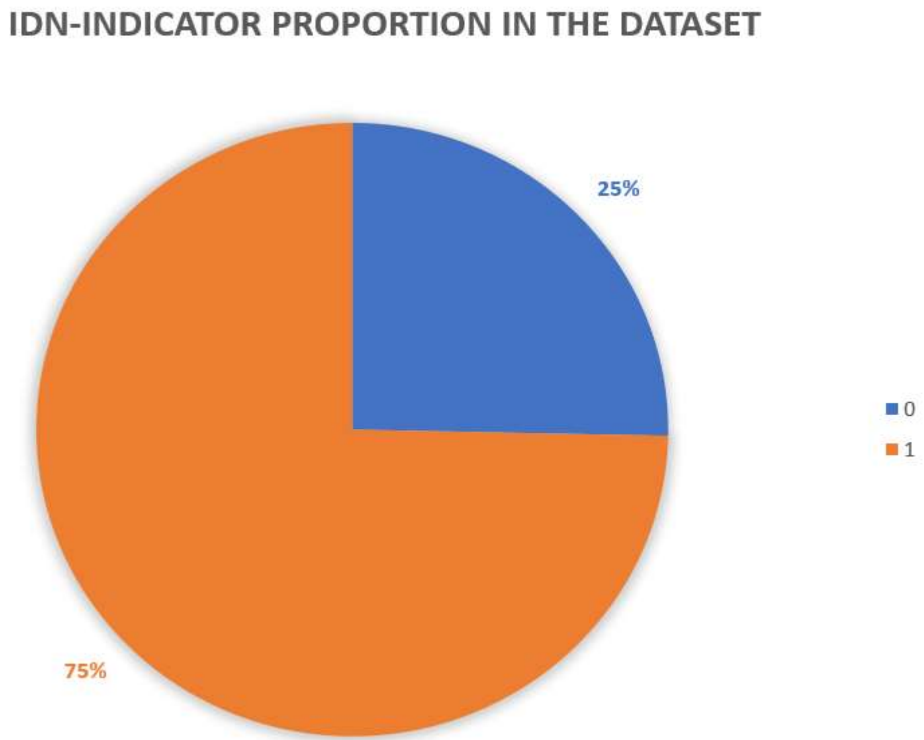


Region vs. Persistency Flag

Demographics Analysis

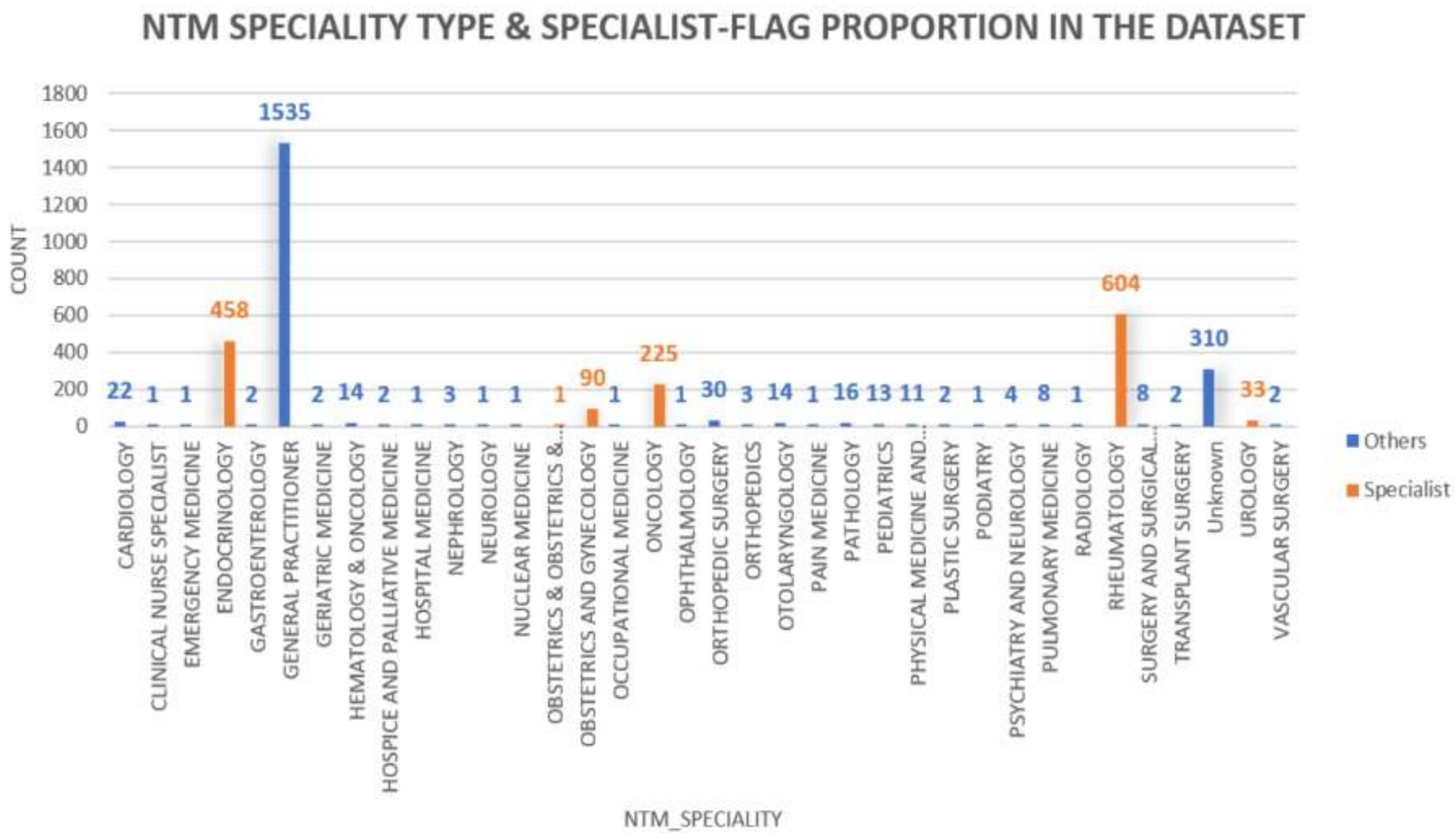


Race Proportion

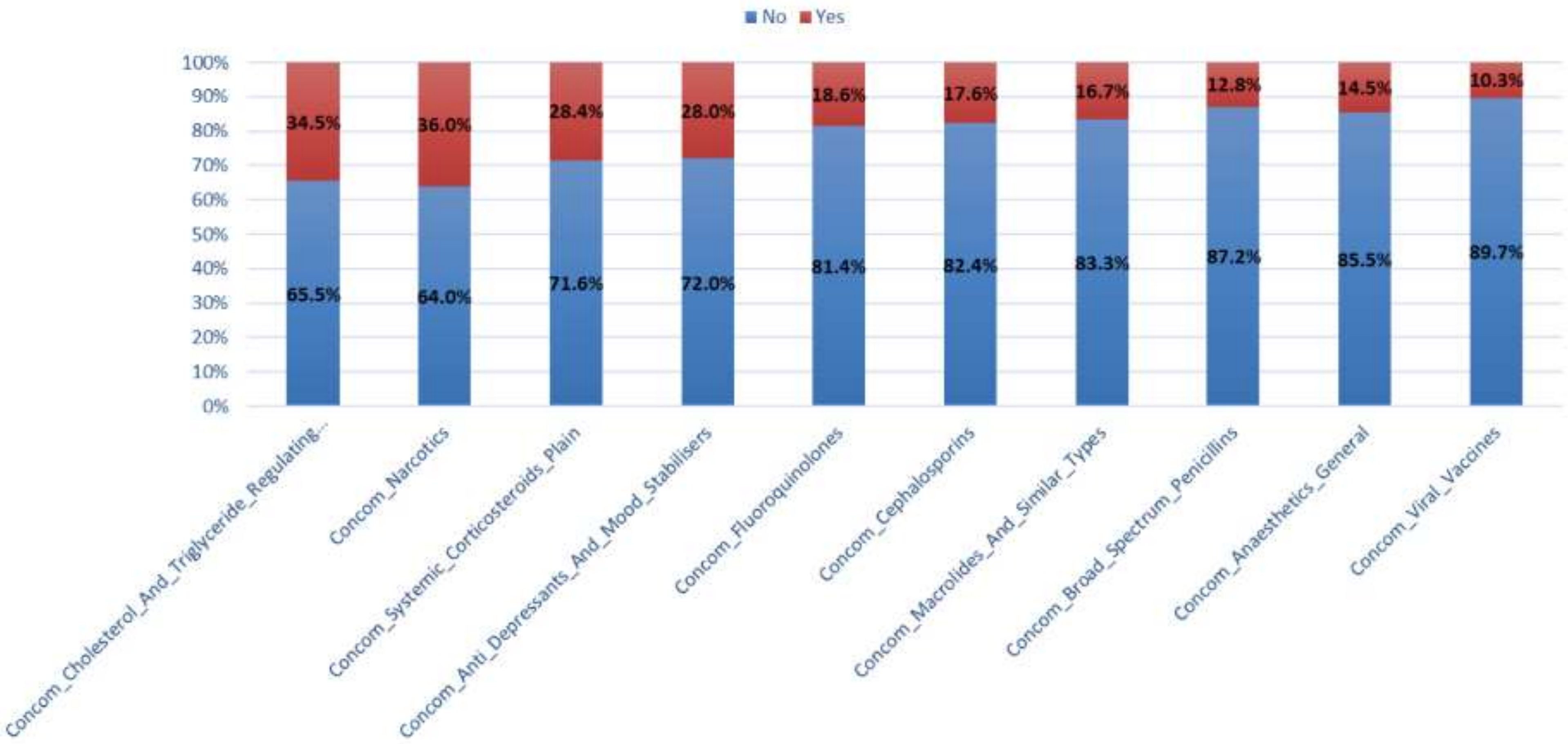


IDN Indicator Ratio

Disease Type and Responsible Physician Specialty Analysis

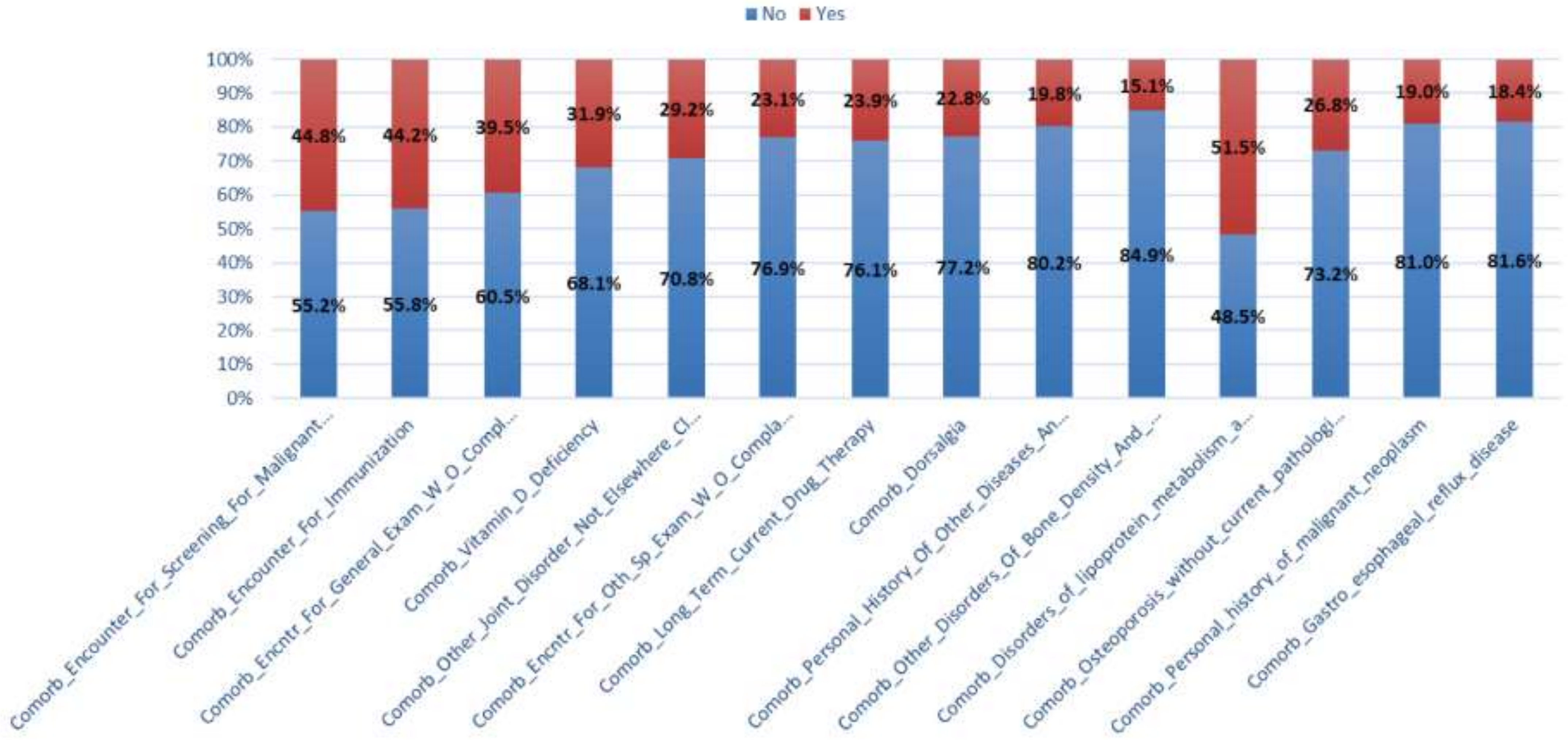


Drug Factor Analysis



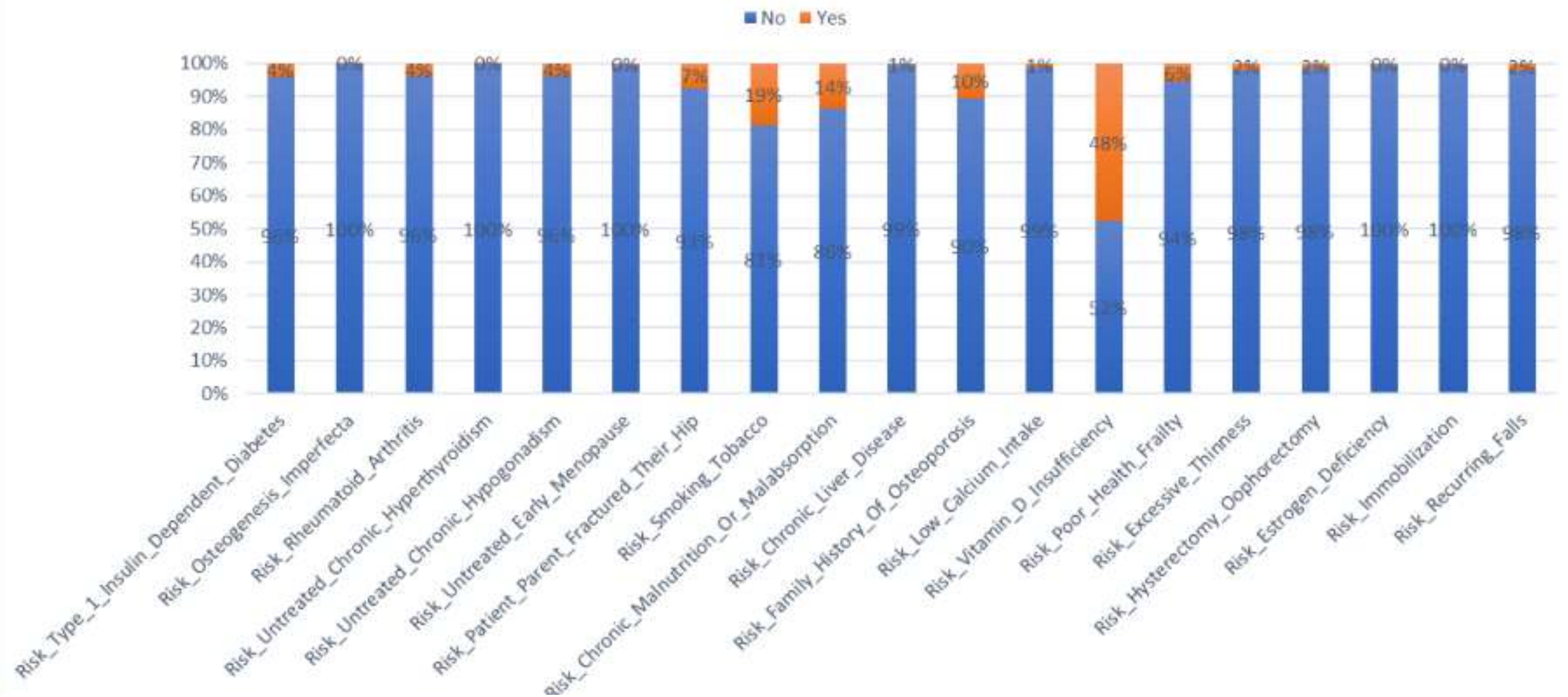
Concomitancy of Drugs

Diseases Factor Analysis



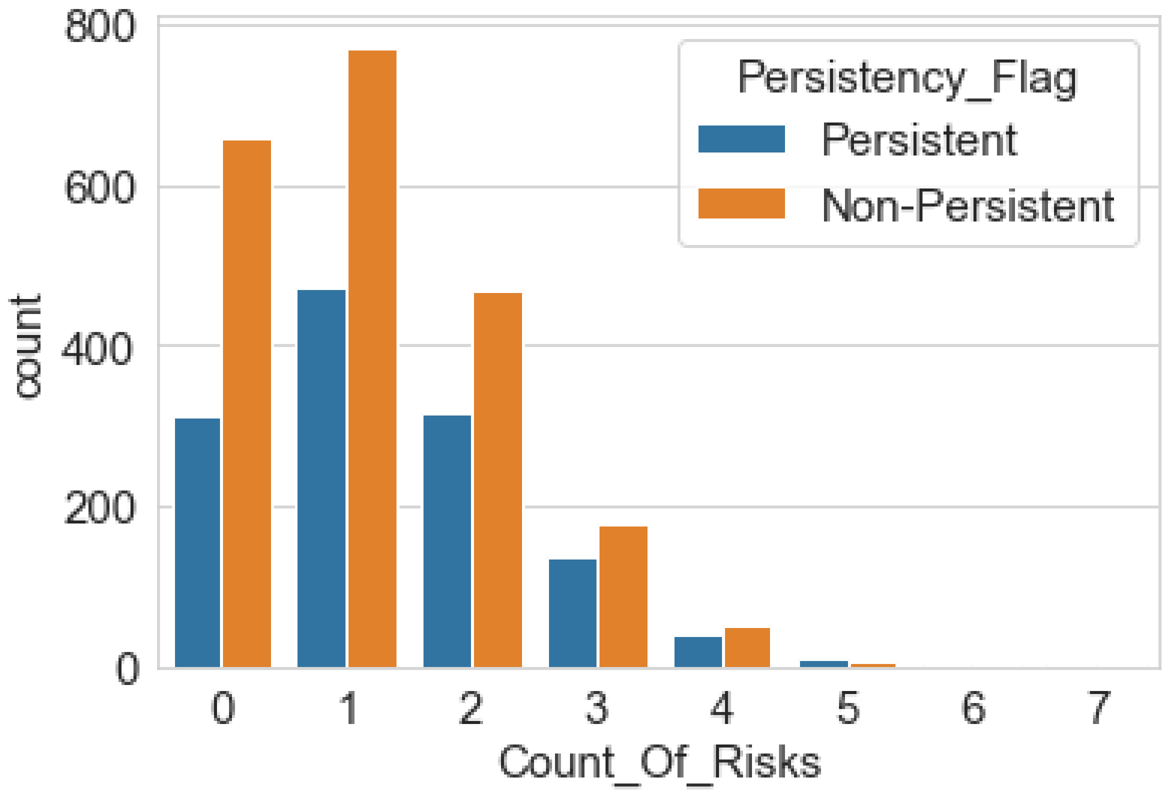
Comorbidity of Diseases

Risk Factor Analysis



Risk Factors

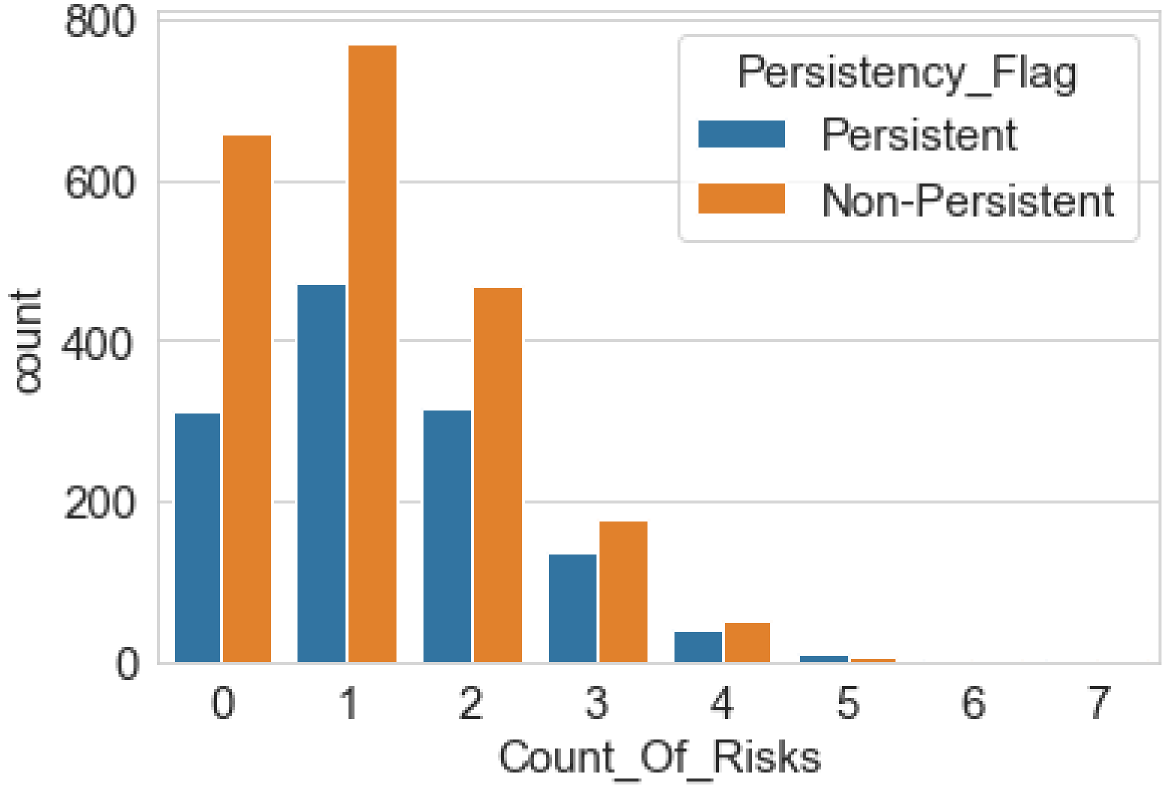
Risk Factor Analysis



Risk Counts Vs Persistency Flag

- High number of non persistent patients has less than 3 count of risks.
- Patients with more than 3 count of risks has the highest percentage of non-persistent cases compared to total registered cases.

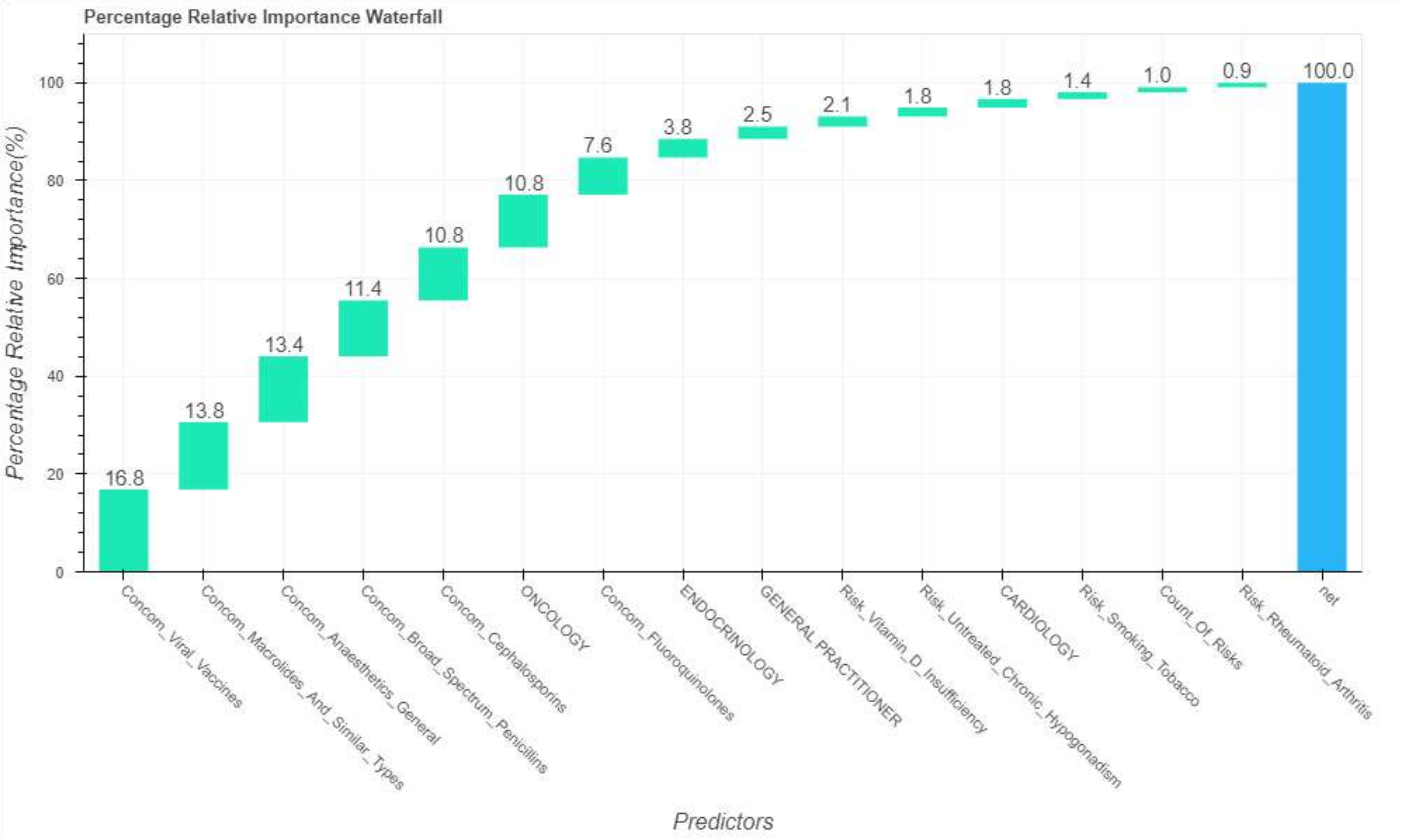
Risk Factor Analysis



Risk Counts Vs Persistency Flag

- High number of non persistent patients has less than 3 count of risks.
- Patients with more than 3 count of risks has the highest percentage of non-persistent cases compared to total registered cases.

Dominance Analysis



Dominance Analysis show most influential features in the data set (Most 15 influential factors).

It can be noticed that clinical parameters were the most influential factors behind persistency of drugs.

Recommendations

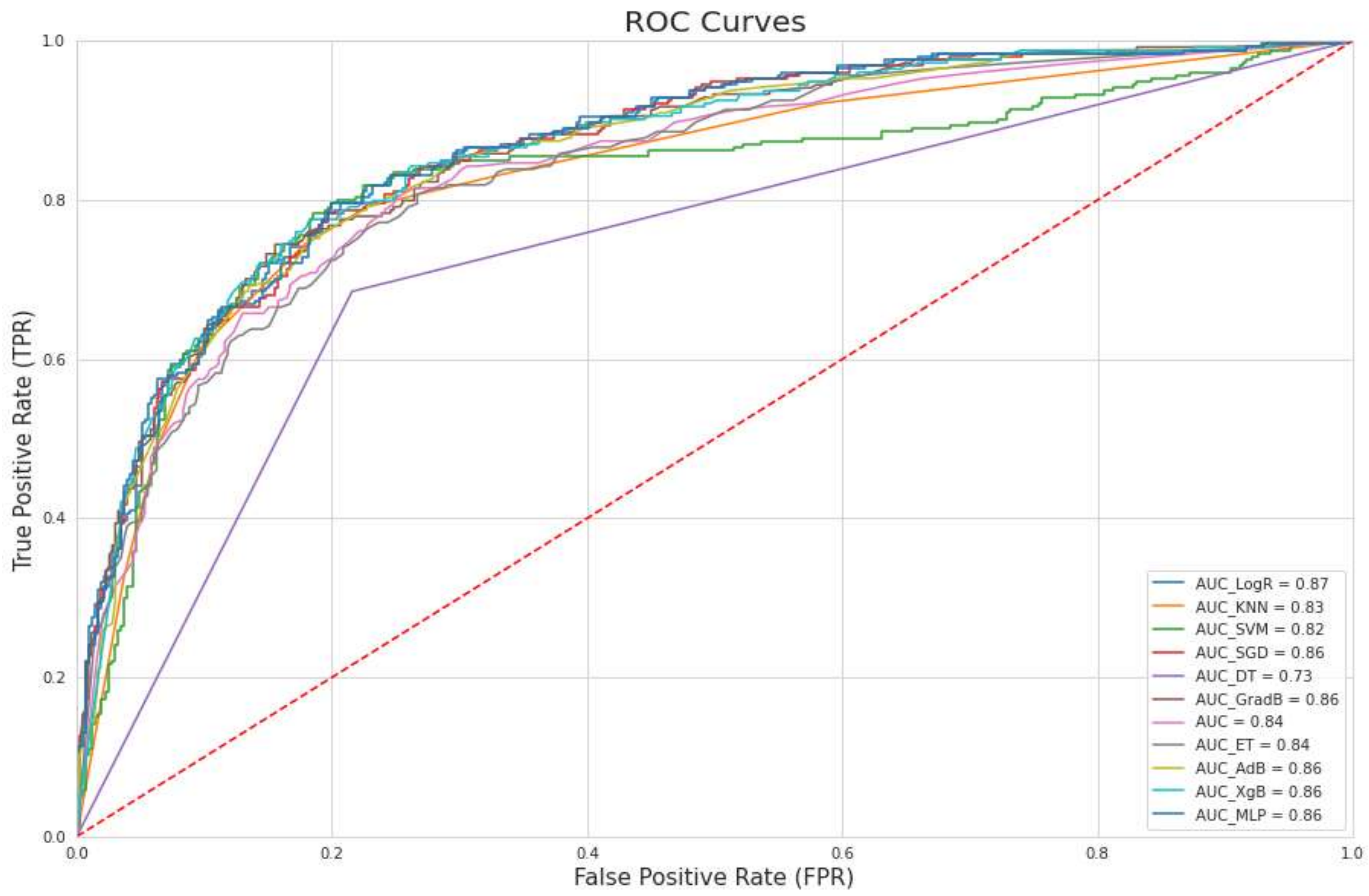
From the Exploratory Data Analysis (EDA) done on the dataset, following recommendations are given to the ABC company’s technical team:

- Demographic Factors provided in the dataset is not strongly related to the “Persistency Level” of the patients.
- NTM Specialist type or Specialist Flag did not show any correlation to the target variable.
- Some important parameters were determined using Dominance Analysis which can be used to transform the dataset into a subset and perform quantitative analysis.
- Clinical Factors such as “Concomitancy of Drugs”, “Comorbidity of Various Diseases” and “Risk Factors” do show some correlations with the target variable “Persistency Level” of the patients which needs to be investigated further through a Quantitative Analysis such as Machine Learning.

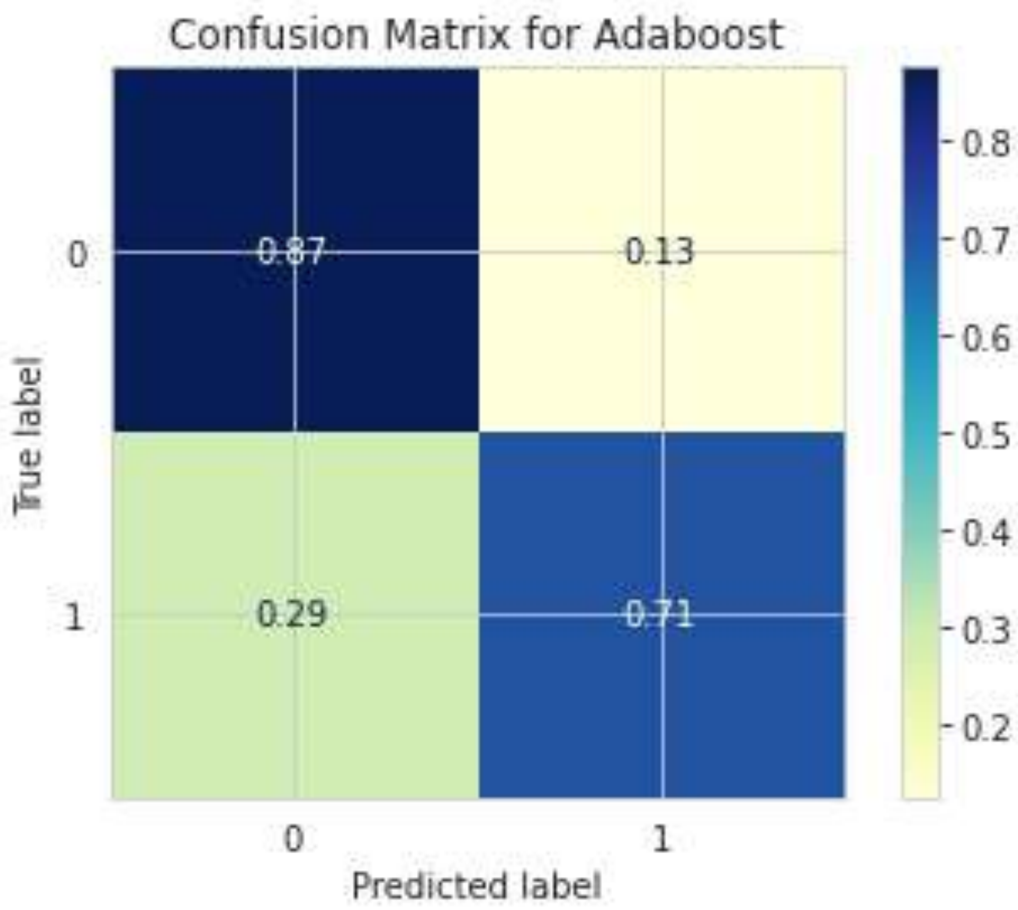
Recommendations

ML Algorithms	Original						Autoencoder						Dominance Analysis					
	MAE	Accuracy	Precision	Recall	f1-Score	AUC	MAE	Accuracy	Precision	Recall	f1-Score	AUC	MAE	Accuracy	Precision	Recall	f1-Score	AUC
Logistic Regression	0.19	0.81	0.81	0.79	0.81	0.88	0.20	0.80	0.79	0.78	0.79	0.87	0.27	0.73	0.73	0.68	0.72	0.76
K-Nearest Neighbour (KNN)	0.22	0.78	0.78	0.73	0.76	0.84	0.21	0.79	0.80	0.78	0.79	0.83	0.31	0.69	0.68	0.63	0.67	0.70
Support Vector Machine (SVM)	0.21	0.79	0.78	0.76	0.78	0.86	0.20	0.80	0.80	0.79	0.80	0.82	0.28	0.72	0.72	0.69	0.72	0.71
Stochastic Gradient Descent (SGD)	0.24	0.76	0.76	0.75	0.76	0.81	0.21	0.79	0.80	0.79	0.79	0.86	0.28	0.72	0.71	0.67	0.71	0.74
Decision Tree	0.27	0.73	0.73	0.71	0.73	0.71	0.25	0.75	0.75	0.73	0.75	0.73	0.35	0.65	0.63	0.59	0.63	0.62
Gradient Boosting	0.19	0.81	0.81	0.78	0.81	0.88	0.20	0.80	0.80	0.79	0.80	0.86	0.28	0.72	0.71	0.68	0.71	0.76
Random Forest	0.19	0.81	0.80	0.78	0.80	0.88	0.22	0.78	0.78	0.76	0.78	0.84	0.33	0.67	0.66	0.63	0.66	0.69
Extra Trees	0.21	0.79	0.79	0.77	0.79	0.87	0.23	0.77	0.77	0.75	0.77	0.84	0.34	0.66	0.64	0.60	0.64	0.67
AdaBoost	0.19	0.81	0.81	0.79	0.81	0.87	0.21	0.79	0.79	0.78	0.79	0.86	0.28	0.72	0.72	0.67	0.71	0.75
XgBoost	0.21	0.79	0.80	0.75	0.78	0.87	0.20	0.80	0.80	0.79	0.80	0.86	0.28	0.72	0.71	0.66	0.70	0.76
Multiple Layer Perceptron (MLP)	0.25	0.75	0.75	0.74	0.75	0.82	0.21	0.79	0.79	0.77	0.79	0.86	0.32	0.68	0.68	0.65	0.68	0.70
ANN Developed with KERAS								0.80	0.78	0.79	0.79							

Recommendations



Recommendations



Recommendations

Based on the provided data (mostly categorical), and the previous analysis we recommend 2 types of model to build for this problem:

- Neural Networks
- Adaptive Boosting (Ensemble)
- Gradient Boosting (Ensemble)

Both should be complex enough to learn well the data and provide high accuracy

Selected Pipeline: Autoencoder based feature extraction

Thank You

Data Glacier
Your Deep Learning Partner