**Group Name**: **Attack on Data**

**Team Members**:

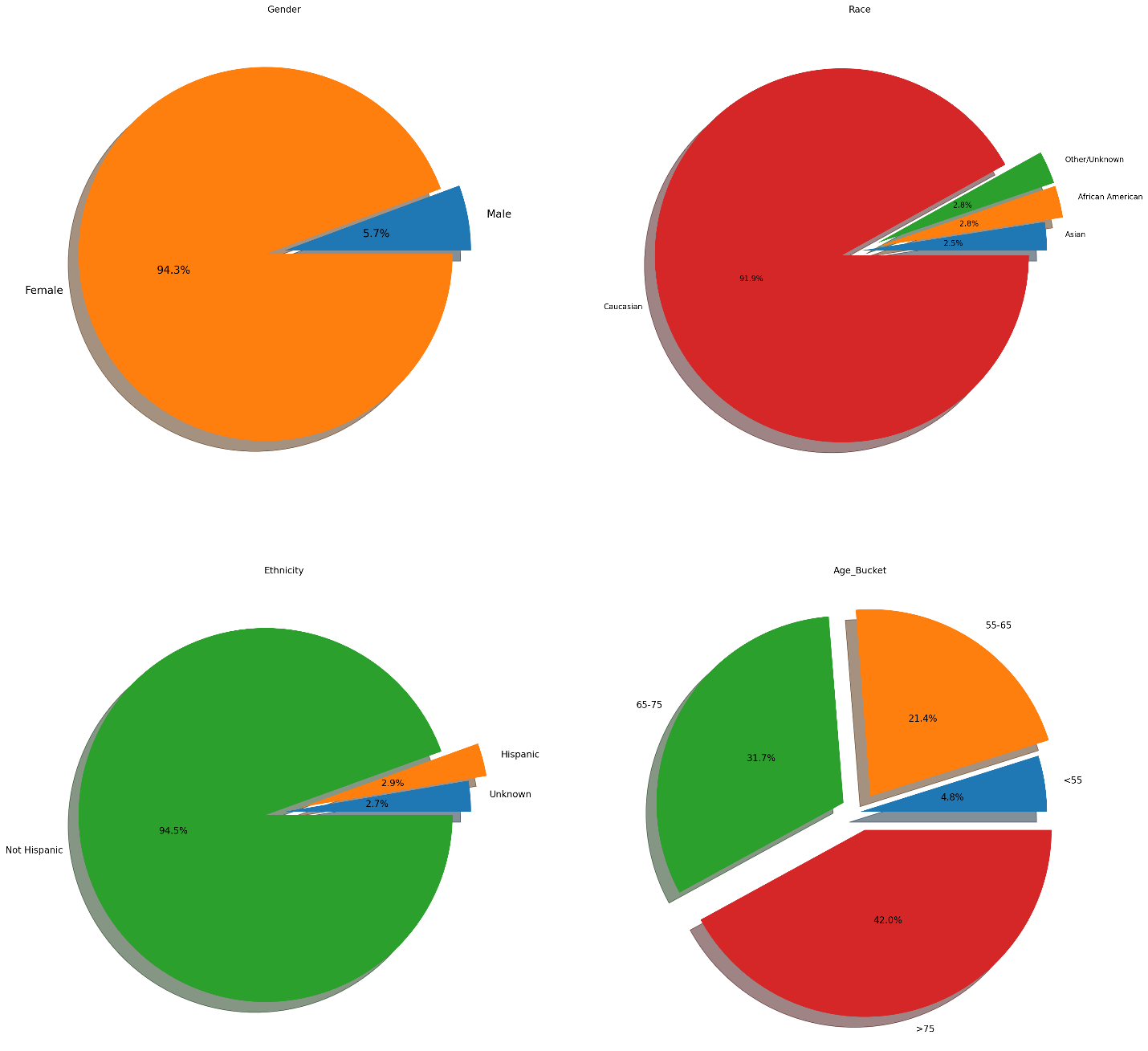
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name | Email | Country | College/Company | Specialization |
| Armin Khayati | [Northatlas@gmail.com](mailto:Northatlas@gmail.com) | UAE | - | Data Science |
| Ezzuldin Zaky | Ezzulding.zaky@gmail.com | UAE | American University of Sharjah | Data Science |

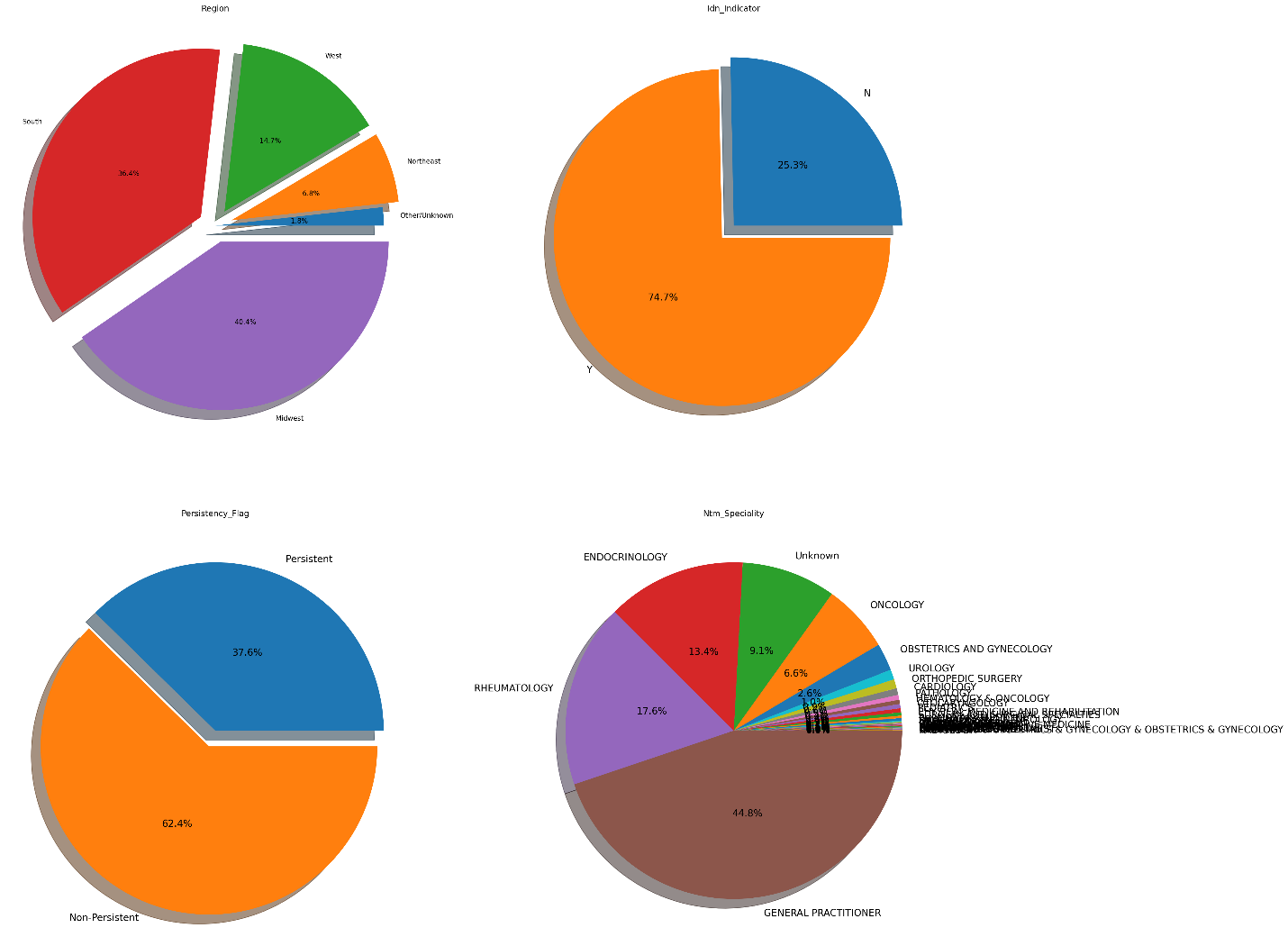
**Problem description**

One of the challenges 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.

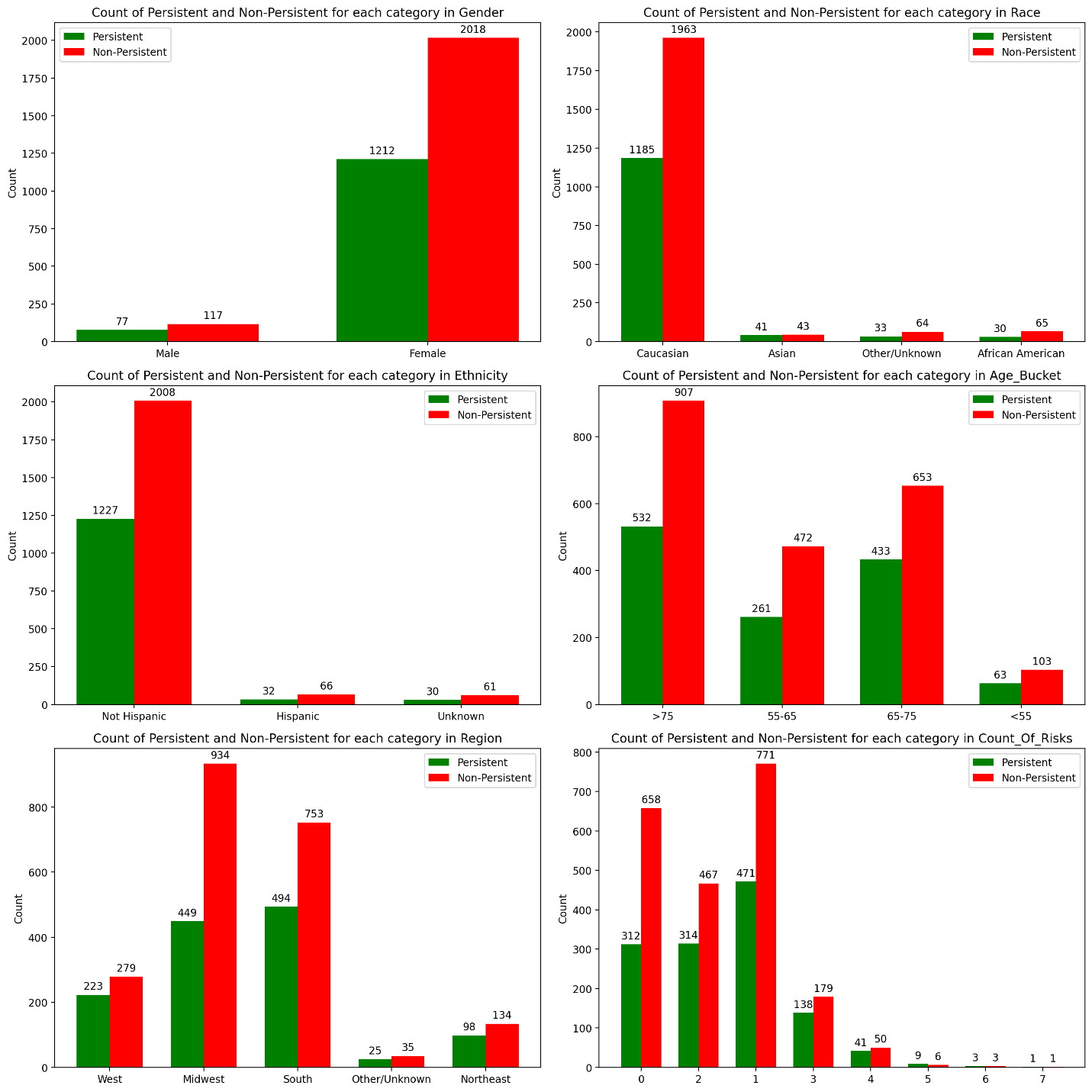
**Exploratory data analysis (EDA)**

Our dataset has 69 columns, 67 of them are categorical and two of them are numerical. There are no nan values in the dataset. The main problem of it is that some columns have rare labels. As we can see in the bellow charts rare categories is the main issue.

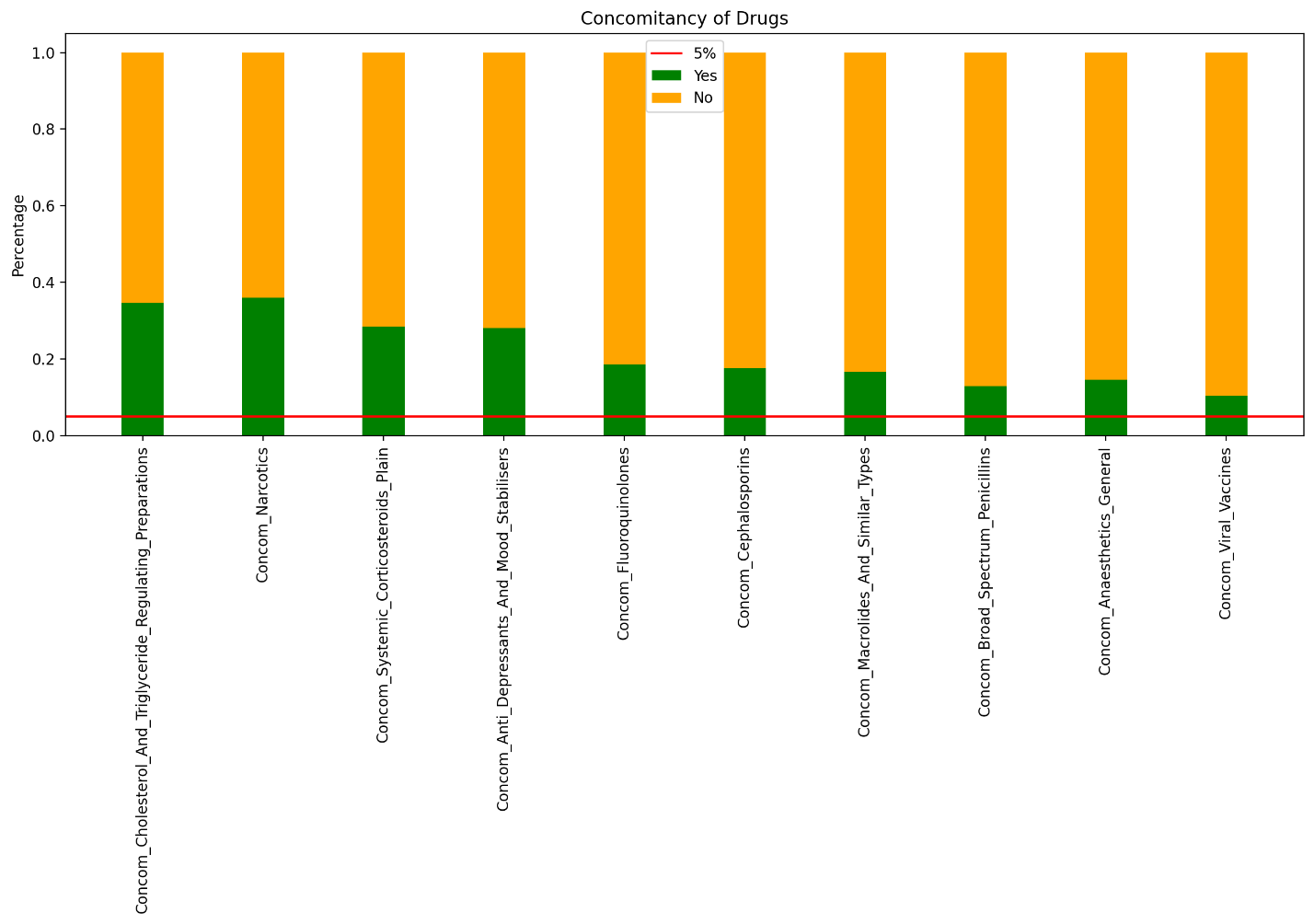




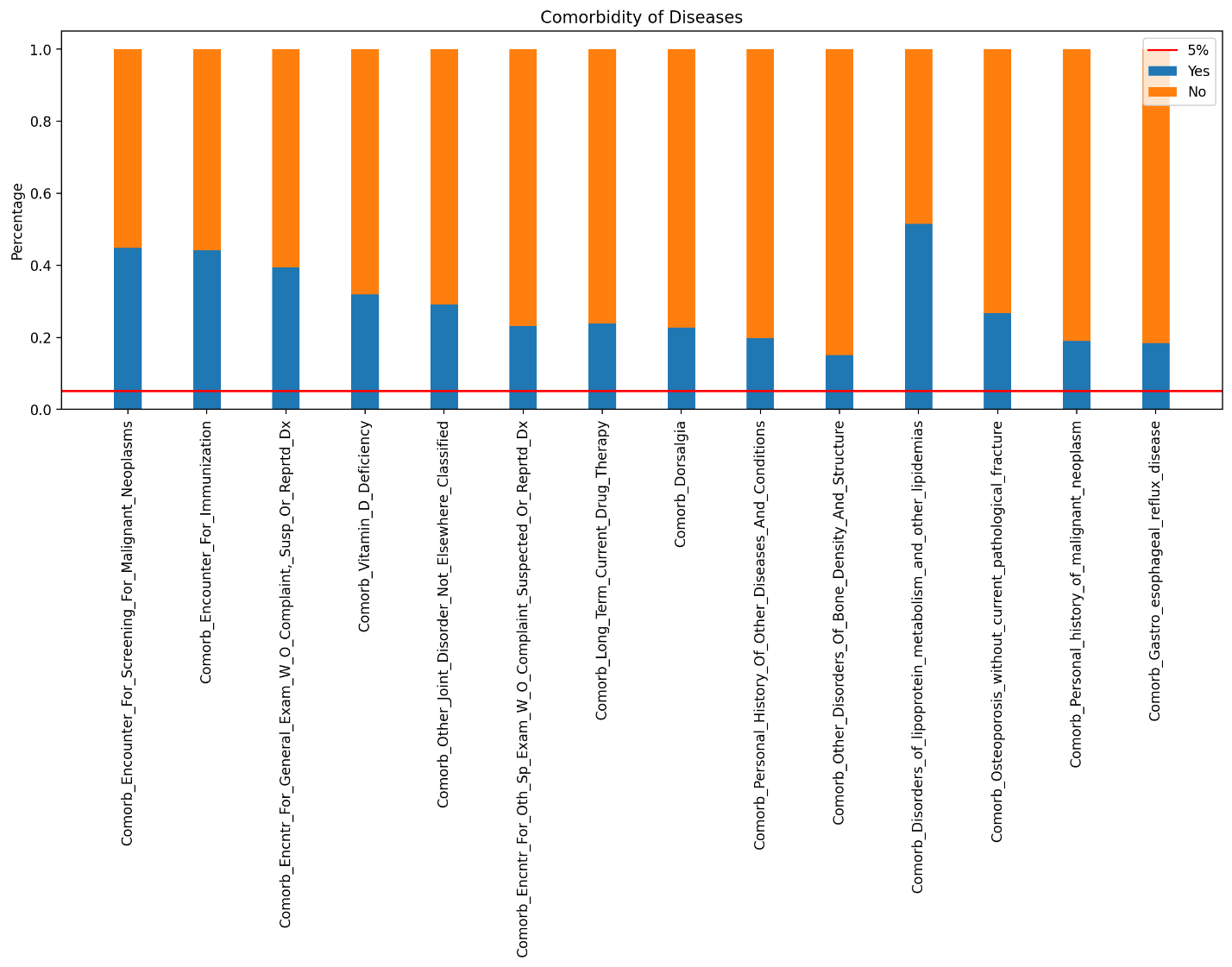
For gender variable 94% of values are female and only 5% are male and it means results for this dataset are mostly accurate for women. It is also the same for Caucasian Race and Not Hispanic Ethnicity. Also we can observe the imbalanceness of age bucket variable for younger patients. It can be seen than most people on the dataset belonged to the aged class, but the non-persistency level (or ratio) is more among the older patients. As discussed earlier, this study has been imbalanced towards female subjects but among females, the non-persistency level is higher than the males. But no concrete conclusion can be drawn due to the data imbalance. Also, low risk patients were found to be less persistent than the high-risk ones.



By a look at Concomitancy of Drugs bellow we can see that Cholesterol drug is more used than others beside the main drug. For other drugs, Concomitancy percentage is very low.



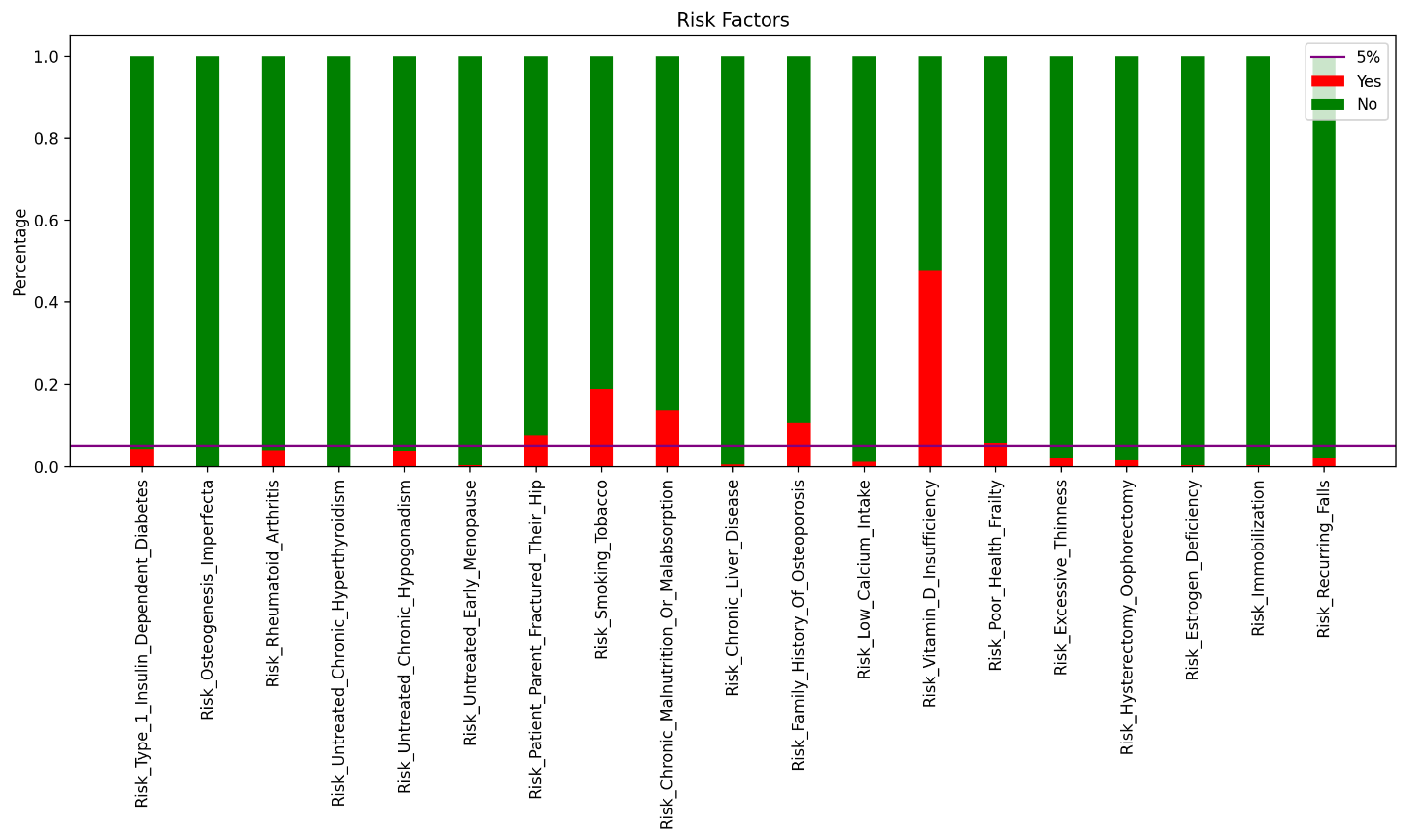
Also for Comorbidity of Diseases, we can see that disorders of lipoprotein have higher percentage to comorbid with main disease than other diseases and disorders of bone density is the lowest among all.



And among the risk factors most of them have less than 5% chance to endanger treatment. The risk factor with highest chance is Vitamin D Insufficiency and others above 5% are:

* Poor Health Frailty
* Family History of Osteoporosis
* Chronic Malnutrition or Malabsorption
* Smoking Tobacco
* Patient Parent Fractured Their Hip

Rest of the factors have less than 5% risk to endanger treatment.

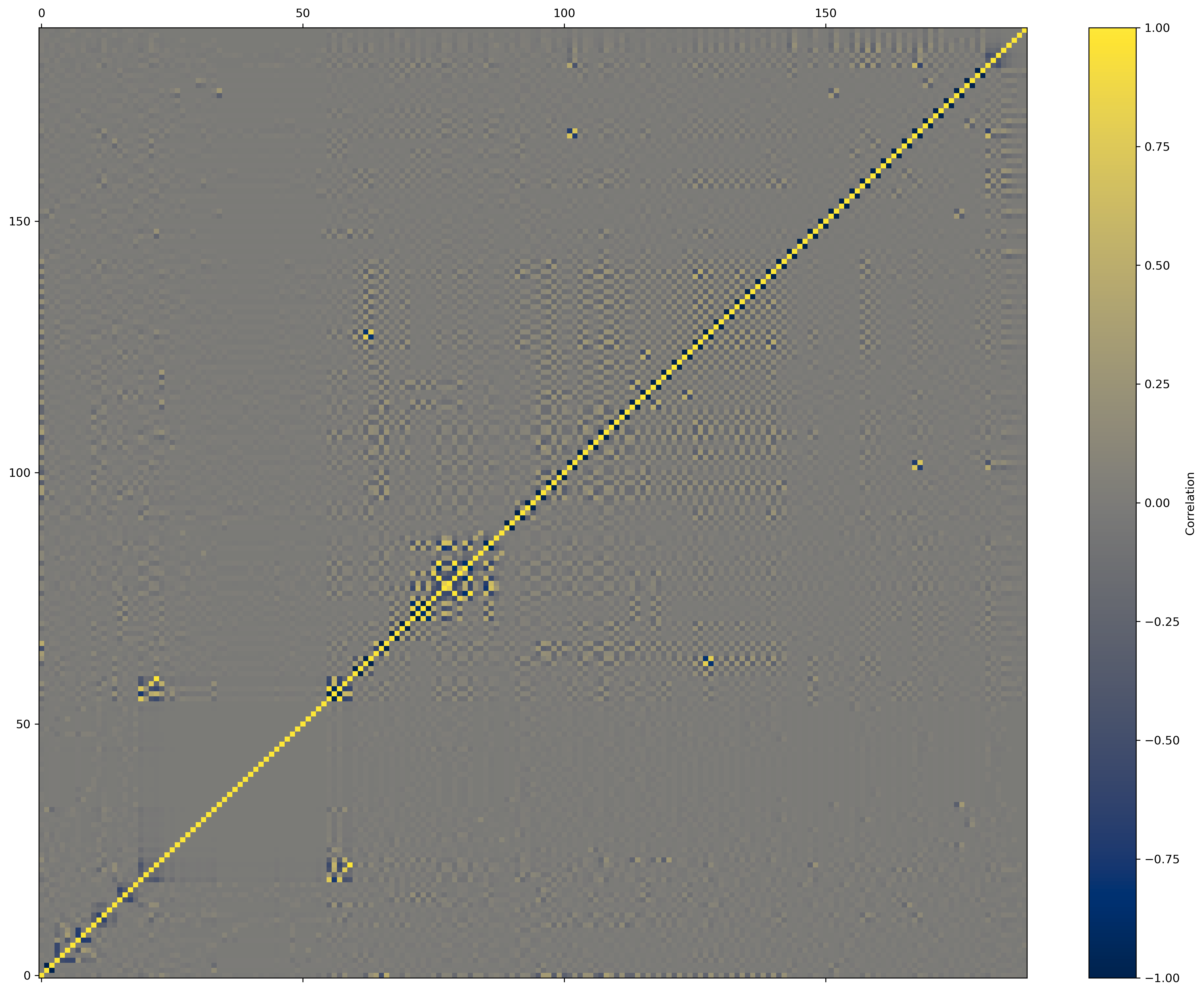


**Recommendation**

1. Handling **Unknown** values for Race, Region, and Ethnicity Variables
   * Using mode as an imputer as an imputer on ***Race*** and ***Ethnicity*** variables.
   * For ***Region*** variable, because most of the people with **Unknown** Region have **Not Hispanic** Ethnicity, and Most of people with Not Hispanic Ethnicity, have Midwest Region, we will replace Unknown Regions with **Midwest**.
2. Handling **Rare Labels**: Finding categories less than 5 percent in each variable, then merging those categories into one or drop them if the variable only has 2 categories (e.g., Y/N) and cardinality of one them is less than 5 percent.
3. Grouping integer values of Count\_Of\_Risks variable into two **bins**.

|  |  |  |
| --- | --- | --- |
| Group | Variable | Categories to Merge |
| Variables Chosen to Merge Categories | Race | African American, Asian |
| Age\_Bucket | <55, 55-65 |
| Ntm\_Speciality | OBSTETRICS AND GYNECOLOGY , UROLOGY , ORTHOPEDIC SURGERY , CARDIOLOGY , PATHOLOGY , HEMATOLOGY & ONCOLOGY , OTOLARYNGOLOGY , PEDIATRICS , PHYSICAL MEDICINE AND REHABILITATION , PULMONARY MEDICINE , SURGERY AND SURGICAL SPECIALTIES , PSYCHIATRY AND NEUROLOGY , NEPHROLOGY , ORTHOPEDICS , PLASTIC SURGERY , VASCULAR SURGERY , HOSPICE AND PALLIATIVE MEDICINE , GERIATRIC MEDICINE , GASTROENTEROLOGY , TRANSPLANT SURGERY , CLINICAL NURSE SPECIALIST , OCCUPATIONAL MEDICINE , HOSPITAL MEDICINE , OPHTHALMOLOGY , PODIATRY , EMERGENCY MEDICINE , RADIOLOGY , OBSTETRICS & OBSTETRICS & GYNECOLOGY & OBSTETRICS & GYNECOLOGY , NEUROLOGY , PAIN MEDICINE , NUCLEAR MEDICINE |
| Change\_T\_Score | Improved, Worsened |
| Change\_Risk\_Segment | Improved, Worsened |
| Count\_Of\_Risks | Bin 1 is [0,1,2,3] and bin 2 is [4,5,6,7] |
| Variables Chosen to Drop | Ethnicity |  |
| Risk\_Type\_1\_Insulin\_Dependent\_Diabetes |  |
| Risk\_Osteogenesis\_Imperfecta |  |
| Risk\_Rheumatoid\_Arthritis |  |
| Risk\_Untreated\_Chronic\_Hyperthyroidism |  |
| Risk\_Untreated\_Chronic\_Hypogonadism |  |
| Risk\_Untreated\_Early\_Menopause |  |
| Risk\_Chronic\_Liver\_Disease |  |
| Risk\_Low\_Calcium\_Intake |  |
| Risk\_Excessive\_Thinness |  |
| Risk\_Hysterectomy\_Oophorectomy |  |
| Risk\_Estrogen\_Deficiency |  |
| Risk\_Immobilization |  |
| Risk\_Recurring\_Falls |  |
| Dexa\_Freq\_During\_Rx |  |

1. **One hot encoding** all the variables after doing above tasks
2. Removing variables with more than **98% correlation**.



1. Try different machine learning approaches and select the best model.

**ML model evaluation and selection**

The following models were trained and tested on the processed dataset:

1. GradientBoostingClassifier
2. ExtraTreesClassifier
3. RandomForestClassifier
4. LogisticRegression
5. SVR (RBF kernel)
6. SVR (Linear kernel)
7. SVR (Polynomial kernel)
8. SVR (Sigmoid kernel)
9. XGBClassifier

Results:

For each model, the accuracy, precision, recall, f1-score and support were noted. The figures below illustrate

Chart, treemap chart

Description automatically generated

Chart, treemap chart

Description automatically generated

Chart, treemap chart

Description automatically generatedChart, treemap chart

Description automatically generated

Chart, treemap chart

Description automatically generatedChart, treemap chart

Description automatically generated

Chart, treemap chart

Description automatically generatedChart, treemap chart

Description automatically generated

Chart, treemap chart

Description automatically generated

**Final Recommendations**

The highest accuracies belonged to Gradient boosting, Extra Trees classifier and the random forest classifiers, having accuracies of 83%, 82% and 81% respectively. The random forest classifier is 1% more likely to make a false negative prediction on the drug persistency, which is favorable to the contrary. A physician can follow up on the patient that is predicted to stop the medication regardless of their actual drug persistency. The models also have the highest F1 scores: 0.85 0.85 and 0.84 respectively.

**Code Review**

|  |  |
| --- | --- |
| Code By | Review By |
| Armin Khayati | Ezzuldin Zaky |
| Ezzuldin Zaky | Orcun Sami Tandogan |
| Orcun Sami Tandogan | Armin Khayati |

**Github Repo**:

<https://github.com/Arminkhayati/dataglacier_internship>