Exploratory Data Analysis

HEALTH CARE – DRUG PERSISTENCY

20TH AUGUST 2022

Background

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 is seeking to automate this process of identification.

Problem Statement

- To understand the persistency of a drug as per the prescription given by the physician is an important question faced by pharmaceutical companies.
- The problem here is to build a classification model to understand the persistency (persistent or not) of a drug for the given dataset.

Data Analysis Approach

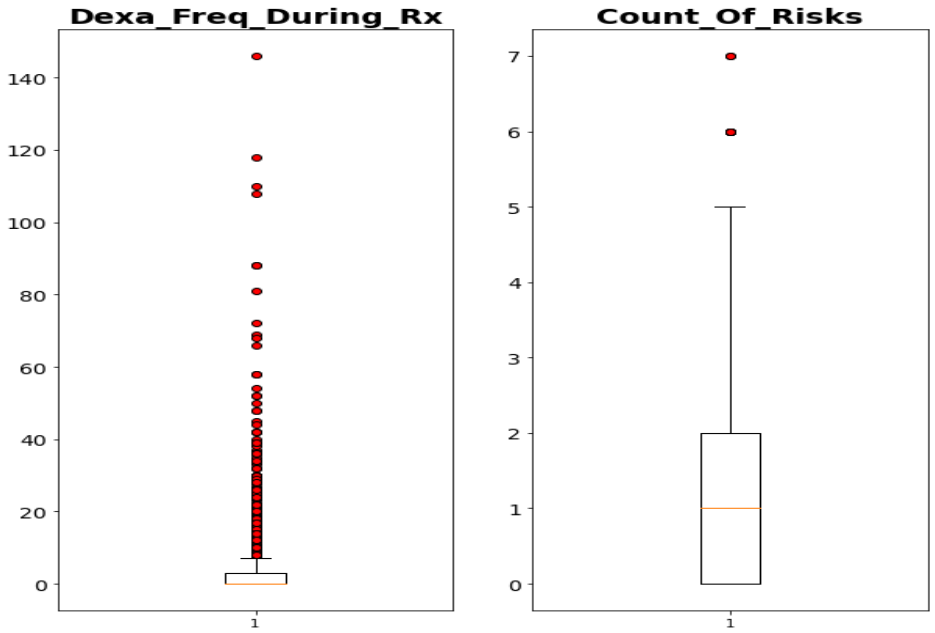
- Explore and Understand the data.
- Prepare and clean the data.
- Analyze the data and find the features/variables that affects drug persistency.
- Give recommendations for the classification model that is to be built

Data Exploration

- One file used for the dataset
- 3,424 data points
- 69 features/variables

Analysis of Data

Outliers



- The red colour dots are outliers.
- Various methods like IQR, Zscore, and Quantile filters are explained and outliers are removed by using IQR

Feature Selection

- Dropping of Unwanted Columns:
  - The columns 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_Patient_Parent_Fractured_Their_Hip ,Risk_Smoking_Tobacco, Risk_Chronic_Malnutrition_Or_Malabsorption, Risk_Chronic_Liver_Disease, Risk_Family_History_Of_Osteoporosis ,Risk_Low_Calcium_Intake, Risk_Vitamin_D_Insufficiency, Risk_Poor_Health_Frailty, Risk_Excessive_Thinness, Risk_Hysterectomy_Oophorectomy, Risk_Estrogen_Deficiency, Risk_Immobilization Risk_Recurring_Falls are summed to Count_Of_Risks column. Hence all the above columns can be dropped.
  - Column count reduced to 50.

Feature Selection

- Grouping of columns together
  - The Disease/Treatment Factor -the column name starts with 'Concom' or 'Comorb', this is grouped together by summing the number of positives for that and dropping the unwanted columns.
  - Column count reduced to 28

Numerical Conversion

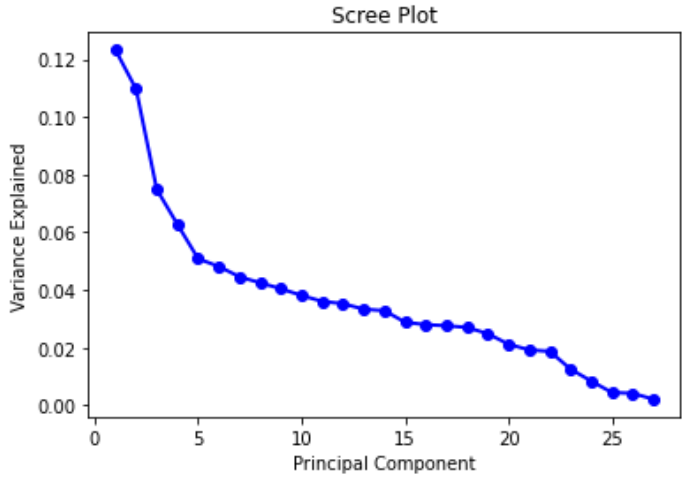
- Convert to Numerical Values:
  - Factorize method is used to convert the categorical columns to numerical.

Preprocessing

- Separate X and Y
- Divide into x_train, x_test, y_train, y_test.
- Apply Standard Scaler

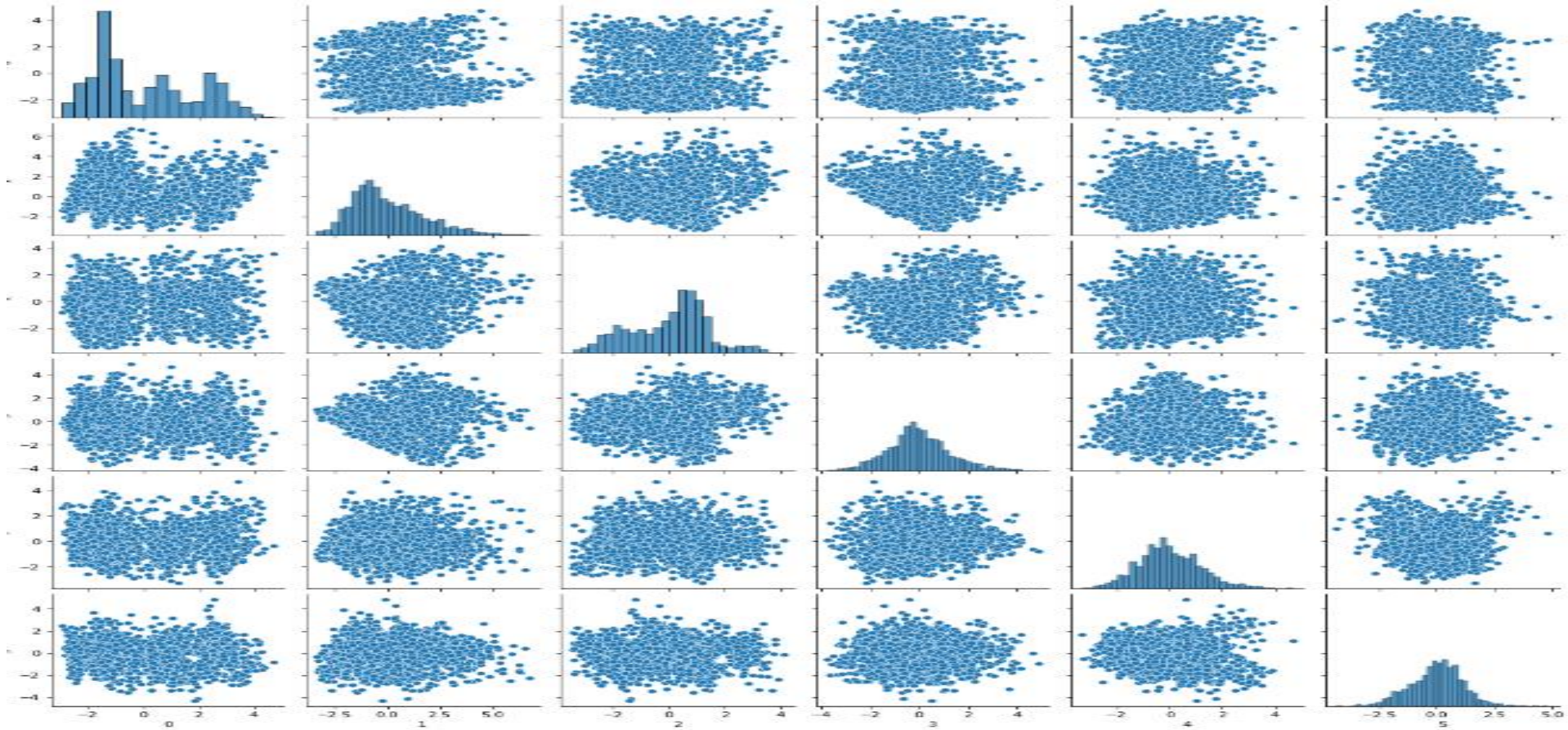
PCA

- Identify the number of components for the PCA by plotting scree plot.



- Apply PCA to reduce the dimensionality to 6.
- Total Column count after PCA is 6

Feature Plotting



EDA Summary

- From the Exploratory Data Analysis done, we are able to find how the different features/variables affects drug persistency.
- Dropped the unwanted columns
- The categorical variables are converted to numerical variables
- A total of 6 principal components can contribute a 0.4692567637481701% of variance.

Recommendations

For the purpose of automating the process of drug persistency identification, the following machine learning models can be used:

- **Logistic regression** – It is a type of linear model that is used for binary classification. It predicts output which is a categorical dependent variable. Such predictions are like yes or no, A or B, etc.
- **Random Forest** – It is a type of Bagging Ensemble Learning Classification. It predicts by the averaging of a number of decision tree classifiers.
- Adaboost and XGBoost – It is a type of Boosting model. It convert weak learners to strong learners.

Thank You