**EDA Report**

**Team Members: Apurwa Bhausaheb Sontakke, Vedant Wagh**

**Group Name - HealthData Innovators**

**Name – Vedant Wagh**

**Email – wagh.v@northeastern.edu**

**Country - USA**

**College– Northeastern University**

**Specialization – Data Science**

**Github Link -**

**Problem Description**

Pharmaceutical companies face a significant challenge in understanding why patients continue or discontinue their prescribed medications. To address this, ABC Pharma has sought the help of an analytics company to automate the identification process of factors influencing drug persistency. The aim is to develop a classification model that predicts whether a patient will persist with a prescribed drug (Persistency\_Flag).

Exploratory Data Analysis (EDA)

**Load the Dataset**

The dataset was successfully loaded and contains 3424 entries and 69 columns. The data includes various demographic, medical, and risk-related attributes.

**Data Cleaning**

Missing Values

An examination of the dataset revealed that there were no missing values present. Each column had the full 3424 entries, indicating that the dataset was complete with no gaps in the data.

**Descriptive Statistics**

Numerical Columns

Descriptive statistics for numerical columns were calculated to understand the central tendency, dispersion, and shape of the dataset’s distribution. Key statistics such as mean, standard deviation, minimum, maximum, and quartiles were obtained.

Key Findings:

Dexa\_Freq\_During\_Rx: Mean = 3.02, Std = 8.14, Max = 146.

Count\_Of\_Risks: Mean = 1.24, Std = 1.09, Max = 7.

Categorical Columns

Frequency counts for categorical columns were analyzed to understand the distribution of categorical data. This helped in identifying the most common categories and potential imbalances in the data.

**Data Visualization**

A graph with blue lines

Description automatically generated

Key Observations

Dexa\_Freq\_During\_Rx

Right-Skewed: Most patients have low Dexa frequency.

Outliers: Some patients have extremely high frequencies.

Count\_Of\_Risks

Multi-modal: Multiple peaks indicate distinct risk groups.

Common Risks: Most patients have 0 to 3 risk factors.

anomaly

Clear Separation: Most data points are normal, few are anomalies.

Further Steps

Investigate Outliers: Examine extreme values in Dexa\_Freq\_During\_Rx and Count\_Of\_Risks.

Cluster Analysis: Identify distinct patient risk groups.

Anomaly Analysis: Explore characteristics of anomalies.

Correlation Matrix:

A screenshot of a graph

Description automatically generated

Key Observations from the Correlation Matrix

Dexa\_Freq\_During\_Rx

Count\_Of\_Risks: Almost no correlation (0.014).

Anomaly: Moderate negative correlation (-0.5), suggesting higher Dexa frequency reduces anomaly likelihood.

Count\_Of\_Risks

Dexa\_Freq\_During\_Rx: Almost no correlation (0.014).

Anomaly

Dexa\_Freq\_During\_Rx: Moderate negative correlation (-0.5).

Summary

Low Correlation: Between Count\_Of\_Risks and Dexa\_Freq\_During\_Rx.

Moderate Negative Correlation: Higher Dexa\_Freq\_During\_Rx is associated with fewer anomalies.

Visualization of Outlier detection (Isolation Forest).

A blue rectangular bar graph

Description automatically generated

Key Observations

Dexa\_Freq\_During\_Rx

Outliers: Several outliers remain above 15.

Count\_Of\_Risks

Range: Most patients have 0 to 4 risk factors, median around 1.

anomaly

Outliers: Few anomalies remain

Visualization of Outlier detection (Local Outlier Factor).

A blue rectangular bar graph

Description automatically generated

Key Observations

Dexa\_Freq\_During\_Rx

Outliers: Several outliers remain, with values above 15 and up to 30.

Count\_Of\_Risks

Range: Most patients have 0 to 4 risk factors, with a median around 1.

anomaly

Outliers: Few anomalies remain, identified below -1.

Summary

Dexa\_Freq\_During\_Rx: Outliers persist, ranging from 15 to 30.

Count\_Of\_Risks: Typically, 0 to 4 risk factors, median at 1.

Anomaly: Isolated anomalies detected below -1.