Team Member Details

Group Name: *SoloVision Analytics*

Name: Ankita Roy
Email: roy.anki39@gmail.com
Country: India
College/Company: Institute: Great Learning, Gurgaon/ Internship: Data Glacier
Specialization: Data Science

Problem Description

Drug persistency is a key metric in determining the long-term success of a treatment regimen. It reflects whether patients continue using a prescribed therapy as advised. Pharmaceutical companies like ABC Pharma often struggle to identify the factors that cause patients to discontinue treatment.

To address this, ABC Pharma has sought the help of data science to automate the detection of persistency patterns among patients using their drug. The aim is to develop a **classification model** that predicts whether a patient is likely to persist with the treatment based on demographic, clinical, and behavioural data.

The binary target variable for this task is **Persistency_Flag**, where the goal is to classify patients as either **persistent** or **non-persistent**.

EDA Performed on the Data

- **Univariate Analysis:**
  - For the target variable Persistency_Flag, created a pie chart to visualize the proportion of persistent vs non-persistent patients
  - For numerical columns like Dexa_Freq_During_Rx, used displot and boxplot to study the distribution, detect outliers, and understand skewness
  - For categorical columns (e.g., Gender, Race, Ethnicity, Region), used countplots to see category counts and detect any data imbalance
- **Bivariate Analysis:**
  - For numeric vs category relationships (e.g., Dexa frequency by persistency group), used displots and boxplots to compare distributions between persistent and non-persistent groups
  - For category vs category (e.g., Persistency with all categorical columns), used grouped bar plots to compare how different categories relate to the target
- **Statistical Testing:**
  - Performed hypothesis testing on Dexa_Freq_During_Rx vs Persistency_Flag to check if higher Dexa scan frequency is linked with higher persistency
  - Tested normality with Shapiro-Wilk – p_value=0.0 – not normal
  - Tested equal variance with Levene's test – p_value=7.2e-97 – unequal variances

- Used Mann-Whitney U test – t_statistics = 2024137.0, p_value = 4.7e-191 – Reject Ho, concluding that higher Dexa frequency is statistically linked to higher persistency

---

Final Recommendation

- The initial EDA suggests that the dataset has clear patterns between clinical activity (like Dexa scan frequency) and drug persistency
- The pie chart confirms a mild class imbalance, which will be accounted for in modelling by monitoring precision/recall
- Based on the results of the Mann-Whitney U test, Dexa scan frequency could be an important predictive feature for the final model
- Going forward, the next steps will include:
  - Finalizing feature selection
  - Applying appropriate scaling or transformations
  - Training multiple classification models
  - And deploying the best performing model

All visualizations and statistical tests are available in the Jupyter notebook and linked GitHub repository

---

GitHub Repository Link
https://github.com/Ankita-ar/PERSISTENCY-OF-A-DRUG