HEALTHCARE---PERSISTENCY-OF-A-DRUG

Business Overview: ABC Pharma, like many other pharmaceutical companies, is facing challenges in understanding whether patients adhere to prescribed medication regimens. This challenge is particularly significant when trying to track how persistently patients follow their physician's prescription for specific drugs. To gain actionable insights, ABC Pharma has decided to leverage data analysis to explore key factors influencing drug persistency.

Problem Definition: ABC Pharma has approached OpenML to analyze factors that influence whether a patient will continue taking the prescribed medication throughout the full treatment period. The aim is to build a model that predicts whether a patient, based on their demographics and medical data, will adhere to their prescribed medication plan. A dataset containing patient information has been provided to support this analysis.

Objective: The core objective of this project is to develop a classification model that accurately predicts whether a patient will adhere to their prescribed drug regimen. By doing so, the company aims to better understand patient behavior and improve treatment outcomes. The analysis will be used to create a predictive model, which will be integrated into a web-based application for ABC Pharma to facilitate real-time predictions.

Data Summary: The dataset provided contains 69 variables, encompassing demographic details, provider attributes, clinical factors, and disease/treatment-related factors. The target variable is the "Persistency Flag," which indicates whether a patient adhered to the prescribed treatment plan. In total, there are 3,424 patient records available for analysis.

Modeling Approach: The aim is to build a classification model that can predict patient adherence based on the given data. The process will involve:

- Cleaning the dataset for issues such as null values, incorrect data formats, outliers, and missing data.
- Conducting exploratory data analysis to identify patterns and trends that could influence model performance.
- Building and evaluating machine learning models to achieve the best prediction accuracy.
- Integrating the final model into a web application to provide ABC Pharma with a tool for predicting patient adherence.

Project Goal: The final deliverable will be a web application that allows ABC Pharma to input patient data and receive predictions on whether the patient is likely to adhere to their prescribed drug treatment. This tool will help the company optimize patient outcomes and refine its strategies for improving drug persistency.

Project Timeline: The project will follow a structured lifecycle, with key milestones including data cleaning, model development, evaluation, and application deployment. Expected deadlines will be set based on the project's progression.

Summary Table: Missing Value Handling

Variable	Imputation Strategy	Missing Value Percentage
Race	Use mode ("Caucasian")	2.83%
Region	Use mode for "Not Hispanic" group	1.75%
Ethnicity	Use mode ("Not Hispanic")	2.66%
Ntm_Speciality	Keep "Unknown" or use mode ("General Practitioner")	9.05%
NTM Variables (Injectable, Risk, Comorbidity, Concomitancy)	Convert "Y" to 1, "N" to 0	No missing values

Risk Segment During Rx, Tscore Bucket, Change T-Score, Change Risk Segment	Eliminate due to high missing value percentage (>40%)	43%+

□ **Race Variable - Missing Values:**

- **Imputation Strategy:** Use the mode to fill in missing values. This approach is justified by the following:
    - Only 2.83% of the data (97 instances out of 3,424) have missing values categorized as "Other/Unknown."
    - The mode ("Caucasian") represents 91.94% of the data, making it a reliable imputation choice. Additionally:
        - When grouped by ethnicity, the mode accounts for 93.45% of the "Not Hispanic" group and 61.22% of the "Hispanic" group.
    - Given the small percentage of missing values and the dominance of the mode, it is safe to treat "Other/Unknown" as the mode.

□ **Region Variable - Missing Values:**

- **Imputation Strategy:** Use the mode for the "Not Hispanic" ethnicity group. Justifications include:
    - All missing values in the "Region" variable correspond to individuals classified as "Not Hispanic."
    - Given this alignment, imputing the missing values with the mode for this group ensures consistency.

□ **Ethnicity Variable - Missing Values:**

- **Imputation Strategy:** Use the mode to fill missing values. This is supported by:
    - The mode accounts for 94.48% of the data (3,235 instances out of 3,424).
    - With only 2.66% of the data having missing values, using the mode is considered a safe and reasonable imputation strategy.

□ **Ntm_Speciality Variable - Missing Values:**

- **Imputation Strategy:** Two potential approaches:
    - **Approach 1:** Keep "Unknown" as a category since it represents less than 10% of the data. This will allow for analysis of how "Unknown" correlates with other variables.
    - **Approach 2:** Impute missing values with the mode, which is "General Practitioner."

□ **Handling Categorical Data for NTM Variables (Injectable Experience, Risk Factors, Comorbidity, and Concomitancy):**

- **Imputation Strategy:** These variables are binary ("Y" or "N"). To handle these:
    - Convert "Y" to 1 and "N" to 0 to standardize the data for modeling purposes.

□ **Risk Segment During Rx, Tscore Bucket During Rx, Change T-Score, and Change Risk Segment - Missing Values:**

- **Imputation Strategy:** Due to the large proportion of missing data (over 40%), these variables will be eliminated from the analysis. This ensures the model is not negatively impacted by excessive missing values.

Github Link: https://github.com/shradhanjalipradhan/HEALTHCARE---PERSISTENCY-OF-A-DRUG