Name: Yusuf Aishat
Email: aishatyusuf12@gmail.com
Country: Nigeria
Specialization: Data Science

PROBLEM DESCRIPTION

ABC Pharma wants to automate the identification of drug persistence as per the physician's prescription. The goal is to build a classification model to predict whether a patient will persist with their treatment and to identify the factors that influence drug persistence.

DATA UNDERSTANDING

The dataset provided tracks patient persistence with their NTM (Non-Tuberculous Mycobacterial Disease) treatment. The dataset contains 3424 records, each representing patients with various demographics, attributes of their health care provider, clinical factors, disease, and treatment factors.

Total number of rows: 3424
Total number of columns: 69

Bucket	**Feature**	**Feature Description**
Unique Row Id	Ptid	Data Type: Object Missing Values: none
Target Variable	Persistency_Flag	Data Type: Object

		Unique Values (in order of frequency): ● ‘Non-Persistent’: 62.35% ● ‘Persistent’: 37.64% Missing Values: none
Demographics	Age_Bucket	Data Type: Object Unique Values (in order of frequency): ● ‘>75’: 42.02% ● ‘65-75’: 31.71% ● ‘55-65’: 21.40% ● ‘<55’: 4.84% Missing Values: none
	Race	Data Type: Object Unique Values (in order of frequency): ● ‘Caucasian’: 91.93% ● ‘Other/Unknown’: 2.83% ● ‘African American’: 2.77% ● ‘Asian’: 2.45% Missing Values: ‘Other/Unknown’
	Region	Data Type: Object Unique Values (in order of frequency): ● ‘Midwest’: 40.39% ● ‘South’: 36.41%

		● ‘West’: 14.66% ● ‘Northeast’: 6.77% ● Other/Unknown: 1.75% Missing Values: ‘Other/Unknown’
	Ethnicity	Data Type: Object Unique Values (in order of frequency): ● ‘Not Hispanic’: 94.48% ● ‘Hispanic’: 2.86% ● ‘Unknown’: 2.65% Missing Values: ‘Unknown’
	Gender	Data Type: Object Unique Values (in order of frequency): ● ‘Female’: 94.33% ● ‘Male’: 5.66% Missing Values: none
	Idn_Indicator	Data Type: Object Unique Values (in order of frequency): ● ‘Y’: 74.67% ● ‘N’: 25.32% Missing Values: none
Provider Attributes	Ntm_Specialty	Data Type: Object Unique Values: 'GENERAL PRACTITIONER', 'Unknown',

		'ENDOCRINOLOGY', 'RHEUMATOLOGY', 'ONCOLOGY', 'PATHOLOGY', 'OBSTETRICS AND GYNECOLOGY', 'PSYCHIATRY AND NEUROLOGY', 'ORTHOPEDIC SURGERY', 'PHYSICAL MEDICINE AND REHABILITATION', 'SURGERY AND SURGICAL SPECIALTIES', 'PEDIATRICS', 'PULMONARY MEDICINE', 'HEMATOLOGY & ONCOLOGY', 'UROLOGY', 'PAIN MEDICINE', 'NEUROLOGY', 'RADIOLOGY', ‘GASTROENTEROLOGY', 'EMERGENCY MEDICINE', 'PODIATRY', 'OPHTHALMOLOGY', 'OCCUPATIONAL MEDICINE', 'TRANSPLANT SURGERY', 'PLASTIC SURGERY',

		'CLINICAL NURSE SPECIALIST', 'OTOLARYNGOLOGY', 'HOSPITAL MEDICINE', 'ORTHOPEDICS', 'NEPHROLOGY', 'GERIATRIC MEDICINE', 'HOSPICE AND PALLIATIVE MEDICINE', 'OBSTETRICS & OBSTETRICS & GYNECOLOGY & OBSTETRICS & GYNECOLOGY', 'VASCULAR SURGERY', 'CARDIOLOGY', 'NUCLEAR MEDICINE' , Missing Values: ‘Unknown’
	Ntm_Specialist_Flag	Data Type: Object Unique Values (in order of frequency): ● ‘Others’: 58.79% ● ‘Specialist’: 41.21% Missing Values: none
	Ntm_Specialist_Bucket	Data Type: Object Unique Values (in order of frequency): ● ‘OB/GYN/Others/PCP/Unknown’: 61.44%

		● ‘Endo/Onc/Uro’: 20.91% ● ‘Rheum’: 17.64 Missing Values: none
Clinical Factors	Gluco_Record_Prior_Ntm	Data Type: Object Unique Values (in order of frequency): ● ‘N’: 76.48% ● ‘Y’: 23.51% Missing Values: none
	Gluco_Record_During_Rx	Data Type: Object Unique Values (in order of frequency): ● ‘N’: 73.65% ● ‘Y’: 26.34% Missing Values: none
	Dexa_Freq_During_Rx	Data Type: Integer , Mean: 3.01 Standard deviation: 8.13 Min: 0 25%: 0 50%: 0 75%: 3 Max: 146 Missing Values: none
	Dexa_During_Rx	Data Type: Object Unique Values (in order of frequency): ● ‘N’: 72.66%

		● ‘Y’: 27.33% Missing Values: none
	Frag_Frac_Prior_Ntm	Data Type: Object Unique Values (in order of frequency): ● ‘N’: 83.87% ● ‘Y’: 16.12% Missing Values: none
	Frag_Frac_During_Rx	Data Type: Object Unique Values (in order of frequency): ● ‘N’: 87.82% ● ‘Y’: 12.17% Missing Values: none
	Risk_Segment_Prior_Ntm	Data Type: Object Unique Values (in order of frequency): ● ‘VLR_LR’: 56.39% ● ‘HR_VHR’: 43.60% Missing Values: none:
	Tscore_Bucket_Prior_Ntm	Data Type: Object Unique Values (in order of frequency): ● ‘>-2.5’: 56.98% ● ‘<=2.5’: 43.01% Missing Values: none
	Risk_Segment_During_Rx	Data Type: Object Unique Values (in order of frequency):

		● ‘Unknown’: 43.72% ● ‘HR_VHR’: 28.18% ● ‘VLR_LR’: 28.09% Missing Values: ‘Unknown’:
	Tscore_Bucket_During_Rx	Data Type: Object Unique Values (in order of frequency): ● ‘Unknown’: 43.72 ● ‘<=2.5’: 29.70% ● ‘>-2.5’: 26.57% Missing Values: none
	Change_T_Score	Data Type: Object Unique Values (in order of frequency): ● ‘No change’: 48.48% ● ‘Unknown’: 43.72% ● ‘Worsened’: 5.05% ● Improved: 2.74% Missing Values: ‘Unknown’:
	Change_Risk_Segment	Data Type: Object Unique Values (in order of frequency): ● ‘Unknown’: 65.09% ● ‘No change’: 30.72% ● ‘Worsened’: 3.53% ● Improved: 0.64% Missing Values: ‘Unknown’:

Disease/Treatment Factor	Injectable_Experience_During_Rx	Data Type: Object Unique Values (in order of frequency): ● ‘Y’: 89.25% ● ‘N’: 10.74% Missing Values: none
	NTM - Risk Factors (19 risk factor columns)	Data Type: Object Unique Values:‘Y’, ‘N’ Missing Values: none
	Count_Of_Risks	Data Type: Integer, Mean: 1.23 Standard deviation: 1.09 Min: 0 25%: 0 50%: 1 75%: 2 Max: 7 Missing Values: none
	NTM - Comorbidity (14 comorbidity columns)	Data Type: Object Unique Values:‘Y’, ‘N’ Missing Values: none
	NTM - Concomitancy (10 concomitancy columns)	Data Type: Object Unique Values:‘Y’, ‘N’ Missing Values: none
	Adherent_Flag	Data Type: Object Unique Values (in order of frequency):

		● ‘Adherent’: 94.94% ● ‘Non-Adherent’: 5.05% Missing Values: none

DATA QUALITY ISSUES

- **Race**: ~2.83% of entries are labeled Other/Unknown. This entries will be relabeled as “Other” because there are other races outside of Caucasian, African American and Asian

- **Ethnicity**: ~2.65% were entered as Unknown. These will be replaced by the mode (“Not Hispanic”) which is about 94.5%

- **Region**: ~1.75% are labeled as “Other/Unknown”. These will also be replaced by the mode (“Midwest”)

- **NTM Speciality**: At approximately 9.05%, the Unknown category is quite a sizeable portion and for now it will be left as is.
  An entry labeled as “Obstetrics & Obstetrics & Gynecology & Obstetrics & Gynecology” seems to be a data entry error and it will be added to the “Obstetrics and Gynecology” category.

- **Risk_Segment_During_Rx, Change_Risk_Segment, Tscore_Bucket_During_Rx, Change_T_Score**:
  These attributes have >40% of their entries as Unknown, as they are not adding much information to the data, they will be removed.

GITHUB REPOSITORY:
https://github.com/aishatyusuf/drug_persistence_abc_pharma