ABC Pharma Persistency Classification Project

Team Member Details

Group Name	Name	Email	Country	College/Company	Specialization
SoloAI	Rishabh Raman	internships.apply1@gmail.com			Data Science

Problem Description

ABC Pharma seeks to identify whether a patient is likely to remain persistent with therapy based on various demographic, clinical, and physician-level features. This predictive task will help streamline physician targeting and improve patient adherence strategies.

- Machine Learning Task: Binary Classification
- Target Variable: Persistency_Flag

Business Understanding

High therapy dropout rates lead to poor patient outcomes and wasted pharmaceutical investment. By modeling persistency:
- ABC Pharma can allocate resources more efficiently
- Physicians can intervene early with at-risk patients
- Marketing strategies can focus on high-impact areas

Project Lifecycle & Deadlines

Phase	Description	Target Deadline
Problem Understanding	Translate business need into ML goal	Completed
Data Understanding	Review structure, missingness, distributions	Completed
Data Cleaning & Engineering	Encode, normalize, impute, feature select	In Progress

Model Development	Train/test multiple classifiers	June 24
Model Evaluation	Compare accuracy, precision, recall, ROC-AUC	June 25
Deployment	Export and serve best model	June 26
Final Report Submission	Submit PDF + GitHub	June 27

Data Intake Report

Dataset Overview:

- Records: 3424
- Features: 68 predictors + 1 target (Persistency_Flag)
- Types: Mostly categorical/binary, 2 numerical

Key Feature Categories:

- Demographics: Age, Race, Region, Gender
- Clinical Factors: T-Score, Risk Segment, Dexa scans, fractures
- Comorbidities & Risks: 30+ binary flags
- Drug History: Concomitant meds, glucocorticoids

Cleaning Actions:

- No missing values found in initial load
- Mapped "Yes"/"No" to binary 1/0
- Planning One-Hot Encoding for multicategory fields

Target Balance (to be confirmed in EDA):

- Persistency_Flag: Binary (distribution yet to be explored)

Model Plan

- Baseline: Logistic Regression
- Tree-Based: Random Forest, XGBoost
- Other: SVM or KNN for benchmarking

Evaluation Metrics:

- Accuracy
- Precision / Recall (for both classes)
- ROC-AUC Score

Model Selection Rationale & Challenges

- High dimensionality from many binary features
- Need to guard against overfitting
- Explainability for pharma industry compliance
- Class imbalance (if found) will require SMOTE or class weights

GitHub Repository

GitHub Link: Pending

Will include:
- Full codebase
- Cleaned dataset reference (if allowed)
- Trained model + deployment script
- README with usage instructions

Prepared by: Rishabh Raman
Email: internships.apply1@gmail.com
Specialization: Data Science