Data Intake Report

Name: Data Science: Healthcare – Persistency of a Drug

Report date: 19/05/2025

Internship Batch: LISUM44

Version:

Data intake by: Ankita Roy

Data intake reviewer:

Data storage location: <https://drive.google.com/file/d/1P_oMc6gOBlhw6dY5PxaqxV2swdHMUooK/view>

**Tabular data details:**

|  |  |
| --- | --- |
| **Total number of observations** | 3424 |
| **Total number of files** |  |
| **Total number of features** | 69 |
| **Base format of the file** | .xlsx |
| **Size of the data** | KB |

**Note: Replicate same table with file name if you have more than one file.**

**Proposed Approach:**

The focus was on understanding the business problem and preparing the dataset for further analysis. The dataset was reviewed for structure, completeness, and potential quality issues.

Key steps taken include:

* Data Loading & Exploration: Imported the dataset and reviewed the number of rows, columns, data types, and summary statistics.
* Missing Value Check: Identified columns with missing values to be addressed in later steps.
* Duplicate Check: Performed checks for duplicate patient records using Patient ID and other identifying columns to ensure data integrity.
* Initial Assumptions: Assumed the dataset is a consolidated and de-identified version of patient data suitable for model development.
* Target Variable Confirmation: Verified the target column Persistency\_Flag and its distribution to understand class balance.

This initial assessment lays the groundwork for cleaning, feature engineering, and modelling in the upcoming weeks.