# Project Report

“Fundamentals Of Programming MOD B”

**Group Members:**

ZAIGHUM JAWAD ASLAM (D03000118), Mona Nouri Mahmoudabad (D03000173)

# Introduction:

The pharmaceutical sector plays a crucial role in public health by ensuring the availability and safety of medicinal products. In Italy, the Italian Medicines Agency (AIFA) oversees the regulation and distribution of pharmaceuticals, maintaining comprehensive data on approved medicines. The aim of this project is to construct a comprehensive and structured dataset of Italian medicines by integrating regulatory information and clinical summaries from official AIFA sources. This dataset serves as a valuable resource for healthcare professionals, researchers, and policymakers by providing detailed insights into medicine classifications, active ingredients, therapeutic indications, side effects, and regulatory statuses. The project involves merging existing medicine lists from AIFA’s Class A and Class H categories, extracting clinical data from official product summaries, and assembling a unified database. Additionally, the project includes the development of interactive visualizations and a relational database to facilitate exploration and analysis of the pharmaceutical landscape in Italy. Through this effort, the project highlights the challenges of data integration from heterogeneous sources and demonstrates practical methods for managing and analyzing pharmaceutical data.

# Data Sources

The primary data sources for this project are official datasets and documents provided by the Italian Medicines Agency (AIFA). Two main lists of medicines were used: Class A and Class H medicines, which encompass reimbursed and hospital-only drugs, respectively. These lists were downloaded directly from the AIFA website and contain detailed information such as active ingredients, medicine names, authorization holders, unique AIC codes, equivalence groups, and classification codes.

In addition to these regulatory lists, clinical information was obtained from the Summary of Product Characteristics (Riassunto Caratteristiche Prodotto) documents available on the AIFA portal. These summaries provide detailed therapeutic indications, posology, contraindications, warnings, interactions, side effects, overdose information, and incompatibilities. The combination of these data sources ensures a comprehensive view of both the regulatory and clinical aspects of each medicine.

# Methodology

The project methodology consisted of three main steps. First, the Class A and Class H medicine lists were downloaded and cleaned by removing duplicates and irrelevant rows. The relevant columns were selected, including active ingredients, group descriptions, medicine denominations, marketing authorization holders, AIC codes, equivalence group codes, and classification labels. These two datasets were then merged into a single unified dataset to facilitate further processing.

Second, for each medicine identified by its unique AIC code, clinical information was manually extracted from the corresponding Summary of Product Characteristics documents on the AIFA portal. Key clinical sections (4.1 to 4.9 and 6.2) were recorded, focusing on therapeutic indications, posology, contraindications, warnings, interactions, fertility considerations, effects on driving, side effects, overdose, and incompatibilities. Missing sections were noted as "Not available." ATC codes were also included when present.

Third, the clinical and regulatory data were combined into a final structured dataset with one row per medicine, integrating all relevant clinical and regulatory columns. This dataset was cleaned and validated for consistency and completeness. Following this, interactive visualizations were created using Power BI to summarize and analyze the data, and a relational database schema was designed to support complex queries and analysis of the medicines.

## Extraction:

To manage and prepare the medicine datasets, Python scripts utilizing the Pandas library were employed for efficient data processing and cleaning. Initially, separate CSV files for Class A and Class H medicines were loaded, with appropriate encoding and delimiters to ensure accurate reading. A new column was added to each dataset to label the class type ('A' or 'H'), and both datasets were concatenated into a single unified DataFrame.

For the most part of dealing with the CSVs , a software named as OpenRefine is used because Microsoft Excel was having problems with the CSV in its limiters and delimiters of text issues and extreme loading times as well.

Subsequent data cleaning steps included the removal of rows with missing AIC codes, elimination of duplicate rows, and dropping rows that were completely empty. The cleaned and consolidated dataset was then saved as a new CSV file for further use.

Additionally, the code handled extraction of AIC codes by reading a dedicated CSV, converting the relevant column into a list for processing. Exception handling was implemented to catch and report errors during file loading and saving, ensuring robustness of the data pipeline.

Overall, this Python code automated critical data integration and cleaning steps, enabling a reliable foundation for subsequent clinical data extraction and analysis.

To efficiently extract clinical information from multiple Summary of Product Characteristics PDFs, a Python script was developed using the PyMuPDF library. The script automates the retrieval of key clinical sections for each medicine identified by its AIC code.

The process begins by scanning a designated folder containing all PDF files. Each PDF corresponds to a unique medicine identified by its AIC code, which is extracted from the filename. The script reads the full text from each PDF using PyMuPDF. A set of predefined clinical sections such as therapeutic indications (4.1), posology and method of administration (4.2), contraindications (4.3), undesirable effects (4.8), and others were specified for extraction. Using regular expressions, the script locates these section headings within the text and extracts the content for each section, capturing the text between consecutive headings. If any section is missing in a document, the script inserts a placeholder "Not available" to maintain consistent dataset structure.

After processing all PDFs, the extracted clinical data are compiled into a structured tabular format using pandas and saved as an Excel file. This automated extraction greatly reduces manual effort and ensures standardized, clean data ready for analysis.

Additionally, the script compares the set of AIC codes from the PDF files against those in the master CSV dataset to identify any missing PDF documents, providing a quality check to ensure data completeness.

# Visualization and Database Work

To facilitate the exploration and analysis of the dataset, interactive visualizations were developed using Power BI. Key visual elements included bar charts for the distribution of medicines by active ingredient, word clouds for therapeutic indications and side effects, pie charts to represent equivalence group distributions, and tree maps to provide hierarchical views of medicine classifications. Slicers and filters were incorporated to enhance interactivity and allow users to focus on specific subsets of data. Parallel to visualization, a normalized relational database was designed and implemented using a lightweight database management system. The database schema separated medicines, clinical sections, equivalence groups, and ATC codes into related tables, linked by the unique AIC identifier. This structure supported efficient querying and analysis. Several SQL queries were developed to provide analytical insights, including grouping medicines by equivalence group, identifying the most frequent contraindications, analyzing side effect prevalence, and exploring dosage patterns. These tools combined to make the dataset both accessible and analytically powerful.

## MySQL:

Relations were made in MySQL (mySQL Workbench) tool to imply queries according to the dataset columns. First proper tables are created then excel sheet has been imported to the necessary table and proceeded with the queries part.

# Conclusion

The project successfully constructed a comprehensive dataset of Italian medicines integrating regulatory and clinical data sourced from AIFA. The methodology balanced manual precision with data processing automation to ensure quality and completeness. Visualizations created with Power BI offered clear and insightful summaries of key medicine characteristics, enhancing understanding of the pharmaceutical landscape.

The relational database enabled sophisticated analysis through structured queries, demonstrating practical applications of the dataset in healthcare research and policy analysis. This project not only fulfilled its objectives but also highlighted real-world challenges in pharmaceutical data integration and the importance of thorough data management practices.