PYTHON SCRIPT EXPLANATION

01_ Data Loading and Integration.ipynb

It **connects to the hospital's SQL database**, which stores information about patient admissions, diagnostics, laboratory results, and clinical risk factors.

Once the connection is established, the notebook **inspects the database** structure to confirm the available entities and their relationships. Each table is then imported into **pandas** DataFrames.

After loading, the notebook performs a series of **merges to integrate the different sources into a single consolidated dataset**. This unified dataset combines demographic, clinical, and laboratory information, making it suitable for subsequent steps

Finally, the **integrated data is exported in a serialized .pickle format**. This allows for faster access in later notebooks, where the dataset will be cleaned, transformed, and used to develop the hospital readmission risk scorecard.

02_ Data Quality and Exploratory Data Analysis .ipynb

The second notebook focuses on **assessing the quality of the integrated dataset** and exploring its main characteristics before model development. This step ensures that the data used to predict hospital readmissions is consistent, complete, and reliable.

The notebook begins with a general inspection of the dataset, displaying **sample records and basic descriptive statistics**. It then **examines variable types, missing values, and potential logical inconsistencies**, such as discharge dates occurring before admission dates or ICU stays longer than total hospital stays. Detected issues are corrected to maintain data integrity.

After validating data quality, the notebook performs an **exploratory data analysis (EDA**) to better understand the structure and behavior of the variables. It **analyzes categorical variable** distributions, numerical statistics, and admission frequencies to detect readmissions and identify relevant patterns.

03_ Minicube and Risk Scorecard.ipynb

The third notebook builds the **hospital readmission risk scorecard**.

It starts by generating a **target variable to identify patient readmissions**, then prepares the data structure used to estimate risk profiles.

A combination of **statistical tests** (**Chi-square** for categorical variables and **T-tests** for continuous ones) identifies the **most significant predictors of readmission**.

Continuous variables are discretized into interpretable ranges, and a **minicube is created to calculate the contribution of each variable-value** pair to the overall readmission likelihood.

Finally, a **risk scoring function is designed so healthcare professionals** can assign points according to hospital-specific criteria.

The **score is normalized to a 0–10 scale**, visualized, and validated to ensure it reflects the expected relationship between scoring levels and readmission probability.

04_Production.ipynb

The final notebook focuses on **transitioning the analytical workflow into a production-ready script.**

It defines **a structured process that can be automated to periodically calculate and update patient readmission risk scores.**

The **production code connects to the original data sources**, applies all required cleaning and transformations, calculates the Risk Score for each patient, and writes the enriched results back into the hospital's readmission data system.

Unlike previous notebooks, this **script contains no exploratory steps or visual outputs**,only input/output processes and data updates.

This ensures that the model **can be executed reliably in an operational environment**, providing clinicians with continuously updated, interpretable, and actionable risk scores.