PROJECT 4: Pima Diabetes - Detecting Diabetes from Imperfect Medical Data

▼ Deliverables

- An `.Rmd` file.
- The corresponding knitted HTML file.
- The presentation slides (submitted as a PDF).

▼ Instructions

For the Rmd/HTML report, explanations must be provided for each step. The layout must be professional and "business-ready," paying close attention to Markdown formatting, the use of a CSS file for styling, and the inclusion of a Table of Contents (TOC).

Regarding the presentation, the total speaking time is 10 minutes (to be shared), followed by a 5-minute Q&A session. The presentation support must be slides in PPTX format; while there is no maximum slide count, there should be at least 6 slides at the minimum. The slides must be submitted in PDF format.

It would be appreciated if you use a Notion/Git or any project tool ; that will give you bonus points.

▼ Recommendations

Pay special attention to the polish and formatting.. The presentation should concentrate on a quick overview of the case study, its analyses (focusing on the **results**, not the technical operations), its visualizations, and its conclusions, including different avenues for planned or future improvements.

▼ Timeline

- Finalisation of the groups : Saturday 18/10 23H59
- Start of the project : Sunday 19/10 00H00
- Next session when you can ask any questions : Wednesday 22/10
- Time limit of the submission of project files (Rmd/HTML) : Monday 10/11 at 23h00
- Time limit of the submission of presentation file (pdf) : Wednesday 12/11 at 7h00
- Presentation of the project : Wednesday 12/11

For each student , you'll have to submit the files : just submit one version of it on Junia Learning ; then for each group, you'll tell me which member has the latest version of the project.

Project Goal : As data analysts in the healthcare sector, your team must build a model to predict the likelihood of diabetes in female patients of Pima Indian heritage. The primary challenge of this dataset is the presence of "impossible" data (zero values for biological measurements) that mask missing values. Your success will depend on your ability to identify and correct these anomalies before modeling.

Source : https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

Step 1: Data Cleaning and Preparation

- **Identifying problematic values and replacing them :** By calculating descriptive statistics for all variables, identify columns where certain values are biologically impossible. Replace them by NA values at first ; then look if some specific value could be chosen to replace the NA values (remember the work done in class)

- **Post-Cleaning Verification:** After imputation, double-check that there are no more missing values or anomalous zeros in the medical columns.

Step 2: Feature Transformation and Engineering

- **Standardizing Variables:** The predictive features have very different scales. Apply standardization (scaling to a mean of 0 and a standard deviation of 1) to all numerical variables.
- **Creating BMI Categories:** Create a new categorical variable `BMICategory` from `BMI` using standard WHO thresholds.
- **Creating Blood Pressure Categories :** Similarly, create a `BPCategory` variable from `BloodPressure` (e.g., 'Normal', 'Elevated').
- **Target Transformation:** Ensure the target variable `Outcome` is a factor with clear labels (e.g., "Diabetic", "Not Diabetic") for interpretability.

Step 3: Visualization and Exploration (with `ggplot2`)

You'll show the following visualisations with the problematic values encountered in the first step, and without it (after cleaning). Put them side by side when it's possible.

- For the `Glucose` variable, create density plots.
- **Comparing Distributions:** For the `BMI` and `Age` variables, show their distributions between diabetic and non-diabetic patients.
- **Correlation Between Predictors:** Calculate the correlation matrix of all numerical variables and visualize it with a heatmap to identify relationships between risk factors.
- **Analyzing Pedigree:** Create a visualisation of `DiabetesPedigreeFunction` vs. `Age`, coloring the points by `Outcome`.
- Add 2 pertinent visualisations of your liking.

Step 4: Modeling and Prediction (Data Science)

- **Logistic Regression Model:** Train a logistic regression model on the standardized data. This model is highly interpretable and will allow you to see

the effect (positive or negative) of each variable.

- **k-Nearest Neighbors (KNN) Model:** Train a KNN model. This model classifies a patient based on the majority class of their closest "neighbors" in the feature space.
- **Cross-Validation:** Use cross-validation to evaluate the robustness of your models and select the best one.
- **Analyzing the Impact of Cleaning:** Train a logistic regression model on the data *before* cleaning (with the zeros). Compare its performance to the model trained on the clean data to demonstrate the critical importance of data cleaning.