Data Processing and Cleaning

TeenTox Team

March 6, 2025

1 Introduction

This document presents a **systematic procedure** for handling missing values. The objective is to:

- Minimize *information loss*.
- Reduce biases due to missing data.
- Maintain the *statistical consistency* of variables.

Throughout the process, different imputation methods (e.g., `MICE` and `kNN`) are employed, chosen based on the percentage of missing data in each variable.

2 Initial Analysis of Missing Values

2.1 Calculating NA Percentages

The first step is to calculate the proportion of missing values (NA) for each variable, in order to:

1. Determine the severity of missing data in each column.
2. Choose the most appropriate imputation method.

An example of `R` code:

```
missing_percent <- df %>%
summarise(across(everything(),
                 ~ mean(is.na(.)) * 100)) %>%
pivot_longer(cols = everything(),
             names_to = "Variable",
             values_to = "Missing_Percentage")
```

This generates a dataframe with two columns: `Variable` and `Missing_Percentage`, identifying variables with high levels of NA (e.g., around 30%) versus others with minimal values (5%–10%).

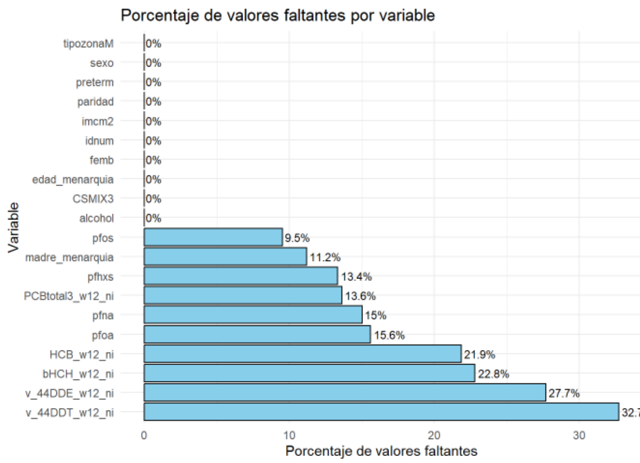

Figure 1: Table of missing values in percentage

2.2 25% Threshold

Based on the previous results, a **25% threshold** for missing values is established, dividing the variables into:

- **Less than 25% missing:** can be imputed with methods like MICE.
- **More than 25% missing:** kNN or additional techniques are used to avoid significant biases.

This division is based on the idea that, beyond a certain level of missingness, some imputation methods (e.g., MICE) may be significantly affected, increasing statistical distortion in the data.

3 Handling Variables with Less than 25% Missing Values

3.1 Multiple Imputation by Chained Equations (MICE)

For variables below the 25% threshold, *Multiple Imputation by Chained Equations* (MICE) is used. The general process:

1. Identify a model for each variable with NA.
2. Perform multiple iterations (chained equations) to *predict* and sequentially re-impute missing values.
3. Combine the results (multiple imputed datasets) into a unified dataset.

We compare various methods within MICE (`pmm`, `rf`, `cart`). The RMSE (*Root Mean Squared Error*) and MAE (*Mean Absolute Error*) were calculated for each case, determining that **`rf`** (Random Forest) generally offers the lowest error. The `R` loop for this comparison is as follows:

```
methods <- c("pmm", "rf", "cart")
results <- list()
```

```
for(m in methods){
  imp <- mice(df[vars_below_25], method = m, ...)
  # Calculate RMSE, MAE, etc.
  results[[m]] <- list(rmse=..., mae=...)
}
```

4 Handling Variables with More than 25% Missing Values

4.1 k-Nearest Neighbors (kNN)

Variables exceeding 25% NA are handled using **kNN** (k-Nearest Neighbors), utilizing the `kNN()` function from the `VIM` package. This procedure involves:

- Testing different values of k (1, 3, 5, 7, 9, ...).
- Evaluating RMSE and MAE for each k.
- Choosing the configuration that best preserves the variable's distribution shape and maintains a low error.

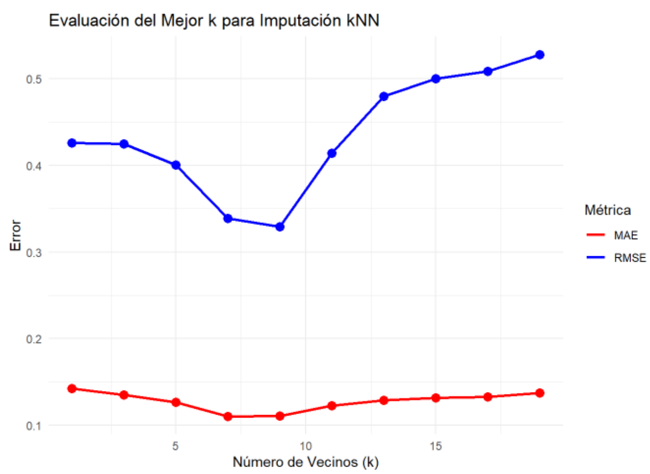

Figure 2: Evaluation graph for the best "k"

5 Validation and Final Evaluation

5.1 Error Metrics (RMSE, MAE)

For MICE and kNN, imputed values and actual observations (or reference methods) are compared using metrics such as:

- **RMSE (Root Mean Squared Error):** Square root of the mean squared errors.
- **MAE (Mean Absolute Error):** Average of absolute differences.

5.2 Distribution Verification

In addition to error metrics, *density plots*, *histograms*, and *boxplots* are examined to confirm that the overall shape of the distribution has not been significantly affected.

5.3 Final Results

Finally, it is verified that the columns no longer contain *NA*.

```
colSums(is.na(df[vars_below_25]))  # Expected to be 0
colSums(is.na(df[vars_above_25]))  # Expected to be 0
```

6 Conclusions

The proposed missing values treatment process is based on:

- An **initial analysis** to identify columns with higher or lower incidence of *NA*.
- The selection of **MICE** (with `rf` method) for variables with less than 25% missing data.
- The use of **kNN** for those with more than 25%, adjusting the value of k based on error metrics.

Note: Additional adjustments could be made depending on the nature of the data (e.g., categorical variables, outliers).