Ejercicio 3.4: Imputación con missForest

Silvia Pineda

Carga de Datos y Librerías

```
library(naniar)
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.1     v stringr   1.6.0
v ggplot2   4.0.0     v tibble    3.3.0
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.2.0
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```
library(rio)

data <- read.csv("students_FP.csv",
  na.strings = c("", "NA", "NaN", "NULL"),
  stringsAsFactors = TRUE
)
```

Imputación con missForest

```r
library(missForest)

set.seed(123)
data_missForest<-select(data,-student_id)
imp <- missForest(data_missForest)
imp$OOBerror
```

```
    NRMSE       PFC
0.1441779 0.4063217
```

```r
set.seed(123)
imp <- missForest(data_missForest,variablewise = TRUE)
imp$OOBerror
```

```
       MSE        MSE        MSE        MSE        MSE        PFC        PFC
 0.0000000 13.1874025 52.6990576  0.4157744 51.1884986  0.6928328  0.5261324
       PFC
 0.0000000
```

```r
# Calcular la Standard Deviation para normalizar solo en las cuantitativas
num_vars <- names(data_missForest)[sapply(data_missForest, is.numeric)]
id<-match(num_vars,names(data_missForest))

mse_num <- imp$OOBerror[id]
sd_num <- sapply(data_missForest[id], sd, na.rm = TRUE)

# Calcular el NMRSE
NRMSE <- sqrt(mse_num) / sd_num
names(NRMSE)<-colnames(data_missForest[id])
NRMSE
```

```
 hours_work_week hours_study_week   attendance_pct              gpa
       0.0000000        0.7111160        0.9260997        0.4147127
      exam_score
       0.3855255
```

De forma global los dos errores tanto para las cuantitativas como para las cualitativas, los errores son pequeños:

NRMSE = 0.14

PFC = 0.41

Si lo sacamos de forma individual por variable, para las cualitativas tenemos:

program (PFC = 0.69)

study_mode (PFC = 0.53)

hours_study_week (NRMSE = 0.71)

attendance_pct (NRMSE = 0.93)

gpa (NRMSE = 0.41)

exam_score (NRMSE = 0.38)

Ninguno parece demasiado alto, pero es curioso que gpa sea uno de los más bajos a diferencia del resto de imputaciones.

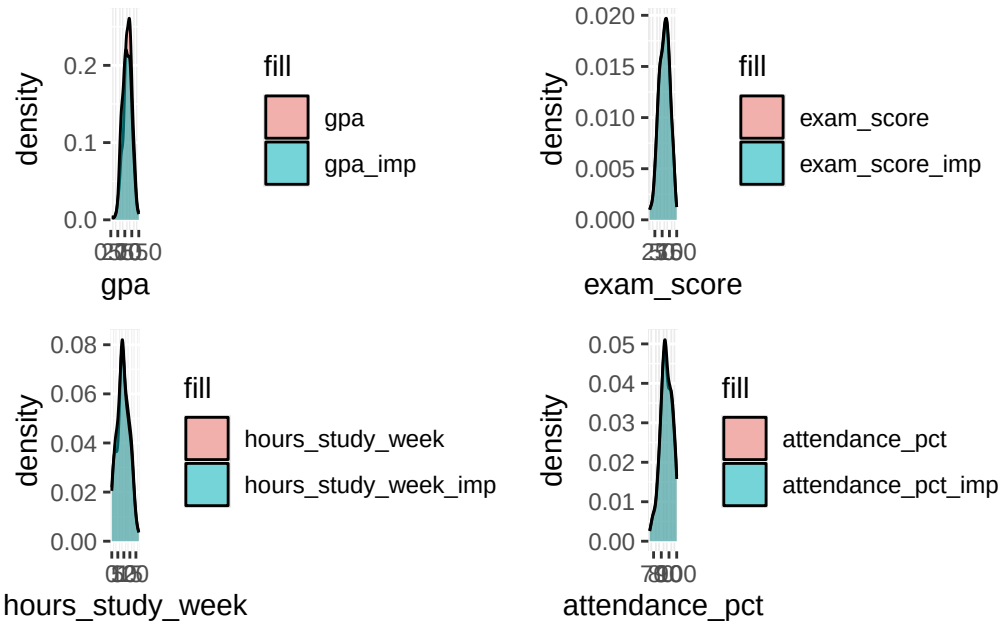
```
g1<-ggplot(data, aes(x =gpa, fill = "gpa")) +
  geom_density(alpha = 0.5, na.rm = TRUE) +
  geom_density(aes(x = imp$ximp$gpa,
                   fill = "gpa_imp"), alpha = 0.5)

g2<-ggplot(data, aes(x = exam_score, fill = "exam_score")) +
  geom_density(alpha = 0.5, na.rm = TRUE) +
  geom_density(aes(x = imp$ximp$exam_score,
                   fill = "exam_score_imp"),alpha = 0.5)

g3<-ggplot(data, aes(x = hours_study_week,
                     fill = "hours_study_week")) +
  geom_density(alpha = 0.5, na.rm = TRUE) +
  geom_density(aes(x = imp$ximp$hours_study_week,
                   fill = "hours_study_week_imp"), alpha = 0.5)

g4<-ggplot(data, aes(x = attendance_pct,
                     fill = "attendance_pct")) +
  geom_density(alpha = 0.5, na.rm = TRUE) +
  geom_density(aes(x = imp$ximp$attendance_pct,
                   fill = "attendance_pct_imp"), alpha = 0.5)

library(patchwork)
g1+g2+g3+g4
```



El gráfico de gpa es el que peor se ajusta a su distribución, pero mejora mucho a lo que veíamos con el resto de imputaciones.