CU18_MODEL_DEVELOPMENT_01_CLUSTER

June 12, 2023

#

CU18_Comportamienta Infra. Eventos extremos

1 IV. Model development

En este anexo se incluye el código utilizado durante el desarrollo de los modelos incluidos en el caso de uso.

1.1 Modelo CLUSTER

1.1.1 Paquetes

[6]:
```
## Paquetes
library(readr)
library(dplyr)
library(tidyr)
library(cluster)
library(recipes)
library(janitor)
library(purrr)
library(FactoMineR)
library(mclust)
```

1.1.2 Datos

1.1.3 Clustering

A dos niveles: distrito y diario

1.1.4 Nivel diario

[7]:
```
NIVEL <- "Diario"


  ## DISTRITOS ----

  df <- read_csv("CU_18_05_16_distritos_variables.csv")

  ## Valores perdidos que tendría que haber solucionado en sus
```

```r
## notebooks pero lo hago aquí

ids <- df |> drop_na(cmun) |> select(1:2)
df |> filter(is.na(cmun))
df <- df |>
  drop_na(cmun) |>
  select(-c("cmun", "cdis", "X", "Y"))

## ¿hay perdidos?

df |>
  map_dbl(~sum(is.na(.x))) |>
  sum()

## ¿hay varianza cero?
df |>
  map_dbl(~var(.x, na.rm = TRUE) == 0) |>
  sum()



# df_complete <- df |>
#   drop_na()

## Imputo por KNN

rec <- recipe(
  ~ .,
  data = df
)

impute_recipe <- rec |>
  step_impute_knn(all_predictors(), neighbors = 3)
impute_recipe2 <- prep(impute_recipe, training = df)
df_imputed <- bake(impute_recipe2, df)

df_imputed |>
  map_dbl(~sum(is.na(.x))) |>
  sum()


## Cluster ----

# https://bradleyboehmke.github.io/HOML/model-clustering.html
```

```r
# dfz <- scale(df_imputed)

df_mc <- Mclust(df_imputed)

summary(df_mc)

# plot(df_mc, what = "density", dimens = 1:4)
# plot(df_mc, what = "BIC")



## PCA para visuals

dfpca <- PCA(df_imputed, graph = FALSE)
# dfpca$ind$coord
# summary(dfpca)

dfout <- ids |>
  bind_cols(dfpca$ind$coord,
            cluster = factor(df_mc$classification),
            df_imputed)


write_rds(dfout, "datos_cluster_distritos.rds")
write_rds(df_mc, "modelo_cluster_distritos.rds")
```

```
Rows: 247 Columns: 143
-- Column specification --------------------------------------------------------
Delimiter: ","
chr   (2): cmun, cdis
dbl (141): consultorios_de_salud, helisuperficies,
centros_de_atencion_a_dro…

i Use `spec()` to retrieve the full column specification for this
data.
i Specify the column types or set `show_col_types = FALSE` to quiet
this message.
```

A spec_tbl_df: 1 x 143

cmun	cdis	consultorios_de_salud	helisuperficies	centros_de_atencion_a_dro
<chr>	<chr>	<dbl>	<dbl>	<dbl>
NA	NA	0	0	0

1543

0

0

```
------------------------------------------------------
```

```
Gaussian finite mixture model fitted by EM algorithm
----------------------------------------------------

Mclust VEI (diagonal, equal shape) model with 3 components:

 log-likelihood   n  df       BIC       ICL
       -43151.3 246 560 -89385.58 -89385.58

Clustering table:
  1   2   3
 88 106  52
```

1.1.5 Nivel distrito

[8]:
```
NIVEL <- "Diario"

  ## DIARIO ----

  df <- read_csv("CU_18_05_20_diario_infra.csv")

  ## Valores perdidos que tendría que haber solucionado en sus
  ## notebooks pero lo hago aquí


  ## ¿hay perdidos?

  df |>
    map_dbl(~sum(is.na(.x))) |>
    sum()

  ## ¿hay varianza cero?
  df |>
    map_dbl(~var(.x, na.rm = TRUE) == 0) |>
    sum()

  ## Quitamos perdidos ya que la imputación no termina

  df_complete <- df |>
    drop_na()

  ids <- df_complete |> select(1:2)
  df_complete <- df_complete |>
    select(-c(1:2))
```

```r
## Cluster ----

# https://bradleyboehmke.github.io/HOML/model-clustering.html



# dfz <- scale(df_imputed)

df_mc <- Mclust(df_complete)

summary(df_mc)

# plot(df_mc, what = "density", dimens = 1:4)
# plot(df_mc, what = "BIC")



## PCA para visuals

dfpca <- PCA(df_complete, graph = FALSE)
# dfpca$ind$coord
# summary(dfpca)

dfout <- ids |>
  bind_cols(dfpca$ind$coord,
            cluster = factor(df_mc$classification),
            df_complete)


write_rds(dfout, "datos_cluster_diario.rds")
write_rds(df_mc, "modelo_cluster_diario.rds")
```

```
Rows: 415370 Columns: 10
-- Column specification --------------------------------------------------------
Delimiter: ","
dbl  (9): id_inf, capacidad, demanda, evento_infra, evento_zona, tmed,
prec,…
date (1): fecha

i Use `spec()` to retrieve the full column specification for this
data.
i Specify the column types or set `show_col_types = FALSE` to quiet
this message.
```

37554

0

```
----------------------------------------------------
Gaussian finite mixture model fitted by EM algorithm
----------------------------------------------------

Mclust VEV (ellipsoidal, equal shape) model with 4 components:

 log-likelihood      n  df       BIC       ICL
       -7449405 377816 158 -14900840 -14903176

Clustering table:
     1      2      3      4
 53579 116459  80263 127515
```