# COMPETICIÓN DE KAGGLE

## Feature Imputation with a Heat Flux Dataset

Trabajo realizado por
ALBA CRUZ

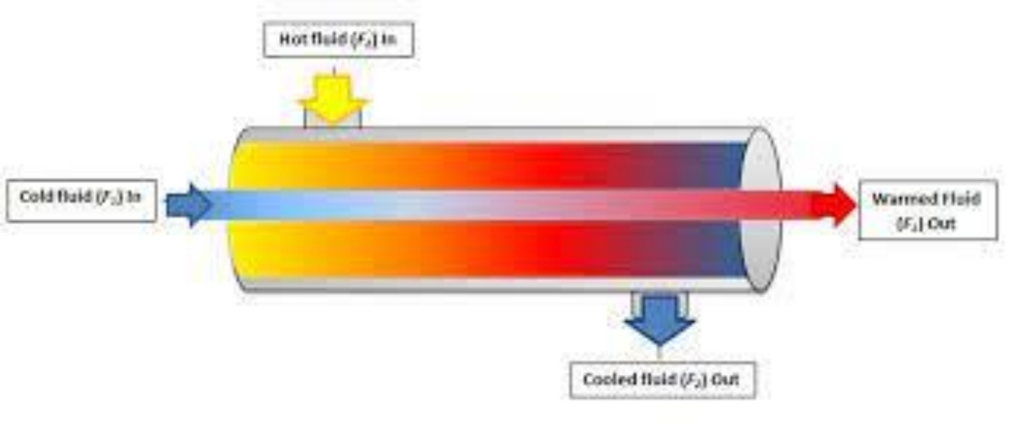
# DESCRIPCIÓN DEL PROBLEMA

# DESCRIPCIÓN DEL PROBLEMA

- El flujo de calor crítico (CHF) se refiere a la cantidad máxima de calor que puedes aplicar al agua antes de que ocurran problemas. (Predicción en problema original).

- Los datos se recopilaron en diferentes configuraciones experimentales, utilizando diferentes tipos de calentadores.

- El objetivo es utilizar estos datos para diseñar sistemas de calentamiento más seguros y eficientes en diferentes industrias.



# EDA

# DATASET GENERADO Y DATASET ORIGINAL

**La incorporación del dataset original en el entrenamiento del dataset generado mejora el modelo**

## DATASET ORIGINAL

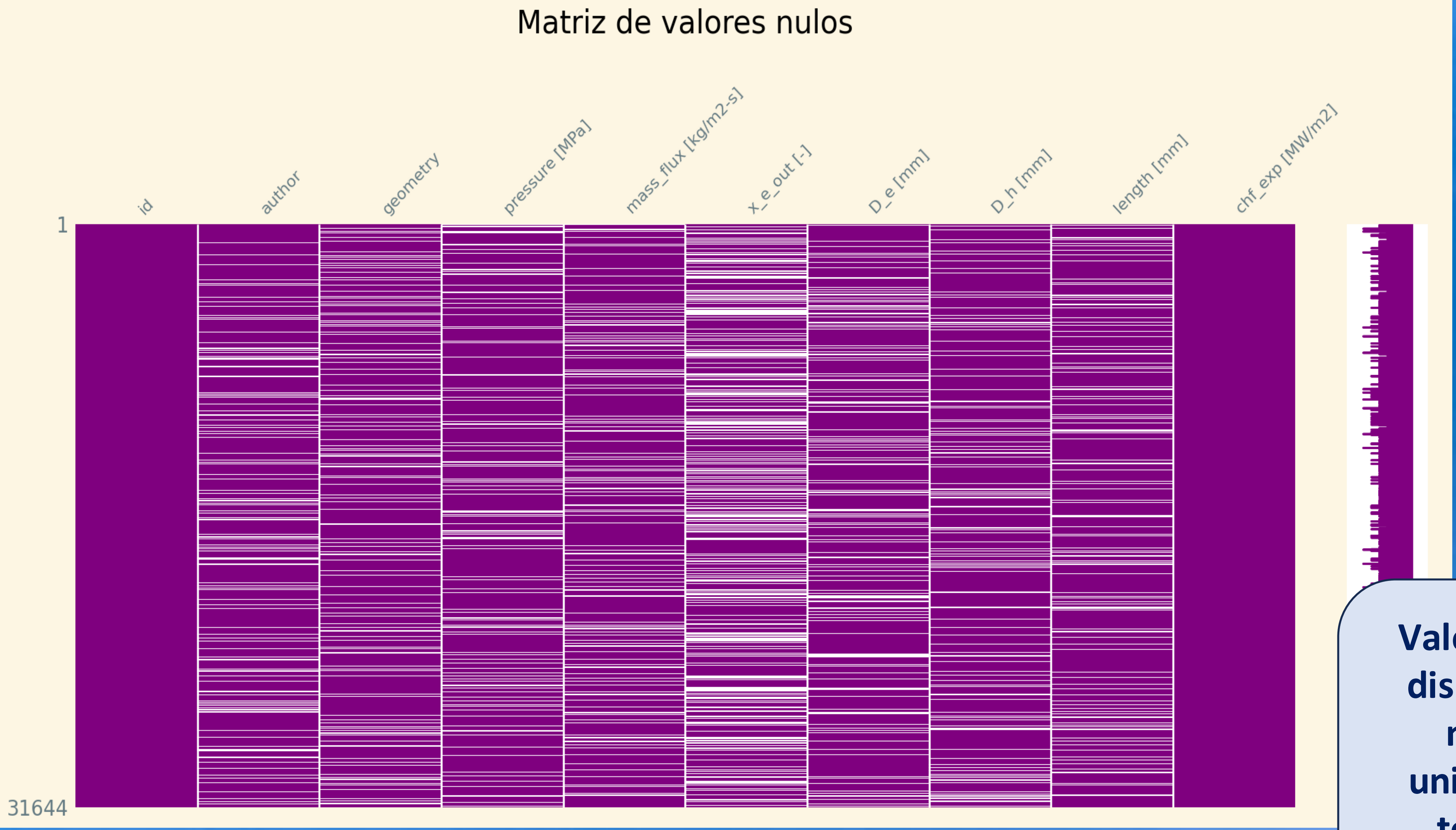
data shape: (1865, 10)

| | %missing | mean | std | median |
|---|---|---|---|---|
| author | 0.0 | NaN | NaN | NaN |
| geometry | 0.0 | NaN | NaN | NaN |
| pressure [MPa] | 0.0 | 10.010949 | 4.282715 | 10.34 |
| mass_flux [kg/m2-s] | 0.0 | 2862.647721 | 1656.412247 | 2590.0 |
| x_e_out [-] | 0.0 | 0.016179 | 0.117575 | 0.0244 |
| D_e [mm] | 0.0 | 9.417212 | 6.333807 | 8.5 |
| D_h [mm] | 0.0 | 16.167721 | 21.18287 | 10.3 |
| length [mm] | 0.0 | 911.340483 | 726.718974 | 625.0 |
| chf_exp [MW/m2] | 0.0 | 3.854638 | 1.985535 | 3.5 |

## DATASET GENERADO

data shape: (31644, 10)

| | %missing | mean | std | median |
|---|---|---|---|---|
| author | 15.876627 | NaN | NaN | NaN |
| geometry | 17.380862 | NaN | NaN | NaN |
| pressure [MPa] | 14.069018 | 10.640747 | 4.333683 | 11.07 |
| mass_flux [kg/m2-s] | 15.140311 | 3068.011023 | 1777.03208 | 2731.0 |
| x_e_out [-] | 32.913032 | -0.000453 | 0.100911 | 0.0038 |
| D_e [mm] | 17.342940 | 8.629255 | 5.185692 | 7.8 |
| D_h [mm] | 14.501959 | 14.17433 | 19.838489 | 10.0 |
| length [mm] | 15.039186 | 832.987391 | 672.299239 | 610.0 |
| chf_exp [MW/m2] | 0.000000 | 3.796985 | 1.983991 | 3.4 |




**Valores nulos dispersos de manera uniforme en todas las columnas**

# MATRICES DE CORRELACIÓN

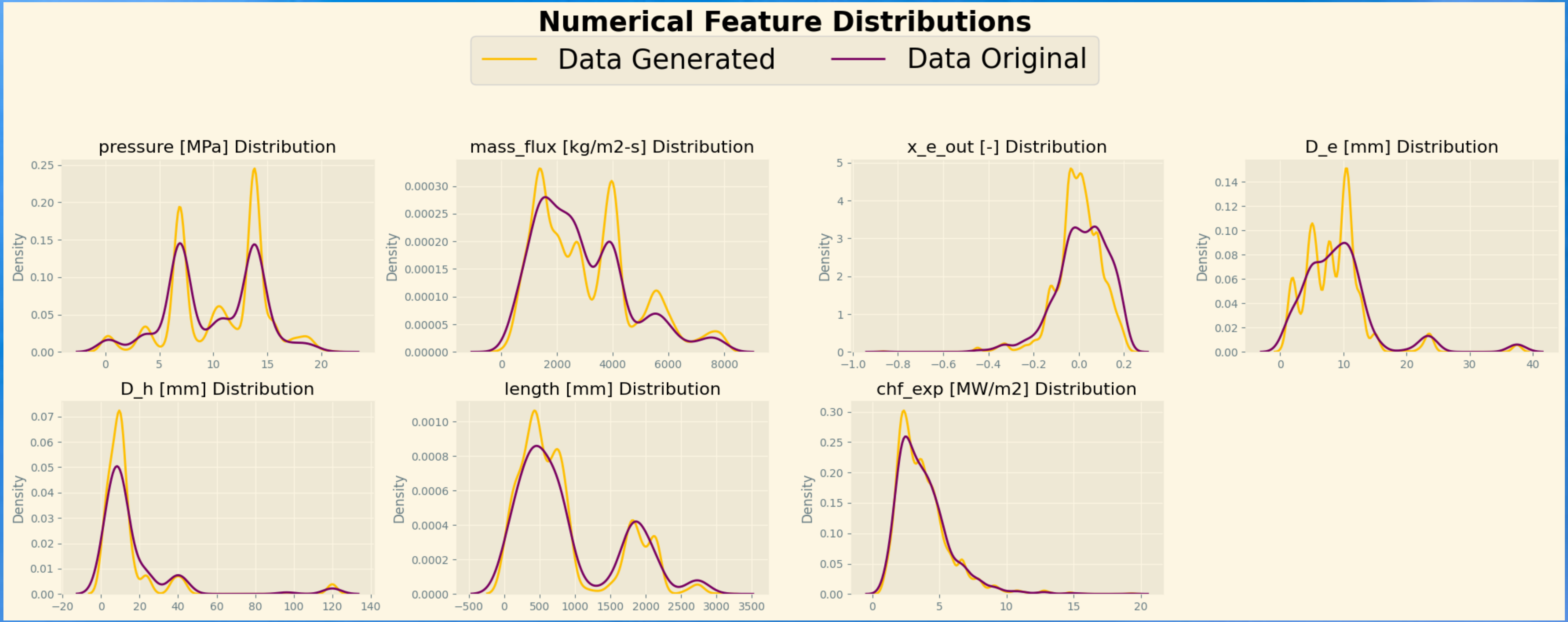
## DATASET ORIGINAL

| | pressure [MPa] | mass_flux [kg/m2-s] | x_e_out [-] | D_e [mm] | D_h [mm] | length [mm] | chf_exp [MW/m2] |
|---|---|---|---|---|---|---|---|
| pressure [MPa] | 1.000000 | -0.165660 | -0.296783 | -0.400600 | -0.514806 | -0.190572 | -0.356977 |
| mass_flux [kg/m2-s] | -0.165660 | 1.000000 | -0.223631 | -0.046866 | -0.242915 | -0.062630 | 0.453562 |
| x_e_out [-] | -0.296783 | -0.223631 | 1.000000 | 0.110438 | 0.080584 | 0.378102 | -0.513687 |
| D_e [mm] | -0.400600 | -0.046866 | 0.110438 | 1.000000 | 0.493515 | 0.373820 | -0.082771 |
| D_h [mm] | -0.514806 | -0.242915 | 0.080584 | 0.493515 | 1.000000 | 0.186977 | 0.099406 |
| length [mm] | -0.190572 | -0.062630 | 0.378102 | 0.373820 | 0.186977 | 1.000000 | -0.423167 |
| chf_exp [MW/m2] | -0.356977 | 0.453562 | -0.513687 | -0.082771 | 0.099406 | -0.423167 | 1.000000 |

## DATASET GENERADO

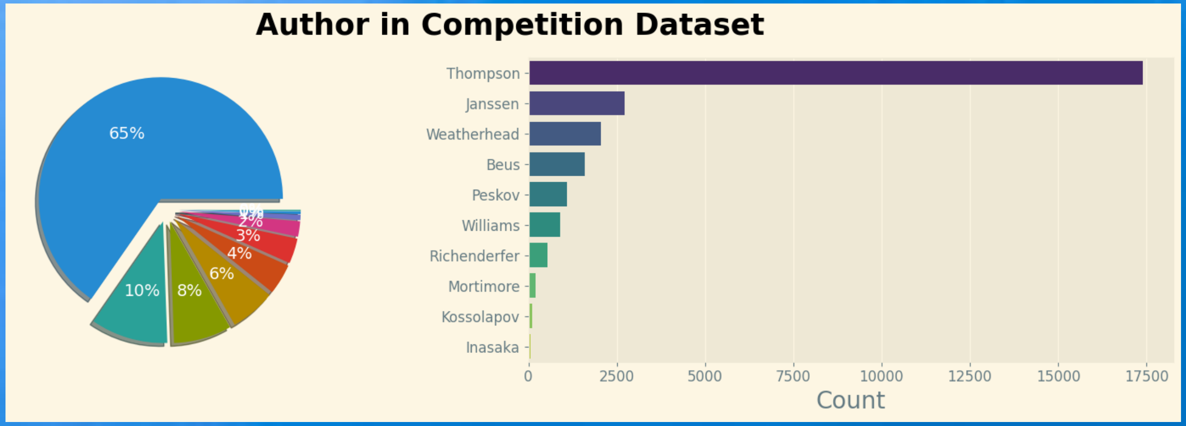
| | pressure [MPa] | mass_flux [kg/m2-s] | x_e_out [-] | D_e [mm] | D_h [mm] | length [mm] | chf_exp [MW/m2] |
|---|---|---|---|---|---|---|---|
| pressure [MPa] | 1.000000 | -0.195332 | -0.193125 | -0.468037 | -0.498645 | -0.090388 | -0.259936 |
| mass_flux [kg/m2-s] | -0.195332 | 1.000000 | -0.168136 | 0.004676 | -0.180331 | -0.055095 | 0.308971 |
| x_e_out [-] | -0.193125 | -0.168136 | 1.000000 | 0.124835 | 0.063367 | 0.336840 | -0.370580 |
| D_e [mm] | -0.468037 | 0.004676 | 0.124835 | 1.000000 | 0.494538 | 0.314969 | 0.019495 |
| D_h [mm] | -0.498645 | -0.180331 | 0.063367 | 0.494538 | 1.000000 | 0.113241 | 0.055734 |
| length [mm] | -0.090388 | -0.055095 | 0.336840 | 0.314969 | 0.113241 | 1.000000 | -0.276146 |
| chf_exp [MW/m2] | -0.259936 | 0.308971 | -0.370580 | 0.019495 | 0.055734 | -0.276146 | 1.000000 |

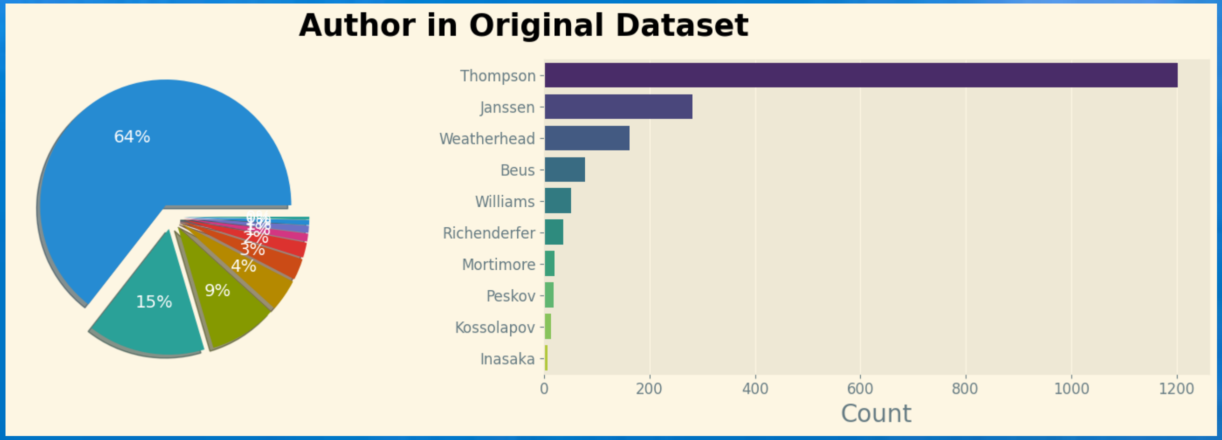
# Distribución de las variables numéricas

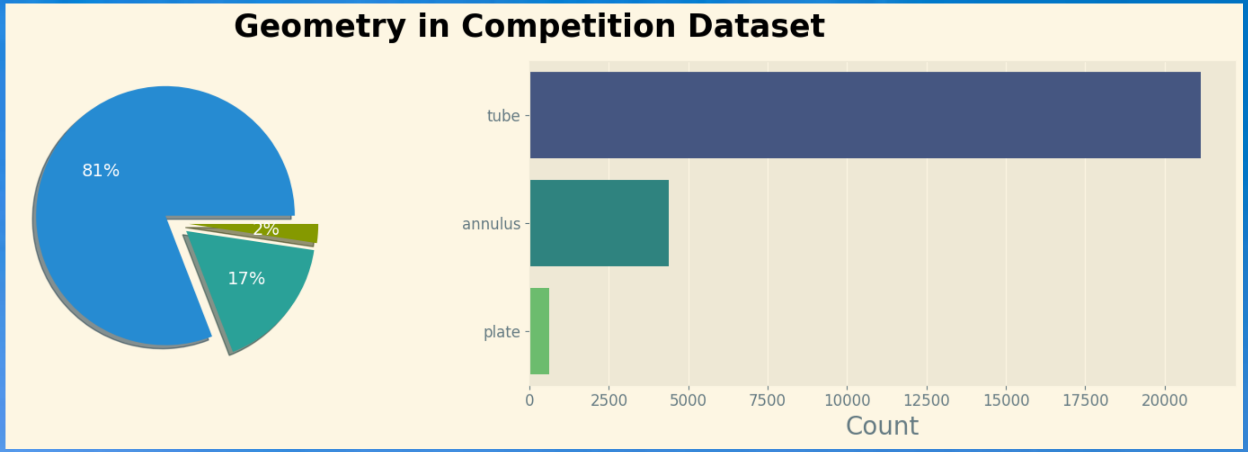


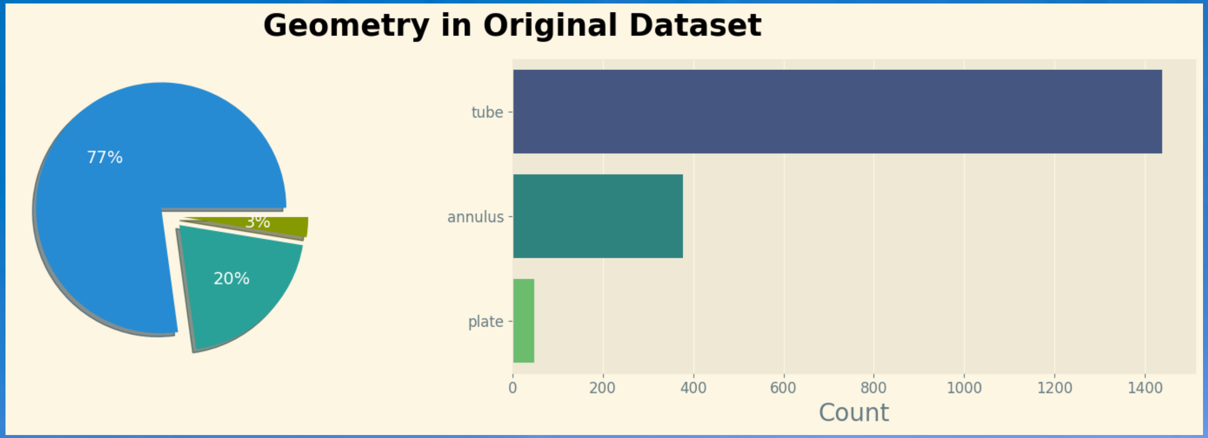
## Distribución de las Variables categóricas

**Tienen una categoría dominante. Sería buena idea imputar con la moda.**

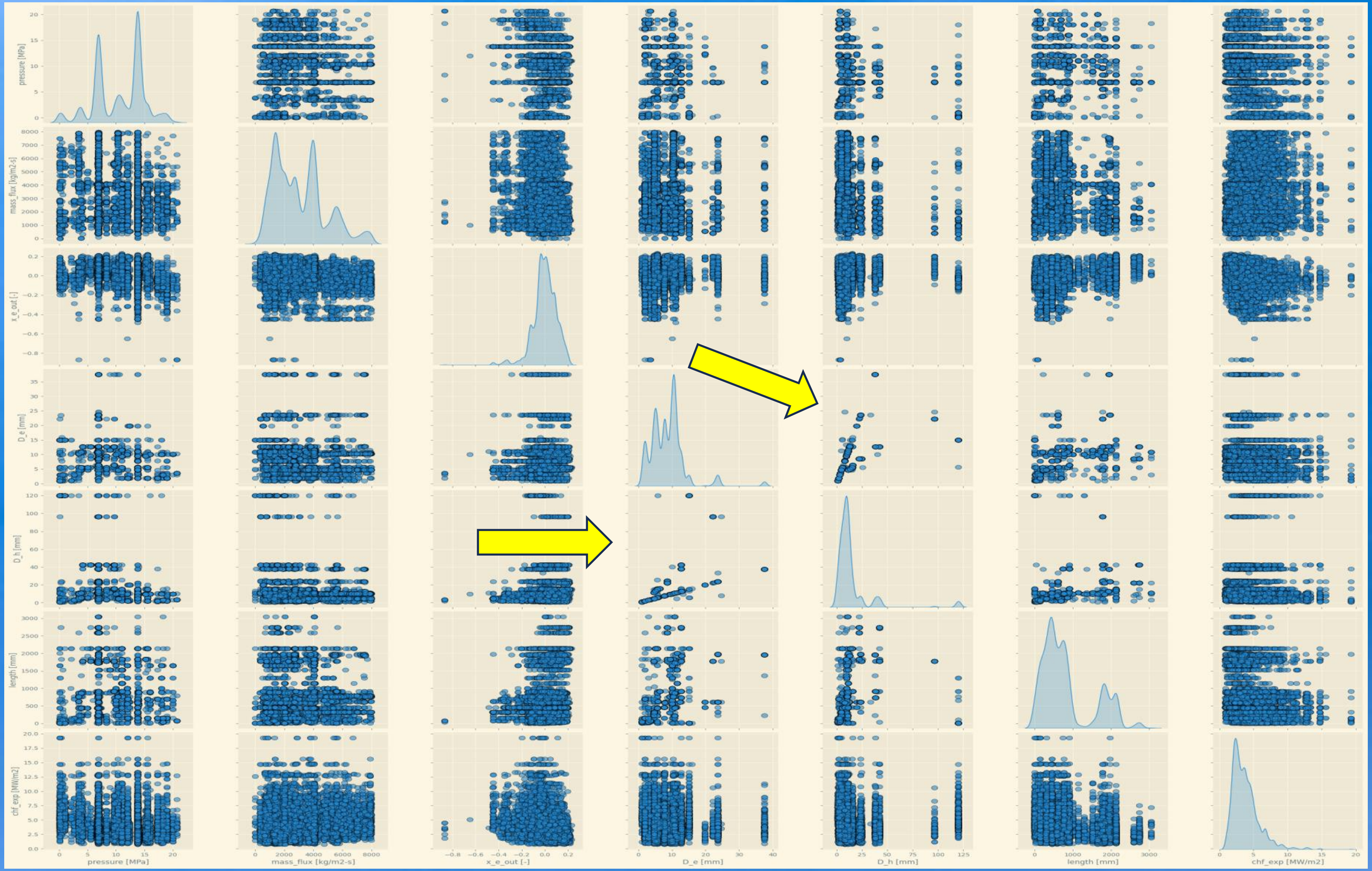


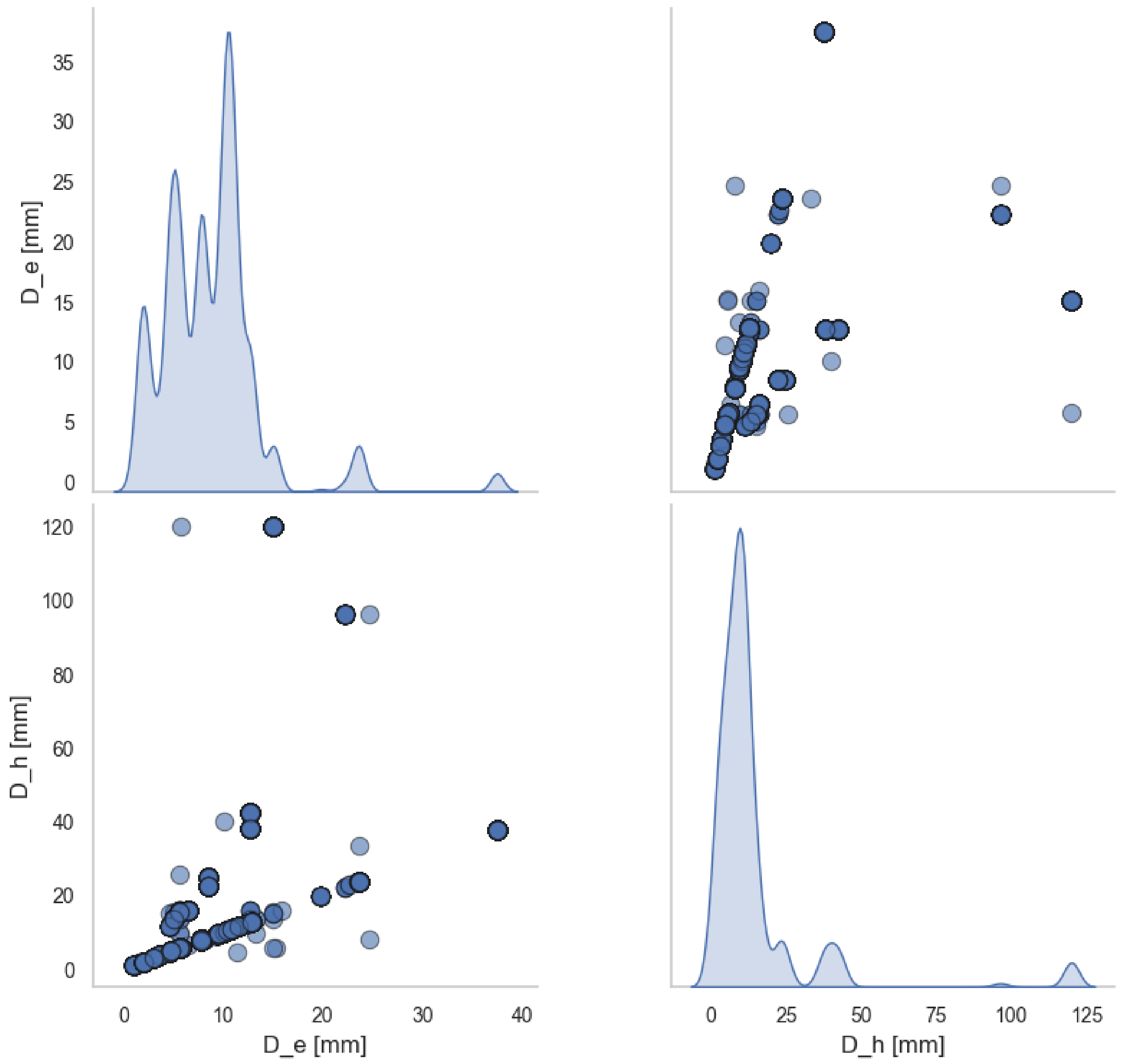



# PATRONES Y CORRELACIONES EN LAS VARIABLES NUMÉRICAS




Cuando las variables no siguen una correlación lineal clara, imputar los valores nulos con medidas de estadística descriptiva puede introducir sesgos en los datos imputados.

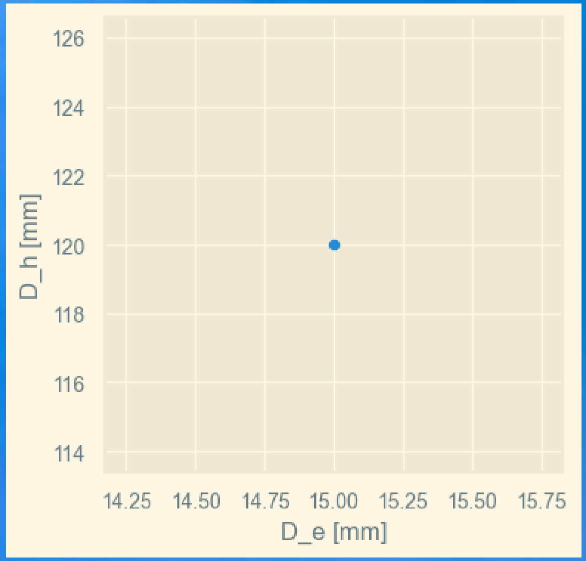


D_e [mm]
D_h [mm]
0
5
10
15
20
25
30
35
0
20
40
60
80
100
120
0
10
20
30
40
0
25
50
75
100
125

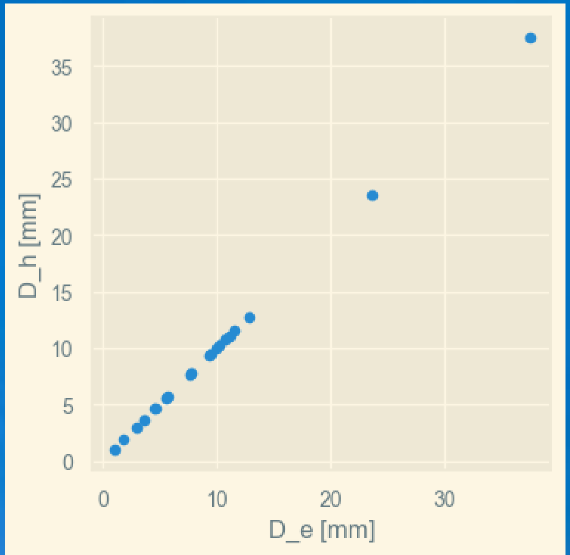
# IMPUTACIÓN DE MISSING VALUES

# D_e [mm] y D_h [mm] en función de 'geometry'
## Dataset ORIGINAL

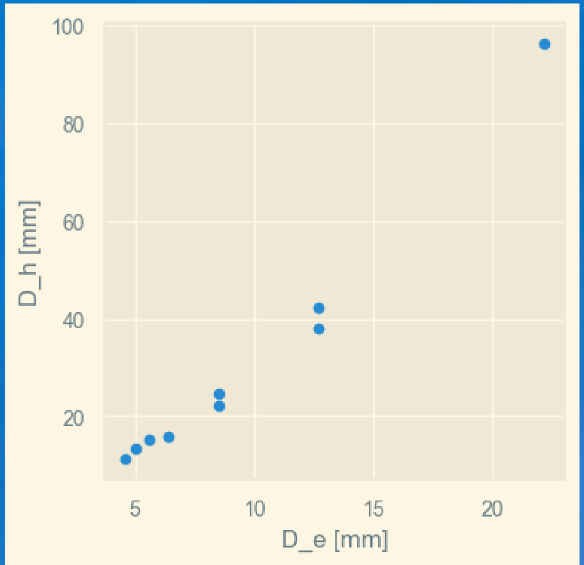
**Geometry = plate**
**D_h [mm] = 120**
**D_e [mm] = 15**




**Geometry = tube**
**D_h [mm] = D_e [mm]**

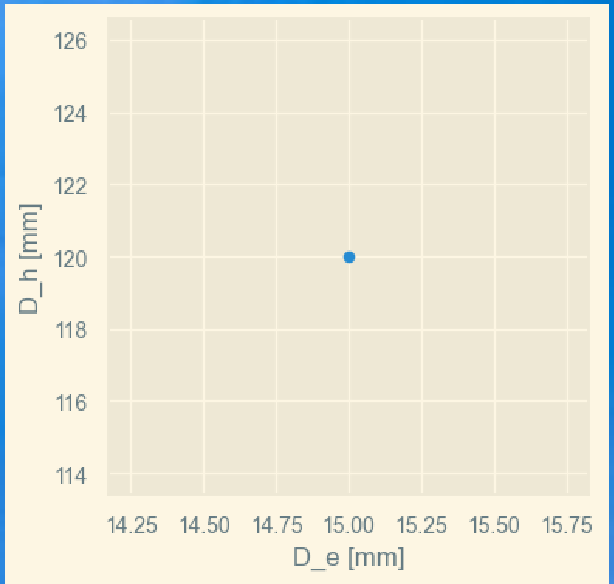



**Geometry = annulus**
**D_e [mm] < D_h [mm]**

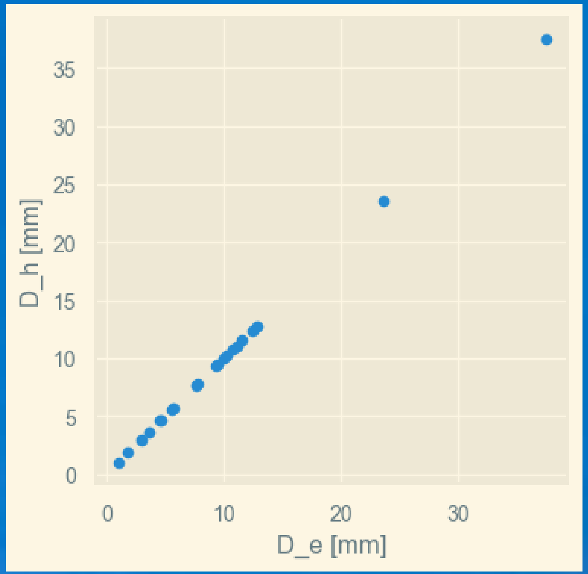


# D_e [mm] y D_h [mm] en función de 'geometry'
# Dataset GENERADO

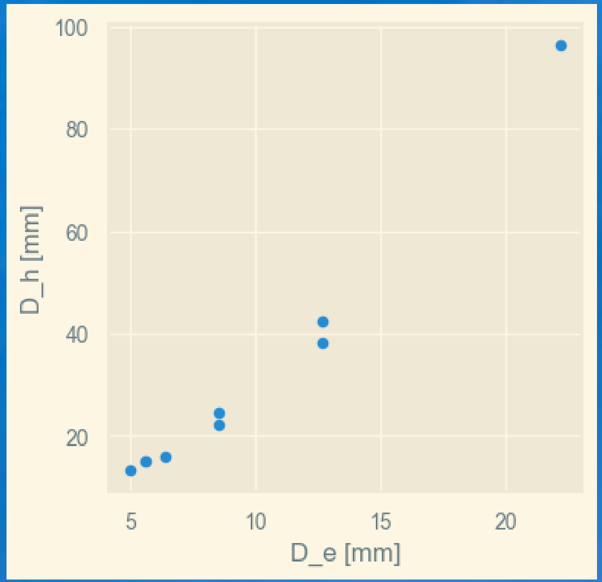
**Geometry = plate**
**D_h [mm] = 120**
**D_e [mm] = 15**



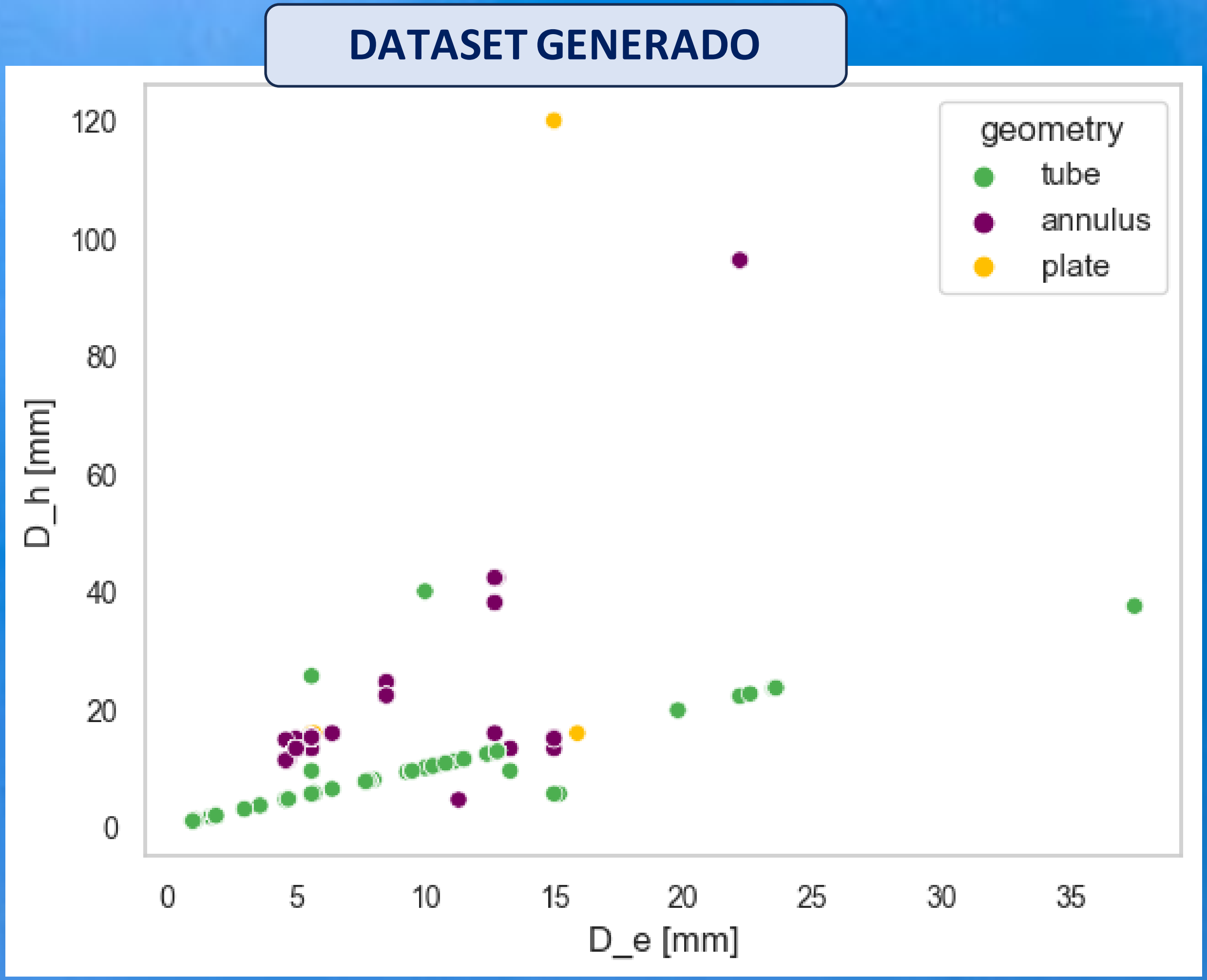

**Geometry = tube**
**D_h [mm] = D_e [mm]**

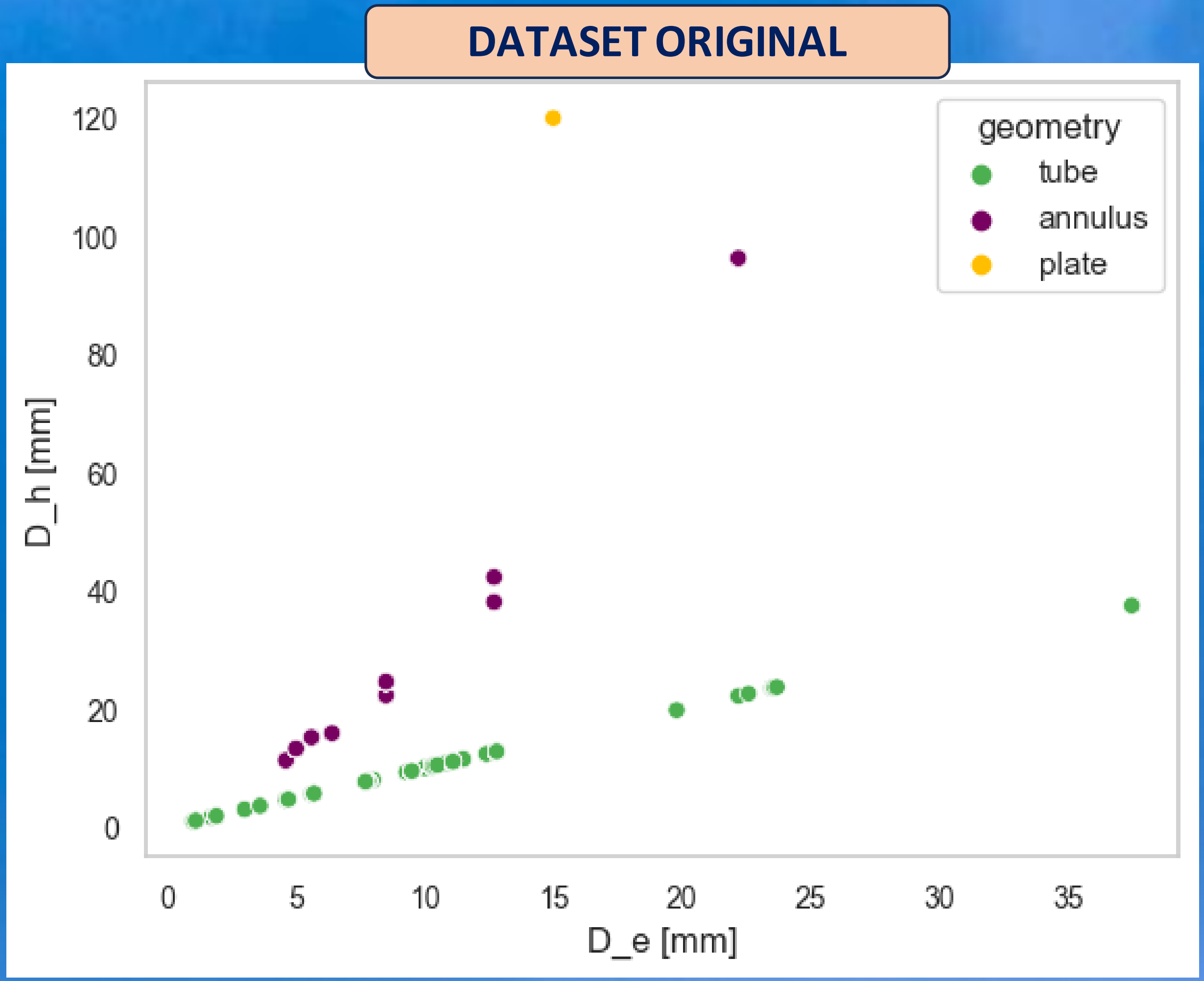



**Geometry = annulus**
**D_e [mm] < D_h [mm]**



# D_e [mm]  y  D_h [mm] en función de 'geometry'



Variables categóricas

'author' → MODA

'geometry'

get dummies


Imputamos nulos de geometry con valores informados de D_e [mm] y D_h [mm]

MODA

Variables Numéricas no relacionadas linealmente

pressure [MPa]

mass_flux [kg/m2-s]

length [mm]

KNN

## Variables Numéricas relacionadas linealmente

D_h [mm]

D_e [mm]

Geometry = plate
D_h [mm] = 120
D_e [mm] = 15

Geometry = tube
D_h [mm] = D_e [mm]

Geometry = annulus
D_e [mm] < D_h [mm]

**1. REGRESIÓN LINEAL**

D_h [mm] = k * D_e [mm]

Entrenamos en dataset orig.

k = 4.307381844408128

**2. REGRESIÓN LINEAL INVERSA**

D_e [mm] = a + b * D_h [mm]

Entrenamos en dataset orig.

a = 2.9750
b= 0.2226

**3. KNN**

D_e [mm] y D_h [mm] = null

# MODELOS

Grid searh →

Best Params →

Entreno por separado

XGBoostReg

LightGBoostRegReg

CatBoostReg

GradientBoostReg

RandomForestReg

**ENSEMBLE**
`VotingRegressor`

RMSE:
0.0746145

SCORE KAGGLE:
0.075681

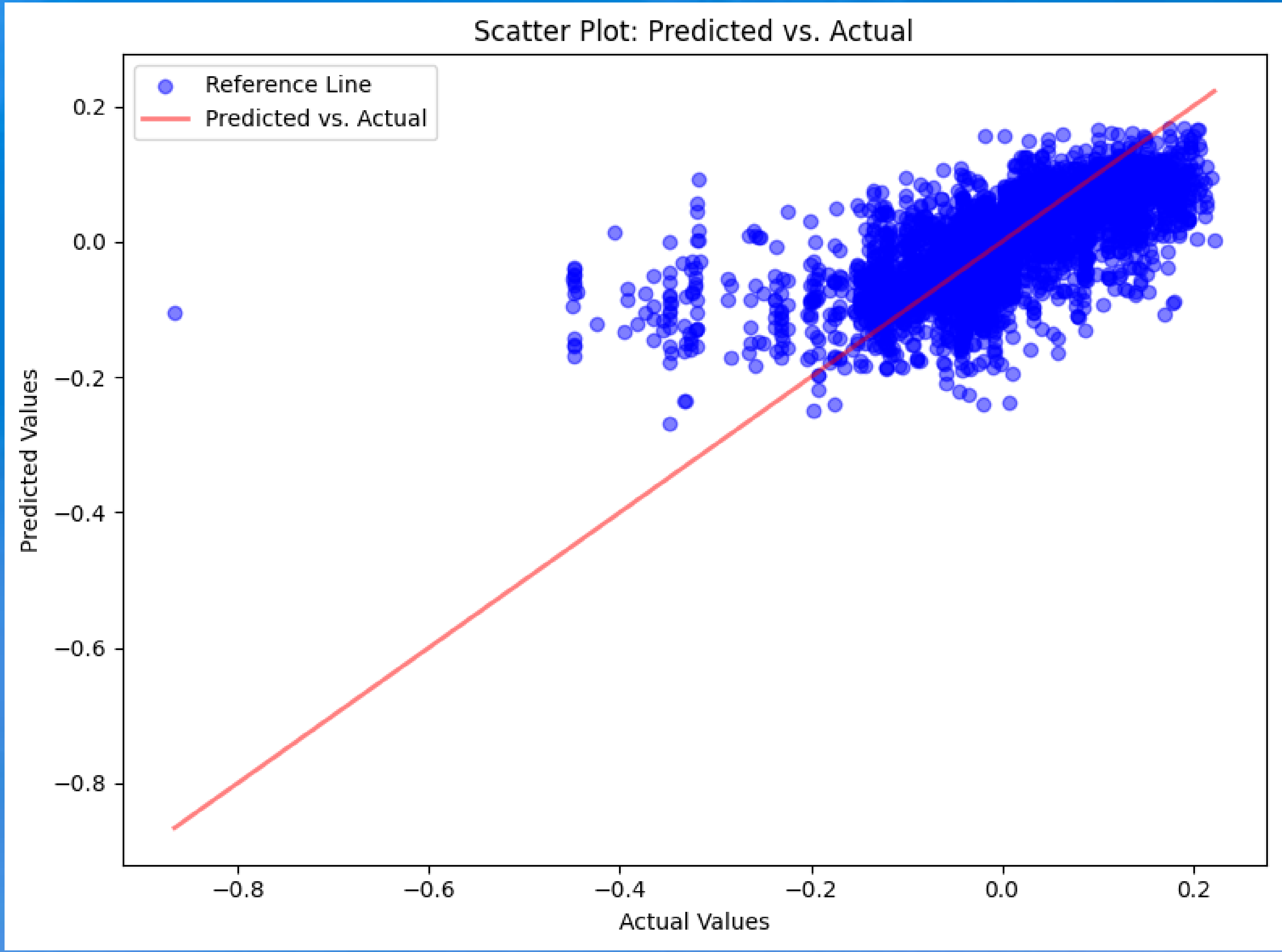
# MEJOR MODELO

XGBoostReg

RMSE:
0.0746145

SCORE KAGGLE:
0.075426

# Mejor modelo : XGBoostReg



¡MUCHAS GRACIAS!