Kaggle: Heat Flux Imputation

Índice

1. Dataframe Inicial
2. EDA
3. Feature Engineering
4. Pipeline & Grid
5. Análisis Modelos
6. Conclusiones

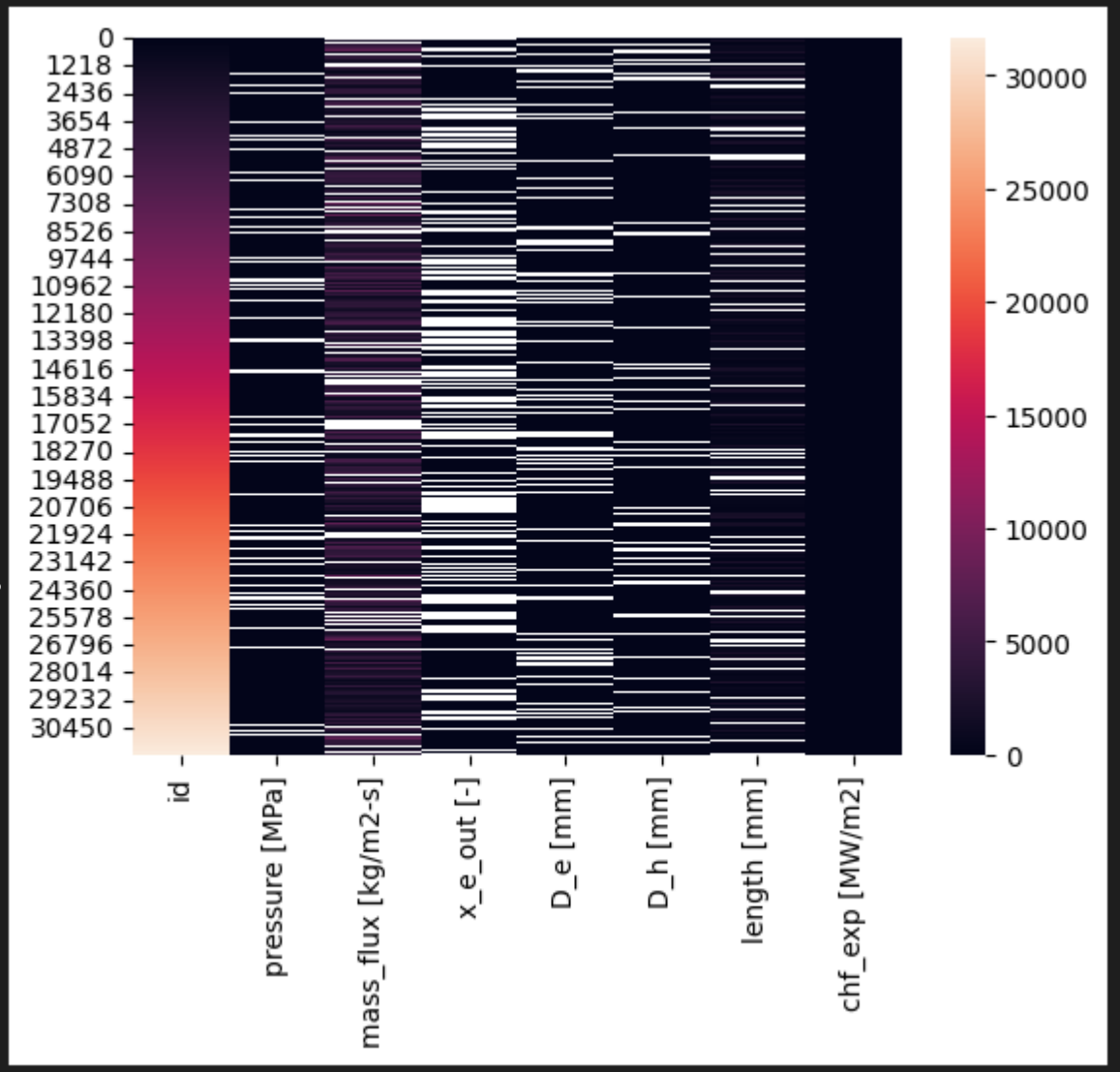
1. DATAFRAME INICIAL

	id	author	geometry	pressure [MPa]	mass_flux [kg/m2-s]	x_e_out [-]	D_e [mm]	D_h [mm]	length [mm]	chf_exp [MW/m2]
0	0	Thompson	tube	7.00	3770.0	0.1754	NaN	10.8	432.0	3.6
1	1	Thompson	tube	NaN	6049.0	-0.0416	10.3	10.3	762.0	6.2
2	2	Thompson	NaN	13.79	2034.0	0.0335	7.7	7.7	457.0	2.5
3	3	Beus	annulus	13.79	3679.0	-0.0279	5.6	15.2	2134.0	3.0
4	4	NaN	tube	13.79	686.0	NaN	11.1	11.1	457.0	2.8

2. EDA

```
Data columns (total 11 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   id                   31644 non-null  int64
 1   author               26620 non-null  object
 2   geometry             26144 non-null  object
 3   pressure [MPa]       27192 non-null  float64
 4   mass_flux [kg/m2-s]  26853 non-null  float64
 5   x_e_out [-]          21229 non-null  float64
 6   D_e [mm]             26156 non-null  float64
 7   D_h [mm]             27055 non-null  float64
 8   length [mm]          26885 non-null  float64
 9   chf_exp [MW/m2]      31644 non-null  float64
 10  geometry_corrected   22382 non-null  object
dtypes: float64(7), int64(1), object(3)
memory usage: 2.7+ MB
```

2. EDA

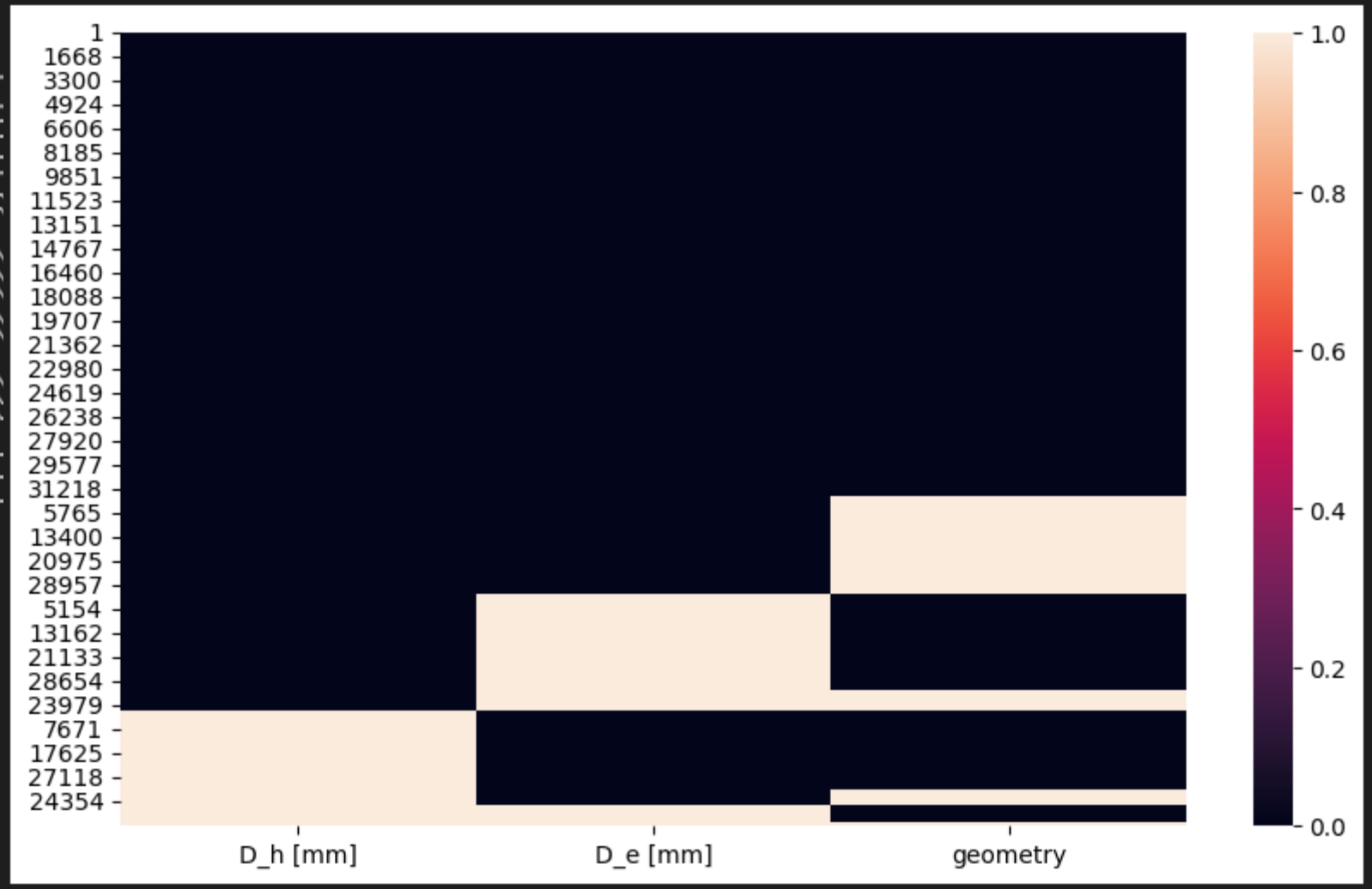

0
1218
2436
3654
4872
6090
7308
8526
9744
10962
12180
13398
14616
15834
17052
18270
19488
20706
21924
23142
24360
25578
26796
28014
29232
30450
id
pressure [MPa]
mass_flux [kg/m2-s]
x_e_out [-]
D_e [mm]
D_h [mm]
length [mm]
chf_exp [MW/m2]
30000
25000
20000
15000
10000
5000
0

2. EDA

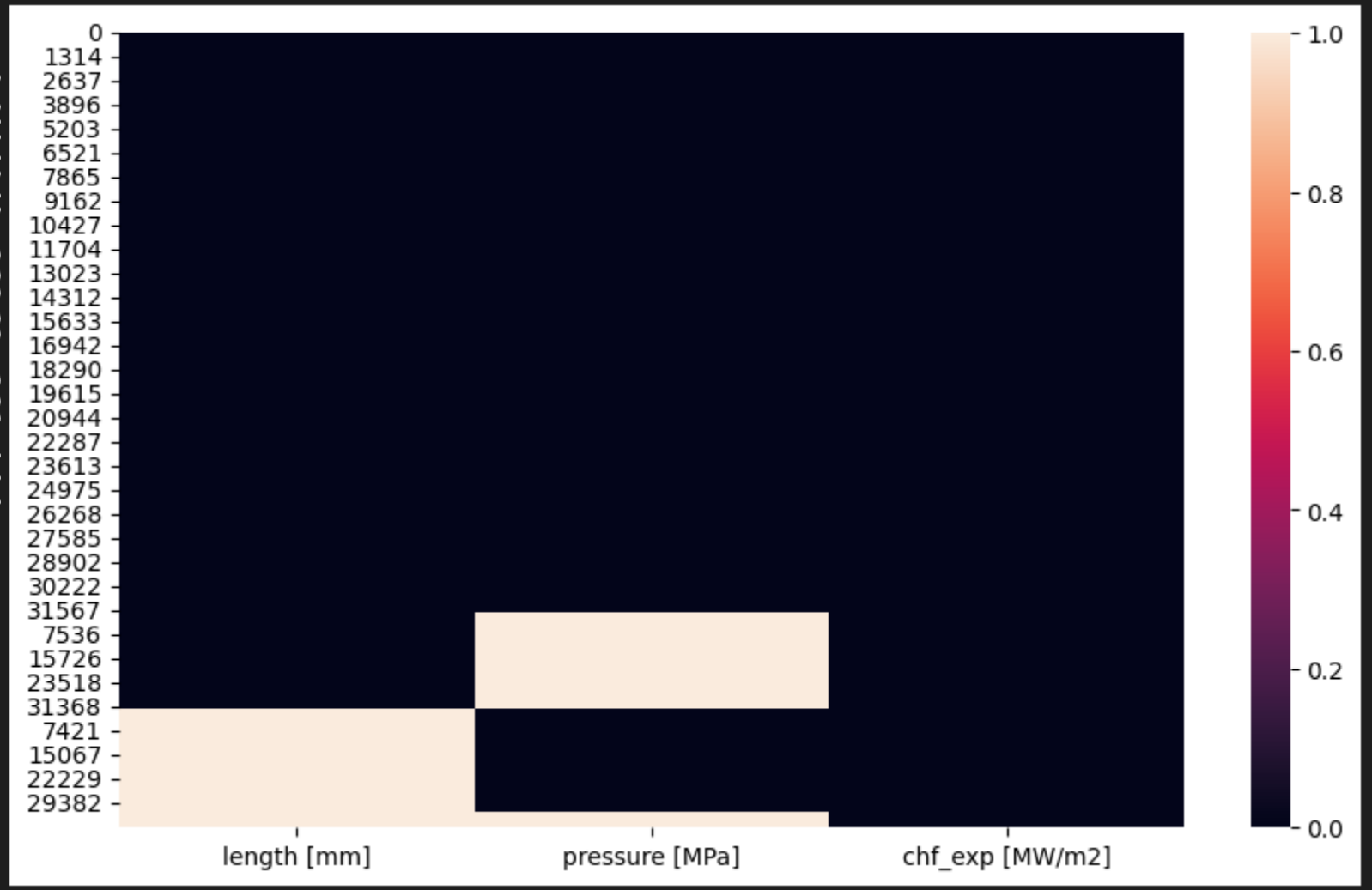
geometry	D_e [mm]	D_h [mm]	
tube	10.3	10.3	1880
	10.8	10.8	1790
	4.7	4.7	1754
	1.9	1.9	1723
	7.7	7.7	1642
			...
	10.0	40.0	1
annulus	12.8	42.3	1
plate	10.8	10.8	1
	10.0	10.0	1
annulus	11.3	4.6	1

Name: id, Length: 89, dtype: int64

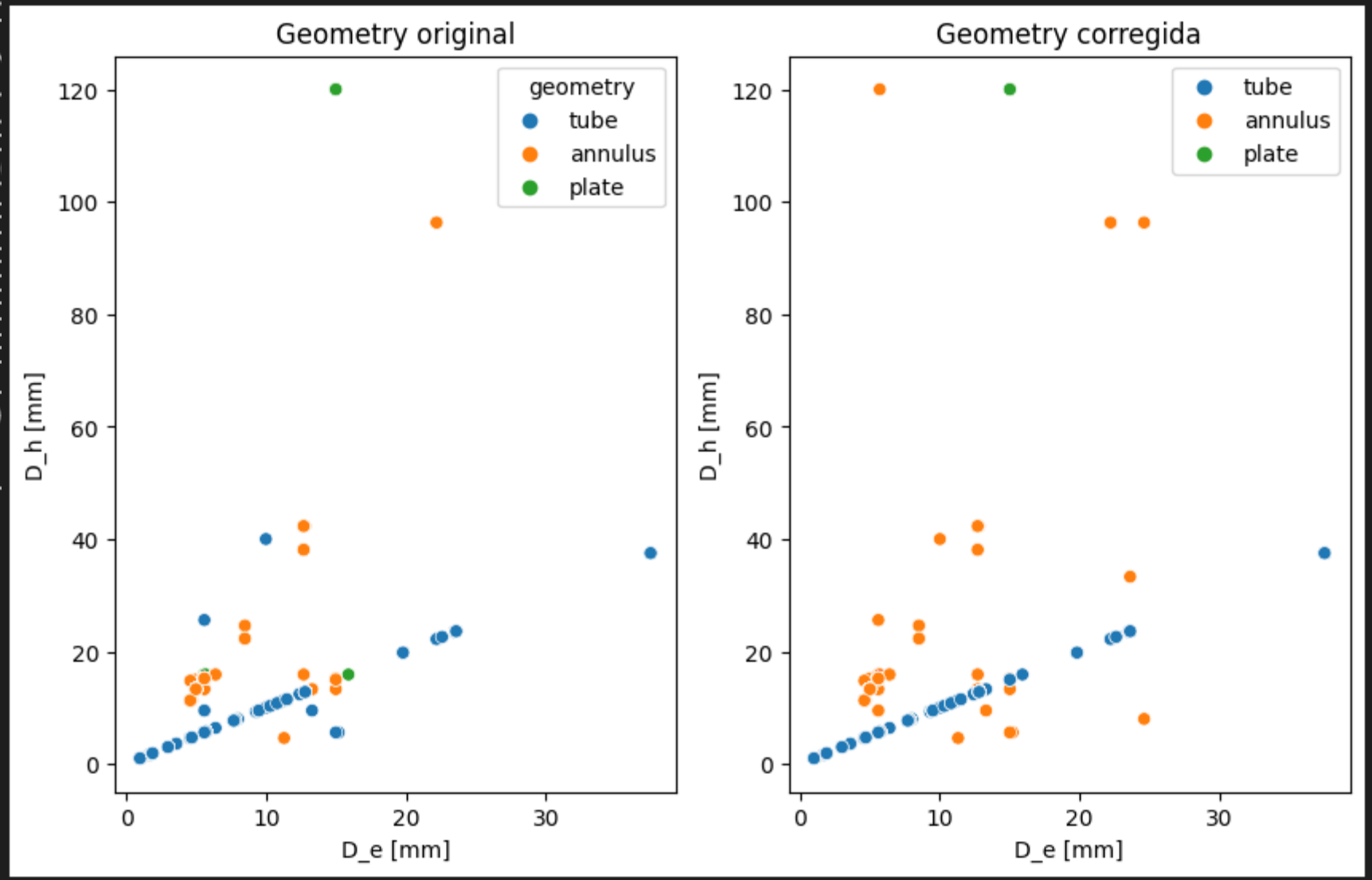
3. Feature Engineering



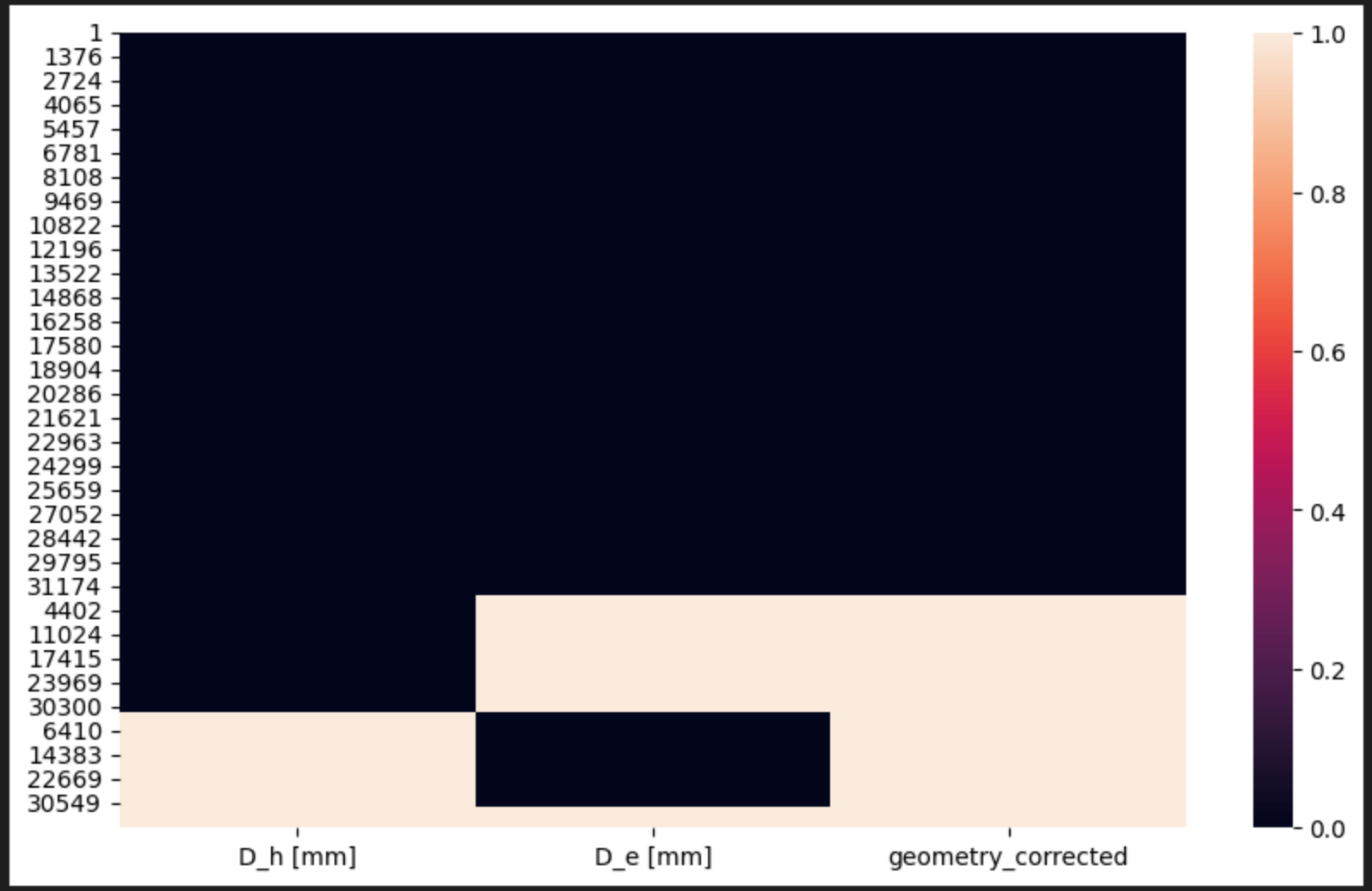
3. Feature Engineering



3. Feature Engineering



3. Feature Engineering



3. Feature Engineering

```
Data columns (total 11 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   id                   33509 non-null  int64
 1   author               28485 non-null  object
 2   geometry             28009 non-null  object
 3   pressure [MPa]       33509 non-null  float64
 4   mass_flux [kg/m2-s]  33509 non-null  float64
 5   x_e_out [-]          23094 non-null  float64
 6   D_e [mm]             33509 non-null  float64
 7   D_h [mm]             33509 non-null  float64
 8   length [mm]          33509 non-null  float64
 9   chf_exp [MW/m2]      33509 non-null  float64
 10  geometry_corrected   30254 non-null  object
dtypes: float64(7), int64(1), object(3)
memory usage: 3.1+ MB
```

4. Pipeline & Grid

Modelos:

- Random Forest
- Gradient Boosting
- XGBoost

- ANN

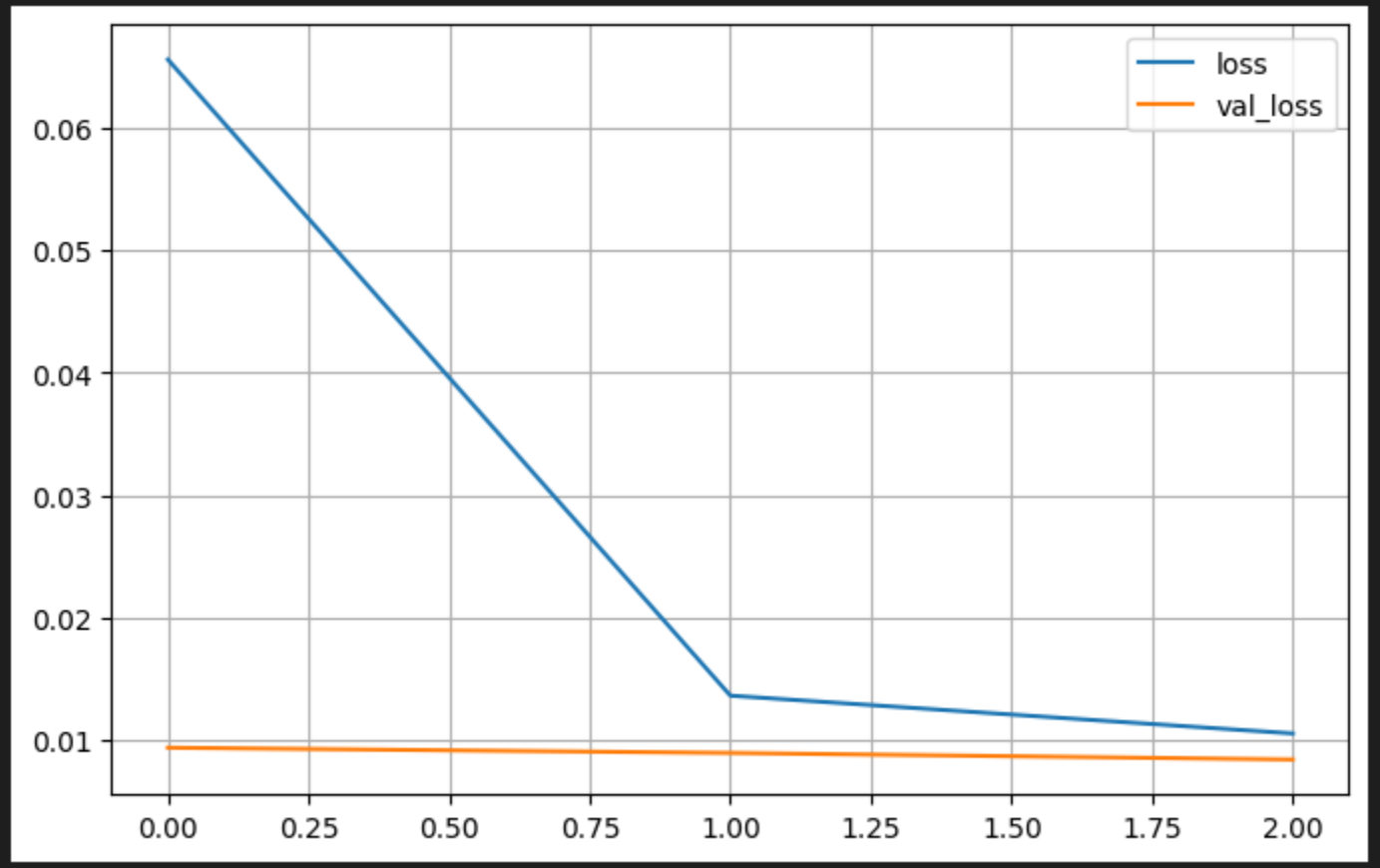
5. Análisis modelos

Mejor modelo:

	Grid	Best score
1	gs_xgb	-0.006169
0	gs_rand_forest	-0.006209
2	gs_gbr	-0.006267

Mejor modelo:
XGBoost

5. Análisis modelos



6. Conclusiones

- Importancia Feature Engineering
- Red Neuronal tiende a Overfitting