


THE BRIDGE



Ignacio Aya Sáenz | Proyecto Machine Learning

 Playground Prediction Competition

Feature Imputation with a Heat Flux Dataset

Playground Series - Season 3, Episode 15

29/05/23



1 - Introducción



2- Workflow

- 1º fase – baseline
- 2º fase - probamos otros modelos
- 3º fase - EDA + feature eng.



3 - *Exploraty Data Analysis*



4 - Métricas

- Random Forest
- XGBoost



5 - Conclusiones



Introducción

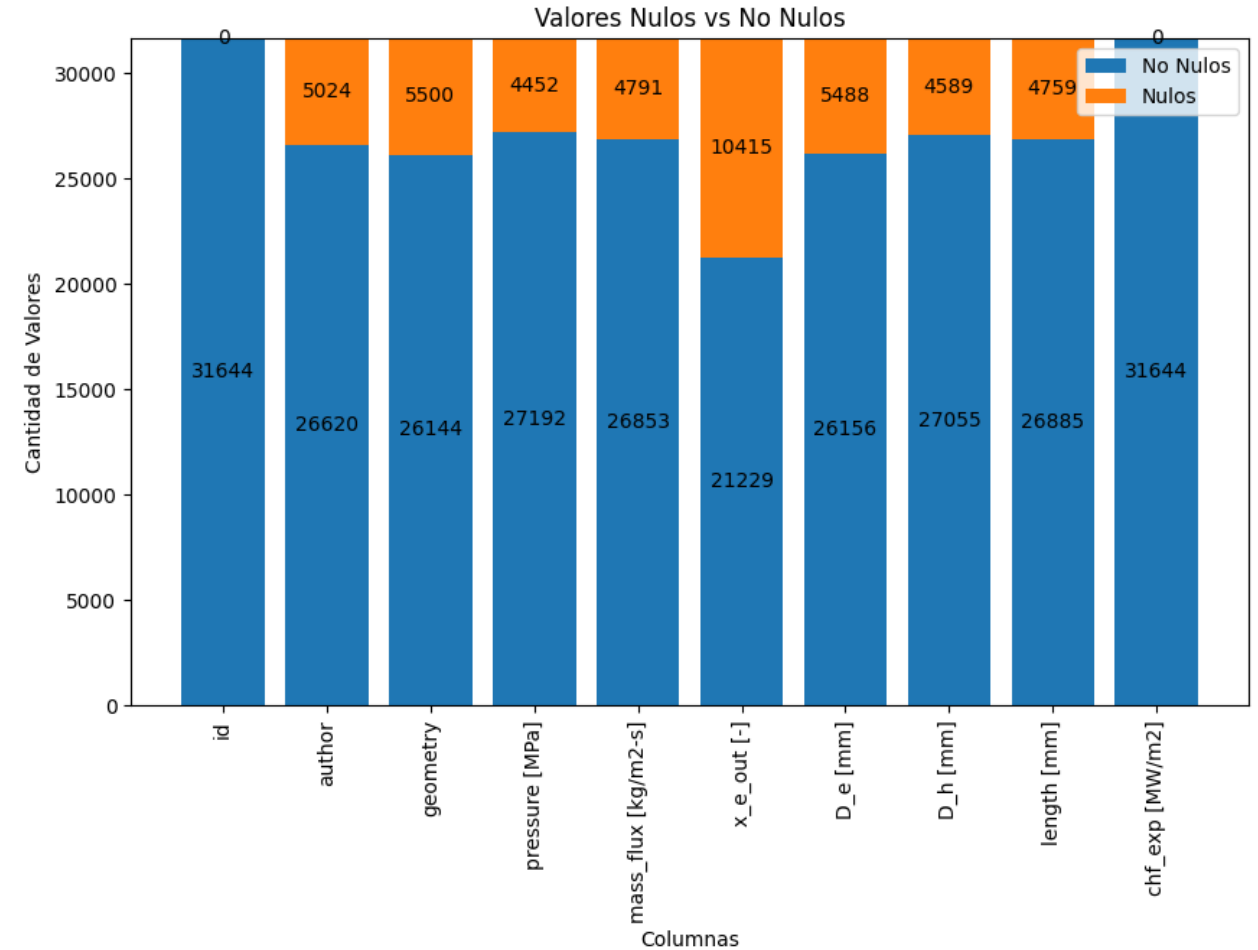
La competición consiste en trabajar con un **conjunto de datos de flujo de calor** muy utilizado en el campo de la ingeniería y la física.

Sin embargo, el conjunto de datos **presentaba ciertas características faltantes**, lo que dificultaba su análisis y modelado preciso.

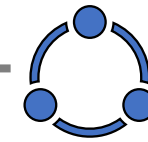
El objetivo del proyecto es **desarrollar modelos de machine learning que puedan imputar (es decir, completar o estimar) los valores faltantes** en un conjunto de datos relacionados con el flujo de calor.

	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

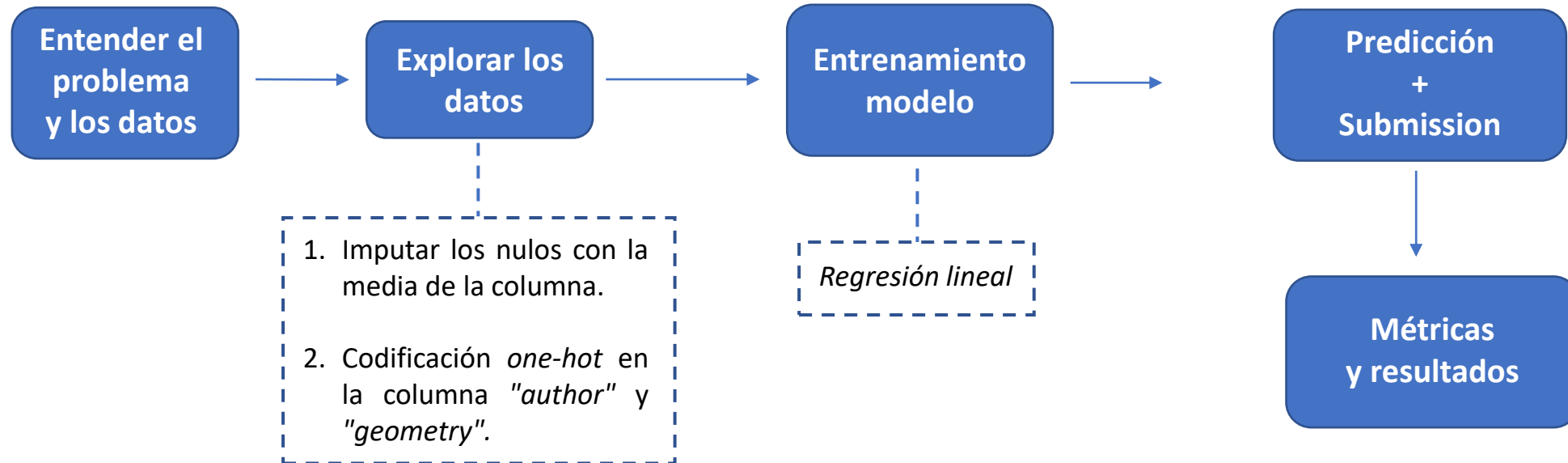
Nº de filas y columnas: (31644, 10)



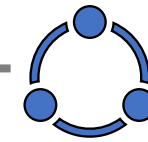
Nota: la variable objetivo $x_{e_out} [-]$



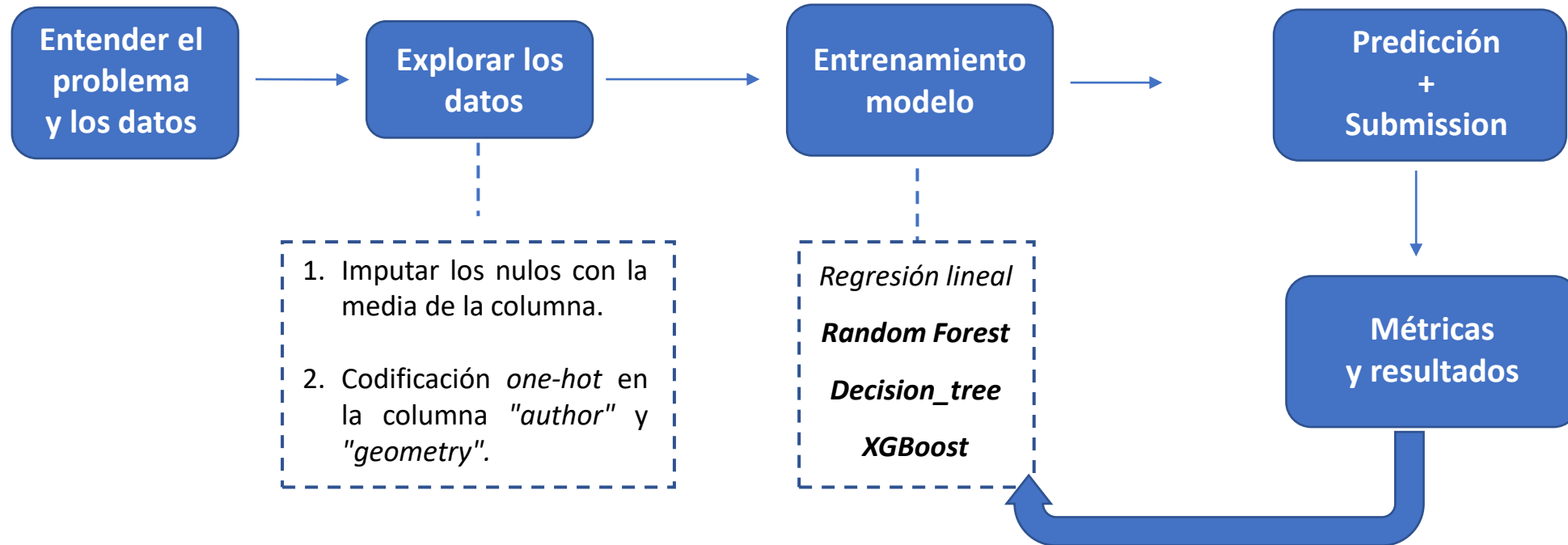
1º fase (*baseline*)



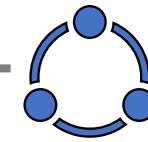
Nº	Modelo	RMSE	R2	MSE	MAE	Ranking Kaggle	position	total
1	Linear Regresión - Baseline	0,08657	0,26211	0,00750	0,06299	88%	356	405



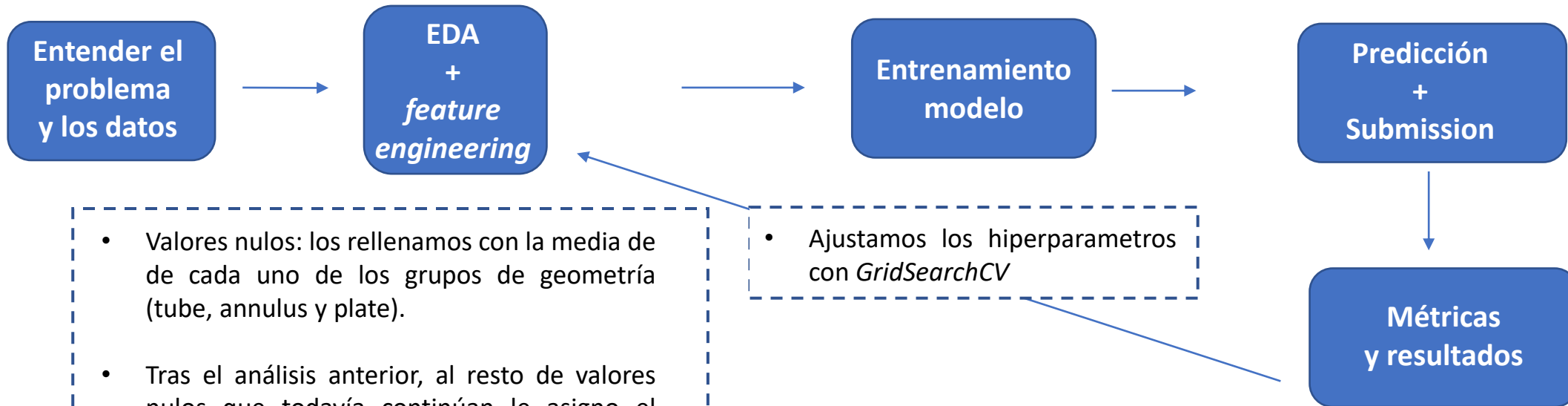
2º fase (*probamos otros modelos*)



Nº	Modelo	RMSE	R2	MSE	MAE	Ranking Kaggle	position	total
1	Linear Regresión - Baseline	0,08657	0,26211	0,00750	0,06299	88%	356	405
2	Ramdom Forest	0,07745	0,40948	0,00600	0,05398	79%	320	405
3	Decisión_tree	0,10388	0,06220	0,01079	0,07216			
4	XGBoost	0,07725	0,41258	0,00597	0,05393	76%	312	412



3º fase (Mejorar XGBoost)



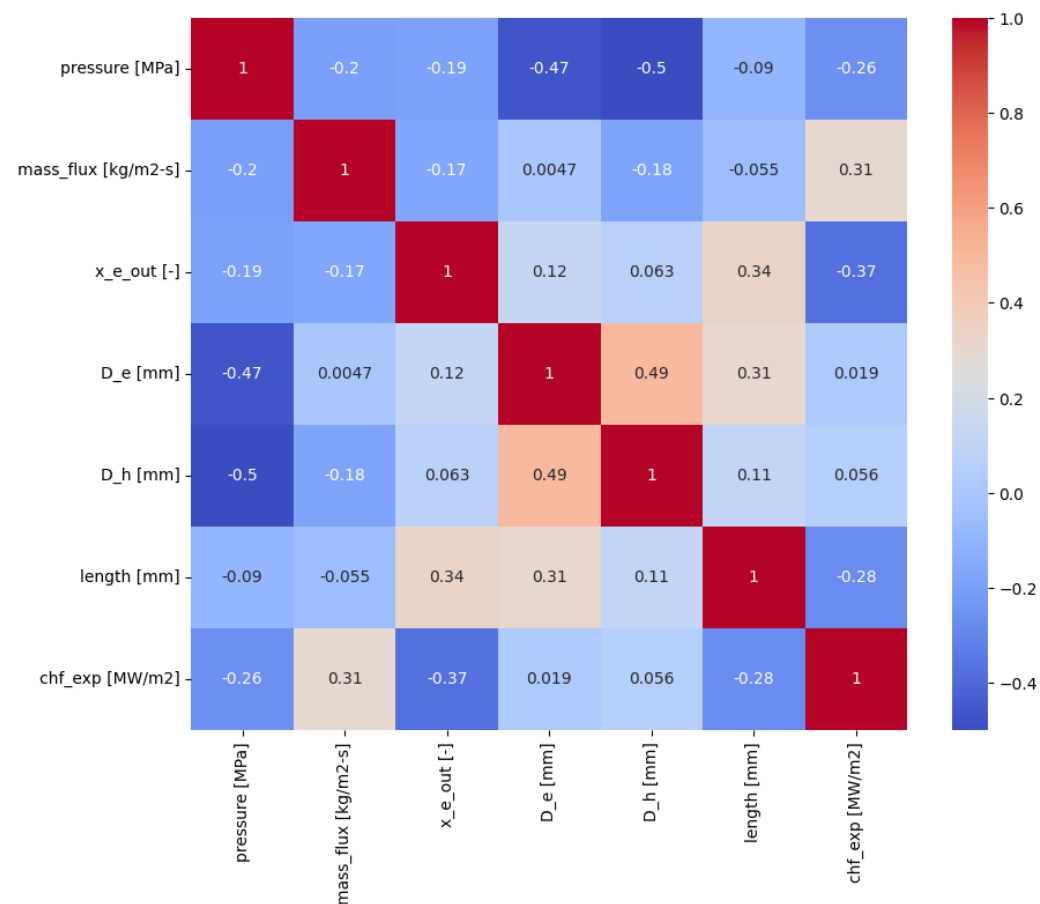
- Valores nulos: los rellenamos con la media de de cada uno de los grupos de geometría (tube, annulus y plate).
- Tras el análisis anterior, al resto de valores nulos que todavía continúan le asigno el valor medio de la columna.
- Se elimina la columna ID.
- Para hacer el modelo más simple y dado que no considero que las columnas de author y geometría sean relevantes, decidimos eliminarlas.

- Ajustamos los hiperparametros con *GridSearchCV*

Nº	Modelo	RMSE	R2	MSE	MAE	Ranking Kaggle	position	total
1	Linear Regresión - Baseline	0,08657	0,26211	0,00750	0,06299	88%	356	405
2	Ramdom Forest	0,07745	0,40948	0,00600	0,05398	79%	320	405
3	Decisión_tree	0,10388	0,06220	0,01079	0,07216			
4	XGBoost	0,07725	0,41258	0,00597	0,05393	76%	312	412
5	XGBoost_2	0,07509	0,44487	0,00564	0,05215	46%	274	600



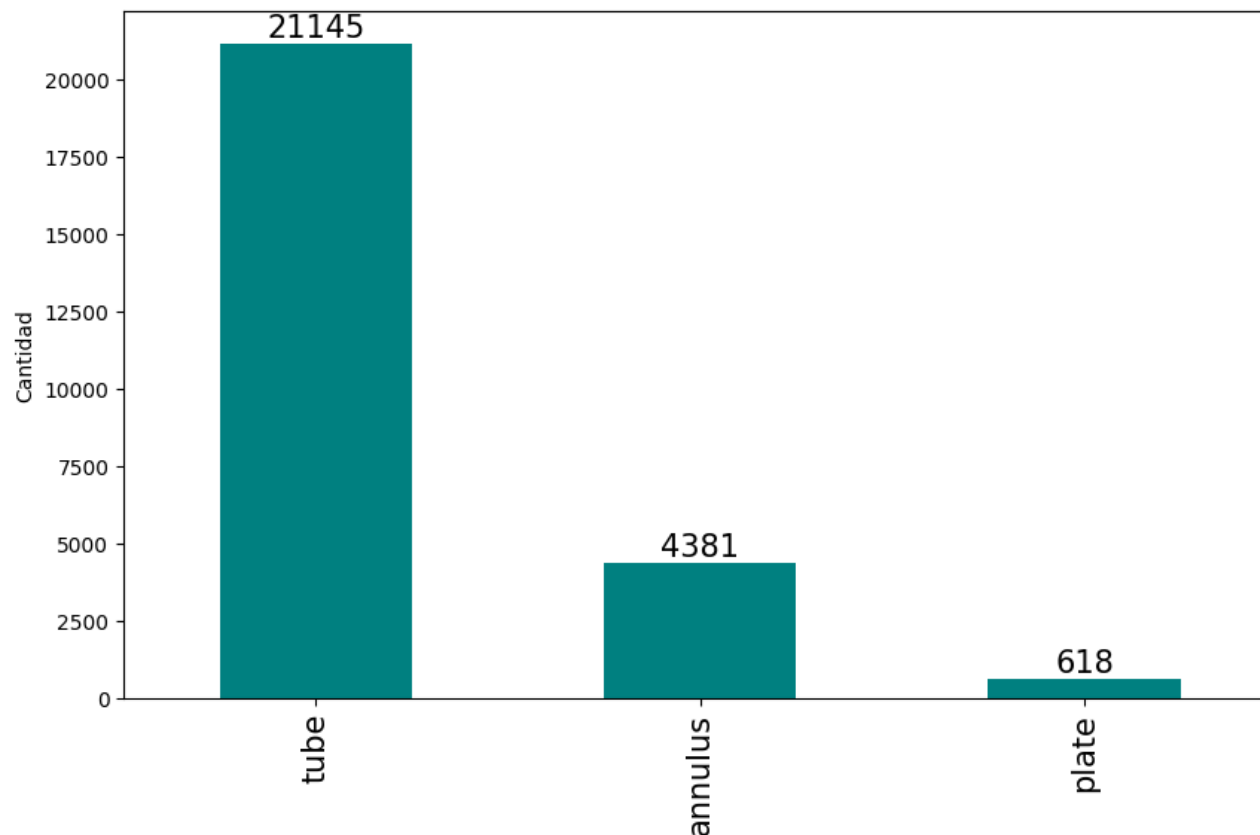
Matriz de correlación



La geometría puede influir en cómo se relacionan estas dos variables

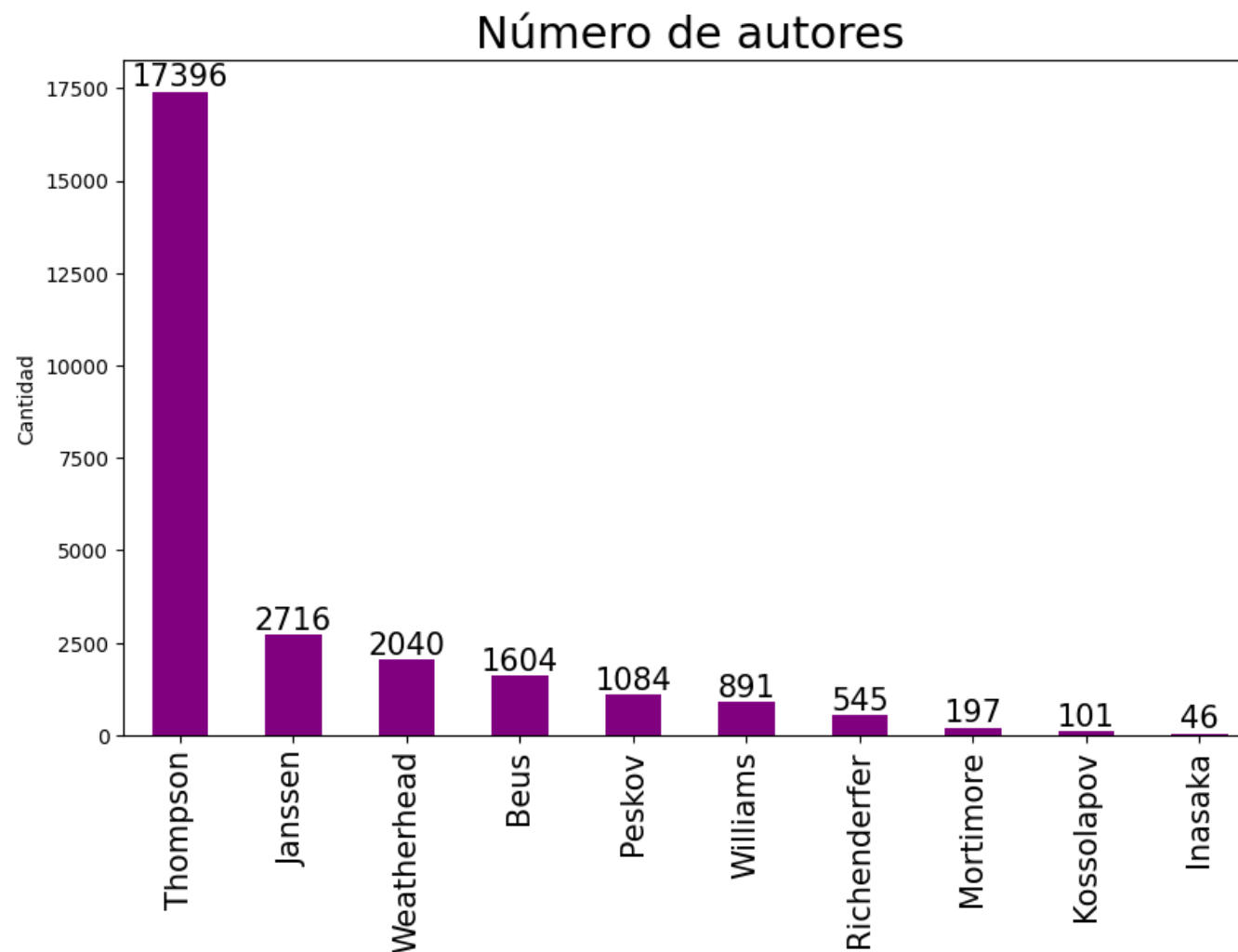


Geometría

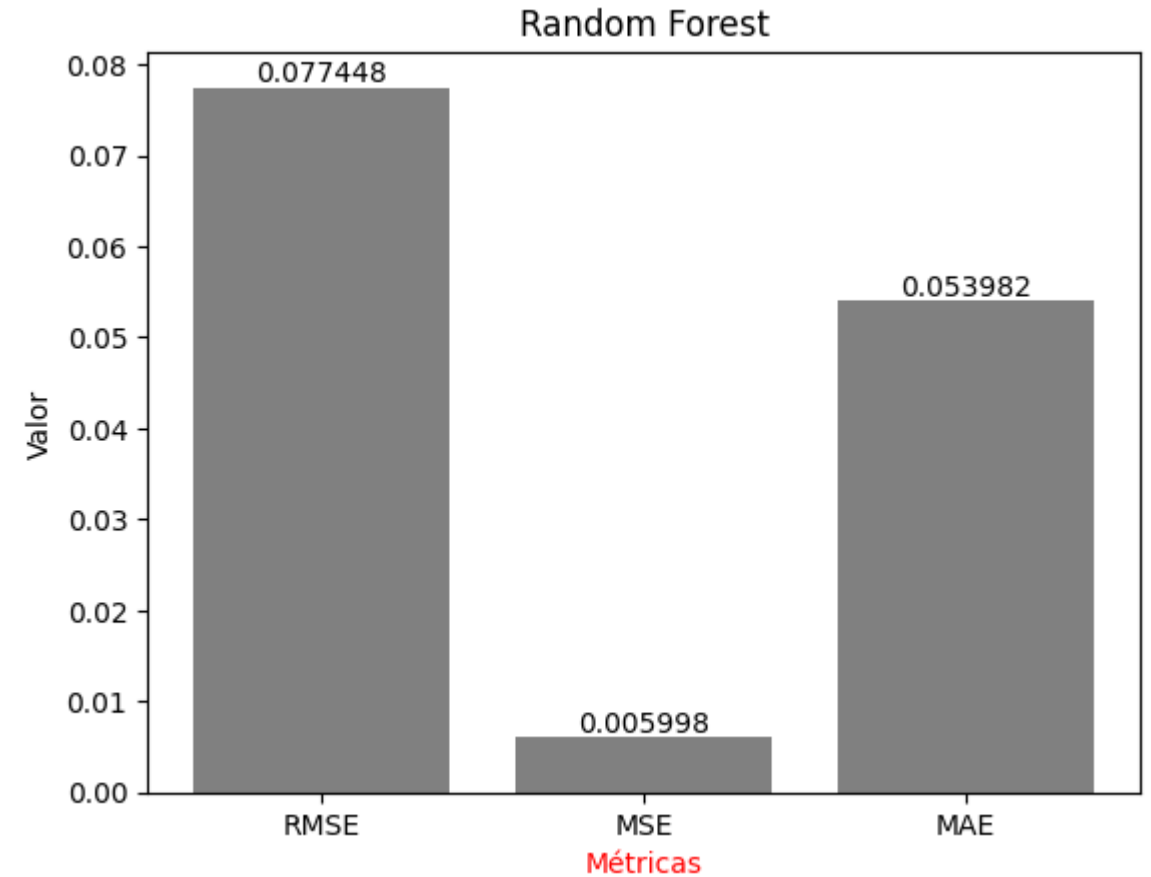
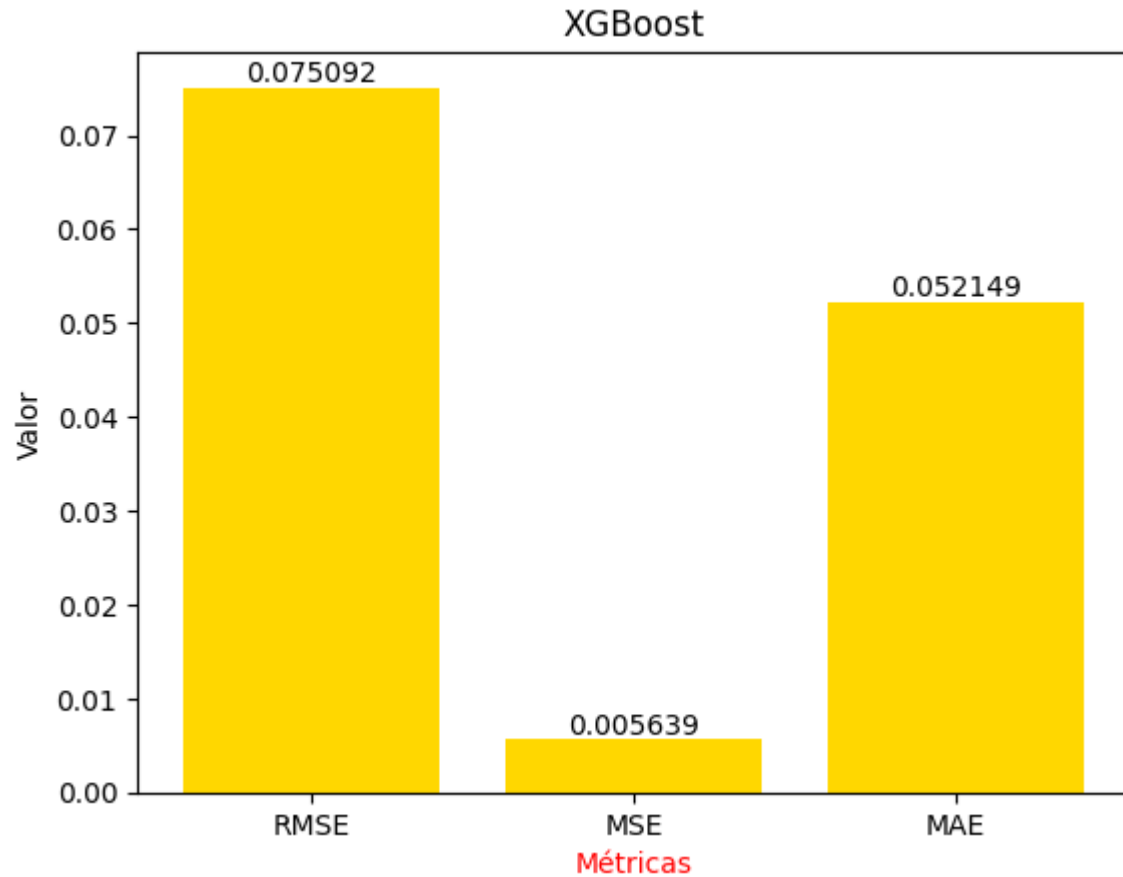


	tube	annulus	plate
id	15821.950674	16016.758959	15707.048544
pressure [MPa]	11.163435	9.472251	0.908117
mass_flux [kg/m2-s]	3234.336062	2456.423397	1563.629344
x_e_out [-]	-0.010808	0.052503	-0.033045
D_e [mm]	8.406165	8.792583	14.750193
D_h [mm]	8.736598	26.455987	117.063269
length [mm]	661.744978	1783.605341	28.539924
chf_exp [MW/m2]	3.928603	2.907669	5.279288
author	NaN	NaN	NaN
geometry	NaN	NaN	NaN

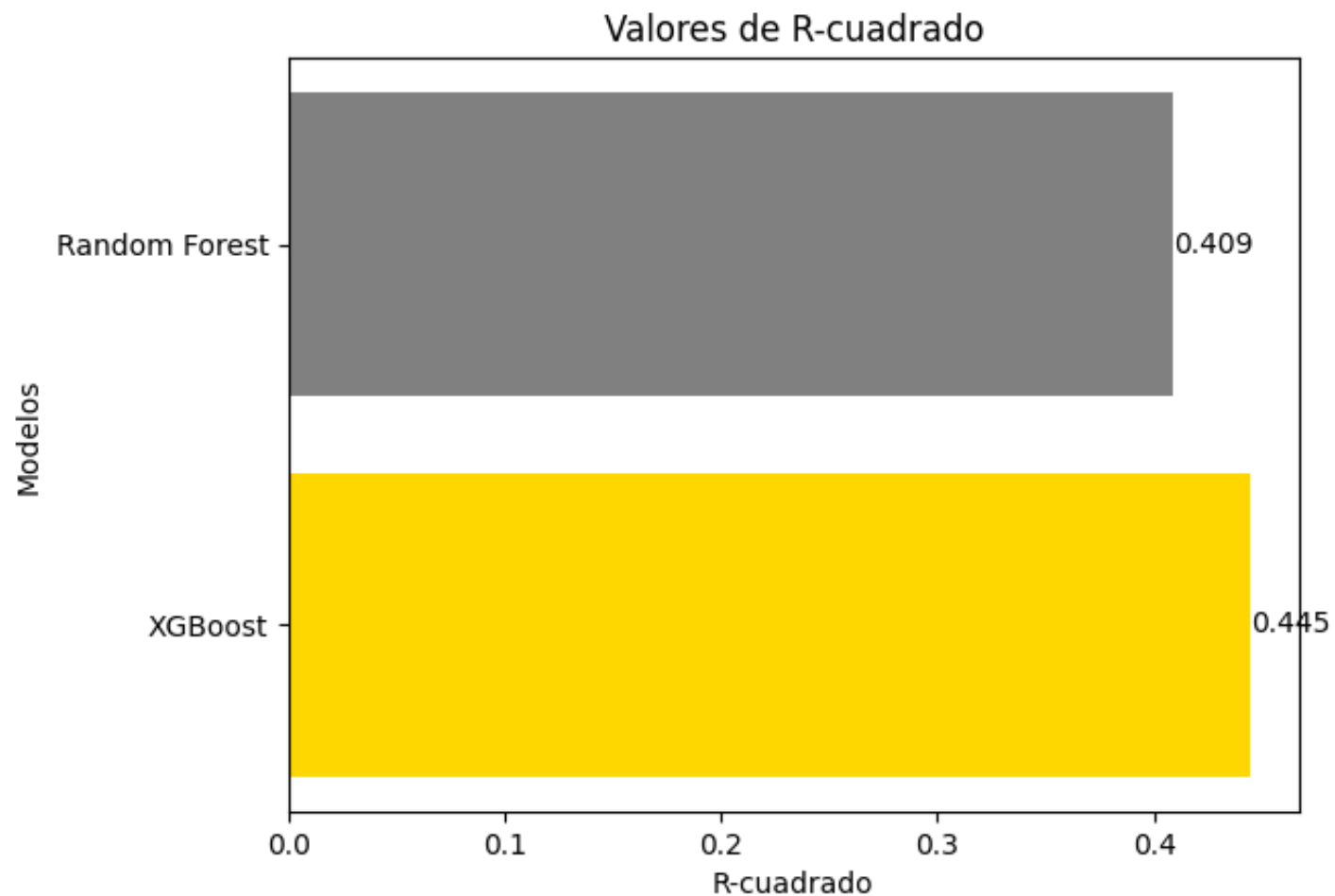
Imputamos los valores nulos con la media, según la geometría a la que pertenecen



Se elimina la columna de autor de nuestro modelo



En las métricas de precisión, XGBoost ha obtenido mejores resultados que Random Forest



El modelo XGBoost ha mostrado mejor ajuste que Random Forest



- **Random forest y XGBoost** son los que **demuestran mejor rendimiento** según las métricas utilizadas (RMSE, MSE, MAE, Ranking de la competición) sobre el resto de modelos utilizados
- La columna **"geometría"** **resultó útil en la imputación de valores nulos, pero se decidió eliminarla** durante el entrenamiento del modelo para simplificarlo.
- Se realizó un análisis de las columnas **"ID"** y **"Author"** en la etapa de exploración de datos, sin embargo, **se eliminaron posteriormente para simplificar el modelo.**
- Durante el **ajuste de los hiperparámetros del modelo**, se identificó que **estábamos cometiendo overfitting**, Lo que esto implicaba era que el rendimiento no era óptimo cuando se trataba de trabajar con datos nuevos.
- Considerando los modelos con mejor rendimiento, como XGBoost y Random Forest, se observó que **XGBoost obtuvo métricas superiores y una mejor puntuación en el ranking público de la competición.**

Gracias