22-11-2024

Inversión en Solar

¿En qué país tendríamos oportunidad de negocio?

GENMA SANZ
ML – THE BRIDGE

Contents

ELEGIR TEMATICA

¿En qué países tendremos oportunidad de negocio?

¿Qué características tienen los países que apuestan por energía solar?

OBTENCION DE DATOS

Año de los datos: 2018

https://globalsolaratlas.info/global-pv-potential-study

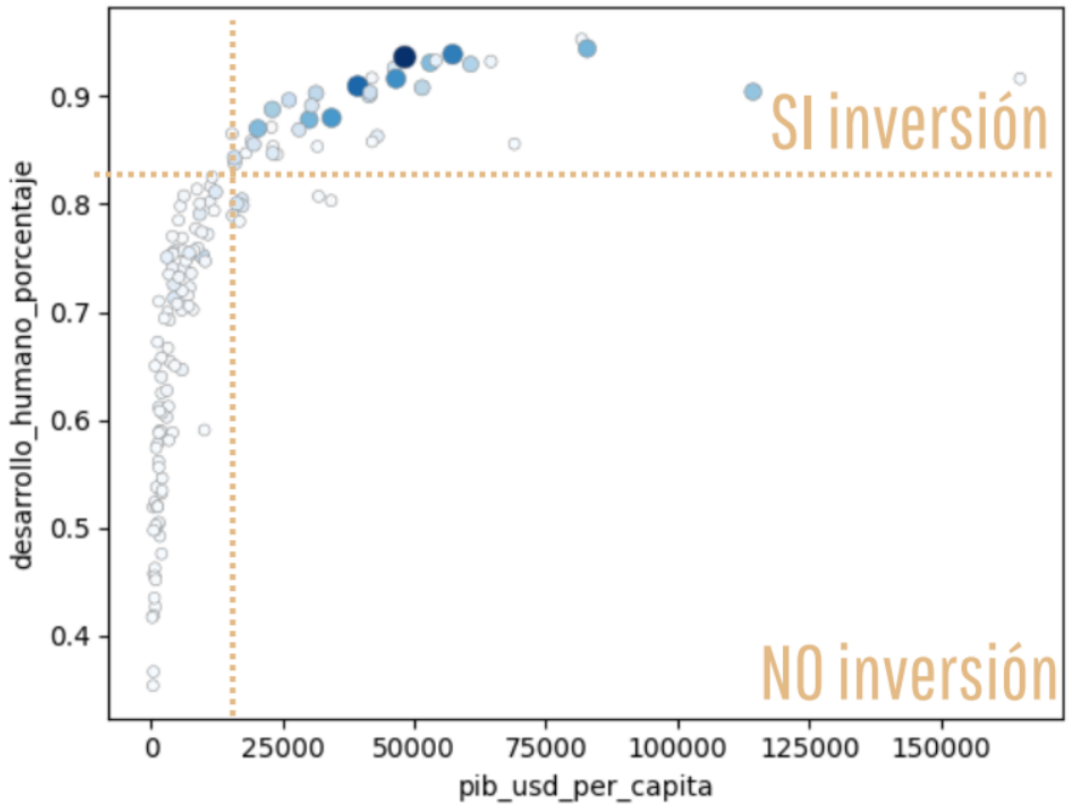
VISUALIZACIÓN DE DATOS

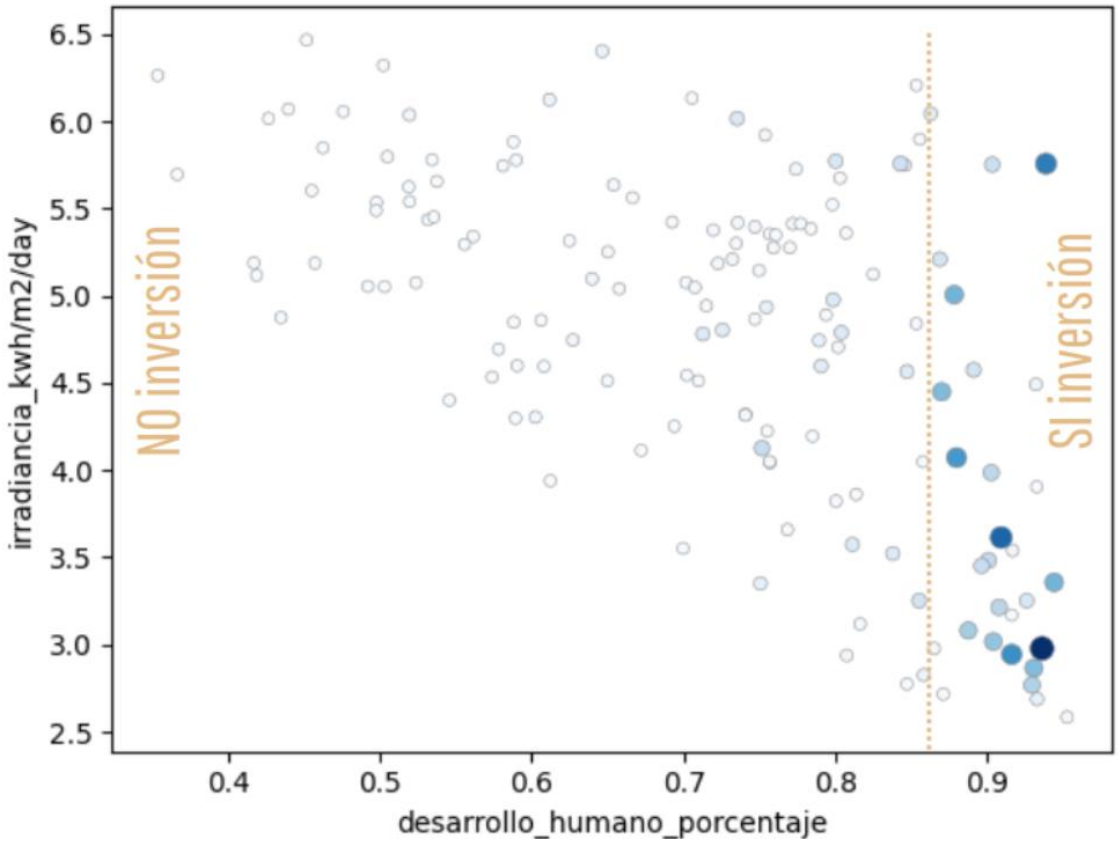
Obtención de muchos gráficos para comprender las características del dataset.

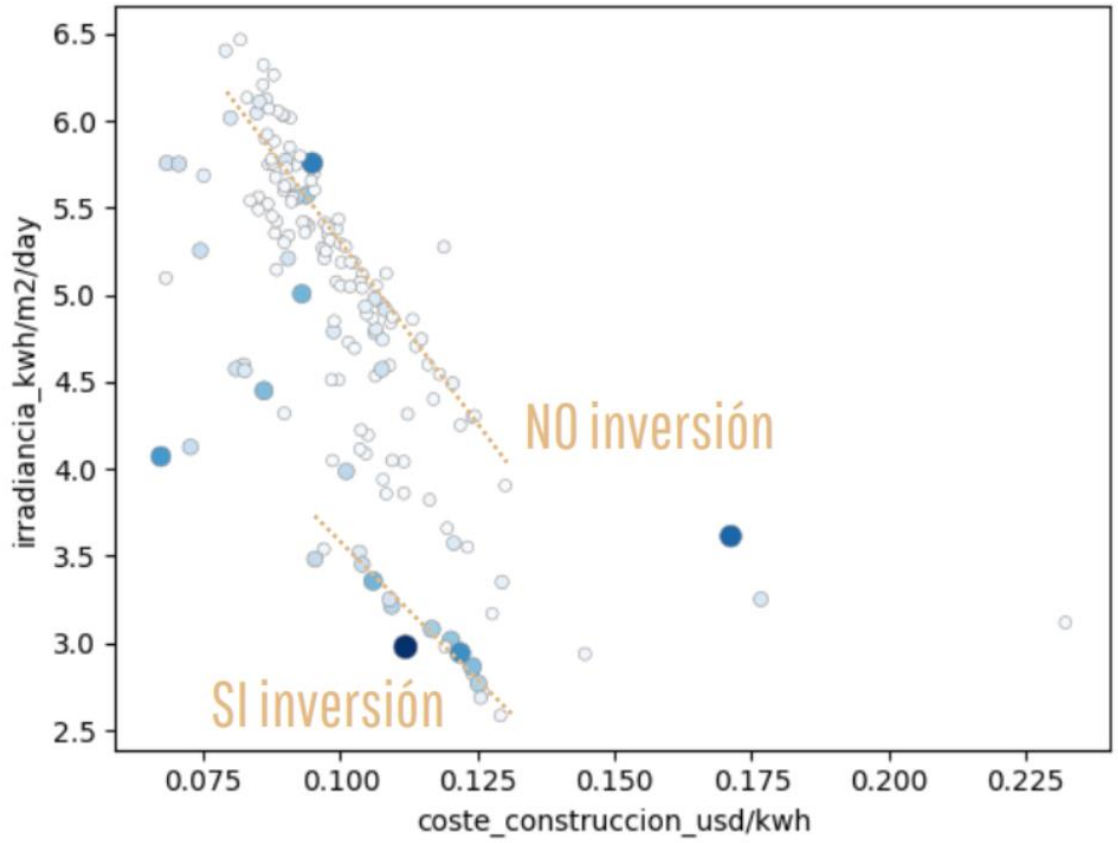
Ver que los datos están desbalanceados.

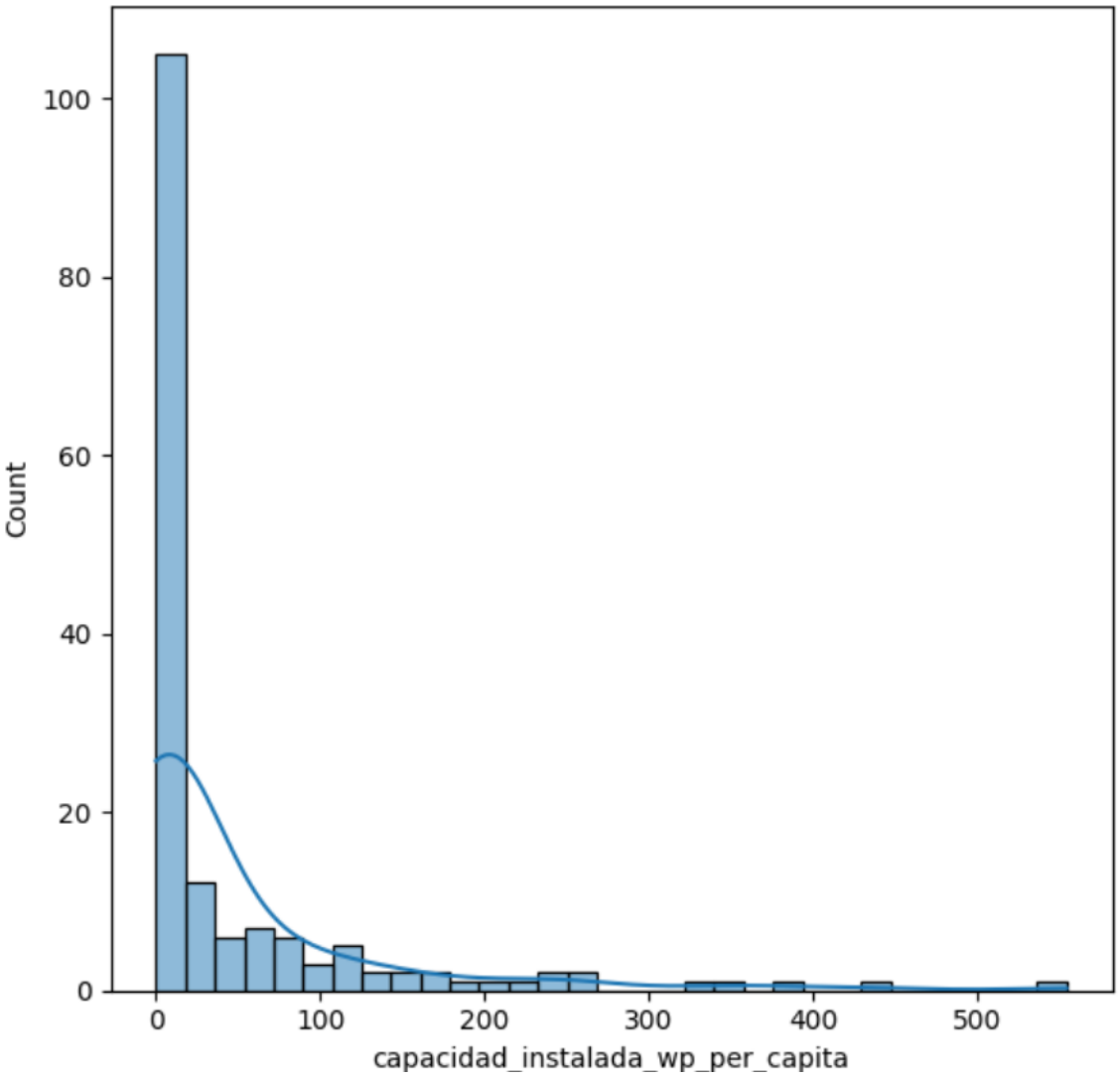
Ver que relaciones hay entre las variables.

Gráficos relevantes:




irradiancia_kwh/m2/day
6.5
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
0.4
0.5
0.6
0.7
0.8
0.9
desarrollo_humano_porcentaje
NO inversión
SI inversión


irradiancia_kwh/m2/day
6.5
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
0.075
0.100
0.125
0.150
0.175
0.200
0.225
coste_construccion_usd/kwh
NO inversión
SI inversión



MODELO SUPERVISADO

1. **TRAIN & TEST SPLIT**

Juntar los datos y ponerlos en formato entendible

2. **DEFINIR EL TARGET SUPERVISADO Y FEATURES**

Los valores iniciales del target son la capacidad instalada, para convertir el target en categórico, hemos aplicado las mismas acciones al Train y Test.

Además de escoger las features: PIB, Irradiancia y desarrollo humano.

3. **TRATAR MISSINGS – KNN IMPUTER**

Después de probar varias opciones, el mejor resultado se ha obtenido con el KNN Imputer n= 5.

Se pueden encontrar las opciones en la tabla inferior.

4. TRATAR EL DESBALANCEO DE DATOS – RANDOM UNDER SAMPLER

Al tener el dataset desbalanceado quitamos casos de “No invertir” para ser capaz de detectar mejor los “Invertir”

5. BASELINE

Se han utilizado muchos modelos iniciales para ver cuales son las mejores métricas. En la tabla solo reflejo los mejores modelos de cada iteracción.

Finalmente AdaBoost ha sido el que mejor métrica y más fiable me parece para este caso.

6. MODELO SUPERVISADO DEFINITIVO – ADABOOST

Se ha aplicado un GridSearchCV y al aplicarlo vimos como disminuía un poco la métrica, síntoma de over/underfitting. Entiendo que es porque el dataset tiene pocas instancias.

El modelo AdaBoost busca asignar pesos en el inicio del modelado, y en función a la primera iteracción junto con el calculo del error, vuelve ha recalcular los pesos, y así sucesivamente.

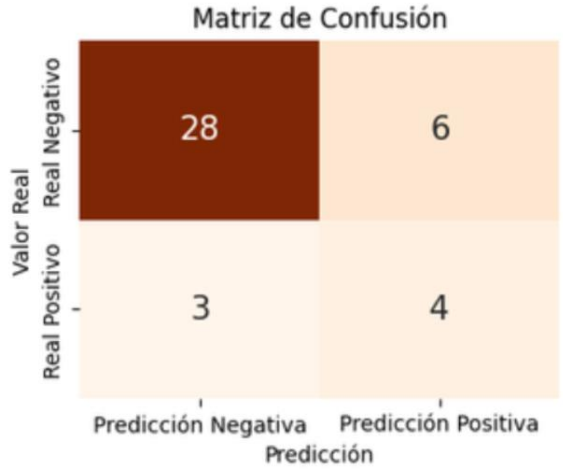
Este modelo es bastante simple, y por tanto, rápido. Se adapta bien a los dataset desbalanceados como es este caso, y no necesita escalas o normalización de los datos.

Como es bastante sensible al ruido, se han dejado solo 3 features: pib, desarrollo humano e irradiancia.

Métricas finales:

	Precisión	Recall	F1 Score
No invertir	0,9	0,82	0,86
Invertir	0,4	0,57	0,47

Accuracy	0,78



PRUEBAS SUPERVISADO CATEGORICO

MISSING	ESCALADO	FEATURES	MODELO	METRICA
Knn 5	Min max	Todas	perceptron	Recall 71%
Knn 10	Min max	Todas	Perceptrón/árbol decisión /ada boost	43% / 57% / 57%
Knn5 + undersampling	Min max	Todas	Ridge lightgbm	71% y 45% de precision
Knn5 + undersampling		Todas	Ridge lightgbm	71% y 50% de precision
Mean		Todas		Menos valor
Knn5 + undersampling		Con las 4 primeras	Adaboost	71% y 56% precisión . 85%

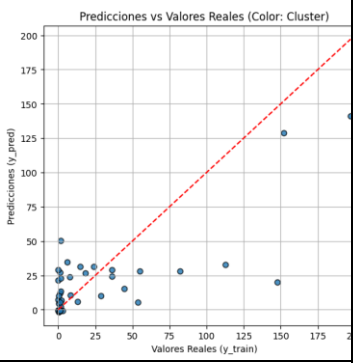
SUPERVISADO NUMÉRICO < Proyecto desechado >

La idea inicial del proyecto fue la predicción de la capacidad instalada por país.

Estos son los modelos estudiados, deseché el proyecto puesto que consideraba que la métrica era baja.

PRUEBAS SUPERVISADO NUMÉRICO

MISSING	ESCALADO	CLUSTERING	FEATURES	MODELO	TEST – r2
Media	Standard Scaler	KMEANS – 6 gen	Todas		4%
Mediana	Standard Scaker	Kmeans – 6 gen	Todas		6.6%
Mediana	MinMax Scaler	Kmeans-6 gen	Todas	Elastic net	22%
Media	MinMax	Kmeans – 6 gen	todas	Elastic net	22%
Media	Min max	Kmeans – 2 + 2 modelos diferentes	todas	+ 2 modelos diferentes	17%
Media	Min Max	K Means – 2	Totas	Elastic	24%
Media	Min Max	K Means – 3	Totas	Elastic	18%
Media	Min Max	Pca KMEANS – 1	Todas	CatBoost	30%
Media	Min max	-	Todas	Catboost	30%
knn 15	Min max		todas	elastic	23%
Knn 5	Min max			Elastric	26%
Knn5	Satandar scaler				8%
Knn5	-		Todas	MLP Regressor	44%
Simple imputer most frequent	Estándar scaler / min max / -		Todas	Catb	36%
Simple imputer most frequent	Estándar scaler / min max / -	Tratando outlayers antes del simple imputer	todas		6%

Simple most fre		Tratando los outlayers después del simple imputer	Todas		5%
Simple most fre	Min max	Tratando los outlayers después del simple imputer	Todas	cat	36%
Simple most fre	Min max		Sin desarrollo	Cat	33%
Simple most fre	Min max		Sin pib	MLP regr	26%
Knn 5	- / min max		Sin pib	Mlp	19%
Knn5	-		Todas	Knn	-12%
Knn5	-		Solo desarrollo	Ridge	19%
Knn5	-		Desarro+pib	Rdige	18%
Knn5			Desa+pib+irradiancia	Catboost	45%
Knn5			Desa+pib+irradiancia+ coste constr	Catboosr	55% 
Knn3			Desa+pib+irradiancia+ coste constr	Cat /árbol	51% /42%
Knn7			Desa+pib+irradiancia+ coste constr	Cat /árbol	46&/42%
Knn5			Pib+consumo+irra +coste con	gradient	50%
Knn5	Min max		Desa+pib+irradiancia+ coste constr	Catboosr	55%

Most freque rt	Min max		Desa+pib+irradian cia+ coste constr	Cat boost	48%