PRONÓSTICOS METEOROLÓGICOS USANDO TIME SERIES

Juan Antonio de la Cuadra
The Bridge School
DS/PT/Sep-2022

PLANTEAMIENTO

Dentro del ámbito del Machine Learning, vamos a intentar hacer unas predicciones meteorológicas, sin más. Y al ser datos diarios vamos a usar la serie temporal. Parece fácil de decidir.

Pero no, no es tan fácil.

- Época del año
- Clima de la ciudad
- Particularidades del modelo

TOMA Y LIMPIEZA DE DATOS

ORIGEN Y FORMATO DE DATOS.

Datos diarios de la estación meteorológica del Aeropuerto de Sevilla entre el 01/01/1990 y el último registro disponible. A día de cierre de proyecto es el 12/03/2023.

API AEMET: OpenData API

IDEMA - 5783

NOMBRE - SEVILLA AEROPUERTO

LOCALIDAD - SEVILLA

PROVINCIA - SEVILLA

ALTITUD - 34 msnm

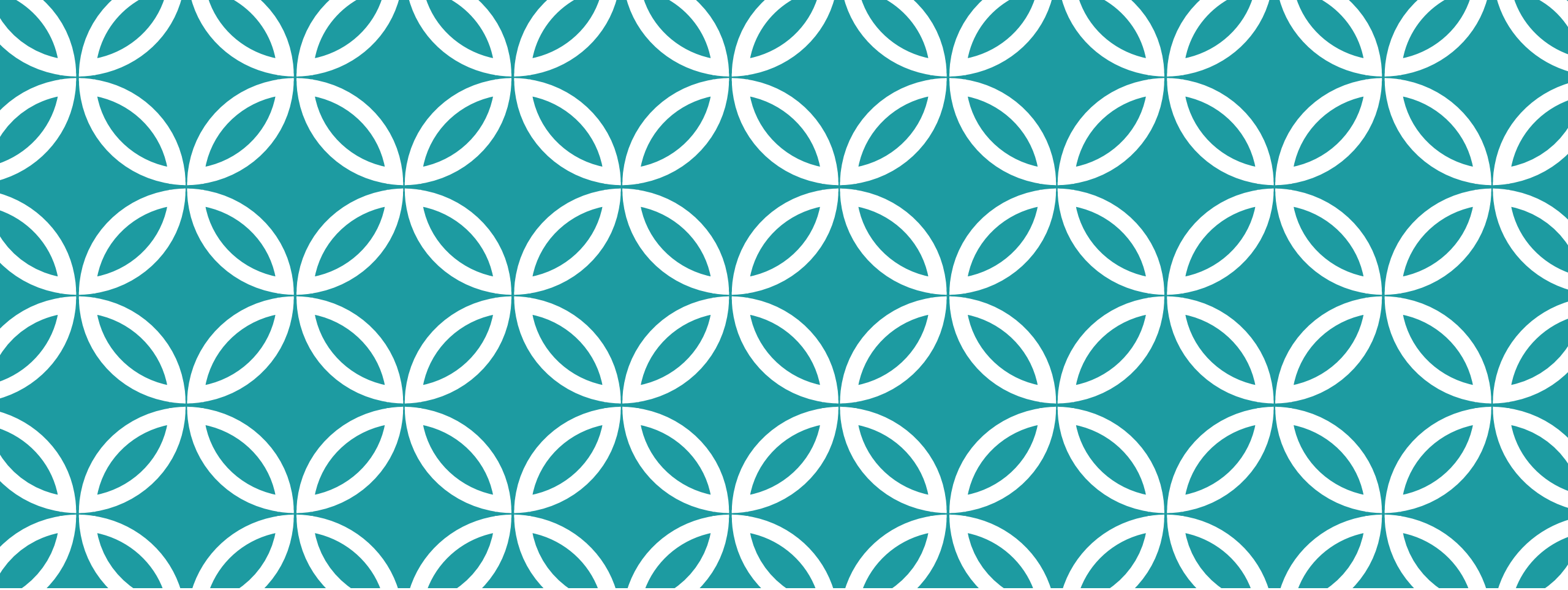
tmin -> Temperatura minima (°C)
tmax -> Temperatura máxima (°C)
tmed -> Temperatura media (°C)
presMin -> Presión mínima (milibares)
presMax -> Presión máxima (milibares)
dir -> Dirección del viento en base a los rumbos principales
velmedia -> Velocidad media del viento (km/h)
racha -> Velocidad de la racha máxima de viento (km/h)
sol -> Indice Ultravioleta
prec -> Precipitación acumulada (l/m2)

ORIGEN Y FORMATO DE DATOS.

COL_N	tmed	prec	tmin	tmax	dir	velmedia	racha	sol	presMax	presMin
DATA_TYPE	float64	float64	float64	float64	float64	float64	float64	float64	float64	float64
MISSINGS (%)	0.23	3.64	0.23	0.21	1.72	0.45	1.72	1.11	0.72	0.72
MISSINGS	28	441	28	26	209	54	209	134	87	87
UNIQUE_VALUES	309	346	290	366	37	42	81	145	371	410
CARDIN (%)	2.55	2.85	2.39	3.02	0.31	0.35	0.67	1.2	3.06	3.38

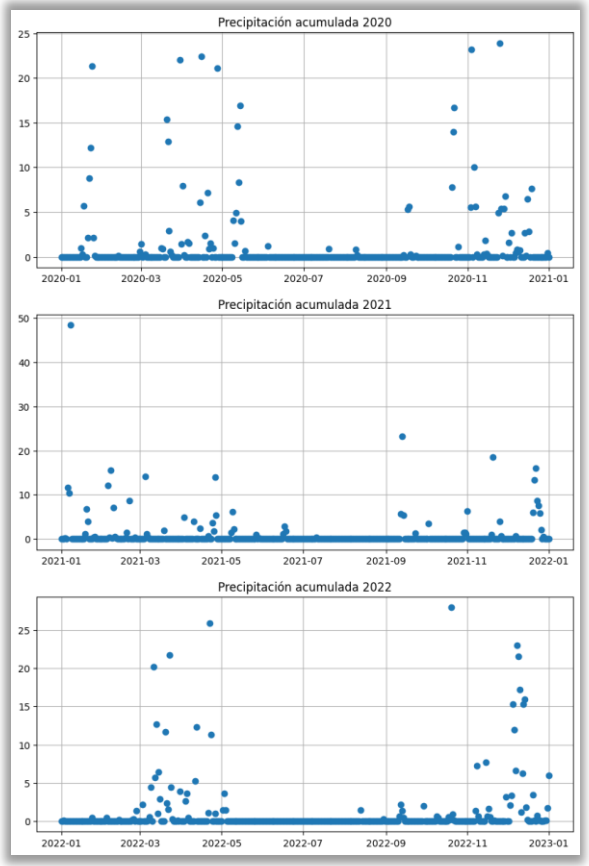
Y tras imputar los nulos con la media móvil…

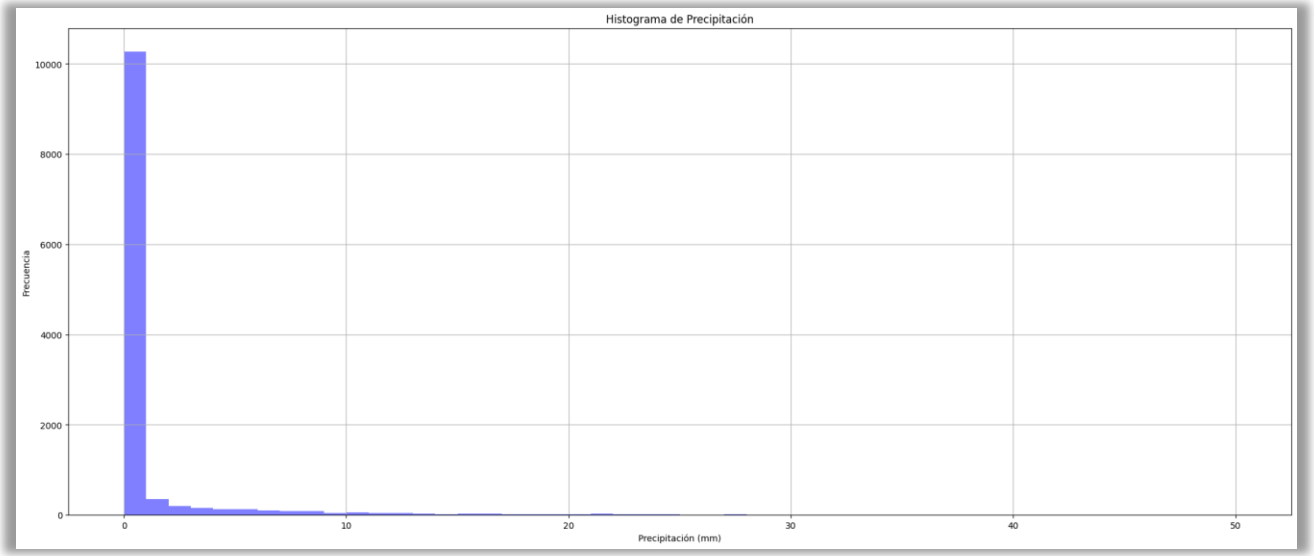
COL_N	tmin	tmax	tmed	presMin	presMax	dir	velmedia	racha	sol	prec
DATA_TYPE	float64	float64	float64	float64	float64	float64	float64	float64	float64	float64
MISSINGS (%)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MISSINGS	0	0	0	0	0	0	0	0	0	0
UNIQUE_VALUES	310	385	330	477	444	216	85	242	251	530
CARDIN (%)	2.56	3.18	2.72	3.94	3.66	1.78	0.7	2.0	2.07	4.37



EDA

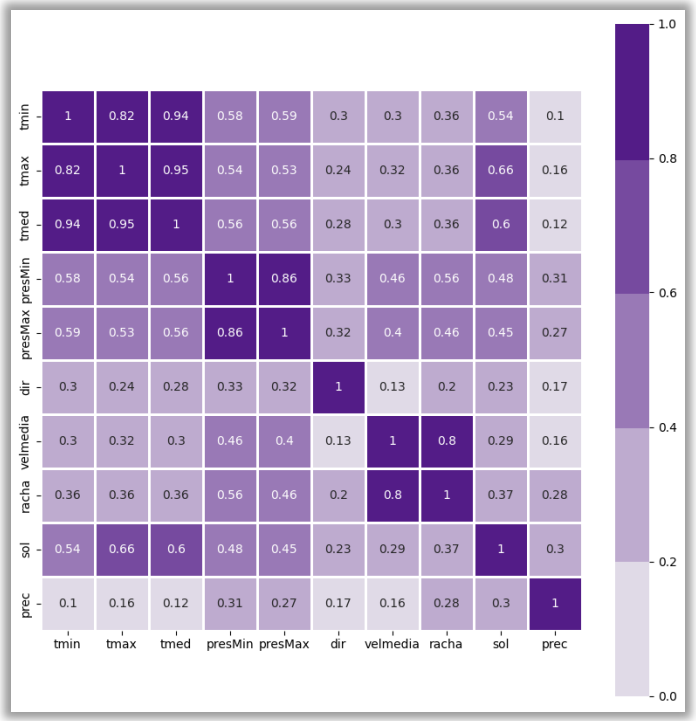
PRIMERAS IMPRESIONES





CORRELACIONES

GRÁFICA PHIK



Las mayores correlaciones se dan entre:

- Rachas de viento y velocidad media del mismo.
- Temperaturas mínimas, máximas y medias.
- Presión mínima y máxima.

FEATURE IMPORTANCE

RandomForest vs. SelectKBest

	Score	Feature
0	0.2768	sol
1	0.2219	presMin
2	0.1089	racha
3	0.0752	presMax
4	0.0712	dir
5	0.0671	tmax
6	0.0647	velmedia
7	0.0642	tmin
8	0.0499	tmed

	column	score
8	sol	16.444918
7	racha	5.528743
3	presMin	5.187728
1	tmax	4.449790
4	presMax	3.447445
6	velmedia	2.757258
2	tmed	2.564513
0	tmin	1.016548
5	dir	0.916736

Conclusiones

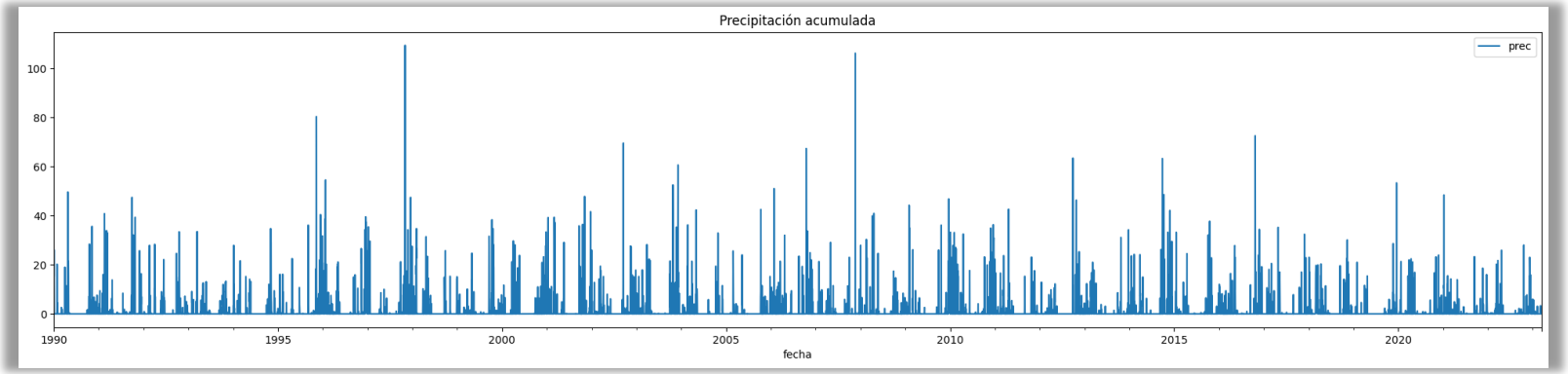
- Los más relevantes según RF son la radiación solar y la presión atmosférica mínima.

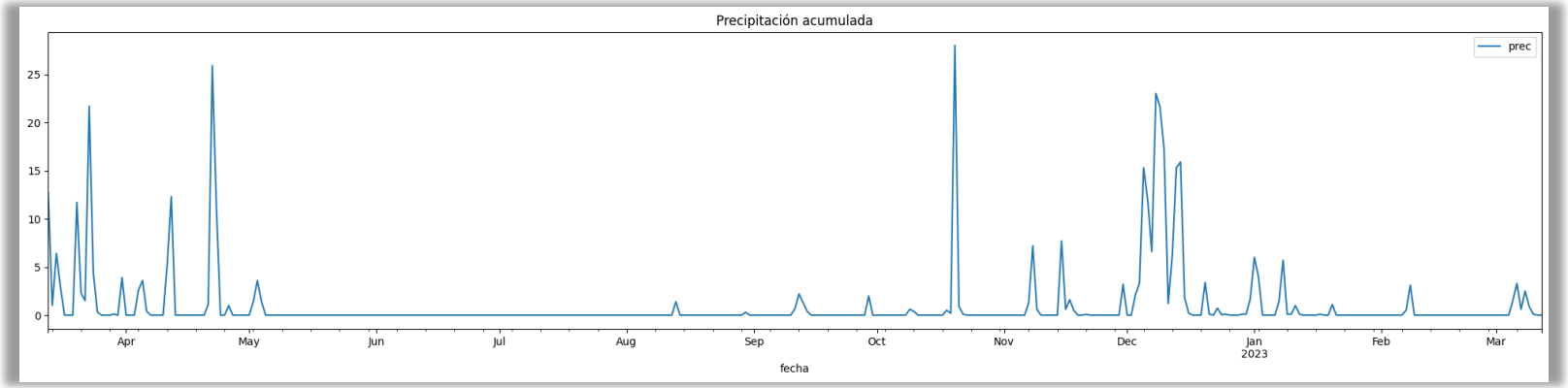
- La presión atmosférica baja cuando se acerca un frente lluvioso, por lo que parece bastante revelador este feature.

- SelectKBest nos ratifica lo que veíamos sobre la radiación solar, aunque también da importancia a las rachas de viento.

- Ante cambios rápidos de presión atmosférica, se dan mayores rachas de viento, por lo que sigue teniendo sentido.

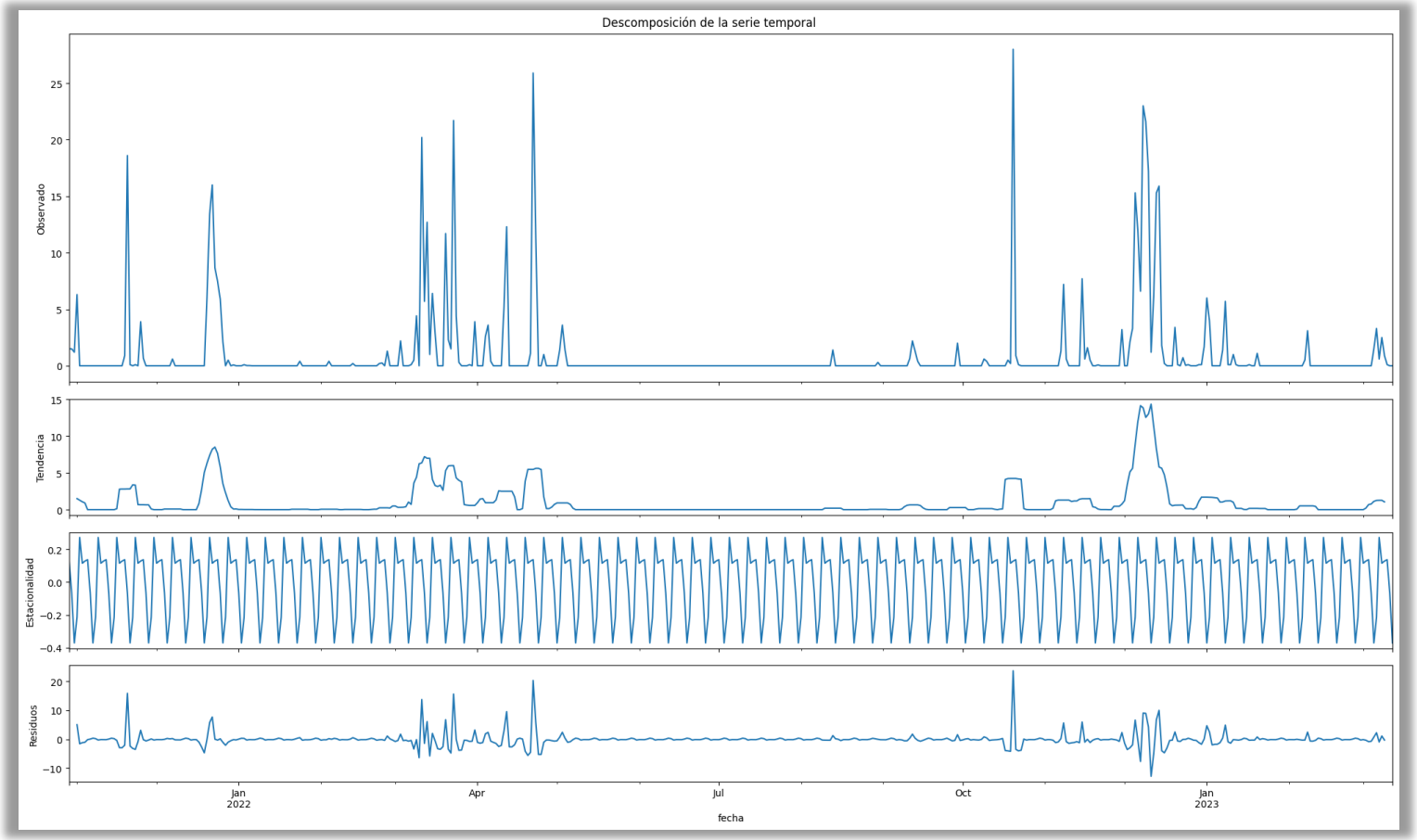
VISTA DEL TARGET



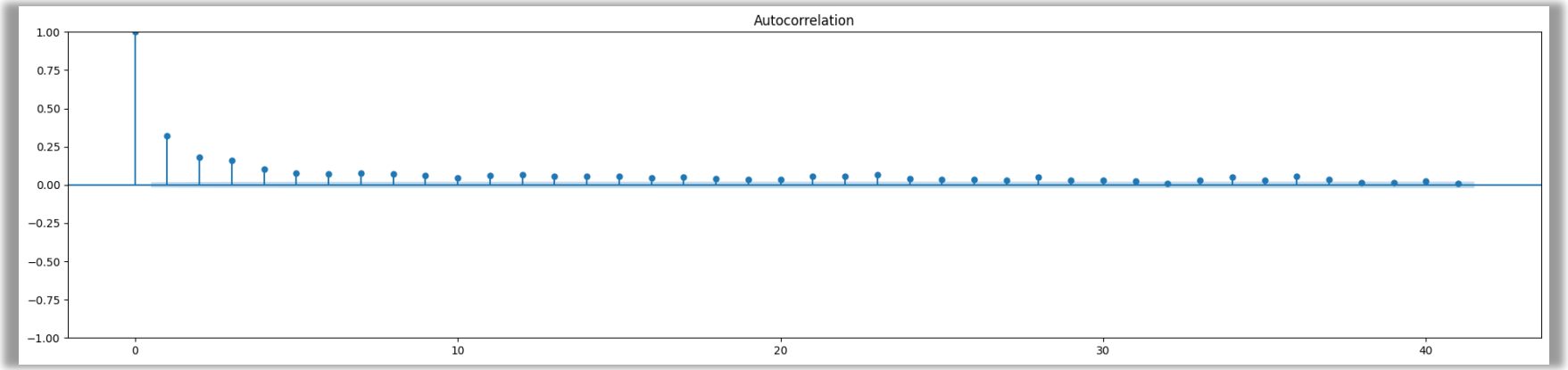


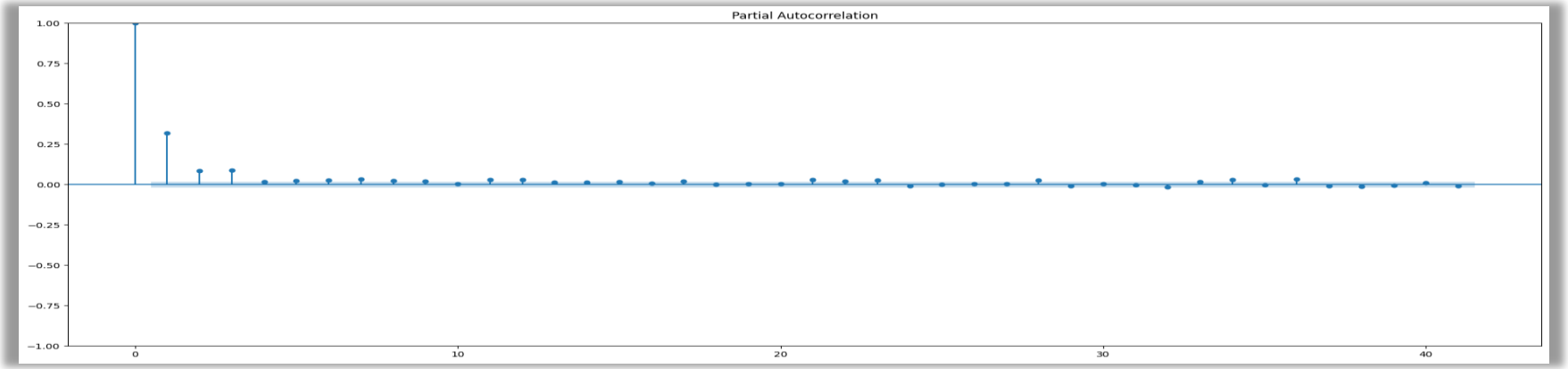
MODELOS

SEASONAL DECOMPOSE

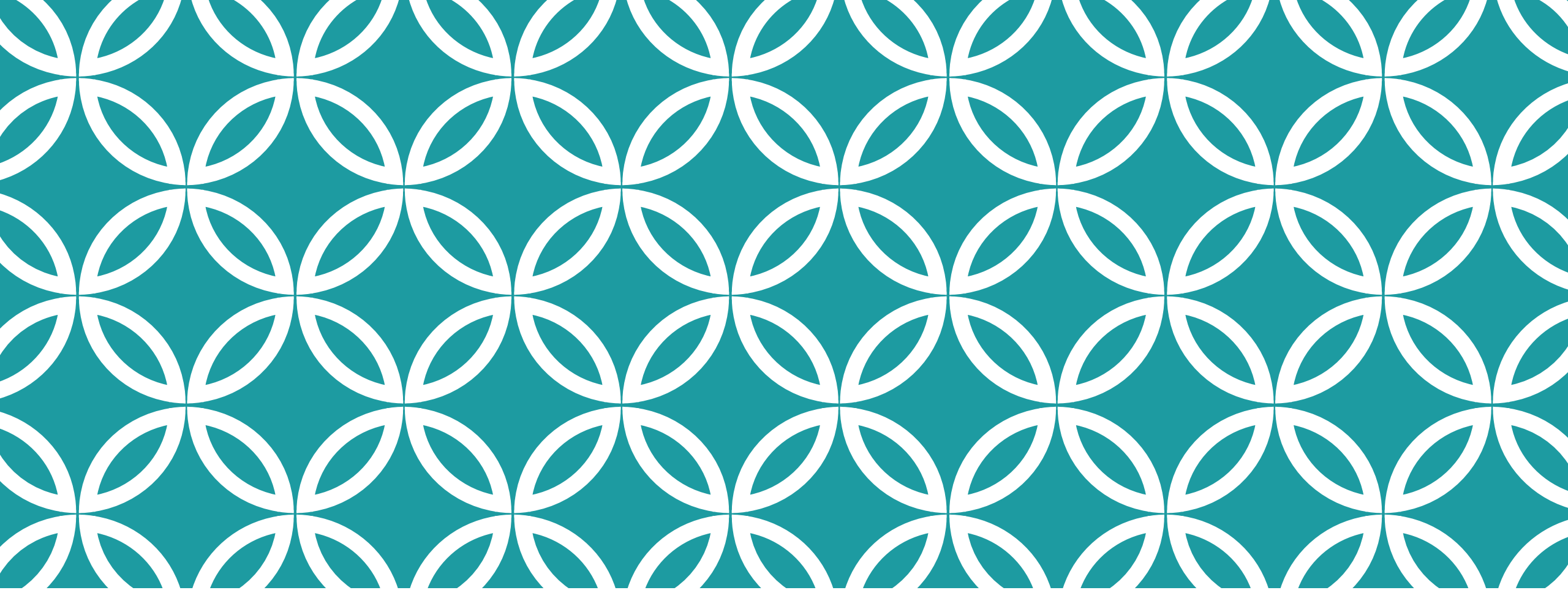


AUTOCORRELACIÓN



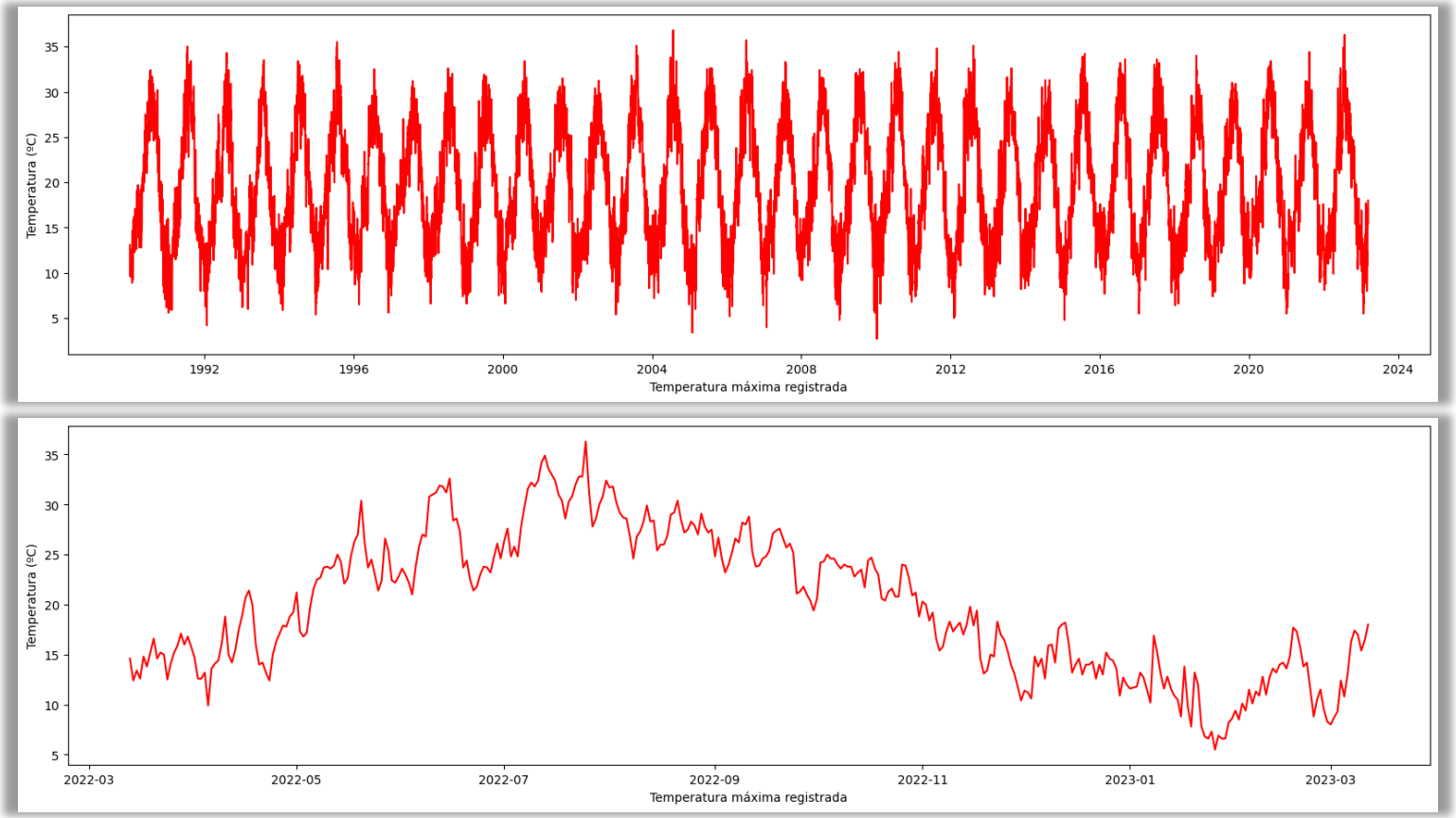



BLITZ
Jason Statham
22:30
Juan, 33
alegre pero no mucho
cuatro HD

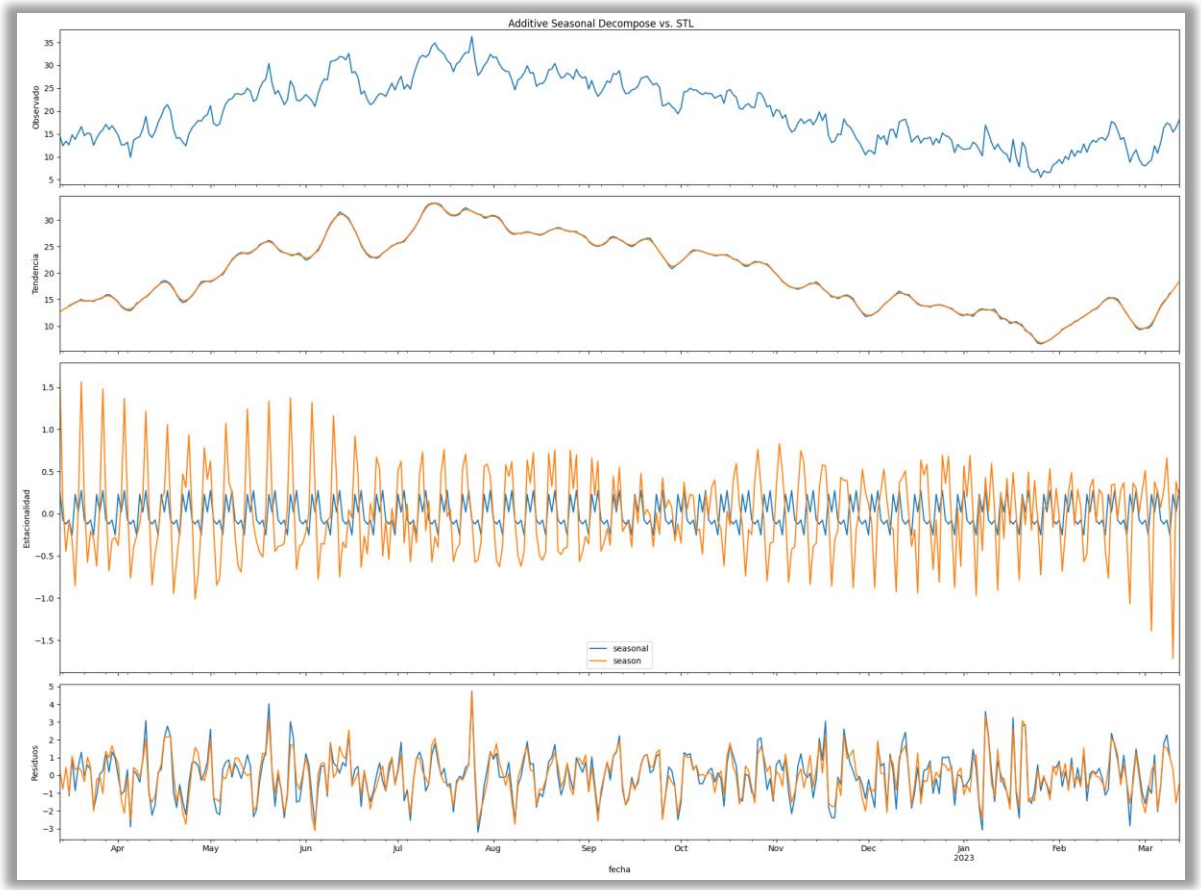


PLAN B

TEMPERATURA MEDIA



SEASONAL DECOMPOSE VS. STL



El algoritmo STL realiza el suavizado de las series temporales utilizando LOESS en dos bucles; el bucle interior itera entre el suavizado estacional y el de tendencia y el bucle exterior minimiza el efecto de los valores atípicos. Durante el bucle interno, se calcula primero el componente estacional y se elimina para calcular el componente de tendencia. El resto se calcula restando los componentes estacional y de tendencia de la serie temporal.

STLFORECAST + ARIMA

Tras un gridsearch nos da los parámetros 3, 0 , 5

STL Decomposition and SARIMAX Results

Dep. Variable:	y	No. Observations:	12124
Model:	ARIMA(3, 0, 5)	Log Likelihood	-19592.315
Date:	Sat, 18 Mar 2023	AIC	39204.629
Time:	01:14:19	BIC	39278.658
Sample:	01-01-1990	HQIC	39229.450
	- 03-12-2023		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
x1	0.0011	0.001	1.148	0.251	-0.001	0.003
ar.L1	1.0533	0.009	121.723	0.000	1.036	1.070
ar.L2	-0.4582	0.013	-34.645	0.000	-0.484	-0.432
ar.L3	0.3998	0.008	47.697	0.000	0.383	0.416
ma.L1	-0.0530	0.008	-6.837	0.000	-0.068	-0.038
ma.L2	0.8200	0.007	110.031	0.000	0.805	0.835
ma.L3	0.1754	0.010	18.259	0.000	0.157	0.194
ma.L4	0.3564	0.007	49.703	0.000	0.342	0.370
ma.L5	0.6597	0.007	91.946	0.000	0.646	0.674
sigma2	1.2358	0.014	89.238	0.000	1.209	1.263

Ljung-Box (L1) (Q):	2.62	Jarque-Bera (JB):	53.17
Prob(Q):	0.11	Prob(JB):	0.00
Heteroskedasticity (H):	0.97	Skew:	-0.02
Prob(H) (two-sided):	0.34	Kurtosis:	3.32

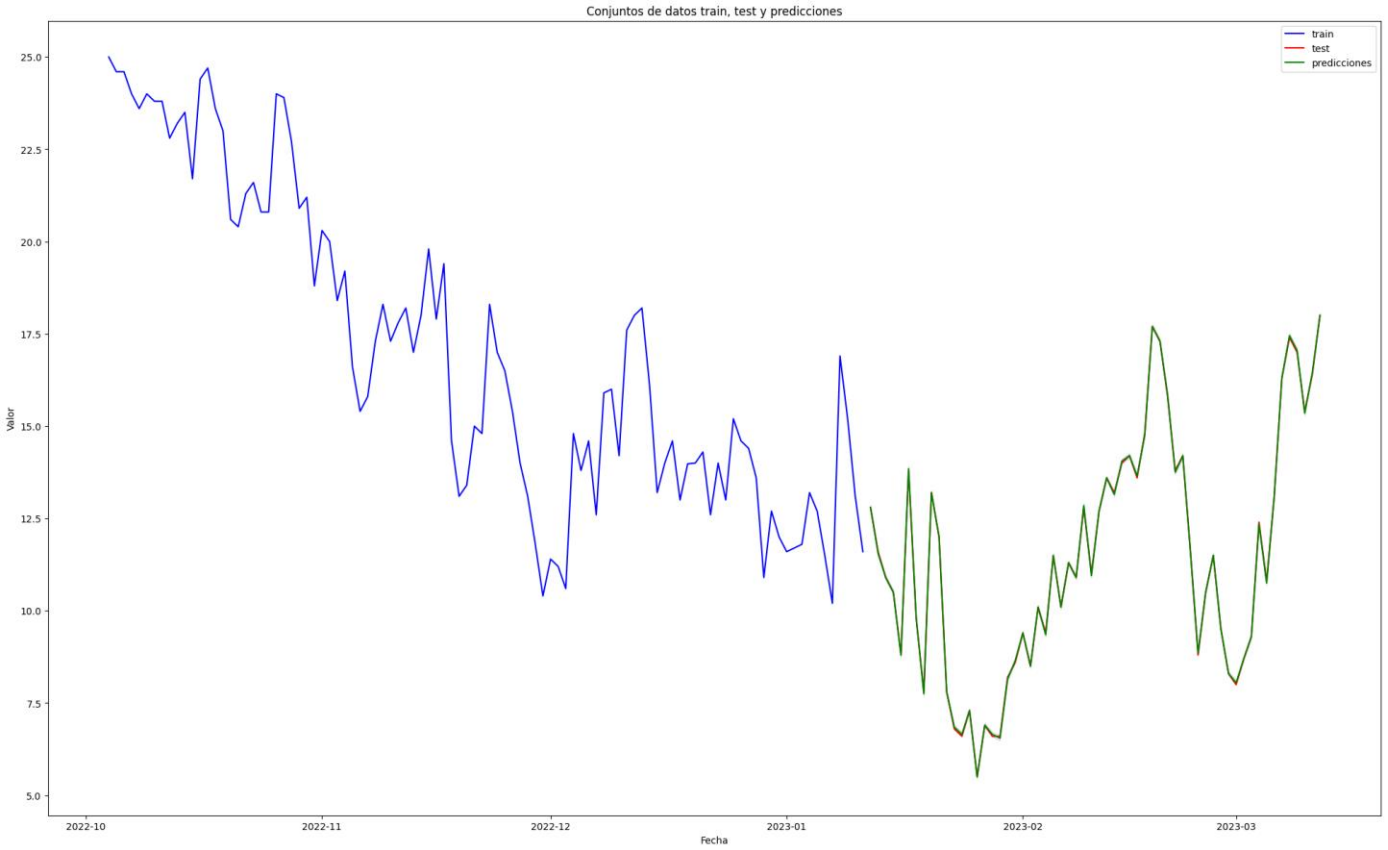
STL Configuration

Period:	7	Trend Length:	15
Seasonal:	7	Trend deg:	1
Seasonal deg:	1	Trend jump:	1
Seasonal jump:	1	Low pass:	9
Robust:	False	Low pass deg:	1

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

MSE: 23.303371307245435

PRUEBA: SARIMAX



SARIMAX Results

Dep. Variable:	tmed	No. Observations:	12124
Model:	SARIMAX(3, 0, 5)	Log Likelihood	24258.826
Date:	Sat, 18 Mar 2023	AIC	-48481.652
Time:	01:13:26	BIC	-48348.399
Sample:	01-01-1990	HQIC	-48436.975
	- 03-12-2023		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
tmin	0.5002	0.000	3416.198	0.000	0.500	0.500
tmax	0.4998	0.000	3773.335	0.000	0.500	0.500
presMin	0.0002	0.000	1.216	0.224	-0.000	0.001
presMax	-0.0002	0.000	-1.189	0.234	-0.001	0.000
dir	-1.703e-05	1.03e-05	-1.647	0.100	-3.73e-05	3.23e-06
velmedia	3.695e-05	0.000	0.116	0.908	-0.001	0.001
racha	-0.0001	0.000	-0.789	0.430	-0.000	0.000
sol	0.0001	0.000	1.030	0.303	-0.000	0.000
prec	-2.569e-06	6.98e-05	-0.037	0.971	-0.000	0.000
ar.L1	-1.4255	0.515	-2.767	0.006	-2.435	-0.416
ar.L2	-1.1529	0.608	-1.897	0.058	-2.344	0.038
ar.L3	-0.2172	0.470	-0.462	0.644	-1.139	0.705
ma.L1	1.4340	0.515	2.784	0.005	0.424	2.444
ma.L2	1.1674	0.612	1.908	0.056	-0.032	2.366
ma.L3	0.2416	0.475	0.509	0.611	-0.690	1.173
ma.L4	0.0306	0.016	1.876	0.061	-0.001	0.063
ma.L5	0.0196	0.010	1.965	0.049	4.68e-05	0.039
sigma2	0.0011	6.6e-06	162.271	0.000	0.001	0.001

Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	94320.52
Prob(Q):	1.00	Prob(JB):	0.00
Heteroskedasticity (H):	1.91	Skew:	-0.71
Prob(H) (two-sided):	0.00	Kurtosis:	16.59

MSE: 0.0010875842794987886

CONCLUSIONES

- Estamos en uno de los peores momentos del año para ponernos a prever temperaturas.
- El modelo STLForecast mejora mucho en rendimiento a ARIMA en este campo, ya que es más óptimo frente a estacionalidades.
- El modelo SARIMAX provoca un overfitting brutal. Además no tiene sentido si tampoco sabemos a priori las condiciones exógenas.
- Al desestimar los condicionantes exógenos, abrimos la puerta a otros modelos de regresión / Deep Learning.
- Sigo sin saber si va a llover en Semana Santa.

GRACIAS POR VUESTRA ATENCIÓN

Juan Antonio de la Cuadra
The Bridge School
DS/PT/Sep-2022