



APRENDIZAJE Y PREDICCIÓN DE LA CALIDAD DEL AIRE A PARTIR DE DATOS ABIERTOS

TFG

Autor: Sergio Sánchez Vallés
Tutor: Holger Billhardt
Curso: 2021/2022

Autor



Sergio Sánchez Vallés



Ingeniería del Software (URJC)

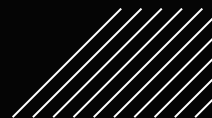


TABLA DE CONTENIDOS



01

Introducción

02

Objetivos

03

Descripción
informática

04

Experimentos y
validación

05

Conclusiones

06

Trabajos futuros





01



Introducción

La contaminación del aire



La contaminación del aire mata a...

+7.000.000

de personas al año



Precedentes



+ HABITANTES



PROBLEMA
COMÚN

+ TRÁFICO RODADO
+ INDUSTRIA



+ NO₂
+ PM_{2.5}

¿Se puede predecir la calidad del aire?

Imágenes
satelitales



Sensores de
tráfico



Estaciones de
calidad del aire



Estaciones
climatológicas





02



Objetivos

Planteamiento y metodología



Planteamiento del problema



Imagen: REUTERS

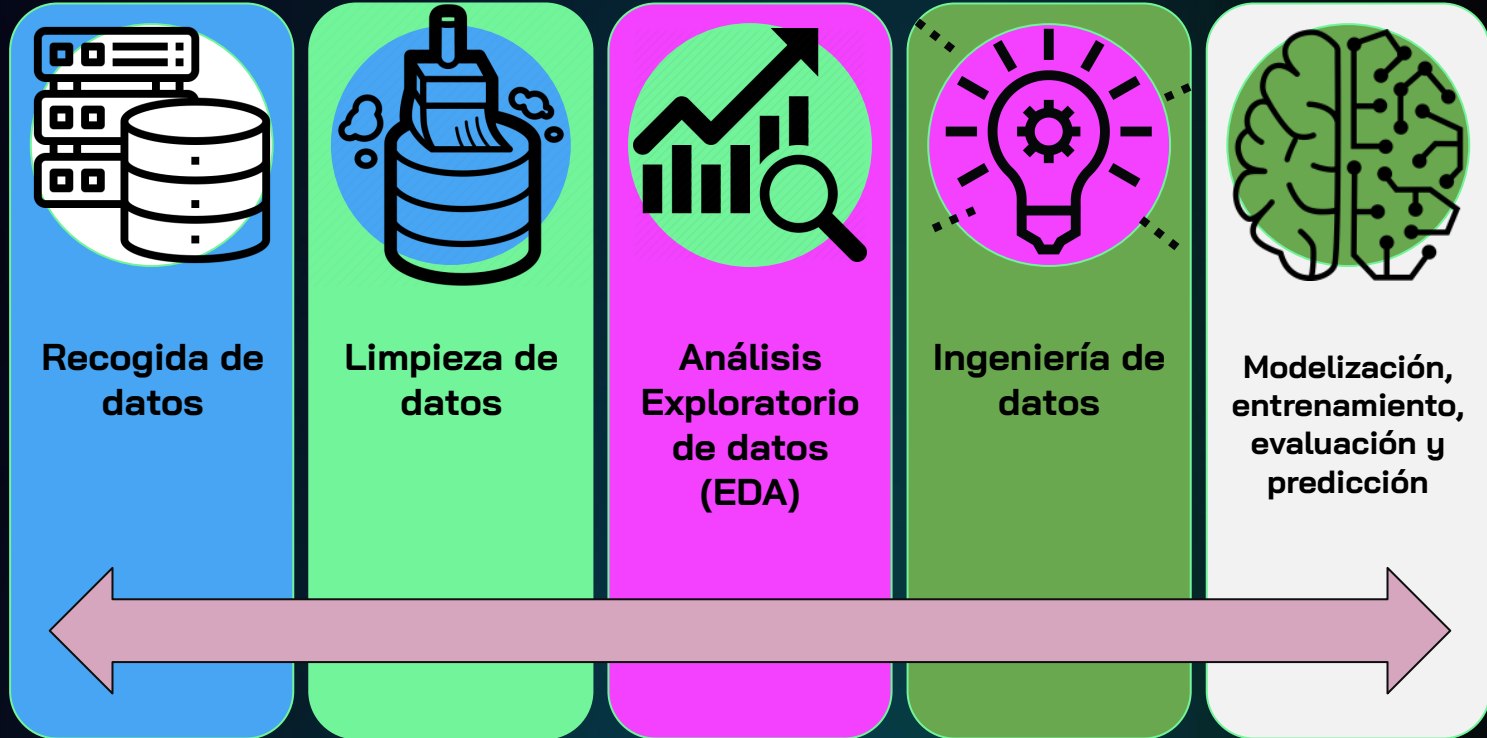


Usamos la ciencia de datos porque queremos responder algunas preguntas que ayudarán a nuestra organización, nuestro entorno y nuestro planeta.



¿CÓMO Y POR QUÉ?

Metodología





▶▶▶ 03 ●

Descripción informática

Tecnologías utilizadas, recogida de datos, limpieza de datos, análisis exploratorio, ingeniería de características, modelización, entrenamientos, evaluaciones y predicciones.

Tecnologías utilizadas





Recogida de datos

- Calidad del aire (API REST del Portal de Datos Abiertos del Ayuntamiento de Madrid)

- Formato: CSV
- Rango de tiempo: [2001, 2020] diarios
- Estaciones: 24
- Contaminantes: 14

```
PROVINCIA;MUNICIPIO;ESTACION;MAGNITUD;PUNTO_MUESTREO;ANO;MES;D01;V01;D02;V02;D03;V03;D04;  
28;079;4;1;28079004_1_38;2001;01;00017;V;00015;V;00015;V;00015;V;00016;V;00020;V;00025;V;  
28;079;4;1;28079004_1_38;2001;02;00040;V;00042;V;00038;V;00029;V;00018;V;00023;V;00016;V;  
28;079;4;1;28079004_1_38;2001;03;00019;V;00018;V;00013;V;00011;V;00015;V;00023;V;00015;V;  
28;079;4;1;28079004_1_38;2001;04;00014;V;00017;V;00013;V;00010;V;00009;V;00017;V;00011;V;  
28;079;4;1;28079004_1_38;2001;05;00010;V;00011;V;00016;V;00018;V;00015;V;00013;V;00013;V;  
28;079;4;1;28079004_1_38;2001;06;00013;V;00010;V;00012;V;00011;V;00012;V;00011;V;00012;V;  
28;079;4;1;28079004_1_38;2001;07;00014;V;00013;V;00012;V;00009;V;00007;V;00007;V;00009;V;  
28;079;4;1;28079004_1_38;2001;08;00007;V;00006;V;00007;V;00006;V;00006;V;00010;V;00009;V;  
28;079;4;1;28079004_1_38;2001;09;00006;V;00007;V;00007;V;00007;V;00007;V;00007;V;00007;V;
```



Recogida de datos

- Climatología (API REST Agencia Estatal de Meteorología)
 - Formato: CSV
 - Rango de tiempo: [2001, 2020] diarios
 - Estaciones: 13
 - Magnitudes: 15

```
fecha,indicativo,nombre,provincia,altitud,tmed,prec,tmin,horatmin,tmax,horatmax,velmedia,sol,presMax,horaPresMax,presMin,horaPres
2001-01-01,2462,PUERTO DE NAVACERRADA,MADRID,1894,"0,3","10,1","-1,0",23:30,"1,6",15:00,"7,5","0,0","806,3",00,"800,1",17,,,
2001-01-02,2462,PUERTO DE NAVACERRADA,MADRID,1894,"-1,2","2,1","-2,4",23:59,"0,0",14:15,"5,8","1,6","808,1",24,"800,3",03,,,
2001-01-03,2462,PUERTO DE NAVACERRADA,MADRID,1894,"-1,1","7,5","-2,8",02:00,"0,6",23:59,"4,2","0,0","809,9",11,"806,9",24,,,
2001-01-04,2462,PUERTO DE NAVACERRADA,MADRID,1894,"1,5","0,5","-0,6",08:00,"3,6",18:30,"7,5","5,9","811,1",12,"805,5",02,,,
2001-01-05,2462,PUERTO DE NAVACERRADA,MADRID,1894,"3,3","35,5","1,0",23:59,"5,6",13:00,"7,2","0,0","810,7",00,"800,3",23,,,
2001-01-06,2462,PUERTO DE NAVACERRADA,MADRID,1894,"-1,4","5,1","-4,4",23:59,"1,6",03:00,"1,9","0,0","805,1",24,"798,7",Varias,,,
2001-01-07,2462,PUERTO DE NAVACERRADA,MADRID,1894,"-4,9","0,0","-5,4",Varias,"-4,4",00:00,"1,9","0,0","810,5",24,"805,1",00,,,
2001-01-08,2462,PUERTO DE NAVACERRADA,MADRID,1894,"-3,6","0,0","-5,6",Varias,"-1,6",14:00,"1,7","6,3","811,8",12,"808,5",24,,,
2001-01-09,2462,PUERTO DE NAVACERRADA,MADRID,1894,"-2,1","0,5","-4,0",12:00,"-0,2",18:00,"8,1","0,0","808,5",00,"802,1",13,,,
```


Limpieza de datos



- Valores nulos
- Formato de fechas
- Pivotar filas y columnas
- Transformación de magnitudes



Magnitud		Abreviatura o fórmula	Unidad medida	Técnica de medida	
01	Dióxido de Azufre	SO ₂	µg/m ³	38	Fluorescencia ultravioleta
06	Monóxido de Carbono	CO	mg/m ³	48	Absorción infrarroja
07	Monóxido de Nitrógeno	NO	µg/m ³	08	Quimioluminiscencia
08	Dióxido de Nitrógeno	NO ₂	µg/m ³	08	Id.
09	Partículas < 2.5 µm	PM2.5	µg/m ³	47	Microbalanza
10	Partículas < 10 µm	PM10	µg/m ³	47	Id.
12	Oxidos de Nitrógeno	NOx	µg/m ³	08	Quimioluminiscencia
14	Ozono	O ₃	µg/m ³	06	Absorción ultravioleta
20	Tolueno	TOL	µg/m ³	59	Cromatografía de gases
30	Benceno	BEN	µg/m ³	59	Id.
35	Etilbenceno	EBE	µg/m ³	59	Id.
37	Metaxileno	MXY	µg/m ³	59	Id.
38	Paraxileno	PXY	µg/m ³	59	Id.
39	Ortoxileno	OXY	µg/m ³	59	Id.
42	Hidrocarburos totales (hexano)	TCH	mg/m ³	02	Ionización de llama
43	Metano	CH ₄	mg/m ³	02	Id.
44	Hidrocarburos no metánicos (hexano)	NMHC	mg/m ³	02	Id.

Magnitudes medidas en la red de calidad del aire de Madrid

Limpieza de datos



date	station	SO ₂	CO	NO	NO ₂	PM _{2,5}	PM ₁₀	NO _x	O ₃	TOL	BEN	EBE	TCH	CH ₄	NMH C
2012-05-27	28079024	2	0.2	1	10	6	13	12	64	0.5	0.3	0.5	1.35	1.1	0.25
2012-05-27	28079027			4	24			30	65				1.23	1.09	0.14
2012-05-27	28079035	2	0.2	7	32			43	55						
2012-05-27	28079036	3	0.2	4	24		19	31							
2012-05-27	28079038	1		10	26	6	13	41		1.5	0.2	0.5			

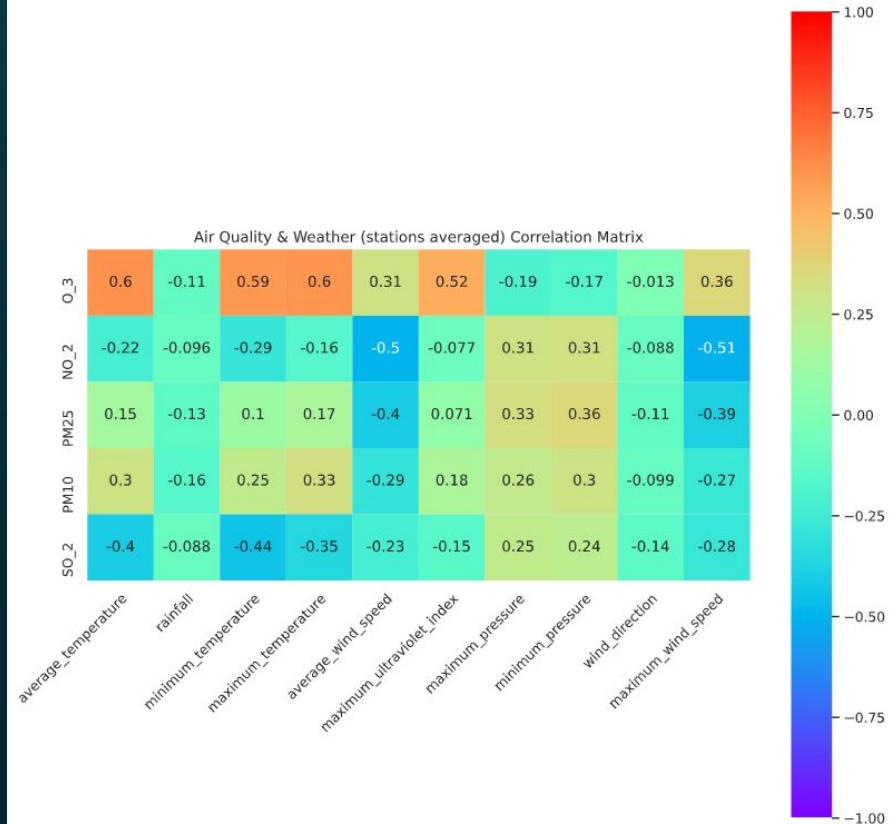
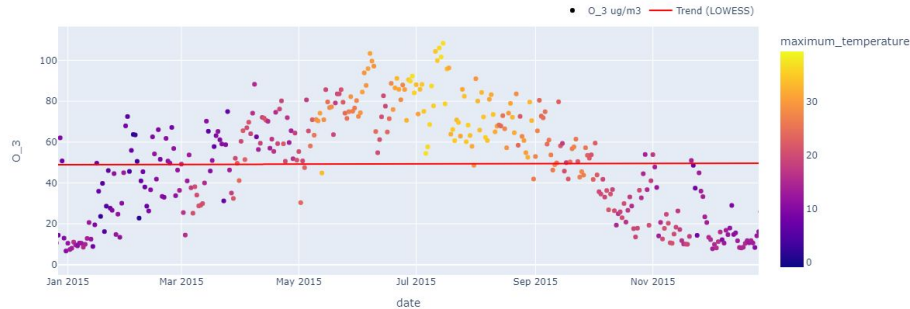
Análisis Exploratorio de Datos (EDA)

→ Correlación de los datos



Se puede apreciar una fuerte correlación entre las temperaturas e índice ultravioleta con el Ozono, las altas temperaturas y los rayos del sol propician reacciones químicas del O_3 con el resto de contaminantes.

Ozone and max. temperatures (2001-2020)



Matriz de correlación de la calidad del aire y la climatología

Análisis Exploratorio de Datos (EDA)

→ Correlación de los datos

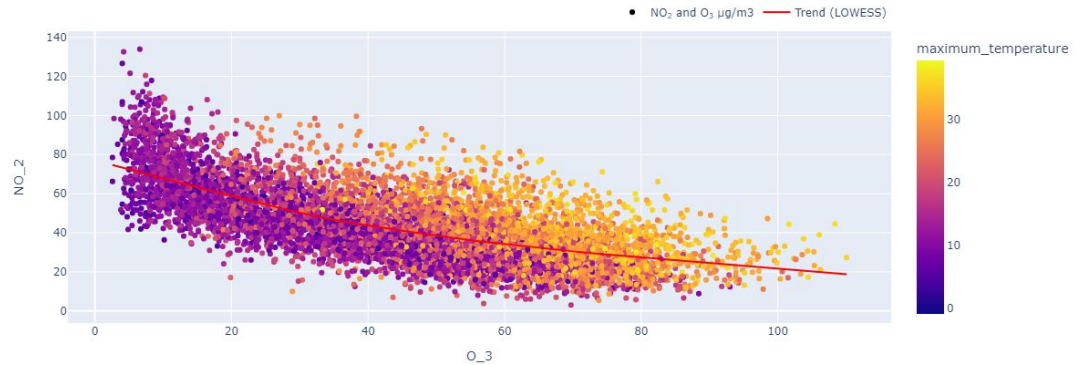


El proceso químico que hemos descrito anteriormente, la mezcla del ozono troposférico con los óxidos nitrosos, sumado a las altas temperaturas son la combinación ideal para desarrollar más contaminantes.

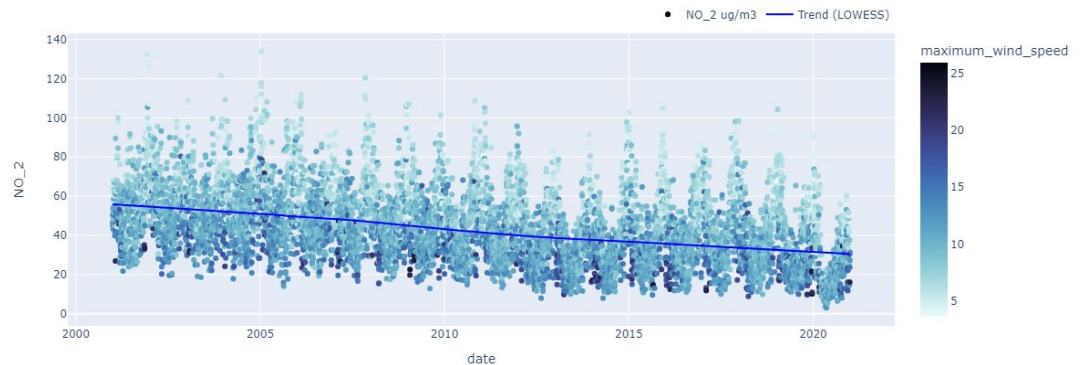
Existe una correlación entre las velocidades máximas de viento y el dióxido de nitrógeno, el viento toma parte en las capas del aire donde se mezcla el dióxido de nitrógeno, es por eso que palia parcialmente sus efectos.



NO₂ and O₃ with max. temperatures (2001-2020)



NO₂ and max. wind speed (2001-2020)



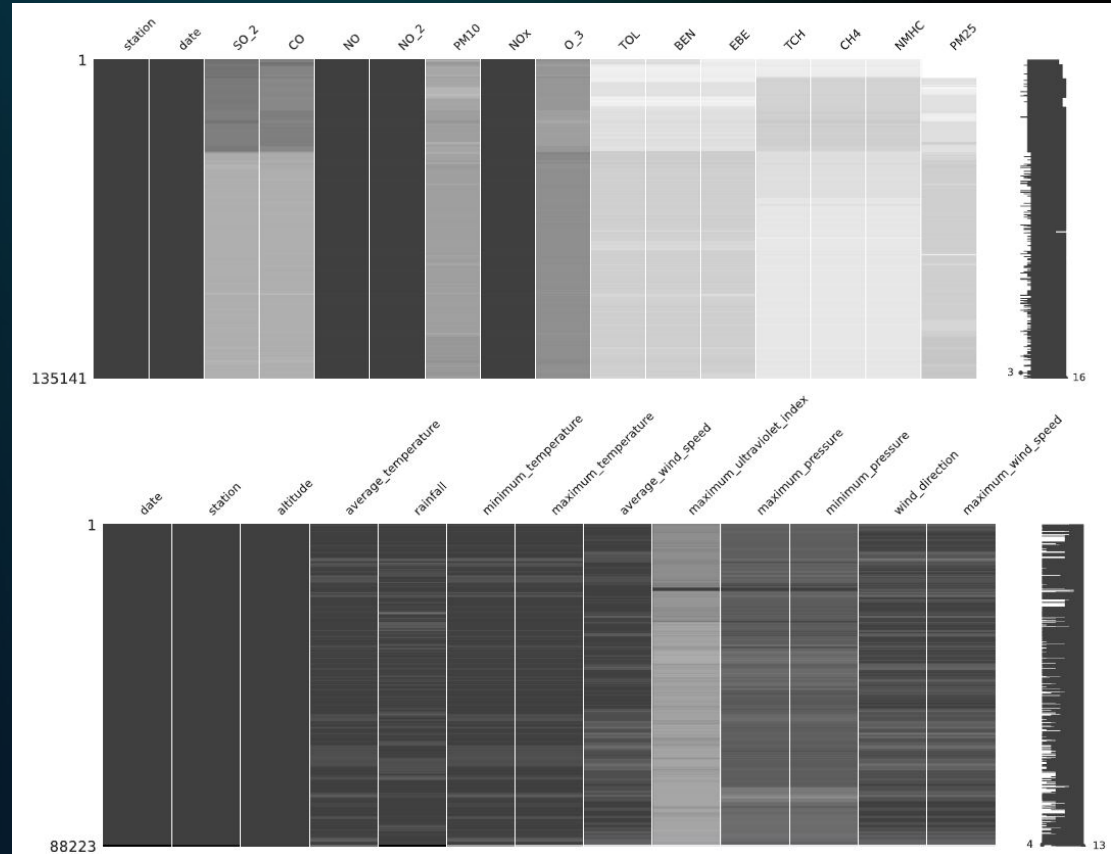
Análisis Exploratorio de Datos (EDA)

→ Valores nulos



Para todas las estaciones y para cada fecha comprobamos la falta de datos, los que están marcados en oscuro quiere decir que existe el dato para ese preciso instante, los huecos indican los valores nulos.

Los contaminantes de los que tenemos más datos son los óxidos de nitrógeno, el dióxido de azufre (SO₂), monóxido de carbono (CO), partículas finas de menos de 10µm (PM₁₀) y Ozono (O₃).



Análisis Exploratorio de Datos (EDA)

→ Índices de calidad del aire



Se hace necesaria una medida estandarizada para llevar un control de la contaminación.

Air Quality Index (AQI) de la Agencia de Protección Ambiental de los Estados Unidos de América [EPA], el European Air Quality Index de la Agencia Ambiental Europea [EEA] o el Índice Nacional de Calidad del Aire [ICA] del Ministerio para la Transición Ecológica y el Reto Demográfico español.

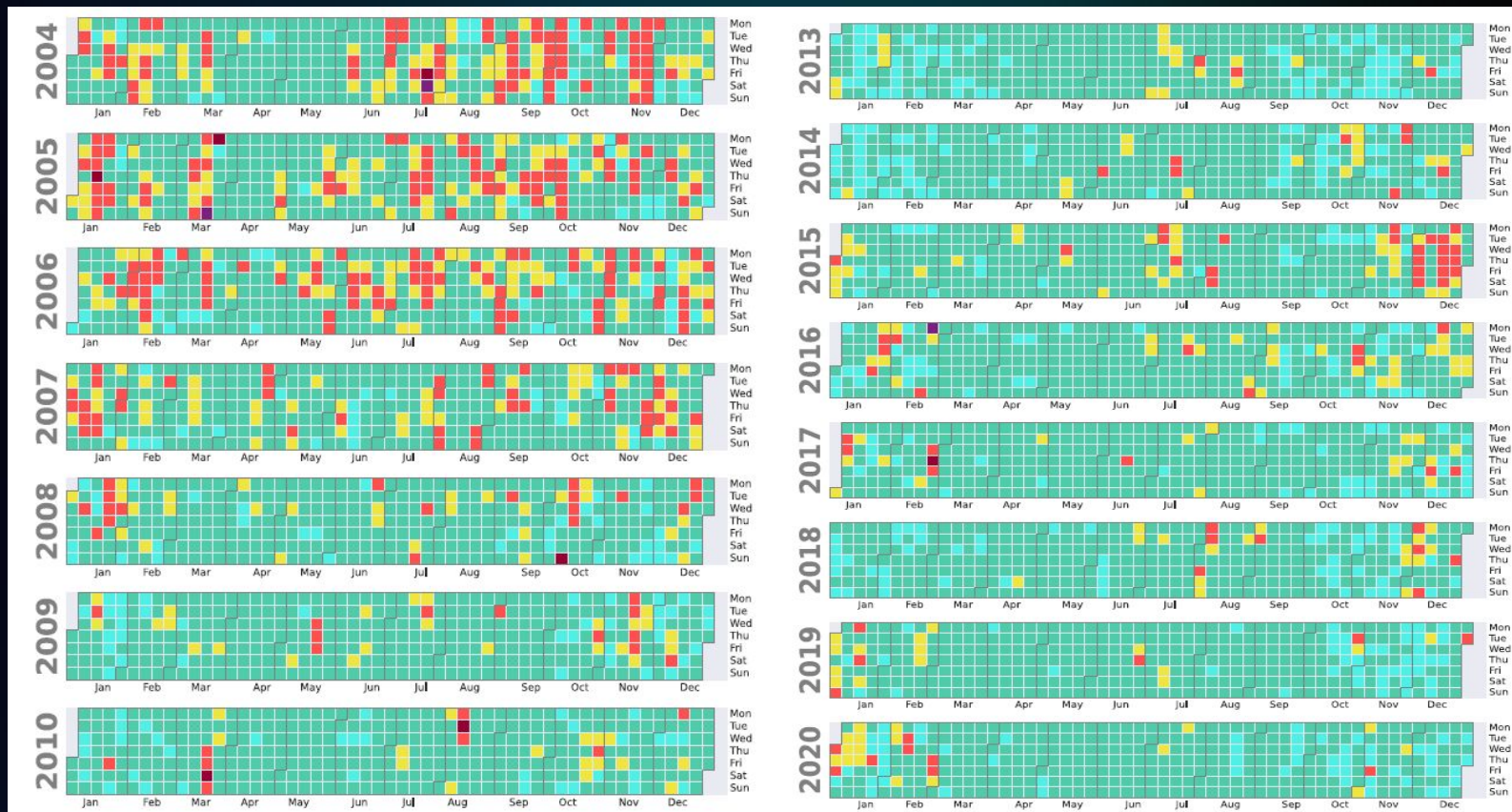


Pollutant	Index level (based on pollutant concentrations in $\mu\text{g}/\text{m}^3$)					
	Good	Fair	Moderate	Poor	Very poor	Extremely poor
Particles less than $2.5 \mu\text{m}$ ($\text{PM}_{2.5}$)	0-10	10-20	20-25	25-50	50-75	75-800
Particles less than $10 \mu\text{m}$ (PM_{10})	0-20	20-40	40-50	50-100	100-150	150-1200
Nitrogen dioxide (NO_2)	0-40	40-90	90-120	120-230	230-340	340-1000
Ozone (O_3)	0-50	50-100	100-130	130-240	240-380	380-800
Sulphur dioxide (SO_2)	0-100	100-200	200-350	350-500	500-750	750-1250

Índice de Calidad del Aire Europeo.

Análisis Exploratorio de Datos (EDA)

→ Índices de calidad del aire



Ingeniería de características (o de datos)



Mapa de las estaciones de calidad del aire (color rosa) y de climatología (color azul).



Los métodos de interpolación valorados han sido los siguientes: Kriging, **distancia inversa ponderada**, función de base radial (ej. gaussiana, spline) e interpolación polinómica.

$$u(\mathbf{x}) = \begin{cases} \frac{\sum_{i=1}^N w_i(\mathbf{x}) u_i}{\sum_{i=1}^N w_i(\mathbf{x})}, & \text{if } d(\mathbf{x}, \mathbf{x}_i) \neq 0 \text{ for all } i, \\ u_i, & \text{if } d(\mathbf{x}, \mathbf{x}_i) = 0 \text{ for some } i, \end{cases}$$
$$w_i(\mathbf{x}) = \frac{1}{d(\mathbf{x}, \mathbf{x}_i)^p}$$

Ingeniería de características (o de datos)

→ Formato final



```
station, date, SO_2, CO, NO, NO_2, PM25, PM10, NOx, O_3, TOL, BEN, EBE,  
TCH, CH4, NMHC, average_temperature, rainfall, minimum_temperature,  
maximum_temperature, wind_direction, average_wind_speed,  
maximum_wind_speed, maximum_ultraviolet_index, maximum_pressure,  
minimum_pressure
```



Modelización, entrenamientos, evaluaciones y predicciones

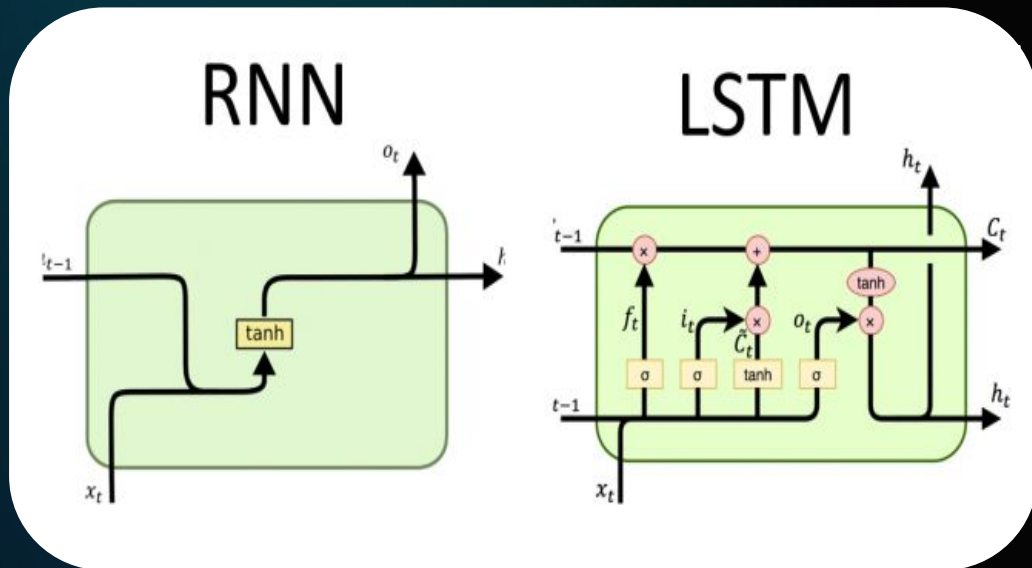


Principales ventajas para RNN:

- Aprende patrones complejos en datos de series temporales (no lineales)
- Aprende la dependencia temporal presente en los datos

Desventajas RNN / Mejoras LSTM:

- Gradiente se desvanece/explota
- Memoria débil
- Alto coste computacional



Recurrent Neural Network and it's variants. Shujaat Hasan. (2020) Medium.

Modelización, entrenamientos, evaluaciones y predicciones

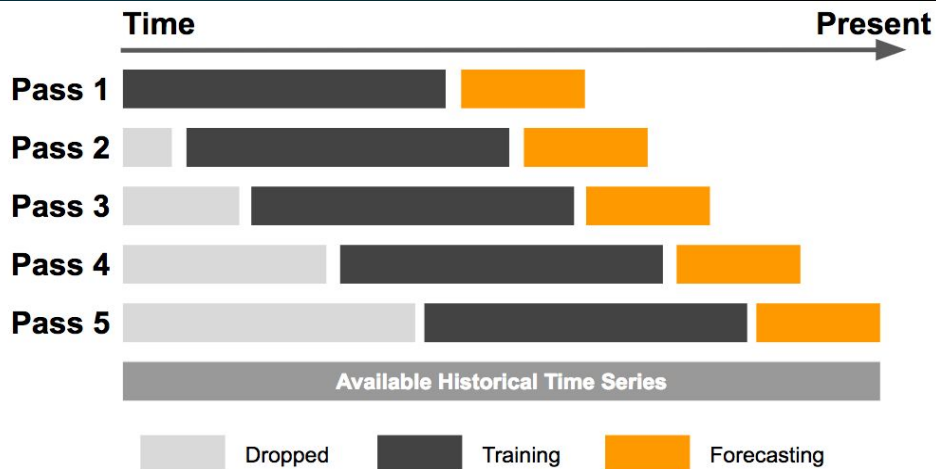
- Datos de **entrenamiento**: [2001, 2015]
- Datos de **test**: [2016, 2019]
 - 2020 ha sido eliminado de los datos por la pandemia del COVID-19

- ❑ **Normalización** de los datos.
- ❑ **One-hot encoding** para las estaciones.
- ❑ Generar datos con el **método de ventana deslizante**, 14 días de datos anteriores, 1 día siguiente a predecir.

id	color
1	red
2	blue
3	green
4	blue

One Hot Encoding

id	color_red	color_blue	color_green
1	1	0	0
2	0	1	0
3	0	0	1
4	0	1	0



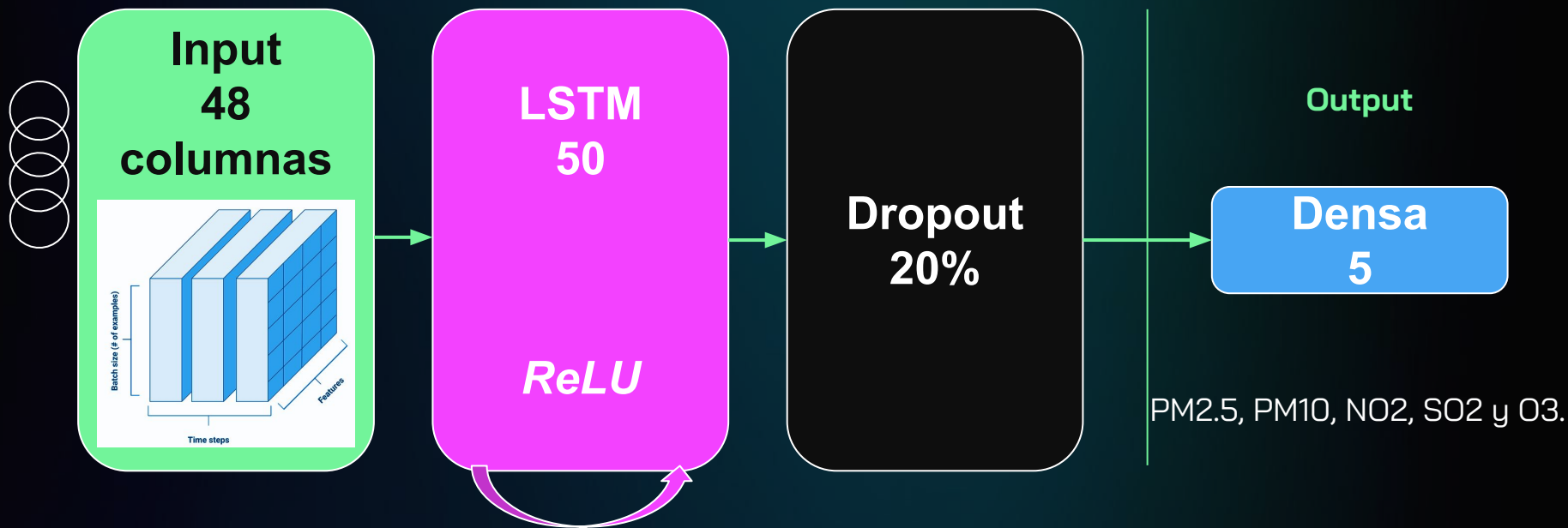


▶▶▶▶ 04 ●

Experimentos y validación

Arquitectura del modelo, capas y optimización
de hiperparámetros

Arquitectura del modelo



Optimizador: **adam**

Función de pérdida: **mae**

Batch size: **32**

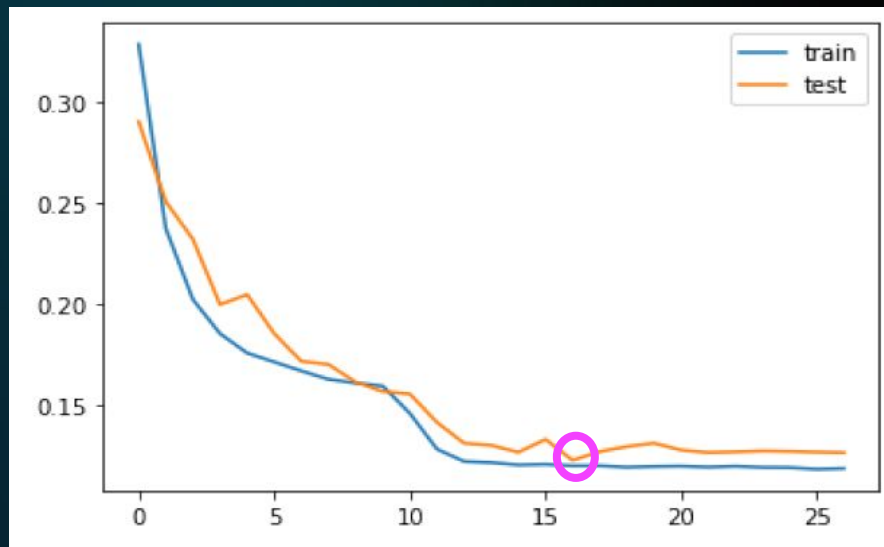
Total de 1.110 parámetros entrenables

Entrenamiento del modelo

Early stopping: **10 épocas**

El **Early Stopping** es una forma de regularización, con la que se decide en qué época se debe dejar de entrenar el modelo para evitar el **overfitting**.

Cuando el modelo **solo mejora en los datos de entrenamiento pero no en los de test**, esto significa que está memorizando en vez de generalizar.



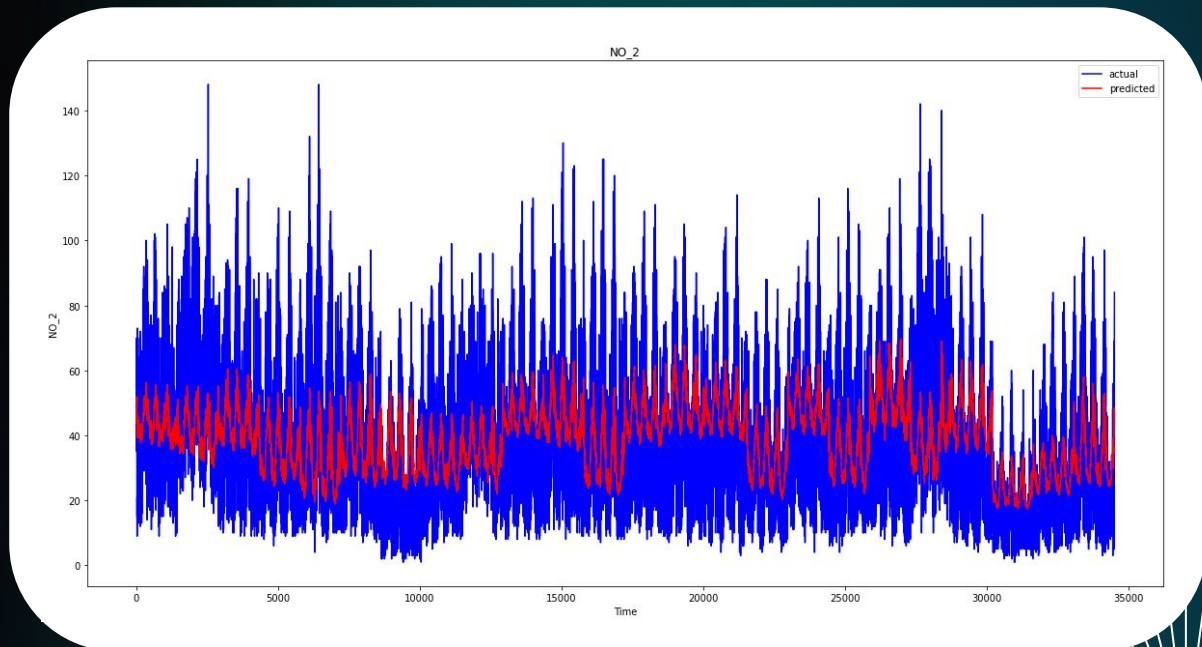
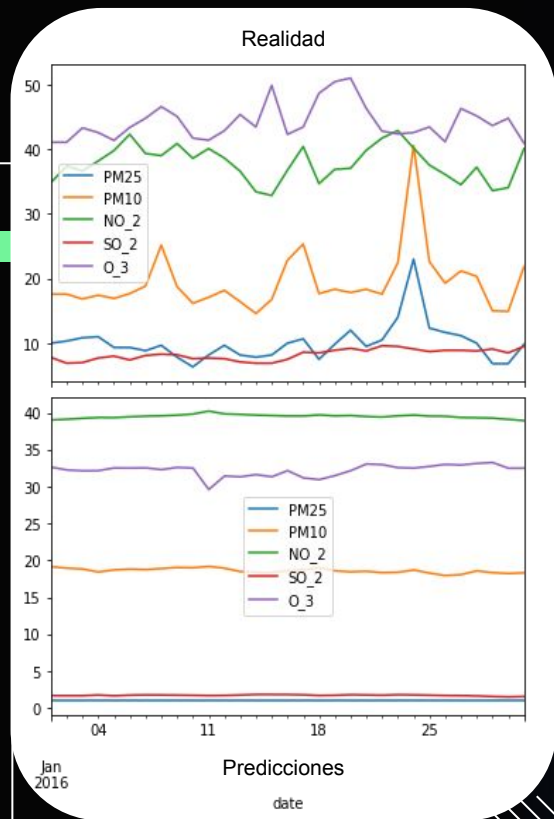
Resultados del modelo

→ Errores medios absolutos

- $PM_{2.5}$: **9.06** [0-800]
- PM_{10} : **8.48** [0-1200]
- NO_2 : **13.74** [0-1000]
- SO_2 : **5.98** [0-800]
- O_3 : **21.17** [0-1250]
- **Media total:** **11.68**

Pollutant	Index level (based on pollutant concentrations in $\mu\text{g}/\text{m}^3$)					
	Good	Fair	Moderate	Poor	Very poor	Extremely poor
Particles less than $2.5 \mu\text{m}$ ($PM_{2.5}$)	0-10	10-20	20-25	25-50	50-75	75-800
Particles less than $10 \mu\text{m}$ (PM_{10})	0-20	20-40	40-50	50-100	100-150	150-1200
Nitrogen dioxide (NO_2)	0-40	40-90	90-120	120-230	230-340	340-1000
Ozone (O_3)	0-50	50-100	100-130	130-240	240-380	380-800
Sulphur dioxide (SO_2)	0-100	100-200	200-350	350-500	500-750	750-1250

Resultados del modelo





05



Conclusiones



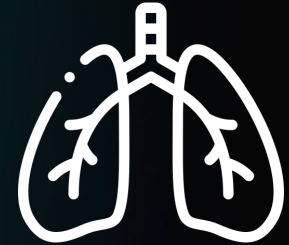
Consistencia de los datos

Predecir la contaminación
con solo 14 días de datos



Establecer buenas bases

Utilidad en la prevención de
problemas de salud



Conocimientos adquiridos



▶▶▶▶ 06 ●

Trabajos futuros

Capas

Unidades

GRU

Dropout

Early Stopping

ReLU

Bi-Directional



Sets de test

Batch size

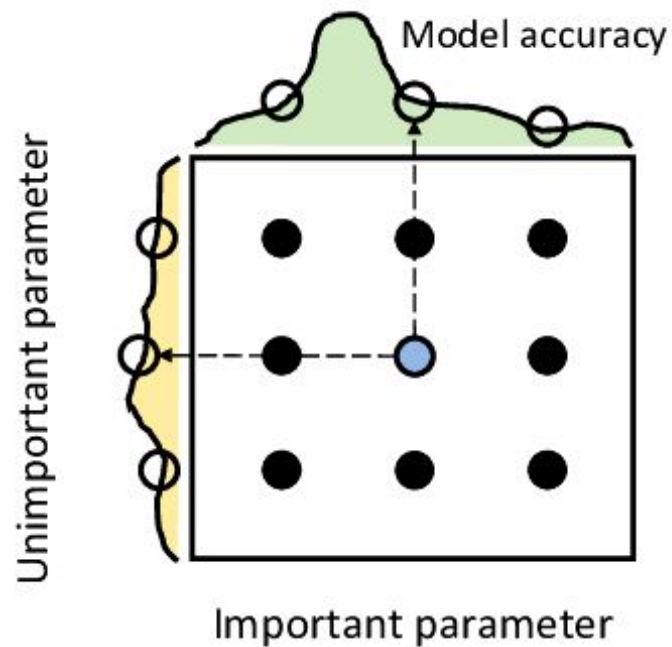
Learning rate

Inicialización de pesos

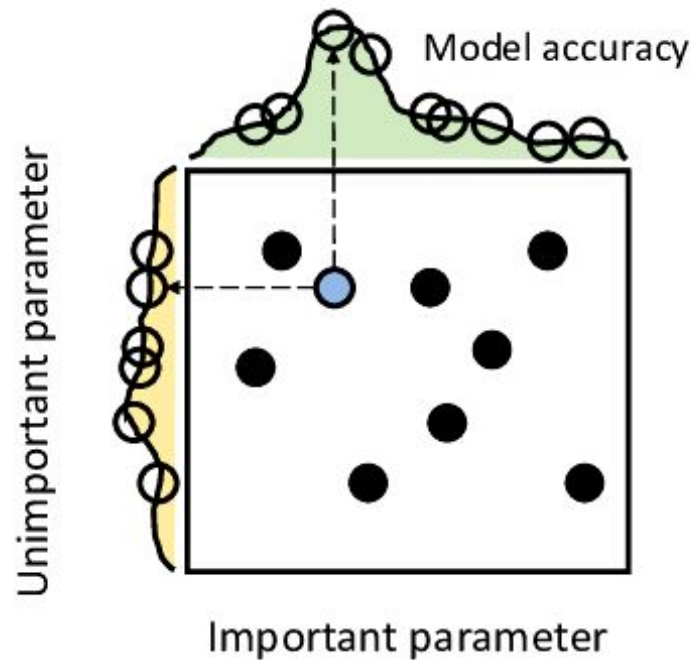


Tuner

Grid Search



Random Search



Otros modelos de
machine learning



Más optimización de
hiperparámetros



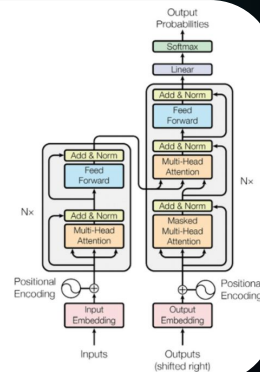
Transformer
(Estado del arte)

XGBoost

Auto-Sklearn



Transformer
Attention Is All You Need





Muchas gracias



[sergio-sanchez-valles](#)



[SergioSV96](#)

