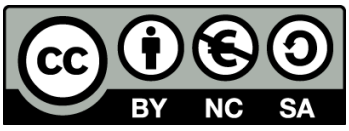
Calidad del aire en Valencia

Mikel Navarro - Tono Villarejo - Santiago Cámara



Mikel Navarro - Tono Villarejo - Santiago Cámara
Curso de especialización
Memoria del Proyecto de Inteligencia Artificial y Big Data
IES Abastos. Curso 2021/22. Grupo 8IA. 9 de junio de 2022
Tutores: Boris Anaya – José Clemente – Elena Tejadillos

ÍNDICE

Tabla de cuadros e imágenes

1 IDENTIFICACIÓN Y OBJETIVOS DEL PROYECTO

1.1 Presentación

Se pretende realizar un estudio de la calidad del aire en la ciudad de Valencia.

Para ello se elaborará una base de datos histórica con la mayor cantidad de información posible que se complementará con información de tipo meteorológico que, a priori, dará un valor añadido a la información disponible sobre calidad del aire. Dicha información podrá visualizarse por diversos criterios.

Por otra parte, se hará un estudio especial de los años 2020 y 2021 en el sentido de comprobar la evolución de los distintos indicadores a pesar del parón que supuso el confinamiento y la pandemia.

Además, se dispondrá de un módulo de predicciones que permitirá predecir indicadores de calidad del aire a futuro.

Todo lo anterior se podrá visualizar en una página web creada para ello.

Se hará una web de datos abiertos para poner a disposición del público los dataset elaborados en el proyecto.

En este proyecto se utilizarán diversas técnicas de Big Data e Inteligencia Artificial para conseguir los objetivos propuestos.

1.2 Indicadores de calidad del aire

La evaluación de la calidad del aire se define como el resultado de aplicar cualquier método que permita medir, calcular, predecir o estimar las concentraciones de un contaminante en el aire ambiente o su depósito en superficies en un momento determinado.

La información disponible está agrupada por la estación donde se recogen los distintos parámetros.

Se dispone de información de los siguientes indicadores:

- PM1 (materia particulada 1): Partículas menores de un (1) micrómetro (µm)
- PM2.5 (materia particulada 2,5): Partículas menores de dos y medio (2,5) micrómetros (µm)
- PM10 (materia particulada 10): Partículas menores de diez (10) micrómetros (µm)
- NO: óxido de nitrógeno(II), óxido nítrico o monóxido de nitrógeno
- NO2: dióxido de nitrógeno u óxido de nitrógeno (IV)
- NOx: Óxidos de nitrógeno
- O3: Ozono
- SO2: dióxido de azufre u óxido de azufre (IV)
- CO: monóxido de carbono u óxido de carbono (II)
- NH3: amoniaco, azano, espíritu de Hartshorn, trihidruro de nitrógeno o gas de amonio
- C7H8: tolueno o Tolveno
- C6H6: Benzol o benceno

- C8H10: xileno, xilol o dimetilbenceno
- Ruido
- Velocidad viento
- Dirección del viento
- Temperatura
- Humedad relativa
- Presión
- Radiación Solar
- Precipitación
- Velocidad Máxima del viento

1.3 Estaciones de Valencia

El ayuntamiento de Valencia dispone de las siguientes estaciones dónde se recoge información:

- Avda. Francia
- Bulevar Sud
- Molí del Sol
- Pista Silla
- Politécnico
- Viveros
- Centro
- Consellería Meteo
- Nazaret Meteo
- Puerto Valencia

2 DISEÑO DEL PROYECTO

2.1 Modelado del proyecto de calidad del aire

Contexto	Información histórica disponible sobre indicadores de calidad del aire y meteorológicos que se van actualizando con la información proporcionada en tiempo real por las estaciones de medición.
Big Data	El sistema se alimentará de la información histórica para entrenar los modelos e irá acumulando las nuevas lecturas diarias para hacer predicciones futuras. Así mismo, estas lecturas podrían utilizarse en el futuro para reentrenar los modelos.
Sistema Cognitivo	Se crearán distintos modelos de aprendizaje automático (por combinaciones de estación e indicador) con algoritmos de series temporales. Se trata de aprendizaje supervisado.
Decisión	En base a los datos de la ventana de entrada los modelos realizarán predicciones de los distintos parámetros de calidad del aire. Mediante una web se podrá visualizar tanto la información histórica como las predicciones realizadas.

Cuadro 1 - Modelado del proyecto

2.2 Herramientas y soluciones tecnológicas

Para la parte de Big Data se utilizan las siguientes herramientas y tecnologías:

- **Elasticsearch**: es un motor de búsqueda y analítica distribuido, gratuito y abierto para todos los tipos de datos, incluidos textuales, numéricos, geoespaciales, estructurados y no estructurados.
- **Logstash**: es una herramienta para la administración de logs. Esta herramienta se puede utilizar para recolectar, analizar y guardar los logs para futuras búsquedas.
- **Kibana**: es una aplicación de frontend gratuita y abierta que proporciona capacidades de visualización de datos y de búsqueda para los datos indexados en Elasticsearch.

Para los modelos de Inteligencia Artificial se utiliza:

- Scripts desarrollados en lenguaje **Python**: Python es un lenguaje de alto nivel de programación interpretado cuya filosofía hace hincapié en la legibilidad de su código, se utiliza para desarrollar aplicaciones de todo tipo y es muy utilizado en proyectos de Inteligencia Artificial.
- **Tensorflow**: es una biblioteca de código abierto para aprendizaje automático a través de un rango de tareas, desarrollada por Google para satisfacer sus necesidades de sistemas capaces de construir y entrenar redes neuronales para detectar y descifrar patrones y correlaciones, análogos al aprendizaje y razonamiento usados por los humanos.

- Diversas librerías complementarias para python: pandas, numpy, mathplotlib, dataprep, datetime, etc. para resto de tareas relacionadas con los scripts necesarios en EDA, ETL y utilidades de apoyo.

2.3 Necesidades de hardware y software

2.3.1 Hardware

Se utiliza el servidor Linux del IES Abastos para almacenar la base de datos de elasticsearch y ejecutar los distintos scripts. En este servidor se ejecuta Ubuntu 20.04.4 LTS.

Para el portal de acceso a los conjuntos de datos procesados en el proyecto y para el acceso a visualizaciones se han utilizado dos Virtual Private Servers (VPS) de la compañía OVH con las siguientes características:

Starter:	Basic:
• 1 vcore • 2 GB RAM • 20 GB SDD	• 1 vcore • 2 GB RAM • 40 GB SDD

Cuadro 2 - Servidores VPS

Durante el desarrollo del proyecto también se han utilizado en modo local ordenadores del instituto y ordenadores particulares de los miembros del equipo.

2.3.2 Software

Se ha realizado la instalación en el servidor de las siguientes herramientas/tecnologías:

- Elasticsearch versión 7.17.2
- Kibana versión 7.17.2
- Logstash versión 7.17.2
- Python versión 3.8.10
- Apache Nifi 1.16.2

2.4 Flujo de trabajo – Procesamiento de datos

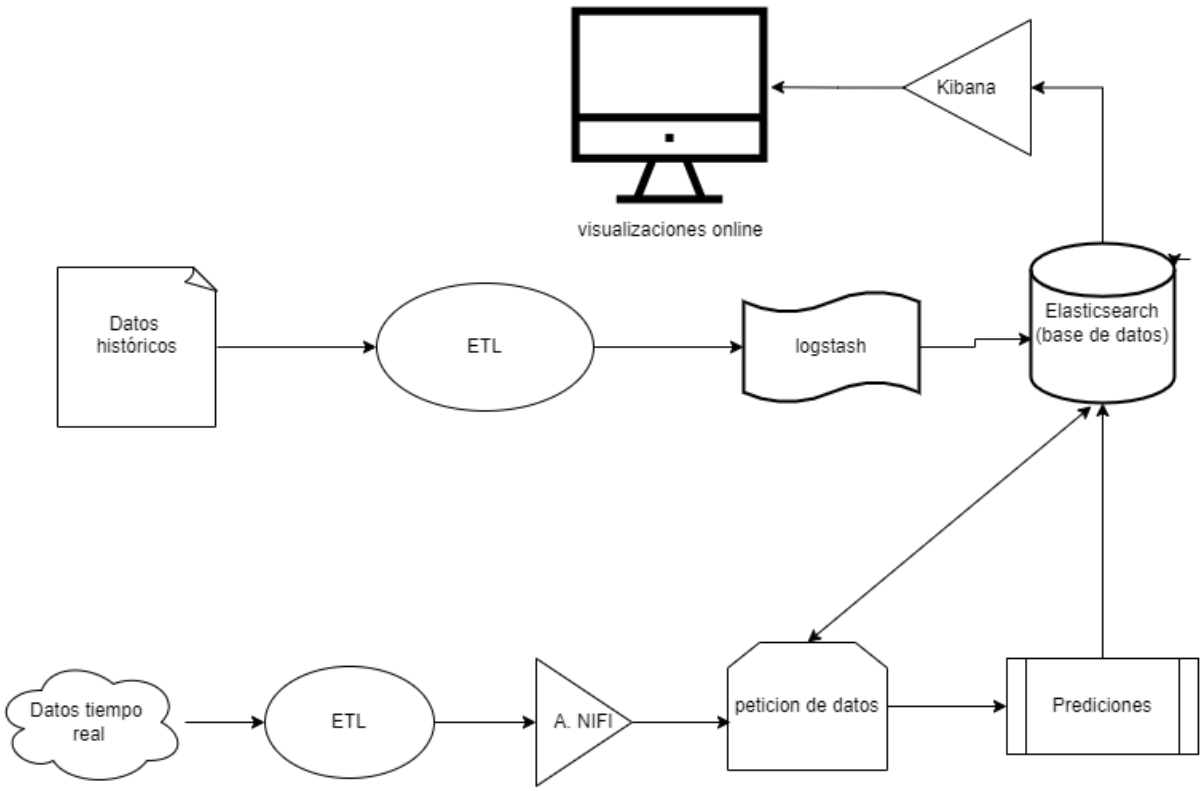


Imagen 3 - Flujo de trabajo

3 DESARROLLO DEL PROYECTO

3.1 Ingesta de datos

3.1.1 Fuentes de datos

En este apartado se abrieron varias líneas de investigación, búsqueda y localización de fuentes que nos permitieran confeccionar un conjunto de datos para el desarrollo del proyecto.

En relación a la calidad del aire se investigó la información aportada por:

Organismos oficiales:

- Datos de calidad del aire que proporciona el portal de datos abiertos de la Generalitat Valenciana (https://portaldadesobertes.gva.es/es)
- Datos de mediciones de calidad del aire de la Consellería de agricultura, desarrollo rural, emergencia climática y transición ecológica (https://agroambient.gva.es/)

Portales especializados:

- Portal de medición de calidad del aire en el mundo (fork) (https://waqi.info/es/)
- Portal de calidad del aire en el mundo (https://aqicn.org/map/world/es/)

En relación a los datos de climatología se investigó la información aportada por:

Organismos oficiales:

- Datos de mediciones meteorológicas proporcionados por la Agencia Estatal de Meteorología (http://www.aemet.es/)

Portales especializados:

- Datos de mediciones meteorológicas proporcionados por la Base de datos Meteorológica (https://datosclima.es/)
- Datos de mediciones meteorológicas proporcionados por la Associació valenciana de meteorología Josep Peinado (https://www.avamet.org/)

Como se puede ver las muestras de posibles datos eran múltiples y cabría esperar que cumplieran con suficiencia el objetivo del proyecto.

Hubo, a la par, un estudio de las APIs que, en algunos casos, acompañaban a las fuentes de datos. Se encontraron algunas, pero como se verá más adelante las estructuras de datos que ofrecían las APIs precisaban múltiples retoques y reformateos, con lo cual dado que era un trabajo ineludible se optó por la descarga directa de los datos que se ofrecían.

3.1.2 EDA (Exploratory Data Analysis)

Una vez detectadas y recopiladas las fuentes de datos se iniciaron las tareas relacionadas con la verificación de utilidad, descarte de aquella información que no permitía ningún tipo de

tratamiento y evaluación de las fuentes que se considerarían aptas para los siguientes procesos.

En principio se contaba con dos tipos de archivos en los conjuntos de datos referidos a la calidad del aire (históricos):

Archivos .csv con múltiples diferencias en los campos, aunque suficientes campos coincidentes para el estudio que podríamos aprovechar, con los datos de todas las estaciones de la Comunidad Valenciana.

Archivos .txt con información de cada una de las estaciones de la comunidad, de las cuales se podían filtrar para descarga las pertenecientes a la ciudad de valencia.

En ambos casos podía contarse con un rango de fechas que abarcaba desde 1994 hasta 2022.

En una primera fase se planificó la limpieza, maquetación y consolidación de los datos en formato txt (Consellería de Agricultura…), construyendo una batería de scripts que ayudaron a que se les diera un formato igual a todos los ficheros. Se eliminaron líneas de texto que no se necesitaban para el conjunto de datos y se ubicaron adecuadamente las columnas para posteriormente poderse convertir sin errores en dataframes y que se grabaran a otros formatos.

Esta primera fase presentó algunas inconsistencias respecto a la información de los valores de calidad que suministraba, así como grandes huecos temporales de información que no constaban.

Esto dio lugar a una nueva fase en la que se trabajó con los ficheros csv (datos abiertos GVA). En este caso se desarrollaron una batería de scripts para recorrer todos los ficheros, filtrar las estaciones correspondientes al objeto de estudio (ciudad de Valencia), verificación de la coherencia de datos en cuanto a las mediciones que mostraban datos y los tramos en que se presentaban huecos con falta de información.

Finalmente se optó por esta segunda fuente de datos (csv) para componer junto a los datos climáticos unos datasets definitivos a partir de los cuales pasar a la ETL.

En cuanto a los datos climáticos la elección de la fuente se basó en que era la única que proporcionaba históricos de la ciudad de Valencia desde el año 2013 de forma gratuita.

El hecho de que fuera gratuita suponía que se tuvieron que hacer varias peticiones a la web para que mostrara los resultados por pantalla y posteriormente copiarlos y trasladarlos a unas hojas de cálculo

Los resultados correspondían a cada una de los items de la medición por separado que era lo que proporcionaba dicha fuente y a partir de las hojas de cálculo montadas se crearon scripts para montar un conjunto de datos que albergara todos los items y todas las fechas disponibles para grabarlos en formato csv.

3.1.3 ETL (Extract Transform and Load)

Una vez completadas las tareas de EDA, el siguiente paso consistió en la creación de scripts que fusionaran los datos climáticos con los de calidad del aire.

Se tuvo que tener en cuenta que los conjuntos de datos de calidad de aire tenían ya algunos campos de algunas estaciones que aportaban lecturas de valores climáticos con lo cual, dado que los datasets climáticos eran de una medición diaria de cada uno de los valores climáticos, se optó por utilizar los datos climáticos únicamente para rellenar los huecos donde no hubiera una medición en una fecha y en una estación concretas del dato climático en cuestión. Lo ilustraremos con un ejemplo:

Si el dataset climático tenía todos los datos (velocidad del viento, precipitaciones…) en una fecha concreta (v.g. 7 de mayo de 2014) y la estación (v.g. Av. Francia) no disponía de lecturas de valores climáticos en esa fecha se cumplimentaban dichos datos con los valores del dataset climático. En caso contrario la prioridad la tenían las mediciones del dataset de calidad del aire.

Una vez consolidados los dos datasets se había obtenido un conjunto de datos que cubría cualitativamente el objeto de estudio, pero cuantitativamente presentaba importantes carencias debido a los porcentajes de ausencia de datos y a los valores nulos del dataset resultante.

Para contrarrestar esto se utilizó un mecanismo denominado interpolación para partir de unos datos conocidos y con limitación respecto al número de huecos o valores nulos seguidos que presentaba el dataset insertar datos basados en la mencionada interpolación.

Posteriormente se hizo necesario un estudio pormenorizado de la composición del dataset obtenido atendiendo a los porcentajes de nulos, que quedaban de las fases previas, las correlaciones entre las diferentes características del dataset, la continuidad de las fechas y el peso, en cuanto a la cantidad de registros y valores, de cada una de las estaciones.

Sobre estos se aplicaron varios mecanismos basados en el desarrollo de scripts específicos que sirvieron para sacar las conclusiones que se buscaban y tomar las decisiones específicas que permitía contextualizar y adaptar definitivamente el conjunto de datos al objetivo del proyecto.

Entre otras cosas estas conclusiones fueron:

Las relacionadas con los datos climáticos de los cuales se prescindió pues, aun entendiendo que son relevantes, respecto a la información recabada no aportaban ningún factor que pudiera resultar necesario para la creación y entrenamiento del modelo.

En cuanto a las estaciones se optó por utilizar aquellas que presentaban un equilibrio entre cantidad de mediciones y valores nulos para disponer de suficientes registros una vez descartados aquellos con valores nulos.

En cuanto al rango de fechas se verificó que las estaciones elegidas disponían de históricos cuyo rango empezaba en 1994 con lo que se podían establecer varios niveles de estudio basados en diferentes rangos de fechas y atendiendo a hechos relevantes y de importancia suficiente como para influir en el objeto del estudio (2008 gran crisis económica mundial, 2019 pandemia covid...).

Una vez dilucidado cual sería el conjunto de datos definitivo se inició un proceso de decisión basado en el tipo de algoritmos que se utilizarían para entrenar y crear el modelo. Esto dio lugar a una investigación que propuso varias vías todas relacionadas con el tratamiento de series temporales.

3.2 Big Data

3.2.1 Almacenamiento de datos

Para el almacenamiento de datos se ha optado por usar Elasticsearch que es una base de datos no sql basada en clave valor. Se crean creado 2 grupos de índices distintos:

- El primero es un histórico de todas las estaciones desde 1994 hasta 2022 con un mapeo de datos estricto, subido usando logstash y distribuidos: cada estación tiene su propio índice.
- En el segundo grupo están las predicciones que se almacenan usando Python sin un mapeo estricto y cada indicador de las estaciones tiene su propio índice.

3.2.2 Visualización de la información

Inicialmente se valoró la posibilidad de utilizar PowerBI y/o Tableau para crear visualizaciones más complejas, pero por la imposibilidad de conectar Tableau con elasticsearch y la gran dificultad de conectar PowerBi Desktop se descartaron, optando por Kibana.

Se ha usado Kibana como herramienta de representación visualizaciones por su perfecta integración con Elasticsearch. Se han creado seis index patterns (un index patterns es un enlace para que Kibana acceda a los datos de Elasticsearch):

- Uno para los datos históricos de todas las estaciones.
- Cuatro más para las predicciones de cada estación.
- Finalmente, uno que engloba las predicciones de todas las estaciones.

Se han creado tres Dashboards distintos:

- Datos históricos con todas las estaciones.

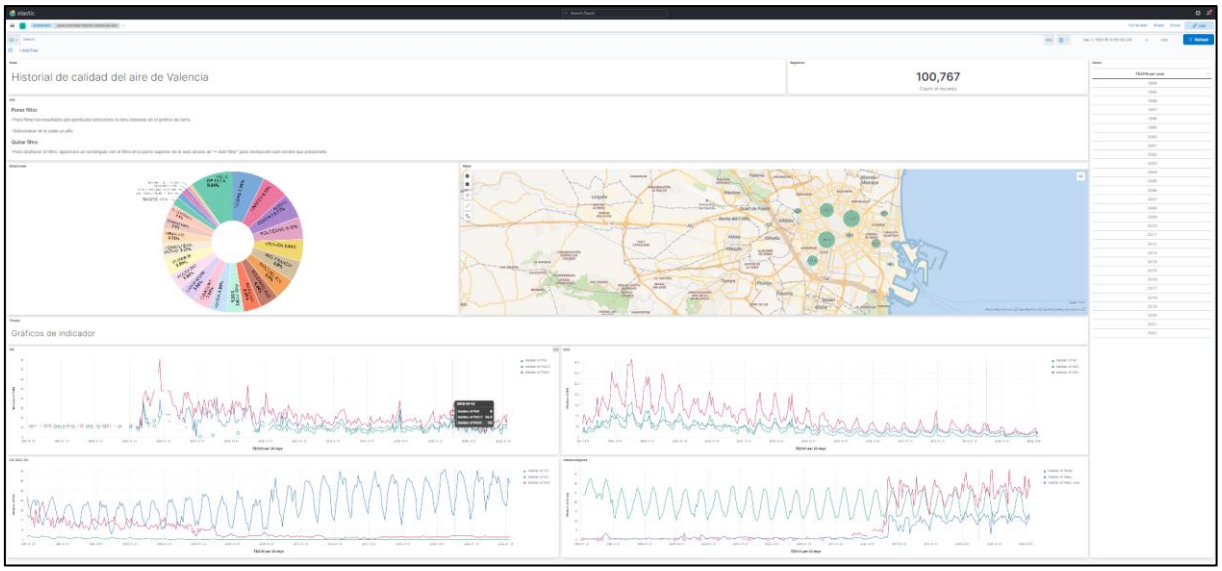


Imagen 4 - Visualización histórico de todas las estaciones

- Predicciones y datos reales de 2020-2021 (pandemia) por estaciones principales

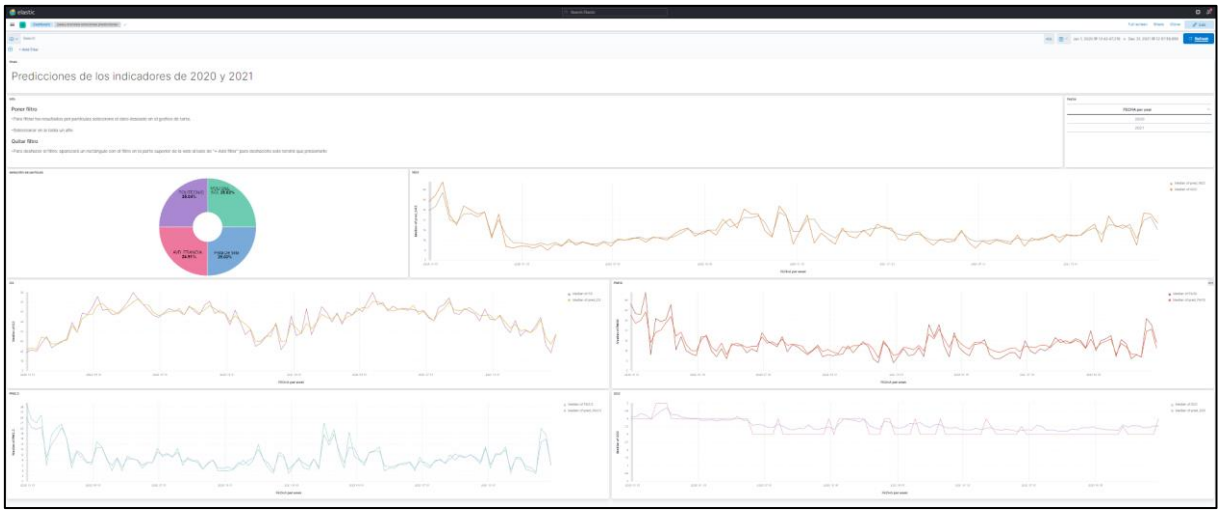


Imagen 5 - Visualización predicciones 2020-2021 por estaciones principales

- Predicciones y datos reales de 2020-2021 (pandemia) por indicador

Imagen 6 - Visualización predicciones 2020-2021 por indicador

3.3 Machine Learning

3.3.1 Selección de datos

Una vez conseguido un dataset con suficiente información histórica, se procede a realizar una selección de los datos a tratar en toda la parte de Machine Learning.

En primer lugar se comprueban los datos disponibles por estación:

Nº de registros totales y PORCENTAJE DE atributos nulos por estación:

===NOM_ESTACION===	Nº reg	=SO2==	==CO==	==NO==	=NO2==	=NOx==	==O3==	=PM10=	PM2.5=	=PM1==	Preci=	=Temp=	Veloc=	Vemax=	TOTAL=
CEMENTERIO	2922	12.66	97.09	92.98	92.98	92.98	94.73	94.83	97.13	99.59	93.57	93.63	93.63	95.21	88.54
LABORATORI	3984	10.22	98.74	97.72	97.72	97.72	98.37	98.87	98.95	99.90	98.49	98.57	98.49	98.80	91.73
VIVERS-M	3984	12.07	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	93.24
POLIGONO	3984	9.69	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	93.05
ARAGÓ-M	3984	9.89	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	93.07
C/SAGUNT	3984	6.50	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	92.81
AVD. PORT	3984	8.58	87.93	88.10	88.10	88.10	88.25	100.00	100.00	100.00	100.00	87.85	87.85	100.00	86.52
PISTA DE SILLA	9919	4.00	4.48	3.40	3.40	3.40	4.74	58.12	71.94	85.44	69.74	2.02	2.02	66.56	29.17
NUEVO CENTRO	5810	2.05	2.01	2.58	2.58	2.58	4.54	99.52	99.52	99.52	100.00	87.61	87.61	100.00	53.09
GRAN VIA	3253	7.41	7.10	10.05	12.88	10.05	19.40	99.97	99.97	99.97	99.97	99.97	99.97	99.97	58.98
ARAGÓN	5079	2.60	25.20	8.49	8.51	7.95	2.48	95.31	95.31	95.31	99.98	99.98	99.98	99.98	57.01
LINARES	6175	3.42	4.71	14.38	14.20	14.36	3.69	95.34	97.80	99.53	99.98	99.21	99.21	99.98	57.37
AYORA	3984	16.82	85.54	86.90	86.90	86.90	85.34	98.29	98.67	99.67	98.67	98.09	98.09	98.57	87.57
N. CENTRO-2	2922	11.26	95.38	95.31	95.31	95.31	95.24	96.89	97.57	98.39	98.25	96.89	96.89	97.98	90.05
NAZARET	731	24.76	35.70	31.19	31.19	31.19	32.56	67.99	67.99	67.99	100.00	100.00	100.00	100.00	60.81
TENDETES	1827	4.98	91.79	91.95	91.95	91.95	91.13	94.03	97.81	98.47	97.26	97.10	97.10	100.00	88.12
NAZARET-MET.	731	78.93	94.66	93.57	93.57	93.57	93.43	95.08	96.17	97.26	41.04	40.22	40.22	96.17	81.07
VIVERS	7396	2.33	39.82	6.79	6.79	6.80	3.66	10.57	34.28	90.89	54.83	52.43	52.43	54.92	32.04
POLITÈCNIC	5205	3.88	83.86	3.55	3.55	3.55	2.19	2.29	2.63	33.39	37.79	37.56	33.49	36.79	21.89
AVD. FRANCIA	4839	1.74	2.52	1.84	1.84	1.84	0.06	23.52	23.52	47.37	33.33	33.33	0.02	25.87	15.14
MOLÍ DEL SOL	4839	3.64	10.35	5.56	5.56	5.56	3.64	10.60	10.85	10.27	32.01	31.43	27.13	31.02	14.43
BULEVARD SUD	4474	4.83	78.07	6.64	6.64	6.64	2.77	31.40	92.27	96.96	25.86	24.41	24.41	25.77	32.82
CONSELLERIA METEO.	3744	74.76	87.87	75.93	75.93	75.93	74.12	78.18	87.58	87.58	4.51	4.54	4.54	4.54	56.62
VALÈNCIA CENTRE	1552	75.26	94.97	14.82	14.82	14.82	81.44	16.49	17.98	98.45	1.80	1.80	1.80	1.80	33.56
ALBUFERA2 (MÒBIL)	731	41.59	35.29	36.80	36.80	36.80	39.26	76.74	84.13	88.10	40.90	40.63	40.63	95.08	53.29
VALENCIA-ALBUFERA	365	33.70	40.00	36.44	36.44	36.44	33.42	44.66	70.96	98.90	0.00	0.00	0.00	0.00	33.15
VALENCIA	365	76.99	92.05	89.86	89.86	89.86	90.68	95.07	100.00	100.00	100.00	100.00	100.00	100.00	94.18
TOTAL ESTACIONES:	100767	11.03	53.91	41.85	41.93	41.83	41.77	69.49	76.19	86.26	74.05	65.80	63.78	74.18	57.08

Cuadro 7 - Porcentaje de atributos nulos por estación

Se decide utilizar el porcentaje total de datos faltantes para ver que estaciones tienen globalmente mayor cantidad de indicadores informados. Las 4 estaciones con mejor ratio total de datos faltantes son:

- 14.43% MOLÍ DEL SOL
- 15.14% AVD. FRANCIA
- 21.89% POLITÈCNIC
- 29.17% PISTA DE SILLA

Este será el subset de estaciones con las que se trabajarán las predicciones.

Nº de registros totales y PORCENTAJE DE atributos nulos por estación:

===NOM_ESTACION===	Nº reg	=SO2==	==CO==	==NO==	=NO2==	=NOx==	==O3==	=PM10=	PM2.5=	=PM1==	Preci=	=Temp=	Veloc=	Vemax=	TOTAL=
PISTA DE SILLA	9919	4.00	4.48	3.40	3.40	3.40	4.74	58.12	71.94	85.44	69.74	2.02	2.02	66.56	29.17
POLITÈCNIC	5205	3.88	83.86	3.55	3.55	3.55	2.19	2.29	2.63	33.39	37.79	37.56	33.49	36.79	21.89
AVD. FRANCIA	4839	1.74	2.52	1.84	1.84	1.84	0.06	23.52	23.52	47.37	33.33	33.33	0.02	25.87	15.14
MOLÍ DEL SOL	4839	3.64	10.35	5.56	5.56	5.56	3.64	10.60	10.85	10.27	32.01	31.43	27.13	31.02	14.43
TOTAL ESTACIONES:	24802	3.46	21.90	3.55	3.55	3.55	3.08	30.38	36.03	52.42	48.57	21.32	13.13	45.44	22.03

Cuadro 8 - Porcentaje de atributos nulos por estación seleccionada

De las cuatro estaciones seleccionadas se descartan las características CO y PM1 pues en algún caso supera el 80% de valores nulos.

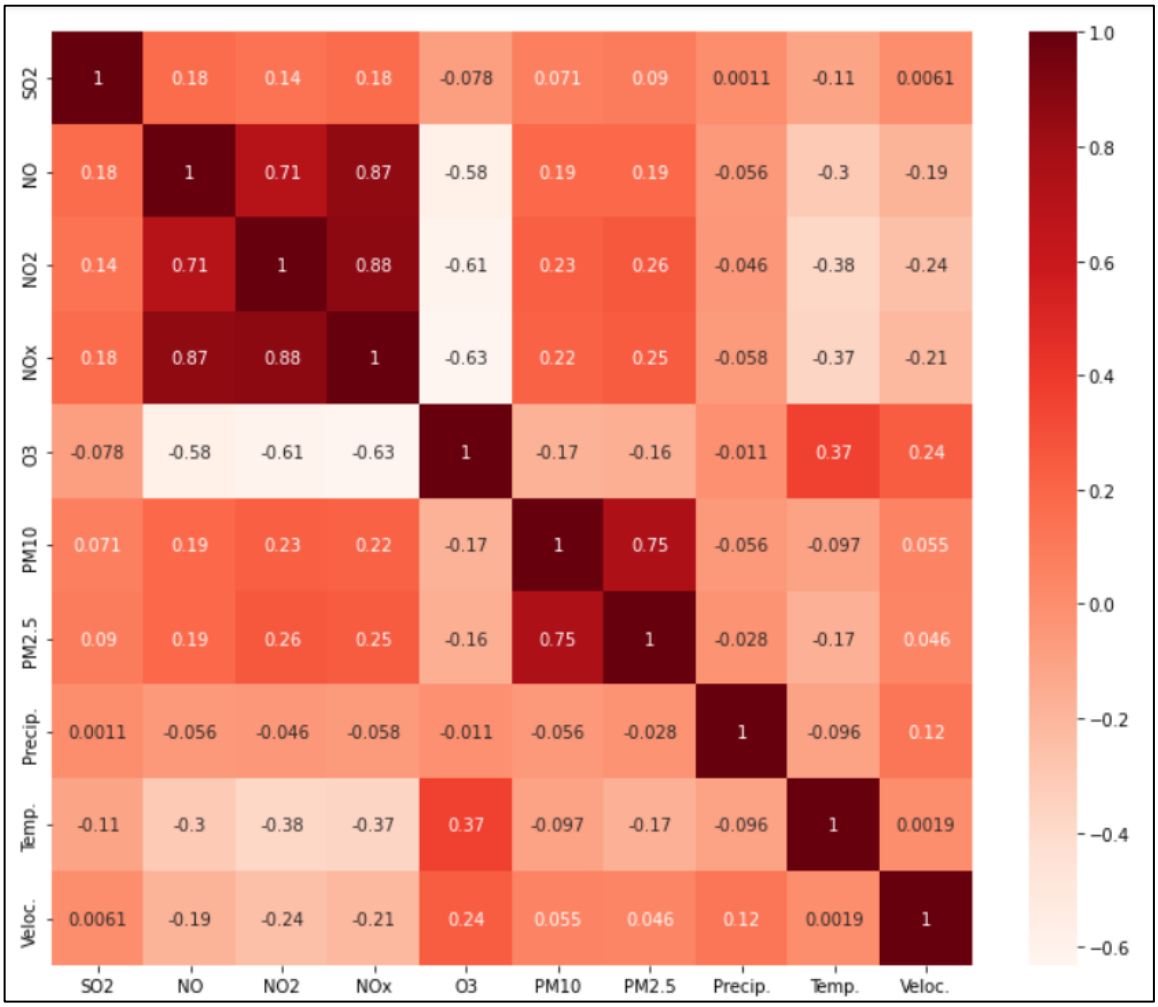
Finalmente se queda un dataset con esta información:

```
Nº de registros totales y rango de fechas por estación:
==CODIGO== ===NOM_ESTACION=== ==Desde=== ==Hasta=== =Nº reg=
[46250030] PISTA DE SILLA     1994-01-01 2022-03-31     9919
[46250046] POLITÈCNIC         2008-01-01 2022-03-31     5205
[46250047] AVD. FRANCIA       2009-01-01 2022-03-31     4839
[46250048] MOLÍ DEL SOL       2009-01-01 2022-03-31     4839
```

Cuadro 9 - Nº de registros y rango de fechas por estación seleccionada

3.3.2 Correlación

Inicialmente se ha hecho un estudio de la correlación entre los distintos atributos para comprobar que influencia tienen entre sí, obteniendo los siguientes resultados:



Cuadro 10 - Correlación entre características

A la vista de los resultados del estudio se concluye que no hay correlación entre los distintos contaminantes y las partículas. Solo influyen entre sí las partículas PM10 con PM2.5 y por otra parte los contaminantes óxidos de nitrógeno con el ozono.

En contra de lo que a priori se podría haber afirmado, el resultado más sorprendente es que no existe correlación entre las variables meteorológicas y las partículas o contaminantes.

3.3.3 Modelos estudiados

Para realizar las predicciones se ha optado por un modelo de **series temporales**. A tal efecto se han realizado diversas pruebas para decidir qué modelo exacto utilizar. Se han probado los siguientes tipos de modelo:

- Series temporales con tensorflow
    - Base line
    - Linear
    - Dense
    - Multi step Dense
    - Convolutional
    - LSTM
- Series temporales con statsmodels
    - Sarima

Los modelos que han dado mejores resultados han sido LSTM (Long Short Term Memory) y SARIMA (Seasonal Autoregressive Integrated Moving Average).

No obstante, se ha optado por la red neuronal recurrente LSTM debido a que ofrece unas ratios de error absoluto medio y error cuadrático medio mejores que SARIMA.

Además, los tiempos de computación de SARIMA para compilar y entrenar el modelo son unas 7 veces peores que los ofrecidos por LSTM.

Una vez elegido el modelo a utilizar (LSTM), se realizaron pruebas con distintos tamaños de ventanas temporales tanto de entrada como de salida. Se llegó a la conclusión de que con una ventana de entrada de 30 días y una de salida de 1 día de predicción se obtenían buenos resultados y una velocidad de entrenamiento muy razonable.

En cuanto al entrenamiento, debido a que los datos reales de los dos años de pandemia (2020 y 2021) podían afectar a las predicciones que hiciera el modelo, se ha utilizado la información histórica disponible desde la fecha más antigua hasta final de 2019 para entrenar la RNN LSTM.

Para delimitar el alcance de los modelos, solo se han tenido en cuenta los indicadores de calidad del aire principales (según el ministerio de Transición Ecológica): PM10, PM2.5, O3, NO2 y SO2)

Dado que cada estación tiene sus peculiaridades, se ha creado un script que automáticamente genera un modelo predictivo LSTM por cada combinación de estación e indicador de calidad del aire, partiendo de los datos históricos disponibles.

3.3.4 Predicciones

Se han hecho predicciones con los 20 modelos para cada estación e indicadores principales de los valores correspondientes a los años de la pandemia. Por ejemplo, para el indicador PM10 se han obteniendo los siguientes resultados:

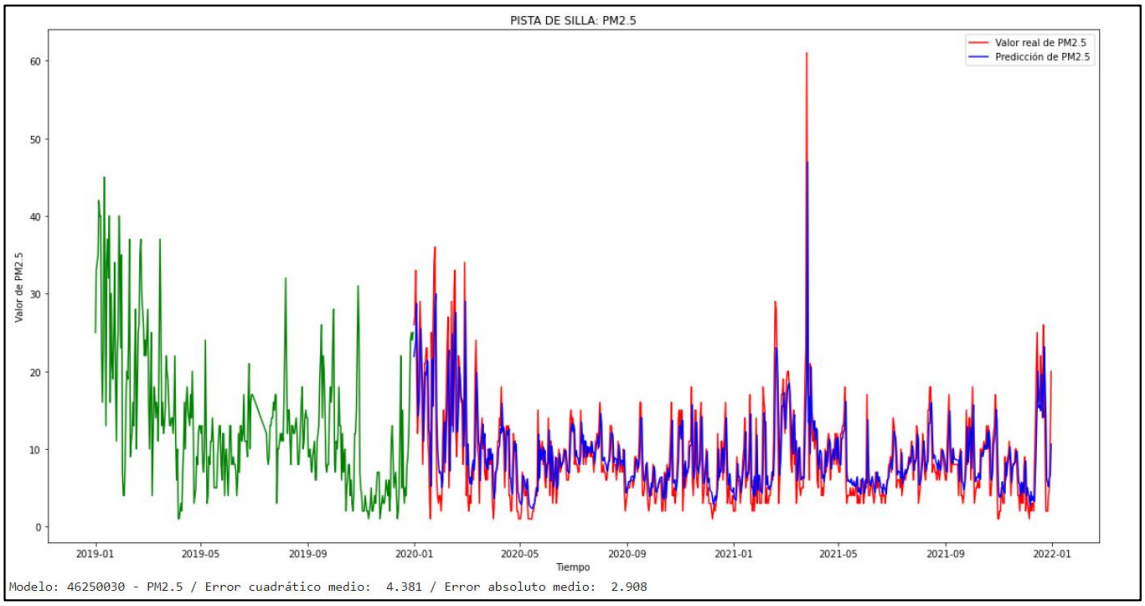


Imagen 11 - Predicción PM2.5 para 2020 y 2021 estación Pista de Silla

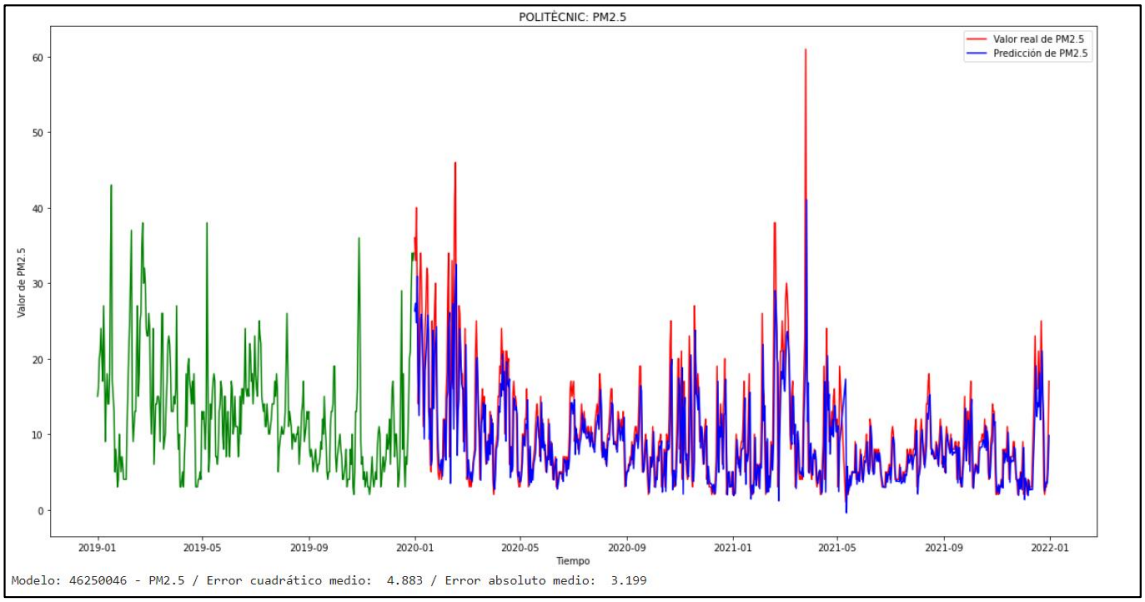


Imagen 12 - Predicción PM2.5 para 2020 y 2021 estación Politécnico

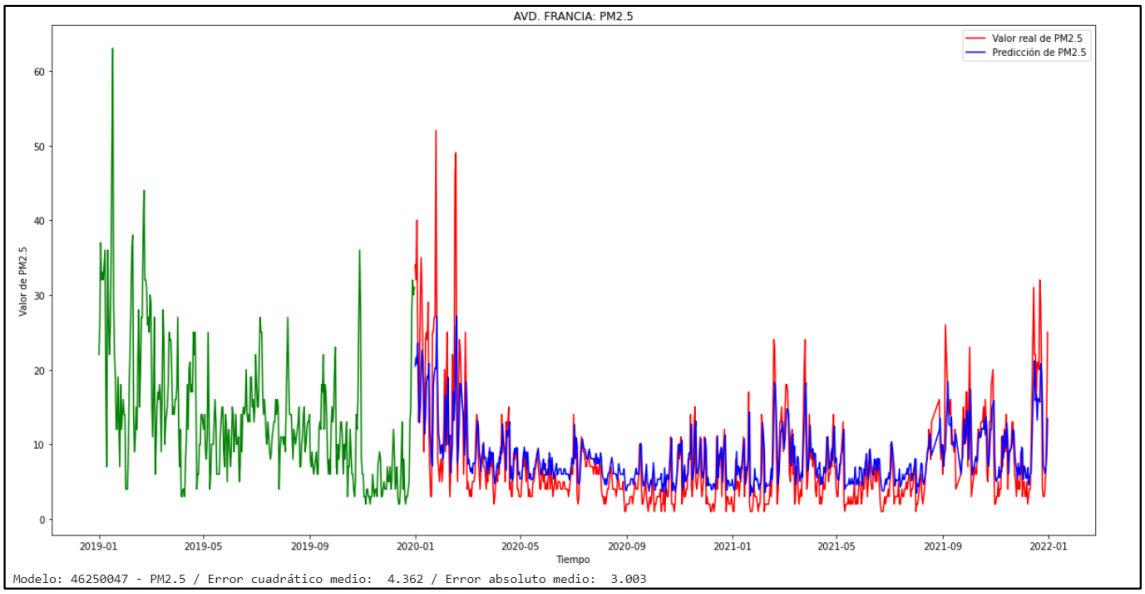


Imagen 13 - Predicción PM2.5 para 2020 y 2021 estación Av. de Francia

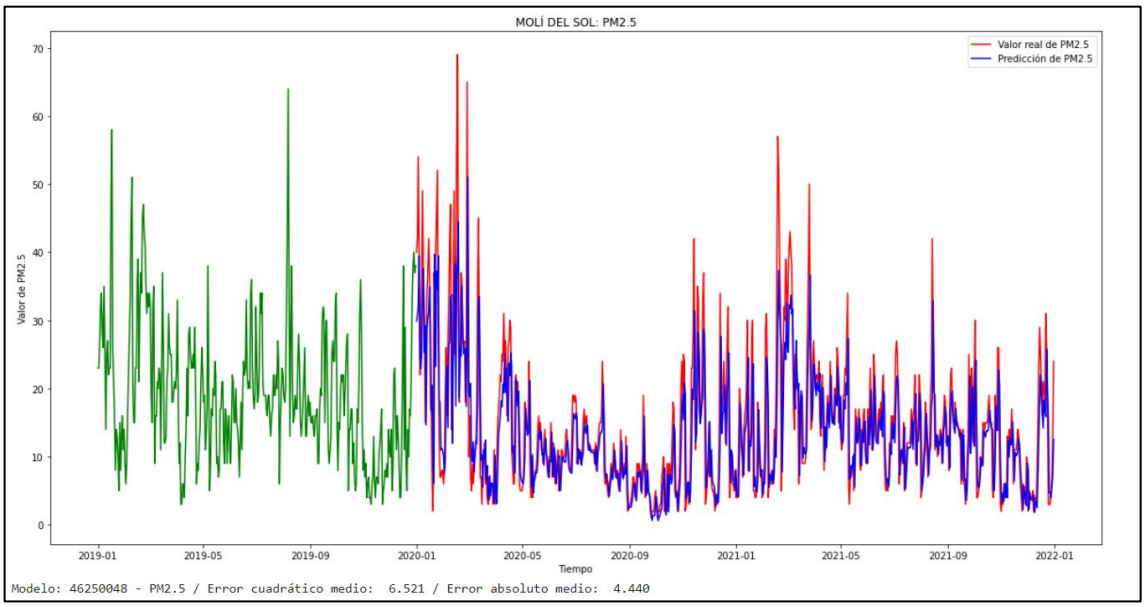


Imagen 14 - Predicción PM2.5 para 2020 y 2021 estación Molí del Sol

3.4 Datos abiertos y visualizaciones online

Además de los recursos para visualizaciones, presentación de datos y cuadros de mandos, que proporcionan las herramientas utilizadas en este proyecto se incorporan elementos online a través de la web:

- Acceso al portal principal del proyecto http://humanobasico.net/

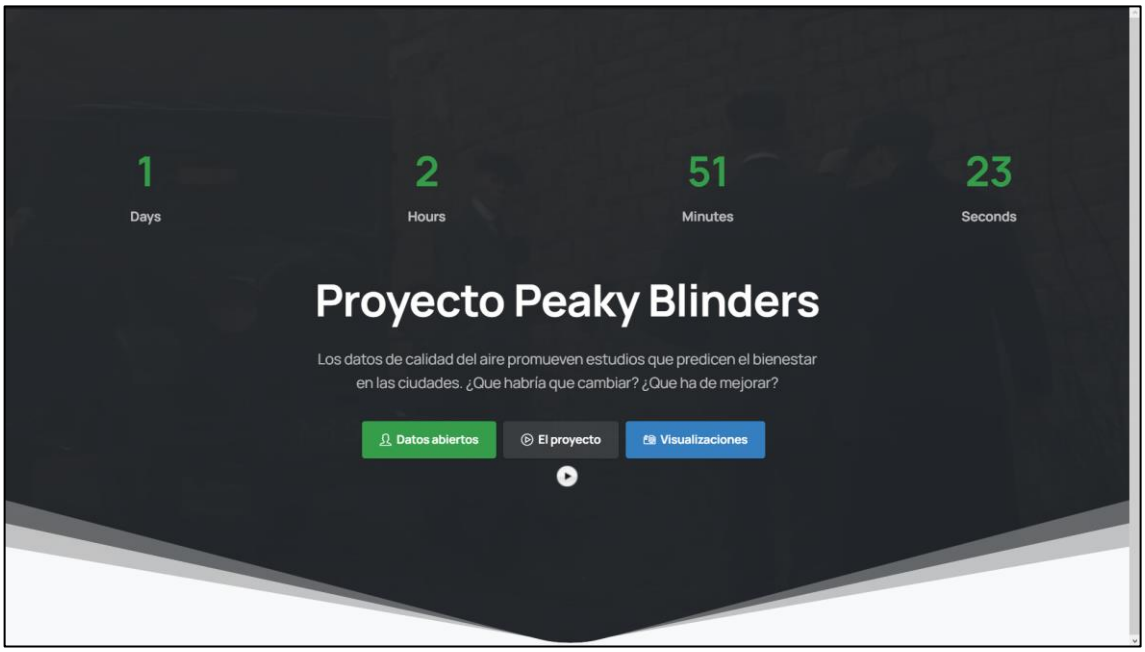


Imagen 15 - Captura de pantalla de la web de acceso

- Acceso a los conjuntos de datos procesados para realizar los entrenamientos y modelos desarrollados en el proyecto (Datos abiertos) http://humanobasico.com/

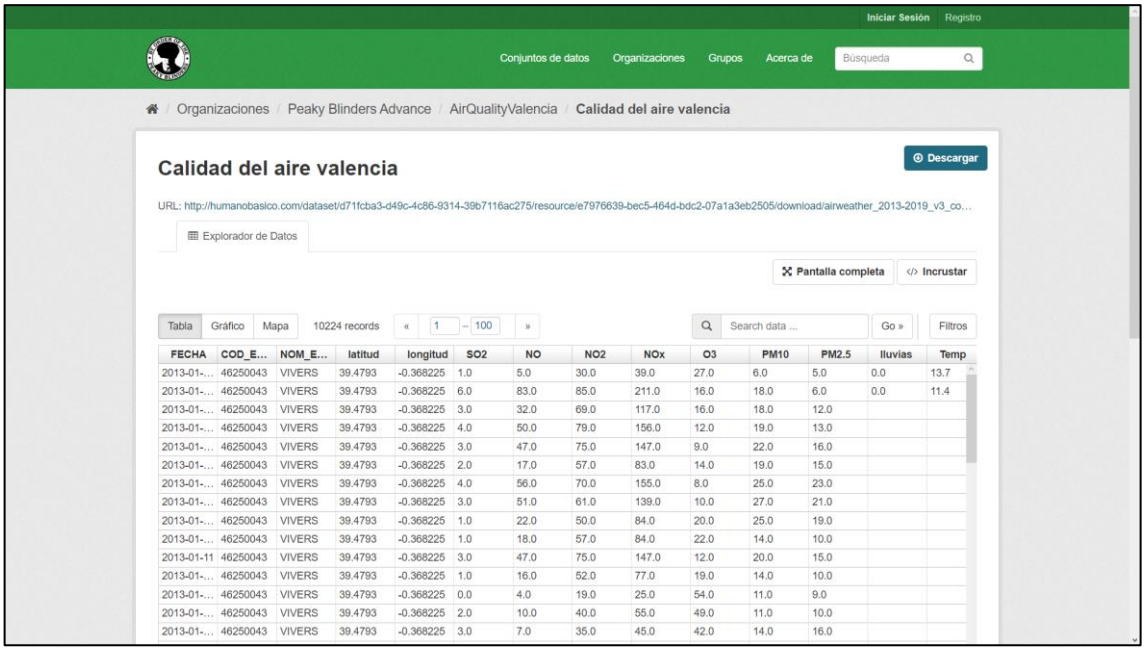


Imagen 16 - Captura de pantalla de la web de datos abiertos

- Acceso a información sobre el proyecto.

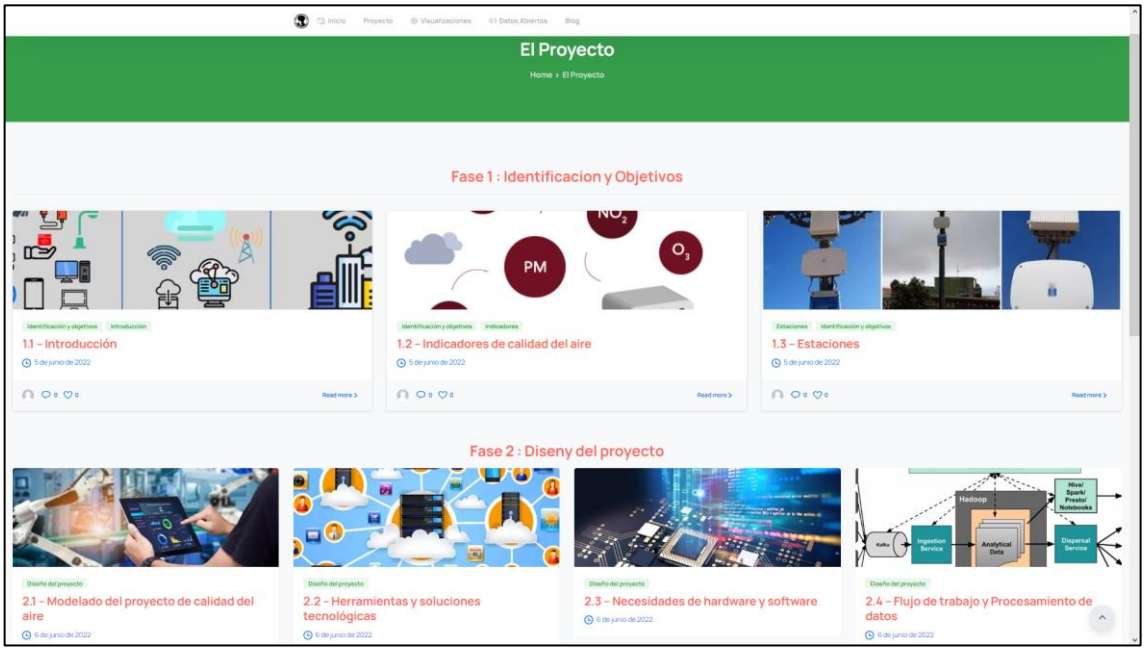


Imagen 17 - Captura de pantalla de información sobre el proyecto

- Acceso a visualizaciones y cuadros de mando.

Imagen 18 - Captura de pantalla de ejemplo de visualizaciones

3.5 Convenciones adoptadas

En todo el proyecto, para los archivos de tipo csv se ha convenido utilizar como separador el punto y coma ";" y para separar decimales el punto "."

3.6 Licencia

Se ha optado por utilizar la licencia Creative Commons Atribución/Reconocimiento-No Comercial-Compartir Igual 4.0 Internacional (CC BY-NC-SA).

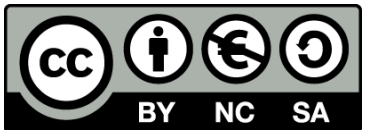


Imagen 19 - Licencia de software

Las características más significativas de esta licencia son:

- Se debe atribuir o reconocer al creador.
- Solo se permiten usos no comerciales de la aplicación.
- Cualquier adaptación del software debe compartirse en los mismos términos.

Más información en https://creativecommons.org/licenses/by-nc-sa/4.0/deed.es_ES

4 EVALUACIÓN Y CONCLUSIONES FINALES

4.1 Evaluación de la ejecución del proyecto

Inicialmente se ha tenido una gran dificultad para recabar los datos, habiendo invertido mucho tiempo en investigar, por una parte, las fuentes disponibles desde donde recuperar la información y, por otro lado, en las tareas de Análisis Exploratorio de Datos y en la Extracción, Transformación y Carga de datos (EDA – ETL).

Ha resultado sorprendente descubrir que no hay correlación significativa entre los indicadores de calidad del aire y la información meteorológica.

Los modelos han dado unos resultados muy satisfactorios en base a las series temporales de datos utilizadas.

4.2 Propuesta de mejoras

Para darle la necesaria continuidad al proyecto se precisará la incorporación de algunos elementos:

- Automatizar la captura e inserción de nuevas mediciones diarias que vayan alimentando y actualizando las fuentes de datos.
- Automatizar los procesos de aplicación de modelos para realizar las nuevas predicciones diarias.
- Automatización de la generación de presentaciones y cuadros de mandos con los datos actualizados.
- Continuar la publicación de los nuevos datos adquiridos de predicción y conjunto de datos en los elementos online dispuestos para el acceso público a los mismos.

5 REFERENCIAS

5.1 Referencias de fuentes de datos

Portal de datos abiertos de la Generalitat Valenciana	https://portaldadesobertes.gva.es/es	GENERALITAT VALENCIANA DADES OBERTES
Consellería de agricultura, desarrollo rural, emergencia climática y transición ecológica	https://agroambient.gva.es/	GENERALITAT VALENCIANA Conselleria de Agricultura, Desarrollo Rural, Emergencia Climática y Transición Ecológica
Portal medición calidad del aire en el mundo (fork)	https://waqi.info/es/	
Portal de calidad del aire en el mundo	https://aqicn.org/map/world/es/	
Agencia Estatal de Meteorología	http://www.aemet.es/	AEMet Agencia Estatal de Meteorología
Base de datos Meteorológica	https://datosclima.es/	
Associació Valenciana de meteorología Josep Peinado	https://www.avamet.org/	avamet associació valenciana de meteorologia Josep Peinado

Cuadro 20 - Referencias a Fuentes de Datos

5.2 Referencias de tecnologías utilizadas en el proyecto

Elasticsearch	https://www.elastic.co/es/	elasticsearch
Logstash	https://www.elastic.co/es/logstash/	logstash
Kibana	https://www.elastic.co/es/kibana/	kibana
Lenguaje Python	https://www.python.org/	python
Tensorflow	https://www.tensorflow.org/	TensorFlow
Statsmodels	https://www.statsmodels.org	statsmodels
Pandas	https://pandas.pydata.org/	pandas
Numpy	https://numpy.org/	NumPy
Mathplotlib	https://matplotlib.org/	matplotlib
Dataprep	https://dataprep.ai/	dataprep
Framework datos abiertos	https://ckan.org/	ckan
Framework para blogs	https://wordpress.com/es/	WordPress

Lenguaje páginas web	https://www.php.net/	
Base de datos MySQL	https://www.mysql.com/	
Servidor web	https://www.nginx.com/	
Servidores	https://www.ovhcloud.com/	
Licencias de Creative Commons	https://creativecommons.org/	

Cuadro 21 - Referencias a tecnologías utilizadas