UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escola Tècnica Superior d'Enginyeria Informàtica

Estudi de la qualitat de l'aire a l'àrea metropolitana de la ciutat de València

Treball Fi de Grau

Grau en Ciència de Dades

AUTOR/A: Llinares Llinares, Anna

Tutor/a: Gómez Adrian, Jon Ander

CURS ACADÈMIC: 2022/2023

Resum

A pesar d'haver-se constatat una reducció de la contaminació de l'aire respecte a les dues dècades anteriors, la qualitat de l'aire segueix sent un problema molt greu. Els contaminants que habitualment es detecten són precursors de malalties respiratòries, cardíaques i cerebrals, a més de nombroses morts prematures.

València és una de les ciutats que s'adhereix als Objectius de Desenvolupament Sostenible de l'agenda 2030 i es troba en la posició 163 de 344 ciutats amb l'aire més net del ranking de l'Agència Europea del Medi Ambient. A més, actualment compta amb 13 estacions de mesuraments de distints contaminants atmosfèrics localitzades en distintes zones de la ciutat.

Així doncs, aquest estudi avalua la qualitat de l'aire a la ciutat de València a partir de les dades recopilades en les diverses estacions, on es disposa d'una mesuració per hora. Al tenir registres cada hora de nombroses estacions durant 6 anys, hi ha moltes variables amb dades faltants. Per això, s'estudia l'evolució de cada indicador individualment, i també conjuntament amb altres variables, en funció de cada estació i per a distints períodes temporals. Després, per determinats indicadors on hi haja fins a un 20% de dades faltants, se'n fa una estimació.

Una vegada feta la predicció, es proposa una eina visual que facilite la comprensió dels perills que comporta la contaminació de l'aire, gràcies a la representació dels nivells dels contaminants als quals estan sotmeses les distintes zones de la ciutat de València. D'aquesta manera, es comparen els valors actuals de cada contaminant, en cada estació i període, amb el seu respectiu valor límit proporcionat per l'Organització Mundial de la Salut.

Paraules clau: Qualitat de l'aire; contaminació; mesuracions atmosfèriques; contaminants; anàlisi de dades

Resumen

A pesar de haberse constatado una reducción en la contaminación del aire respecto a las dos décadas anteriores, la calidad del aire sigue siendo un problema muy grave. Estos contaminantes, son precursores de enfermedades respiratorias, cardíacas y cerebrales, además de numerosas muertes prematuras.

Valencia es una de las ciudades que se adhiere a los objetivos de desarrollo sostenible de la agenda 2030 y se encuentra en la posición 163 de 344 ciudades con el aire más limpio en el ranking de la Agencia Europea del Medio Ambiente. Además, actualmente cuenta con 13

estaciones de medición de distintos contaminantes atmosféricos localizadas en distintas zonas de la ciudad.

Así pues, este estudio evalúa la calidad del aire en la ciudad de Valencia a partir de los datos horarios recolectados de las distintas estaciones. Al tener registros horarios de muchas estaciones a lo largo de 6 años, hay muchas variables con datos faltantes. Por eso, se ha estudiado la evolución de cada indicador individualmente, y también conjuntamente con otras variables, en función de cada estación y para distintos periodos temporales. Después, para determinados indicadores donde se supere un 20% de datos faltantes, se hace una estimación.

Hecha la predicción, se propone una herramienta visual, que facilita la comprensión del peligro que conlleva la contaminación del aire y los niveles a los cuales las distintas zonas de la ciudad de Valencia se ve sometida, comparando los valores actuales de cada indicador, en cada estación y periodo, con su respectivo valor límite proporcionado por la Organización Mundial de la Salud.

Palabras clave: Calidad del aire; contaminación; medidas atmosféricas; contaminantes; análisis de datos

Abstract

Despite a reduction in air pollution compared to the previous two decades, air quality remains a very serious problem. These pollutants are precursors of respiratory, cardiac and brain diseases, in addition to numerous premature deaths.

Valencia is one of the cities that joins to the Sustainable Development Goals of the 2030 agenda and is number 163 in the ranking of 344 cities with the cleanest air. This ranking is published by the European Environment Agency. In addition, it currently has 13 stations measuring different air pollutants located in different areas of the city.

Thus, this study evaluates the air quality in the city of Valencia from the data collected from the different stations. Having records every hour from many stations during 6 years, there are many variables with missing data. Therefore, the evolution of each indicator has been studied individually, and also together with other variables, depending on each season and for different time periods. Then, for certain indicators where missing data is lower or equal 20%, an estimate is done.

Once the prediction be available, a visual tool had been developed to facilitate the understanding of the danger posed by air pollution, thanks to the representation of the levels of

the pollutants measured in the different areas of the city of Valencia, comparing the current values of each indicator, at each station and period, with their respective limit value provided by the World Health Organization.

Keywords: Air quality; pollution; weather analytics; pollutants; data analytics

Índex general

Índex de taules

Índex de figures

1. Introducció

Hui en dia la qualitat de l'aire és una de les majors preocupacions pels efectes directes en la salut humana i ambiental, sent la causa de moltes malalties i morts prematures. L'Organització Mundial de la Salut (OMS) ha estimat recentment que 3,7 milions de morts anuals al món podrien atribuir-se a la contaminació atmosfèrica. A més a més, l'exposició a curt o a llarg termini a la contaminació de l'aire danya la salut durant la infància i agreuja els riscos de patir malalties futures, sent precursora de malalties respiratòries, cardiovasculars, mentals i també càncers.

Per suposat, la ciutat de València no és aliena a aquest problema, i cada vegada es veu més afectada amb els alts nivells de contaminants atmosfèrics. D'aquesta manera, d'acord amb l'article 50, competència 6 de l'Estatut d'Autonomia de la Comunitat Valenciana, li correspon a la Generalitat el desenvolupament legislatiu i l'execució de les matèries en la protecció del medi ambient [1].

Per a portar-lo a terme, compta amb la Xarxa Valenciana de Vigilància i Control de la Contaminació Atmosfèrica, la qual fa un seguiment continu dels nivells dels diferents contaminants atmosfèrics al llarg de les principals àrees urbanes i industrials de la Comunitat Valenciana.

Aquesta xarxa compta amb nombroses estacions de mesurament operatives al llarg de la Comunitat Valenciana, les quals capten fins 24 paràmetres contaminants i meteorològics. En aquest treball, sols s'analitzaran les dades procedents de les 12 estacions disponibles en el municipi de València, centrant-se en l'avaluació i comparació de la qualitat atmosfèrica a la ciutat durants els períodes temporals compressos entre l'any 2017 i 2022.

1.1. Objectius

L'objectiu principal és analitzar els nivells dels diferents contaminants atmosfèrics. A més, es busca identificar possibles diferències entre els anys, especialment en l'any 2020, dins del context de la pandèmia del Covid-19 on la ciutat de València es va aturar quasi per complet.

Altres objectius importants són (a) l'estudi de les correlacions entre variables i (b) comprovar si la temperatura, humitat relativa o precipitació tenen un paper rellevant en la qualitat de l'aire. Es tracta de trobar una relació entre els nivells de contaminació atmosfèrica i

les variacions estacionals, fet que permet comprendre millor els factors que influeixen en la qualitat de l'aire.

Així mateix, cal tenir en compte que els sensors capten cada hora un nou registre per cada contaminant o paràmetre meteorològic. D'aquesta manera, pot haver moments en els quals un sensor falla i no és capaç de recopilar dades, és per això que altre objectiu és predir les dades faltants per als diferents paràmetres en determinats períodes temporals, en els quals es disposa dels valors mesurats de les altres variables, que s'utilitzaran com a entrada al model predictiu.

Per últim, amb un conjunt de dades més complet gràcies a la predicció de les dades faltants amb models de machine learning, es proposa desenvolupar una ferramenta visual (dashboard) que permeta una observació clara i comprensible dels diferents paràmetres atmosfèrics.

Aquest dashboard té la finalitat de proporcionar una visió completa, clara i accessible als usuaris sobre la contaminació atmosfèrica a les diferents zones de la ciutat de València, facilitant la comprensió i comunicació dels resultats obtinguts. Així com, contribuir en el coneixement científic i aportar informació útil per a la presa de decisions en matèria de salut pública i polítiques ambientals a València per la qualitat de vida i la consciència pública sobre la importància de la qualitat de l'aire.

1.2. Motivació

Deia el poeta francés Victor Hugo que produeix una inmensa tristesa pensar que la natura ens parla mentre el gènere humà no l'escolta. Estar atents al nostre entorn, a tots els missatges que envia la natura és una obligació que té la societat.

En un moment de grans reptes mediambientals davant del canvi climàtic no es pot mirar cap a un altre costat, s'ha d'observar el que ocurreix al nostre voltant i actuar segons les possibilitats de cada individu. Per això aquest estudi naix amb l'objectiu de conèixer la situació de la qualitat de l'aire de València i intentar ajudar en aquesta problemàtica, contribuint a una societat i un món més sostenible per a tots els éssers vius i l'ecosistema en general.

Tenint en compte la responsabilitat social que els individus tenen en l'acció mediambiental tant local com global, des del primer curs d'aquest grau s'ha treballat amb dades relacionades amb el medi ambient, tot i que en petita escala, anàlisis de dades sense transcendència a la societat, simplement per aprendre més sobre ciència de dades.

Arribat el moment d'afrontar el TFG es va considerar que calia seguir amb la mateixa línia de treball però de manera més directa i concreta. Per això, una forma de contribuir de manera

real a millorar l'entorn més proper i evitar la seua 'degradació mediambiental' era estudiar i analitzar la contaminació de l'aire a la ciutat de València.

Així, posant en pràctica habilitats i coneixements adquirits al llarg del grau, s'ha abordat un problema real per buscar-li les solucions adients. De fet aplicant tot allò aprés als anys d'estudi amb aquest treball s'ajuda a conèixer i comprendre la situació actual de la qualitat de l'aire de la ciutat de València.

1.3. Estructura

Aquest treball de fi de grau està estructurat en els següents nou capítols fonamentals.

Primer es descriuen els motius que han conduït a la selecció d'aquest tema per al seu estudi, així com els objectius per desenvolupar el projecte i l'estructura del treball per a proporcionar una visió general de tot el contingut. El segon capítol, l'Estat de l'art, on s'ofereix una revisió actualitzada de l'estat de les tecnologies, els mètodes i la situació actual. En tercer lloc, s'explica els materials i mètodes emprats al llarg de tota la investigació. Després, al quart capítol es descriuen de manera detallada les fonts i la naturalesa dels conjunts de dades utilitzats. En cinqué lloc, es comenta l'anàlisi exploratori de dades: realitzant una anàlisi descriptiva i de correlacions entre les variables. Es considera especialment l'estudi dels valors faltants per a diferents estacions i períodes d'anàlisi seleccionats. Al sisé capítol es mostra la representació visual de l'anàlisi de dades. En seté lloc s'esmenta la reconstrucció de dades mitjançant l'aprenentatge automàtic, on s'explica l'experimentació realitzada i els resultats obtinguts. També s'inclou el càlcul de les prediccions realitzades amb aquests models. Al vuité capítol, es crea altra representació visual amb les noves dades reconstruïdes amb Power BI. Per últim, al nové capítol es comenten les conclusions, s'analitza el marc legal i ètic relacionat amb l'estudi i els futurs passos del projecte.

2. Estat de l'art

El 25 de setembre de 2015, 193 països membres de Nacions Unides van acordar abordar conjuntament el pla de l'Agenda 2030 i els objectius de Desenvolupament Sostenible (ODS). Aquest acord és un compromís comú i universal per promoure la prosperitat humana, la lluita contra la pobresa, la protecció del medi ambient i la disminució de les desigualtats. Convertint-se en un referent fonamental per a la búsqueda d'un desenvolupament sostenible en tots els aspectes de la societat [2].

La qualitat de l'aire i protecció del medi ambient és un dels desafiaments més apremiants a nivell mundial, ja que el risc medioambiental afecta a la salut de totes les persones i a la biodiversitat. Sobretot amb el creixement urbà desmesurat i l'empitjorament de la contaminació de l'aire com a resultat de barris marginals, infraestructures i serveis inadequats, ja que des del 2007 més de la meitat de la població mundial viu a ciutats [3].

S'estima que al 2019 la contaminació de l'aire va causar 4.2 milions de morts prematures, de les quals el 89% d'aquestes ocorregueren en països d'ingressos baixos [4]. A pesar que la major càrrega de contaminació es registra en les regions d'Àsia Sud Oriental i del Pacífic Occidental, la ciutat de València no és aliena a aquest problema.

València ja ha emprés el camí per a complir amb l'Agenda urbana València 2030 i convertir-se en una Smart City, és a dir una Ciutat Intel·ligent i Sostenible que aprofita les Tecnologies de l'informació i la Comunicació (TIC) i altres mitjans per millorar la qualitat de vida, la competitivitat, l'eficiència del funcionament i els serveis urbans, alhora que s'assegura que respon a les necessitats de les generacions present i futures pel que fa als aspectes econòmics, socials, mediambientals i culturals, segons UNE 178201:2016 [5].

Per tant, per a abordar aquests desafiaments de manera global i col.laborativa, el Quadripartit format per la FAO, l'OMS, la WOAH i el PNUMA va formular l'estratègia One Health, "Una sola salut". La qual fomenta un enfocament integrat i unificador per optimitzar de manera sostenible la salut humana, dels animals i dels ecosistemes. Mobilitzant professionals de múltiples sectors, disciplines i comunitats per treballar conjuntament per prendre mesures sobre el canvi climàtic i contribuir al desenvolupament sostenible [6].

Els estudis més recents utilitzen dades de qualitat de l'aire i variables socioeconòmiques per entrenar i validar models d'aprenentatge automàtic, amb la finalitat de predir la qualitat de l'aire de diferents contaminants, centrant-se en diferents àrees urbanes d'Àsia i Europa. Tenint en

compte variables com les emissions dels contaminants, el trànsit rodat, la densitat poblacional, entre altres variables rellevants [7].

Respecte a una investigació amb dades horàries de l'estat de Califòrnia a Estats Units que es centra en predir l'ozó, CO, SO_2, PM2.5, PM10 i també l'índex de qualitat de l'aire, utilitzant com a regressor un model de Màquines de Vectors Suport. El resultats obtinguts demostren que l'experimentació ha obtingut una precisió satisfactòria, encara que en alguns casos com per al monòxid de carboni, els models s'han vists afectats pels valors anòmals. Aquest estudi pot tenir implicacions molt importants a l'hora de la presa de decisions relacionades amb la qualitat de l'aire de Califòrnia i altres regions similars [8].

Altres estudis rellevants per entendre aquesta problemàtica, analitzen les diferents fonts de contaminació de l'aire en àrees urbanes, així com les emissions dels vehicles i activitats industrials. On es destaca la importancia de crear polítiques i mesures per moderar i millorar els efectes nocius de la contaminació atmosfèrica en l'ambient urbà [9].

És important destacar que aquest estudi té un enfocament significativament diferent als esmentats prèviament, ja que l'objectiu principal no és predir els paràmetres atmosfèrics ni meteorològics en el futur. Sinó centrant-se en les pròpies dades ja que aquestes tenen un número elevat de dades faltants perquè aquesta problemàtica pot tindre una greu manca de representativitat i fiabilitat en estudis futurs si no s'aborda de manera adequada.

Així doncs, aquest estudi es centra en la predicció dels valors faltants amb la finalitat d'imputarlos al conjunt de dades original, per a completar la informació disponible la qual cosa permet tindre anàlisis més representatius i precisos.

Aquesta aproximació innovadora és fonamental per garantir la integritat en les dades. Al superar aquest repte predint les dades amb models de regressió de ML, la investigació té un gran potencial per proporcionar una visió més completa, transparent i representativa de la situació actual de la contaminació de l'aire a les distintes zones de la ciutat de València.

3. Material i mètodes

3.1. Tipus de contaminants

Existeixen dos tipus de contaminants atmosfèrics segons el seu origen, primaris o secundaris. A continuació s'explicaran els distints contaminants i els llindars definits per les Directrius mundials de l'OMS sobre la qualitat de l'aire, actualitzades al 2021 [10] i els valors legislats pel Reial decret 34/2023, de 24 de gener, pel qual es modifiquen el Reial decret 102/2011, de 28 de gener, relatiu a la millora de la qualitat de l'aire (BOE-A-2023-2026), llei que aborda la gestió de la qualitat de l'aire i la protecció de l'atmosfera [11].

3.1.1. Contaminants primaris

Els contaminants primaris són les substàncies emeses directament de fonts específiques, com poden ser la crema de combustibles fòssils, la indústria, activitats agrícoles, vehicles rodats, etc. Aquests contaminants representen més del 90% de la contaminació atmosfèrica. Entre ells destaquen:

Diòxid de Sofre (SO_2)

L'SO_2 és un gas que es forma en la combustió de carburants fòssils, els quals contenen sofre. Sobretot els processos industrials són els primers emissors.

Aquest contaminant té uns efectes molt nocius en la salut, sobretot problemes respiratoris, i en l'ecosistema, sent precursor de la formació de sulfat d'amoníac, el qual augmenta els nivells de PM10 i PM2.5, que també ocasionen dràstiques conseqüències per a la salut [12].

		Període	**Valor**
Diòxid de sofre			
Llindar per a la protecció de la salut humana	BOE	1 hora	Llindar d'informació = 350 µg/m³ No deu de superar-se en més de 24 ocasions per any civil
			Llindar d'alerta = 500 µg/m³ No deu de superar-se en més de 24 ocasions per any civil

		24 hores	Llindar = 180 µg/m³ No deu de superar-se en més de 3 ocasions per any civil
	OMS	24 hores	Llindar = 40 µg/m³ No deu de superar-se en més de 3 ocasions per any civil

Taula 1: Llindars per a la protecció de la salut humana en relació al diòxid de sofre.

Monòxid de Carboni (CO)

El CO es forma a partir de la combustió incompleta de qualsevol compost que continga carboni. Els principals emissors de CO són els processos de combustió en sectors no industrials, és a dir, activitats del sector agropecuari i el trànsit rodat.

En concentracions molt elevades pot arribar a ser mortal, reduint la capacitat de la sang per a transportar oxígen [13].

		Període	Valor
Monòxid de carboni			
Llindar per a la protecció de la salut humana	BOE	24 hores	Llindar = 10 mg/m³ No deu de superar-se en més de 3 ocasions per any civil
	OMS	24 hores	Llindar = 4 mg/m³ No deu de superar-se en més de 3 ocasions per any civil

Taula 2: Llindars per a la protecció de la salut humana en relació al monòxid de carboni.

Òxids de nitrogen (NO, NO_2, NO_X)

Els òxids de nitrogen es produeix amb l'oxidació del nitrogen atmosfèric durant la combustió a altes temperatures. A les grans ciutats més del 75 % d'NO_2 en l'aire és emès pel trànsit rodat.

A més, l'NO_2 és la principal substància química amb efectes nocius en la salut i el medi ambient per sí mateixa. Però també és precursora de partícules inorgàniques, de l'ozó i d'altres compostos fotoquímics al reaccionar amb compostos orgànics volàtils (COVs), la qual cosa agrava els efectes negatius tant en la salut com en l'ecosistema [14].

		Període	Valor
Òxids de nitrogen (NO_2 i NO_X)			
Llindar per a la protecció de la salut humana	BOE	1 hora	Llindar = 200 µg/m³ No deu de superar-se en més de 18 ocasions per any civil
		1 any civil	Llindar = 40 µg/m³
	OMS	1 any civil	Llindar = 10 µg/m³ No deu de superar-se en més de 3 ocasions per any civil

Taula 3: Llindars per a la protecció de la salut humana en relació als Òxids de nitrogen (NO_2 i NO_X).

Partícules en suspensió (PM2.5 i PM10)

Aquestes partícules es classifiquen segons el seu diàmetre aerodinàmic en PM10 (diàmetre aerodinàmic inferior a les 10 micres) o PM2.5 (diàmetre aerodinàmic inferior a les 2.5 micres). Cal destacar que en Espanya sempre ha hagut un nivell elevat de partícules, les quals s'incrementen de forma natural per les intrusions de pols africana.

Les PM10 solen tenir un origen natural, com les partícules d'aerosol formades per l'oceà o el desert. En canvi, les PM2.5 inclouen tot tipus de combustions, provinents del trànsit rodat, la crema de fusta o agrícoles, incendis forestals i alguns processos industrials. Aquestes últimes, són més perilloses ja que pel seu diminut tamany, poden ser inhalades i poden arribar fins als alvèols provocant dificultats respiratòries [15].

		Període	Valor
Partícules en suspensió de diàmetre inferior a 10 micres (PM10)			
Llindar per a la protecció de la salut humana	BOE	24 hores	Llindar = 50 µg/m³ No deu de superar-se en més de 35 ocasions per any civil
		1 any civil	Llindar = 40 µg/m³
	OMS	1 any civil	Llindar = 15 µg/m³

Partícules en suspensió de diàmetre inferior a 2.5 micres (PM2.5)			
Llindar per a la protecció de la salut humana	BOE	1 any civil	Llindar = 25 µg/m³
	OMS	1 any civil	Llindar = 5 µg/m³

Taula 4: Llindars per a la protecció de la salut humana en relació a les partícules PM10 i PM2.5.

3.1.2. Contaminants secundaris

Els contaminants secundaris són els que es produeixen com a conseqüència de reaccions químiques entre els contaminants primaris, entre sí i amb altres compostos.

Ozó (O_3)

L'ozó és un gas què es forma a la troposfera a partir de reaccions químiques on participen altres gasos contaminants.

L'O_3 té un efecte positiu en la estratosfera ja que protegeix de la radiació ultravioleta. Tanmateix, altes quantitats d'ozó poden ser molt perjudicials per a la salut i l'ecosistema. Així com, és un gas que contribueix en el calfament de l'atmosfera, ja que és d'efecte hivernacle [16].

		Període	**Valor**
Ozó			
Llindar per a la protecció de la salut humana	BOE	1 hora	Llindar d'informació = 180 µg/m³
			Llindar d'alerta = 240 µg/m³
		24 hores	Llindar = 120 µg/m³
	OMS	8 hores	Llindar = 100 µg/m³

Taula 5: Llindars per a la protecció de la salut humana en relació a l'Ozó.

3.2. Software utilitzat

3.2.1. Pandas

Pandas és una eina ràpida, potent, flexible i fàcil d'utilitzar d'anàlisi i manipulació de dades de codi obert, construïda sobre el llenguatge de programació Python [17].

3.2.2. NumPy

NumPy és el paquet fonamental per a la informàtica científica a Python. És una biblioteca que proporciona un objecte de matriu multidimensional, diversos objectes derivats (com ara matrius i matrius emmascarades) i un assortiment de rutines per a operacions ràpides en matrius, incloses les matemàtiques, lògiques, manipulació de formes, ordenació, selecció, transformades discretes de Fourier, àlgebra lineal bàsica, operacions estadístiques bàsiques, simulació aleatòria i molt més [18].

3.2.3. Matplotlib

Matplotlib és una biblioteca de programari per a generar gràfiques a partir de dades contingudes en llistes, o vectors, en el llenguatge de programació Python i en la seva extensió matemàtica NumPy. Proporciona una API, pylab, dissenyada per ser similar a les funcions gràfiques de MATLAB [19].

3.2.4. Requests

Requests és una biblioteca HTTP elegant i senzilla per a Python, creada per a éssers humans. Per això, permet enviar sol·licituds HTTP/1.1 amb molta facilitat, sense necessitat d'afegir manualment cadenes de consulta als URL ni codificar amb formularis les dades POST [20].

3.2.5. Microsoft Excel

Microsoft Excel és un programa de full de càlcul desenvolupat per Microsoft per a Windows, macOS, Android i iOS.

Compta amb càlcul, eines gràfiques, taules dinàmiques i un llenguatge de programació macro anomenat Visual Basic per a aplicacions [21].

3.2.6. Dash i Plotly

Plotly Python és una llibreria de traçat interactiva i de codi obert que admet més de 40 tipus de gràfics únics que cobreixen una àmplia gamma de casos d'ús estadístics, financers, geogràfics, científics i tridimensionals. A més, permet als usuaris de Python crear visualitzacions interactives basades en web que es poden mostrar als quaderns de Jupyter,

desar-se en fitxers HTML autònoms o servir com a part d'aplicacions web creades per Python mitjançant Dash [22].

Dash Open Source, crea aplicacions de dades en Python pur, sense necessitat de JavaScript. Oferint resultats més ràpids i impactants en iniciatives d'intel·ligència artificial i ciència de dades [23].

3.2.7. Scikit-learn

Scikit-learn és una extensió del llenguatge Python en forma de biblioteca informàtica que agrega suport en l'àmbit de l'aprenentatge automàtic. Scikit-learn és de codi obert i disposa d'algoritmes de classificació estadística, regressió i clustering per a implementar màquines de vector de suport, random forests, gradient boosting, K-Means [24].

3.2.8. Keras i TensorFlow

TensorFlow facilita la creació de models d'aprenentatge automàtic per a ordinadors d'escriptori, dispositius mòbils, web i al núvol, sense importar si l'usuari és principiant o expert [25].

Keras és l'API d'alt nivell de TensorFlow per construir i entrenar models d'aprenentatge profund. S'utilitza per a la creació ràpida de prototips, la recerca d'avantguarda i en producció, amb tres avantatges clau: amigable a l'usuari; modular i configurable; fàcil d'estendre [26].

3.2.9. XGBoost

XGBoost és una biblioteca optimitzada per augmentar el gradient distribuït dissenyada per ser altament eficient, flexible i portàtil. Implementa algoritmes d'aprenentatge automàtic sota el marc Gradient Boosting [27].

3.2.10. Power BI

Power BI és un servei d'analítica empresarial de Microsoft. Té com a objectiu proporcionar visualitzacions interactives i capacitats d'intel·ligència empresarial amb una interfície prou senzilla perquè els usuaris finals puguen crear els seus propis informes i taulers [28].

3.3. Mètriques d'avaluació

Tots els models de regressió, s'avaluen en les següents mètriques:

3.3.1. Error Quadràtic Mitjà (MSE)

L'Error Quadràtic Mitjà és una mètrica utilitzada per mesurar la qualitat d'un model de regressió. Es calcula prenent la mitjana dels errors quadrats entre els valors predits pel model i els valors reals del conjunt de dades [29].

$$\mathrm{MSE}(y, \hat{y}) = \frac{1}{n_{\mathrm{mostres}}} \sum_{i=0}^{n_{\mathrm{mostres}}-1} (y_i - \hat{y}_i)^2.$$

3.3.2. Arrel de l'Error Quadràtic Mitjà (RMSE)

L'RMSE és una mesura utilitzada comunament en Estadística i Machine Learning per avaluar la precisió d'un model de regressió. Es calcula prenent l'arrel quadrada de l'MSE, fet que permet interpretar la mètrica a la mateixa escala que els valors originals. L'RMSE proporciona una mesura de la diferència mitjana entre els valors predits i els valors reals del conjunt de dades [30].

$$\mathrm{RMSE}(y, \hat{y}) = \sqrt{\mathrm{MSE}}$$

3.3.3. Error Absolut Mitjà (MAE)

L'Error Absolut Mitjà és una mètrica per mesurar la discrepància mitjana entre els valors predits i els valors reals en un conjunt de dades. A diferència de l'MSE, el MAE no dona més pes als errors més grans i és més robust davant de valors atípics [31].

$$\mathrm{MAE}(y, \hat{y}) = \frac{1}{n_{\mathrm{mostres}}} \sum_{i=0}^{n_{\mathrm{mostres}}-1} |y_i - \hat{y}_i| \,.$$

3.3.4. Error Mitjà Absolut Percentual (MAPE)

L'Error Mitjà Absolut Percentual és una mètrica utilitzada per avaluar la precisió d'un model de regressió. Es calcula prenent la mitjana dels valors absoluts dels errors percentuals individuals, que són la diferència entre els valors predits i els valors reals, dividida pels valors reals. El MAPE permet avaluar el rendiment relatiu del model a diferents escales de dades [32].

$$\mathrm{MAPE}(y, \hat{y}) = \frac{1}{n_{\mathrm{mostres}}} \sum_{i=0}^{n_{\mathrm{mostres}}-1} \frac{|y_i - \hat{y}_i|}{\max(\epsilon, |y_i|)}$$

3.3.5. Coeficient de Determinació (R^2)

El Coeficient de Determinació, també conegut com a R quadrat, és una mesura estadística utilitzada per avaluar la qualitat d'un model de regressió. El R quadrat indica la proporció de la variància de la variable dependent que pot ser explicada pel model. Té un valor entre 0 i 1, on 1 indica que el model explica tota la variància i 0 indica que el model no explica cap variància [33]. És interessant destacar que aquesta mètrica pot ser negativa per als casos en els quals el rang de les variables és molt prop a zero.

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

4. Descripció del conjunt de dades

4.1. Font de dades

En aquest estudi s'han emprat 3 conjunts de dades horàries sobre la qualitat de l'aire a la ciutat de València que han sigut recopilades a partir de la Xarxa Valenciana de Vigilància i Control de la contaminació atmosfèrica.

Aquesta xarxa compta amb diversos punts de mesurament, els quals s'ubiquen en 78 estacions de mesurament al llarg de les tres províncies de la Comunitat Valenciana. Aquests punts de mesurament, fan un seguiment continu dels nivells de diferents paràmetres ambientals i meteorològics: PM1, PM2.5, PM10, NO, NO_2, NO_X, O_3, SO_2, CO, NH_3, C_7H_8, C_6H_6, C_8H_{10}, soroll, velocitat del vent, direcció del vent, temperatura, humitat relativa, pressió, radiació solar, precipitacions, velocitat màxima del vent, arsènic, níquel, cadmi i plom.

Els dos primers conjunts s'han obtingut del portal de dades obertes de l'Ajuntament de València. Aquests conjunts de dades contenen informació recopilada per les distintes estacions de mesurament de la ciutat de València des de l'any 2018 fins a l'any 2021.

A la Taula 6 es pot observar la localització de les 13 estacions de mesurament ubicades a la ciutat de València i els contaminants que capta segons la Xarxa Valenciana de Vigilància i Control de la Contaminació Atmosfèrica [34].

Estació	Codi	Emplaçament	Latitud (en graus decimals DD)	Longitud (en graus decimals DD)	CONTAMINANTS ATMOSFÉRICS							
					SO2	CO	NO	NO2	NOx	O3	PM2.5	PM10
València - Avd. França	46250047	Avinguda de França,60	39.45750439L	-0.3426899	X	X	X	X	X	X	X	X
València - Bulevard Sud	46250050	Bulevar Sur s/n (Parking cementiri de València)	39.45037852	-0.39631399	X		X	X	X	X		
València - Centre	46250054	Plaça de l'Ajuntament	39.47071883	-0.37648469			X	X	X		X	X
València - Molí del Sol	46250048	Avinguda Pio Baroja, s/n	39.48113875	-0.40855865	X	X	X	X	X	X	X	X
València - Nazaret Met-2	46250030	Placa Aras de alpuente	39.44855408	-0.33328931								
València - Pista de Silla	46250030	C/ Filipines s/n front nº 37	39.45806013	-0.37665323	X	X	X	X	X	X	X	X
València - Politècnic	46250046	Cami de Vera, s/n	39.47962193	-0.3374074	X		X	X	X	X	X	X
València - Vivers	46250043	Jardins de Vivers	39.47948825	-0.36955032	X		X	X	X	X		
València Olivereta	46250055	Avinguda Cid / Avinguda Tres Creus	39.46923859	-0.40603766			X	X	X	X		X
València Port llit antic Túria	46250302	Recinte Portuari antic llit del riu Turia	39.45051894	-0.32894501	X	X	X	X	X	X	X	X
València Port Moll Trans. Ponent	46250301	València Port Moll Trans. Ponent	39.45926486	-0.32321741	X	X	X	X	X	X	X	X
València-Conselleria Meteo	46250049	C/ Castan Tobeñas 77, Ciutat Adinistrativa 9 d'octubre	39.4720341	-0.40487846	X	X	X	X	X	X	X	X

Taula 6: Estacions de mesurament de la ciutat de València i contaminants que capta (Xarxa Valenciana de Vigilància i Control de la Contaminació Atmosfèrica). Elaborada a partir de les dades de la Conselleria d'Agricultura, Desenvolupament Rural, Emergència Climàtica i Transició Ecològica.

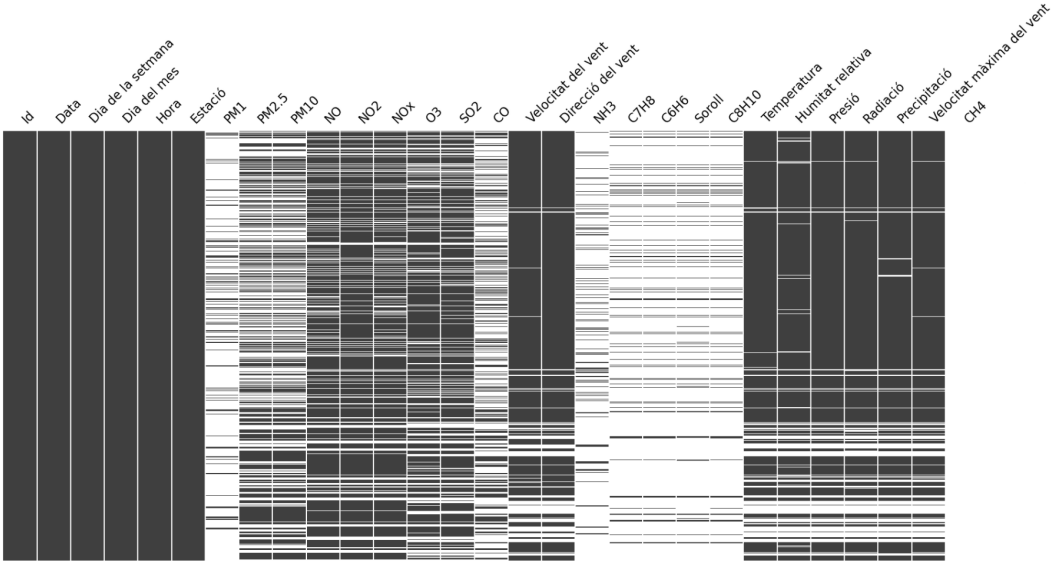
Les dues primeres bases de dades es van descarregar del portal de dades obertes de l'Ajuntament de València, la primera recopilava dades entre els períodes compresos entre el 2016 i 2020 [35] i la segona des del 2021 fins a 2022 [36]. Tanmateix, la tercera base de dades prové de la pàgina web de la Conselleria d'Agricultura, Desenvolupament Rural, Emergència Climàtica i Transició Ecològica de la GVA (Agroambient GVA) [37]. Per a obtindre les dades de totes les estacions i tots els anys (2017-2022) es va crear un script de Python per automatitzar la descàrrega de dades, mitjançant tècniques de web scraping utilitzant la llibreria Requests de Python, i també per a unificar tots el fitxers de text, en un sol.

4.2. Descripció del conjunt de dades

Inicialment, es va utilitzar com a base els conjunts de dades del portal de dades obertes de l'Ajuntament de València. Aquest conjunt de dades compta amb un total de 352666 registres i 32 variables. De les quals capten registres horaris durant els períodes compresos entre el 2017 al 2022 de 13 estacions de mesurament de València.

Tanmateix, el número de dades faltants en general era prou elevat en determinades variables. Per tant, per completar la informació faltant i reduir el percentatge de dades faltants del conjunt de dades, es va emprar el tercer conjunt de dades, procedent d'Agroambient GVA.

El qual compta amb 521373 registres, 26 variables i registra dades horàries durant els anys 2017 i 2022 per a les mateixes estacions, excepte Port València, la qual no registra.



Taula 7: Matriu de dades faltants després de la unió dels datasets.

D'aquesta manera, el conjunt de dades resultant de la unió dels tres esmentats anteriorment compta amb 554146 registres i 32 variables.

A pesar de no haver una disminució considerable dels valors faltants per a les variables, sí s'observa una lleugera reducció en el percentatge de valors faltants després d'unir ambdós conjunts de dades, veure Taula 8.

Variable	Percentatge de dades faltants dataset inicial (%)	Percentatge de dades faltants dataset final (%)	Reducció de faltants (%)
Id	0	0	0
Data	0	0	0
Dia de la setmana	0	0	0
Dia del mes	0	0	0
Hora	0	0	0

Estació	0	0	0
PM1	85.03	84.96	0.07
PM2.5	47.67	47.03	0.64
PM10	47.67	47.09	0.58
NO	24.09	23.58	0.51
NO_2	20.96	20.44	0.52
NO_X	24.09	23.58	0.51
O_3	29.05	28.49	0.56
SO_2	29.23	28.78	0.45
CO	65.34	64.78	0.56
Velocitat del vent	14.08	13.20	0.88
Direcció del vent	13.58	13.20	0.38
NH_3	89.85	89.79	0.06
C_7H_8	90.02	89.86	0.16
C_6H_6	90.09	89.95	0.14
Soroll	89.50	89.36	0.14
C_8H_{10}	89.98	89.85	0.13
Temperatura	15.84	15.57	0.27
Humitat relativa	18.64	18.24	0.40
Pressió	15.21	14.94	0.27
Radiació	15.66	15.65	0.01
Precipitació	15.33	15.08	0.25
Velocitat máxima del vent	15.6	15.31	0.29
Data de creació	0	0	0
Data de baixa	100	100	0
CH_4	–*	99.998	99.998

R.Sol.	–*	81.73	81.73

***Variable no registrada**

Taula 8: Percentatge de valors faltants abans i després de la unió dels conjunts de dades.

4.3. Pre-processament de dades

4.3.1. Selecció de variables

Tenint en compte que pràcticament tots els paràmetres atmosfèrics i meteorològics tenen dades faltants. S'ha establert un llindar màxim del 65% de dades faltants per tenir en compte eixa variable en les anàlisis ja que el seu estudi no és representatiu. Per això, s'han eliminat les variables que superen eixe llindar. Així com també s'han eliminat les variables que no són importants per a l'estudi, amb l'objectiu de centrar-se únicament en aquelles que tenen un impacte significatiu en la investigació.

A la Taula 9 es mostren les variables que s'han seleccionat, en funció del seu impacte en la salut humana i en l'ecosistema. Així com per les àmplies regulacions i estàndards per al seu monitoreig.

Data	Estació	NO_2	CO
Dia de la setmana	PM2.5	NO_X	Temperatura
Dia del mes	PM10	O_3	Humitat relativa
Hora	NO	SO_2	Precipitació

Taula 9: Llistat de variables d'estudi.

4.3.2. Creació de noves variables

Es van crear dues noves variables: *Mes* i *Any*, a partir de la columna *Data*, la qual conté informació sobre el dia, mes i any en què es capten els paràmetres. La generació d'aquestes dues noves columnes es va fer de manera senzilla i eficient emprant la llibreria *Datetime* de *Python.*

5. Anàlisi exploratori de dades

5.1. Anàlisi descriptiu de les variables

Primer, es va realitzar una anàlisi descriptiva de les variables d'estudi per així conèixer la naturalesa de les dades, la seua tendència, dispersió i distribució, excloent valors faltants.

Com podem veure a la Taula 10, les variables CO i les Partícules de Suspensió (PM2.5, PM10) són les variables amb menor recompte de dades.

Quant a la desviació estàndard de les variables, en general, els valors són elevats, la qual cosa indica que les dades estan més disperses al voltant de la mitjana. Excepte per al CO, el qual té valors propers a 0, la qual cosa significa que són més similars entre sí.

A més, cal destacar que per a totes les variables hi ha valors atípics, ja que la comparativa entre el valors màxims de cada variable respecte a la mitjana és alarmant. Sobretot, per a la Precipitació ja que els tres primers percentils són 0, amb una mitjana de 0.04 i després té un valor màxim de 317. Aquests valors atípics tan elevats tenen sentit si ho connectem amb la desviació estàndard tant distant al 0, en la majoria de les variables.

	PM2.5	PM10	NO	NO2	NOx	O3	SO2	CO	Temperatura	Humitat Relativa	Precipitació
Total	295391	295304	425952	443307	425953	399192	398227	197471	468355	454020	470972
Mitjana	11.69	19.13	9.93	23.71	38.55	51.36	3.69	0.159	18.42	65.44	0.04
Desviació	9.70	16.98	20.55	20.41	47.39	28.22	1.62	0.09	6.26	19.12	0.89
Mínim	0.0	0.0	1.0	0.0	2.0	0.0	0.0	0.1	0.7	11.0	0.0
25%	5.0	9.0	2.0	8.0	11.0	29.0	3.0	0.1	13.5	52.0	0.0
50%	9.0	15.0	3.0	18.0	23.0	53.0	3.0	0.1	18.2	66.0	0.0
75%	15.0	25.0	8.0	33.0	47.0	73.0	4.0	0.2	23.5	80.0	0.0
Màxim	309.0	460.0	532.0	229.0	987.0	172.0	167.0	9.0	201.0	107.0	317.0

Taula 10: Percentatge de valors faltants abans i després de la unió dels conjunts de dades.

5.2. Estudi de faltants per estació

Com s'ha pogut veure, aquest conjunt de dades compta amb un número elevat de dades faltants, en general. Per això, s'ha fet una anàlisi exhaustiva de totes les variables per a cada una de les estacions de mesurament, per a cada mes comprés entre els anys 2018 al 2022.

Aquesta anàlisi, s'ha fet amb la finalitat de seleccionar períodes temporals per a cada estació i variable. De manera que, s'ha establert un llindar màxim del 20% de valors faltant per a cada

mes de cada any. Aquesta decisió s'ha pres amb l'objectiu de garantir la fiabilitat de les dades emprades.

En les 11 estacions disponibles, hem vist alguns casos pels quals en totes les variables hi ha més d'un 20% de dades faltants, com és el cas de Nazaret Meteo, València Olivereta, Port de València i Conselleria Meteo, aquesta última sols capta CO, veure Figura 1.

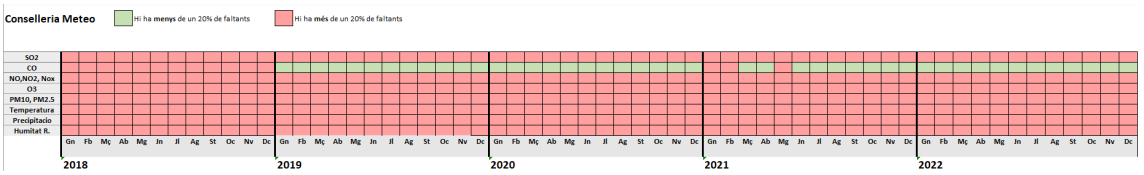


Figura 1: Percentatge de dades faltants per mes i any per a l'estació Conselleria Meteo.

Tanmateix, també hi ha casos pels quals totes les variables tenen registres mensuals on hi ha menys d'un 20% de dades faltants, com és el cas de les estacions Avinguda França, veure Figura 2, i Molí del Sol, aquesta última amb registres des del 2019 al 2022.

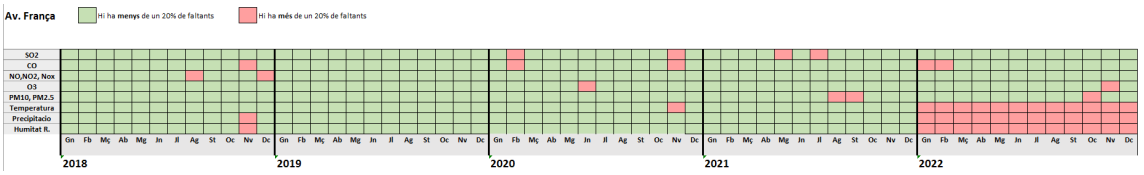


Figura 2: Percentatge de dades faltants per mes i any per a l'estació Avinguda França.

Per al Port Antic Llit Túria i Port Moll Trans. Ponent és capten totes les variables però sols durant els anys 2021 i 2022. Cosa que dona a entendre que són dues estacions molt recents, veure Figura 3.

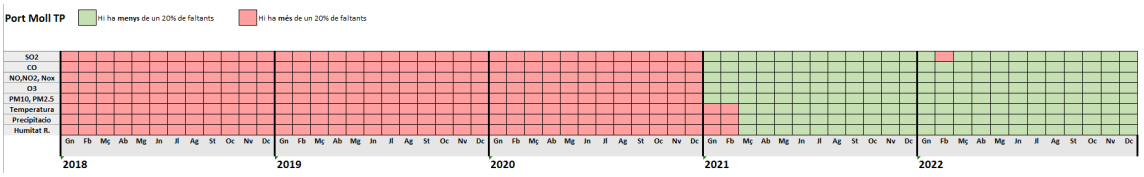


Figura 3: Percentatge de dades faltants per mes i any per a l'estació Port Moll Trans. Ponent.

També hi ha estacions per les quals els sensors no capten determinades variables com és el cas de Bulvedar Sud, que no capta CO, PM2.5 ni PM10, veure Figura 4. L'estació situada al centre no capta SO_2, CO ni O_3. El politècnic no capta la variable CO i l'estació Pista de Silla no capta cap òxid de nitrogen (NO, NO_2, NO_X). Per últim, l'estació Vivers no registra dades de CO ni de les partícules de suspensió (PM2.5 i PM10).

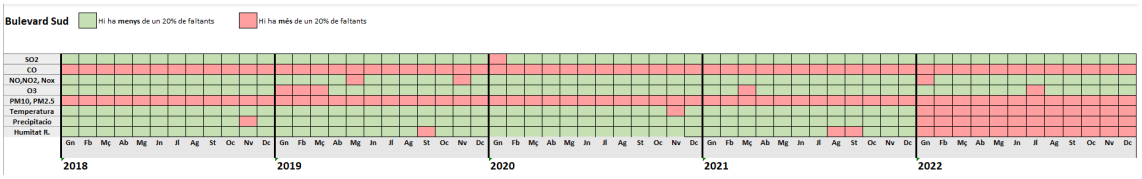


Figura 4: Percentatge de dades faltants per mes i any per a la estació Bulevard Sud.

A l'Annex II es troben la resta de figures per als períodes temporals en les altres estacions.

Cal destacar que durant l'any 2022 per a moltes estacions ha hagut una falla ens els sensors i no s'han pogut captar les variables Temperatura, Humitat Relativa ni Precipitació. Tanmateix, per a les estacions més recents, Port Llit Antic Turia i Port Moll Trans. Ponent no ha hagut cap error a l'hora de recol.lectar les dades.

5.3. Períodes d'anàlisi

Com s'ha comentat anteriorment, un dels objectius és predir les dades faltants. Però sols podem predir les dades faltants de cada estació en determinats períodes temporals.

Per això, les estacions Conselleria Meteo, Nazaret Meteo, València Olivereta i Port de València no formen part de l'anàlisi ja que no recopilen el percentatge de dades suficient per aportar informació rellevant a l'estudi. Així com tampoc té sentit predir variables per a una estació on la majoria de dades són faltants.

Els períodes temporals seleccionats han sigut tots els mesos de cada any per als quals no hi ha dades faltats de la variable explicada, ni tampoc per a les variables d'entrada. Dins d'aquest context quan es refereix a que no hi ha dades faltants per a cap variable, s'han d'exceptuar les variables que en tota una estació els sensors no capten. Com és el cas de Bulevard Sud que no capta CO ni les partícules de suspensió de diàmetre 2.5 micres ni les de diàmetre 10 micres, ja que sinó s'eliminaria una estació d'estudi.

5.4. Correlacions entre variables

Per avaluar les relacions lineals i no lineals entre les variables, s'ha realitzar un estudi de correlacions entre variables segons l'Índex de correlació de Pearson.

S'observa un comportament semblant entre la correlació de les variables i les 11 estacions de mesuració, a pesar d'estar ubicades en diferents localitzacions de la ciutat de València. Així com, també s'observen uns patrons de correlació similars que es mantenen durant els diferents mesos i anys. Açò indica que encara que les condicions atmosfèriques varien estacionalment, les interaccions d'aquests contaminants i paràmetres atmosfèrics són robustes, en termes generals.

D'aquesta manera, per comentar les distintes relacions entre variables es comenta l'estació Molí del Sol per al mes de febrer de 2019, veure en Annex III les correlacions per als anys 2020, 2021 i 2022 d'aquesta estació.

S'observa en la Figura 5 una clara relació positiva entre les variables PM2.5 i PM10. A més, s'observa un altra clara relació positiva entre les variables NO i NO_X.

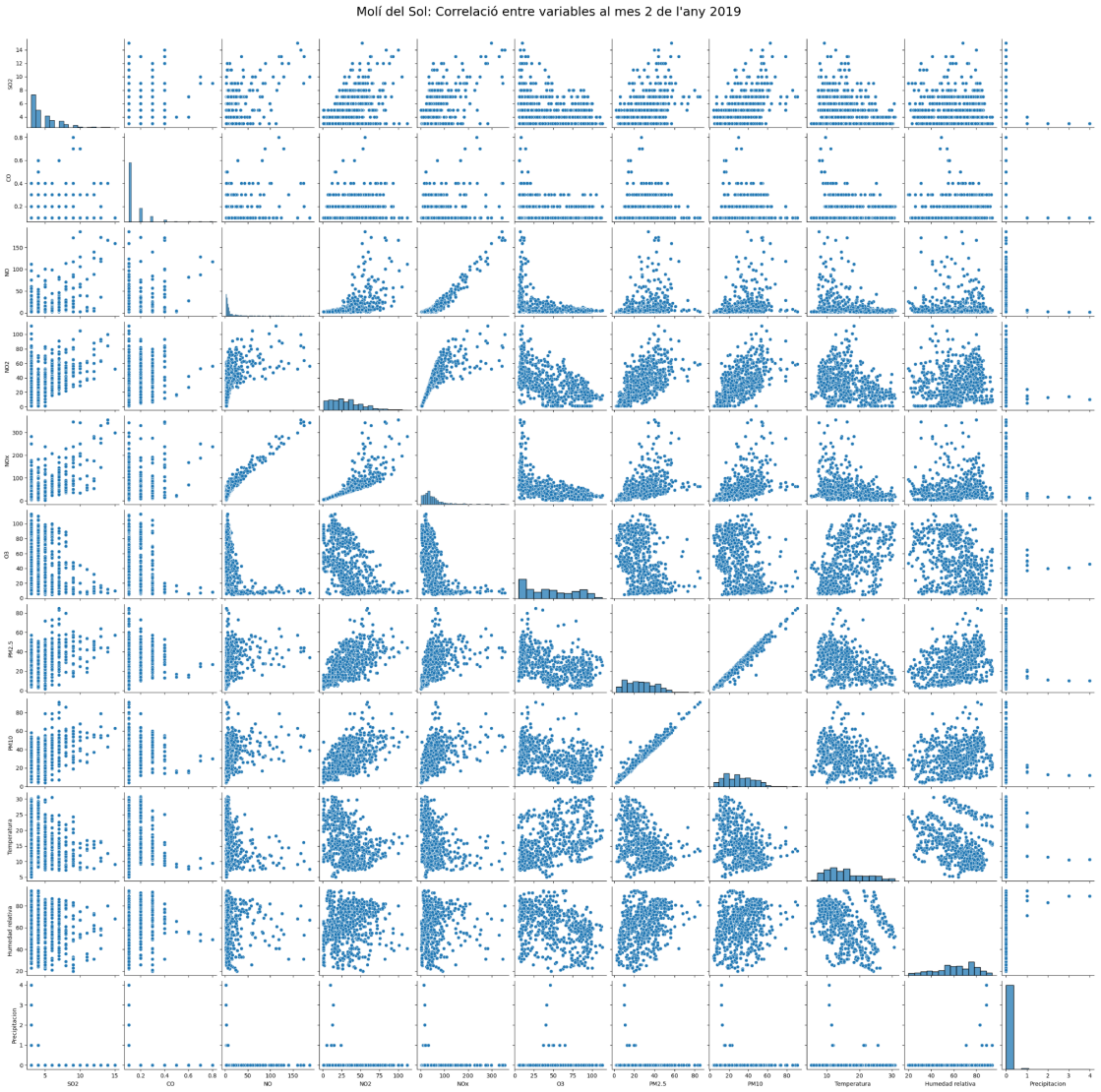


Figura 5: Gràfica Pairwise, relacions entre les parelles de variables per a febrer de 2019.

Com s'ha pogut veure a les gràfiques, no totes les variables tenen correlacions clares. D'aquesta manera, és interessant estudiar-ho amb el coeficient de correlació de Pearson, el qual estudia el grau d'associació lineal entre dues variables sense tenir en compte l'escala de mesura d'aquestes, prenent valors entre -1 i 1. A partir d'una correlació major que 0.4, es diu que existeix una correlació baixa i a partir de 0.7 una correlació forta [38].

A la Taula 4, s'observa que les partícules de suspensió PM2.5 i PM10 estan estretament relacionades ja que existeix una forta correlació positiva entre les dues. D'aquesta manera, que l'augment d'una d'elles generalment, va acompanyada de l'augment de l'altra. Açò suggereix que ambdós paràmetres es veuen influenciats per factors o condicions atmosfèriques similars.

A més, s'ha identificat una correlació forta entre l'NO_X i l'NO_2. També s'observa una correlació forta entre l'NO i l'NO_X, la qual cosa indica una relació significativa entre les variables.

Per l'altra banda, s'ha trobat una correlació baixa negativa entre l'ozó i els òxids de nitrogen (NO, NO_2, NO_X). La qual cosa implica que un augment dels nivells dels òxids de nitrogen, pot estar associada a una disminució dels nivells d'O_3.

Addicionalment, s'observa una correlació baixa positiva entre la temperatura i l'ozó. Açò suggereix que conforme augmenta la temperatura, els nivells d'O_3 també tendeixen a augmentar, la qual cosa indica que la temperatura influeix en la formació i estabilitat de l'O_3.

Per últim, s'ha identificat una correlació baixa positiva entre el CO i l'NO_2. Així mateix, també existeix una correlació baixa negativa entre el CO i NO_X.

Aquestes correlacions entre parelles de variables, es poden veure influenciades per factors ambientals i reaccions químiques específiques.

	PM2.5	PM10	NO	NO2	NOx	O3	SO2	CO	Temperatura	Humitat Relativa	Precipitació
PM2.5	1.0	0.7491	0.2658	0.3316	0.317	-0.3208	0.1271	0.2199	-0.1453	0.2522	-0.0156
PM10	0.7491	1.0	0.3522	0.3861	0.3974	-0.243	0.1367	0.1855	-0.0428	0.0198	-0.0199
NO	0.2658	0.3522	1.0	0.6586	0.9466	-0.4083	0.3749	0.4093	-0.185	-0.0323	-0.0123
NO2	0.3316	0.3861	0.6586	1.0	0.8657	-0.5804	0.2481	0.3735	-0.1995	-0.0518	-0.0182
NOx	0.317	0.3974	0.9466	0.8657	1.0	-0.5258	0.3554	0.4291	-0.2108	-0.0404	-0.0159
O3	-0.3208	-0.243	-0.4083	-0.5804	-0.5258	1.0	-0.1718	-0.2811	0.4043	-0.1547	0.0146
SO2	0.1271	0.1367	0.3749	0.2481	0.3554	-0.1718	1.0	0.203	-0.0912	-0.0656	-0.012
CO	0.2199	0.1855	0.4093	0.3735	0.4291	-0.2811	0.203	1.0	-0.1733	-0.0175	-0.0105
Temperatura	-0.1453	-0.0428	-0.185	-0.1995	-0.2108	0.4043	-0.0912	-0.1733	1.0	-0.0545	-0.0299
Humitat Relativa	0.2522	0.0198	-0.0323	-0.0518	-0.0404	-0.1547	-0.0656	-0.0175	-0.0545	1.0	0.0471
Precipitació	-0.0156	-0.0199	-0.0123	-0.0182	-0.0159	0.0146	-0.012	-0.0105	-0.0299	0.0471	1.0

Taula 11: Matriu de correlació de Pearson.

6. Representació de l'anàlisi de dades

Després de l'anàlisi exploratori de dades, s'ha creat un dashboard amb l'objectiu d'estudiar de manera interactiva i detallada tres mesures diferents per a cada estació de mesurament: percentatge de faltants, variància i mitjana aritmètica, el qual s'ha creat amb Dash de Plotly. A més a més, s'ha utilitzat codi CSS per a l'estructuració i composició de l'aplicació i el seu estil.

Per a facilitar la comprensió i l'anàlisi de dades, el dashboard ofereix diverses funcionalitats interactives. Una d'elles és el selector de dades, més conegut com datepicker, el qual permet escollir un mes i any concret per a la visualització. Després, s'ha implementat un *dropdown* o menú desplegable on es llisten els diferents contaminants, la qual cosa proporciona flexibilitat a l'usuari, ja que es pot centrar en eixa variable en particular. Per últim, es pot elegir el tipus de mesura desitjada.

El dashboard compta amb un *datepicker*, el qual permet seleccionar una data, és a dir, mes i any concret. A continuació, es pot seleccionar el contaminant que es vol estudiar i, per últim, es pot elegir el tipus de mesura desitjada.

El dashboard presenta un scatter mapbox de la ciutat de València, amb les seues respectives estacions de mesurament, veure Figura 6. En aquest cas, es representa el percentatge de dades faltants. Per facilitar la comprensió a l'usuari, s'utilitza una escala de colors per a resaltar visualment els diferents rangs de valors faltants.

- Verd fosc: representa un percentatge de dades faltants menor o igual al 5%.
- Verd clar: indica un percentatge de dades faltants menor o igual al 10%.
- Groc: denota un percentatge de dades faltants menor o igual al 20%.
- Taronja: senyala un percentatge de dades faltants menor o igual al 30%.
- Vermell: representa un percentatge de faltants major al 30%.

Per últim, s'ha customitzat el *hover info*, és a dir, quan es passa el cursor per damunt de cada bombolla es pot veure el percentatge de faltants en eixa determinada estació de mesurament.

En la Figura 6, es pot veure de forma ràpida i intuitiva les variables disponibles i el percentatge de dades faltants concret per a eixa variable, estació i moment temporal.

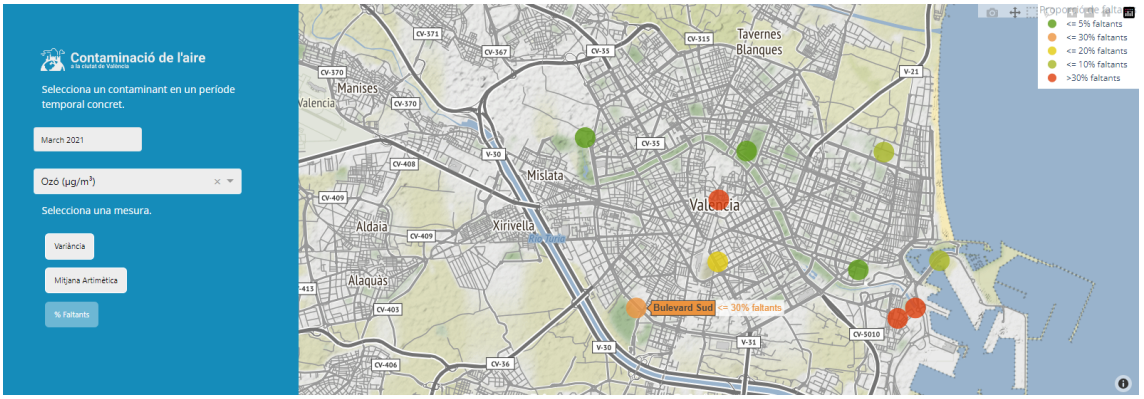


Figura 6: Dashboard, percentatge de dades faltants per a l'O3 a març de 2021.

Quant a la variància dels distints contaminants en determinats períodes temporals, el dashboard és molt similar al de les dades faltants, veure Figura 7. Tanmateix, en aquest el tamany de la bombolla fa referència a la variància, és a dir, major tamany de la bombolla significa que la variància és major, cosa que significa que per a eixa determinada variable hi haurà una major variabilitat respecte a la mitjana. A més, canvia el *hover info*, perquè en aquest cas al passar per damunt el cursor, es pot observar l'estació i la seua variància per a la variable seleccionada.

D'aquesta manera, la variància és una mesura estadística que proporciona informació sobre la dispersió de les dades i amb aquesta representació visual és molt fàcil identificar patrons, tendències i comportaments atípics entre estacions en el mateix període temporal.

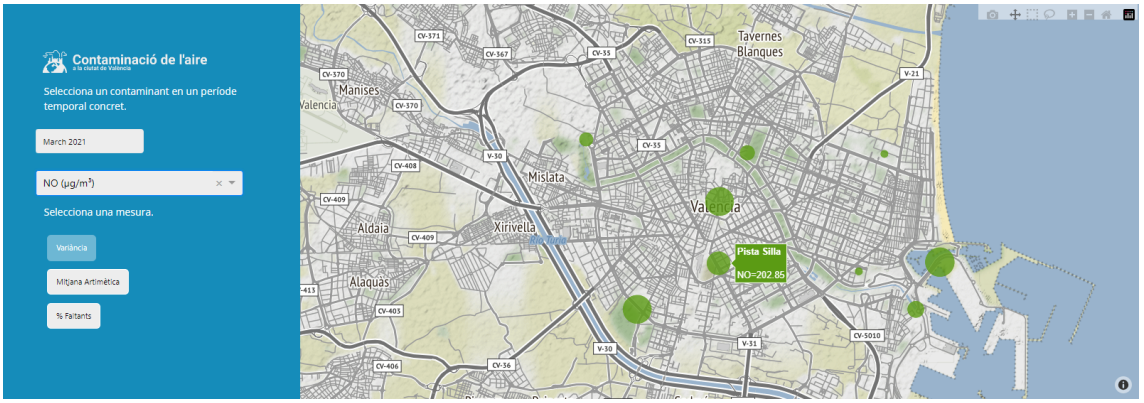


Figura 7: Dashboard, variància per a l'NO a març de 2021.

Per últim, la Figura 8 mostra les diferents mitjanes aritmètiques per a cada estació, el tamany de la bombolla varia en funció dels valors de la mitjana i el *hover info*, mostra l'estació i la seua mitjana aritmètica per a la variable seleccionada.

La mitjana proporciona una visió general dels valors mitjans dels contaminants, facilitant la detecció de patrons estacionals, canvis a llarg termini o estudiar diferències significatives entre les estacions de mesurament.

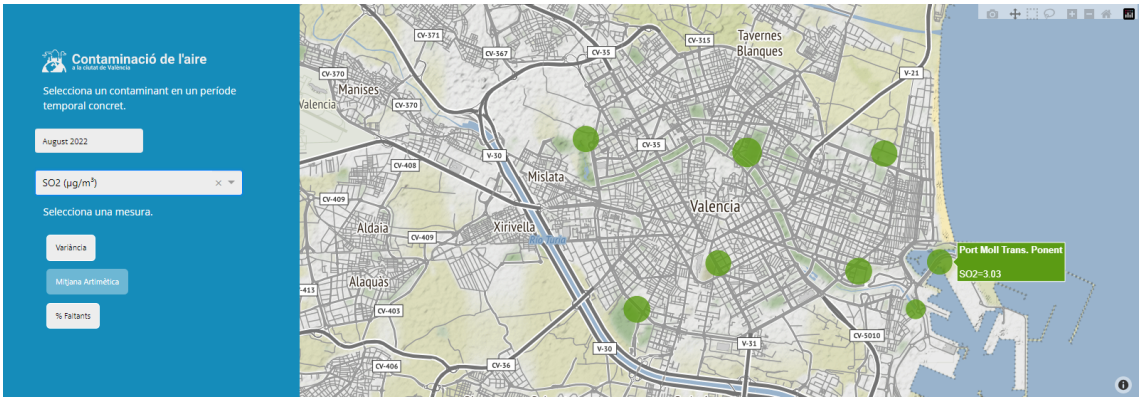


Figura 8: Dashboard, mitjana aritmètica per a l'SO_2 a agost de 2022.

Mitjançant la combinació de la selecció de períodes temporals, variables i tipus de mesura, els usuaris del dashboard poden explorar de forma interactiva les dades de qualitat de l'aire de València, permetent identificar patrons, tendències i anomalies en la disponibilitat i qualitat de les dades. Permentent una eina ràpida i senzilla per estudiar les diferències entre estacions dins d'un mateix període temporal per a una variable concreta, proporcionant una visió realista amb la ubicació de les distintes estacions de mesurament al llarg de València.

Tanmateix, s'ha de tenir en compte que aquesta representació sols és útil per a estudiar el comportament de les variables en les distintes estacions de mesurament per a moments temporals concrets. Per això, s'ha estudiat la variància i mitjana aritmètica per a cada una de les estacions.

A la Taula 12, s'observa de manera general que la variància de cada contaminant és similar entre les distintes variables. Però la variància en totes les estacions és prou elevada, la qual cosa indica que existeix una gran variabilitat entre les variables estudiades. Excepte per al CO, el qual té uns valors molt propers a zero, el que suggereix que els valors són molt propers entre sí.

És important destacar que per a l'estació Pista Silla es mostra una variabilitat moderadament major respecte a la resta de variables. Esta discrepància pot deure's a la ubicació geogràfica de de l'estació, ja que aquesta estació està ubicada en una autovia per accedir a la ciutat de València, en la qual hi ha molt trànsit.

Estació	**SO_2**	**CO**	**NO**	**NO_2**	**NO_X**	**O_3**	**PM2.5**	**PM10**
Avda.	1.28	0.14	17.85	20.16	42.71	26.30	8.53	16.31

França								
Bulevard Sud	1.61	- *	24.90	21.80	54.63	29.66	- *	-*
Centre	- *	-*	20.06	19.24	45.70	- *	10.67	16.65
Molí del Sol	1.30	0.08	15.31	15.67	35.22	51.60	11.70	12.88
Pista Silla	2.02	0.11	25.80	22.58	58.19	27.04	9.38	22.84
Politècnic	1.32	- *	11.91	17.07	31.70	28.53	8.57	12.19
Port Antic Llit Turia	1.60	0.13	21.60	19.79	49.13	27.19	8.07	14.85
Port Moll Tras Ponent	2.11	0.09	24.55	21.08	54.42	28.18	7.78	17.39
Vivers	1.26	- *	16.91	18.75	40.41	29.17	- *	-*

***Variable no registrada**

Taula 12: Variància de cada contaminant per estació de mesurament.

Respecte a la mitjana aritmètica, les estacions també presenten valors similars entre les diferents estacions. Però cal destacar que els òxids de nitrogen (NO, NO_2 i NO_X) per a les estacions Molí del Sol, Politècnic i Vivers tenen una mitjana considerablement menor en comparació amb la resta d'estacions. Com aquestes estacions estan ubicades en zones amb poc trànsit directe, els nivells d'òxid de nitrogen són menors, ja que s'ubiquen en un entorn universitari o àrees ajardinades on els vehicles tenen el pas restringit, encara que això no significa que no hi ha trànsit a les proximitats dels sensors de mesurament.

Estació	SO_2	CO	NO	NO_2	NO_X	O_3	PM2.5	PM10
Avda. França	3.73	0.14	17.85	20.16	42.71	51.83	8.53	16.31
Bulevard Sud	4.05	- *	12.28	28.17	46.82	50.38	- *	-*
Centre	- *	-*	12.81	24.78	44.22	- *	12.76	21.97
Molí del Sol	3.63	0.15	6.97	18.73	29.22	51.60	14.60	16.74
Pista Silla	4.03	0.16	25.80	22.58	54.77	46.52	11.04	20.76

Politècnic	3.57	- *	4.53	16.18	22.86	55.73	10.99	16.03
Port Antic Llit Turia	2.20	0.17	12.85	27.37	46.98	52.82	11.05	21.49
Port Moll Tras Ponent	3.01	0.14	13.53	27.32	47.86	57.71	10.62	22.47
Vivers	1.27	- *	7.27	21.50	32.47	53.05	- *	-*

***Variable no registrada**

Taula 13: Mitjana aritmètica per a cada estació.

7. Reconstrucció de dades a partir de l'aprenentatge automàtic

Com s'ha comentat anteriorment, un dels grans reptes ha sigut les dades faltants. És per això, que s'ha fet una reconstrucció de dades a partir de distints models d'aprenentatge automàtic.

L'objectiu és automatitzar l'entrenament per a cada variable i estació, amb un sol regressor per als distints períodes temporals i variables que estan disponibles. Així, es podrà entrenar el regressor quan les variables estiguen completes, i després aquest reconstruirà la variable objectiu quan no estiga disponible en un determinat període temporal.

7.1. Models de regressió

El problema de regressió és fonamental en l'estadística i machine learning. En estudis de regressió, s'infereix una función amb valors continus, la qual correspon a la mitjana d'una variable dependent (denominada variable resposta o variable d'eixida) condicionada a una o més variables independents (o variables d'entrada) [39].

D'aquesta manera, l'objectiu dels models és poder fer prediccions o inferències sobre els valors de la variable depenent a partir de les variables independents, utilitzant diferents tècniques en funció de la naturalesa i característiques de les dades.

7.2. Estandarització

Pel que fa a l'estadística, estandaritzar les variables és el procés d'ajustar els valors a un rang determinat i posar totes les variables a la mateixa escala, és a dir, els valors tindran la mitjana centrada a 0 i norma unitària.

La següent fórmula calcula el valor estandaritzat per a una mostra x, on μ és la mitjana de les mostres d'entrenament i σ és la desviació estàndard.

$$z = \frac{x - \mu}{\sigma}$$

Aquesta tècnica és molt comuna en els estimadors de ML perquè si les característiques individuals no varíen en rangs similars aquests estimadors poden tenir un mal rendiment [40], ja que alguna variable pot dominar més que altra per culpa d'una alta variància.

Per això, en aquest projecte s'han estandaritzat les variables d'entrada per a utilitzar-les en el model de regressió. Després, una vegada entrenat el model i obtingudes les prediccions, s'ha realitzat una normalització inversa a la variable d'eixida perquè tornen a tindre el rang corresponent.

7.3. Expansió polinòmica

Els models de regressió més simples, com un model de regressió lineal, assumeixen que la relació entre les variables d'entrada i la variable d'eixida és lineal. Tanmateix, hi ha casos en els quals les variables no tenen una relació lineal entre elles.

Així doncs, per afegir complexitat al model considerant les característiques no lineals de les variables d'entrada, s'utilitza l'expansió polinòmica. Aquesta tècnica consisteix en crear característiques noves a partir de les originals, a partir de les combinacions polinomials de les característiques en funció del grau especificat. En aquest projecte s'ha considerat un grau màxim d'aquesta manera, si s'observa una mostra d'entrada bidimensional de la forma (X_1, X_2), les característiques polinomials de graus 2 són: $(1, X_1, X_2, X_1^2, X_1X_2, X_2^2)$. Dins del llenguatge de programació de models de machine learning aquest hiperparàmetre és conegut com *Interaction_only=False*.

En el cas anterior, es consideren tant els termes d'ordre superior com les interaccions entre les característiques. Tanmateix, es pot especificar una expansió polinòmica sols amb les interaccions, excloent els termes amb potència superior de 2 o superiors a la del grau especificat, coneguda com *Interaction_only=True*. Amb l'exemple esmentat anteriorment seria: $(1, X_1, X_2, X_1X_2)$ [41].

En aquest projecte després de realitzar l'estandarització s'ha aplicat l'expansió polinòmica, per afegir-li complexitat al model.

7.4. Models emprats

Respecte als models d'aprenentatge automàtic utilitzats, aquests van ser elegits per la seva popularitat i eficàcia en problemes de regressió, amb l'objectiu d'avaluar tant tècniques més senzilles com un model de regressió lineal fins a models complexos, com les xarxes neuronals.

7.4.1. Ridge Regression amb i sense expansió polinòmica.

El model Ridge Regression és una millora de la regressió basada en mínims quadrats. Aquesta millora consisteix en imposar una penalització als valors absoluts dels coeficients, tal

que, els coeficients obtinguts mitjançant la tècnica Ridge Regression, minimitzen l'error quadràtic al temps que mitiguen el problema del sobreaprenentatge gràcies a que controlen que els valors dels coeficients en valor absolut és proper a 0. La funció a optimitzar per mínims quadrats amb Ridge Regression és la següent:

$$\min_w ||Xw - y||_2^2 + \alpha||w||_2^2$$

El paràmetre α controla la quantitat de contracció quant més gran és el valor de α més gran és la quantitat de contracció i, per tant, els coeficients es tornen més robustos a la colinealitat [42] .

Per aquest estudi s'ha utilitzat un valor de α = 0.001.

7.4.2. Random Forest Regression

Random Forest és una combinació d'arbres predictors de manera que cada arbre depén dels valors d'un vector aleatori testejat de manera independent i amb la mateixa distribució per a tots els arbres del bosc. L'error de generalització del Random Forest, convergeix a mesura que el número d'arbres és més elevat. En les tasques de regressió, la variable d'eixida és la mitjana dels valors predits per tots els arbres [43].

En aquesta investigació, aquest model s'ha provat amb diferents profunditats màximes de l'arbre {3, 5, 7, 9, 11}.

7.4.3. K Nearest Neighbour Regression

KNN regression té com a objectiu predir un valor numèric per a una nova mostra basant-se en els valors de les seues mostres veïnes, és a dir, busca els K veïns més propers i fa la mitjana aritmètica dels seus valors per predir el valor de la nova mostra. La selecció d'una K adequada és crucial per obtenir bones prediccions ja que afecta al nivell de suavitat de la predicció, d'aquesta manera, una K molt petita pot ser molt sensible a valors anòmals i una K gran pot suavitzar massa la predicció [44].

En aquest estudi s'ha experimentat amb diversos números de veïns {3, 5, 7, 11}.

7.4.4. Xarxes Neuronals: Feedforward neural network

Les Deep Feedforward Network, també anomenades Feedforward Neural Network o Perceptró Multicapa (MLP), són un tipus de xarxes neuronals les quals tenen l'objectiu

d'aproximar alguna funció f*, és a dir, defineix un mapeig $y = f(x; \theta)$ i aprén el valor dels paràmetres θ que donen lloc a la millor aproximació de la funció [45].

Aquests models s'anomenen feedforward perquè la informació flueix a través de la funció que s'està avaluant des de x, a través dels càlculs intermedis utilitzats per definir i, finalment, cap a l'eixida. No hi ha connexions de retroalimentació en les quals les eixides del model es retornen a sí mateix, és a dir, en aquesta xarxa, la informació només es mou en una direcció, cap endavant.

Per aquesta investigació, es defineix una Feedforward neural network, amb la capa d'entrada, a la qual se li aplica un soroll gaussià, amb capes ocultes i una capa d'eixida, utilitzant la funció d'activació ReLU i la normalització per lots, per a millorar el rendiment del model.

Els diferents hiper paràmetres que s'han provat per aquest model són:

- Número d'epochs del model, és a dir, la quantitat de vegades que el model accedeix a la totalitat de les dades d'entrenament: 100 i 200.
- Número de neurones a les capes ocultes: [128, 128, 128], [256, 256], [512, 512].
- Tassa d'aprenentatge de l'optimitzador: 1e-03, 1e-04.
- Optimitzador: RMSprop.

7.4.5. Gradient Boosted Trees

Gradient Boosting és una tècnica d'aprenentatge automàtic utilitzada en tasques de regressió i classificació, que produeix un model de predicció en forma d'un conjunt de models febles, típicament arbres de decisió. És a dir, va construint el model de manera seqüencial, on cada nou arbre intenta corregir els errors dels arbres anteriors apuntant a direccions oposades als gradients.

L'enfocament del Gradient Boosting és millorar gradualment el model combinant múltiples models febles. Cada arbre de decisió es construeix de forma iterativa, on es dona més èmfasi a les dades que han estat mal predites pels arbres anteriors. D'aquesta manera, l'algoritme cerca aprendre dels errors i reduir-los en cada etapa [46].

Aquest model s'ha experimentat amb el número de *gradient boosted trees* {100, 200}, amb un coeficient d'aprenentatge igual a 1, i amb diferents profunditats màximes {3, 7}.

7.5. Experimentació

En aquesta secció, es descriu l'experimentació realitzada per avaluar el rendiment dels models d'aprenentatge automàtic per a la reconstrucció de dades. Per això, l'objectiu principal és

avaluar i comparar els distints models, en funció dels resultats obtinguts segons les mètriques d'avaluació i el seu cost temporal.

S'han seleccionat aquestes mètriques ja que proporcionen una visió global del rendiment dels models des de distintes perspectives.

- MSE: Mean Squared Error
- RMSE: Root Mean Squared Error
- MAE: Mean Absolute Error
- MAPE: Mean Absolute Percentage Error
- R^2: Coeficient de determinació

L'experimentació es va fer seguint una metodologia rigorosa i sistemàtica. D'aquesta manera, per a entrenar els models es van utilitzar sols les dades per a les quals totes les variables tenen menys d'un 20% de dades faltants, en tots els períodes temporals i per a totes les estacions disponibles. Una vegada seleccionades, el conjunt de dades es va dividir en conjunts d'entrenament i test, utilitzant un 70% a les dades entrenament i el 30% a les dades de test. El conjunt de dades d'entrenament és va emprar per entrenar i ajustar els models i les de test van ser reservades per avaluar el rendiment. Cal destacar que l'any 2022 no s'ha utilitzat per a l'entrenament ja que per a la majoria d'estacions, hi ha hagut una fallada sistemàtica a l'hora de captar les dades de la Temperatura, Humitat Relativa i Precipitació.

El procés d'experimentació ha sigut complex ja que s'han utilitzat una gran varietat de models de machine learning. Per a cada variable i estació de mesurament, s'ha entrenat un model sempre i quan les dades han estat disponibles. La qual cosa significa que s'han entrenat els models de forma individual, en funció de les particularitats de cada variable i estació.

A més a més, s'ha fet una cerca exhaustiva dels hiper paràmetres òptims per a cadascun dels models, buscant la configuració amb el millor rendiment per a cada model.

D'aquesta manera, el número de models generats ha sigut considerablement alt, així com el cost temporal d'entrenar nombrosos models de machine learning també ha sigut considerablement elevat. Per això, s'ha seleccionat l'estació Molí del Sol com a referència a l'hora d'avaluar els resultats, ja que és una de les estacions més completes, la qual recol.lecta dades de totes les variables d'estudi i té disponibles la majoria de períodes temporals. A més que durant l'estudi de correlacions entre variables esmentades en la Secció 5.4, no s'observen diferències significatives entre estacions durant els diferents períodes temporals.

En la Taula 14 es pot observar el model i la combinació d'hiper paràmetres òptims per a cada variable, en funció dels resultats obtinguts amb l'avaluació de l'estació Molí del Sol. En la Taula

14 es mostren sols el MAPE i l'R^2, ja que han sigut les mètriques més decisives per a seleccionar els models, però també s'ha tingut en compte els valors del MAE, MSE i RMSE.

Variable	**Model**	**Hiper paràmetres**	**MAPE**	**R^2**	**Temps d'entrenament (segs)**
SO_2	Feed Forward	Epochs = 200 Tassa d'aprenentatge =$1e^{-3}$ Número capes ocultes = [512, 512]	0.153529	0.430968	433.42
CO	Feed Forward	Epochs = 200 Learning Rate =$1e^{-3}$ Número capes ocultes = [128,128,128]	0.252413	0.582300	204.12
NO	Ridge Regression	Interaction_only = False	0.160053	0.999200	0.004
NO_2	Ridge Regression	Interaction_only = False	0.075832	0.998223	0.004
NO_X	Ridge Regression	Interaction_only = False	0.045964	0.999652	0.004
O_3	Feed Forward	Epochs = 100 Learning Rate = $1e^{-3}$ Número capes ocultes = [512,512]	0.230948	0.879357	209.09
PM2.5	Radom Forest	Max depth = 11	0.083758	0.977508	132.17
PM10	Radom Forest	Max depth = 11	0.083174	0.968437	133.12
Temperatura	Feed Forward	Epochs = 200 Learning Rate =$1e^{-3}$ Número capes ocultes = [128,128,128]	0.121542	0.824584	164.77
Humitat Relativa	Feed Forward	Epochs = 200	0.120090	0.778932	383.39

		Learning Rate =$1e^{-3}$ Número capes ocultes = [512, 512]			
Precipitació	-	-	-	-	-

Taula 14: Models i combinació d'hiper paràmetres òptims per a cada variable.

7.6. Resultats

Una de les coses que diferencien clarament els models, és el temps d'entrenament de cadascun d'ells. Com s'ha pogut observar en la Taula 14, els models més complexos com les xarxes neuronals, gradient boosted trees i random forest, tenen un temps d'entrenament dels models molt més elevat comparant-los amb els més simples com són el Ridge Regression o K veïns més propers, que tenen un temps d'execució molt baix.

Tot i encara que els temps d'execució pot arribar a ser molt elevats en alguns casos, s'ha prioritzat el rendiment del model al temps de predicció, ja que com l'objectiu és imputar dades faltants, es necessita que siguen el més precisos possible.

Respecte a les variables SO_2 i CO, en general s'ha obtingut un coeficient de determinació baix. Tanmateix, cal destacar que s'ha de tenir en compte la naturalesa de les dades, ja que tant l'SO_2 com el CO solen ser una constant en la majoria dels casos, prenent valors entre 3 o 4, l'SO_2, i valors molt propers al 0 el CO, excepte algunes dades anòmales. Atés que les variables explicatives són constants en molts casos, el coeficient de determinació (R^2) pot no ser una mètrica fiable, ja que pot tornar resultats molt baixos i fins i tot, resultats negatius, ja que l'R^2 representa la proporció de variància que ha estat explicada per la variable independent del model [47]. Aleshores, si no hi ha molta variabilitat entre les variables, l'R^2 tindrà uns resultats baixos.

Per això, s'ha considerat l'error MAPE com una mètrica més adequada per avaluar el rendiment dels nostres models. Per consegüent, s'han obtingut resultats molt similars entre les xarxes neuronals i KNN amb el nombre de veïns igual a 3. No obstant això, finalment s'han seleccionat les xarxes neuronal com al model més adequat per a capturar les característiques de les variables SO_2 i CO, ja que al ser un model més complex, tornaven unes prediccions molt més precises que el veí més proper, el qual sempre prediu els mateixos valors.

A la Figura 9 s'observa a l'esquerra, com la Feedforward neural network captura millor les característiques del monòxid de carboni en comparació amb les prediccions obtingudes amb el KNN, ambdues prediccions per a les mateixes instàncies.

Feed Forward			KNN	
Id	CO_pred		Id	CO_pred
255035	0,12639		255035	0,10000
255043	0,10295		255043	0,10000
255044	0,12077		255044	0,13333
255047	0,12550		255047	0,10000
255051	0,24342		255051	0,33333
255052	0,38913		255052	0,36667
255054	0,24948		255054	0,36667
255431	0,19720		255431	0,20000
255432	0,15916		255432	0,16667
273709	0,10926		273709	0,13333
273711	0,10317		273711	0,10000
241029	0,17759		241029	0,16667
241030	0,17782		241030	0,13333

Figura 9: Prediccions del CO amb xarxes neuronals (esquerra) i prediccions CO amb KNN (dreta).

El mateix ocorre per a la variable SO_2, on podem observar a l'esquerra les prediccions resultants, molt més precises emprant xarxes neuronals i a la dreta les del veí més proper.

Feed Forward			KNN	
Id	SO2_pred		Id	SO2_pred
269304	2,90282		269304	3,00000
269306	2,86111		269306	3,00000
269307	2,78646		269307	3,00000
269308	2,70419		269308	3,00000
269310	2,72273		269310	3,00000
269301	2,85220		269301	3,33333
269303	2,83849		269303	3,66667
269309	2,81476		269309	3,00000
269311	2,70524		269311	3,00000
269312	2,85084		269312	3,00000
269313	2,82771		269313	3,00000
269531	4,91455		269531	6,00000
70057	3,09265		70057	3,00000

Figura 10: Prediccions del SO_2 amb xarxes neuronals (esquerra) i prediccions CO amb KNN (dreta).

Pel que fa als òxids de nitrogen (NO, NO_2 i NO_X), s'han obtingut uns resultats molt interessants, ja que amb un model senzill com és el Ridge Regression, s'han assolit errors més baixos i un coeficient de determinació més elevat en comparació amb models més complexos com les xarxes neuronals o Gradient Boosted Trees.

De fet, el model Ridge Regression amb expansió polinòmica i *Interaction_only = False*, tenint en compte les interaccions i incloent els termes de potència, ha obtingut uns resultats excel.lents, sent les variables amb menor error i major R^2 de les 11 estudiades.

Respecte als models emprant Feedforward neural network, com és el cas de les variables O_3, Temperatura i Humitat Relativa, els models amb millors resultats, de manera general, han sigut amb un coeficient d'aprenentatge d'$1e^{-3}$. A més, s'ha observat una notable diferència temporal notable entre els models amb 100 epochs i els models amb 200, menor número d'epochs, menor temps d'entrenament. Així com, també s'ha observat que a mesura que augmenten el número de capes ocultes del model, el cost temporal és major.

En relació a les partícules de suspensió (PM2.5 i PM10), ambdós models han obtingut uns errors molt baixos i un R^2 elevat emprant un model de Random Forest amb una profunditat màxima d'11. Respecte al temps d'entrenament, és elevat si el comparem amb models més senzills com Ridge Regression o KNN, però el cost temporal de l'entrenament és menor comparant-lo amb les xarxes neuronals.

Per últim, cal destacar que de la variable de la precipitació no se n'ha escollit cap model òptim perquè els models han mostrat uns errors significatius i un coeficient de determinació (R^2) molt baix, i fins i tot en alguns casos negatiu. Com s'ha comentat en la Secció 5.1, sobre les distribucions de les variables, la precipitació té un rang de dades molt limitat, sent pràcticament sempre 0, excepte en alguns casos concrets. Aquest fet és normal ja que la ciutat de València no és un lloc on la pluja siga característica. Aquesta particularitat de la variable, ha fet arribar a la conclusió de que no té sentit predir una variable que en la majoria dels casos té un valor constant, 0. És per això, que els regressors no han sigut capaços de capturar correctament les variacions i patrons d'aquesta variable, la qual cosa es reflecteix en el baix rendiment dels models.

A la Taula 15, es poden observar els millors resultats obtinguts per a la variable precipitació, els quals tenen uns errors extremadament elevats i un coeficient de determinació molt baix.

Model	**MAPE**	**R^2**
Random Forest	123062640829370	0,0339
KNN	78912705737550,9	0,1072
Ridge Regression	283152242785045	0,0458
Gradient Boosted Trees	315225533669213	0,0124

Taula 15: MAPE i R^2 dels millors models per a la variable Precipitació.

A pesar que els millors resultats, la majoria dels quals s'han obtés amb les xarxes neuronals, cal destacar que el veí més proper (KNN), també ha proporcionat uns bons resultats. D'altra banda, els models basats en Gradient Boosted Trees amb la llibreria XGBoost han obtingut

resultats més baixos dels esperats, ja que al ser un model més complex s'esperava obtenir millors resultats.

Aleshores, a la vista d'aquests resultats obtinguts en els models de regressió es pot dir que confiem en aquests models per a reconstruir dades faltants assumint cert percentatge d'error, aquest percentatge d'error difereix en funció de cada variable ja que com es pot observar a la Taula 14, les partícules de suspensió PM2.5 i PM10 junt amb els òxids de nitrogen tenen uns errors molt baixos, fins i tot aquests últims esmentats es prediuen a un error mínim. La resta de variables tenen un error més accentuat en comparació a aquestes últimes, tot i això, segueixen tenint un error menut.

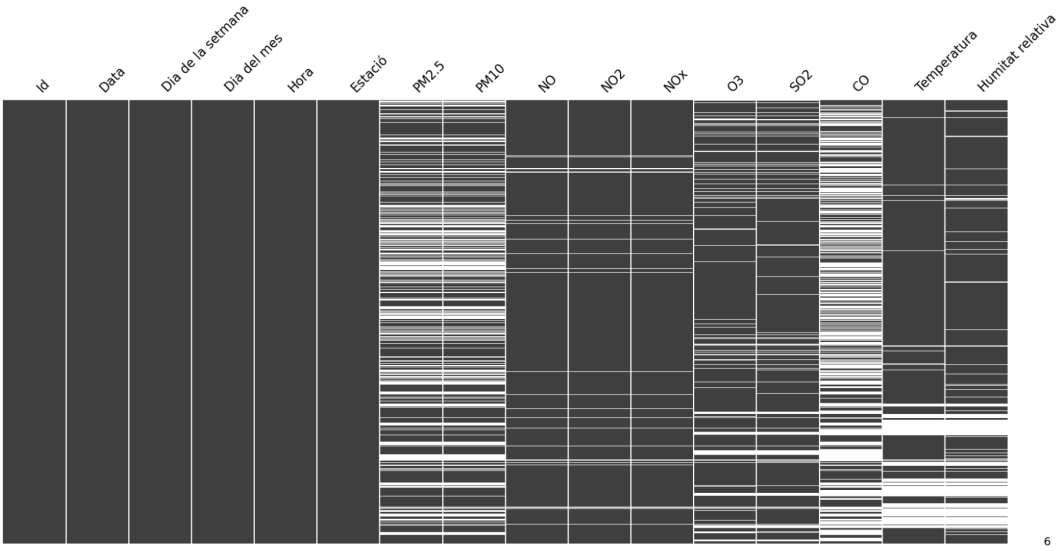
7.7. Imputació de dades faltants

Una vegada s'han seleccionat els models de regressió òptims per a cadascuna de les 11 variables d'estudi, es porta a terme el procés de predicció de dades faltants, sempre i quan siga possible. Amb l'objectiu de maximitzar la qualitat i integritat de les dades, per a utilitzar-lo en les posteriors ferramentes visuals, sempre que s'assumisca que existeix un cert percentatge d'error, concret per a cada model.

Cal destacar que hi ha moments complexos per a determinar quina de totes les variables té dades faltants a causa de la naturalesa de l'automatització de les dades del conjunt de test, ja que per a predir un model de machine learning no poden haver dades faltants per a les variables resposta. Per això, cal destacar que encara que totes les variables tinguen menys d'un 20% de faltants, no implica que aquesta variable no tinga faltants, ja que poden coexistir dues variables amb faltants durant el mateix període temporal. Per a variables que tenen més d'un 20% de dades faltants durant un mateix període és més senzill, ja que una de les variables s'omet quan es prediu l'altra, i viceversa. Però quan tenen menys d'un 20% de dades faltants no es pot saber quines són les variables que coexisteixen amb dades faltants en una determinada instància.

A la Taula 16 es pot observar la matriu de valors faltants obtinguda amb la imputació de dades faltants. A partir dels identificadors de cada instància s'han imputat aquests valors per la predicció resultant del model de regressió per a cada determinada variable.

Es pot observar que els òxids de nitrogen estan pràcticament complets, després de la imputació de dades amb les prediccions.



Taula 16: Matriu de dades faltants després de la imputació de valors faltants.

A la Taula 17 es mostra més en detall el percentatge de dades faltants després d'haver realitzat la imputació de valors faltants i la reducció assolida.

Es pot observar que hi ha variables com els òxids de nitrogen, l'SO_2 i l'O_3 on s'ha produït una reducció notable respecte a la resta de variables.

Variable	**Percentatge de dades faltants dataset (%)**	**Percentatge de dades faltants després de la imputació de dades (%)**	**Reducció de faltants (%)**
Id	0	0	0
Data	0	0	0
Dia de la setmana	0	0	0
Dia del mes	0	0	0
Hora	0	0	0
Estació	0	0	0
PM2.5	47.03	34.06	12.97
PM10	47.09	34.08	13.01
NO	23.58	4.14	19.44
NO_2	20.44	4.14	16.30
NO_X	23.58	4.14	19.44

O_3	28.49	11.99	16.50
SO_2	28.78	12.48	16.30
CO	64.78	54.94	9.84
Temperatura	15.57	15.21	0.36
Humitat relativa	18.24	18.07	0.17

Taula 17: Percentatge de valors faltants abans i després de la imputació de dades faltants amb les prediccions.

A la vista dels resultats assolits, amb la imputació de dades, es constata que els models han pogut reduir satisfactòriament el percentatge de valors faltants en la majoria de les variables d'estudi.

8. Dashboard amb les dades reconstruïdes

8.1. Desenvolupament del dashboard

Per a finalitzar amb les anàlisis, s'ha realitzat un conjunt de dashboards en Power Bi per estudiar la contaminació de l'aire a la ciutat de València. S'han dissenyat un total de 8 dashboards, un per cadascun dels contaminants atmosfèrics.

A la Figura 11 es pot veure el dashboard per l'ozó en l'estació de mesurament ubicada a la Universitat Politècnica durant gener de 2019, on avisa que la dashboard compta amb dades reconstruïdes utilitzant models de predicció de ML.

El tauler de dades proporciona diverses funcionalitats interactives, com són el selector del mes, any i estació de mesuració amb els quals es poden visualitzar els resultats corresponents.

A la part superior del dashboard, es pot observar una taula ordenada en funció de les concentracions màximes d'O_3, a més, permet visualitzar la data i hora concreta en què es va recopilar cada registre amb la temperatura i humitat relativa d'eixe moment concret.

Després, hi ha dos medidors dels nivells màxims d'O_3. Aquests nivells es comparen amb el nivell d'informació, per al qual les concentracions d'O_3 suposen un risc per a la salut per als grups vulnerables, i també amb el llindar d'alerta, per al qual les concentracions d'ozó suposen un risc per a la salut humana, en general. D'aquesta manera, els medidors fluctuen en funció dels llindars establerts per cada variable. És per això, que en cada dashboard es comenten els diferents llindars, la institució que el defineix (directrius de l'OMS 2021 [10] o BOE-A-2023-2026 [11]) i a quin tipus de registre afecta, és a dir, horaris, diaris o anuals.

Seguidament, s'observa la mitjana porcentual de l'O_3 en funció dels dies de la setmana i també la distribució de la mitjana de l'O_3 a les distintes hores del dia, per a un mes, any i estació concrets.

Per últim, s'ha desenvolupat un gràfic d'àrees que mostra l'evolució diària de l'O_3. Aquesta gràfica varia en funció de la variable, ja que en alguns casos es mostren els nivells mitjans diaris en un mes concret i en altres variables es mostra els nivells acumulats diaris de la variable esmentada, per a poder fer una comparativa amb els llindars diaris.

Figura 11: Dashboard de la variable O_3 per a gener de 2019.

Cal destacar que s'ha explicat la composició del tauler de comandaments per a l'ozó, però l'organització i l'estètica són completament iguals per a la resta de variables, excepte algunes funcionalitats que són concretes de cada contaminant.

8.2. Resultats obtinguts

S'ha realitzat un dashboard per a cadascun dels contaminants. D'aquesta manera es comenten els resultats obtinguts de forma individual per a cada una d'elles.

8.2.1. SO_2

En referència als llindar diaris, hi ha una gran diferència entre els definits per les directrius de l'OMS [10] i els del BOE-A-2023-2026 [11], respectivament 40 µg/m³ i 180 µg/m³.

Fixant-se en l'evolució dels nivells diaris de SO_2 acumulat, es pot observar que la majoria dels dies superen els límits establerts per l'OMS, però estan per sota del llindar del BOE-A-2023-2026, excepte en determinats casos.

A més a més, pel que fa al patró horari observat en el gràfic dels nivells mitjans d'SO_2 en funció de les hores, es pot apreciar que entre les 9 i les 10 del matí fins a les 14-15 hores de la vesprada s'assoleixen unes concentracions lleugerament superiors en comparació amb les altres hores del dia.

Figura 12: Dashboard de la variable SO_2 en febrer de 2019 a l'estació Avinguda França.

8.2.2. Òxids de Nitrogen

Pel que fa als òxids de nitrogen, s'observa que tant l'NO, NO_2 i NO_X tenen un comportament similar en el sentit que tots ells presenten fluctuacions en els seus nivells. No obstant això, les proporcions entre aquests gasos varien. És important destacar que els nivells de NO són més baixos en comparació amb els d'NO_2 i NO_X, i l'NO_2 presenta nivells inferiors a NO_X. Aquest fet és normal ja que el diòxid de nitrogen es forma amb la combinació entre l'NO i oxigen, i l'NO_X es forma a partir de NO_2 i NO_X.

Tant per a l'NO, NO_2 i NO_X s'observa un patró horari, veure a la Figura 13, al gràfic de la mitjana per hores. A partir de les 4-5 del matí, les concentracions de NO_2 pugen, aplegant als nivells màxims. A les 9-10 del matí comencen a baixar els nivells fins a les 15 de la vesprada que assoleixen els valors mínims. A partir d'ací, les concentracions segueixen creixent fins a les

00 del dia següent on tornen a baixar els nivells durant la matinada, i a les 5 del matí es torna a repetir aquest cicle. Veure a l'Annex IV aquest mateix dashboard per a les variables NO i NO_X.

Figura 13: Dashboard de la variable NO_2 en octubre de 2019 a l'estació Avinguda França.

Després, a la Figura 14 s'observa que durant els mesos d'estiu, els nivells dels òxids de nitrogen són notablement inferiors en comparació amb els mesos freds, per a totes les estacions. Per això, és als mesos d'hivern on es sobrepassen els llindars establerts per l'OMS i al BOE-A-2023-2026 de manera més recurrent.

Per últim, s'aprecia que les estacions Molí del Sol, Vivers i Politècnic presenten nivells més baixos d'aquestes variables en comparació amb les altres estacions. Aquesta observació pot ser el resultat de diferents factors, com la ubicació geogràfica, la densitat de trànsit o la presència de fonts d'emissió més baixes en aquestes àrees.

Figura 14: Nivells d'NO, NO_2 i NO_X per mesos per a cada estació.

Finalment, es comparen a la Figura 15 els nivells mitjans anuals de l'NO_2 establerts per la OMS, 10 µg/m³ i els del BOE-A-2023-2026, 40 µg/m³.

S'aprecia una reducció dels nivells anuals des de l'any 2017, any en què gran part de les estacions de mesurament es troben per damunt dels llindars establerts pel BOE-A-2023-2026, i òbviament superen també els de les directrius de l'OMS 2021. Tanmateix, a mesura que passen els anys, s'observa una disminució de les concentracions de NO_2, fins i tot algunes estacions estan per sota els llindars del BOE-A-2023-2026. Especialment, és al 2021 on es registren els valors més baixos per a totes les estacions, fins i tot l'estació Politènic està per sota dels llindars de l'OMS, sent la primera vegada que alguna estació registra uns valors tan baixos. Encara que aquests valors mínims tornen a augmentar a l'any 2022 de manera general per a totes les estacions de mesurament.

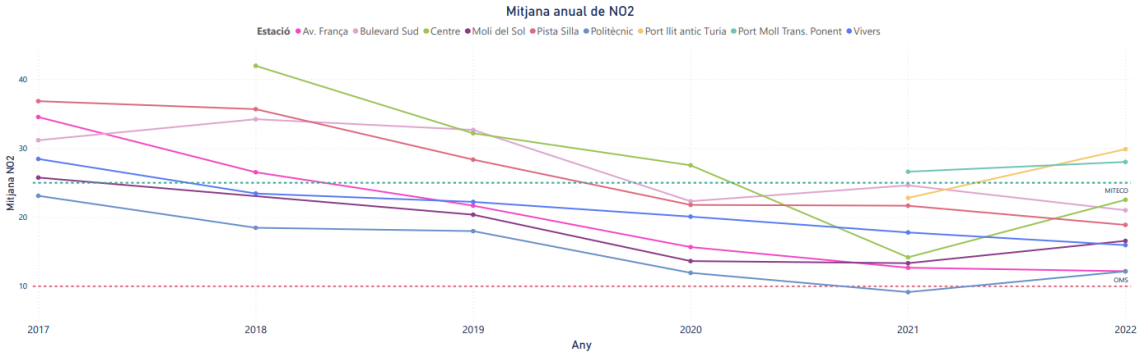


Figura 15: Mitjana anual de NO_2 per estació (2017-2022).

8.2.3. CO

Després d'analitzar el quadre de comandaments del monòxid de carboni (CO) s'han observat diversos aspectes rellevants.

En primer lloc, no s'ha detectat cap patró significatiu en els nivells de CO al llarg dels diferents mesos. S'han obtingut valors mitjos similars per a cada mes, la qual cosa indica que no hi ha una variació consistent en els nivells de CO al llarg d'un mateix any. D'altra banda, no s'ha observat cap patró clar en els nivells de CO en relació amb el dia de la setmana.

No obstant això, s'ha detectat un patró temporal interessant en els nivells de CO al llarg de les hores del dia, veure Figura 16. Es pot observar que a partir de les 4 o 5 del matí, comencen a augmentar els nivells de CO, assolint els seus valors màxims fins a les 6-7 del matí. A partir d'aquest moment, els nivells de CO comencen a disminuir fins a les 10 del matí, on es mantenen constants fins a les 7-8 de la vesprada. Després, tornen a augmentar fins a assolir una altra vegada els valors màxims cap a les 23 hores de la nit, i posteriorment disminueixen i es mantenen constants fins que es repeteix el cicle.

Figura 16: Dashboard de la variable NO_2 en maig de 2019 a l'estació Molí del Sol.

A més, s'ha observat que hi ha diferències significatives entre les estacions en termes de nivells diaris de CO de manera general en els distints períodes temporals. Es pot veure a la Figura 17, que les estacions Pista de Silla i Port Llit Antic Turia presenten nivells diaris molt més elevats de CO, fins i tot superant el llindar màxim establert per l'OMS (línia discontínua vermella), 4 mg/m³, i acostant-se al límit del BOE-A-2023-2026 (línia discontínua verda), 10 mg/m³. En canvi, altres estacions com Molí del Sol registren valors més baixos, fins i tot, dins del rang de valors permesos per l'OMS.

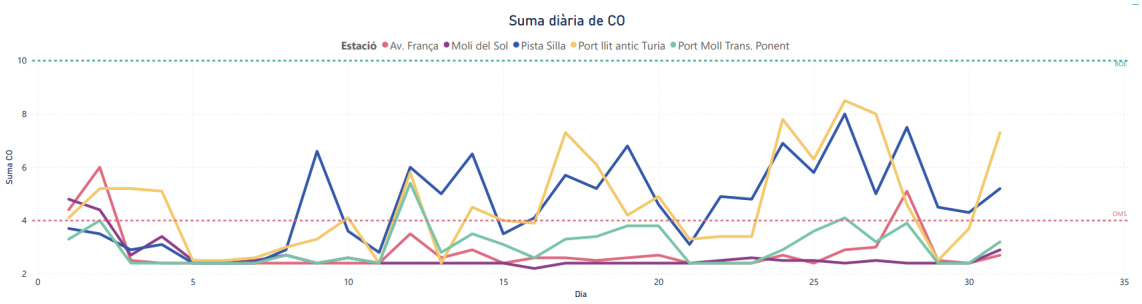


Figura 17: Nivells diaris de CO per estació, al desembre de 2022.

Cal destacar que en la majoria de casos, els valors diaris per a les diverses estacions no superen el llindar establert pel BOE-A-2023-2026, però hi ha alguns casos excepcionals en què sí es supera.

8.2.4. O_3

Pel que fa a l'ozó, es pot observar a la Figura 11, esmentada anteriorment per a explicar les distintes funcionalitats del tauler de comandaments, que de manera general no es sobrepassa el llindar d'informació establert pel BOE-A-2023-2026.

A més a més, s'han trobat uns patrons horaris molt accentuats en general per a tots els períodes temporals i totes les variables. A partir de les 9-10 del matí les concentracions d'O_3 creixen fins assolir el seu màxim sobre les 15 de la vesprada. Després es redueixen fins a les 20-21 de la nit on es mantenen constants fins a les 5 del matí que decreixen i apleguen a obtenir els valors mínims durant la matinada, fins que es torna a repetir aquest cicle.

Després d'estudiar el comportament horari de les concentracions d'ozó de manera individual, s'ha pogut apreciar que l'O_3 es comporta de manera totalment contrària a l'NO_2.

A la gràfica 18 es pot observar que l'O_3 assoleix valors màxims durant el dia i l'NO_2 és el que obté durant la matinada i la vesprada/nit, veure Figura 18.

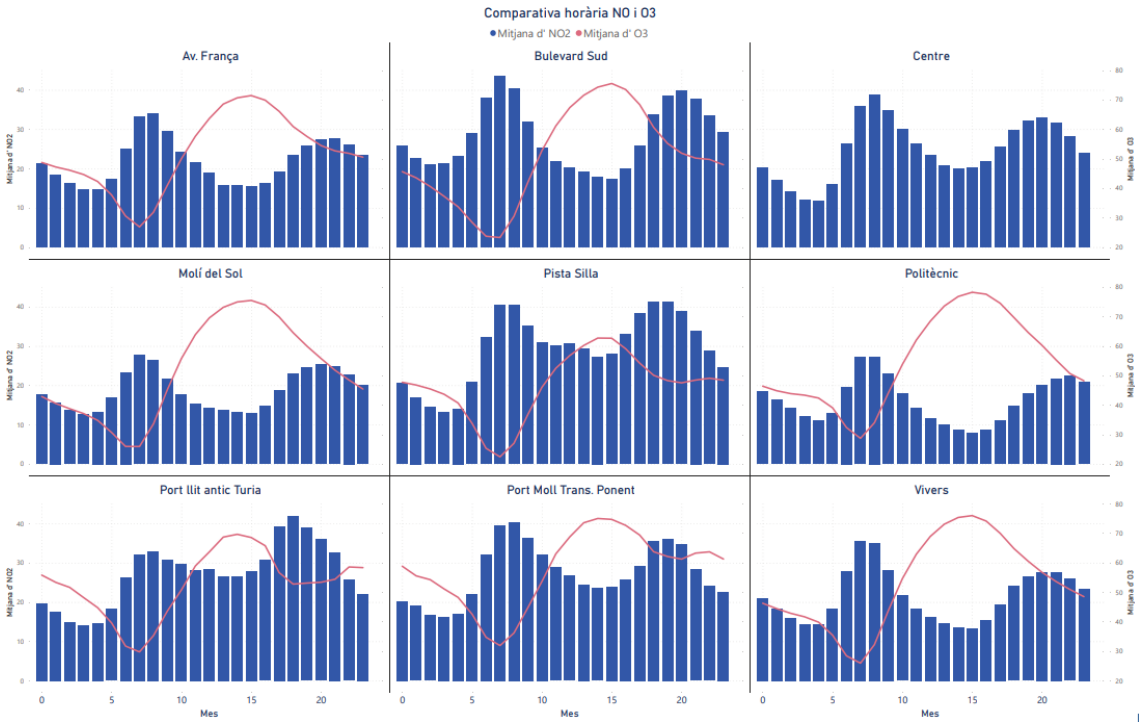


Figura 18: Nivells de diaris de CO per estació, al desembre de 2022.

Per últim, amb la possibilitat d'analitzar els diferents períodes temporals al llarg dels mesos, s'ha observat que durant els mesos freds, és a dir, a l'hivern, les concentracions d'O_3 són més baixes que durant la primavera i l'estiu.

A la vista dels resultats obtinguts amb les anàlisis mensuals i horaris dels nivells d'ozó, es pot dir que quan hi ha més presència del sol, és a dir al migdia i durant l'estiu, les concentracions d'O_3 són majors. D'aquesta manera, es podria dir que la radiació solar té un paper rellevant en la formació d'aquest contaminant.

8.2.5. PM2.5

En relació a les PM2.5, s'observa un patró consistent en tots els mesos, anys i estacions, veure a la Figura 19 per a l'estació Av. França. A partir de les 8-9 del matí, els nivells de partícules de suspensió amb un diàmetre inferior a 2.5 micres, comencen a disminuir. A mesura que avança el matí, els nivells continuen disminuint fins a arribar a un punt mínim al voltant de les 14-15 hores. Després d'assolir el valor mínim, els nivells de PM2.5 comencen a augmentar de nou. Aquest augment es manté durant la vesprada i la nit, estabilitzant-se al voltant de les 23 hores de la nit, on s'observen nivells màxims de partícules en suspensió. Durant aquesta franja horària, els nivells de PM2.5 es mantenen constants fins a les primeres hores del matí següent, quan el patró es reinicia amb la disminució dels nivells.

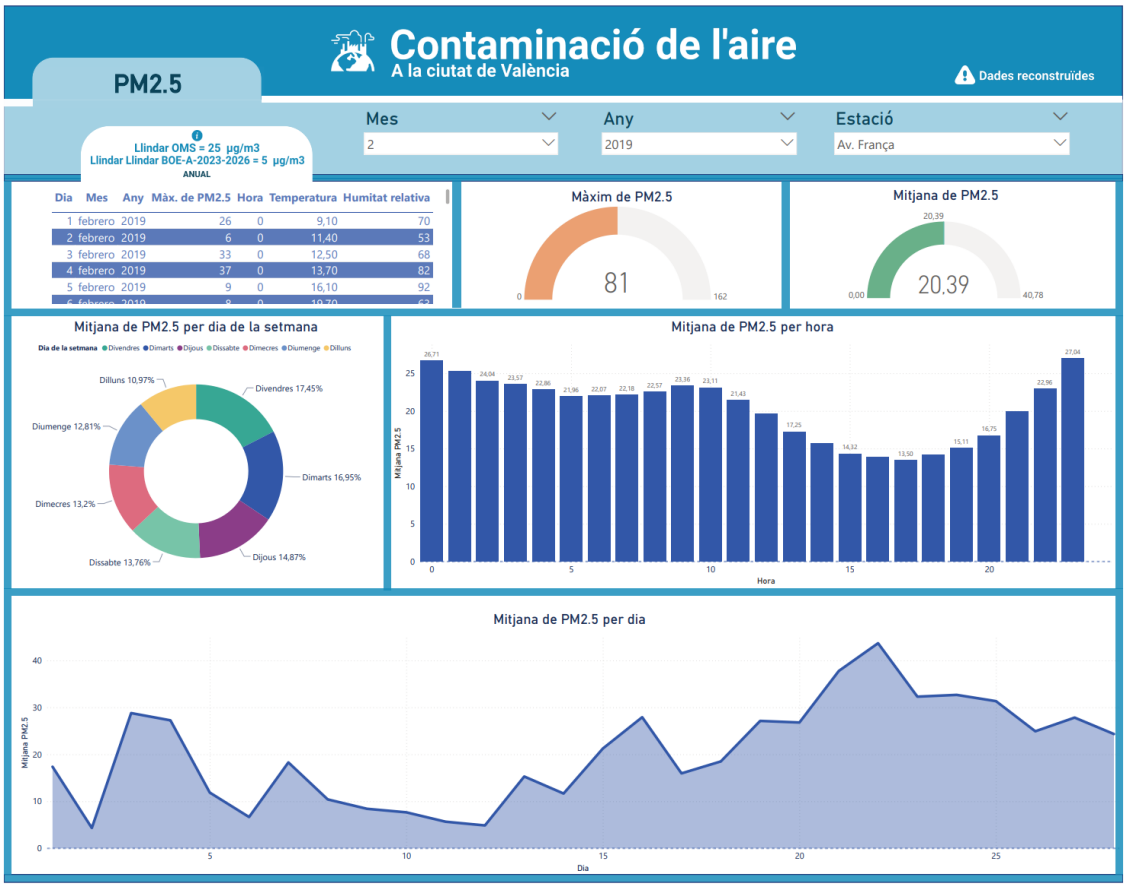


Figura 19: Dashboard de la variable PM2.5 en febrer de 2019 a l'estació Av. França.

A més a més, segons l'OMS, el valor anual de PM2.5 és 5 µg/m³, què és un límit molt més restrictiu. No obstant això, en els valors límits establerts pel BOE-A-2023-2026, aquest límit és més flexible, fixant-se en 25 µg/m³. Aquesta diferència entre els dos estàndards ha generat controvèrsia i debat en el marc de la regulació de la qualitat de l'aire.

En la Figura 20, l'evolució anual mitja de les PM2.5 amb el valor límit legislat pel BOE-A-2023-2026, representat en la línia discontínua verda i el de l'OMS en la línia discontinua vermella, es pot observar que el valor mitjà de PM2.5 establert per l'OMS sempre és inferior als valors mitjans anuals de cada any en totes les estacions. Això indica que els valors mitjans de PM2.5 es mantenen per sobre del límit de l'OMS, reflectint una major restricció i precaució en la protecció de la salut pública.

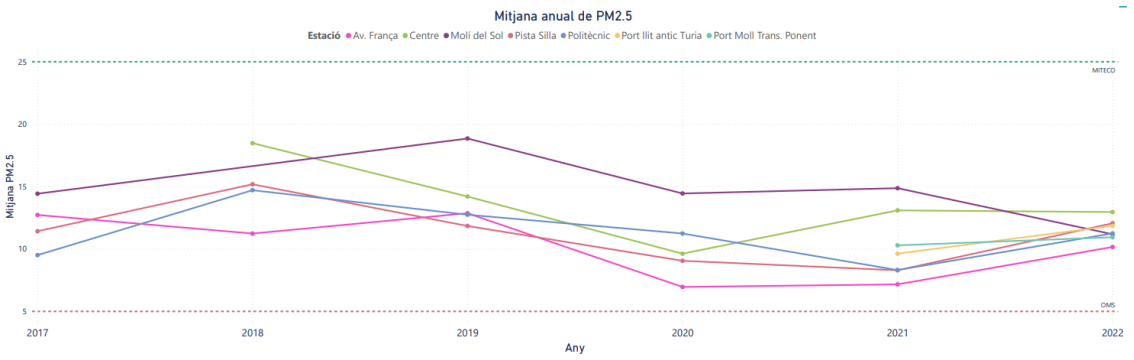


Figura 20: Nivells de PM2.5 anuals per estació.

Aquesta discrepància entre els dos llindars planteja qüestions importants sobre la protecció de la salut pública i la necessitat de revisar i harmonitzar els estàndards de qualitat de l'aire. Els resultats d'aquest TFG evidencien la importància de tenir en compte les recomanacions més restrictives de l'OMS per garantir una millor qualitat de l'aire i protegir la salut de la població.

8.2.6. PM10

Pel que fa a les partícules en suspensió amb diàmetre inferior a 10 micres, s'ha detectat un patró horari consistent, en què s'assoleixen els nivells màxims d'aquest contaminant entre les 6 i les 10 del matí, independentment de l'estació. La resta de les hores del dia, els nivells d'aquest contaminant es mantenen relativament constants, veure Figura 21.

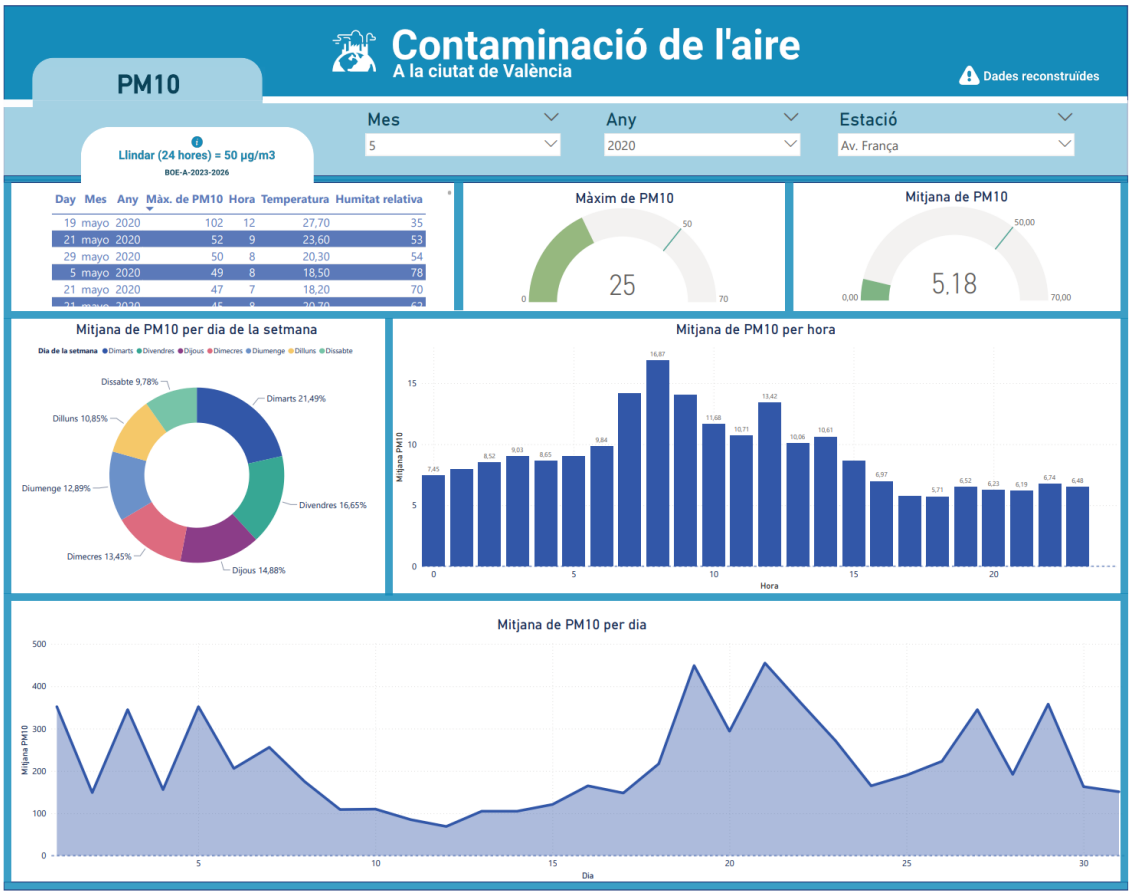


Figura 21: Dashboard de la variable PM10 en maig de 2020 a l'estació Av. França.

A més, s'ha comparat el llindar establert per l'Organització Mundial de la Salut (OMS), que és de 15 µg/m³ com a límit mitjà anual, amb el llindar establert pel BOE-A-2023-2026, que és de 40 µg/m³. Es nota una clara diferència entre aquests dos llindars, sent el de l'OMS molt més restrictiu. S'ha realitzat una anàlisi dels valors mitjans de PM10 a totes les estacions durant els anys 2017 a 2022.

A la Figura 21, es pot veure l'evolució dels nivells mitjans de les PM2.5. En cap any s'ha superat el llindar establert pel BOE-A-2023-2026 (línia discontínua verda), tot i que entre el 2017 i el 2019 quasi totes les estacions superaven el llindar establert per l'OMS. És destacable la diferència en l'any 2020, en què totes les estacions es trobaven per sota d'aquest llindar tan restrictiu. Aquesta observació és coherent, ja que l'any 2020 va ser marcat per la pandèmia de la COVID-19, durant la qual la ciutat de València es va aturar completament. Es pot atribuir a aquesta paralització l'absència d'activitats emissores importants com l'agricultura, la construcció o altres activitats industrials, que solen ser fonts rellevants de partícules en suspensió.

Posteriorment, durant l'any 2021, algunes estacions continuen mantenint-se per davall d'aquest límit, però en l'any 2022 es detecta un augment progressiu dels nivells de PM10 a totes

les estacions. Sols l'estació Molí del Sol es troba per davall del llindar de l'OMS (línia discontínua vermella) amb 14.80 µg/m³ .

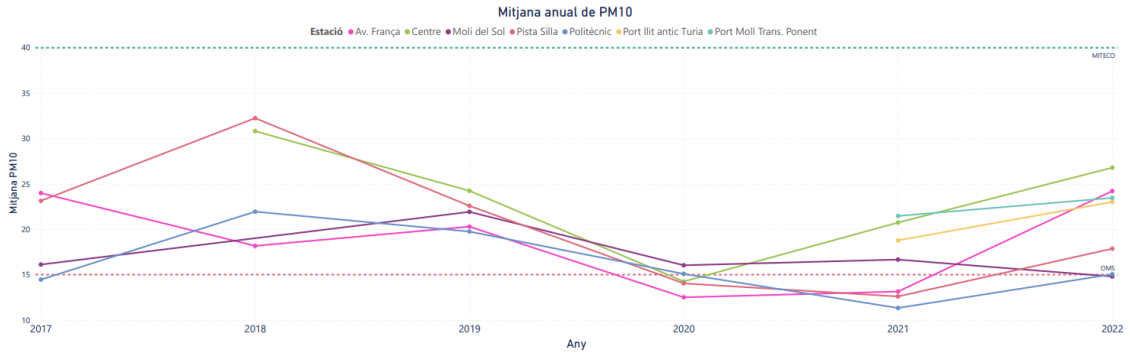


Figura 22: Nivells de PM2.5 anuals per estació.

9. Conclusions

9.1. Conclusions finals

En referència a les correlacions entre variables, s'ha observat una forta correlació dels òxids de nitrogen (NO, NO_2, NO_X) entre sí i les partícules de suspensió també entre elles mateixa (PM2.5 i PM10). Cal destacar, que existeixen correlacions no tan accentuades com les esmentades anteriorment, com és el cas de la temperatura i l'ozó, la qual s'ha pogut veure també reflectida en el dashboard on es veu que durant l'estiu i primavera i també, al matí i migdia, els nivells d'O_3 són majors, així com la temperatura també és més elevada,

Respecte a l'experimentació i selecció dels models de regressió i mètriques d'avaluació, és molt important basar-se en el context i les característiques específiques de cada variable, com ha sigut el cas de l'SO_2 i CO, per a les quals el R^2 no era una mètrica d'avaluació fiable a causa de la naturalesa de les dades amb un rang molt menut. És per això que cal destacar que és molt important considerar tant la complexitat dels models com la capacitat per adaptar-se a les dades disponibles. En alguns casos un model més senzill és més efectiu i apropiat que altre amb una major complexitat, com ha ocorregut en el cas dels òxids de nitrogen, on el Ridge Regression ha obtingut els millors models per a aquestes variables.

Per consegüent, a la vista dels resultats obtinguts en els models de predicció es pot dir que es poden reconstruir satisfactòriament les dades faltants sempre i quan siga possible, és clar, assumint un marge d'error. A pesar que la reconstrucció de dades no ha sigut perfecta, ja que no ha estat possible predir tots els períodes temporals, sí que s'han pogut imputar un gran percentatge de dades faltants, especialment per als òxids de nitrogen que compten sols amb aproximadament un 4% de faltants, front a un 24% de faltants quan es començà l'estudi.

Una de les coses que més ha destacat en referència a les visualitzacions de cada variable, ha sigut la discrepància entre els llindars establerts per les directrius de l'OMS (2021) [10] i les establertes en el BOE-A-2023-2026 [11], les primeres molt més restrictives que les últimes. Per això, no té sentit comparar dues mesures que difereixen tant una respecte l'altra, perquè, a la fi, si es vol ser transparent front a la situació actual de l'aire, és necessari tenir uns llindars i valors d'alerta comuns. Perquè, en la majoria de les estacions, sempre es tendeix a superar els llindars de l'OMS però al mateix temps, aquests valors es troben per sota dels definits pel BOE-A-2023-2026, la qual cosa fa dubtar a l'hora de prendre un d'ells com a referència, perquè arriben a ser contradictoris.

Per últim, durant tot aquest estudi s'ha destacat l'alt percentatge de dades faltants i s'ha observat que en quatre de les tretze estacions no es registra cap paràmetre meteorològic. A més, hi ha estacions que es suposa que capten certs paràmetres contaminants però després en les dades no hi ha cap registre.

És per això que sembla una fal.làcia promoure una València sostenible i transparent, quan es comprova que l'Administració Pública es vanaglòria d'uns serveis de mesurament i recopilació de dades que no es corresponen amb la realitat, ja que s'ha pogut constatar una gran manca de dades, per a les distintes estacions i períodes temporals estudiats amb aquests conjunts de dades.

9.2. Anàlisi del marc legal i ètic

Les dades emprades per a la realització d'aquest treball de fi de carrera han sigut obtingudes del portal de dades obertes de l'Ajuntament de València i dels registres històrics de la qualitat de l'aire de Conselleria d'Agricultura, Desenvolupament Rural, Emergència Climàtica i Transició Ecològica.

Ambdós portals són de caràcter obert i tenen l'objectiu de facilitar a tots els ciutadans l'accés lliure i gratuït a les dades públiques, conforme estableix la Llei 39/2015, d'1 d'octubre, del Procediment Administratiu Comú de les Administracions Públiques [48], que en el Títol II-Capítol I-Article 13 diu que "Els ciutadans tenen dret a accedir a la informació pública, arxius i registres, d'acord amb el que preveu la Llei 19/2013, de 9 de desembre, de transparència, accés a la informació pública i bon govern [49] i la resta de l'ordenament jurídic".

Cal destacar la influència del conegut com el Conveni d'Aarhus (Aarhus el 25 de juny de 1998), en la Llei 27/2006, de 18 de juliol, per la qual es regulen els drets d'accés a la informació, de participació pública i d'accés a la justícia en matèria de medi ambient [50], que parteix del següent postulat: "perquè els ciutadans puguen gaudir del dret a un medi ambient saludable i complir el deure de respectar-lo i protegir-lo, han de tenir accés a la informació mediambiental rellevant, han d'estar legitimats per participar en els processos de presa de decisions de caràcter ambiental i han de tenir accés a la justícia quan aquests drets els siguin negats" [51]. Per això, per a aquest projecte de fi de carrera, és molt important el dret a cercar i obtenir informació que estiga en poder de les autoritats públiques, i el dret a rebre informació ambientalment rellevant per part d'aquestes, que l'han de recollir i fer pública sense necessitat que hi haja una petició prèvia.

És per això que no hi ha problemes de protecció de dades o protecció intel.lectual a l'hora d'utilitzar, reutilitzar i distribuir aquestes dades.

A més, cal destacar que cap dels models emprats discriminen en base a les diferents ètnies, sexe, estereotips o qualsevol tipus d'atribut protegit ja que no hi ha cap variable que puga crear alguna d'aquestes controvèrsies morals.

9.3. Relació del treball amb els estudis cursats

La investigació de la qualitat de l'aire a la ciutat de València s'ha convertit en una recapitulació completa del material estudiat al llarg de la carrera de Ciència de Dades. Com s'ha treballat amb dades en cru, és a dir, dades obtingudes directament dels sensors de mesurament de cada estació, s'ha realitzat una anàlisi de dades exhaustiva utilitzant diferents llibreries de Python, a més que s'ha realitzat un profund enteniment de les dades per a poder manipular i prendre decisions clares i coherents sobre el conjunt de dades, la qual cosa ha sigut una part fonamental del projecte.

Per altra banda, s'han aplicat tècniques de processament de les dades i d'aprenentatge automàtic per a fer les prediccions. A més d'haver realitzat una àmplia cerca de hiper paràmetres òptims en funció dels models i les variables d'estudi.

Per últim, una vegada es van obtenir els resultats, s'han emprat diverses tècniques de visualització de dades. Aprofitant el coneixement aprés de les ferramentes com Power BI i Dash de Plotly. Amb aquests dashboards s'ha pogut fer un storytelling del projecte. Primer amb el dashboard de Dash, el qual mostra la primera fase del projecte que fou l'anàlisi de dades, el qual permet explorar la situació de cada variable en un determinat moment temporal, incloent-hi la variància, el percentatge de faltants i la mitjana aritmètica. Després, amb dashboards fets amb Power BI mostren el comportament i l'evolució per a cada variable una vegada s'han reconstruït els valors faltants.

En definitiva, aquest projecte ha sigut una oportunitat per aplicar els coneixements teòrics i pràctics adquirits al llarg de la carrera. Utilitzant diferents tècniques estadístiques per a l'anàlisi de dades, diferents eines de ML per a realitzar prediccions i per últim usant diferents eines de visualització de dades per representar els resultats obtinguts.

9.4. Llegat

Amb l'objectiu de fomentar la col·laboració i l'ús compartit d'aquest treball de fi de carrera, perquè qualsevol persona puga usar-lo com a punt de partida per a fer una possible col.laboració, tot el codi relacionat amb el pre processament de dades, la creació dels models de regressió, el script del Dash de Python i el fitxer de Power BI que conté les visualitzacions i els informes, es troben publicats al següent repositori de GitHub:

https://github.com/Anniellinn/TFG.

9.5. Limitacions del treball

Una de les limitacions més importants d'aquest projecte ha sigut l'elevat número de dades faltants, la qual cosa ha tingut un impacte considerable en el desenvolupament de la investigació. De les 13 estacions de mesurament, només dues d'elles, Molí del Sol i Av. França, recopilaven dades de tots els contaminants atmosfèrics que s'han estudiat.

Una altra limitació significativa ha estat la presència d'estacions que, tot i que es suposava que estaven actives, no registraven dades de cap contaminants atmosfèrics, per tant van ser eliminades de l'estudi, reduint el nombre d'estacions de 13 a 9.

A més, altre aspecte important a destacar és que, fins i tot en les estacions restants, no es disposava de totes les dades de totes les variables. Això va generar limitacions addicionals, ja que, per exemple, l'estació del centre de València no registrava dades d'ozó, una variable d'interès per al seu estudi en una zona tan concorreguda.

Per últim, altre desafiament considerable ha estat la impossibilitat de reconstruir completament el conjunt de dades. Tot i que es va aconseguir reduir significativament el número de faltants, encara hi va haver variables per a les quals el percentatge de valors mancants seguia sent elevat.

En resum, la principal limitació d'aquest projecte ha sigut la incompletitud de les dades disponibles.

9.6. Pròxims treballs

Un del futurs passos per a continuar amb aquest projecte, per a perfilar les visualitzacions finals, per a millorar la interpretabilitat, s'incorporaria una distinció en la gràfica que mostra la mitjana per dia, en un mes i any concret, quan en eixe dia hi ha dades reconstruïdes. Actualment, en el tauler de comandaments no es diferencien de cap forma les dades originals de les predites amb models de ML. Aquesta distinció seria acolorir amb altre color els dies que compten amb dades predites amb un color, i els que no amb altre.

Altra millora seria la inclusió de més variables a l'estudi per a poder explorar i analitzar altres contaminants, com és la radiació la qual, s'ha pogut veure, té un paper rellevant en la formació de l'ozó. Per això, contemplar més variables d'estudi proporcionaria una visió més completa de la qualitat de l'aire perquè es contemplaria un ventall més ample de factors ambientals

Referències

[1] «Competències i traspassos de funcions - Generalitat Valenciana», *Gva.es*. [En línia]. Disponible en: https://gvaoberta.gva.es/va/competencias-y-traspaso-de-funciones. [Accedit: 29-jun-2023].

[2] United Nations, «La Cumbre de Desarrollo Sostenible de la ONU: 17 Objetivos para transformar nuestro mundo», 19-ene-2016. [En línia]. Disponible en: https://www.youtube.com/watch?v=bk9Z6OWh_34. [Accedit: 29-jun-2023].

[3] M. Moran, «Objetivo 11: Lograr que las ciudades sean más inclusivas, seguras, resilientes y sostenibles», *un.org*, 07-ene-2015. [En línia]. Disponible en: https://www.un.org/sustainabledevelopment/es/cities/. [Accedit: 29-jun-2023].

[4] O. M. de la Salud, «Contaminación del aire ambiente (exterior)», *Who.int*. [En línia]. Disponible en: https://www.who.int/es/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health. [Accedit: 29-jun-2023].

[5] «UNE 178201:2016 Ciudades inteligentes. Definición, atributos y requisitos», *Une.org*. [En línia]. Disponible en: https://www.une.org/encuentra-tu-norma/busca-tu-norma/norma?c=N0056504. [Accedit: 29-jun-2023].

[6] WHO Media Team, «Tripartite and UNEP support OHHLEP's definition of "One Health"», *Who.int*. [En línia]. Disponible en: https://www.who.int/news/item/01-12-2021-tripartite-and-unep-support-ohhlep-s-definition-of-one-health. [Accedit: 29-jun-2023].

[7] N. I. Molina-Gómez, J. L. Díaz-Arévalo, y P. A. López-Jiménez, «Air quality and urban sustainable development: the application of machine learning tools», *Int. J. Environ. Sci. Technol. (Tehran)*, vol. 18, n.o 4, pp. 1029-1046, 2021. [Accedit: 29-jun-2023].

[8] M. Castelli, F. M. Clemente, A. Popovič, S. Silva, y L. Vanneschi, «A machine learning approach to predict air quality in California», *Complexity*, vol. 2020, pp. 1-23, 2020. [Accedit: 29-jun-2023].

[9] J. Fenger, «Urban air quality», *Atmos. Environ. (1994)*, vol. 33, n.o 29, pp. 4877-4900, 1999. [Accedit: 29-jun-2023].

[10] O. M. de la Salud, «Directrices mundiales de la OMS sobre la calidad del aire»,

Who.int, 2021. [En línia]. Disponible en: https://apps.who.int/iris/bitstream/handle/10665/346062/9789240035461-spa.pdf?sequence=1&isAllowed=y. [Accedit: 29-jun-2023].

[11] R. C. las C. y. M. D. Ministerio de la Presidencia, «Documento BOE-A-2023-2026», *Boe.es*, 25 de enero de 2023. [En línia]. Disponible en: https://www.boe.es/buscar/doc.php?id=BOE-A-2023-2026. [Accedit: 29-jun-2023].

[12] MITECO, «Dióxido de azufre», *Gob.es*. [En línia]. Disponible en: https://www.miteco.gob.es/es/calidad-y-evaluacion-ambiental/temas/atmosfera-y-calidad-del-aire/calidad-del-aire/salud/dioxido-azufre.aspx. [Accedit: 29-jun-2023].

[13] MITECO, «Monóxido de carbono», *Gob.es*. [En línia]. Disponible en: https://www.miteco.gob.es/es/calidad-y-evaluacion-ambiental/temas/atmosfera-y-calidad-del-aire/calidad-del-aire/salud/monoxido-carbono.aspx. [Accedit: 29-jun-2023].

[14] MITECO, «Óxidos de Nitrógeno», *Gob.es*. [En línia]. Disponible en: https://www.miteco.gob.es/es/calidad-y-evaluacion-ambiental/temas/atmosfera-y-calidad-del-aire/calidad-del-aire/salud/oxidos-nitrogeno.aspx. [Accedit: 29-jun-2023].

[15] MITECO, «Partículas», *Gob.es*. [En línia]. Disponible en: https://www.miteco.gob.es/es/calidad-y-evaluacion-ambiental/temas/atmosfera-y-calidad-del-aire/calidad-del-aire/salud/particulas.aspx. [Accedit: 29-jun-2023].

[16] MITECO, «Ozono», *Gob.es*. [En línia]. Disponible en: https://www.miteco.gob.es/es/calidad-y-evaluacion-ambiental/temas/atmosfera-y-calidad-del-aire/calidad-del-aire/salud/ozono.aspx. [Accedit: 29-jun-2023].

[17] «pandas», *Pydata.org*. [En línia]. Disponible en: https://pandas.pydata.org/. [Accedit: 29-jun-2023].

[18] «What is NumPy? — NumPy v1.8 Manual», *Scipy.org*. [En línia]. Disponible en: https://docs.scipy.org/doc/numpy-1.8.0/user/whatisnumpy.html. [Accedit: 29-jun-2023].

[19] Wikipedia contributors, «matplotlib», *Wikipedia, The Free Encyclopedia*. [En línia]. Disponible en: https://ca.wikipedia.org/w/index.php?title=Matplotlib&oldid=32044282. [Accedit: 29-jun-2023].

[20] «Requests: HTTP for humansTM — requests 2.31.0 documentation», *Readthedocs.io*. [En línia]. Disponible en: https://requests.readthedocs.io/en/latest/. [Accedit: 29-jun-2023].

[21] Wikipedia contributors, «Microsoft Excel», *Wikipedia, The Free Encyclopedia*. [En línea]. Disponible en:

https://es.wikipedia.org/w/index.php?title=Microsoft_Excel&oldid=151963426. [Accedit: 29-jun-2023].

[22] «Getting Started with Plotly in Python», *Plotly.com*. [En línea]. Disponible en: https://plotly.com/python/getting-started/#:~:text=Plotly%20is%20a%20free%20and,to%20some%20Basic%20Charts%20tutorials. [Accedido: 29-jun-2023].

[23] «Overview of Dash & Dash Apps», *Plotly.com*. [En línia]. Disponible en: https://plotly.com/dash/. [Accedit: 29-jun-2023].

[24] Wikipedia contributors, «Scikit-learn», *Wikipedia, The Free Encyclopedia*. [En línea]. Disponible en: https://ca.wikipedia.org/w/index.php?title=Scikit-learn&oldid=32102111. [Accedit: 29-jun-2023].

[25] J. Soto, «Introducción a TensorFlow», *Pharmacoecon. Span. Res. Artic.*, vol. 4, n.o S1, pp. 1-2, 2007. [Accedit: 29-jun-2023].

[26] S. Rae, «Keras», 2014. . [Accedit: 29-jun-2023].

[27] «XGBoost Documentation — xgboost 1.7.6 documentation», *Readthedocs.io*. [En línea]. Disponible en: https://xgboost.readthedocs.io/en/stable/. [Accedido: 29-jun-2023].

[28] Wikipedia contributors, «Power BI», *Wikipedia, The Free Encyclopedia*. [En línia]. Disponible en: https://es.wikipedia.org/w/index.php?title=Power_BI&oldid=148103813. [Accedit: 29-jun-2023].

[29] Wikipedia contributors, «Mean squared error», *Wikipedia, The Free Encyclopedia*, 15-dic-2022. [En línia]. Disponible en: https://en.wikipedia.org/w/index.php?title=Mean_squared_error&oldid=1127519968. [Accedit: 29-jun-2023].

[30] Wikipedia contributors, «Root-mean-square deviation», *Wikipedia, The Free Encyclopedia*, 20-jun-2023. [En línia]. Disponible en: https://en.wikipedia.org/w/index.php?title=Root-mean-square_deviation&oldid=1161024991. [Accedit: 29-jun-2023].

[31] Wikipedia contributors, «Mean absolute error», *Wikipedia, The Free Encyclopedia*, 19-may-2023. [En línia]. Disponible en: https://en.wikipedia.org/w/index.php?title=Mean_absolute_error&oldid=1155831004. [Accedit: 29-jun-2023].

[32] Wikipedia contributors, «Mean absolute percentage error», *Wikipedia, The Free Encyclopedia*, 18-may-2023. [En línia]. Disponible en:

https://en.wikipedia.org/w/index.php?title=Mean_absolute_percentage_error&oldid=1155495457. [Accedit: 29-jun-2023].

[33] Wikipedia contributors, «Coefficient of determination», *Wikipedia, The Free Encyclopedia*, 10-jun-2023. [En línia]. Disponible en: https://en.wikipedia.org/w/index.php?title=Coefficient_of_determination&oldid=1159534280. [Accedit: 29-jun-2023].

[34] «Introducció - Calidad Ambiental - Generalitat Valenciana», *Calidad Ambiental*. [En línia]. Disponible en: https://agroambient.gva.es/va/web/calidad-ambiental/introduccion. [Accedit: 29-jun-2023].

[35] A. d. València, «Datos por horas calidad aire 2016-2020». 15-feb-2023. [Accedit: 29-jun-2023].

[36] A. d. València, «Datos por horas calidad aire 2021-2022». 15-feb-2023. [Accedit: 29-jun-2023].

[37] «Dades històriques - Calidad Ambiental - Generalitat Valenciana», *Calidad Ambiental*. [En línia]. Disponible en: https://agroambient.gva.es/va/web/calidad-ambiental/datos-historicos. [Accedit: 29-jun-2023].

[38] Wikipedia contributors, «Coeficiente de correlación de Pearson», *Wikipedia, The Free Encyclopedia*. [En línia]. Disponible en: https://es.wikipedia.org/w/index.php?title=Coeficiente_de_correlaci%C3%B3n_de_Pearson&oldid=150164801. [Accedit: 29-jun-2023].

[39] P. Flach, *Machine learning: The art and science of algorithms that make sense of data*. Cambridge, England: Cambridge University Press, 2012. [Accedit: 29-jun-2023].

[40] «Sklearn.Preprocessing.StandardScaler», *scikit-learn*. [En línia]. Disponible en: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.htm. [Accedit: 29-jun-2023].

[41] «Sklearn.Preprocessing.PolynomialFeatures», *scikit-learn*. [En línia]. Disponible en: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.PolynomialFeatures.html. [Accedit: 29-jun-2023].

[42] «Sklearn.Linear_model.Ridge», *scikit-learn*. [En línia]. Disponible en: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html. [Accedit: 29-jun-2023].

[43] L. Breiman, *Random Forests*, vol. 45. 2001. [Accedit: 29-jun-2023].

[44] C. M. Bishop, *Pattern Recognition and Machine Learning*, 1.a ed. New York, NY: Springer, 2006. [Accedit: 29-jun-2023].

[45] Ian Goodfellow and Yoshua Bengio and Aaron Courville, *Deep learning*, vol. 26. MIT Press, 2016. [Accedit: 29-jun-2023].

[46] Wikipedia contributors, «Gradient boosting», *Wikipedia, The Free Encyclopedia*, 24-jun-2023. [En línia]. Disponible en: https://en.wikipedia.org/w/index.php?title=Gradient_boosting&oldid=1161733016. [Accedit: 29-jun-2023].

[47] «3.3. Metrics and scoring: quantifying the quality of predictions», *scikit-learn*. [En línia]. Disponible en: https://scikit-learn.org/stable/modules/model_evaluation.html. [Accedit: 29-jun-2023].

[48] «Documento consolidado BOE-A-2015-10565», *Boe.es*. [En línea]. Disponible en: https://www.boe.es/eli/es/l/2015/10/01/39/con. [Accedido: 29-jun-2023].

[49] «Documento consolidado BOE-A-2013-12887», *Boe.es*. [En línia]. Disponible en: https://www.boe.es/eli/es/l/2013/12/09/19/con. [Accedit: 29-jun-2023].

[50] «Documento consolidado BOE-A-2006-13010», *Boe.es*. [En línia]. Disponible en: https://www.boe.es/eli/es/l/2006/07/18/27/con. [Accedit: 29-jun-2023].

[51] «Documento BOE-A-2005-2528», *Boe.es*. [En línia]. Disponible en: https://www.boe.es/eli/es/ai/1998/06/25/(1). [Accedit: 29-jun-2023]

Annexes

Annex I. Relació i reflexió del treball amb els Objectius de Desenvolupament Sostenible de l'Agenda 2030.

Aquest projecte de fi de carrera està estretament relacionat amb dos Objectius de Desenvolupament sostenible de l'Agenda 2030: ODS 11, Ciutats i comunitats sostenibles, i ODS 13, Acció pel clima.

Aquest estudi es centra en l'avaluació de la qualitat de l'aire a la ciutat de València, amb els resultats obtinguts i les ferramentes visuals s'intenta generar visibilitat tant per a les administracions públiques com a la població de València, perquè es comencen a prendre mesures, evidenciant la necessitat de combatre contra els alts nivells de contaminants atmosfèrics.

A més, no sols es tracta la millora de la qualitat de l'aire, sinó també està molt lligat amb assolir una València sostenible i de qualitat per a tots els valencians i valencianes.

Per això, amb aquest projecte es busca aportar informació rellevant i útil per ajudar en la implementació de mesures que contribuisquen en assolir els ODS 11 i 13, especialment per a la ciutat de València.

Objectius de Desenvolupament Sostenible	**Baix**	**Mig**	**Alt**	**No procedeix**
ODS 1. Fi de la pobresa.				X
ODS 2. Fam cero.				X
ODS 3. Salut i benestar.		X		
ODS 4. Educació de qualitat.				X
ODS 5. Igualtat de gènere.				X
ODS 6. Aigua neta i sanejament.				X
ODS 7. Energia asequible i no contaminant.	X			
ODS 8. Treball decent i creixement econòmic.				X

ODS 9. Indústria, innovació i infraestructures.				X
ODS 10. Reducció de les desigualtats.				X
ODS 11. Ciutats i comunitats sostenibles.			X	
ODS 12. Producció i consum responsable.				X
ODS 13. Acció pel clima			X	
ODS 14. Vida submarina.				X
ODS 15. Vida d'ecosistemes terrestres.		X		
ODS 16. Pau, justícia i institucions sòlides.				X

Taula 18: Grau de relació del treball amb els objectius de desenvolupament sostenible de l'Agenda 2030.

Annex II. Estudi de dades faltants en funció de l'estació de medició i la variable per als distints períodes temporals.

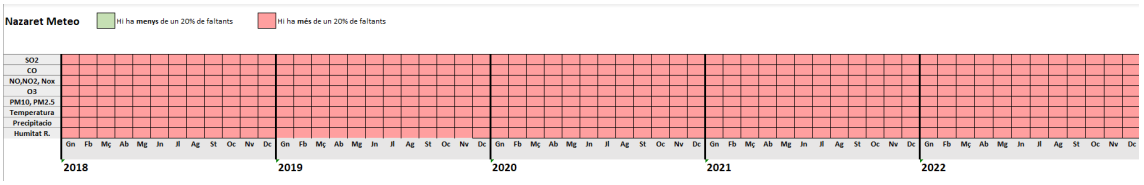


Figura 23: Percentatge de dades faltants per mes i any per a la estació Nazaret Meteo.

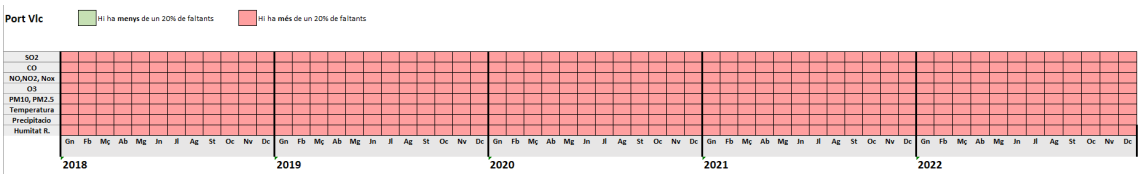


Figura 24: Percentatge de dades faltants per mes i any per a la estació Port València.

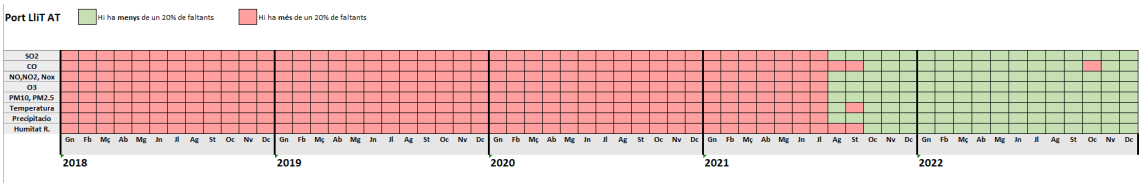


Figura 25: Percentatge de dades faltants per mes i any per a la estació Port Llit Antic Turia.

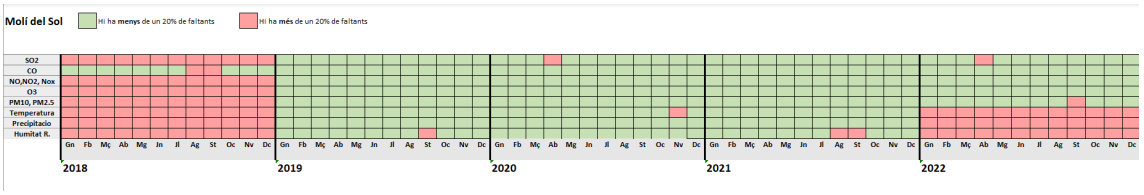


Figura 26: Percentatge de dades faltants per mes i any per a la estació Molí del Sol.

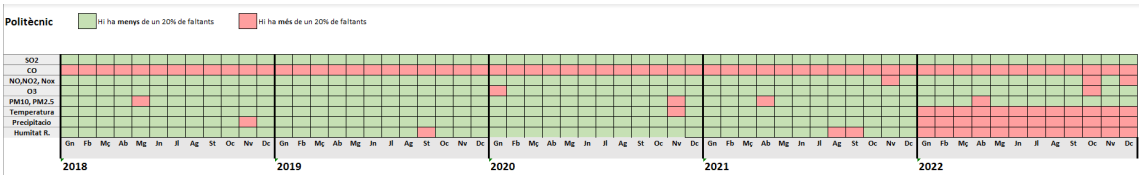


Figura 27: Percentatge de dades faltants per mes i any per a la estació Politècnic.

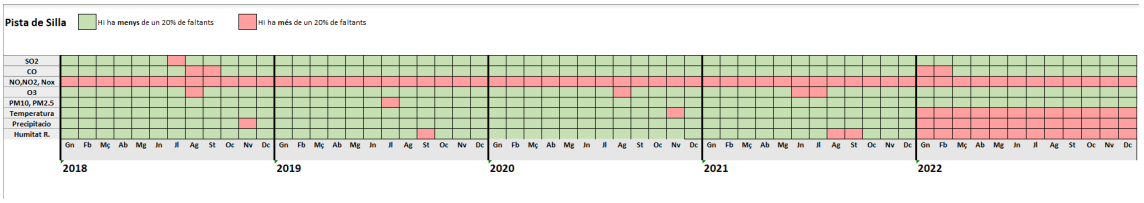


Figura 28: Percentatge de dades faltants per mes i any per a la estació Pista de Silla.

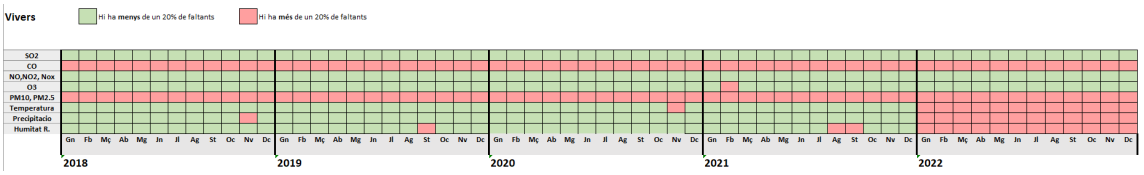


Figura 29: Percentatge de dades faltants per mes i any per a la estació Vivers.

Annex III. Correlacions entre variables per a l'estació Molí del Sol durant febrer desde 2020 fins a 2021.

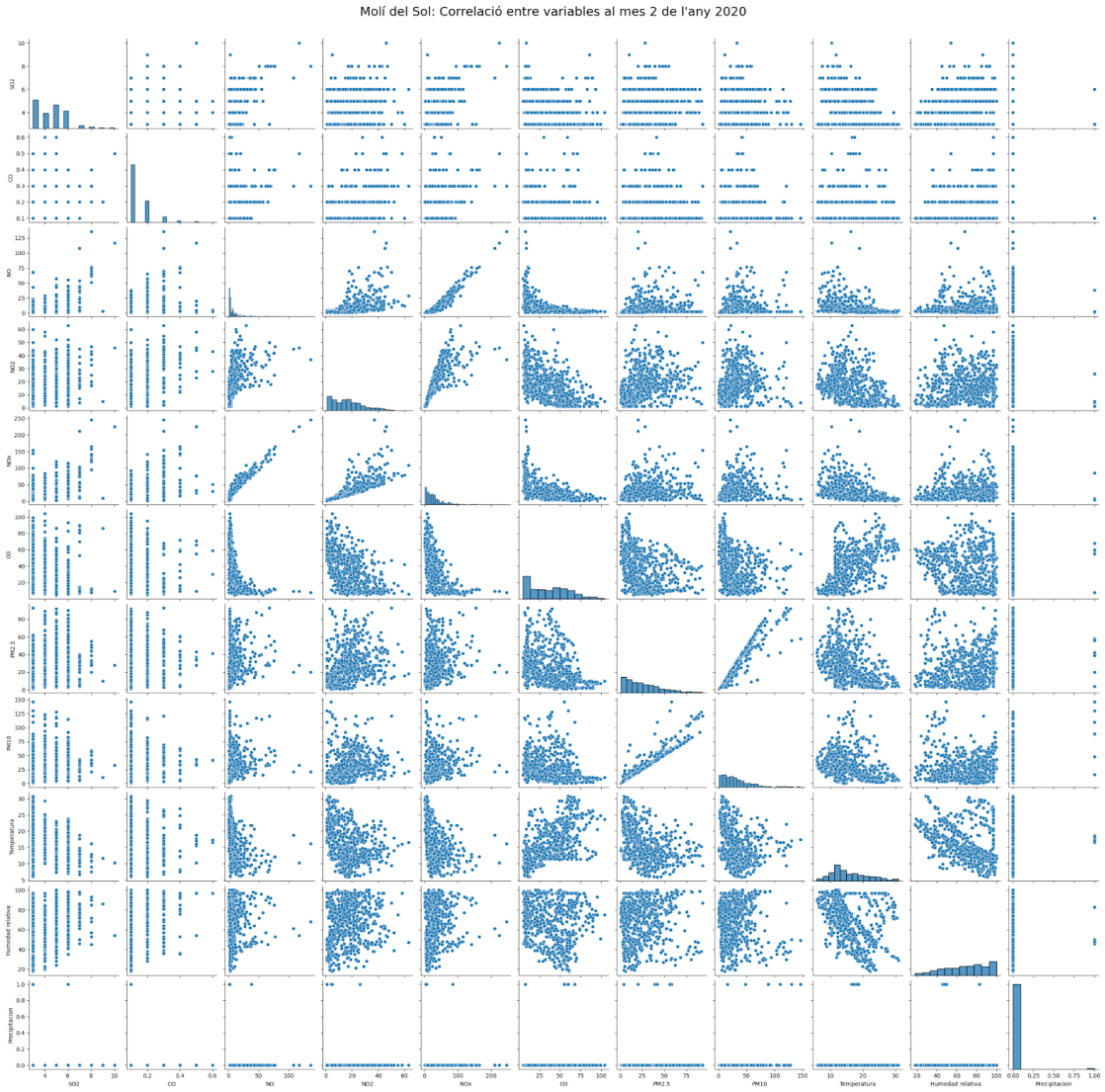


Figura 30: Gràfica Pairwise, relacions entre les parelles de variables per a febrer de 2020.

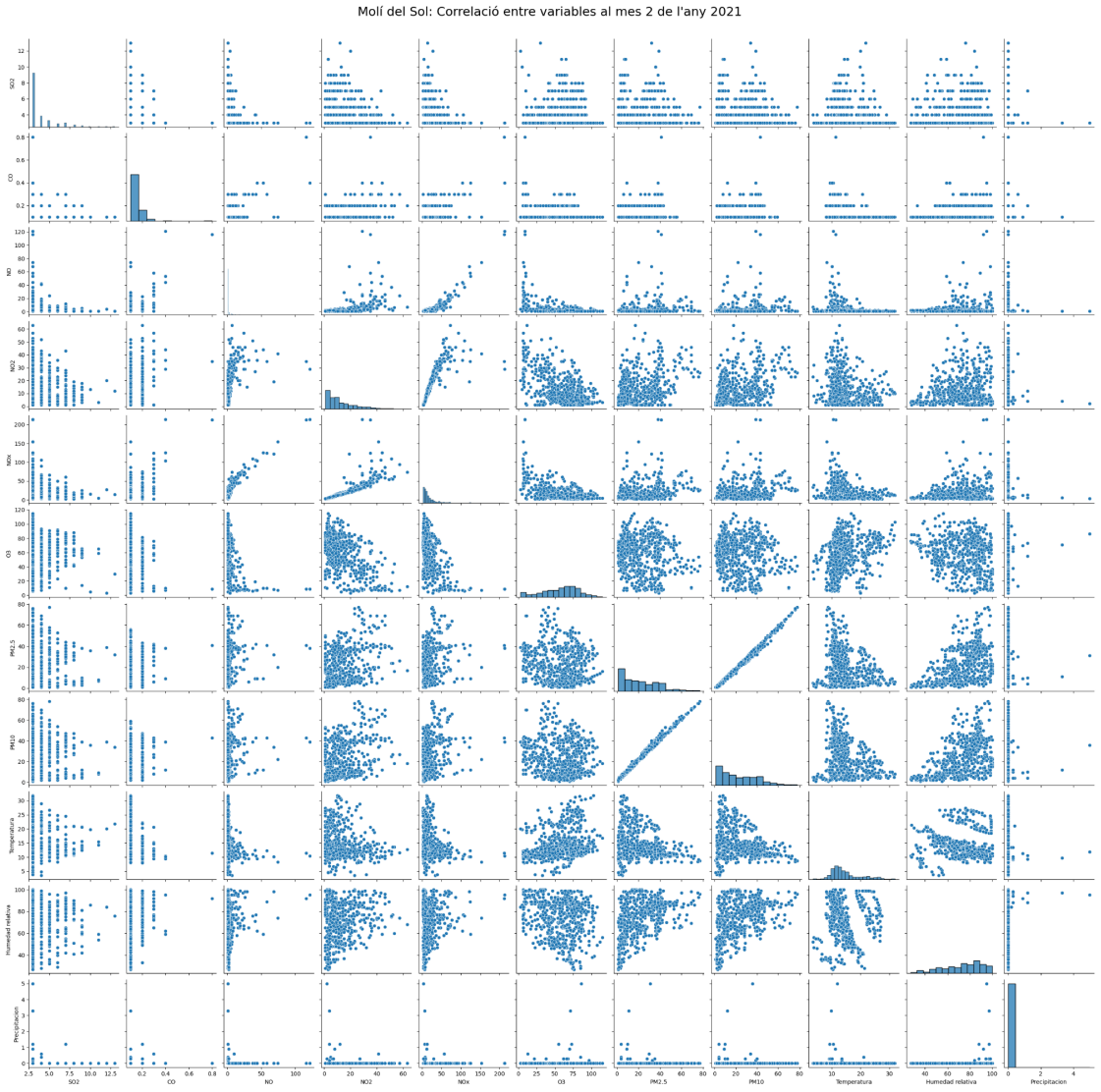


Figura 31: Gràfica Pairwise, relacions entre les parelles de variables per a febrer de 2021.

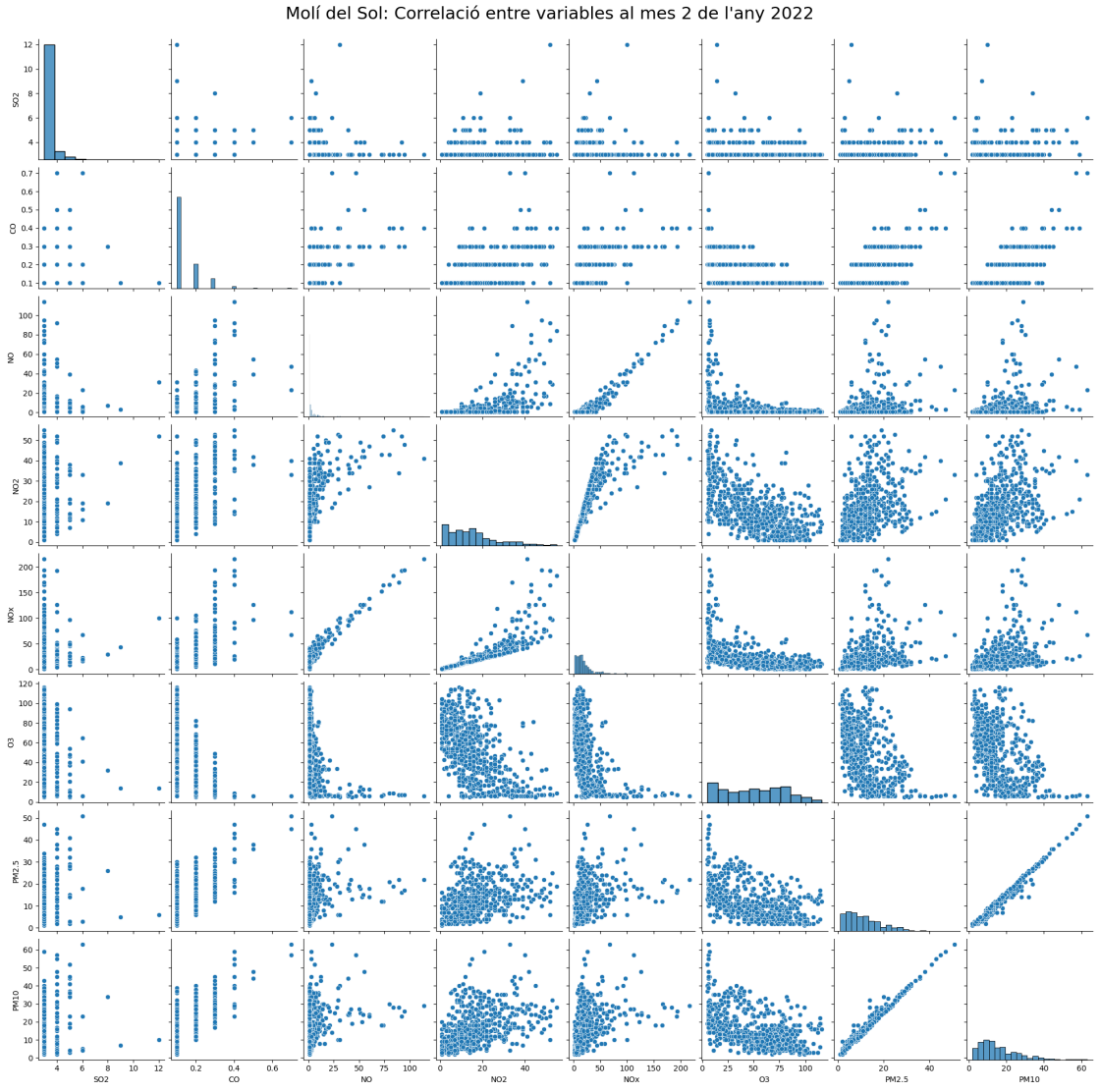


Figura 32: Gràfica Pairwise, relacions entre les parelles de variables per a febrer de 2022.

Annex IV. Dashboard amb dades reconstruïdes per al monòxid de carboni (NO) i els òxids de nitrogen (NO_X).

Figura 33: Dashboard de la variable NO en gener de 2019 a l'estació València Centre.

Figura 34: Dashboard de la variable NO_X en gener de 2019 a l'estació Av. França.

Annex V. Glossari terminològic

BOE: Boletí Oficial de l'Estat

CO: Monòxid de carboni

COVs: Compostos orgànics volàtils

FAO: Organización de las Naciones Unidas para la Alimentación y la Agricultura

GVA: Generalidad Valenciana

KNN: K veïns més propers

MITECO: Ministeri per a la Transició Ecològica o el Repte Demogràfic

ML: Machine learning (Aprenentatge automàtic)

mg/m³: Mil·ligrams per metres cúbics

µg/m³: Micrograms per metres cúbics

NO: Monòxid de nitrogen

NO_2: Diòxido de nitrógeno

NO_X: Òxids de nitrogen

ODS: Objectiu de desenvolupament sostenible

OMS: Organización Mundial de la Salud

O_3: Ozó

PM2.5: Partícules en suspensió de diàmetre inferior a 2.5 µg

PM10: Partícules en suspensió de diàmetre inferior a 10 µg

PNUMA: Programa de las Naciones Unidas para el Medio Ambiente

SO_2: Diòxid de sofre

WOAH: Organización Mundial de Sanidad Animal