Trabajo Final de Máster

Máster en Ingeniería Industrial

MODELO PREDICTIVO DE LA VARIACIÓN DE LOS NIVELES DE CONTAMINACIÓN EN FUNCIÓN DEL TRÁFICO URBANO UTILIZANDO HERRAMIENTAS DE INTELIGENCIA ARTIFICIAL

Memoria

Autor: Federico Aguilar Calvo
Directora: Mª Antonia de los Santos López
Codirector: Vicente César de Medina Iglesias

Convocatoria: 09/2017

Escuela Técnica Superior de Ingeniería Industrial de Barcelona

Resumen

Los elevados niveles de contaminación del aire en grandes ciudades como Barcelona, han provocado que surja la necesidad de contar con herramientas que permitan predecir la concentración de estos contaminantes, y poder así tomar medidas antes de que estos alcancen niveles peligrosos para la salud.

Por ello, en el presente trabajo, se detalla el procedimiento seguido para crear unos modelos, que mediante herramientas de Inteligencia Artificial, permitan predecir la concentración de los principales contaminantes, presentes en el aire: Dióxido de Sulfuro, Óxidos de Nitrógeno, Ozono, Monóxido de Carbono y Partículas en suspensión. Para elaborar estos modelos, se han utilizado los datos disponibles meteorológicos, de flujo de tráfico y de concentraciones de contaminantes.

Se ha partido de las predicciones hechas por regresión lineal simple, con las variables más correlacionadas con los contaminantes. A continuación, se ha elaborado una serie de Redes Neuronales Artificiales, constituidas de perceptrones multicapa y multivariable. Estas han sido entrenadas a partir de los datos utilizando *Machine Learning*. Y una vez entrenadas, se han utilizado para predecir la concentración de los contaminantes, obteniéndose con estos modelos, predicciones sustancialmente mejores.

En este proyecto, el potencial de los modelos ha estado limitado por el volumen de datos y la capacidad de computación disponibles. Aun así, se han obtenido muy buenos resultados, y aplicando mayores recursos, podría escalarse para obtener predicciones de gran fiabilidad.

Abstract

The high levels of pollution in the air of big cities such as Barcelona, have led to the need for tools to predict the concentration of these pollutants, to be able to take action before they reach levels that are dangerous to health.

Therefore, the present work describes the procedure followed to create models, which using Artificial Intelligence tools, predict the concentration of the main pollutants present in the air: Sulfide Dioxide, Nitrogen Oxides, Ozone, Carbon Monoxide and Particles in suspension. To elaborate these models, the available meteorological, traffic flow and concentration of pollutants data have been used.

Firstly, predictions by simple linear regression, with the variables most correlated with the pollutants, were made. Secondly, a series of Artificial Neural Networks, made up of multilayer and multivariate perceptrons, have been elaborated. These have been trained from the data using Machine Learning. And finally, they have been used to do predictions of the concentration of pollutants, obtaining with these models, substantially better predictions.

In this project, the potential of the models has been limited by the volume of data and the available computing capacity. Even so, very good results have been obtained, and by applying more resources, it could be scaled to obtain predictions of great reliability.

Abreviaturas

ECM	Error Cuadrático Medio
Conc	Concentración
HRM	Humedad Relativa Media
ML	Machine Learning
PM10	Partículas con diámetro aerodinámico menor que 10 µm
Ppt	Precipitaciones
RNA	Red Neuronal Artificial
t-X	X instantes de tiempo anteriores
Tª	Temperatura
TNR	True Negative Ratio: Ratio de negativos verdaderos
TPR	True Positive Ratio: Ratio de positivos verdaderos

Contenido

Índice de figuras

Índice de tablas

1. Introducción

1.1. Motivación

El aumento de los niveles de contaminación en Barcelona ha provocado una creciente preocupación de los impactos de esta en la salud. Otras ciudades de España, como Madrid, ya han tenido que tomar medidas drásticas, como limitar la circulación cuando la concentración de contaminantes ha llegado a sobrepasar ciertos niveles críticos.

En una ciudad como Barcelona, que obtiene gran parte de sus recursos económicos del turismo, resulta especialmente crítico conservar unos niveles aceptables de calidad del aire. Ya que de lo contrario, además de resultar dañada la salud de los habitantes podría repercutir negativamente en la economía local.

Esta creciente preocupación de la población ha provocado la toma de conciencia desde las instituciones para tratar de controlarlo. Para lo cual resulta necesario contar con modelos de predicción que permitan identificar cuando resulta conveniente alguna actuación.

1.2. Justificación

Nos encontramos en un punto en el que la creciente preocupación por el impacto para la salud provoca que desde el Ayuntamiento busquen nuevas formas de poder combatir los niveles de contaminación en la ciudad.

El desarrollo del concepto de Smart City en ciudades como Barcelona, implica que la ciudad esté cada vez más digitalizada y la cantidad de datos fiables y estandarizados recogidos a lo largo de toda la ciudad es cada vez mayor.

Esto unido al creciente desarrollo de modelos predictivos, utilizando herramientas de Inteligencia Artificial, nos permite entender estos datos y su interrelación. Esto justifica el interés por desarrollar modelos que aprovechen la información de estos datos para crear potentes herramientas predictivas.

1.3. Objetivos

El objetivo del presente estudio, es desarrollar una serie de modelos que permitan predecir, con la mayor exactitud posible, la concentración de contaminantes en el aire en la ciudad.

En primer lugar, se analizarán los datos disponibles con el fin de identificar aquellos que resultan más relevantes para este estudio y prepáralos para poder utilizarlos en los modelos. A continuación, se desarrollará un modelo predictivo utilizando técnicas convencionales.

Por último, se crearan unas máquinas predictivas utilizando Redes Neuronales Artificiales, entrenadas utilizando *Machine Learning*, con las que obtener predicciones más exactas que con las técnicas anteriores.

2. Contaminación, problemática e impacto

Numerosos estudios demuestran la responsabilidad de los altos niveles de contaminación en el aire en el aumento de la mortalidad en ciudades con una concentración de contaminantes elevada.

Se estima [1], que los niveles diarios de Partículas en suspensión menores de 10 micras, en adelante PM10, por encima de 50 µg/m3 son responsables de 1,4 muertes prematuras por 100.000 habitantes y año debido a sus efectos a corto plazo, y de 2,8 muertes/100.000 en un periodo de hasta 40 días tras la exposición. A largo plazo, el número de muertes prematuras atribuibles a la contaminación media anual de PM10 por encima de 20 µg/m3 es 68/100.000.

Además de estas partículas, los habitantes de las ciudades también están expuestos a otra serie de contaminantes, que dañan a largo plazo seriamente su salud.

Cierto nivel de exposición es inevitable, ya que en un entorno urbano a día de hoy es imposible contar con un aire completamente libre de contaminantes. Pero mientras que niveles bajos de concentración de estos componentes en el aire, producen un bajo impacto, si alcanzan concentraciones por encima de los límites prudenciales, el impacto para la salud aumenta a niveles realmente preocupantes [2].

Por ello resulta crítico contar con modelos que alerten de situaciones en las que estos límites van a ser rebasados, y así poder actuar en consecuencia.

En este estudio se pretende desarrollar un modelo predictivo para los siguientes contaminantes: Dióxido de Sulfuro, Monóxido de Nitrógeno, Dióxido de Nitrógeno, Ozono, Monóxido de Carbono y Partículas en suspensión menores de 10 micras. Estos son los contaminantes más dañinos presentes en el aire de las ciudades. En la Figura 1, pueden verse las principales fuentes de estos contaminantes [3].

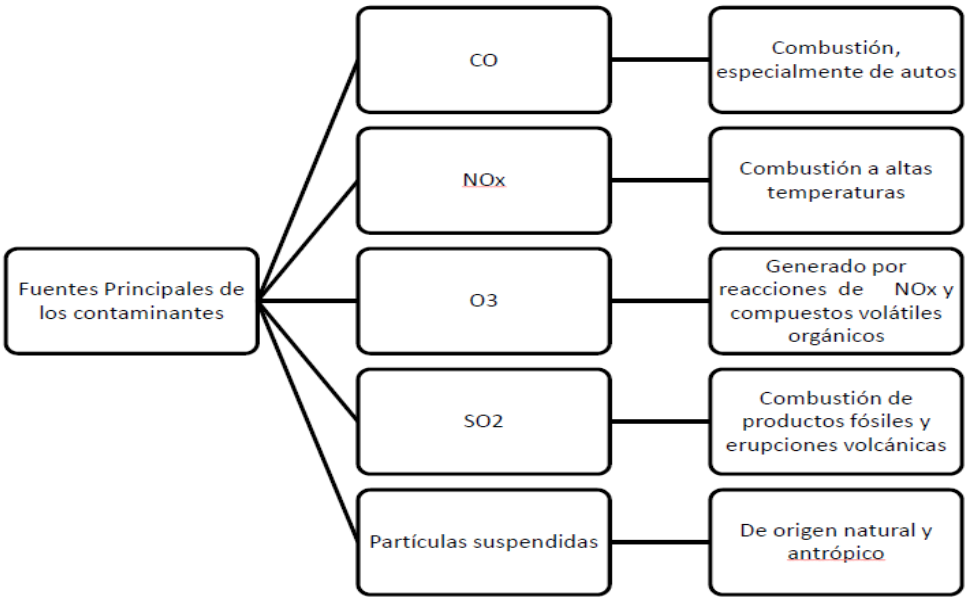


Figura 1 Principales fuentes de contaminantes

Partículas PM10

Son una mezcla de partículas sólidas y liquidas, que se encuentran en suspensión en el aire, formadas por diferentes compuestos, tanto orgánicos como inorgánicos.

Todos estos compuestos en el aire tienen la capacidad de ser absorbidos a través del aparato respiratorio, por lo que tienen un gran potencial para causar daños a la salud.

Dióxido Sulfúrico (SO_2)

El dióxido sulfúrico es un gas irritante y tóxico. Afecta principalmente a las mucosas y los pulmones. La exposición de altas concentraciones por cortos periodos de tiempo puede irritar el tracto respiratorio, causar bronquitis y congestionar los conductos bronquiales de quienes sufren de problemas respiratorios como el asma.

El SO_2 es uno de los causantes de la lluvia ácida, cuyos efectos son especialmente dañinos para las ciudades.

Óxidos de Nitrógeno (NOx)

Los compuestos NO_2 y NO constituyen los dos óxidos de nitrógeno más importantes desde el punto de vista toxicológico, siendo el primero de ellos el más nocivo.

La exposición aguda a NO_2 puede provocar lesiones en las vías respiratorias y pulmones, ocasionando una reducción de la capacidad pulmonar y un aumento en la sensibilidad a

alérgenos. Y en el caso de exposiciones prolongadas puede provocar cambios irreversibles en la estructura y función de los pulmones.

En los países desarrollados, la contaminación por NOx se debe fundamentalmente a los vehículos de combustión interna, por lo que la concentración de estos es especialmente preocupante en los puntos donde hay mayor concentración, como en las ciudades.

Ozono (O_3)

El ozono es un gas incoloro, producido por reacciones de hidrocarburos y óxidos de nitrógeno bajo la influencia de la luz solar. Es un contaminante que agrede a las mucosas e irrita el tracto respiratorio y los ojos, facilitando la acción de virus y bacterias.

Es uno de los principales contaminantes atmosféricos presentes en zonas altamente industrializadas y en las ciudades con un número alto de automóviles.

Monóxido de Carbono (CO)

El monóxido de carbono es un gas inodoro, incoloro, inflamable y altamente toxico que en concentraciones altas puede llegar a ser letal.

La principal fuente antropogénica de monóxido de carbono es la quema incompleta de combustibles como la gasolina por falta de oxígeno.

3. Obtención de los datos

3.1. Datos disponibles

Para la realización de este estudio se han analizado los diferentes datos a los que se ha podido tener acceso, y que a su vez resultasen relevantes para predecir la concentración de los distintos contaminantes estudiados.

Estos contaminantes proceden de distintas fuentes como Ayuntamiento de Barcelona, la Generalidad de Cataluña y la AEMET. Este estudio presta principal atención a la influencia del tráfico urbano, ya que la emisión de otros focos, son parámetros que difícilmente se pueden llegar a controlar desde un Ayuntamiento.

Los datos utilizados se agrupan en 3 clases, por su naturaleza y a la vez por la procedencia de estos:

- Datos de contaminantes
- Datos meteorológicos
- Datos de circulación de tráfico

Datos de contaminantes

Los contaminantes, que con este trabajo se quieren predecir, son los medidos por los aforadores de la Generalidad de Cataluña [4]. Hay varios medidores repartidos por la ciudad, pero para este estudio se ha elegido aquel que se encontraba más cerca de los medidores de flujo de tráfico, de los que se hablará más adelante. En la Figura 2 podemos ver su situación.

Figura 2 Localización de los medidores de tráfico y contaminación

El aforador toma medidas de la concentración de Dióxido de Sulfuro, Monóxido de Nitrógeno, Dióxido de Nitrógeno, Ozono, Monóxido de Carbono y Partículas menores de 10 micras. Las unidades en las que se toman estas medidas se muestran en la Tabla 1.

Tabla 1 Unidades de medida de los diferentes contaminantes

SO_2	NO	NO_2	O_3	CO	PM10
$\mu g/m^3$	$\mu g/m^3$	$\mu g/m^3$	$\mu g/m^3$	mg/m^3	$\mu g/m^3$

Datos meteorológicos

Se dispone de datos meteorológicos obtenidos de la AEMET [5]. Se han obtenido los siguientes parámetros: temperatura máxima, temperatura mínima, temperatura media, índice de humedad relativa, medida de precipitaciones, velocidad media del viento, rachas máximas de viento, dirección del viento, presión atmosférica e irradiación solar.

Todos estos parámetros se han utilizado a priori. Más adelante, tras analizar la dependencia de estos sobre los niveles de contaminantes en el aire, se han descartado aquellas variables meteorológicas que introducían más ruido al sistema que información aportaban.

Datos de circulación de tráfico

En cuanto a las medidas del tráfico, se han utilizado los datos aportados por el Ayuntamiento de Barcelona de sus aforadores. Estos miden la intensidad del tráfico, en vehículos por hora, y la Ocupación de la vía, en tanto por mil.

Se han seleccionado los medidores del tráfico y de contaminantes que estaban más concentrados en torno a un punto. En la Figura 2 puede verse la posición de cada uno de estos. Todos ellos cerca del cruce de la Avenida de Roma con la Calle de Aragón.

A la hora de utilizar estos datos para el modelo predictivo, después de analizar diferentes alternativas, la más conveniente ha resultado ser tomar los datos del medidor de flujo de tráfico situado en la calle que conduce al medidor de contaminación (4071) y un valor medio de todos los medidores.

De los 3 grupos, el limitante en cuanto a volumen de datos ha sido el conjunto de datos de tráfico. Ya que estos no estaban disponibles públicamente, pero el Ayuntamiento ha facilitado los datos de todos los medidores de esta zona para un intervalo de 4 semanas.

Se tiene en cuenta en este tipo de estudios predictivos, mientras más datos se tengan mejores resultados se obtendrán. Tras analizar el conjunto disponible se ha observado que el volumen de la muestra era lo suficientemente extenso como para poder elaborar el modelo.

En un principio se valoraron dos posibilidades a la hora de elegir la escala temporal a la que trabajaría el modelo. Una opción era diseñar el modelo para que predijese el valor del parámetro a determinar, en el siguiente rango de tiempo disponible. Como hay un registro de todas las variables cada hora, la predicción sería a una hora. Por otro lado, estaba la opción de predecir por días, buscando el valor máximo del día siguiente. Esta última opción tenía más sentido desde el punto de vista de la utilidad práctica de este modelo, pero el inconveniente, es que al tener datos de 28 día, solo se contaría con 28 registros, lo cual resulta insuficiente para poder elaborar un buen modelo predictivo. Para la opción horaria se tienen 672 registros, con lo que ya resultaría viable.

3.2. Limpieza de los datos

La primera parte del trabajo con estas medidas ha consistido en realizar la tarea de minado y limpieza de datos. Se han sacado los datos de las distintas fuentes homogeneizando formatos y unificando los rangos temporales de las medidas, ya que no todos los registros estaban tomados con los mismos intervalos.

Ha sido necesario reparar algunos vacíos que había en los datos. Estos eran pocos y en zonas no críticas, principalmente en horas nocturnas de días de bajo tráfico, por lo que se puede considerar que no perjudican la elaboración del modelo predictivo. La Figura 3a muestra un ejemplo de un caso con vacíos. Para solventarlo, se toma como referencia la misma distribución de otros días con las mismas características en la misma franja horaria y siendo consistente con las medias anteriores y posteriores (ver Figura 3b), y se rellenan los huecos aplicando dicha distribución (ver Figura 3c).

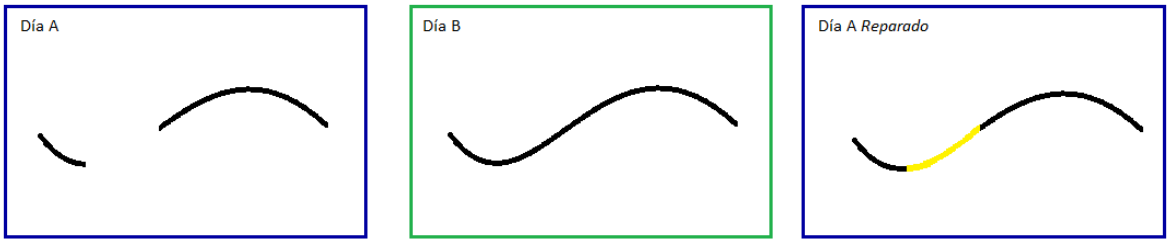


Figura 3a Caso A con vacíos *Figura 3b Caso B completo* *Figura 3c Caso A reparado*

Figura 3 Reparación de datos

3.3. Normalización de los datos

Para poder trabajar con los datos, comparando su evolución temporal y sus interrelaciones, ha sido necesario normalizarlos, ya que al ser medidas de distinta naturaleza y orden de magnitud, no eran comparables. Unas variables tienen unidades de concentración, mientras que otras son grados, velocidades, porcentajes, etc.

Una primera posibilidad era normalizar los datos originales (Figura 4a) de manera que estos quedasen con valores entre 0 y 1, pero manteniendo las misma distribución, como sucede en la Figura 4b. Se ha utilizado un set de datos homogéneos de ejemplo DatosA.

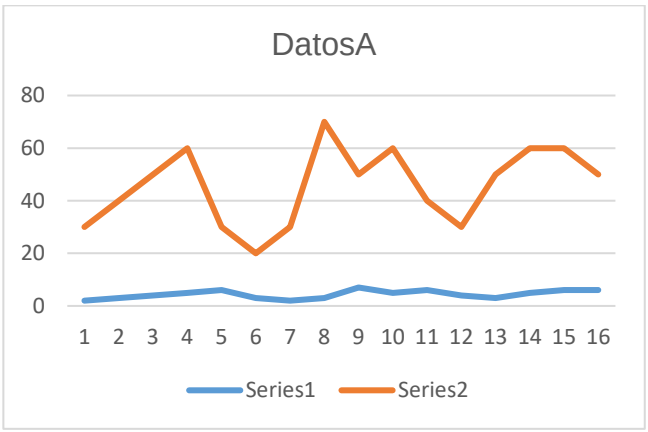


Figura 4a DatosA

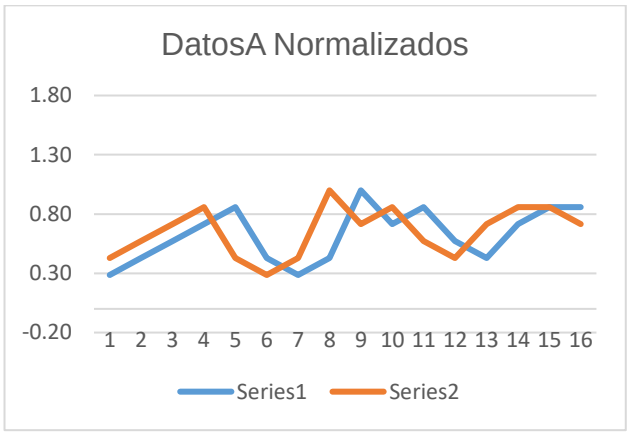


Figura 4b DatosA normalizados 01

Figura 4 Normalización de datos homogéneos entre 0 y 1

Esto resultó no ser idóneo ya que al ser medidas experimentales y muy variables, algunos puntos muy por encima o debajo del resto, distorsionaban completamente las escalas de las series .En la Figura 5a, puede verse un ejemplo de esto analizando un conjunto de datos de ejemplo DatosB, que a diferencia del conjunto de datos de ejemplo DatosA, cuentan con un registro notablemente superior al resto de puntos. Aplicando esta normalización se perdía la correlación entre las dos variables (ver Figura 5b):

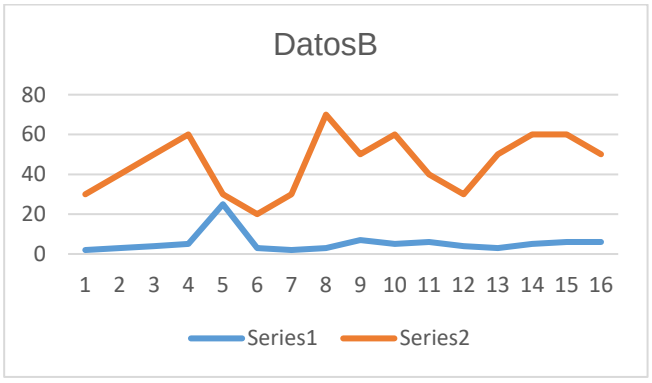


Figura 5a DatosB

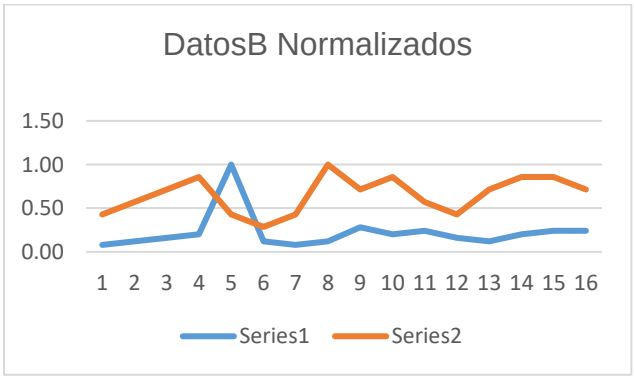


Figura 5b DatosB normalizados 01

Figura 5 Normalización de datos no homogéneos entre 0 y 1

El mismo problema ocurría al normalizar los datos entre 1 y -1.

En su lugar, se han normalizado los datos dividiendo todos ellos entre su promedio. De esta manera, como puede verse en la Figura 6 las medidas extremas se salen del rango, pero el resto de puntos resulta comparable entre distintos datos. Y también sigue siendo útil para comparar datos homogéneos sin distorsionar.

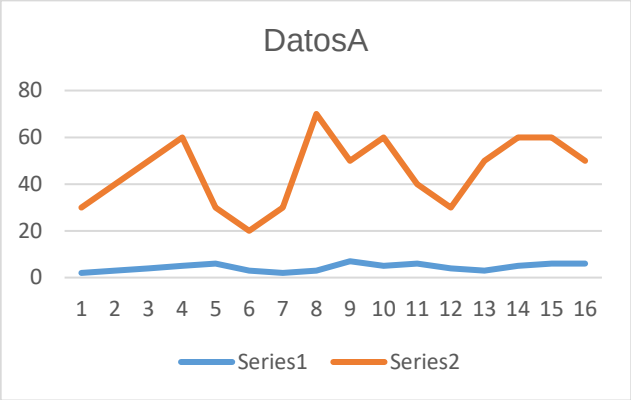


Figura 6a DatosA

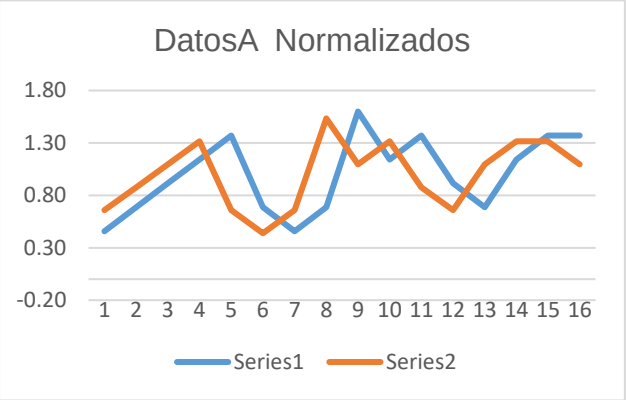


Figura 6b DatosA normalizados por promedio

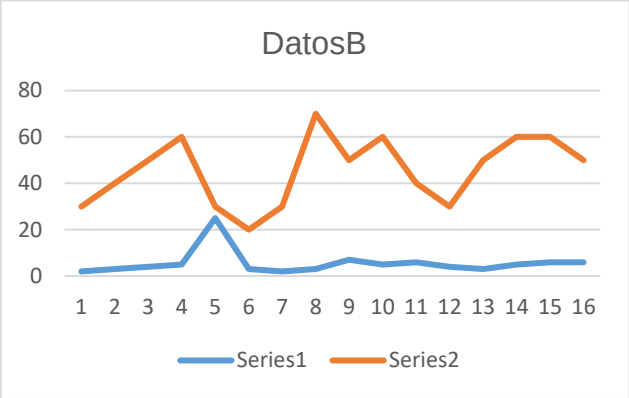


Figura 6c DatosB

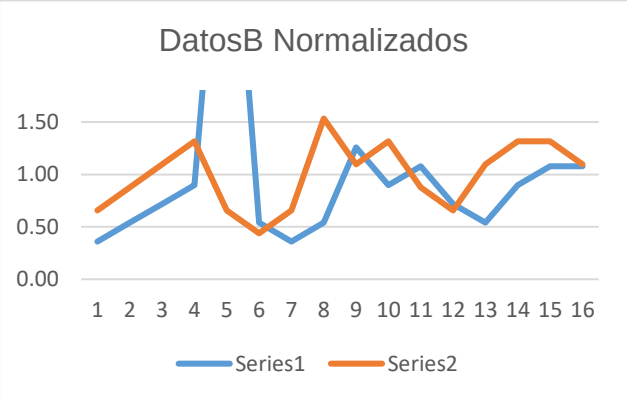


Figura 6d DatosB normalizados por promedio

Figura 6 Normalización de datos con valores no homogéneos por promedio

4. Métodos de predicción convencionales

Una vez que se ha tenido los datos preparados para su análisis, se ha comenzado a analizar las correlaciones entre estos. En este punto se busca encontrar aquellos datos que principalmente determinen los niveles de contaminantes. Y si es posible establecer una relación entre estos, que nos permita mediante una fórmula obtener las predicciones deseadas sin necesidad de recurrir a métodos más complejos.

El objetivo de esta parte del estudio, es determinar si para algún contaminante, su concentración en el aire está completamente determinada siguiendo alguno de los datos disponibles. Por lo que para ese contaminante, no aportaría valor el desarrollar un modelo de mayor complejidad, como una Red Neuronal Artificial (RNA), ya que se predeciría con una relación directa con el dato del que dependiese.

4.1. Modelos específicos ya desarrollados

Existen modelos de predicción desarrollados específicamente para medir la concentración de algún contaminante, y la evolución de esta [6]. Pero estos han sido desarrollados para localizaciones concretas, con determinados focos emisores y factores ambientales. Por lo que estos no serían apropiados para este caso.

Además, todos estos métodos trabajan principalmente sobre los datos de los contaminantes y los parámetros ambientales, pero no con datos que controlen el estado de los focos emisores [7]. Para este estudio estos resultan ser los de mayor interés, ya que se cuenta con datos de los medidores de intensidad de circulación y ocupación de carril. Y estos están directamente relacionados con la emisión de contaminantes por los vehículos con motor de combustión interna. Por ello, resulta más apropiado aplicar modelos empíricos estadísticos, creados específicamente a partir de estos datos.

4.2. Evaluación de las Predicciones

Un aspecto importante cuando se mide el desempeño de un método de predicción es tener una métrica adecuada que permita comparar los resultados de diferentes métodos. Se separarán las predicciones de los diferentes modelos en dos grupos, según sean de regresión logística o lineal.

En los predictores para regresión logística, se obtienen resultados binarios, que se compararán con los valores empíricos en una matriz de confusión.

Verdadero Positivo	Falso Positivo
Falso Negativo	Verdadero Negativo

Por otro lado, en las predicciones para regresión lineal, se ha tomado el error medio cuadrático como parámetro para comparar los resultados de diferentes modelos:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(Y_p - Y_0\right)^2$$

4.3. Predicción con regresión lineal simple

El primer modelo predictivo desarrollado está basado en las correlaciones entre los contaminantes y los diferentes parámetros disponibles, tanto para el mismo instante de medida, como con registros anteriores de estos.

Se ha comenzado por obtener la matriz de correlaciones de todas las variables. Tras lo cual, se han analizado con mayor profundidad los resultados con índices de correlación significativos. La matriz completa puede encontrarse en el Anexo A.

Se observa que los distintos contaminantes están correlacionados con varios de los parámetros disponibles. En la Tabla 2 pueden verse los datos más relevantes de la matriz de correlaciones:

Tabla 2 Matriz de correlaciones abreviada

	SO_2	NO	NO_2	O_3	CO	PM10
SO_2	1.00					
NO	0.30	1.00				
NO_2	0.32	0.69	1.00			
O_3	0.08	-0.51	-0.68	1.00		
CO	0.35	0.74	0.73	-0.32	1.00	
PM10	0.47	0.46	0.45	-0.05	0.57	1.00
Tª media	0.49	0.09	0.02	0.32	0.19	0.55
HRM	-0.27	-0.18	-0.18	-0.11	-0.24	-0.21
Ppt	-0.16	-0.02	-0.08	-0.09	-0.17	-0.23
Velocidad Viento	0.15	0.03	-0.16	0.47	0.01	0.31
Dirección Viento A	-0.03	-0.16	-0.07	0.18	-0.06	-0.04
Dirección Viento B	0.19	0.08	0.12	0.13	0.16	0.32
Dirección Viento C	0.08	0.19	0.06	-0.16	0.06	0.05
Dirección Viento D	-0.29	-0.23	-0.12	-0.08	-0.26	-0.32
Máx. viento	0.11	0.01	-0.19	0.46	-0.02	0.27
Presión Atmosférica	0.20	0.06	0.14	0.01	0.26	0.09
Irradiancia Solar	0.37	0.31	0.09	0.08	0.20	0.39
Intensidad Media	0.31	0.38	0.37	0.15	0.53	0.49
Ocupación Media	0.26	0.32	0.31	0.11	0.42	0.39
Intensidad 4071	0.30	0.39	0.37	0.14	0.54	0.48
Ocupación 4071	0.11	0.25	0.28	0.02	0.38	0.24

Se observa que algunas medidas no son muy relevantes o son redundantes debidas a su fuerte correlación con otras. La presión atmosférica no se considera que aquí aporte valor, por lo que no se seguirá teniendo en cuenta. La racha máxima de viento tampoco aporta mucha información y está fuertemente correlaciona con la velocidad media, por lo que tampoco se utilizará para el modelo predictivo.

Se tienen 4 valores de la dirección del viento, estos son medidas de cuanto sopla el viento en 4 direcciones desplazadas 90 grados entre ellas.

Por otro lado las precipitaciones, aunque afecten en gran medida a la concentración de contaminantes en el aire, no está muy correlacionada con estos. Esto es debido a que las lluvias durante la franja temporal estudiada estuvieron muy concentradas en tres puntos, y para la mayor parte del registro estas medidas son igual a cero independientemente del nivel de contaminantes, de ahí la baja correlación.

Partículas PM10

Para las PM10 se han analizado las cuatro variables más correlacionadas, tal y como se observa en la Tabla 3.

Tabla 3 Principales correlaciones de las PM10

PM10	
CO	0,57
Tª media	0,55
Intensidad Media	0,49
Intensidad 4071	0,48

La concentración de las partículas está principalmente correlacionada con la concentración de CO, con la Temperatura media y con las intensidades del flujo de tráfico. En la Figura 7 pueden observarse estos parámetros graficados frente a la concentración de PM10, con la línea de tendencia en azul oscuro, y el valor de límite de la concentración del contaminante en rojo.

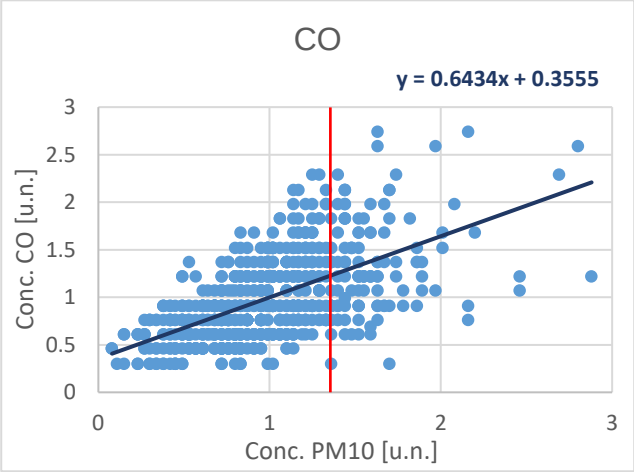


Figura 7a Correlación entre la concentración de CO y la concentración de PM10

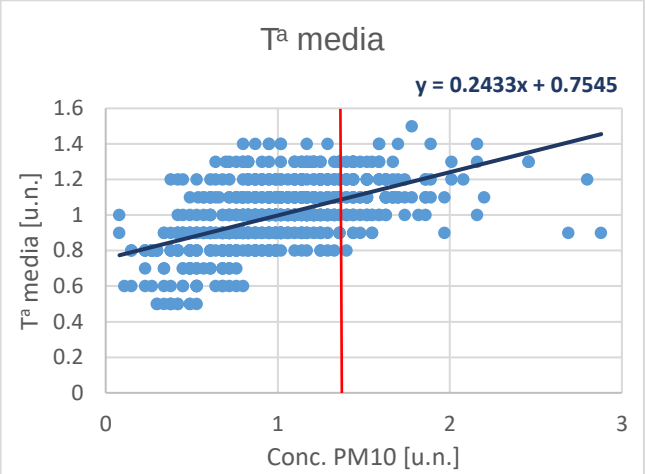


Figura 7b Correlación entre la temperatura media y la concentración de PM10

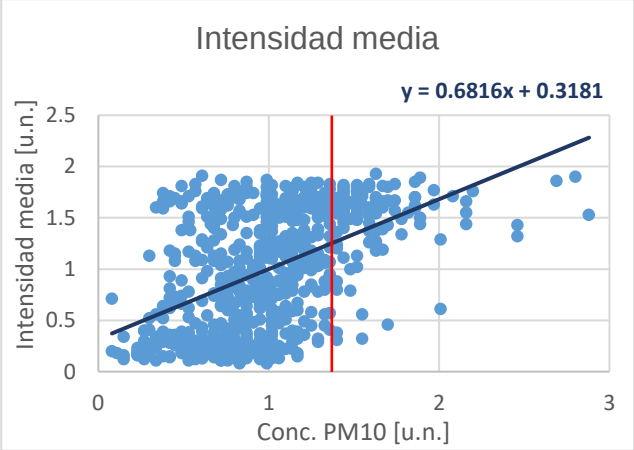


Figura 7c Correlación entre la intensidad media del tráfico y la concentración de PM10

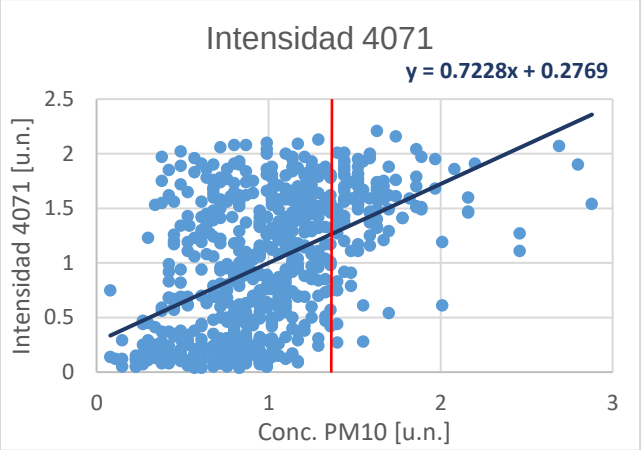


Figura 7d Correlación entre la intensidad del tráfico (4071) y la concentración de PM10

Figura 7 Gráficas de las principales correlaciones de la concentración de PM10

Utilizando las fórmulas obtenidas ajustando una recta a la dispersión de los datos, se ha elaborado predicciones del contaminante analizado en este punto. Para valorar la precisión de este método, se ha obtenido la matriz de confusión reflejada en Tabla 4 y el error medio cuadrático para cada variable (Tabla 5). En la matriz de confusión, puede analizarse la precisión de este método para predecir si el contaminante estará por encima o debajo del valor límite, 35 µg/m3 en este caso. Este límite corresponde a 1,33 en unidades normalizadas, marcado con una línea roja en la Figura 7.

Tabla 4 Matriz de confusión de las variables correlacionadas con las PM10

	VERDADERO				FALSO			
	CO	Tª media	Intensidad Media	Intensidad 4071	CO	Tª media	Intensidad Media	Intensidad 4071
POSITIVO	79	91	100	97	151	164	176	180
NEGATIVO	397	384	372	368	45	33	24	27

Tabla 5 ECM de las variables correlacionadas con la concentración de PM10

Tª media	0,40
Intensidad Media	0,53
Intensidad 4071	0,56
CO	0,35

Una limitación de este estudio es que los modelos han sido entrenados a partir de un conjunto de datos empíricos. Por lo que estos están limitados por el volumen de las medidas. Para poder realizar un buen clasificador para cada contaminante estudiado, son necesarios suficiente datos para casos tanto superiores, como inferiores al límite, de manera que la RNA sea capaz de abstraer las circunstancias que tienen que darse para cada caso.

De los contaminantes estudiados, solo las PM10 tienen suficientes registros superiores e inferiores a los límites como para poder desarrollar el clasificador.

Esto no quiere decir que no se sobrepasen los límites de lo demás contaminantes en la ciudad de Barcelona. Si no que para este rango concreto de tiempo, en este punto específico de la ciudad, no se han sobrepasado lo suficiente como para poder extraer de estos datos las condiciones generales, con las que predecir en otros casos si se sobrepasar los límites.

Por otro lado, existen distintos límites para la concentración de estas partículas, ya que hay varios límites pudiéndose superar cada uno ellos una serie de veces al año, y además esto varía según el organismo. Como el objetivo de este estudio es analizar la posibilidad de desarrollar un predictor, y no estudiar si los niveles de contaminación de la ciudad cumple las normativas, se ha tomado el valor límite para el clasificar en 35 $\mu g/m^3$.

Dióxido de azufre (SO_2)

El análisis de las correlaciones del SO_2 se muestra en la Tabla 6.

Tabla 6 Principales correlaciones del SO_2

SO_2	
Tª media	0,49
PM10	0,47
Irradiancia	0,37
CO	0,35

La concentración del Dióxido de azufre está principalmente correlacionada con las concentraciones de PM10 y de CO, y con la Temperatura media y la Irradiación Solar. A continuación, en la Figura 8 pueden observarse estos parámetros graficados frente a la concentración de SO_2.

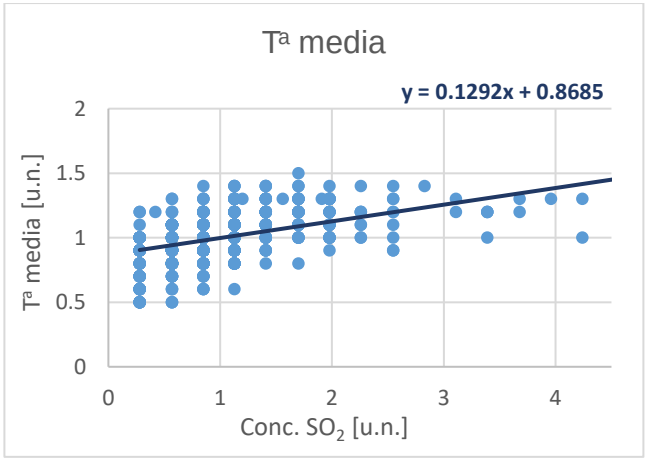


Figura 8a Correlación entre la temperatura media y la concentración de SO_2

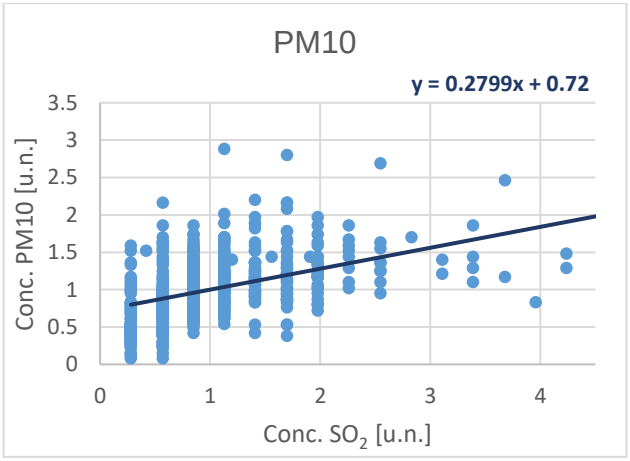


Figura 8b Correlación entre la concentración de PM10 y la concentración de SO_2

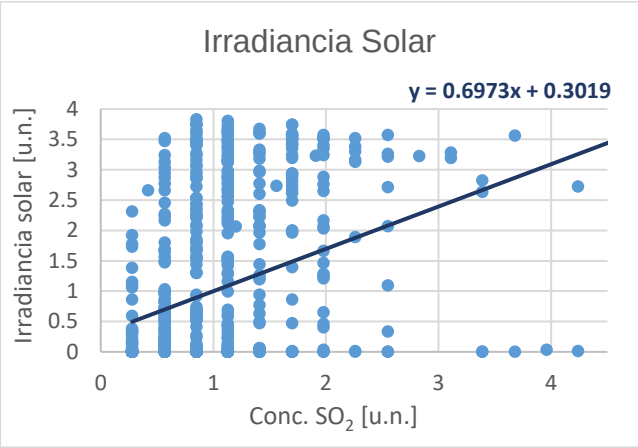


Figura 8c Correlación entre la Irradiancia solar y la concentración de SO_2

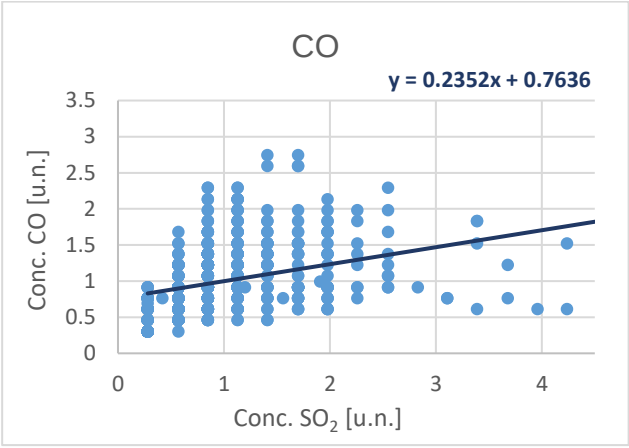


Figura 8d Correlación entre la concentración de CO y la concentración de SO_2

Figura 8 Gráficas de las principales correlaciones de la concentración de SO_2

Utilizando las fórmulas obtenidas ajustando una recta a la dispersión de los datos, se han elaborado predicciones del contaminante analizado en este punto. Para valorar la precisión de este método se ha calculado el error medio cuadrático para cada parámetro. Los resultados se pueden consultar en la Tabla 7.

Tabla 7 ECM de las variables correlacionadas con la concentración de SO_2

Tª media	1,53
PM10	1,70
Irradiancia Solar	2,95
CO	3,44

Monóxido de Nitrógeno (NO)

La Tabla 8 presenta los resultados del análisis de correlaciones para el NO.

Tabla 8 Principales correlaciones del NO

NO	
CO	0,74
NO_2	0,69
Intensidad 4071	0,39
Intensidad Media	0,38

La concentración de NO principalmente se correlaciona con las de CO y NO_2, y además con la intensidad del tráfico, tanto para el medidor 4071, como para la Media. En la Figura 9 pueden verse estos parámetros frente a la concentración de NO.

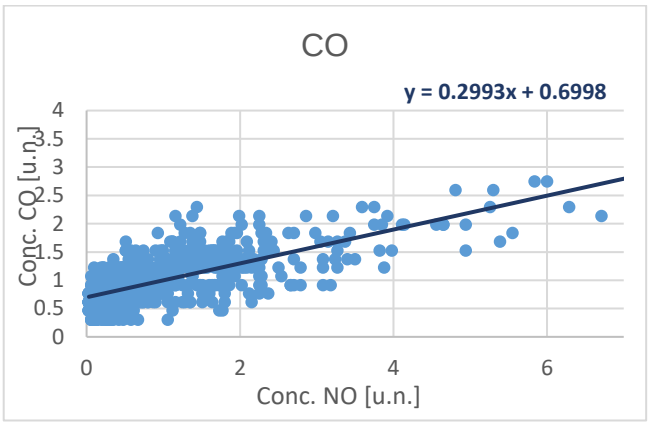


Figura 9a Correlación entre la concentración de CO y la concentración de NO

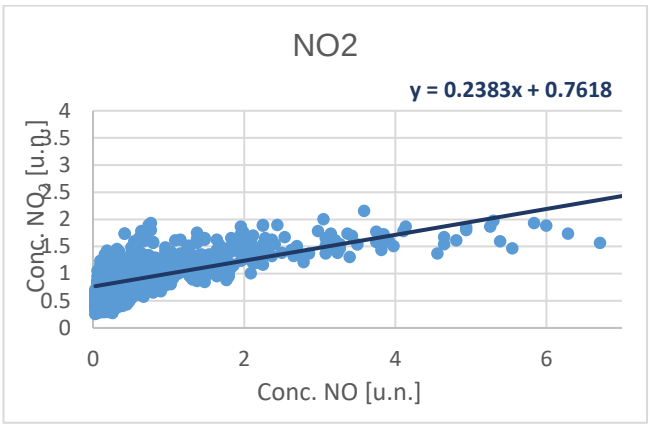


Figura 9b Correlación entre concentración de NO_2 y la concentración de NO

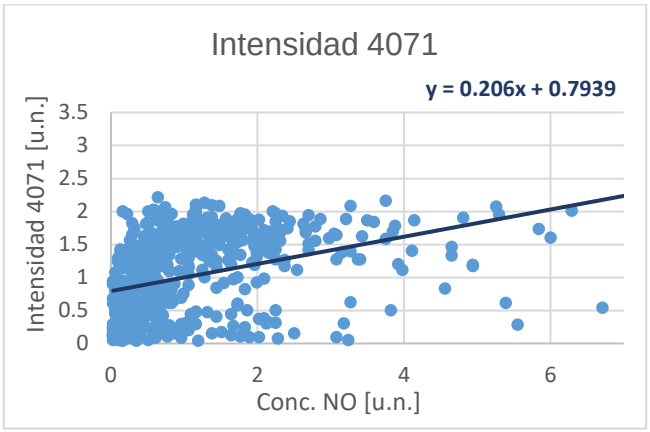


Figura 9c Correlación entre la intensidad 4071 y la concentración de NO

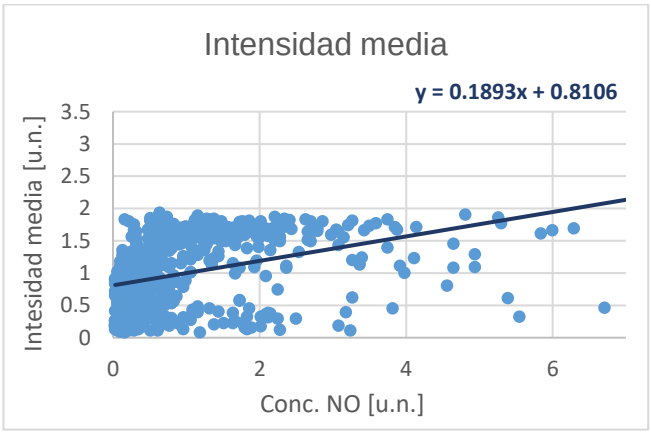


Figura 9d Correlación entre la intensidad media y la concentración de NO

Figura 9 Gráficas de las principales correlaciones de la concentración de NO

Se han elaborado asimismo, predicciones de los valores de NO para cada instante a través de las fórmulas obtenidas a partir de la recta que ajusta la dispersión de datos. Y a partir de estas se han calculado los errores medios cuadráticos de la Tabla 9.

Tabla 9 ECM de las variables correlacionadas con la concentración de NO

CO	1,08
NO_2	1,51
Intensidad 4071	7,61
Intensidad Media	7,81

Dióxido de nitrógeno (NO_2)

La Tabla 10 resume los resultados del análisis de correlaciones para el NO_2.

Tabla 10 Principales correlaciones del NO_2

NO_2	
CO	0,73
NO	0,69
O_3	-0,68
Intensidad 4071	0,37

La concentración de Dióxido de Nitrógeno está principalmente correlacionada con la concentración de CO, la de NO y la de O_3, y con la Intensidad del tráfico en el medidor 4071. En la Figura 10 se observan estos parámetros graficados frente a la concentración de NO_2.

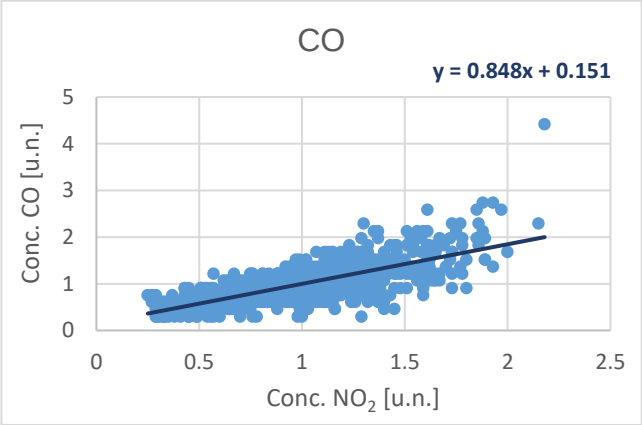


Figura 10a Correlación entre la concentración de CO y la concentración de NO_2

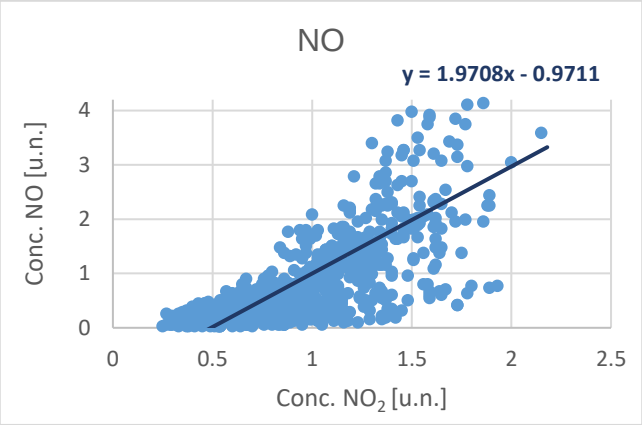


Figura 10b Correlación entre la concentración de NO y la concentración de NO_2

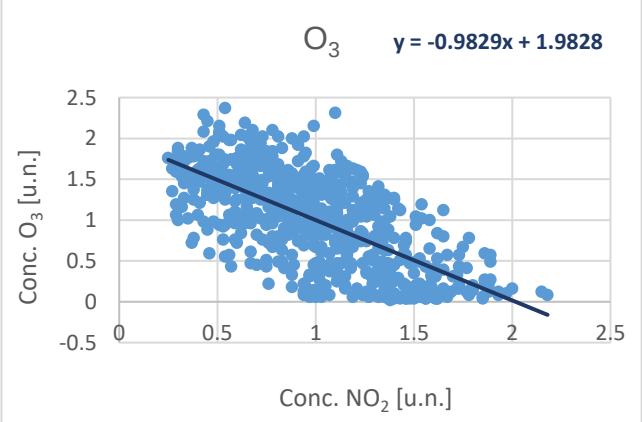


Figura 10c Correlación entre la concentración de O_3 y la concentración de NO_2

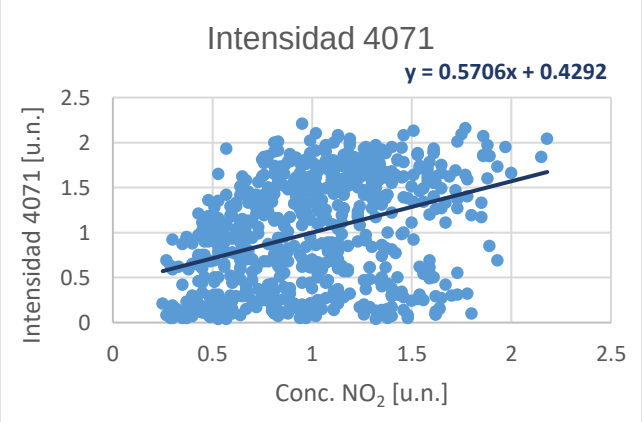


Figura 10d Correlación entre la intensidad 4071 y la concentración de NO_2

Figura 10 Gráficas de las principales correlaciones de la concentración de NO_2

Utilizando las fórmulas obtenidas ajustando una recta a la dispersión de los datos, se han elaborado predicciones del contaminante analizado en este punto. Para valorar la precisión de este método se ha calculado el error medio cuadrático medio para cada parámetro (ver Tabla 11).

Tabla 11 ECM de las variables correlacionadas con la concentración de NO_2

CO	0,14
NO	0,18
O_3	0,19
Intensidad 4071	1,00

Ozono (O_3)

La Tabla 12 presenta los resultados del análisis de correlaciones para el O_3.

Tabla 12 Principales correlaciones del O_3

O_3	
NO_2	-0,68
NO	-0,51
Velocidad Viento	0,47
Tª media	0,32

La concentración de Ozono está principalmente correlacionada con las de NO y NO_2, y con la Velocidad del viento y la Temperatura media. En la Figura 11 pueden verse estos parámetros representados frente a la concentración de O_3.

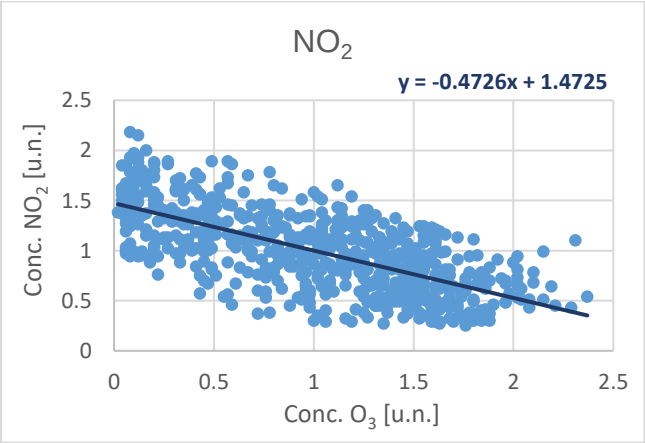


Figura 11a Correlación entre la concentración de NO_2 y la concentración de O_3

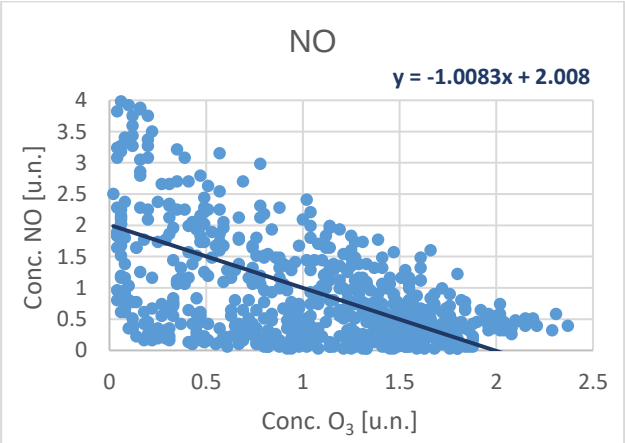


Figura 11b Correlación entre la concentración de NO y la concentración de O_3

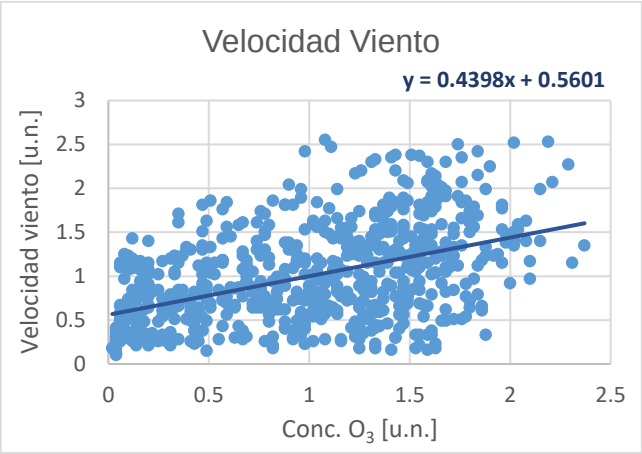


Figura 11c Correlación entre la velocidad del viento y la concentración de O_3

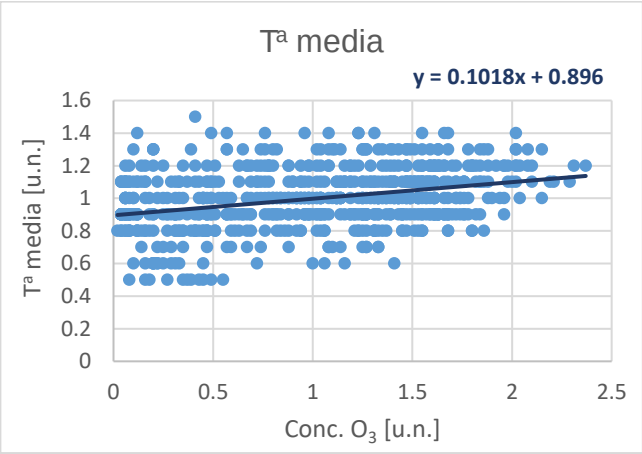


Figura 11d Correlación entre la temperatura media y la concentración de O_3

Figura 11 Gráficas de las principales correlaciones de la concentración de O_3

A continuación, se han elaborado unas predicciones de los valores de la concentración de O_3 para cada instante a través de las fórmulas de las rectas de regresión. Y a partir de estas se han calculado los errores cuadráticos medios resumidos en la Tabla 13.

Tabla 13 ECM de las variables correlacionadas con la concentración de O_3

NO_2	0,39
NO	0,98
Velocidad Viento	1,20
Tª media	2,90

Monóxido de carbono (CO)

En la Tabla 14 pueden verse los resultados del análisis de correlaciones para el CO.

Tabla 14 Principales correlaciones del CO

CO	
NO	0,74
NO_2	0,73
Intensidad 4071	0,54
Intensidad Media	0,53

La concentración de Monóxido de Carbono está principalmente correlacionada con las de NO y NO_2, y además con la intensidad del tráfico, tanto para el medidor 4071, como para la

Media. A continuación pueden observarse estos parámetros graficados frente a la concentración de CO en la Figura 12.

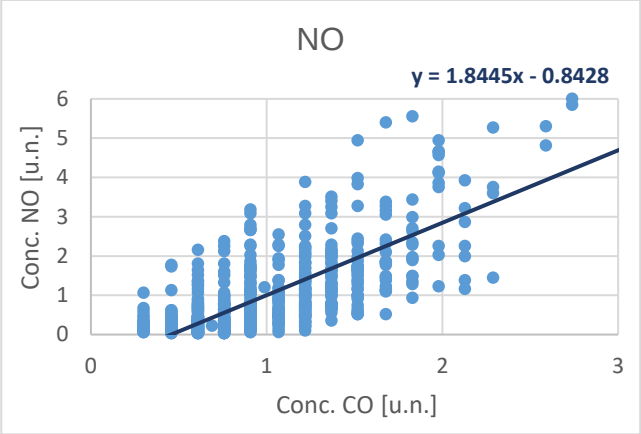


Figura 12a Correlación entre concentración de NO y la concentración de CO

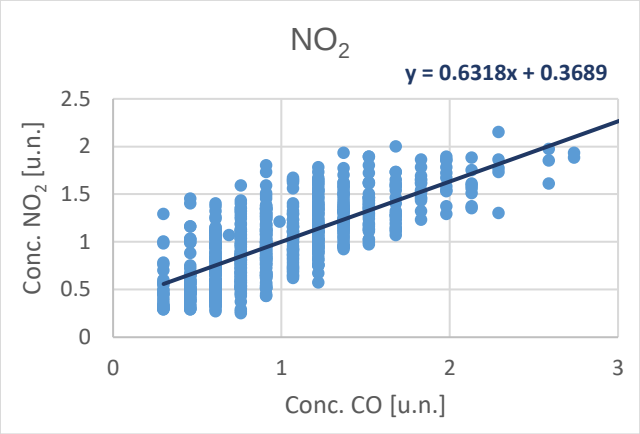


Figura 12b Correlación entre la concentración de NO_2 y la concentración de CO

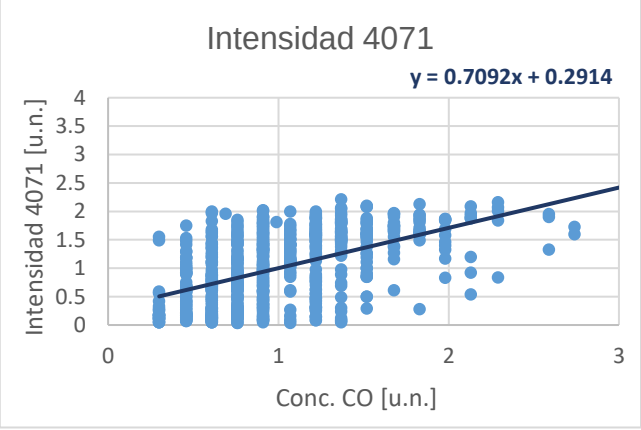


Figura 12c Correlación entre la intensidad 4071 y la concentración de CO

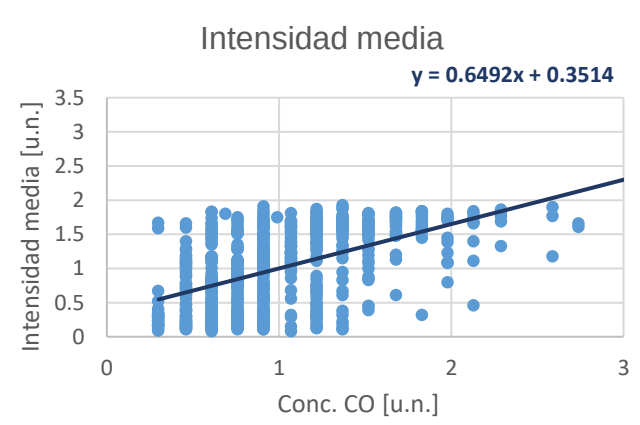


Figura 12d Correlación entre la intensidad media y la concentración de CO

Figura 12 Gráficas de las principales correlaciones de la concentración de CO

A partir de las fórmulas obtenidas ajustando una recta a la dispersión de los datos, se han elaborado predicciones del contaminante analizado en este punto. Y a partir de estas se han calculado los errores medios cuadráticos de la Tabla 15.

Tabla 15 ECM de las variables correlacionadas con la concentración de CO

NO	0,18
NO_2	0,19
Intensidad 4071	33,83
Intensidad Media	6,26

4.4. Autocorrelaciones

Además de todo lo anterior, para valorar estas primeras predicciones, hay que tener en cuenta que estas regresiones están hechas con datos del mismo registro temporal. A la hora de realizar predicciones, se calculará el valor de la concentración del contaminante en un instante t, partiendo de los datos en un t-1.

Por lo anterior, otro aspecto a tener en cuenta es cuantificar como de importante es para predecir una variable, su valor en registros temporales previos. Es decir, se debe analizar la autocorrelación de los parámetros.

Se han estudiado estas autocorrelaciones para, en primer lugar, determinar si podemos predecir alguna variable utilizando sus registros anteriores. Y en segundo lugar, para decidir para que datos resulta interesante introducir como variables del modelo predictivo, los medidas anteriores, además de las del momento desde el que se realiza la predicción.

En la Tabla 16 puede verse la autocorrelación de cada contaminante con valores 60, 120, 180 y 240 minutos anteriores. Esto mismo puede observarse en la Figura 13.

Tabla 16 Autocorrelación de los contaminantes con instantes anteriores

	SO_2	NO	NO_2	O_3	CO	PM10
-1t	0,70	0,76	0,82	0,88	0,78	0,79
-2t	0,57	0,51	0,63	0,69	0,57	0,64
-3t	0,45	0,34	0,49	0,54	0,43	0,52
-4t	0,37	0,23	0,38	0,39	0,32	0,42

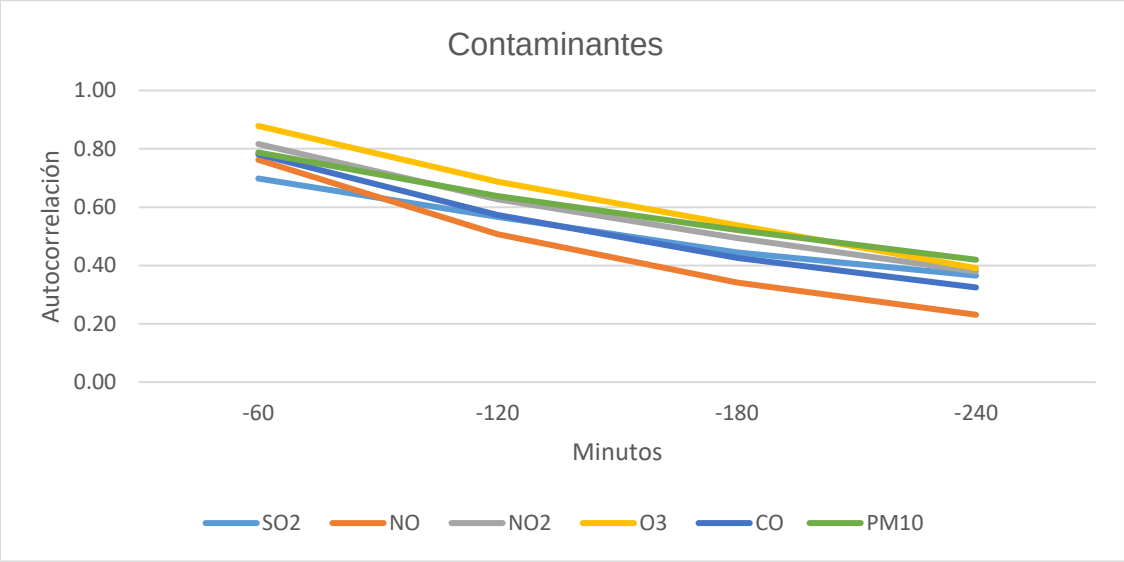


Figura 13 Gráfica de la autocorrelación de los contaminantes con instantes anteriores

En la Tabla 17 puede verse la correlación de las condiciones ambientales para 60, 120, 180 y 240 minutos anteriores. Esto mismo puede observarse en la Figura 14.

Tabla 17 Autocorrelación de las condiciones ambientales con instantes anteriores

	Tª media	HRM	Velocidad Viento	Dirección Viento	Irradiancia Solar
-1t	0,96	0,95	0,89	0,74	0,95
-2t	0,89	0,88	0,75	0,53	0,81
-3t	0,79	0,79	0,59	0,42	0,61
-4t	0,69	0,70	0,43	0,35	0,39

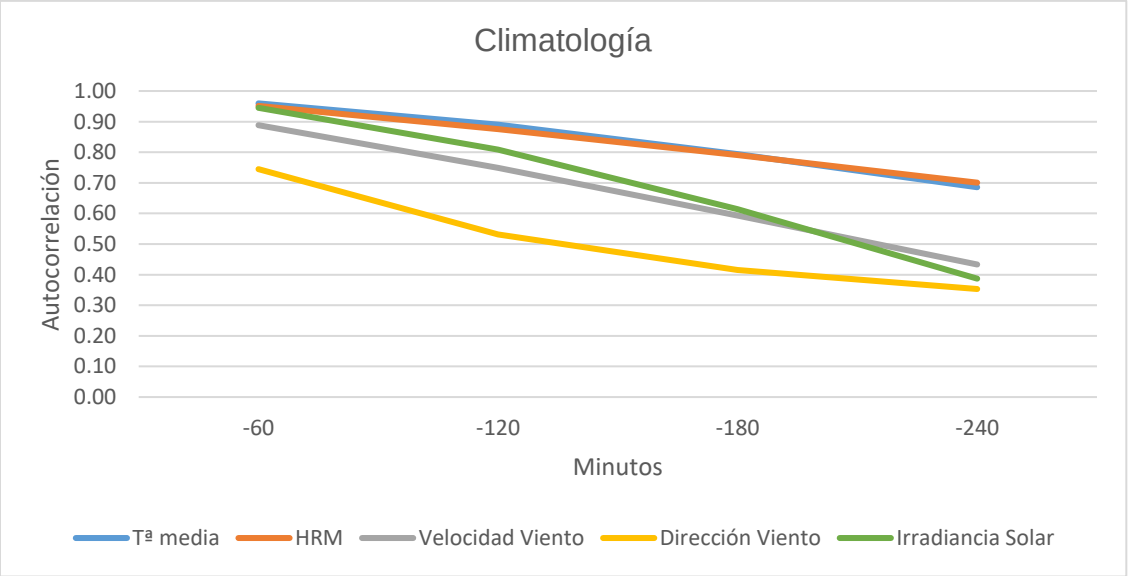


Figura 14 Gráfica de la autocorrelación de las condiciones ambientales con instantes anteriores

En la Tabla 18 puede verse la correlación de cada contaminante para un tiempo t con su medida 60, 120, 180 y 240 minutos anteriores. Esto mismo puede observarse en la Figura 15.

Tabla 18 Autocorrelación de las condiciones del tráfico con instantes anteriores

	Intensidad Media	**Intensidad 4071**	**Ocupación Media**	**Ocupación 4071**
-1t	0,92	0,88	0,73	0,59
-2t	0,76	0,71	0,57	0,37
-3t	0,58	0,52	0,43	0,18
-4t	0,38	0,35	0,29	0,14

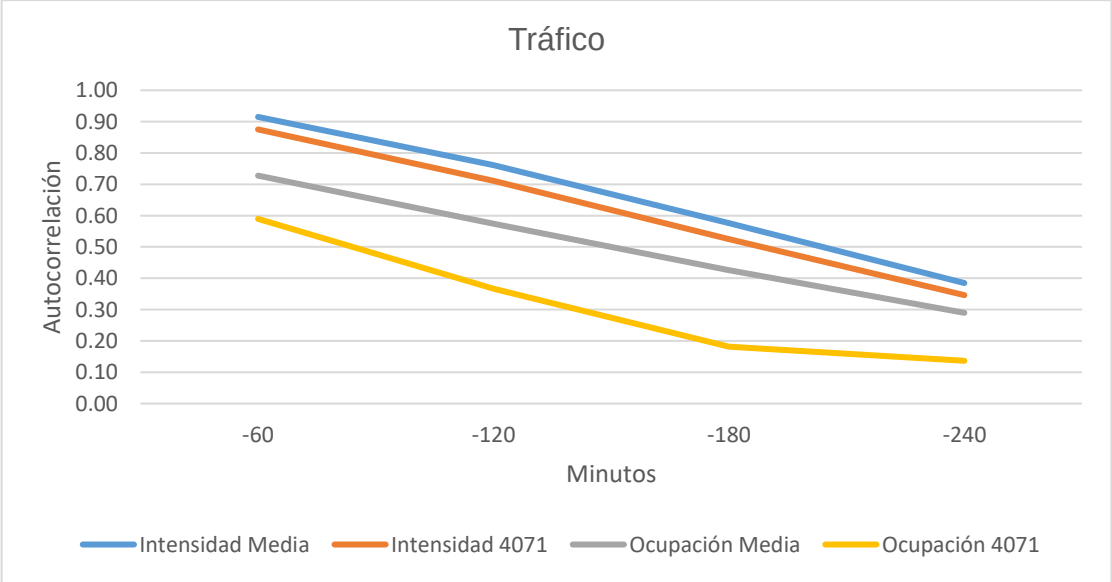


Figura 15 Gráfica de la autocorrelación de las condiciones del tráfico con instantes anteriores

En el apartado anterior, se ha expuesto la cuantificación de las desviaciones de las predicciones basada en correlaciones de las variables una a una. El error en estas predicciones se acentúa al partir de datos anteriores al instante a predecir.

4.5. Conclusión de predicciones por regresión lineal simples

Tras analizar las correlaciones entre las distintas variables y las autocorrelaciones de estas, se observa que aunque estén relacionadas entre sí, no hay una manera directa de determinar la concentración de ninguno de los contaminantes partiendo de otra variable.

Existe dependencia entre las variables ambientales y de tráfico, y los contaminantes, como era previsible. Pero ninguno de ellos puede determinarse a partir de una formula y uno o dos de los otros. Ya que los concentración de los contaminantes depende de bastantes parámetros.

A esto hay que añadirle, que las predicciones hechas con regresión lineal simple, son en realidad interpolaciones, ya que relacionan el valor del contaminante con el resto de parámetros para el mismo instante temporal. Se ha asumido que la precisión calculando con datos de registros temporales anteriores, será siempre peor que calculando con los datos para ese mismo instante de tiempo. Por lo que no se han recalculado estas precisiones partiendo de registros anteriores, ya que se llegaría a la misma conclusión del párrafo anterior sin aportar nada más.

Este resulta ser un problema multivariable, dependiente de demasiados parámetros para trabajar con métodos convencionales. Por ello se decide crear una Red Neuronal Artificial, que resulta el método más adecuado para problemas como este. Ya que es un problema de predicción donde se desconoce la relación física exacta que relaciones los parámetros, pero para el que se cuenta con bastantes datos como para extraerla por métodos empíricos estadísticos y donde los parámetros a calcular dependen de un gran número de variables.

5. Estado del arte

En los últimos años ha crecido el interés por desarrollar modelos predictivos de la contaminación. Esto es debido al aumento de los niveles de concentración de contaminantes en el aire de las ciudades, a la mayor disponibilidad de datos, y a los avances en técnicas para desarrollar modelos predictivos.

A continuación, pueden encontrarse algunos estudios donde se valoran las diferentes alternativas a la hora de desarrollar estos modelos a partir de los conjuntos de datos disponibles, y sobre su aplicación. Concluyendo todos ellos la conveniencia de aplicar Redes Neuronales Artificiales para realizar este tipo de predicciones.

Modelado de un sistema de supervisión de la calidad del aire usando Técnicas de Fusión de Sensores y Redes Neuronales (2010) [8].

Sistema desarrollado para la ciudad de Madrid, en el marco de un proyecto de investigación. Este modela el comportamiento de los contaminantes presente en el aire, tomando datos de un conjunto de sensores y utilizando Redes Neuronales Artificiales.

Air pollutants concentrations forecasting using back propagation neural network based on wavelet decomposition with meteorological conditions (2016) [9].

En este artículo, se desarrolla una RNA utilizando un optimizador con *Back-propagation*. Este modelo fue desarrollado con el objetivo de predecir los niveles de concentración de PM10, SO_2, and NO_2, en la ciudad de Chongqing, China.

A comparative study of computational intelligence techniques applied to PM2.5 air pollution forecasting (2016) [10].

Este artículo presenta los resultados de la comparativa del desempeño de dos modelos elaborados con técnicas de inteligencia computacional: con una Red Neuronal Artificial, y con un sistema neuro-fuzzy adaptativo. Los dos modelos han sido desarrollados para predecir la contaminación del aire debida a las partículas en suspensión. La valoración de los resultados, se hizo con el error medio cuadrático y con el error medio absoluto.

Air Quality Forecasting in Madrid Using Long Short-Term Memory Networks (2017) [11].

En este artículo, se propone un sistema de predicción de contaminación del aire basado en una RNA. Utiliza datos de calidad del aire y meteorológicos para entrenar la red y realizar predicciones. Para medir la precisión del modelo se ha utilizado el error medio cuadrático. Los

resultados de la RNA han sido comparados con los proporcionados por un modelo basado en sistema CALIPE

Todos estos estudios tienen en común que han sido desarrollados a partir de datos de una localización concreta, y modelan la interacción de las variables para ese lugar concreto. Esto quiere decir, que aunque aporten muy buenos resultados, no son aplicables a otros lugares, ya que las particularidades de cada lugar provocan que las variables que influyen en los niveles de contaminación, lo hagan de un modo distinto. Esto justifica la necesidad de elaborar este estudio para obtener un modelo predictivo de la contaminación para la ciudad de Barcelona.

6. Herramientas utilizadas para desarrollar los modelos predictivos

Existen multitud de herramientas y lenguajes de programación con los que poder crear y entrenar Redes Neuronales Artificiales. Para el desarrollo de este trabajo, se ha decidido utilizar como lenguaje de programación Python. Se ha elegido este programa, por ser hoy día el más utilizado para RNA, contando con gran cantidad de librerías y documentación para la elaboración de estas.

Dentro de Python, se ha utilizado el módulo Tensorflow, la librería Open-Source desarrollada por Google orientada a la construcción de modelos a través de RNA [12]. Este trabaja internamente en C++, por lo que el programa unirá las ventajas de Python para este tipo de modelos, con la potencia y velocidad de C++.

Además, se ha trabajado con la herramienta de visualización de modelos Tensorboard, que ha permitido analizar la RNA creada, y el comportamiento de todas las variables del sistema durante el entrenamiento. Esto ha sido enormemente útil, para poder desarrollar con éxito este modelo predictivo. En la Figura 16 se muestra un esquema con todos los programas empleados en el proyecto.

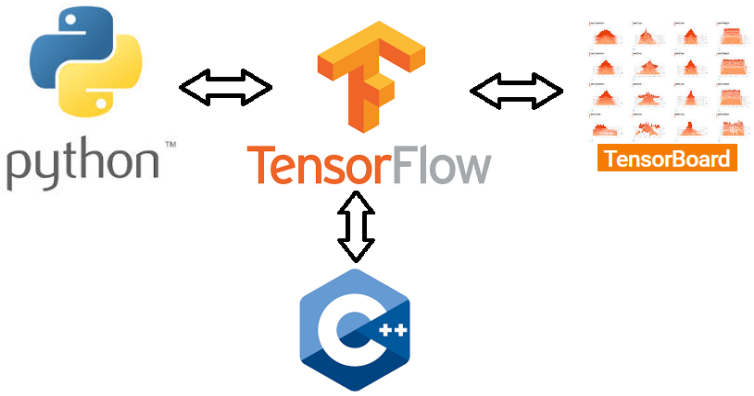


Figura 16 Diagrama de programas utilizados

Con estas herramientas se han desarrollado los programas de las Redes Neuronales Artificiales para crear, entrenar y producir predicciones. En el Anexo B puede encontrarse los códigos de los programas creados para cada caso. En esta memoria podrán verse las estructuras de las redes desarrolladas, así como los resultados de las predicciones de estas

7. Preparación de los datos

Al igual que en análisis inicial, a la hora de utilizar los datos para entrenar los modelos, estos han tenido que ser normalizados, como se detalló anteriormente.

Para verificar que las predicciones del modelo sean correctas, se ha dividido el conjunto de datos en un grupo para entrenar la red, y en otro para validación. De este modo puede comprobarse que la RNA realmente está aprendiendo de los datos y extrayendo información generalizable a otros casos y no simplemente memorizando los datos y creando una curva que se ajuste a estos. Esta capacidad de generalización de las RNA, es una de las principales ventajas de estas frente a modelos convencionales, por lo que resulta muy importante verificarla.

El número de datos de entrada para cada caso, se ha seleccionado cogiendo las variables que el análisis inicial indicaba que estaban más correlacionadas con los parámetros a predecir. Se ha llegado a una situación de compromiso, ya que es conveniente introducir la mayor cantidad de información posible, pero exceder en el número de variables de entrada provoca un aumento del ruido y de la dificultad de convergencia de la red, por lo que no es conveniente introducir parámetros que no sean sustancialmente relevantes.

Además de los parámetros para el instante de tiempo justamente anterior al que se desea predecir, se han introducido en los modelo valores para registros anteriores, en los casos en los que aportaba valor a la predicción. Y para los datos meteorológicos de Temperatura, Humedad, Viento e Irradiancia, también se han añadido el valor de estos para el instante de tiempo donde se realiza la predicción. Ya que existen modelo meteorológicos capaces de predecir estos parámetros para este margen de tiempo con gran precisión, y de esta manera se reduce cierta carga predictiva innecesaria de este modelo.

8. Red Neuronal Artificial multicapa para regresión logística

En primer lugar se ha desarrollado una RNA multicapa para regresión logística. Esta clasificará los casos en dos clases, en función de si predice que la concentración del contaminante medido para el instante de tiempo a calcular, será superior o no al valor límite seleccionado.

8.1. Configuración del modelo

Para desarrollar el clasificador, se ha creado un modelo conformado por una red con tres capas, Figura 17, una de entrada(x), una oculta interna(h), y una de salida(o).

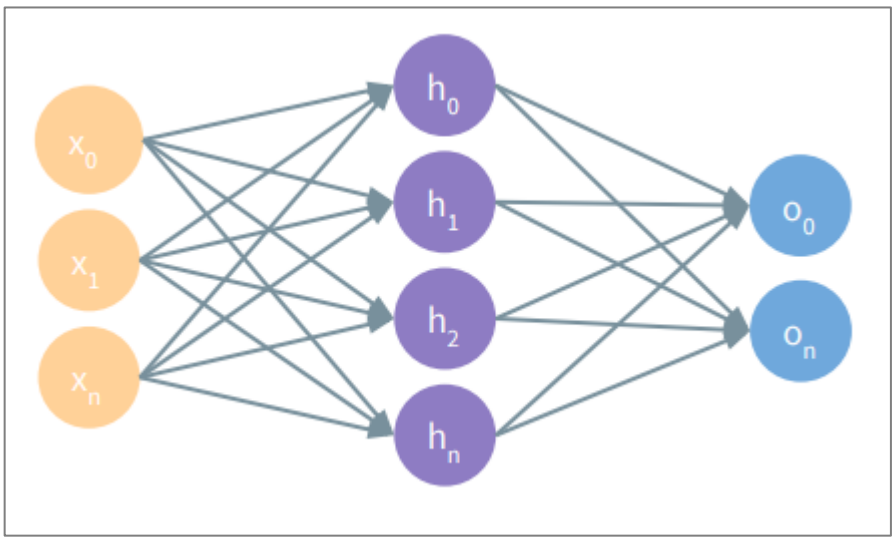


Figura 17 Modelo de RNA multicapa para regresión logística

En cada una de estas capas hay una serie de perceptrones, que son las neuronas que forman la red. Todos estos están conectados con el resto de la red a través de sus respectivos pesos (w), ajustados con sus sesgos (b) y luego pasan por su función de activación. En la Figura 18, puede verse la estructura de un perceptrón/neurona.

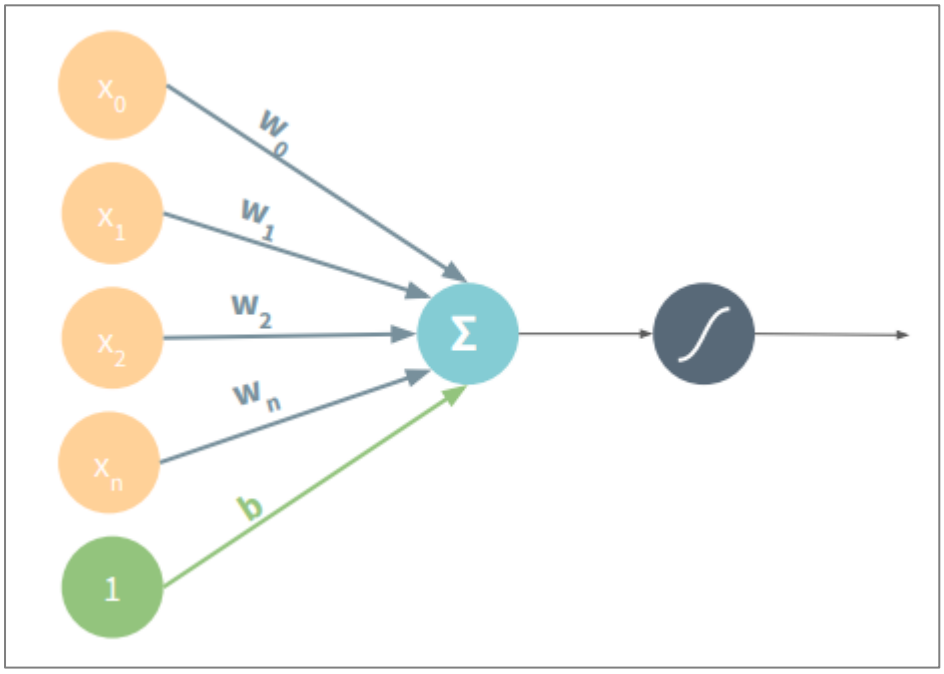


Figura 18 Esquema de perceptrón para la red de clasificación

Todos estos nodos y conexiones de la red conforman el modelo, que tras el entrenamiento permitirá realizar las predicciones. Para llevar a cabo este entrenamiento el programa varía aleatoriamente los distintos parámetros siguiendo las indicaciones impuestas en su configuración. Las predicciones obtenidas son comparadas buscando minimizar una función de coste. Dicha función es una medida del error entre las predicciones obtenidas con el modelo y los datos empíricos disponibles, calculada a partir del EMC.

El programa creado para generar y entrenar la red, además de los componentes estructurales de la red, cuenta con diferentes módulos dedicados a la inicialización de los distintos tensores del sistema, bucle de optimización y actualización de las variables internas, y todos los demás componentes para preparar, visualizar y guardar los datos, tanto de entrada como de resultados del modelo.

Todos estos componentes de la red, quedan conectados entre ellos como puede verse en el esquema de la Figura 19. Este ha sido generado con Tensorboard, y en él puede observarse como queda finalmente la red.

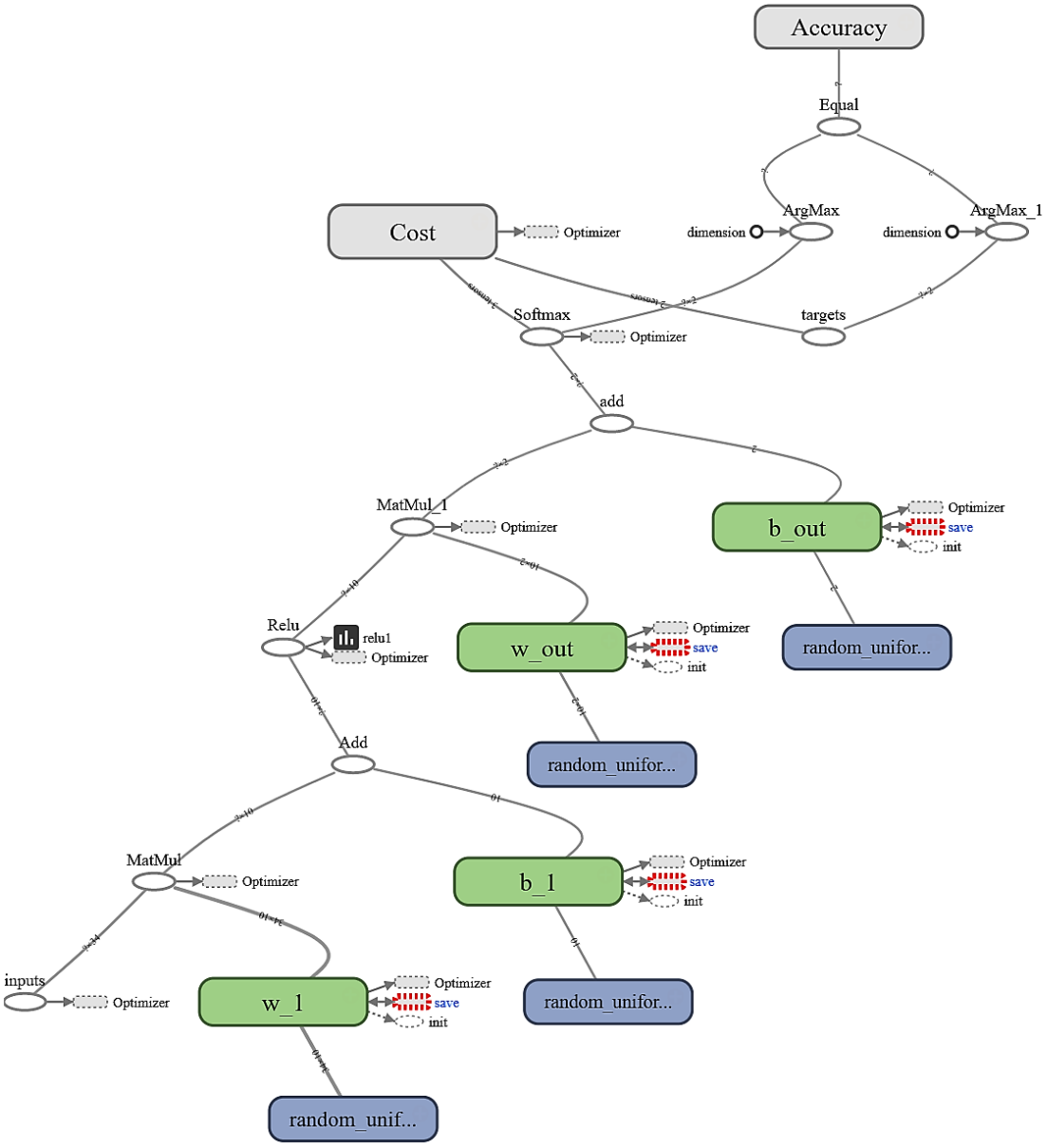


Figura 19 Esquema básico de red neuronal multicapa para regresión logística

La herramienta Tensorboard, permite analizar la red tanto de un punto de vista genérico, como entrando más en detalle. En la Figura 19 aparecía la estructura básica de la red, mientras que en la Figura 20 puede verse esta estructura principal y los elementos auxiliares de entrenamiento. Por último, en la Figura 21 puede observarse un nivel más detallado de la red. Esta herramienta permite entrar en niveles de todavía mayor detalle, lo cual resulta muy útil en la etapa de diseño del modelo.

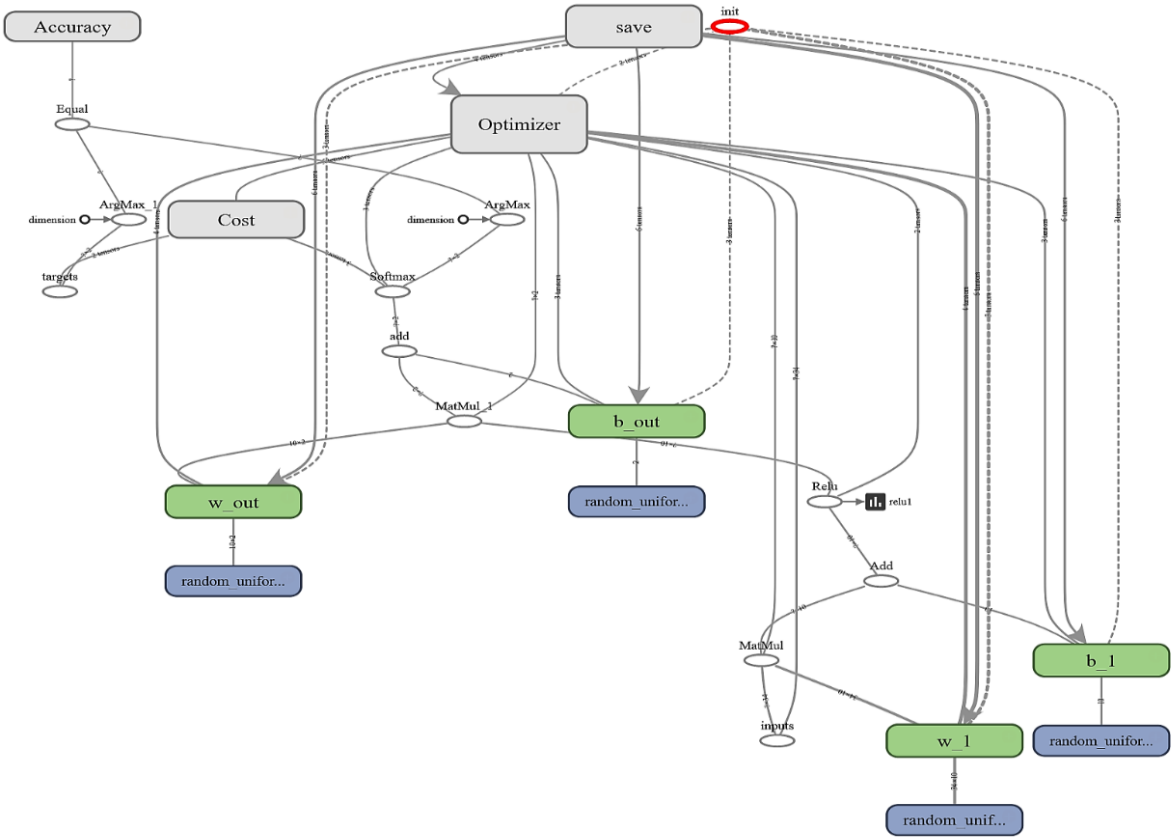


Figura 20 Esquema completo de RNA multicapa para regresión logística

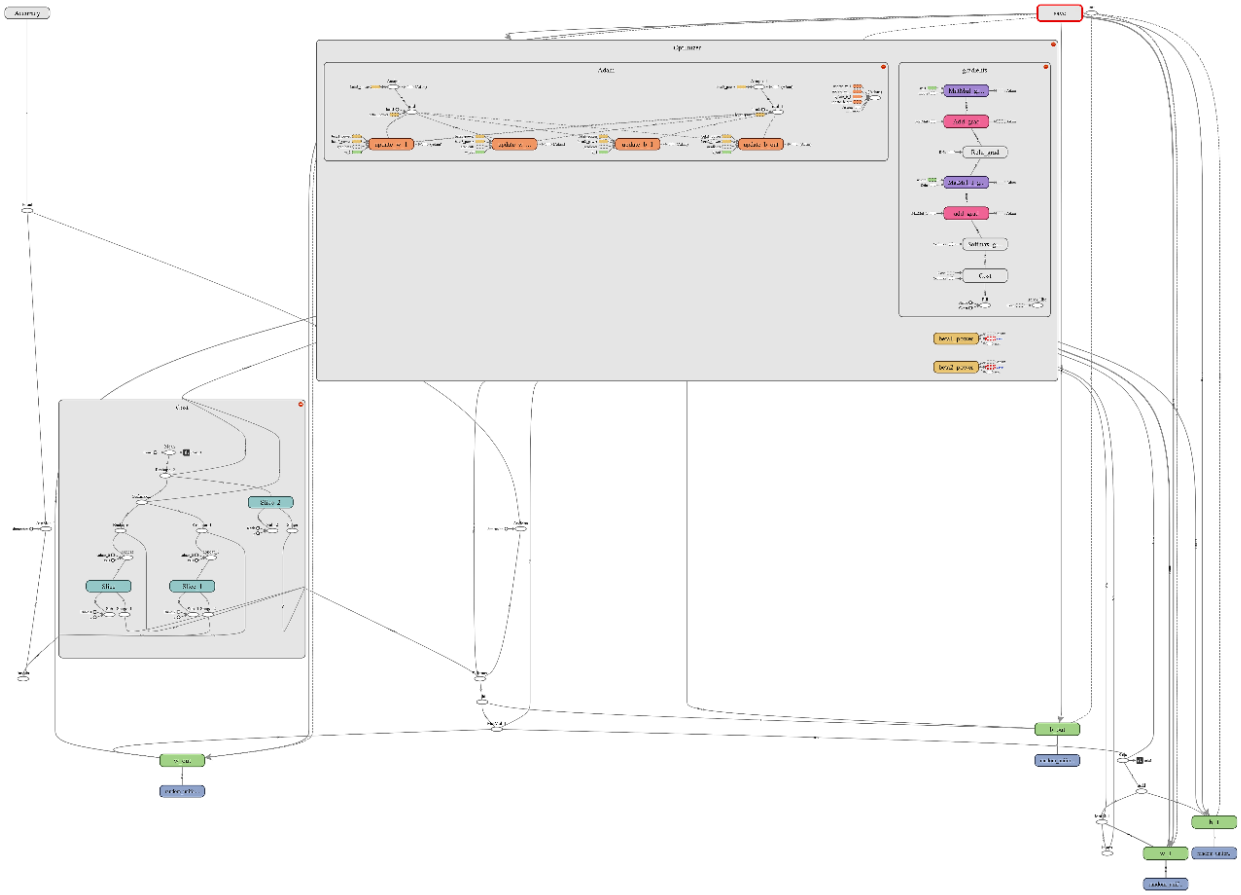


Figura 21 Detalles de esquema de RNA multicapa para regresión logística

Además del esta estructura, se ha probado con otras redes de mayor y menor profundidad. Se ha determinado, que sin capa oculta intermedia el modelo daba predicciones muy pobres. Por otro lado, al aumentar las capas de este no se conseguía mejorar los resultados y aumentaba notablemente el tiempo de entreno del sistema y aparición de divergencias fatales en alguna de sus componentes. En la Figura 22 pueden verse los esquemas utilizados para las pruebas con 0 y 2 capas intermedias.

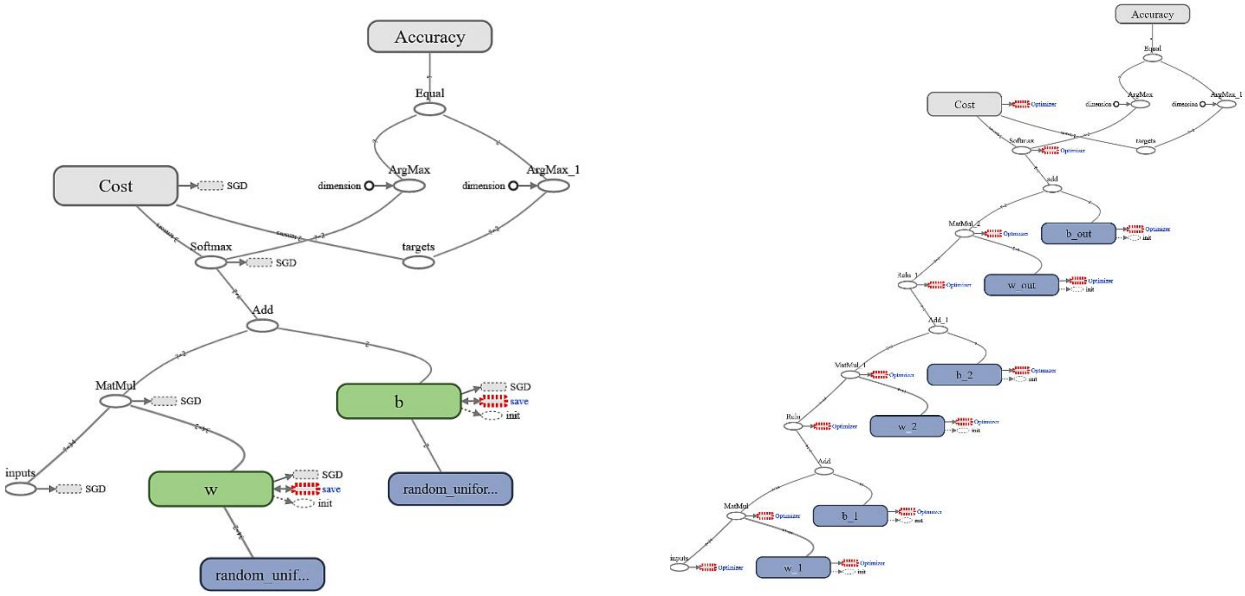


Figura 22 Esquemas de RNA para regresión logística con 0 y 2 capas

En el Anexo B, puede encontrarse el código utilizado para crear y entrenar esta red

8.2. Resultados

Resultados para las PM10

Una vez creada la RNA y el programa para entrenar esta, se ha ido haciendo una serie de pruebas para determinar la configuración de los hiperparámetros óptima. En la Tabla 19 se muestran algunas de las pruebas llevadas a cabo:

Tabla 19 Tabla de pruebas

Ratio de aprendizaje	Variables entrada	Nodos en capa oculta	Precisión entrenamiento	Precisión validación
0,0000001	24	10	**80%**	**76%**
0,0000001	24	20	**84%**	**77%**
0,00000001	24	20	**80%**	**76%**
0,00001	24	20	**92%**	**76%**
0,00001	24	15	**91%**	**77%**
0,000001	24	20	**91%**	**77%**
0,000001	24	10	**92%**	**80%**
0,000001	24	5	**90%**	**77%**

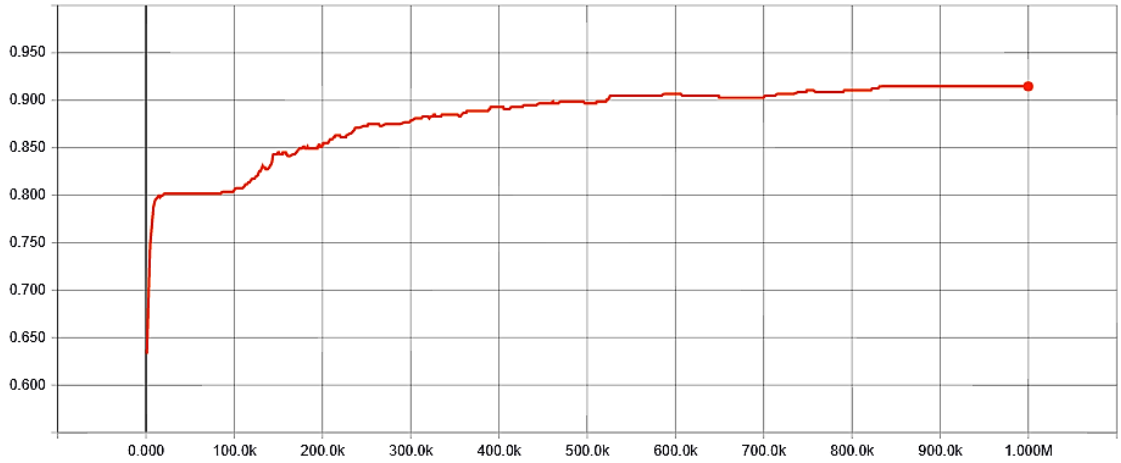
Tras realizar las pruebas, la configuración óptima para este caso ha resultado ser, trabajar con 24 inputs, 10 nodos en la capa oculta intermedia, una ratio de aprendizaje de 0,000001 y entrenar la red 1.000.000 de iteraciones.

La elección del número de neuronas de la capa oculta y del número de iteraciones, se ha tomado buscando que el modelo consiguiese la mejor precisión posible sin producir un sobreajuste a los datos.

El sobreajuste a los datos, más conocido como *overfitting*, ocurre cuando el modelo deja de extraer generalidades/patrones de los datos, y empieza a memorizar la estructura de los datos [13]. Para controlar la aparición de este, se prueba el modelo con el set de datos de validación, que es el conjunto de datos que el modelo no ha visto en ningún momento del entrenamiento. Cuando la precisión respecto a estos mejora, quiere decir que el modelo está generalizando adecuadamente, mientras que si los resultados empiezan a empeorar es un indicador de que el modelo está sufriendo *overfitting*

El ratio de aprendizaje es un indicador de cuánto va variando el optimizador de la red, el valor de las variables cuando trabaja en minimizar el error. Si este resulta demasiado grande el modelo no conseguirá acceder a zonas de mínimos y saltará continuamente en torno a estos. Por el contrario, si es demasiado pequeño, el entrenamiento sería demasiado lento y podría quedarse atascado en un mínimo local.

En la Figura 23 puede verse la evolución de la precisión del modelo conforme la red va entrenándose en una de las pruebas. La precisión, es una métrica que utiliza el programa para cuantificar la veracidad de la clasificación. Este computa la media de todos los elementos de una matriz que contiene un registro de todos los casos clasificados correcta o incorrectamente.



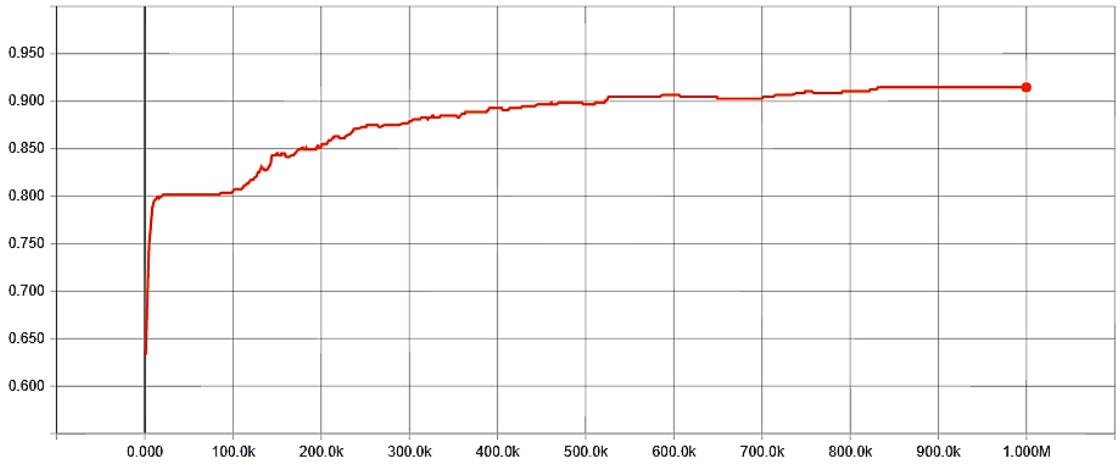


Figura 23 Evolución de la precisión del clasificador para PM10

Y en la Figura 24, puede verse la evolución de la función de coste, que es el parámetro utilizado para ver la proximidad a la solución óptima, e intenta minimizar conforme avanza el entrenamiento. El coste está calculado con el EMC.

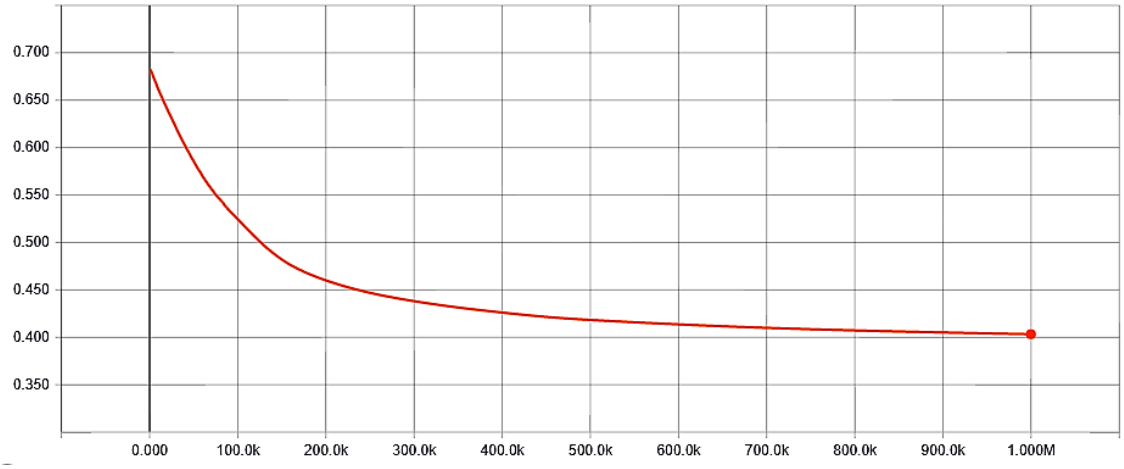


Figura 24 Evolución del coste del clasificador para PM10

Una vez entrenada la red, se ha obtenido la matriz de confusión de las predicciones hechas por el modelo, tanto para los datos de entrenamiento, como para el conjunto de datos de validación. En la Tabla 20 y Tabla 21 pueden verse estos resultados.

Tabla 20 Matriz con datos de entrenamiento

	VERDADERO	FALSO
POSITIVO	70	13
NEGATIVO	391	30

Tabla 21 Matriz con datos de validación

	VERDADERO	FALSO
POSITIVO	12	18
NEGATIVO	126	12

El resultado está lejos de lo deseado para los Positivos, ya que hay un ratio 12/18 entre Verdaderos Positivos y Falsos Positivos.

Esto se debe en parte a que al haber muchos más resultados negativos que positivos, la red ha aprendido mejor a identificar los negativos. Para solventar esto se han repetido los cálculos forzando a que el modelo se fije más en los casos positivos. Esto se ha conseguido ponderando el error de estos, de manera que pesasen más y por tanto que la red tendiese a una configuración que los clasificase mejor.

Con este cambio se repite el entrenamiento y las pruebas de la red obteniendo una mejora en los resultados para los caso positivos, a costa de reducir la precisión en los negativos, como puede observarse en la Tabla 22.

Tabla 22 Matriz con datos de validación (con ponderación de positivos)

	VERDADERO	FALSO
POSITIVO	16	23
NEGATIVO	121	8

En la Tabla 23, pueden verse los resultados del modelo comparados con los resultados del análisis de correlaciones, tomando los parámetros más correlacionados con las PM10. (Para poder compararlos, se ha utilizado el set completo de datos)

Tabla 23 Comparativa de matrices de confusión

	VERDADERO					FALSO				
	RNA	CO	Tª media	Intensidad Media	Intensidad 4071	RNA	CO	Tª media	Intensidad Media	Intensidad 4071
POSITIVO	**86**	79	91	100	97	**36**	151	164	176	180
NEGATIVO	**512**	397	384	372	368	**38**	45	33	24	27

Para poder valorar mejor estos resultados, se ha calculado los ratios de verdaderos positivos y negativos, conocidos como TPR y TNR respectivamente, por sus siglas en inglés. Estos ratios pueden verse en la Tabla 24.

$$TPR = \frac{TP}{TP + FN}$$

$$TNR = \frac{TN}{TN + FP}$$

Tabla 24 Comparativa de TPR y TNR

	RNA	CO	Tª media	Intensidad Media	Intensidad 4071
TPR	**69%**	64%	73%	81%	78%
TNR	**93%**	72%	70%	68%	67%
PROMEDIO	**81%**	68%	72%	74%	73%

Aun habiendo ponderado el peso de los casos positivos, se observa que el TRP sigue siendo bajo. Se podría sobreponderar más estos casos, pero la RNA tiende a sufrir overfitting. La solución sería entrenar el modelo con un mayor volumen de datos.

En general, el promedio de los ratios indica una mejoría de la RNA respecto a los otros casos. Algunos de los otros modelos de referencia tienen mejor TPR sin tener un TNR mucho peor, pero esto es debido a que por azar han clasificado muchos más puntos como positivos, mejorando el TPR. El TNR no se ve muy empeorado en estos casos, por diluirse los falsos positivos en el volumen mucho mayor de casos negativos.

9. Red Neuronal Artificial multicapa para regresión lineal multivariable

Además de por la falta de datos en torno al límite, el modelo de clasificación no es el mejor método para hacer un seguimiento de los contaminantes, ya que ese modelo solo aprende sobre si los valores están por encima o debajo de cierto punto, pero no cuanto lo sobrepasan, ni el valor concreto que tendrán los contaminantes.

Por todo esto, se ha decidido desarrollar una RNA multicapa para regresión lineal multivariable. Con este modelo se podrá predecir el valor para cada contaminante independientemente del nivel de concentración en que se encuentre. Y al no tener la limitación de entrenar el modelo en torno a un límite fijado, se podrá desarrollar para todos los contaminantes.

9.1. Configuración del modelo

Este modelo, al igual que en el caso anterior, cuenta con 3 capas, una de entrada(x), una oculta interna (h), y una de salida(o). Pero a diferencia del clasificador, aquí solo hay una salida, correspondiente al contaminante que se desea predecir. En la Figura 25 puede verse el esquema de la red

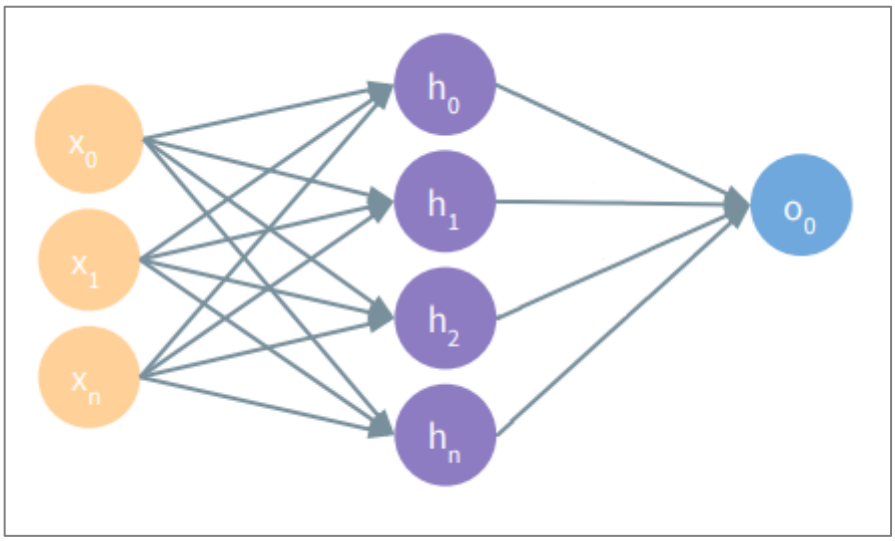


Figura 25 Modelo de RNA multicapa para regresión lineal

Y en este caso también hay una serie de perceptrones, que conforman el modelo. La diferencia con el modelo anterior están en la función de activación. En el caso anterior utilizamos una función sigmoide a la salida del modelo para obtener una señal binaria,

mientras que en este, como queremos el valor del parámetro utilizamos una función de activación lineal. En la Figura 26 aparece la estructura del perceptrón.

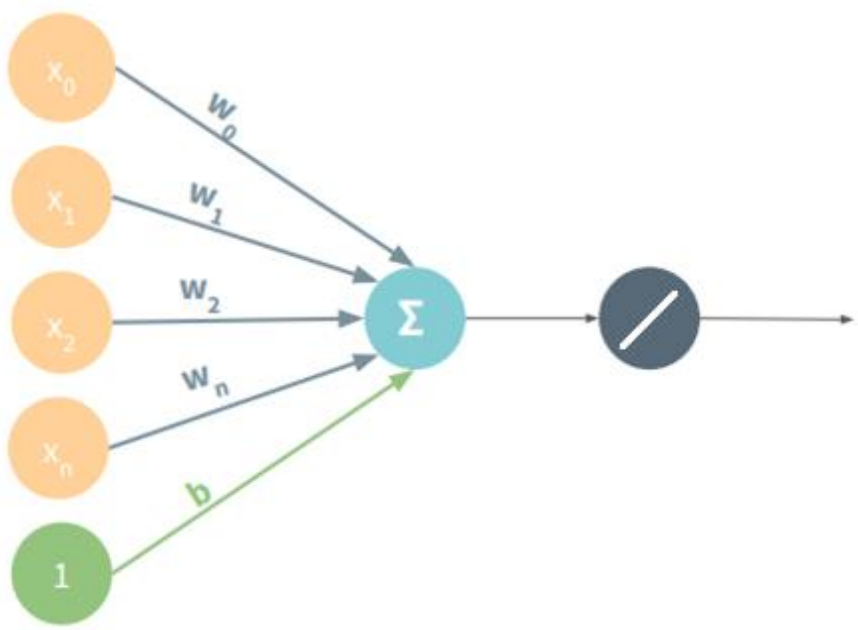


Figura 26 Esquema de perceptrón para red de regresión lineal

A continuación, puede verse como quedaría la red en este caso utilizando Tensorboard (ver Figura 27). Como puede observarse, la diferencia con el modelo logístico está en la función de activación a la salida, por el motivo comentado anteriormente. También se diferencia en el pre y postprocesado de los datos, la función de optimización, el número de neuronas de la capa oculta y en el valor de los hiperparámetros.

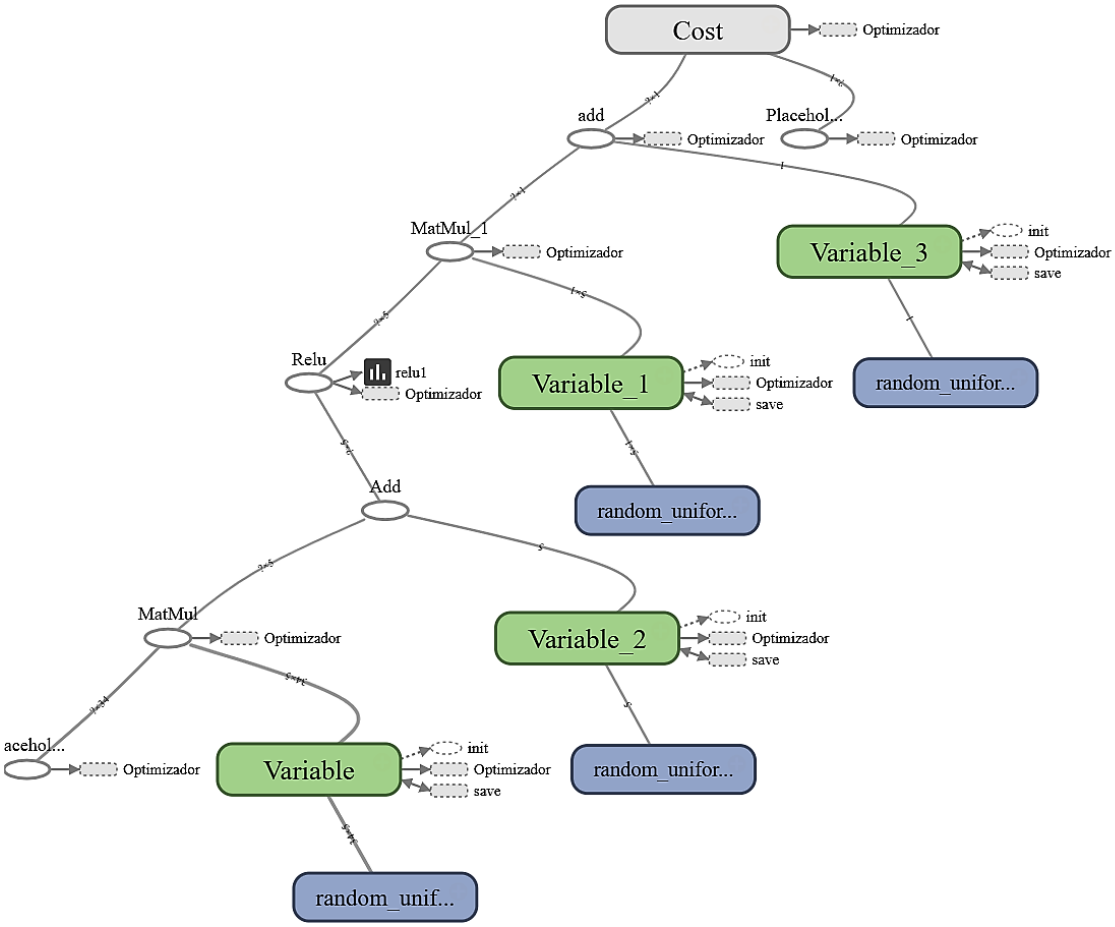


Figura 27 Esquema básico de RNA multicapa para regresión lineal

En el Anexo B, puede verse el código utilizado para crear y entrenar esta red.

9.2. Resultados

Resultados para PM10

En este caso, el modelo predice los valores que tomará la concentración de PM10, para un determinado instante de tiempo, partiendo del conjunto de variables expuestas anteriormente.

El modelo ha aprendido a partir del conjunto de datos de entrenamiento, y a continuación se han comprobado los resultados haciendo predicciones para el conjunto de datos de validación. Esto implica que si se observa que el modelo predice adecuadamente los valores de PM10, no es porque sea una curva ajustada a unos datos, ya que estos no se han utilizado para el desarrollo de esta, sino que el modelo ha conseguido aprender la relación entre los parámetros y con ello ha realizado la predicción.

En la Figura 28, puede verse la evolución del coste a lo largo del entrenamiento. Como puede observarse, en un primer momento se reduce el error rápidamente, pero a partir de un punto la convergencia es mucho más lenta. Esto es debido a que el set de datos es bastante limitado y tras aprender las interrelaciones y patrones más sencillos le cuesta avanzar, ya que tiene pocos ejemplos.

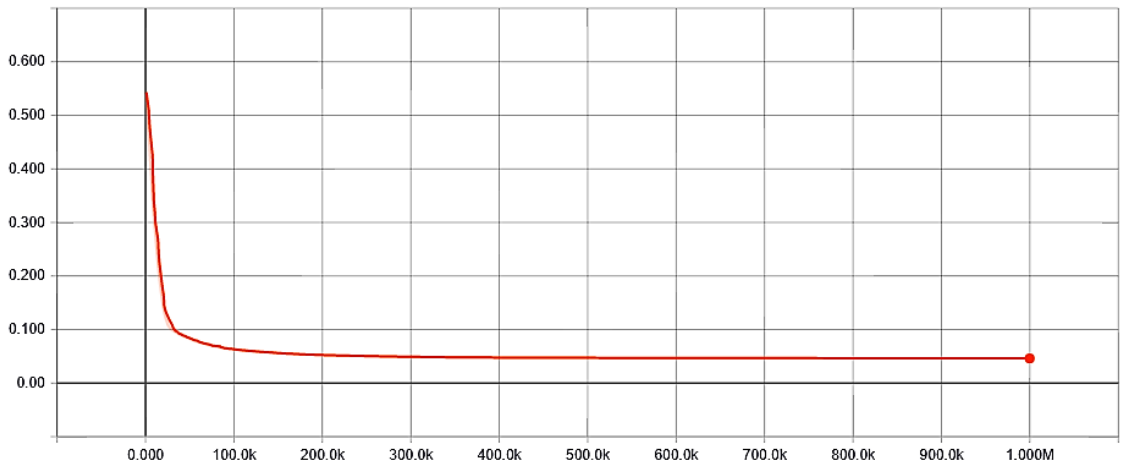


Figura 28 Evolución del coste del modelo regresivo para PM10

Se ha elaborado una serie de pruebas variando los hiperparámetros de la red, hasta encontrar la configuración óptima. A continuación se utilizado el modelo para realizar predicciones.

En la Figura 29 y Figura 30, pueden verse comparados los valores predichos por la RNA, con los valores reales, para el conjunto de datos de entrenamiento y validación respectivamente.

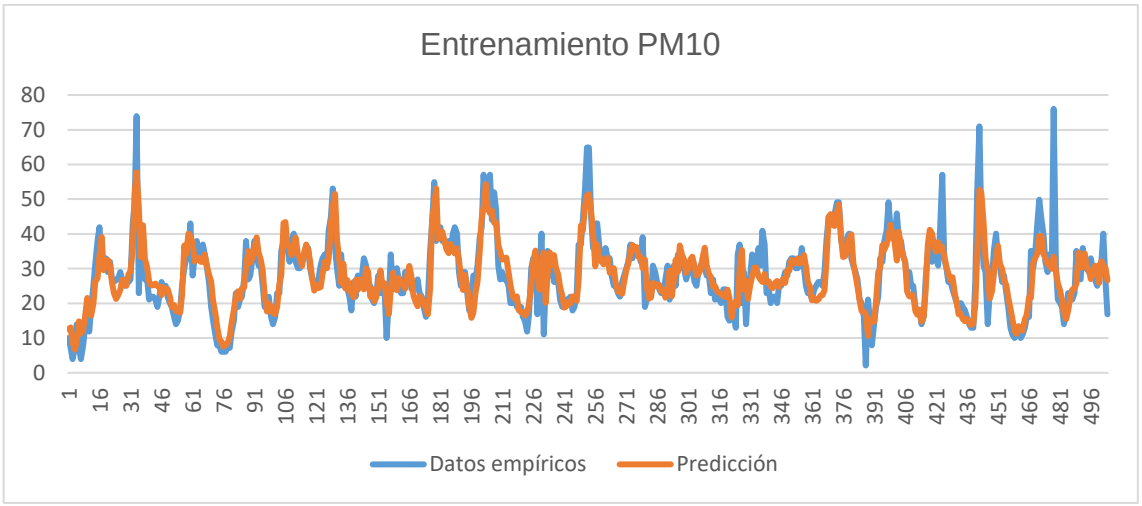


Figura 29 Predicciones para PM10 a partir de datos de entrenamiento comparadas con datos empíricos

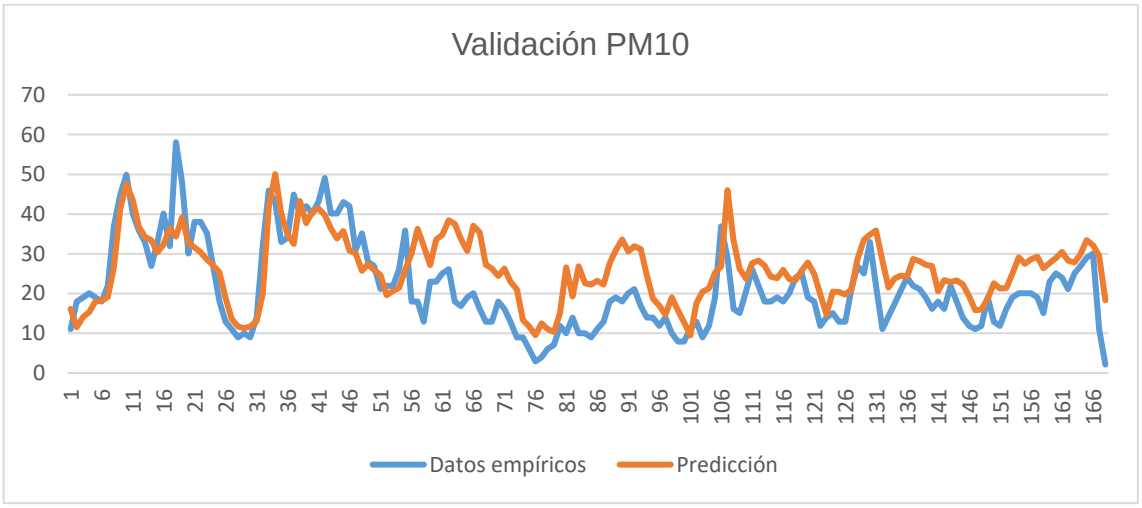


Figura 30 Predicciones para PM10 a partir de datos de validación vs datos comparadas con datos empíricos

Y en la Figura 31 puede verse la predicción hecha a partir de la fórmula obtenida por regresión lineal simple para los datos del conjunto de validación. Se ha utilizado la variable más correlacionada, en este caso la concentración de CO.

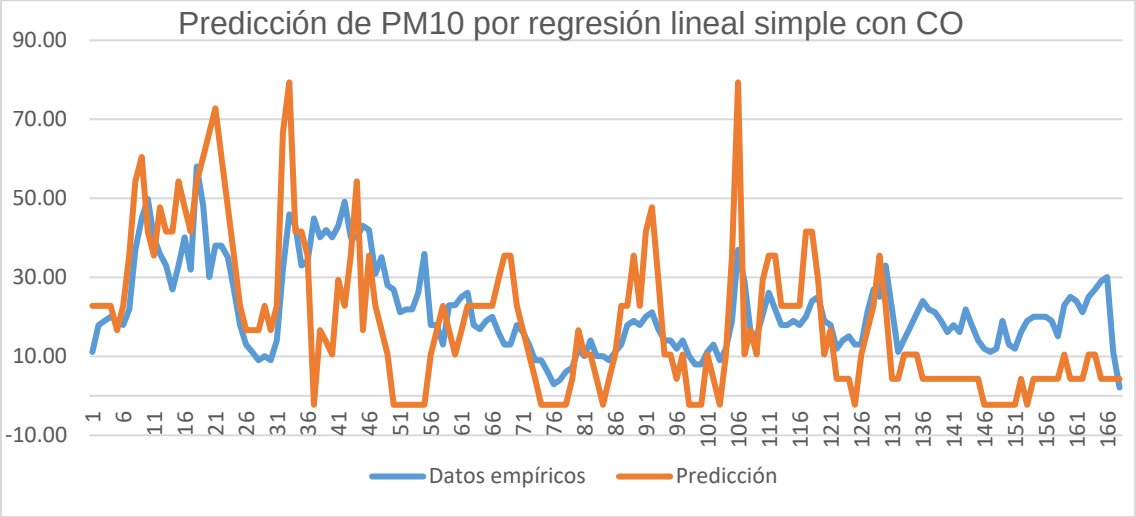


Figura 31 Predicciones para PM10 con regresión lineal simple partiendo del CO comparadas con datos empíricos

Para cuantificar la mejora de la RNA frente a las predicciones con regresiones lineales simples, se ha comparado el error medio cuadrático. En la Tabla 25, puede verse la comparativa de estos.

Tabla 25 Comparativa de resultados de EMC para las PM10

RNA	**0,062**
Tª media	0,40
Intensidad Media	0,53
Intensidad 4071	0,56
CO	0,35

Como puede observarse, el modelo desarrollado aporta una mejora de casi un orden de magnitud del EMC y en los gráficos puede observarse como el comportamiento descrito se aproxima mucho más a la realidad.

Resultados para SO_2

A continuación se muestran los resultados del modelo desarrollado para predecir la concentración del SO_2.

Se ha trabajado del mismo modo que en el caso anterior. En la Figura 32 y Figura 33, pueden verse comparados los valores predichos por la RNA, con los valores reales, para el conjunto de datos de entrenamiento y validación respectivamente.

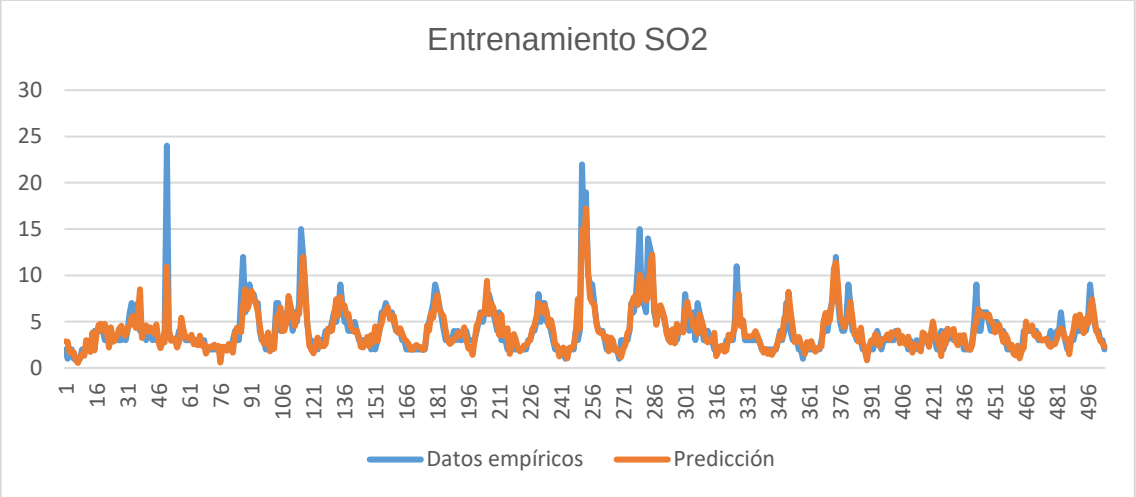


Figura 32 Predicciones para SO_2 a partir de datos de entrenamiento comparadas con datos empíricos

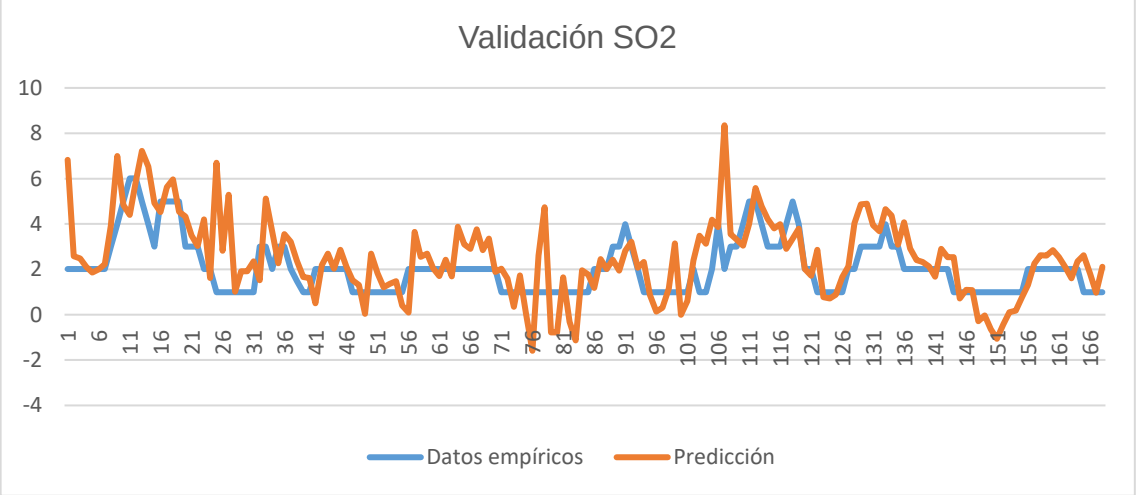


Figura 33 Predicciones para SO_2 a partir de datos de validación comparadas con datos empíricos

Y en la Figura 34, podemos ver las predicciones para el mismo conjunto de datos de validación, utilizando regresión lineal simple con el parámetro más correlacionado, la temperatura media

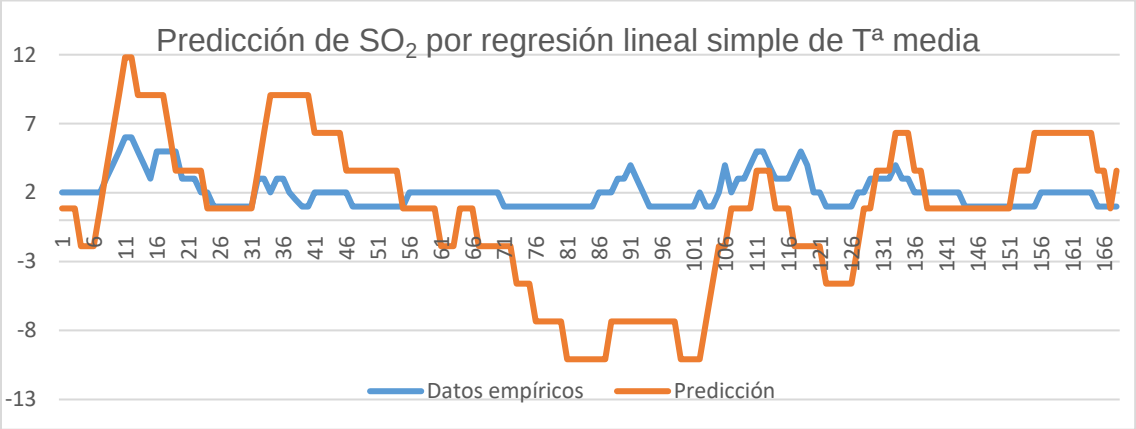


Figura 34 Predicciones para SO_2 con regresión lineal simple partiendo de la Temperatura media comparadas con datos empíricos

En la Tabla 26, puede verse la comparativa del error medio cuadrático de esta red, con los de las regresiones monovariables simple para el SO_2.

Tabla 26 Comparativa de resultados de EMC para SO_2

RNA	**0,149**
Tª media	1,53
PM10	1,70
Irradiancia Solar	2,95
CO	3,44

En este caso, las predicciones para el set de validación son peores que en el modelo de las PM10. Esto es debido a que la mayoría de zonas del set de dato “difíciles”, han resultado estar en los datos de validación, por lo que esto se solventaría reentrenando el modelo con un conjunto de datos más grande, para que apareciesen todo tipo de posibles configuraciones en los datos de entrenamiento. Pero aun así los resultados son notablemente mejores utilizando la Red Neuronal Artificial, con más de un orden de magnitud de mejora.

Resultados para NO

En este punto aparecen expuestos los resultados para el NO. Las comparativas de las predicciones pueden verse en la Figura 35 y Figura 36.

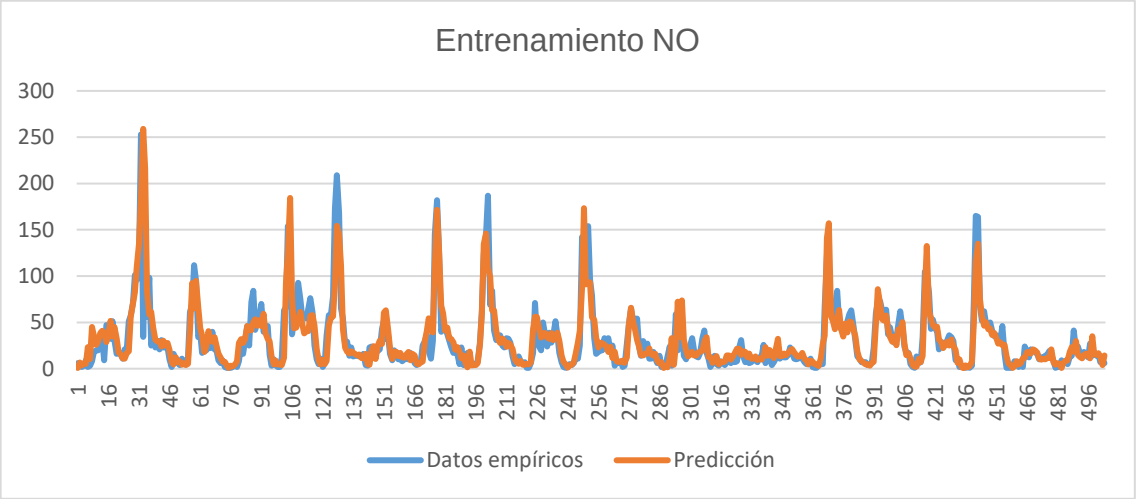


Figura 35 Predicciones para NO a partir de datos de entrenamiento comparadas con datos empíricos

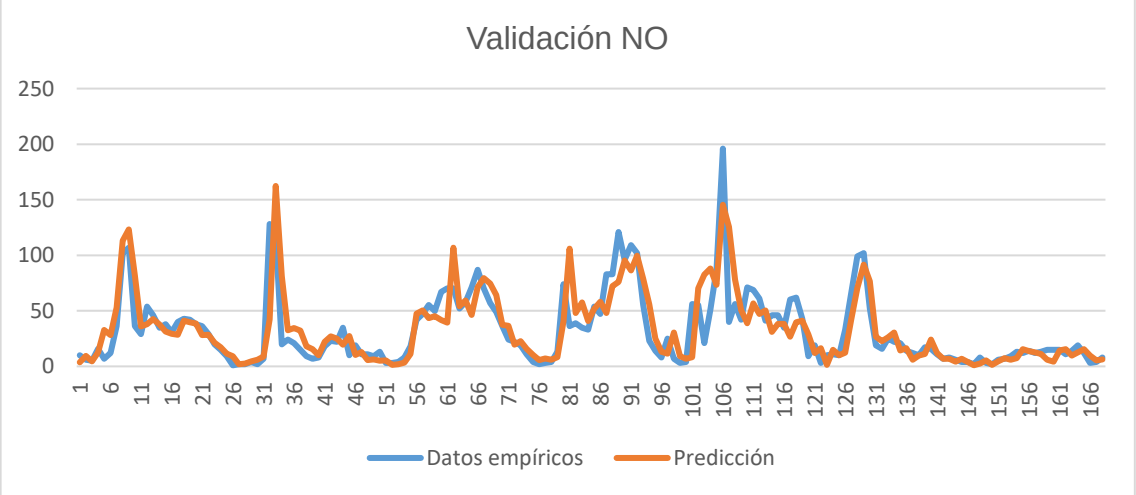


Figura 36 Predicciones para NO a partir de datos de validación comparadas con datos empíricos

Y a continuación, en la Figura 37, puede verla predicción obtenida utilizando el método por regresión lineal simple, con el parámetro más correlacionado con el NO, que como pudo verse anteriormente, es el CO.

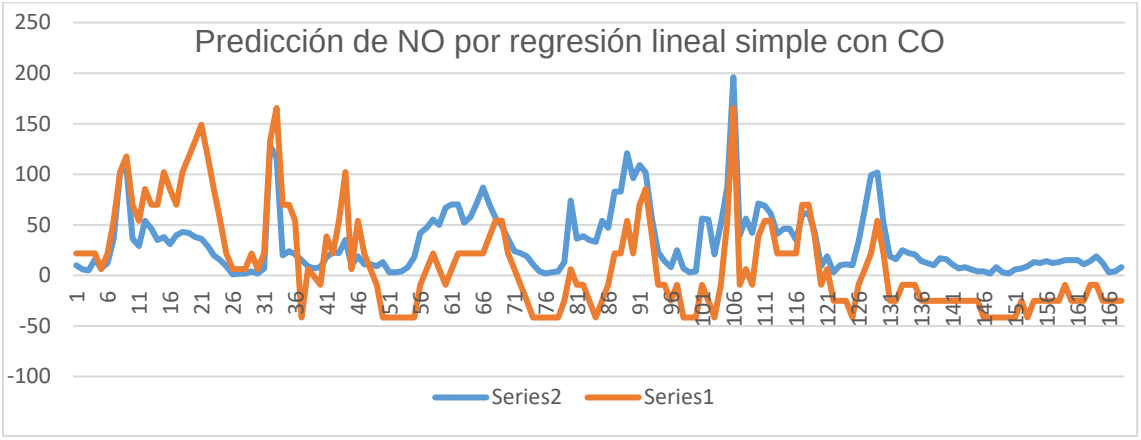


Figura 37 Predicciones para NO con regresión lineal simple partiendo del CO comparadas con datos empíricos

En la Tabla 27, pueden verse los resultados de los EMC, tanto para esta red, como para las predicciones hechas por regresión monovariable.

Tabla 27 Comparativa de resultados de EMC para NO

RNA	**0,39**
CO	1,08
NO_2	1,51
Intensidad 4071	7,61
Intensidad Media	7,81

En este caso, puede observarse la mejoría de la RNA frente al modelo de referencia, tanto en el comportamiento visualizado en los gráficos anteriores, como en el EMC.

Resultados para NO_2

A continuación se muestran los resultados del modelo desarrollado para predecir la concentración del NO_2.

Se ha trabajado del mismo modo que en los casos anteriores. En la Figura 38 y Figura 39, pueden verse comparados los valores predichos por la RNA, con los valores reales, para el conjunto de datos de entrenamiento y validación respectivamente.

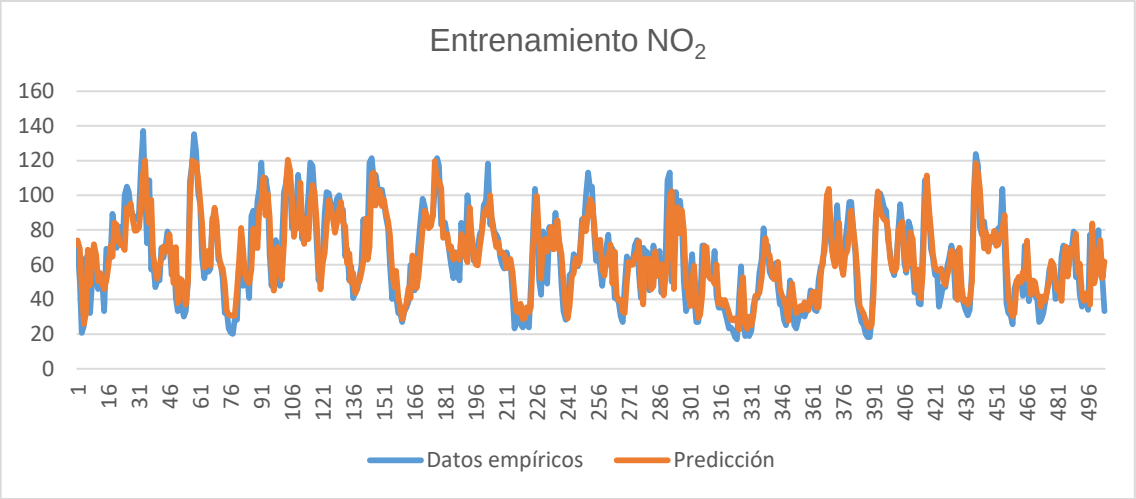


Figura 38 Predicciones para NO_2 a partir de datos de entrenamiento comparadas con datos empíricos

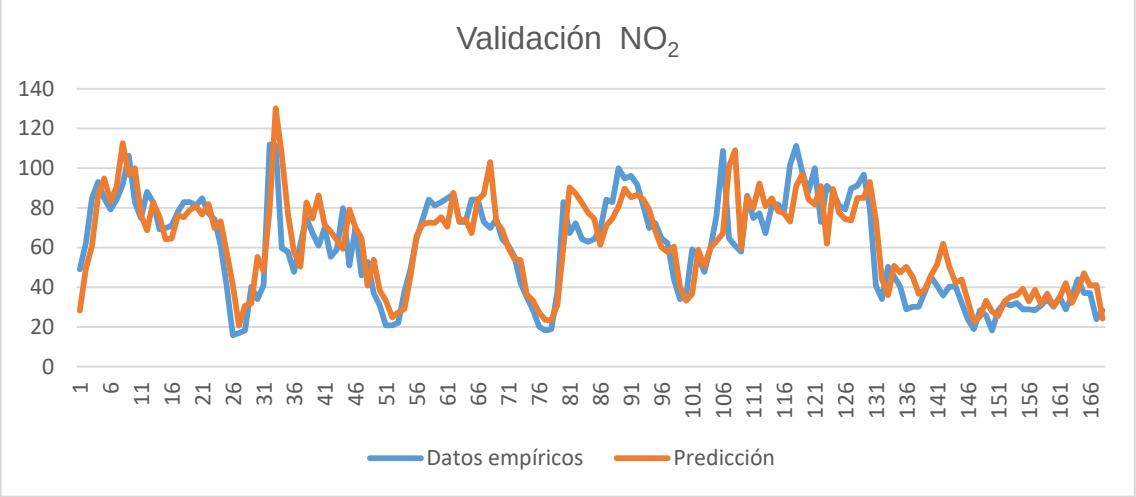


Figura 39 Predicciones para NO_2 a partir de datos de validación comparadas con datos empíricos

Para poder valorar estos resultados, puede verse a continuación en la Figura 40 la predicción hecha utilizando regresión lineal con la concentración de CO, que es el parámetro que está más correlacionada con la concentración de NO_2.

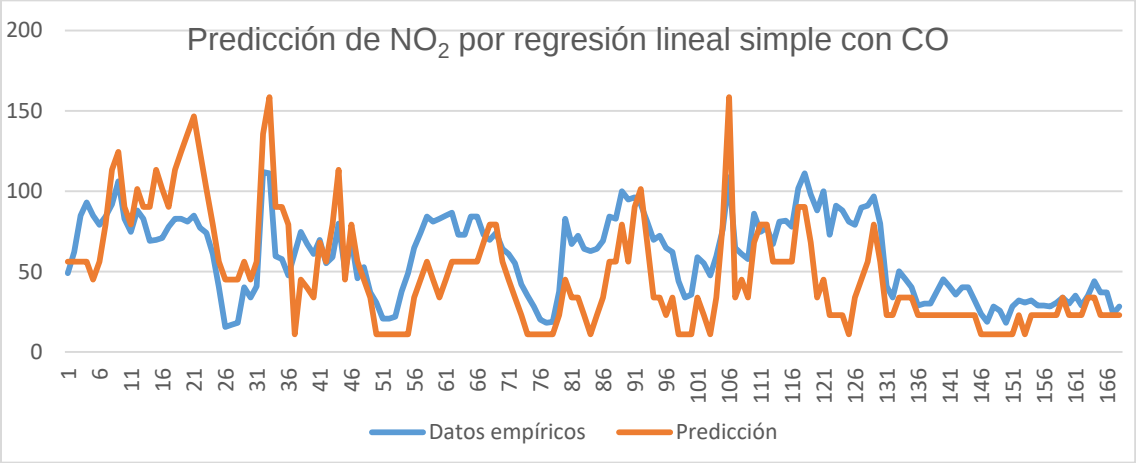


Figura 40 Predicciones para NO_2 con regresión lineal simple partiendo del CO comparadas con datos empíricos

En la Tabla 28, puede verse la comparativa del error medio cuadrático de esta red, con los de las regresiones monovariables simple para el NO_2.

Tabla 28 Comparativa de resultados de EMC para NO_2

RNA	**0,044**
CO	0,14
NO	0,18
O_3	0,19
Intensidad 4071	1,00

Como puede observarse en el gráfico del conjunto de datos de validación, el modelo predice con una precisión significativamente mejor que con el modelo de referencia. Y también se obtiene una notable mejoría en el EMC.

Resultados para O_3

En este punto aparecen expuestos los resultados para el O_3. Las comparativas de las predicciones pueden verse en la Figura 41 y Figura 42.

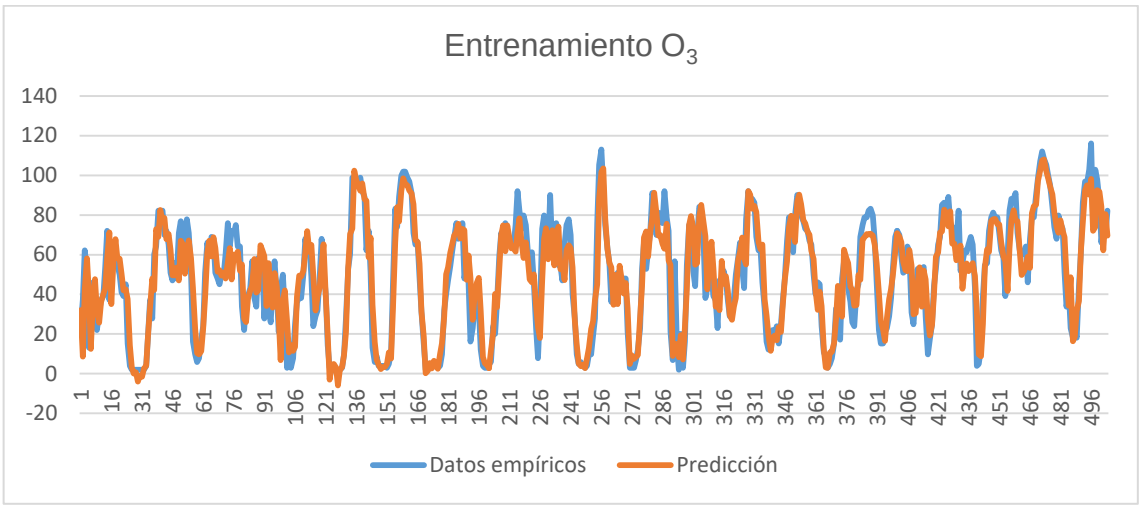


Figura 41 Predicciones para O_3 a partir de datos de entrenamiento comparadas con datos empíricos

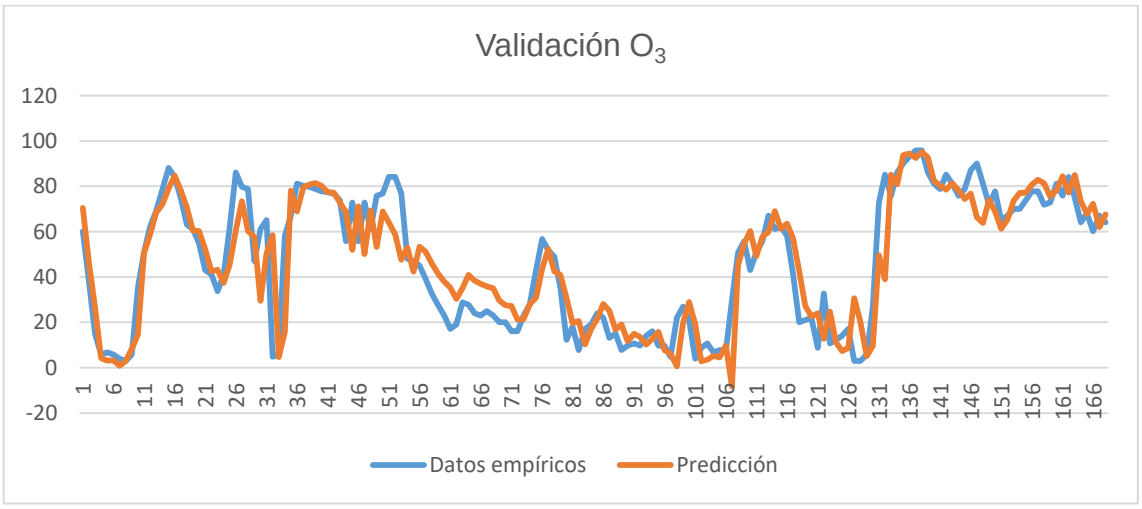


Figura 42 Predicciones para O_3 a partir de datos de validación comparadas con datos empíricos

En la Figura 43 puede verse las predicciones hechas utilizando regresión lineal simple con el parámetro más correlacionado para este caso, la concentración de NO_2.

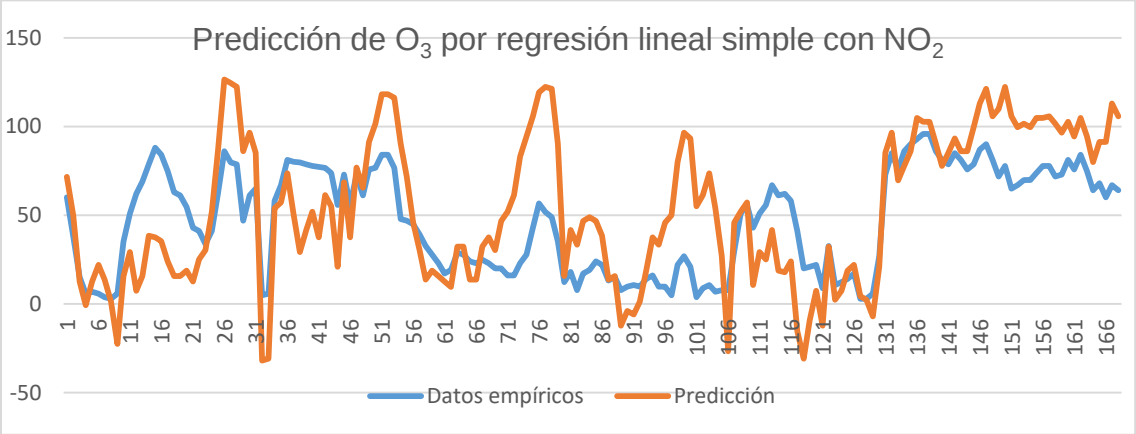


Figura 43 Predicciones para NO_2 con regresión lineal simple partiendo del CO comparadas con datos empíricos

En la Tabla 29, pueden verse los resultados de los EMC, tanto para esta red, como para las predicciones hechas por regresión monovariable.

Tabla 29 Comparativa de resultados de EMC para O_3

RNA	**0,056**
NO_2	0,39
NO	0,98
Velocidad Viento	1,20
Tª media	2,90

En este caso la red ha aprendido con mucha facilidad el comportamiento del Ozono, ha sido el modelo que ha convergido con mayor facilidad, sin tender a sobreajustar los datos en ninguna de las pruebas. Con respecto a los EMC de referencia, se ha obtenido una gran mejoría y en los gráficos puede observase la notable mejoría de la RNA frente al modelo de referencia.

Resultados para CO

Por último, en este apartado, aparecen los resultados para el CO. En la Figura 44 y Figura 45, pueden verse las comparativas de las predicciones, para el conjunto de datos de entrenamiento y de validación respectivamente

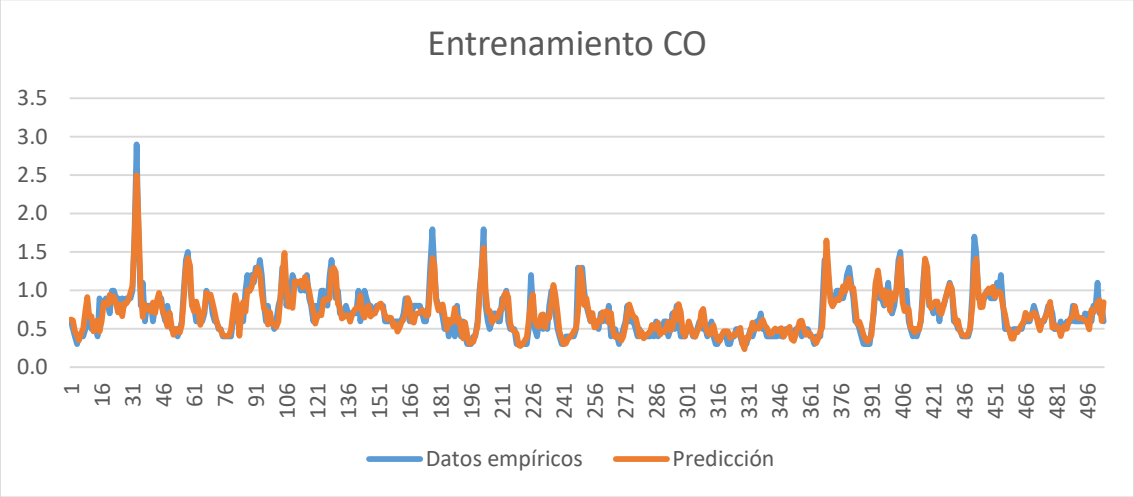


Figura 44 Predicciones para CO a partir de datos de entrenamiento comparadas con datos empíricos

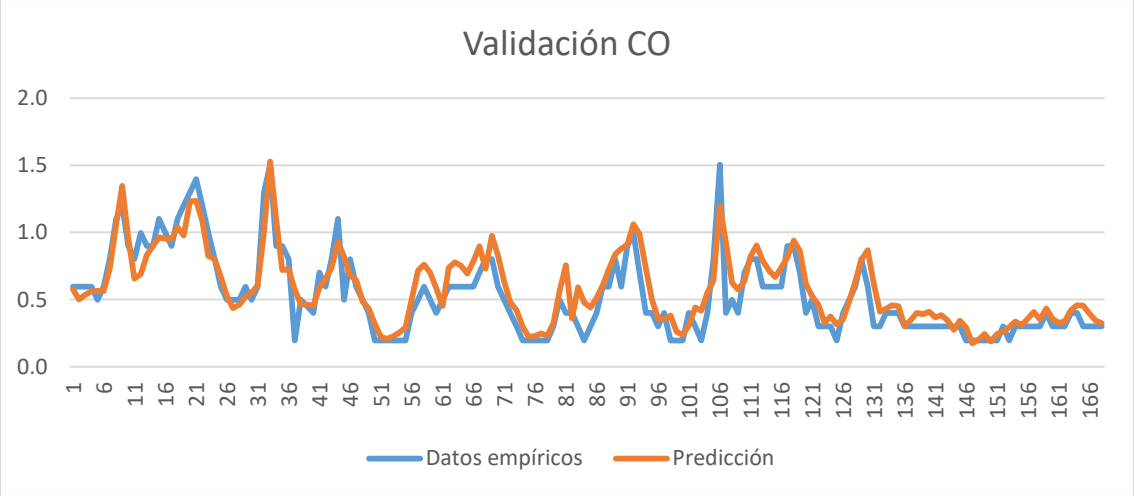


Figura 45 Predicciones para CO a partir de datos de validación comparadas con datos empíricos

Para poder valorar estos resultados, puede verse a continuación en la Figura 46 la predicción hecha utilizando regresión lineal con la concentración de NO, que es el parámetro que está más correlacionada con la concentración de CO.

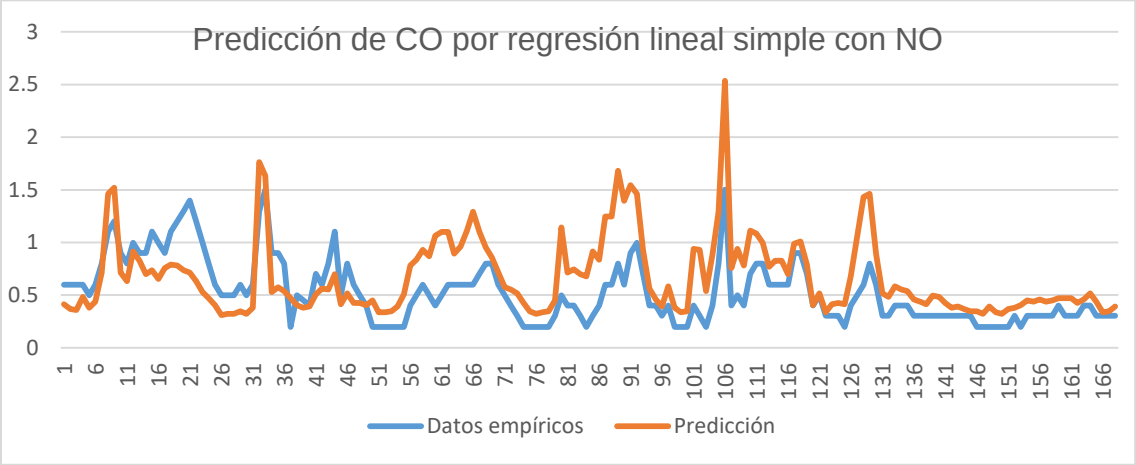


Figura 46 Predicciones para CO con regresión lineal simple partiendo del NO comparadas con datos empíricos

Para cuantificar la mejora de la RNA frente a las predicciones con regresiones monovariables simples, se ha comparado el error medio cuadrático. En la Tabla 30, puede verse la comparativa de estos.

Tabla 30 Comparativa de resultados de EMC para CO

RNA	**0,051**
NO	0,18
NO_2	0,19
Intensidad 4071	33,83
Intensidad Media	6,26

Como puede observarse, en este caso la Red Neuronal Artificial también consigue una notable mejoría respecto a los resultados del ajuste a los datos correlacionados por regresión lineal simple.

10. Conclusiones

En primer lugar se ha conseguido preparar los datos de manera que fueran utilizables para desarrollar modelos predictivos con ellos. Esto se ha conseguido limpiándolos, reparándolos y normalizándolos, obteniéndose los conjuntos de datos necesarios para entrenar y validar los modelos.

Realizando un análisis de correlaciones se han encontrado los parámetros que eran relevantes para las predicciones de cada contaminante analizado en este estudio. Se han hecho predicciones utilizando regresión lineal simple. Estas han resultado ser muy imprecisas, lo cual ha motivado el desarrollo de Redes Neuronales Artificiales y ha proporcionado unos resultados de referencia con los que poder valorar la mejora aportada por la RNA.

Como se ha mostrado, la aplicación de una RNA entrenada a partir de los datos aplicando *Machine Learning* ha generado modelos predictivo de gran precisión, produciendo predicciones mucho mejores que con los modelos de referencia.

Además, los modelos han sido puestos a prueba realizando predicciones con datos que no habían visto durante el entrenamiento. Validando con esto que realmente han conseguido generalizar el comportamiento de los contaminantes frente a los parámetros a partir de los ejemplos disponibles, en lugar de ajustar una curva a los datos sin realmente aprender nada.

11. Futuros trabajos

Como se observó en el estudio inicial de las correlaciones y autocorrelaciones, las distintas variables no solo estaban correlacionadas entre ellas, sino que también lo estaban, algunas en mayor o menor medida que otras, con registros anteriores de estas mismas.

Esta correlación con registros anteriores, motiva que para desarrollar el modelo con mayor precisión este trabaje no solo con los parámetros a t-1, sino que también tenga en cuenta, cuando sea oportuno, medidas a t-2, t-3, etc. En las Redes Neuronales Artificales, tanto en la regresiva lineal como en la logística, esto se ha realizado introduciendo en el modelo los registros a t-1, más los registros a t-2 y t-3 relevantes, como variables independientes de entrada a los modelos.

Esta forma de introducir información anterior a t-1, por un lado tiene los beneficios de poder trabajar con el mismo modelo que solo con t-1 y poder observar las mejorías que introducen, y sin tener que añadir más complejidad a este. Pero por otro lado tiene los inconvenientes, de que aumente el número de nodos de entrada, incrementando con ello la cantidad de variables que el sistema tiene que procesar y optimizar, aumentando con ello el tiempo de cálculo y dificultando la convergencia del modelo.

Además, este método solo aporta al modelo memoria a corto plazo, perdiéndose información que podría resultar relevante para el entrenamiento de este, de registros muy anteriores al t-1. Esto resulta especialmente inconveniente en un problema como este, donde se quiere predecir una variable, que tiene un comportamiento cíclico, tanto a nivel diario, como semanal, e incluso mensual y anual.

Por todo esto, podría ser interesante investigar la aplicación de una Red Neuronal Artificial Recurrente LSTM. Esta contiene los elementos de la red desarrollada, pero añadiendo un nuevo elemento, las células LSTM (*Long Short Term Memory*), las cuales mediante un conjunto de puertas lógicas son capaces de almacenar dependencias a corto y largo plazo [14]. Configurando este tipo de neuronal, probablemente mejorarían los resultados del modelo, aunque sería necesario reprogramar el modelo que construye y entrena las redes, y volver a realizar los ciclos de entrenamiento para cada contaminante.

Además de esta posible mejora del modelo, sería conveniente entrenar la red con un conjunto de datos considerablemente mayor. Por un lado, al introducirle mayor información a la RNA, esta sería entrenada sobre una mayor cantidad de ejemplos, y con estos se podría separar el ruido de los datos de la información relevante para realizar predicciones. Por otro lado, haría el modelo utilizable para predecir a lo largo de todo el año, ya que al contar solo

con datos del mes de Abril, este solo sería aplicable a esta época del año, ya que durante los otros meses el tráfico y las variables ambientales tienen un comportamiento diferente.

12. Agradecimientos

A todos los que me han acompañado a lo largo de estos dos años en los que he realizado este máster. Por haber hecho amena esta travesía vital y haberme hecho sentir en Barcelona como en casa. Y en especial a mi familia y amigos, que siempre me han estado apoyando desde la distancia. Y a Beatriz Sanaú Hornero, mi gran compañera en este viaje, conocerla es sin duda lo mejor que me llevo de mi paso por esta escuela.

A mis profesores Vicente César de Medina Iglesias, Mª Antonia de los Santos López y Jesús Álvarez Flórez, por apoyarme en el desarrollo de este trabajo.

Mención especial al Ayuntamiento de Barcelona, por haber proporcionado los datos necesarios para este estudio.

13. Bibliografía

[1] E. Alonso, T. Martínez, K. Cambra, L. López, E. Boldo, B. Zorrilla, A. Daponte, I. Aguilera, S. Toro, C. Iñiguez, F. Ballester, F. García, A. Plasencia, L. Artazcoz y S. Medina, "Evaluación en cinco ciudades españolas del impacto en salud de la contaminación atmosférica por partículas. Proyecto europeo Apheis. Rev Esp Salud Pública 2005; 79: 297-308.

[2] M. Restrepo, M. Vélez, E. Vallejo, L. Martínez, "Impacto clínico de la contaminación aérea" Archivos de Medicina (Col), vol 16, núm. 2, julio-diciembre, 2016, pp.373-384.

[3] INE-SEMARNAT, "Guía Metodológica para la estimación de emisiones vehiculares en ciudades mexicanas", 2009.

[4] Generalitat de Catalunya, "Dades de qualitat de l'aire" [online]. Disponible en: http://dtes.gencat.cat/icqa/

[5] Agencia Estatal de Meteorología [online]. Disponible en: https://opendata.aemet.es/

[6] S. Finardi, R. De Maria, A. D'Allura, C. Cascone, G. Calori, F. Lollobrigida, "A deterministic air quality forecasting system for Torino urban area, Italy", Environmental Modelling and Software, 23(3):344–355, 2008.

[7] M.R. Cañada, A. Moreno-Jiménez, "El contraste intraurbano de la contaminación del aire por NO_2 y O_3: Estudio en grandes ciudades españolas con datos observados e interpolados con SIG", Geofocus. Revista Internacional de Ciencia y Tecnología de la Inofrmación Geográfica, vol. 19, 27-53, 2017

[8] J.M. Barrón, "Modelado de un sistema de supervisión de la calidad del aire usando técnicas de fusión y redes neuronales", Universidad Politécnica de Madrid, 2010.

[9] Y. Bai, Y. Li, X. Wang, J. Xie, C. Li, "Air pollutants concentrations forecasting using back propagation neural network based on wavelet decomposition with meteorological conditions", Atmos Pollut Res 7:557e566, 2016.

[10] M. Oprea, S.F. Mihalache, M. Popescu, "A comparative study of computational intelligence techniques applied to PM2.5 air pollution forecasting", Proceedings of the 6th International Conference on Computers Communications and Control (ICCCC), Baile Felix, Oradea, Romania, pp. 103-108, May 2016.

[11] E. Pardo, N. Malpica, “Air Quality Forecasting in Madrid Using Long Short-Term Memory Networks”, Biomedical Applications Based on Natural and Artifical Computing, pp.232-239, 2017.

[12] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G.S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mane, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, X. Zheng, “Large-Scale Machine Learning on Heterogeneous Distributed Systems”, 2015

[13] T. Hastie, R. Tibshirani, J. Friedman, “The elements of statistical learning. Data Mining, Inference and Prediction” Springer Series in Statistics, Second Edition, 2013.

[14] S. Hochreiter, J. Schmidhuber, “Long short-term memory”, Neural computation 9(8): 1735-1780, 1997.