Instituto Tecnológico y de Estudios Superiores de Monterrey

Campus Monterrey

Aplicación de métodos multivariados en ciencia de datos (Gpo 101)

Impacto de factores no antropogénicos en los niveles de contaminación del aire

Equipo 4 | Integrantes:

Daniel Eduardo Arana Bodart	A01741202
Isis Yaneth Malfavón Díaz	A01705838
Santiago Juárez Roaro	A01705439
Ericka Sofía Rodríguez Sánchez	A01571463
Alfredo André Durán Treviño	A01286222

Profesores:

Dra. Blanca Rosa Ruiz Hernández

&

Mtro. Rodolfo Fernández de Lara Hadad

7 de septiembre del 2024

Índice

Resumen

Durante las últimas cinco semanas se realizó el proyecto que en las siguientes páginas se detalla. Este tiene como eje principal la calidad del aire en el estado de Nuevo León y los diversos factores que intervienen en la medición y monitoreo relacionado a este índice. El objetivo de este informe es contestar la pregunta de investigación ¿Existe un impacto significativo de los factores no antropogénicos en los niveles de contaminantes del aire? Para ello se empleó la base de datos proporcionada por el socio formador del bloque centrada en los años 2023 y 2024. Además, se emplearon diversos métodos vistos durante el bloque aplicación de métodos multivariados en ciencia de datos con el propósito de responder la pregunta de investigación con un fundamento y sustento estadístico.

Para lograr esto fue necesario realizar un análisis de los datos históricos proporcionados por el socio formador, en donde se encontró que existían datos nulos que eran significativos, por lo que no era adecuado para el análisis eliminarlos. Para tratar la situación se elaboró un clasificador de aprendizaje supervisado no paramétrico llamado K-Nearest Neighbors (KNN), esto debido a que los datos presentan dependencia entre sí al pertenecer a una serie de tiempo y otras técnicas de imputación no eran representativas. Utilizando el modelo KNN se rellenaron los datos faltantes de interés. Posterior a esto se escalaron los datos utilizando el Standard Scaler de la paquetería Scikit Learn para obtenerlos en las unidades requeridas para el modelo. Una vez escalados los datos se buscó normalidad en las variables, al no encontrarla en ninguna, se tomó la decisión de transformar los datos para asegurar la normalidad del modelo. Se probaron varias técnicas de normalización como Box-Cox y Yeo-Johnson, sin embargo, ninguna funciono correctamente para poder realizar la modelación, finalmente se utilizó la transformación de normalización de cuantiles ordenados de la librería bestNormalize y se utilizaron las variables que sí lograron seguir una distribución normal después de la transformación.

Con los datos analizados y transformados se procedió con el modelo, en donde se realizó una regresión lineal multivariada a través de RStudio, mediante la cual se obtuvo un modelo para cada contaminante en base a cuatro variables meteorológicas: presión atmosférica, humedad relativa, velocidad del viento y temperatura. Una vez con las variables dependientes y las variables predictoras elegidas se realizaron los ocho modelos que posteriormente fueron validados al unísono por medio de normalidad en los residuos, homocedasticidad, máxima correlación canónica, coeficientes de sesgo y curtosis e histogramas de densidad. Todos estos supuestos fueron conseguidos adecuadamente asegurando la validez de los modelos. Sin embargo, el desempeño de los modelos fue deficiente, llegando a un valor máximo de R^2 ajustada de 0.46, por lo que se optó por realizar otro modelo para contestar adecuadamente la pregunta de investigación: Redes neuronales.

La red neuronal es una forma de inteligencia artificial que asemeja el funcionamiento del cerebro humano. Por la naturaleza de su funcionamiento, no es necesario que las variables cumplan supuestos tan rigurosos como lo son para modelos estadísticos, por lo que para este modelo se utilizaron todas las variables no antropogénicas sin necesidad de cambios más allá del escalamiento. La arquitectura de la red consiste en dos capas escondidas con 80 y 64 neuronas, ambas capas tienen una función de activación ReLu y utilizan el optimizador Adam, como función de perdida se midió el error medio absoluto (MAE), y se integraron algunas técnicas para evitar el sobreajuste. El mejor resultado para la predicción de Ozono (O3) fue un MAE de 0.01014 y en general un ajuste correcto al comportamiento del componente.

Finalmente, se encontró que la respuesta a la pregunta inicial es que los factores no antropogénicos si ejercen un impacto en la contaminación del aire significativo, pero es claro que no pueden explicar completamente el comportamiento de los contaminantes por sí solos.

Introducción y justificación

El socio formador con el que se trabajó para este proyecto fue el Sistema Integral de Monitoreo Ambiental (SIMA), el organismo encargado del monitoreo de los diferentes niveles de contaminantes del aire a través de 15 centros de monitoreo en el estado de Nuevo León. Esta red de monitoreo se actualiza cada hora para determinar el estado de la calidad del aire en la zona metropolitana de Monterrey. De acuerdo con la Organización Mundial de la Salud, la calidad del aire representa un gran problema para la sociedad en el sentido que es uno de los principales causantes de problemas respiratorios, enfermedades de corazón y cáncer de pulmón, además se estima que aproximadamente 6.7 millones de personas pierden la vida anualmente debido a los efectos de contaminación del aire en el ambiente y en el hogar, siendo el primero de estos el principal causante de muertes, es por esto que el análisis de la calidad del aire es de suma importancia ya que con un análisis adecuado se pueden establecer medidas de prevención certeras para mantener a la población a salvo.

Ahora bien, existen diferentes tipos de contaminantes que pueden encender las alarmas y dictaminar el nivel de calidad del aire en el ambiente, entre los cuales la Organización Mundial de la Salud menciona en el capítulo 3 del libro en el que se establecen las pautas para la calidad del aire. En ellos se menciona que hay un total de 6 contaminantes diferentes con los que se tienen que tener cuidado, entre los cuales se encuentran: PM2.5, PM10, Ozono, Dióxido de Nitrógeno, Dióxido de Azufre y Monóxido de Carbono, los cuales presentan diferentes daños a la población sensible que se encuentre expuesta a ellos por tiempo prolongado, tales como problemas respiratorios en caso de las partículas PM2.5 y PM10; inflamación de las vías respiratorias por medio del Ozono; ataques cardíacos y problemas cardiovasculares por medio del dióxido de nitrógeno; repercusiones en el área ocular tales como ardor en los ojos por medio del dióxido de azufre; y daños cerebrales irreparables por exposición prolongada al monóxido de carbono.

Tomando en cuenta la importancia de estos contaminantes y el impacto que tienen en la vida de los seres humanos, es de suma importancia el monitoreo de su comportamiento, pero también lo es la capacidad predictiva de su comportamiento. La modelación matemática predictiva de estos compuestos permite que se puedan crear planes de acción a largo plazo de acuerdo al comportamiento de estos, o que pueda haber un mayor entendimiento de las interrelaciones y dependencias que existen entre los compuestos y el mundo, lo que finalmente

ayuda para la toma de decisiones de cómo tratar con el problema de la contaminación del aire de la mejor manera.

Problemática y objetivo

Como se mencionó anteriormente, la contaminación del aire es un tema que afecta a toda la humanidad. De acuerdo con la ONU "En 2019, el 99% de la población mundial estaba viviendo en lugares donde no se cumplían los lineamientos de la OMS para la calidad del aire" (2021). No resulta extraño que en ciudades grandes o industrializadas como Monterrey no se cumplan los lineamientos de calidad del aire, ya que es sencillo e incluso intuitivo encontrar razones para explicarlo: la cantidad de tráfico y coches, la refinería, la densidad de población, entre otras. Sin embargo, estos factores no existen en todas las partes del mundo y aun así los lineamientos no se cumplen. Siguiendo esta línea de pensamiento, se realizó la hipótesis de que los factores no antropogénicos, los cuales suelen ser percibidos como ruido en el análisis de la calidad del aire, tienen un rol importante en cuanto al comportamiento de los contaminantes, y se quiso observar qué tanto pueden explicar la calidad del aire sin tomar en cuenta los factores antropogénicos. Más específicamente, se busca contestar a la pregunta ¿Existe un impacto significativo de los factores no antropogénicos en los niveles de contaminantes del aire? por medio de dos modelos, uno estadístico y una red neuronal.

Preparación de la base de datos

Esta etapa se realizó por medio de Python. Antes de realizar cualquier modificación a la base de datos, es necesario analizarla. Para el objetivo, se decidió trabajar con la base de 2023 y 2024, la cual cuenta originalmente con 13873 filas y 239 columnas, que son las variables. La primera columna contiene los datos del momento en el que se captó la información, incluyendo la fecha y hora del registro. Las demás columnas pertenecen a la información de los valores capturados de contaminantes, partículas, o datos climatológicos. La elevada cantidad de columnas se debe a que las mismas variables son registradas de manera separada para cada zona, por lo que se explican en general en la tabla a continuación.

Variable	**Unidades**	**Descripción**	**Tipo de dato**
CO	ppm (partes por millón)	Monóxido de carbono	Numérico
NO	ppb (partes por billón)	Monóxido de nitrógeno	Numérico
NO2	ppb (partes por billón)	Dióxido de nitrógeno	Numérico
NOX	ppb (partes por billón)	NO + NO2	Numérico
O3	ppb (partes por billón)	Ozono	Numérico
PM10	microgramos/metro cúbico	Material Particulado menor a 10 micrómetros	Numérico

PM2.5	microgramos/metro cúbico	Material Particulado menor a 2.5 micrómetros	Numérico
PRS	milímetros de mercurio	Presión atmosférica	Numérico
RAINF	milímetros por hora	Precipitación	Numérico
RH	porcentaje	Humedad relativa	Numérico
SO2	ppb (partes por billón)	Dióxido de azufre	Numérico
SR	kilovatio por metro cuadrado	Radiación solar	Numérico
TOUT	grados Celsius	Temperatura	Numérico
WSR	kilómetros por hora	Velocidad del viento	Numérico
WDV	grados	Dirección del viento	Numérico

Tabla 1. Información general de las variables en la base de datos

La base cuenta con un total de 3,314,930 datos de los cuales 374285 son faltantes, lo cual representa el 11.3% de los datos. En base al menor número de valores nulos fue que se decidió en qué zona se realizaría el proyecto, en general, la zona con menor cantidad de valores nulos es la sureste 2, es decir, el medidor ubicado en la zona Juárez. Posterior a esto se implementaron las zonas centro y norte, esto debido a que son zonas alejadas de la sureste 2, por lo que, si esta llega a fallar para el objetivo por cuestiones estadísticas, se tienen zonas que por ubicación geográficas tendrían menos dependencia a la zona Juárez. Para manejar los valores nulos se decidió emplear un método de clasificador de aprendizaje supervisado no paramétrico llamado K-Nearest Neighbors, el cual se encarga de realizar predicciones de valores basado en la cercanía de un dato desconocido con otros datos que si se conocen, por lo que este método es adecuado debido a que se busca imputar datos considerando la dependencia de estos datos con sus predecesores cercanos. Con el método K-Nearest Neighbors los datos quedaron de la siguiente manera:

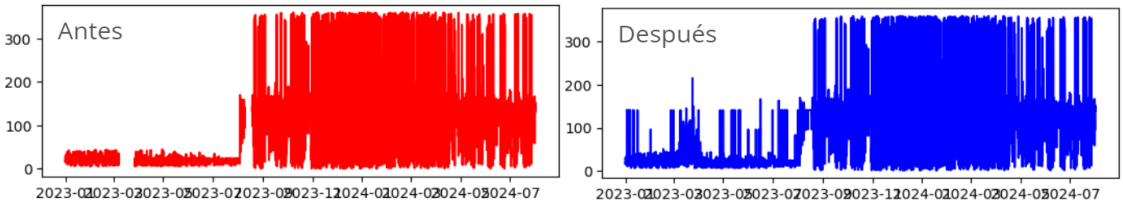


Figura 1. Datos atípicos antes y después de KNN

En la figura 1 se puede observar cómo en los datos originales se encuentran secciones considerables con datos faltantes y una vez aplicado el método K-Nearest Neighbors.

De manera similar, fue necesario realizar un escalamiento de ciertas variables cuyas unidades no corresponden con lo necesario para realizar un análisis adecuado, para mayor información de cómo se escalaron estas variables visitar los anexos. Las variables escaladas fueron: NO, NO2, NOX y O3 para que en lugar de estar representadas en ppb, fuera en ppm.

Otro problema que se presentó con los datos se encuentra en función de los datos atípicos, como se muestra en la siguiente gráfica:

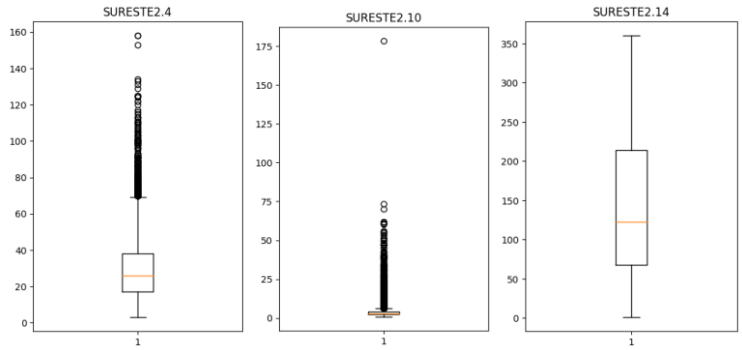


Figura 2. Boxplots de tres variables de la zona sureste.

En la figura 2 se puede observar que dependiendo de las variables se tienen más o menos datos atípicos y con mayor o menor nivel de influencia en el modelo. Es en base a estos datos atípicos que no se obtiene normalidad en ninguna de las variables originales y es necesario realizar una transformación de los datos. Primeramente, se intentó transformar las variables por medio de métodos vistos durante el curso; sin embargo, ninguno de los métodos resultó adecuado debido a la naturaleza de los datos, por lo que se empleó otro método que asegura normalidad por medio de la librería bestNormalize, en donde se obtuvo que todas las variables predictoras a excepción de la radiación solar y la velocidad del viento presentan un comportamiento normal. Un ejemplo de la transformación Best-Normalize se encuentra a continuación.

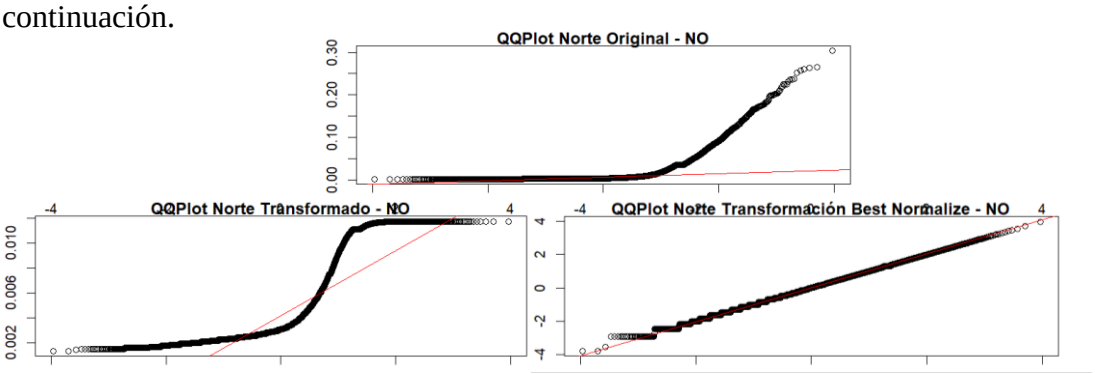


Figura 3. Comparación QQPlot del contaminante NO sin transformación (centro), con la transformación Power transform (izquierda) y la transformación bestNormalize (derecha)

Como se puede observar en la figura 3 luego de aplicar la transformación BestNormalize se observa que, de estar bastante sesgados, los datos se convierten en una distribución normal.

Es importante clarificar que el método Best-Normalize funciona utilizando la técnica de normalización de cuantiles ordenados (Ordered Quantile Normalization), el cual es un

proceso semi paramétrico que utiliza la función de transformación normal usando los cuantiles y la interpolación y extrapolación. El método utiliza una función diferente dependiendo del dato; de acuerdo a Peterson, R. A. y Cavanaugh, J. E. (2019) se define **x** como los datos originales, z como el vector de los ranks de **x**, y **x*** como una nueva observación que puede o no pertenecer a **x**. La transformación se define de la siguiente manera:

$$g(x^* \mid \mathbf{x}) = \begin{cases} f(x^*) & \text{if } x^* \in \{\mathbf{x}\} \\ \dfrac{f(x_u) - f(x_l)}{x_u - x_l} & \text{if } x^* \notin \{\mathbf{x}\} \text{ and } \min \mathbf{x} < x^* < \max \mathbf{x} \\ r(x^*; \mathbf{x}) & \text{if } x^* < \min \mathbf{x} \text{ or } x^* > \max \mathbf{x} \end{cases}$$

Para más información visitar el anexo Etapa 3.

Otro punto importante de mencionar es que a partir de este punto se empleó únicamente la base de datos del centro de medición norte ya que es el único cuyos datos se pudieron transformar para asegurar normalidad.

Modelación y validación

Estas etapas se realizaron por medio de RStudio. Una vez con los datos normalizados se empezó con la realización del modelo de regresión lineal multivariada. Para ello fue necesario eliminar aquellas variables predictoras que previamente fueron clasificadas como no normales a pesar de la preparación rigurosa que se les dio a los datos. Posteriormente se realizó el modelo en donde todas las variables resultaron significativas para todos los modelos (exceptuando el intercepto). Los modelos obtenidos son los siguientes:

CO = 0.2135X1 + 0.0508X2 - 0.1717X3 - 0.4419X4 adj R^2 = 0.1889

NO = 0.0006 - 0.1234X1 + 0.0726X2 - 0.0336X3 - 0.4011X4 adj R^2 = 0.2216

NO2 = -0.1588X1 -0.1811X2 - 0.1410X3 - 0.5031X4 adj R^2 = 0.3297

NOX = -0.2013X1 + 0.0127X2 - 0.1422X3 -0.5176X4 adj R^2 = 0.3758

O3 = 0.0012 + 0.4820X1 -0.2154X2 + 0.1695X3 + 0.3245X4 adj R^2 =0.464

PM10 = -0.0414X1 - 0.2882X2 - 0.24X3 - 0.2747X4 adj R^2 = 0.207

PM2.5 = 0.0007 - 0.0448X1 - 0.0973X2 - 0.3265X3 - 0.3146X4 adj R^2 = 0.1854

SO2 = 0.0002 + 0.1641X1 + 0.1374X2 - 0.0261X3 + 0.0346X4 adj R^2 = 0.05619

Donde:

X1 = Temperatura

X2 = Humedad Relativa

X3 = Presión Atmosférica

X4 = Velocidad del Viento

Una vez con los modelos se realizó un análisis de varianza multivariado (MANOVA) para revisar si las variables predictoras en los modelos son significativas en conjunto donde se obtuvo el siguiente resultado.

	Df	Wilks	approx F	num Df	den Df	Pr(>F)
X1	1	0.48962	1805.67	8	13858	< 2.2e-16
X2	1	0.76089	544.36	8	13858	< 2.2e-16
X3	1	0.82181	375.61	8	13858	< 2.2e-16
X4	1	0.70238	734.01	8	13858	< 2.2e-16
Residuals	13865					

Tabla 2. Pruebas Wilks para efecto significativo de las variables predictoras en los modelos.

En la tabla 2 se observa como la prueba Wilks tiene un valor-p muy cercano a 0, lo que da evidencia estadística para rechazar la hipótesis nula que dictamina que las variables dependientes no tienen un efecto significativo. Entonces se concluye que sí que tienen un efecto significativo en conjunto para el modelo.

Por otro lado, en cuestión de la validación del modelo se realizó un diagrama de dispersión de los residuos con el objetivo de demostrar el supuesto de homocedasticidad de los residuos, el resultado se muestra a continuación.

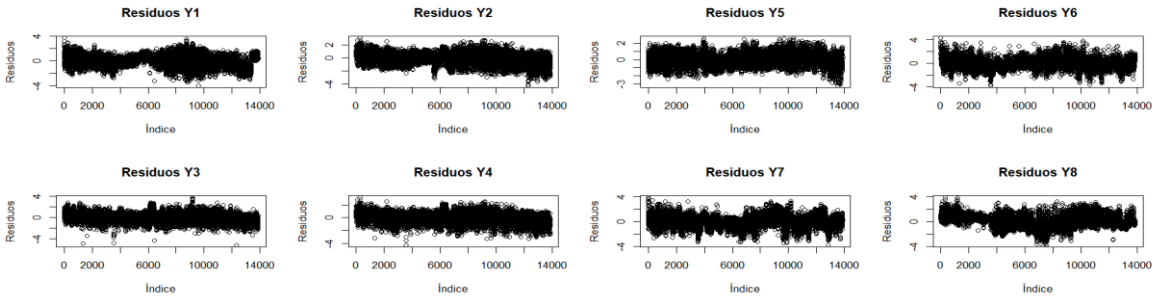


Figura 4. Gráficas de dispersión de los residuos de los modelos de regresión lineal multivariada.

En la figura 4 se observa que los residuos se encuentran centrados en 0 bastante dispersos y sin demostrar un patrón claro, lo que indica que los modelos cumplen con el supuesto de homocedasticidad de los residuos.

También se revisó la normalidad de los residuos por medio de gráficos QQPlot e histogramas de densidad de los residuos, en estos casos solo se muestran los resultados del Ozono, pero las demás gráficas se pueden observar en el anexo del código.

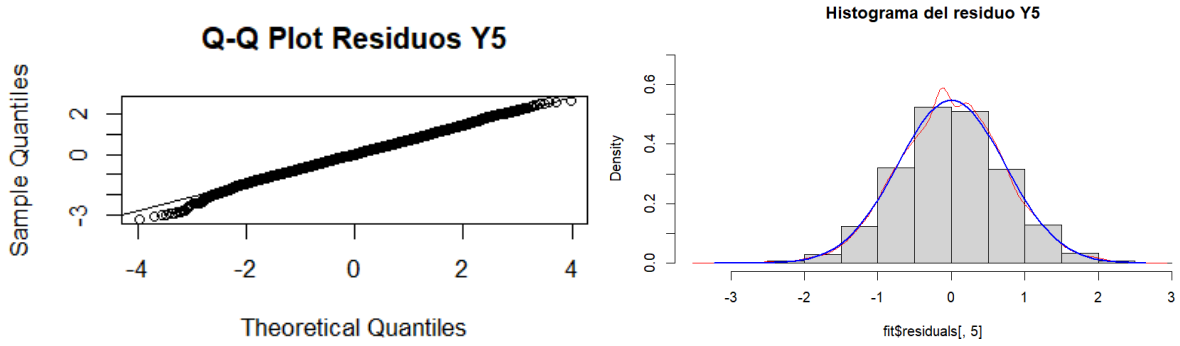


Figura 5. Demostración de normalidad de los residuos en el modelo del Ozono.

En la figura 5 se observan dos pruebas visuales que se realizaron para demostrar la normalidad de los residuos, en donde se demuestra que los residuos se encuentran bastante bien alineados con la recta de normalidad (izquierda), mientras que también se demuestra que la normalidad teórica se encuentra muy cerca de la distribución real (derecha), demostrando que los residuos tienen una distribución normal.

Otra cuestión que se toma en cuenta para la normalidad de los residuos son los coeficientes de sesgo y curtosis de los modelos, cuyos resultados se muestran a continuación.

	Skew	Kurtosis
Y1	-0.20955212	0.05842771
Y2	-0.61655437	0.40840748
Y3	0.01577122	1.14213161
Y4	-0.40812273	0.69004199
Y5	-0.06138478	0.31030358
Y6	0.09562501	0.83478191
Y7	-0.04679880	0.50511800
Y8	-0.01887610	0.11691338

Tabla 3. Coeficientes de sesgo y curtosis de los modelos de regresión lineal multivariada.

En la tabla 3 se muestran los diferentes coeficientes de sesgo y curtosis de los modelos creados y se puede observar que existen variables sin ningún tipo de alerta como lo son CO (Y1), Ozono (Y5) y SO2 (Y8), mientras que las demás variables tienen un poco de sesgo o curtosis; sin embargo, no existe una sola variable que tenga tanto sesgo como curtosis, por lo que con esto también se puede demostrar en cierta manera la normalidad de los residuos.

Por último, se buscó demostrar la linealidad de las variables involucradas, para esto se empleó la correlación canónica, en donde se obtuvo un resultado de 0.759, este valor al ser relativamente cercano a 1, indica que existe una relación lineal aceptable entre las combinaciones lineales que se crearon para este supuesto, por lo tanto, se tiene una linealidad entre variables aceptable, por lo que se terminan cumpliendo casi todos los supuestos.

El único supuesto que no se logró cumplir adecuadamente es la cuestión de multicolinealidad entre las variables para representar el modelo, pues la determinante de la matriz de correlación entre las variables es de 0.0027, inferior al límite permitido de 0.01, por lo que se concluye que existe una multicolinealidad entre las variables que no es recomendada, es por esto que se decidió emplear otro modelo con el objetivo de mejorar los resultados obtenidos.

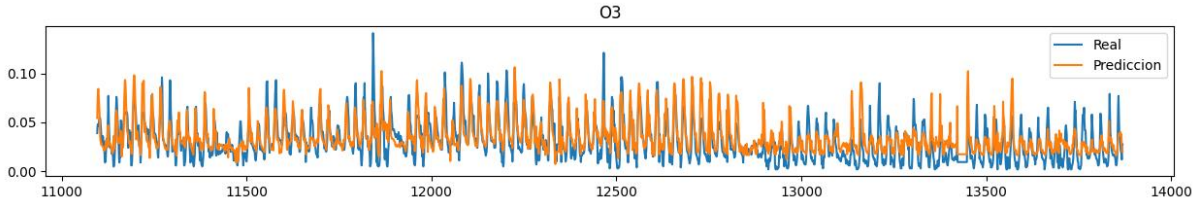
Por otra parte, también se realizó una red neuronal recurrente, para que tomara en cuenta la dependencia de los datos y el tiempo. Para la red se utilizaron los datos sin transformar, por lo que se incluyeron todas las variables a excepción de las que se eliminaron al inicio debido al ruido que podían causar. Se escalaron los datos utilizando el Standar Scaler de Scikit Learn que utiliza el z-score y se dividieron los datos en dos conjuntos: 80% datos de entrenamiento y 20% datos de prueba.

El tipo de red utilizada es una red de memoria a corto-largo plazo (LSTM) para que pudiera procesar la dependencia del tiempo. Su arquitectura consiste de una capa inicial, una capa de regularización tipo dropout, una capa escondida completamente conectada (o densa) de 64 neuronas y función de activación ReLu, otra capa dropout, otra capa densa de 32 neuronas y la misma función de activación, y una capa de salida de una neurona. Además, se utilizó la función de pérdida la medición del error medio absoluto (MAE) usando el optimizador Adam, la tasa de aprendizaje inicial de 0.00001 y 150 épocas.

Algunas de las medidas tomadas para prevenir el sobreajuste fueron el uso de la función EarlyStopping para entrenar el modelo solo con el número de épocas adecuado y el reduceLROnPlateau para reducir automáticamente la tasa de aprendizaje gradualmente y lograr que el gradiente descendiente converja.

Resultados

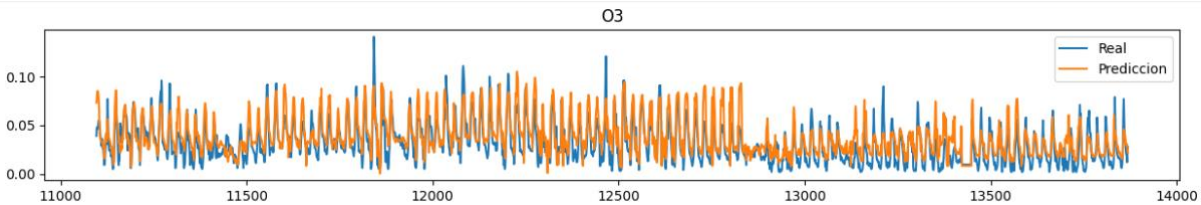
Se realizaron varios experimentos con la red neuronal. Primero se hizo una red que predijera todos los contaminantes, sin embargo, había una pérdida de validación del 6.1481. Para reducir la complejidad del modelo y eliminar el ruido centramos la predicción de la red en un solo componente, el Ozono. Con este nuevo enfoque realizamos un modelo con solo los factores no antropogénicos, el cual obtuvo un MAE de 0.0116. La siguiente gráfica muestra el comportamiento de los valores reales de O3 (línea azul) contra los valores que la red neuronal predice (línea naranja).



Gráfica 1. Rendimiento de la red neuronal con factores no antropogénicos

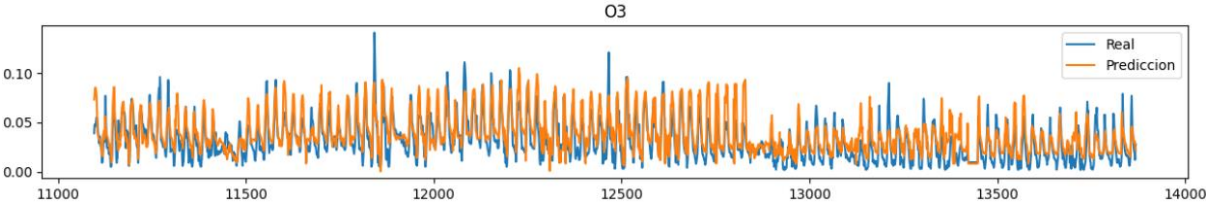
En la gráfica 1 se observa que en general el modelo reconoce el patrón de los datos y sigue la tendencia de los valores reales teniendo un error medio absoluto (MAE) de 0.01014; sin embargo en varios casos se queda corto y no logra asemejar los valores extremos muy altos ni muy bajos. Esto implica que no logra captar la variación total de los datos.

Para contestar la pregunta de investigación se decidió probar la misma red neuronal pero agregando variables antropogénicas como lo son NO, NO2 y NOX. A continuación se muestra el desempeño de las demás redes agregando una a una las variables precursoras del Ozono.



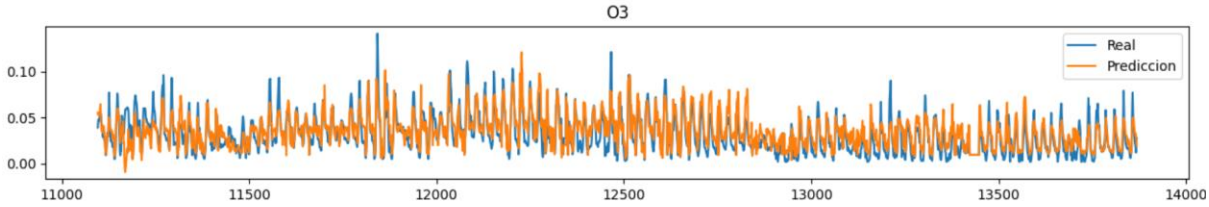
Gráfica 2. Rendimiento de la red neuronal con factores no antropogénicos + NO.

En la gráfica 2 se observa que al agregar la variable NO, el modelo sigue sin poder predecir adecuadamente estos picos extremos que se presentan de manera aleatoria, además se obtuvo un MAE de 0.01611.



Gráfica 3. Rendimiento de la red neuronal con factores no antropogénicos + NO + NO2.

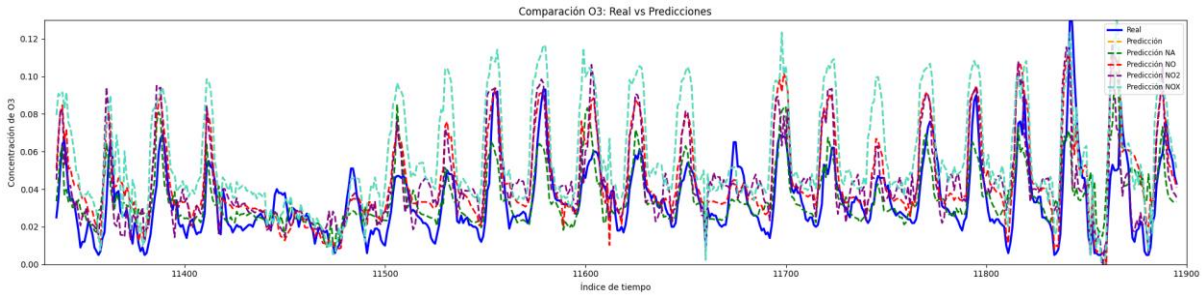
En la gráfica 3 se observa que al agregar NO y NO2 al modelo original se siguen sin poder predecir adecuadamente las variaciones de los datos reales, además, se obtuvo un MAE de 0.01023.



Gráfica 4. Rendimiento de la red neuronal con factores no antropogénicos + NO + NO2 + NOX

Por último, en la gráfica 4 se observa que al agregar todas las variables antropogénicas se sigue sin poder predecir adecuadamente todos los valores extremos de los datos reales, además que se obtuvo un MAE de 0.0196.

A modo de resumen, los resultados gráficos de estas tres redes neuronales se muestran en la siguiente gráfica junto con el modelo original y los datos reales.



Gráfica 5. Comparación del rendimiento de las diferentes redes neuronales.

La línea azul representa los valores reales, como se mencionó anteriormente, el modelo con solo factores no antropogénicos (línea verde punteada) es el que mejor se ajusta. Por otra parte, es importante notar que los modelos que utilizan los precursores del O3 (líneas punteadas rojo, morado y turquesa) parecen quedarse cortas a los valores reales, a diferencia del mejor modelo.

Discusión y conclusiones

Retomando la pregunta de investigación ¿Existe un impacto significativo de los factores no antropogénicos en los niveles de contaminantes del aire? En base a los resultados obtenidos y al proceso realizado durante las últimas 5 semanas es que se considera que se logró encontrar una respuesta concluyente de acuerdo con los datos que se tenían a disposición.

Inicialmente se consideró que los factores no antropogénicos no iban a tener un impacto significativo para predecir los niveles de contaminación del aire esto debido a que los modelos obtenidos por medio de la regresión lineal multivariada no presentaron buenos resultados, más bien se obtuvieron resultados bastante deficientes con valores de R^2 ajustada tan bajos como 0.05 para el contaminante SO2, mientras que el contaminante con el valor de R^2 ajustada más alto fue el Ozono con un valor de apenas 0.46,

Es por esto que la respuesta para la pregunta de investigación iba a ser que los factores no antropogénicos no presentan un impacto significativo en los niveles de contaminantes del aire; sin embargo, después de un poco de análisis de resultados se llegó a la conclusión de que

el desempeño del modelo original podía ser tan deficiente por todas las transformaciones que tuvieron que sufrir los datos, mientras que variables muy importantes en el análisis inicial se quedaron fuera debido a que a pesar de las transformaciones no se logró demostrar normalidad en ellas, tales como la radiación solar y la dirección del viento, de hecho, antes del análisis de los datos, en la etapa de lluvia de ideas se consideraba que estas dos variables justamente junto con la presión atmosférica serían las que más podrían llegar a afectar, y dejar fuera del modelo a dos de ellas fue un golpe bajo para los modelos.

Una vez que se llegó a este pensamiento se consideró la posibilidad de implementar una red neuronal como alternativa del modelo de regresión lineal, esto debido a que son más robustas que los modelos simples en cuestión de la normalidad con los datos originales. Por lo que se decidió crear una red neuronal de memoria a largo plazo con la cual se pudieron identificar patrones no lineales que otros modelos más tradicionales simplemente no pueden detectar, esto resultó ser un rotundo éxito para el modelo, pues como ya se demostró en los resultados, las predicciones que se hacían con la red neuronal para el Ozono resultaban sumamente fieles a los datos reales, a excepción de aquellos registros atípicos que se tenían de vez en cuando.

Se decidió emplear el análisis en el Ozono, pues de acuerdo con el socio formador el Ozono es precisamente un contaminante que se genera por medio de la combinación de factores meteorológicos con reacciones químicas de otros contaminantes, por lo que se consideró como el contaminante adecuado para contestar la pregunta de investigación. Para contestar la pregunta se tuvo que realizar el mismo modelo con los factores no antropogénicos, sumando factores antropogénicos como lo son la adición de los contaminantes NO, NO2 y NOX, por lo que al introducirlos en la red neuronal y entrenarla se esperaba que los resultados mejoraran considerablemente; sin embargo, al obtener el valor del error medio absoluto de cada uno de los modelos creados y al ser comparados a la par se observa que el modelo que mejor se ajusta a los datos reales es aquel que cuenta solamente con las variables de factores no antropogénicos. **Demostrando que se puede predecir significativamente el nivel de un contaminante solamente con factores no antropogénicos.** Dando validez y sentido lógico a los datos proporcionados por la Organización de las Naciones Unidas donde se menciona que el 99% de la población vivió en zonas donde no se cumplían con los estándares de nivel de contaminantes del aire establecidos por la Organización Mundial de la Salud.

Ahora bien, es evidente que el modelo presenta ciertas fallas notorias, como la incapacidad de predecir adecuadamente los picos extremos presentados en los datos reales, esto se puede deber a que las redes neuronales funcionan por medio del aprendizaje de patrones presentes en los datos, y se llegó a la conclusión de que esos picos extremos se debían a "errores" de medición por medio de los centros de SIMA, en donde se obtenían valores extremadamente altos de contaminantes debido a cuestiones que no presentan un peligro inminente a la calidad del aire, como lo puede ser una carne asada realizada cerca de un centro de medición, por lo que es evidente que la red neuronal no debería de ser capaz de notar estos "errores" en la medición de los contaminantes ya que como tal no son patrones. Determinando que el modelo obtenido realiza su función de la mejor manera posible dando más seguridad a la conclusión obtenida.

Anexos

https://drive.google.com/drive/folders/1n8xeJCoJLUbla0iFVKPEUHZIQIs49O5u?usp=sharing

Referencias

Brownlee, J. (2020). *Deep Learning Models for Multi-Out Regression.* https://machinelearningmastery.com/deep-learning-models-for-multi-output-regression/

Peterson, R. A., & Cavanaugh, J. E. (2019). Ordered quantile normalization: a semiparametric transformation built for the cross-validation era. *Journal of applied statistics*, *47*(13-15), 2312–2327. https://doi.org/10.1080/02664763.2019.1630372

Peterson, R. A. (2023). *Using the BestNormalize package*. https://cran.r-project.org/web/packages/bestNormalize/vignettes/bestNormalize.html#other-ot-included-transformations

Organización Mundial de la Salud (OMS). (2021a) Ambient (outdoor) air pollution. Fact sheet.https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health

World Health Organization: WHO. (2022, December 19). *Contaminación del aire ambiente.* https://www.who.int/es/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health