RETO: Análisis de la Calidad del Aire

MA2003B: Aplicación de métodos multivariados en ciencia de datos

Gpo. 201
Domingo 22 de ocutbre de 2023

Docentes:
Dra. Blanca Rosa Ruiz Hernández
Dra. Monica Guadalupe Elizondo Amaya
Tecnológico de Monterrey, Campus Monterrey

Axel Quiroga Caldera - **A00832676**
Andres Morones Navarro - **A00833287**
Valeria María Serna Salazar **A01284960**
David Alejandro Acuña Orozco - **A00571187**

Tecnológico de Monterrey

Índice

Resumen

El presente trabajo reconoce la importancia de tomar acción sobre la calidad del aire en Nuevo León, ya que se ha convertido en uno de los estados con peor calidad del aire. El cual se desarrollo de la mano con SIMA, se encontraron relaciones interesantes entre los comportamientos de algunos contaminantes; CO, NOx y el O3 es por ello por lo que se realizó una regresión logística para predecir la categoría de O3 y observar la influencia de ciertas variables para comprobar su efecto en O3. También, se tuvo siempre presente la idea de predecir, por ese motivo se realizaron series de tiempo y se encontró una función que modela la cantidad de O3 en promedio por año. Finalmente, se realizó un modelo prophet para predecir la cantidad de O3 por hora en cada estación, y utilizando streamlit se desplegaron las predicciones.

La calidad del aire es una preocupación creciente en muchas partes del mundo, y Nuevo León no es la excepción. Este estudio se adentra en el análisis detallado de la calidad del aire en el estado, poniendo especial énfasis en la problemática que representa la contaminación atmosférica y el efecto de diferentes componentes en la cantidad de O3 presente. La investigación se llevó a cabo utilizando avanzados métodos multivariados estadísticos, lo que permitió la comprensión de entre las relaciones de contaminantes y su efecto en la calidad del aire.

A lo largo del estudio, se exploraron las interacciones y relaciones entre diversos compuestos presentes en la atmósfera y varios factores climatológicos. Uno de los hallazgos más significativos fue la identificación de patrones cíclicos diarios en los niveles de ozono (O3). Este compuesto, esencial para la protección de la vida en la Tierra, mostró estar influenciado en gran medida por precursores como el monóxido de carbono (CO) y los óxidos de nitrógeno (NOx).

Dada la complejidad del fenómeno, el estudio propuso dos enfoques metodológicos distintos para predecir los niveles futuros de O3. El primero se basa en la identificación y modelado de patrones cíclicos, mientras que el segundo utiliza técnicas de regresión logística. Ambos enfoques ofrecen perspectivas valiosas y complementarias sobre la evolución esperada de la calidad del aire en Monterrey.

Un aspecto crucial de la investigación fue la implementación del modelo Prophet, una herramienta poderosa para la predicción de series temporales, y su inclusión en el estudio permitió una modelización más precisa y adaptable de los niveles de O3 y el impacto de la radiación solar en los mismos. Esta herramienta, desarrollada por Facebook, es especialmente útil para capturar patrones estacionales en datos con tendencias fuertes y múltiples estacionalidades.

El objetivo principal de esta investigación es proporcionar herramientas robustas y modelos predictivos que puedan ser utilizados por las autoridades y organismos pertinentes para mejorar el monitoreo y gestión de la calidad del aire en Nuevo León. Con el fin de lograr una mejor comprensión y gestión, este proyecto proporciona herramientas y modelos predictivos que pueden ser utilizados por las autoridades y organismos pertinentes para mejorar el monitoreo y gestión de la calidad del aire en Nuevo León. Con este aporte, se pueden tomar medidas efectivas para proteger la salud de los ciudadanos y preservar el medio ambiente.

I Introducción

La calidad del aire es fundamental para la salud y el medio ambiente, ya que la exposición constante a contaminantes como el ozono, dióxido de azufre y dióxido de nitrógeno pueden provocar graves enfermedades, incluyendo el cáncer y la muerte. Esta contaminación no solo afecta la salud y el bienestar de las personas, sino que también contribuye al cambio climático y afecta la biodiversidad [5].

Para evaluar la calidad del aire y detectar posibles riesgos para la salud y el medio ambiente, se utilizan redes de estaciones de monitoreo que recopilan datos en tiempo real sobre contaminantes atmosféricos. Estos datos son esenciales para comparar con normativas y regulaciones ambientales, y si los límites se superan, se toman medidas para reducir emisiones y mejorar la calidad del aire. El análisis implica diversas técnicas y parámetros, como la medición de partículas y gases con monitores y analizadores, y se consideran datos meteorológicos para entender la dispersión de contaminantes. Además, se emplean modelos de dispersión para prever la propagación de contaminantes desde fuentes conocidas y estimar concentraciones en áreas sin estaciones de monitoreo.

El Sistema Integral de Monitoreo Ambiental (SIMA) forma parte del Gobierno de Nuevo León en el sector de Medio Ambiente, inició su operación a partir del 20 de noviembre de 1992 con la finalidad de contar con información continua y fidedigna de los niveles de contaminación ambiental en el estado de Nuevo León. Como su deber es informar sobre la calidad del aire en el estado, se tienen que tomar en cuenta distintas normas para evaluar los niveles en el aire y, si los sobrepasan, emitir comunicados y notificar a la población de ello. El más importante siendo el Índice de Aire y Salud y algunas otras que se localizan en la parte A de los Anexos.

II Problemática y Justificación

En las últimas décadas, el aumento acelerado de la industrialización, urbanización y actividades humanas ha elevado, exponencialmente, las emisiones de contaminantes en la atmósfera, resultando así en una disminución significativa de la calidad del aire. Estos cambios no solo tienen efectos directos en la salud pública, manifestadas en enfermedades respiratorias, cardiovasculares y otros problemas de salud, sino que también tienen un impacto significativo en problemas ecológicos y de sostenibilidad del planeta.

Nuevo León al ser uno de los estados más industrializados de México, presenta estos desafíos frecuentemente; por lo tanto, es necesario tener sistemas efectivos de monitoreo que proporcionen datos precisos y en tiempo real sobre la calidad del aire. El adecuado análisis y respuesta a estos datos no sólo beneficia a la población actual, previniendo enfermedades y mejorando la calidad de vida, sino que también contribuye a la protección y preservación del medio ambiente para las futuras generaciones [3].

SIMA representa uno de esos esfuerzos tan importantes que hace el estado de Nuevo León para abordar estos retos. Aunado a esto y con base en todo lo anterior, nuestro trabajo se compromete a facilitar la comprensión y análisis de los datos en relación con la variable del Ozono, pues, de acuerdo con información disponible sobre la calidad del aire en el Área Metropolitana de Monterrey para el año 2018, el segundo contaminante que con mayor frecuencia determinó una condición de mala calidad del aire fue el ozono (O3). Además, el ozono en la atmósfera baja (troposférico) es un contaminante del aire y un componente del smog, lo que puede ser perjudicial para la salud humana. Si bien en la estratosfera protege de la radiación ultravioleta, a nivel del suelo es un gas de efecto invernadero y componente importante del cambio climático [8].

III Objetivo

Tomando en cuenta que la presencia del ozono en el ambiente es de las problemáticas más urgentes por abordar, el objetivo de este proyecto es hacer un modelo predictivo que tome en cuenta las horas en que el ozono es mayor o menor, para predecir el promedio de cada hora por año. Además, se busca también encontrar los contaminantes que más o menos le afectan para su generación y presencia en la atmósfera, para saber si hay precursores o antecesores que puedan sugerir que si el precursor tiene niveles altos, el ozono lo estará después. Por último, para dar una propuesta de valor a SIMA, se busca hacer una interfaz que permita al usuario seleccionar la estación y el rango de días a predecir, y que le regrese una predicción junto al promedio calculado y una gráfica que compare los valores reales a las predicciones (si las predicciones son a futuro, entonces el programa las compara con los valores del año anterior). De tal manera, para SIMA será de valor el facilitar la comprensión de los datos y tener un contexto más completo de la problemática que es el ozono, así como tomar medidas en función de las predicciones, es decir, si el ozono está alto, entonces alertar a la población, cambiar de hora los eventos que estaban estipulados, etc. Todo lo anterior, con el fin de salvaguardar la salud y bienestar de la población general, así como facilitar la interpretación y el análisis de los datos para el Socio Formador.

IV Preparación de los Datos

Se eligió la estación de San Nicolás (UANL) por que, de acuerdo al Estudio para el rediseño de la red de monitoreo de la calidad del aire de Monterrey (2021), la Av. Fidel Velázquez está a 72 metros al norte de la estación y tiene un aforo de 92,142 autos por día. Además, Gustavo Adolfo Bécquer está a 55 metros al este y tiene un aforo de 26,586 autos por día [7]. Considerando que el componente químico liberado por los automóviles es NOx y CO, que son precursores del O3, se eligió la estación esperando un comportamiento más notorio del componente en la atmósfera.

Además, tomando en cuenta que la Universidad frecuenta los eventos al aire libre, se podrán tomar decisiones informadas al respecto, para así modificarlos en función de la calidad del aire y de la presencia del ozono en horas específicas.

4.1 Limpieza de Datos

Utilizando Python para facilitar el proceso, primero se verificaron fechas y horas faltantes y se encontraron que habían 8 correspondientes a 2020, 1 a 2021 y 1 a 2023. Para tratarlos, se agregaron a la base de datos con valores nulos. Además, se revisó y no se encontró ningun valor duplicado.

Después, se calcularon los Z-scores para cada columna del conjunto de datos, agrupando los datos por día de la semana y hora (para ser "justos" y comparar cada día y hora, evitando así la comparación de datos que son diferentes). Los Z-scores se utilizan para medir la variabilidad de un punto de datos con respecto a la media y su desviación estándar y, en este caso, se consideró que un valor es atípico si el valor supera a 3 (significando que está tres desviaciones estándar lejos de la media). Se encontró que tenía 365 valores atípicos o el 2.5% del total de los datos. Estos valores también pasaron a ser nulos.

Los datos nulos fueron rellenados con los promedios de los componentes no nulos correspondientes a la misma hora y día de la semana, intentando así mantener la coherencia en los datos al rellenarlos con estimaciones basadas en el patrón de variación de las variables. Por último,

se juntaron ambas bases de datos en una que resultó tener 20 columnas y 31,790 observaciones, iniciando el 1 de enero de 2020 y terminando el 17 de agosto de 2023. La explicación a detalle con cada variable dentro de la base de datos se encuentra en la parte B de los anexos.

4.2 Exploración de los Datos

Una vez teniendo los datos limpios y siguiendo con Python, el primer paso fue hacer un análisis de correlación para comprender un poco más el contexto de las variables y la manera en que se relacionan entre sí. Es posible notar que las variables que tienen la mayor correlación con O3 son radiación solar, humedad relativa, temperatura, velocidad del viento y NOx.

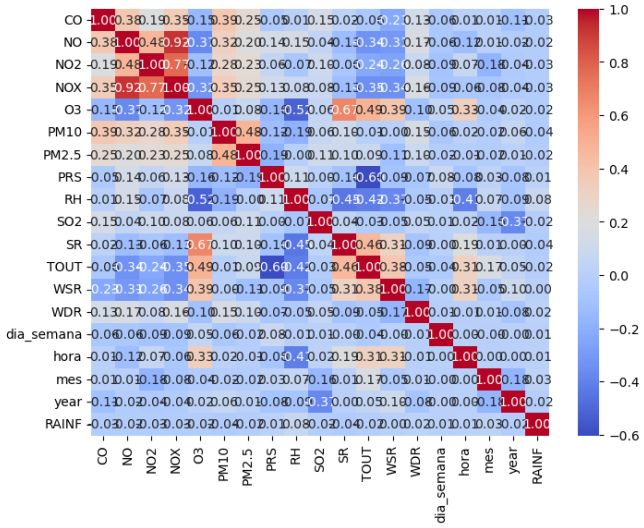

Figura 1: Correlación entre las variables

Además de los resultados de la matriz de correlación, se consideró el capítulo 13 del libro titulado *Atmósfera: una introducción a la meteorología*, que indica que los precursores del O3 son NOx y CO, mientras que uno de los factores más importantes al analizarlo es la hora [6], por lo que la exploración de los datos se hace considerando estas variables y al hacer box plots para ver el comportamiento de cada variable se obtiene que:

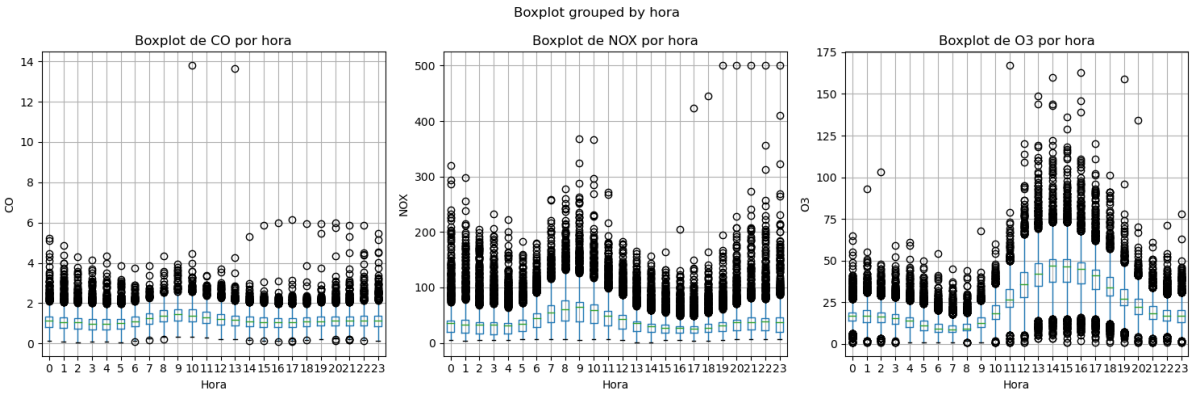

Figura 2: CO, NOX y O3 por hora

El incremento de CO y NOX sucede en la mañana (entre 6 am y 1 pm), mientras que el de O3 sucede entre la 1 pm y 7 pm, lo que muestra que el comportamiento de los primeros define el nivel del segundo.

Además, se explora haciendo una división por mes (3) en donde es posible ver que, aunque CO sigue un comportamiento bastante estable, NOx aumenta en los meses de frío y disminuye en los de calor, mientras que el O3 sigue el comportamiento contrario.

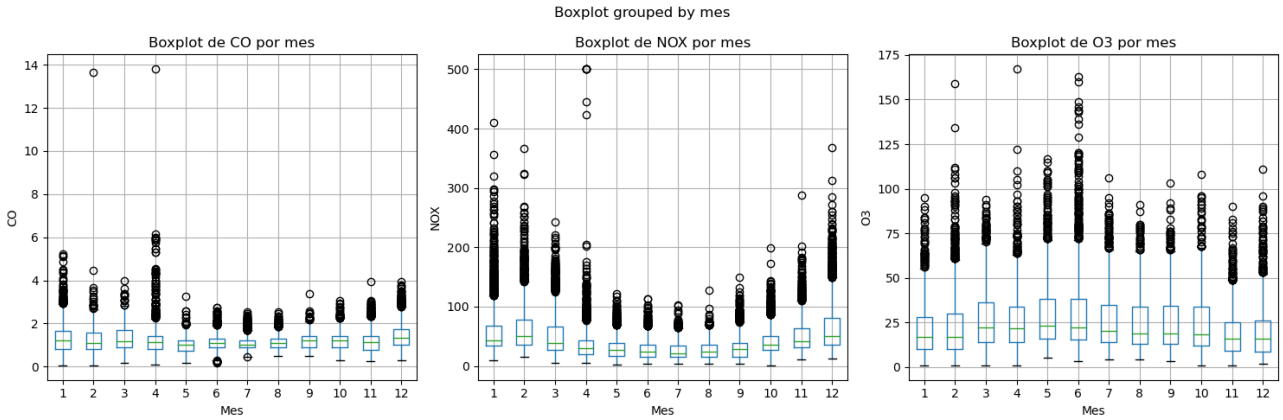

Figura 3: CO, NOX y O3 por mes

Finalmente, se hace un análisis por día de la semana (4), en donde es posible ver que todas las variables siguen un comportamiento bastante estable, solo O3 muestra un incremento ligero el domingo, mientras que NOx disminuye el domingo.

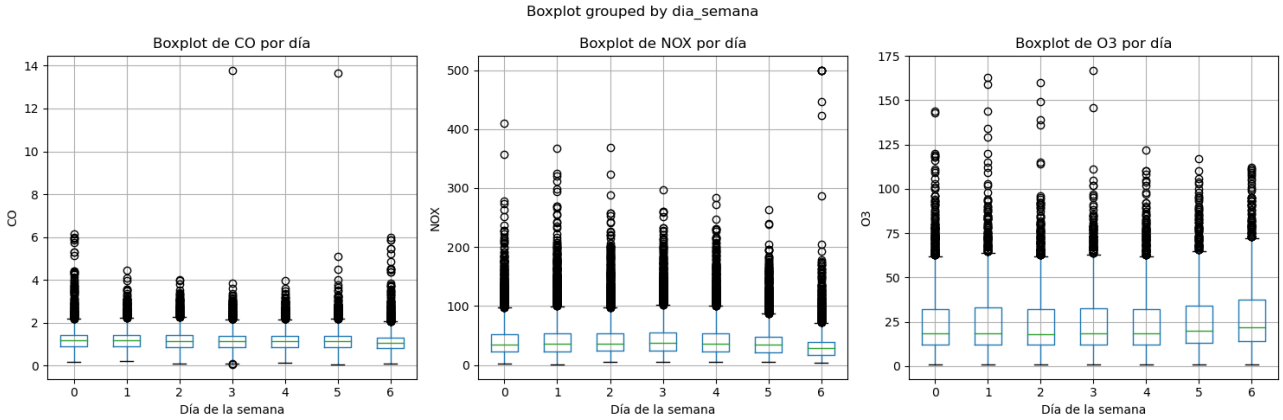

Figura 4: CO, NOX y O3 por día de la semana

Entonces, los boxplots demostraron que, aunque CO sigue un comportamiento poco variable, NOx sí se comporta de una manera contraria al O3, lo que indica que en efecto es su precursor. El ozono aumenta a las altas horas del día, que es cuando la radiación solar aumenta, y esto se debe a que la radiación solar tiene la suficiente energía para romper las moléculas de oxígeno (O2) en átomos de oxígeno (O), y estos se combinan con otras moléculas de oxígeno (O2) para formar moléculas de ozono (O3), mostrandose esencial para su aparición en la atmósfera. De hecho, la figura 5 colorea en rojo las horas de 12 pm a 6 pm (alta cantidad de radiación solar) y en azul las demás horas del día (baja cantidad de radiación solar) y es posible ver de una manera muy clara que los puntos más altos de O3 ocurren cuando hay una mayor radiación solar.

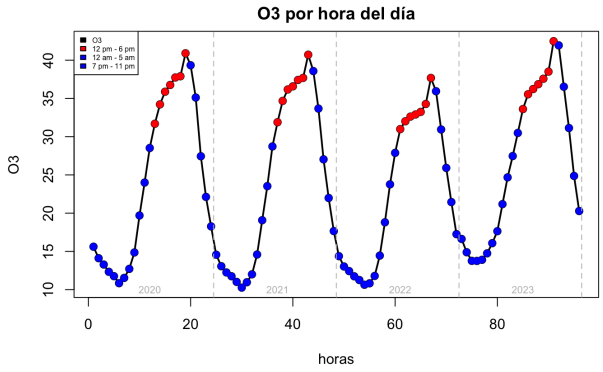

Figura 5: O3 por hora coloreada segun la radiación solar

Para hacer la comparación entre el comportamiento del O3 y sus precursores (CO y NOx) señalando a la radiación solar como principal punto de análisis, se compararon solamente las gráficas de 2022 y 2023, que es cuando la pandemia terminó y las vidas regresaron a la normalidad, significando un tráfico normal y un comportamiento ordinario en el día a día. Con esto, es posible ver en las figuras 6 y 7 que los puntos altos de O3 ocurren después de que CO y NOx los tienen, comprobando así la existencia de una correlación previa. Lo anterior, pues cuando la radiación solar empieza a crecer y, junto a ella, la cantidad de O3 en el ambiente crece también, los niveles de CO y NOx empiezan a disminuir.

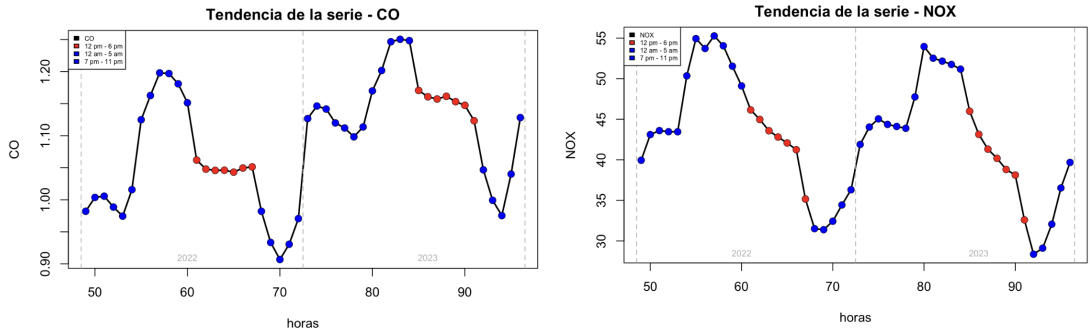

Figura 6: CO y NOX en 2022-2023 por hora coloreada segun la radiación solar

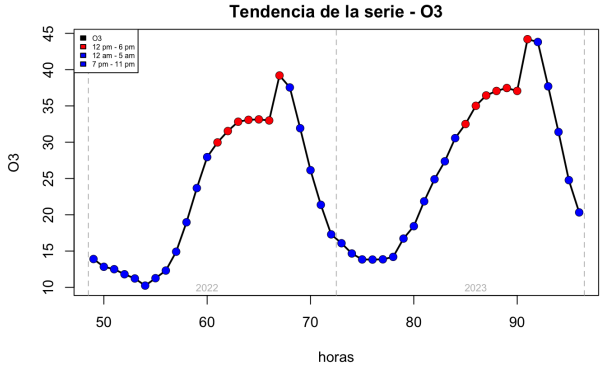

Figura 7: O3 en 2022-2023 por hora coloreado segun la radiación solar

Finalmente, se confirma que las variables necesarias de analizar para observar el comportamiento del O3 son sus precursores (CO y NOx), y los factores que lo forman (radiación solar y la hora).

V Modelación y Validación

5.1 Relaciones de Interdependencia - Serie de Tiempo

Se busca realizar una serie de tiempo para predecir los valores del O3 conforme la hora de la semana, pues resultó que la variable a predecir es estacionaria. Para esto se utilizó R, pues muestra un mayor poder al trabajar con este tipo de procesos. Al descomponer la serie de tiempo y analizar únicamente su tendencia, se descubrió que el comportamiento que sigue es similar a una gráfica senoidal o cosenoidal, por lo que se eligió la ecuación $sin(a * \frac{hora}{b}) + cos(a * \frac{hora}{b})$ para la regresión y se hizo un bucle encargado de encontrar los valores para a y b que minimizaran el error cuadrático medio (CME). La ecuación que mejor modeló el comportamiento de la variable a través de los años fue $sin(1.38 * horas/16.63) + cos(1.38 * horas/16.63)$, que mostró un EPAM del 21.2% y un CME de 6, lo que indica un comportamiento satisfactorio al tener un acierto de casi el 80%. En la figura 8 se muestra la predicción realizada para el 2024.

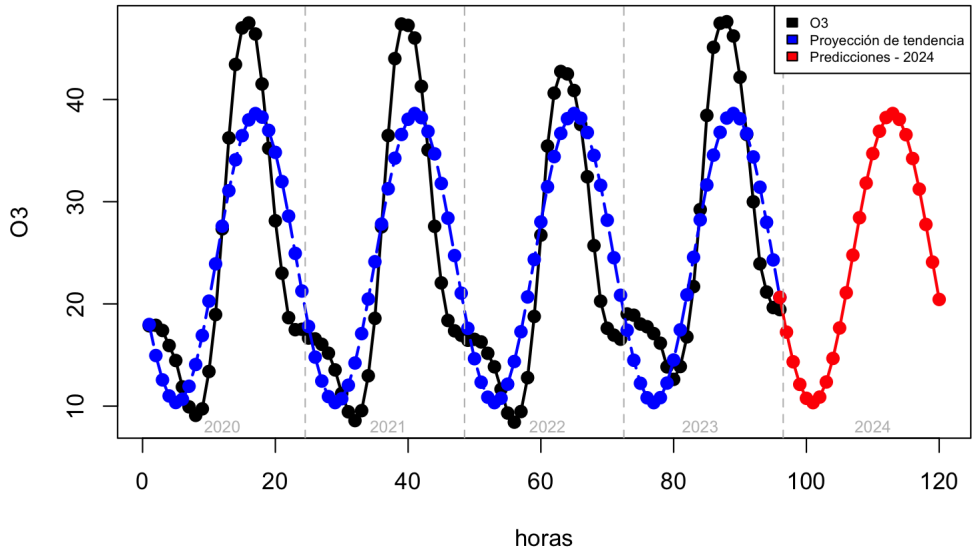

Figura 8: Predicción de O3 en 2024

5.2 Relaciones de Dependencia - Regresión Logística

Al observar que el O3 no seguía un comportamiento normal y por más transformaciones que se hicieron (Boxcox y Yeo Johnson) no se pudo normalizar, se optó por hacer una regresión logística en R para predecir el nivel de O3 (bajo o alto) dependiendo de los valores que toman los compuestos precursores (CO y NOx) y los factores que se comprobó que le afectan durante la exploración de los datos. Al realizarlo y revisar la significancia de los coeficientes, se encontró

que además de CO, NOx, SR (radiación solar) y la hora, RH (humedad relativa) también se mostró significativo, lo cual tiene sentido, pues en la matriz de correlación mostró una correlación relativamente alta con el O3. De hecho, la ecuación que resultó al hacer el modelo fue:

$$2.06 - 0.032 \cdot NOX - 0.91 \cdot CO + 9.016 \cdot SR + 0.081 \cdot hora - 0.037 \cdot RH \tag{1}$$

Es posible ver que, aunque todas son significativas para el modelo (valor de p muy cercano a 0), la que tiene un mayor peso para la clasificación del valor de O3 (alto o bajo) es la radiación solar, lo cual es interesante porque es más de 9 veces mayor que las demás, aunque tiene sentido pues durante la exploración de los datos se había encontrado que la radiación solar era un factor importante para el O3.

De 100 observaciones, el modelo clasifica correctamente 84.28, con una sensibilidad (la capacidad del modelo para identificar correctamente los casos positivos entre todos los casos reales positivos) del 92.19% y una especificidad (la capacidad del modelo para identificar correctamente los casos negativos entre todos los casos reales negativos) de 76.56%. La matriz de confusión que demuestra el buen trabajo de clasificación que resulta del modelo se muestra en la figura 9.

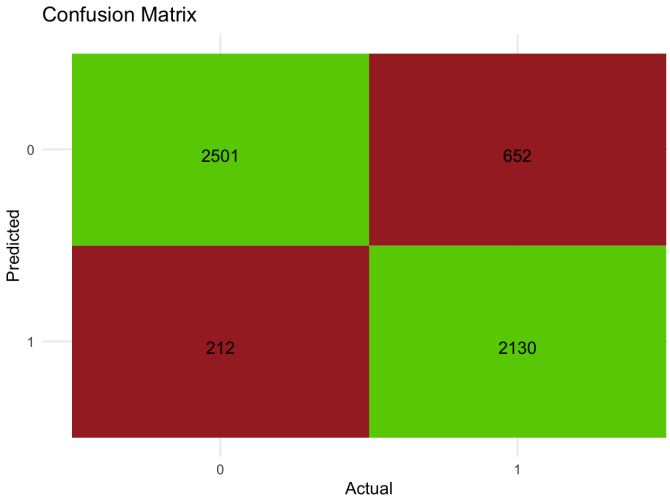

Figura 9: Matriz de confusión - Regresión logística

Cabe aclarar que, en cuanto a los supuestos del modelo, no se cumple el de independencia pero sí se cumple el de colinealidad, significando que las variables están correlacionadas entre sí. Aunque estos resultados podrían explicar la razón por la que la especificidad no es tan alta, el modelo hace un buen trabajo de clasificación y sí se recomienda su uso. La validación de supuestos más a fondo está en la parte C de los anexos.

5.3 Propuesta de Valor (Interfaz)

El proyecto tenía como objetivo generar una propuesta de valor, no solamente estar orientado al análisis estadístico, sino contribuir a SIMA con una herramienta sencilla para generar predicciones de O3 en un determinado período de tiempo.

Para lograr el objetivo, se utilizó el modelo Prophet de Facebook, un modelo que mezcla SARIMA e inteligencia artificial para detectar tendencia, estacionalidad y días festivos y poder generar predicciones acertadas [2]. Es un modelo aditivo que tiene la siguiente fórmula en función

del tiempo:

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t \tag{2}$$

Donde $y(t)$ es la predicción, $g(t)$ es el componente tendencia, $s(t)$ es la estacionalidad, $h(t)$ es el componente de eventos que impactan la serie (vacaciones, días festivos, asuetos) y ϵ_t es el error irreducible.

Usando Python y tras familiarizarnos con el modelo, se realizó una prueba inicial para poder medir la calidad de predicciones del modelo Prophet, para predecir la cantidad de Ozono por hora en un rango de días dado. La prueba está en la parte D de los Anexos y consistió en utilizar los datos de la estación Norte2 y esconder los datos de los últimos 8 días antes de entrenar al modelo, después del entrenamiento, se realizó la predicción correspondiente a los 8 días escondidos para comparar el resultado actual contra el predicho. En la figura 10, se ven los resultados de la prueba, se puede apreciar que el modelo es capaz de detectar las horas pico de Ozono en el lapso de una semana. Posterior a la prueba se calculó el MSE y el EPAM, se obtuvo un $MSE = 109.74$ y $EPAM = 0.35$, con esos resultados se decidió que Prophet es una buena opción para predecir la cantidad de Ozono por hora.

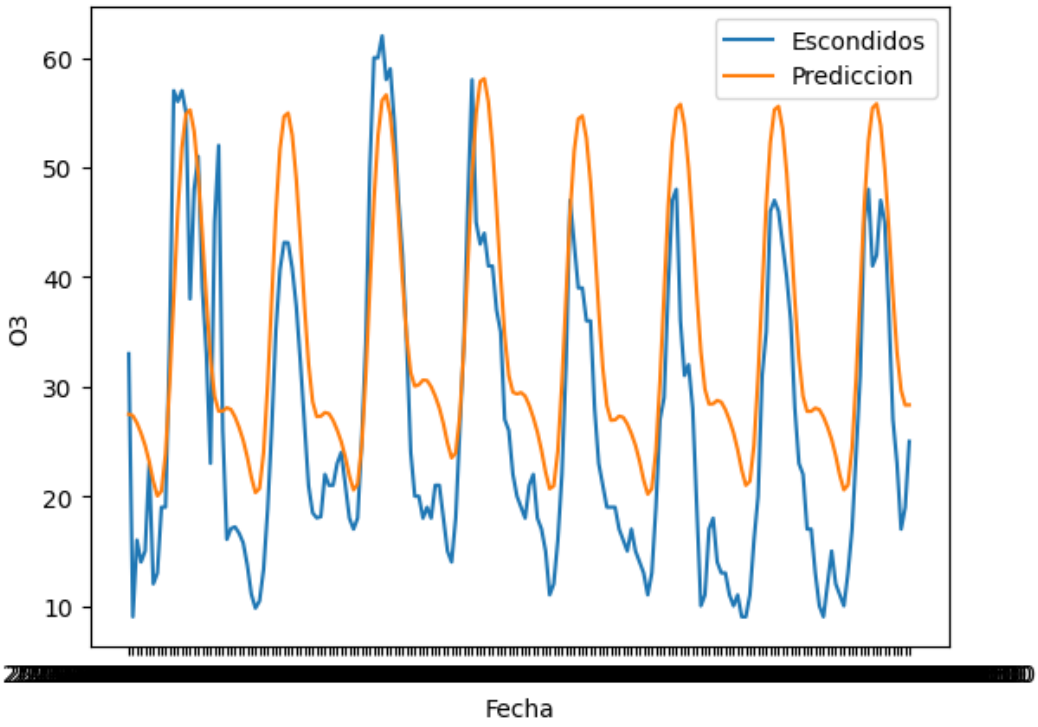

Figura 10: Comparación datos reales contra modelo Prophet

Después, se realizaron los modelos correspondientes a cada estación, para predecir por estación la cantidad de Ozono. Para el despliegue de la información, se realizó un interfaz con Streamlit. En la interfaz se puede escoger la estación y el rango de fechas por predecir. El resultado es un conjunto de gráficas fáciles de interpretar, como:

- **Visualización de predicción:** Donde se puede ver la predicción de Ozono por hora en el rango de fechas escogido.

- **Comparación datos históricos vs. Predicción:** Esta gráfica consiste en visualizar los datos predichos contra los datos del mismo número de semana y mismo día de la semana para observar un comportamiento similar y comparar fechas adecuadamente.
- **Visualización históricos:** Es para visualizar como se ha comportado históricamente el Ozono en la estación escogida.
- **Visualización históricos y modelo:** Consiste en la comparación del Ozono histórico contra el modelo prophet de la estación correspondiente.

En la parte inferior se tiene la sección comportamiento y tendencia que resume la cantidad de Ozono en promedio en la base de datos y como se comportará el ozono en los días predichos. También esta el valor del EPAM de ese modelo en la estación escogida.

Las últimas dos secciones consiste en visualizar los datos de la predicción y un botón para poder descargarlos. El vídeo con el funcionamiento y el código para desplegar la página está en la parte E de los Anexos.

VI Resultados

Al final, se tienen 3 entregables finales que cumplen con lo propuesto inicialmente:

1. **Serie de Tiempo para predecir O3 por hora:** Hecho en R, ayudaría a saber exactamente que horas son las más peligrosas y, considerando que está ubicada en la universidad, ayudaría a planear eventos al aire libre y salvaguardar la salud de todos. Con un 80% de acierto, brinda una seguridad de que los resultados son confiables.
2. **Regresión Logística para predecir O3 dependiendo de precursores y factores climatológicos:** Hecho en R, si en lugar querer saber la predicción del promedio por hora en el año se busca hacer una predicción más actual, la regresión logística recibe la cantidad de radiación solar, humedad relativa, hora, cantidad de Co y NOx en el día a predecir y, con eso, determina si los niveles de O3 serán altos o bajos. Este modelo tiene un accuracy del 84%, lo cual es bastante positivo.
3. **Interfaz para ver las predicciones de O3 por hora en cualquier estación y cualquier fecha:** Hecho en Python y al utilizar una librería de Meta, las predicciones consideran días festivos, vacaciones, factores climatológicos, etc. Lo que brinda una predicción más certera sobre cada estación. Tambien, el usuario puede elegir el rango de días a predecir, lo que puede servir para planear eventos o, si se sabe que a cierta hora estará más alto de lo normal, brindar advertencias y alertar a la población.

VII Discusión y Conclusiones

Durante el desarrollo del proyecto, se llevaron a cabo análisis detallados sobre los datos, centrándonos específicamente en el ozono (O3). La investigación previa a la creación del modelo reveló que los principales precursores del O3 incluyen al monóxido de carbono (CO) y los óxidos de nitrógeno (NOx), junto con factores climatológicos como la radiación solar y la humedad relativa; descubrimientos que se mostraron esenciales para el desarrollo de modelos predictivos que

mostraron una precisión bastante buena. Una pregunta que surge del análisis es ¿qué propiedades químicas hacen que CO y NOx causen O3?

Se idearon dos enfoques principales para predecir los niveles de O3 en la atmósfera; el primero se basa en los patrones cíclicos diarios (al dividir la base de datos en promedios de cada hora por año) y este modelo específico proporcionó una visión general de las tendencias que sigue la variable al largo y corto plazo. De tal manera, se encontró un comportamiento cíclico similar a funciones senoidales o cosenoidales. Al final, la ecuación utilizada en la regresión y que mejor simulaba el comportamiento de la variable al paso de las horas fue $sin(1.38 * horas/16.63) + cos(1.38 * horas/16.63)$, que mostró un EPAM del 21.2% y un CME de 6, lo que indica un comportamiento satisfactorio al tener un acierto de casi el 80%. Este modelo se utilizó para predecir el valor promedio del ozono en cada hora para el año 2024.

El segundo enfoque fue hacer una regresión logística al integrar a los precursores y condiciones climáticas que logró ofrecer predicciones específicas y actuales. Logró tener un accuracy del 84% al definir si el ozono sería alto o bajo dependiendo de el CO, NOx, hora del día, humedad relativa (HR) y radiación solar (SR).

Ambos modelos no solamente son cruciales para la planificación y gestión de eventos al aire libre, sino también para salvaguardar la salud pública y cuidar a la población neolonesa de una exposición a altos niveles de ozono.

Además de los modelos, se creó una interfaz interactiva que permite al usuario seleccionar una estación de monitoreo y un rango de fechas para lograr obtener pronósticos detallados de los niveles de O3. Esta herramienta, además de ser valiosa para las autoridades y personas que toman decisiones, empodera al público en general para tomar decisiones informadas sobre la contaminación del aire y tomar medidas preventivas cuando sea necesario. Sería interesante ver qué ajustes se le pueden hacer para mejorar predicciones, pero actualmente tiene un comportamiento bastante bueno.

Este proyecto no es solamente un logro técnico, sino también una lección sobre la importancia de la estadística en la vida cotidiana. La habilidad para analizar datos complejos y descubrir patrones significativos entre los datos es esencial para comprender los problemas que se presentan y tomar medidas efectivas para resolverlos. Desde decisiones gubernamentales hasta precauciones individuales, la estadística juega un papel fundamental en la protección y cuidado del medio ambiente y la salud humana.

Por último, es crucial destacar el papel fundamental del trabajo en equipo en el éxito de este proyecto, pues la colaboración fue efectiva entre los integrantes del equipo. De tal manera, se logran aprovechar las fortalezas individuales de cada miembro, logrando así una diversidad de perspectivas en el equipo y, a su vez, facilitando la resolución de desafíos al enriquecer la calidad de análisis y modelos desarrollados.

VIII Referencias Bibliográficas

[1] Calidad del aire en nuevo león. `http://aire.nl.gob.mx/map_calidad.html`.

[2] Marina Alonso-Cortamp;eacute;s amp; Victoria Arribas. Análisis y predicción de series temporales con fb prophet python. `https://www.modeldifferently.com/2022/04/analisis_prediccion_ts_prophet/`, Apr 2022.

[3] Alejandro Del Toro. Nuevo león tiene un 2021 más contaminado. https://abcnoticias.mx/local/2021/10/22/nuevo-leon-tiene-un-2021-mas-contaminado-149405.html, Oct 2021.

[4] Documentation. Using r for time series analysis - time series. https://a-little-book-of-r-for-time-series.readthedocs.io/en/latest/src/timeseries.html, 2017.

[5] EPA. Ground-level ozone basics. https://www.epa.gov/ground-level-ozone-pollution/ground-level-ozone-basics, 2019.

[6] Frederick Lutgens and Edward Tarbuck. *13: Air Pollution*. Pearson Education, 2020.

[7] SIMA. Estudio para el rediseño de la red de monitoreo de la calidad del aire de monterrey (2021)., 2021.

[8] Bellio Vocci. How does ozone act? how and why can we avoid ozone toxicity? *Ozone*, page 19–28, 2016.

Anexos

A Normas a Considerar

Además de las mencionadas, otras normas que se consideran son:

Contaminante		Norma	Concentración	Tiempo de Exposición (horas)
Monóxido de Carbono	CO	NOM-021-SSAI-2021	26.0 ppm	1
Monóxido de Carbono	CO	NOM-021-SSAI-2021	9.0 ppm	8
Bióxido de Azufre	SO2	NOM-022-SSAI-2019	0.075 ppm	1
Bióxido de Azufre	SO2	NOM-022-SSAI-2019	0.04 ppm	24
Ozono	O3	NOM-020-SSAI-2021	0.090 ppm	1
Ozono	O3	NOM-020-SSAI-2021	0.065 ppm	8
Bióxido de Nitrógeno	NO2	NOM-023-SSAI-2021	0.106 ppm	1
Bióxido de Nitrógeno	NO2	NOM-023-SSAI-2021	0.021 ppm	Promedio Anual
Partículas Menores a 10 Micras	PM 10	NOM-025-SSAI-2021	70 µg/m3	24
Partículas Menores a 10 Micras	PM 10	NOM-025-SSAI-2021	36 µg/m3	Promedio Anual
Partículas Menores a 2.5 Micras	PM 2.5	NOM-025-SSAI-2021	41 µg/m3	24
Partículas Menores a 2.5 Micras	PM 2.5	NOM-025-SSAI-2021	10 µg/m3	Promedio Anual

Figura 11: Norma

CONTAMINANTE	NORMA MEXICANA	DESCRIPCIÓN
ÍNDICE	NOM-172-SEMARNAT-2019	Lineamientos para la obtención y comunicación del Índice de Calidad del Aire y Riesgos a la Salud
	NOM-156-SEMARNAT-2016	Establecimiento y operación de sistemas de monitoreo de la calidad del aire

Figura 12: Norma

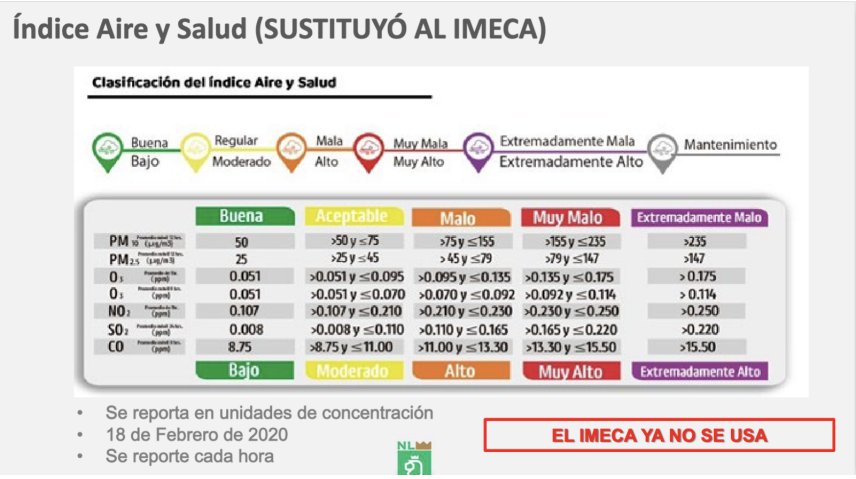

Figura 13: Norma

B Variables en la base de datos

- CO (monóxido de carbono): gas tóxico producido por procesos de contaminación, variable numérica, rango de valores: ($-0.13 \geq CO \leq 7.46$), número de valores nulos: 9785, promedio: 1.3318, moda:1.41, Desviación estándar: 0.6631.
- NO (óxido de nitrógeno): gas tóxico producido por procesos de contaminación, variable numérica, rango de valores: ($0.03 \geq NO \leq 945.1$), número de valores nulos: 9538, promedio: 10.8251, moda: 2.7, Desviación estándar: 19.6056.
- NO2 (dióxido de nitrógeno): gas tóxico producido por procesos de contaminación, variable numérica, rango de valores: ($0 \geq NO2 \leq 167.8$), número de valores nulos: 7507, promedio: 14.9077, moda: 5.9, Desviación estándar: 11.5823.
- NOX (familia de gases óxidos de nitrógeno): familia de gases tóxicos producidos por procesos de contaminación, variable numérica, rango de valores: ($0.5 \geq NOX \leq 971.8$), número de valores nulos: 7476, promedio: 25.5791, moda: 9.3, Desviación estándar: 27.1377.
- O3 (ozono): gas tóxico producido por procesos de contaminación, variable numérica, rango de valores: ($0.7 \geq O3 \leq 171$), número de valores nulos: 8730, promedio: 26.2334, moda: 17, Desviación estándar: 17.8983.
- PM10 (partículas gruesas): Las partículas gruesas PM10 son partículas suspendidas en el aire con un diámetro aerodinámico menor o igual a 10 micrómetros, que pueden representar un riesgo para la salud humana al ser inhaladas y pueden contener diversas sustancias perjudiciales. Variable numérica, rango de valores: ($2 \geq PM10 \leq 1001$), número de valores nulos: 8020, promedio: 61.5738, moda: 39, Desviación estándar: 43.4884.
- PM2.5 (Partículas pequeñas): Partículas muy pequeñas en el aire que tiene un diámetro de 2.5 micrómetros (aproximadamente 1 diezmilésimo de pulgada) o menos de diámetro, representan un riesgo para la salud. Variable numérica, rango de valores: ($0 \geq PM2.5 \leq 442$), número de valores nulos: 44952, promedio: 20.8807, moda: 12, Desviación estándar: 15.5509.
- PRS (partículas respirables en suspensión): Son partículas finas y pequeñas presentes en el aire que pueden ser inhaladas y representar un riesgo para la salud humana. Variable numérica, rango de valores: ($9.6729 \geq PRS \leq 747.6$), número de valores nulos: 5683, promedio: 716.4944, moda: 713.1, Desviación estándar: 9.6729.
- RAINF (Lluvia): Cantidad de lluvia que se presenta en cada estación. Variable numérica, rango de valores: ($0 \geq RAINF \leq 67.8$), número de valores nulos: 3747, promedio: 0.0048, moda: 0, Desviación estándar: 0.2338.
- RH (humedad relativa): Nocivo para la salud debido a su potencial de liberar compuestos tóxicos y al riesgo de inhalación de partículas finas. Variable numérica, rango de valores: ($0 \geq RAINF \leq 714.2$), número de valores nulos: 14810, promedio: 54.6909, moda: 72, Desviación estándar: 37.1079.
- SO2 (dióxido de azufre): Gas tóxico producido por procesos de contaminación. Variable numérica, rango de valores: ($0.5 \geq SO2 \leq 222.9$), número de valores nulos: 15301, promedio: 5.0052, moda: 3.3, Desviación estándar: 4.9018.

- SR (radiación solar): Representa la radiación solar en la hora específica. Variable numérica, rango de valores: ($-0.0360 \geq SR \leq 72.6265$), número de valores nulos: 7355, promedio: -0.36604, moda: 0, Desviación estándar: 72.6265.
- TOUT (Temperatura): Indica la temperatura que corresponde a la hora en que fue tomada la muestra. Variable numérica, rango de valores: ($-9999 \geq SR \leq 112.39$), número de valores nulos: 4742, promedio: 23.4194, moda: 25.3, Desviación estándar: 22.3003.
- WSR (Velocidad del viento): Indica la velocidad del viento que se presentó, en promedio, en esa hora. Variable numérica, rango de valores: ($0.1 \geq SR \leq 188.8$), número de valores nulos: 8042, promedio: 8.7938, moda: 1.4, Desviación estándar: 5.6207.
- WDR (Dirección del viento): Indica la velocidad del viento que se presentó, en promedio, en esa hora. Variable numérica, rango de valores: ($-9999 \geq SR \leq 360$), número de valores nulos: 9741, promedio: 130.6449, moda: 105, Desviación estándar: 98.94883.

C Supuestos de la Regresión Logística

3.1 ANOVA

Se comparó el modelo nulo (la variable a predecir solamente consigo misma) y el modelo con las demás variables en un ANOVA y se encontró que los residuales eran menores con el segundo modelo, tambien, el valor p resultó ser prácticamente 0 y menor a alfa = 0.05, indicando que las variables predictoras son estadísticamente significativas, pues, a comparación del modelo 1, se adapta mejor a los datos.

3.2 Independencia

H_0: Los residuales son independientes.
H_1: Los residuales no son independientes.

Se hizo un test de Durbin-Watson y se obtuvo un valor de p de 0.797, que es mayor a alfa=0.05, lo que indica que hay evidencia estadística para rechazar la hipótesis nula, significando que los residuales no son independientes.

3.3 Multicolinealidad

Se calcularon los eigenvalores, el factor de inflación de la varianza (VIF) y los valores de tolerancia (1/VIF), y resultó que para el orden de NOX, CO, SR, hora y RH:

- **Eigenvalores:** 1.7454243, 1.2697800, 0.8288919, 0.6844486, 0.4714552
- **Valores VIF**: 1.105137, 1.109127, 1.188364, 1.120077, 1.135638
- **Valores de Tolerancia:** 0.9048649, 0.9016097, 0.8414928, 0.8927958, 0.8805622

Las variables pasan la prueba de multicolinealidad porque todos los valores propios son mayores a cero, los VIF son cercanos a 1 y los valores de tolerancia son mayores que 0.1.

D Prueba de Prophet

Notebook con prueba inicial prophet

E Despliegue de Prophet

Folder con todo lo relacionado a la interfaz
Video de su funcionamiento y despliegue de la interfaz

F Códigos utilizados para la realización de todo el proyecto

Códigos de Python y R