Instituto Tecnológico y de Estudios Superiores de Monterrey

Aplicación de Métodos Multivariados en Ciencia de Datos
MA2003B, Grupo 101

Etapa 4. Evidencia Final del Reto

Equipo 4:
Juan Pablo Sada San José - A01722098
Sarah Dorado Romo - A01540946
Kevin Antonio González Díaz - A01338316
Adrian Pineda Sánchez - A00834710

Blanca Rosa Ruiz Hernández
Mónica Guadalupe Elizondo Amaya

Monterrey, Nuevo León, México. 9 de septiembre 2023

Índice

1. Resumen

Este proyecto se divide en etapas clave y cronológicas. En la primera etapa, nos dedicaremos a comprender la función del socio formador SIMA y cuales son sus objetivos además de su función con la medición de la calidad del aire en Nuevo León. Lo siguiente es la comprensión de los datos proporcionados, realizando la limpieza correspondiente según el tipo de análisis requerido y las preguntas que buscamos responder. Esto asegurando que nuestro análisis sea lo más preciso y enfocado posible. A continuación con los datos preparados, se realizan diversos análisis estadísticos, incluyendo el análisis factorial, el modelo de regresión y el modelo predictivo. Seguidamente, validamos estos modelos para confirmar su utilidad y se presentarán los resultados obtenidos. Con esto observar los contaminantes más importantes en las épocas de verano y cuales causan mala calidad del aire en el estado de Nuevo León. Finalmente, discutiremos los resultados obtenidos que nos llevan a las conclusiones significativas basadas en nuestros hallazgos.

2. Conociendo el Negocio

2.1. Introducción

El Sistema Integral de Monitoreo Ambiental (SIMA) en Nuevo León toma un rol esencial en la gestión ambiental del estado al ser un componente central del gobierno dedicado a la supervisión y evaluación continua de los niveles de contaminantes presentes en el entorno en cierto tiempo. Este sistema refleja el compromiso del gobierno con la protección y mejora de la calidad del aire y del ambiente en general, al proporcionar datos concretos y confiables que son fundamentales para la toma de decisiones informadas en materia de políticas públicas y regulaciones ambientales.

Como parte de su labor, el SIMA emplea una red estratégicamente ubicada de estaciones de monitoreo ambiental en todo el estado de Nuevo León. Estas estaciones, equipadas con sensores y dispositivos de alta precisión, capturan una amplia gama de datos relacionados con contaminantes atmosféricos clave, como partículas suspendidas, gases nocivos y otros elementos perjudiciales para la salud humana y el entorno natural.

Un aspecto distintivo que mencionó el personal del SIMA es su enfoque en la recopilación continua y en tiempo real de información. Gracias a la programación con intervalos regulares de una hora, el sistema logra capturar cambios y fluctuaciones en los niveles de contaminantes a lo largo del día, lo que permite obtener una representación más completa y precisa de la calidad del aire en diferentes momentos y ubicaciones. Esta capacidad de monitoreo constante es crucial para detectar patrones y tendencias en la contaminación, identificar fuentes específicas de emisión y evaluar la eficacia de las medidas de mitigación implementadas.

La filosofía de trabajo del SIMA se basa en la premisa fundamental de que la toma de decisiones informadas en políticas públicas relacionadas con la calidad del aire debe ser respaldada por datos objetivos y confiables. Esta filosofía se apoya en la idea de que los datos duros y verificables sobre los contaminantes atmosféricos son esenciales para comprender el estado actual de la calidad del aire y para diseñar estrategias efectivas que mitiguen los riesgos asociados a la contaminación.

Para lograr este objetivo, el SIMA se dedica a la recolección sistemática y constante de

información relacionada con la calidad del aire en diferentes ubicaciones y momentos. Este enfoque implica la instalación de una red de estaciones de monitoreo estratégicamente ubicadas en áreas urbanas y suburbanas, así como en zonas industriales y rurales. Estas estaciones están equipadas con instrumentos de alta precisión que pueden medir diversos contaminantes atmosféricos, como partículas en suspensión (PM2.5 y PM10), dióxido de nitrógeno (NO2), ozono (O3), dióxido de azufre (SO2), monóxido de carbono (CO), entre otros.

La recolección de datos en diferentes momentos y en múltiples ubicaciones permite al SIMA crear una imagen completa y detallada de cómo los niveles de contaminantes varían a lo largo del tiempo y en diferentes áreas geográficas. Esta información es esencial para identificar patrones de contaminación, determinar áreas críticas con niveles alarmantes de contaminantes y evaluar el impacto de diversas fuentes de emisión.

La filosofía del SIMA también subraya la importancia de comunicar los resultados de monitoreo de manera transparente y accesible al público en general, así como a los responsables de la toma de decisiones.

Para comunicar la calidad del aire de manera comprensible para el público, se suele utilizar un índice de calidad del aire (ICA). Este índice combina varios contaminantes en una escala numérica y cromática (por ejemplo, de verde a rojo) para indicar la calidad del aire, donde el verde representa el aire limpio y el rojo el aire altamente contaminado.[15]

El Índice de Calidad del Aire (ICA) emplea una escala que abarca desde 0 hasta 500. A medida que el número aumenta, la polución atmosférica se intensifica y se torna más nociva para la salud. [15] Esta escala se segmenta en seis rangos, cada uno identificado por un color, un nivel de impacto en la salud y una evaluación de la calidad del aire:

- Verde: Buena para la salud (ICA de 0 a 50)
- Amarillo: Moderada (ICA de 51 a 100)
- Naranja: Dañina a la salud para grupos sensibles (ICA de 101 a 150
- Rojo: Dañina a la salud (ICA 151 a 200)
- Morado: Muy dañina a la salud (ICA 201 a 300)
- Marrón: Peligrosa (ICA superior a 300)

Y sin duda una de las variables que más ayudan en este paradigma de la mejora de la calidad del aire son por supuesto: Reducir las emisiones de contaminantes primarios como dióxido de azufre (SO2), dióxido de nitrógeno (NO2), partículas (PM2.5 y PM10), el Ozono (O3) y compuestos orgánicos volátiles (COVs) provenientes de fuentes industriales, vehículos, construcción y otros procesos. [15]

2.2. Problemática y Justificación

Podemos dimensionar la importancia para la ciudadanía del análisis de la calidad del aire a través de realizar una descripción de las consecuencias en términos de salud en torno a una calidad de aire en estado de alerta, por medio de la presencia de gases tóxicos en proporciones

nocivas para la salud humana.

Según la Organización Panamericana de la Salud (OPS) menciona que la presencia de una mala calidad de aire puede crear un aumento en el riesgo de ciertas enfermedades e infecciones respiratorias, enfermedades cardíacas, accidentes cerebrovasculares y cáncer de pulmón, entre otras; y su gravedad aumenta con respecto del nivel y tiempo de exposición que se tenga.[1]

Además, que personas que ya cuentan con una condición médica preestablecida o enfermedad relaciona con el sistema cardiovascular, cerebral o respiratorio predisponen una mayor afección y se encuentran más susceptibles a padecer consecuencia de salud aún más severa. Donde los grupos más vulnerables se encuentran conformados por: niños, los ancianos y los pobres.[1]

Uno de los aspectos fundamentales que se consideran en torno al análisis de la calidad del aire, según la Organización Mundial de la Salud (OMS) son las concentraciones de gases nocivos mayormente causados por: aparatos domésticos de combustión, los vehículos de motor, las instalaciones industriales y sus procesos y construcciones, así como los incendios forestales. [12]

Dentro de los gases que la misma (OMS) considera más preocupantes a la hora de hablar de la salud pública en torno a la contaminación del aire son: las partículas en suspensión (PM2.5 y PM10), el monóxido de carbono, el ozono, el dióxido de nitrógeno y el dióxido de azufre [12], debido a las cuestiones de salud explicadas anteriormente por medio de la (OPS) además que se señala que un 99 % de la población mundial respira un aire que supera los límites aceptados por la (OMS). [12]

Estos son los límites de cada uno de las concentraciones que emite la (OMS) como recomendación:

Contaminante	Tiempo promedio	Meta intermedia				Nivel de las directrices sobre la calidad del aire
		1	2	3	4	
$MP_{2,5}$, µg/m³	Anual	35	25	15	10	5
	24 horas[a]	75	50	37,5	25	15
MP_{10}, µg/m³	Anual	70	50	30	20	15
	24 horas[a]	150	100	75	50	45
O_3, µg/m³	Temporada alta[b]	100	70	–	–	60
	8 horas[a]	160	120	–	–	100
NO_2, µg/m³	Anual	40	30	20	–	10
	24 horas[a]	120	50	–	–	25
SO_2, µg/m³	24 horas[a]	125	50	–	–	40
CO, mg/m³	24 horas[a]	7	–	–	–	4

[a] Percentil 99 (es decir, 3-4 días de superación por año).
[b] Promedio de las concentraciones máximas diarias de O_3 (medias octohorarias) en los seis meses consecutivos con la concentración media móvil de O_3 más alta.

Figura 1: Niveles recomendados de las directrices sobre la calidad del aire y metas intermedias por parte de la OMS. [14]

A continuación se presentan algunos de los daños y consecuencias que estos gases contaminantes provocan en concentraciones nocivas para la salud humana:

1. Monóxido de carbono (CO). El CO, liberado comúnmente por actividades industriales y vehículos, en concentraciones elevadas pueden ser letales [2], afectando la absorción

de oxígeno en los pulmones y causando asfixia. Según la OMS [3], el CO2 es perjudicial para la salud si su concentración en el aire es mayor a 10 mg/m^3 durante 8 horas o 40 mg/m^3 durante 1 hora.

2. Dióxido de Nitrógeno (NO2). El NO2 es de especial importancia ya que es precursor de oxidantes fotoquímicos que impactan directamente en la salud humana' [6]. Desde una perspectiva biológica, la exposición a niveles altos de NO2 se vincula con efectos perjudiciales en el sistema respiratorio [4]. Estos efectos incluyen la contracción de los músculos en las vías respiratorias, lo cual puede causar molestias y dificultades para respirar. Además, el NO2 puede provocar irritación y enrojecimiento en los tejidos de la garganta y las vías respiratorias superiores. El nivel aceptable de este gas es menos de 100 μg/m^3 más de un año [5].

3. Dióxido de azufre (SO2). Este se emite al aire desde fuentes como generadores de energía y actividades mineras que involucran petróleo o carbón con ácidos sulfúricos [2]. Si la cantidad máxima de ácido sulfúrico en el aire supera los límites de 80 μg/m^3 durante un año o 365 μg/m^3 durante 24 horas, puede tener efectos perjudiciales en la salud humana [5].

De forma nacional podemos consultar las normativas encargadas de la regulación de los gases contaminantes, así como la determinación del nivel en proporción en donde se considera un riesgo para la salud:

Concentración	**Requerimiento**
Promedio de 24 horas	Para el monitoreo automático o con métodos equivalentes, el cálculo del promedio aritmético de 24 horas (de la 0 a las 23 horas) requerirá un mínimo del 75 % de las concentraciones horarias válidas (18 registros). Para el muestreo por el método gravimétrico (método de referencia), la medición se realiza en periodos de 24 horas, de la 0 a las 23 horas con una frecuencia de cada 6 días y éste se considera como dato diario. Para asegurar la representatividad de los datos en el año calendario, es necesario contar con al menos tres trimestres válidos que cumplan con el número de muestras o registros y el cálculo incluirá a todos los registros del año. En caso contrario no podrá evaluarse el cumplimiento de este indicador.
Anual	Se requiere de un mínimo de datos en un año calendario. Este mínimo se evalúa a partir de la cantidad de muestras o registros válidos en 24 horas, obtenidos en cada uno de los cuatro trimestres del año. Para cada trimestre se requerirá un mínimo de 75 % de muestras o registros válidos. Para asegurar la representatividad de los datos en el año calendario, es necesario contar con al menos tres trimestres válidos que cumplan con el número de muestras o registros y el cálculo incluirá a todos los registros del año. En caso contrario no podrá evaluarse el cumplimiento de este indicador.

Cuadro 1: Requerimientos de suficiencia para validación de datos en torno a la medición por las normativas establecidas. [7]

1. **PM10**: La Norma Oficial Mexicana (NOM-025-SSA1-2021) con respecto a las partículas suspendidas PM10 y PM2.5 establece que en torno a las particulas PM10 no debe

sobrepasar los límites en el aire ambiente de 70 ($\mu g/m^3$) en 24 horas y 36 ($\mu g/m^3$) en promedio de forma anual. [7]

2. **PM2.5**: La Norma Oficial Mexicana (NOM-025-SSA1-2021) con respecto a las partículas suspendidas PM10 y PM2.5 establece que en torno a las partículas PM2.5 debe sobrepasar los límites en el aire ambiente de 41 ($\mu g/m^3$) en 24 horas y 10 ($\mu g/m^3$) de forma anual. [7]

3. **Dióxido de Azufre (SO_2)**: La Norma Oficial Mexicana (NOM-022-SSA1-2019) establece que el dioxido de azufre como contaminante atmosférico no debe rebasar el límite máximo normado de 0.2 ppm (20 ppb) en 24 horas una vez al año, y 0.03 ppm (3 ppb) en una media aritmética anual. [8]

4. **Ozono (O_3)**: La norma Oficial Mexicana (NOM-020-SSA1-2020) para evaluar la calidad del aire ambiente, con respecto al ozono (O_3). Ha determinando que el Cumplimiento gradual para valores límite de O_3 en el aire ambiente es de 0.090 ppm (90 ppb) en una hora, y 0.065 ppm (65 ppb) en un periodo de 8 horas. [9]

5. **Monóxido de carbono (CO)**: La Norma Oficial Mexicana (NOM-021-SSA1-2020), para evaluar la calidad del aire ambiente, con respecto al monóxido de carbono (C_O) establece que los valores límite de concentración en un periodo de 1 hora son de 26.0 ppm y de 9.0 ppm dentro de 8 horas.[10]

6. **Dioxido de Nitrogeno (NO_2)**: La Norma Oficial Mexicana (NOM-023-SSA1-2021) establece dos valores límite para el NO_2, 0.106 ppm para el máximo del promedio horario y 0.021 ppm para el promedio anual. [11]

Este análisis e investigación se llevó a cabo con un enfoque crítico y urgente debido al profundo impacto que la calidad del aire tiene en la salud pública. Reconociendo la importancia de abordar las amenazas ambientales para la salud, este proyecto se centró en los datos del Sistema Integral de Monitoreo Ambiental (SIMA) para comprender y mitigar los efectos de los contaminantes atmosféricos. La salud de la población y la calidad de vida de las comunidades se encuentran en juego, lo que motiva la necesidad de un análisis riguroso y exhaustivo de los datos para informar políticas y medidas que contribuyan a la protección y mejora de la salud pública a través de la gestión eficaz de la calidad del aire.

2.3. Objetivos

El objetivo principal del SIMA en Nuevo León conlleva más allá de la mera recopilación de datos; su enfoque radica en la divulgación y el fomento de la conciencia ambiental a través de la información recabada. Una parte fundamental de esta estrategia es empoderar a los ciudadanos con el conocimiento necesario para participar activamente en los esfuerzos de mitigación y en la búsqueda de soluciones que mejoren la calidad del aire y, por ende, la calidad de vida, pues se trata de la salud de la población.

En este contexto, nosotros los científicos de datos trabajaremos en el análisis de los datos proporcionados por el SIMA. Se nos da un papel crucial en la misión de comprender en profundidad la dinámica de la contaminación atmosférica en la región. Al descubrir patrones y tendencias ocultas en los datos, estos contribuyen a identificar factores clave que podrían

estar influyendo en la calidad del aire, tales como fuentes de emisión específicas, patrones climáticos, eventos estacionales, etc.

El impacto social de esta problemática es muy explícito a nivel poblacional. La calidad del aire tiene una relación directa con la salud y el bienestar de la población. Los niveles elevados de contaminantes atmosféricos están relacionados con una serie de problemas de salud, incluidas enfermedades respiratorias, cardiovasculares y problemas en el sistema inmunológico. El acceso a datos precisos y oportunos sobre la calidad del aire permite a los ciudadanos tomar decisiones informadas para proteger su salud y la de sus familias, como ajustar actividades al aire libre en función de los niveles de contaminación.

La difusión de la información a través de plataformas accesibles y comprensibles para el público general, como sitios web, aplicaciones móviles y redes sociales, es una vía fundamental para alcanzar un amplio espectro de audiencias. Al compartir de manera clara y accesible la situación actual de la calidad del aire y los hallazgos relevantes, el SIMA y los científicos de datos trabajan en conjunto para promover una mayor conciencia pública y un mayor compromiso con las políticas de mitigación. La combinación de la información proporcionada por el SIMA y el análisis profundo de los científicos de datos genera un enfoque holístico para abordar la problemática de la calidad del aire en Nuevo León. Este enfoque no solo respalda la toma de decisiones informadas por parte de las autoridades, sino que también empodera a los ciudadanos para convertirse en agentes activos de cambio, participando en la promoción de medidas de mitigación, la adopción de prácticas más sostenibles y la promoción de un entorno más saludable y habitable para todos. Es por ello que se plantean como objetivos, el análisis de datos y la difusión de los hallazgos y datos descriptivos.

2.4. Pregunta de Investigación

¿Qué gases contaminantes o factores meteorológicos afectan la clasificación de la calidad del aire de Nuevo León en el período de verano?

Se plantea esta pregunta para tener una dirección del análisis a elaborar, buscando hallazgos significativos que pueden responder esta pregunta y que pueden otorgar valor informativo. Este objetivo se centrará en las variables contaminantes PM10, PM2.5, CO, SO2, O3 y NO2. De igual manera en la variable del tiempo para los años 2022 y hasta el 17 de Agosto del 2023.

3. Comprensión y Preparación de los Datos

3.1. Comprensión de los Datos del Negocio

La entidad SIMA ha proporcionado un archivo en formato xsls que almacena los registros de datos históricos correspondientes a las mediciones de compuestos químicos presentes en la atmósfera. Estos datos han sido registrados en intervalos de una hora, abarcando desde enero de 2022 hasta agosto de 2023.

Este archivo contiene registros de las mediciones realizadas en cada una de las 13 estaciones de monitoreo con las que SIMA evalúa la calidad del aire en los municipios de Guadalupe, San Nicolás de los Garza, Monterrey, Santa Catarina, García, Escobedo, Apodaca, Juárez, San Pedro Garza García y Cadereyta, ubicados en el estado de Nuevo León.

En esta fase inicial de abordaje de la problemática, hemos optado por enfocarnos en la estación situada en San Pedro Garza García, específicamente la denominada 'Suroeste 2'. Nuestra intención es desarrollar y perfeccionar la solución en esta ubicación antes de extenderla a las demás estaciones.

En las siguientes secciones se presenta un análisis detallado del conjunto de datos asociado a la mencionada estación, con el propósito de obtener una comprensión profunda de los patrones y características de las mediciones registradas.

Para la ejecución del análisis de la base de datos, se utilizó la plataforma colaborativa Google Colaboratory, haciendo uso del lenguaje de programación Python como fundamento. Para efectuar el procesamiento de los datos y la implementación de las metodologías requeridas, se importaron diversas librerías fundamentales, entre ellas 'Pandas', la que posibilitó una administración efectiva de los datos en forma de tablas, lo cual simplificó las tareas de depuración, transformación y exploración analítica de los datos. Además de 'Pandas', se aprovecharon otras herramientas cruciales como 'scipy.stats' para análisis estadísticos, 'matplotlib.pyplot' para la generación de gráficos y visualizaciones, y 'seaborn' para mejorar la estética y la presentación visual de los resultados.

3.1.1. Dimensión de la base de datos

Inicialmente, la base de datos correspondiente a la estación 'Suroeste 2', situada en San Pedro Garza García, presenta un total de 14,255 registros. El conjunto de datos abarca desde la fecha del primer registro, que corresponde al 01 de enero de 2022 a las 0:00 horas, hasta el último registrado el 17 de agosto de 2023 a las 23:00 horas. Cada una de estas observaciones está compuesta por 16 columnas, las cuales proporcionan descripciones detalladas de los componentes registrados en las mediciones efectuadas.

Filas	14255
Columnas	16

Cuadro 2: Dimensión de la base de datos 'Suroeste 2'

3.1.2. Descripción de variables

El conjunto de datos abarca una diversidad de variables que capturan de manera precisa una serie de aspectos químicos y meteorológicos en la estación mencionada. Las variables se detallan en el Cuadro 3, en el cual se brinda una descripción de cada una, se especifica su tipo de dato y se proporciona información sobre las unidades en las que dichos valores son presentados. Anteriormente se realizó la descripción detallada de nuestras variables de interés.

Etiqueta	**Tipo**	**Descripción**	**Unidades**
date	object	Fecha y hora de registro	d/m/a y h:m
PM10	float64	Material Particulado menor a 10 micrómetros	$\frac{\mu g}{m^3}$
PM2.5	float64	Material Particulado menor a 2.5 micrómetros	$\frac{\mu g}{m^3}$
O3	float64	Ozono	ppb
SO2	float64	Dióxido de Azufre	ppb
NO2	float64	Dióxido de Nitrógeno	ppb
CO	float64	Monóxido de Carbono	ppm
NO	float64	Monóxido de Nitrógeno	ppb
NOx	float64	Es la suma de NO + NO2	ppb
TOUT	float64	Temperatura	°C
RH	float64	Humedad Relativa	%
RAINF	float64	Precipitación	$\frac{mm}{Hr}$
PRS	float64	Presión Atmosférica	$\frac{mm}{Hg}$
WSR	float64	Velocidad del Viento	$\frac{Km}{hr}$
WDR	float64	Dirección del Viento	°

Cuadro 3: Descripción de etiquetas de la base de datos

3.1.3. Exploración de datos

En el Cuadro 4 se obtuvieron las medidas estadísticas de cada variable a través del método '.describe()' de la biblioteca de Python llamada Pandas. Este método se utiliza para generar estadísticas descriptivas de un DataFrame o una Serie. Proporciona un resumen de varias estadísticas importantes para cada columna numérica en el DataFrame, como el conteo, la media, la desviación estándar, los valores mínimo y máximo, y los percentiles.

Etiqueta	**media**	**std**	**min**	**25 %**	**50 %**	**75 %**	**max**	**var**
CO	1.324	0.628	0.050	0.790	1.310	1.820	4.740	0.39
NO	7.377	9.470	1.700	2.900	4.000	8.000	162.100	89.69
NO2	16.069	9.673	0.700	9.000	14.400	21.000	97.400	93.58
NOX	23.415	16.655	4.000	12.200	19.200	29.200	207.200	277.40
O3	25.577	19.068	1.000	11.000	22.000	35.000	138.000	363.59
PM10	65.341	43.261	6.000	40.000	57.000	79.000	999.000	1871.53
PM2.5	17.938	11.232	1.000	10.000	16.000	24.000	105.000	126.17
PRS	713.808	1.551	708.000	712.800	713.700	714.700	719.600	2.40
RAINF	0.001	0.027	0.000	0.000	0.000	0.000	1.290	0.00
RH	51.416	19.815	1.000	36.000	52.000	68.000	89.000	392.65
SO2	4.873	2.530	2.100	3.500	4.300	5.300	83.800	6.40
SR	0.176	0.251	-0.002	0.000	0.005	0.326	0.904	0.06
TOUT	22.913	7.594	-2.280	18.085	23.860	28.220	41.990	57.67
WSR	11.173	5.002	1.600	7.000	11.200	14.900	30.400	25.02
WDR	128.982	62.542	1.000	85.000	110.000	150.000	359.000	3911.5

Cuadro 4: Estadísticas de las variables

En la Figura 2 Se graficaron los histogramas de las variables que representan un riesgo para la salud que se mencionaron anteriormente en la investigación y que son las que usamos en esta fase del proyecto. A través de estos gráficos se puede analizar la distribución de cada variable. En un escenario ideal las variables deberían presentar un sesgo hacia el lado izquierdo para indicar cantidades bajas en los compuestos químicos dañinos a la salud. Esto indica una buena calidad del aire en general.

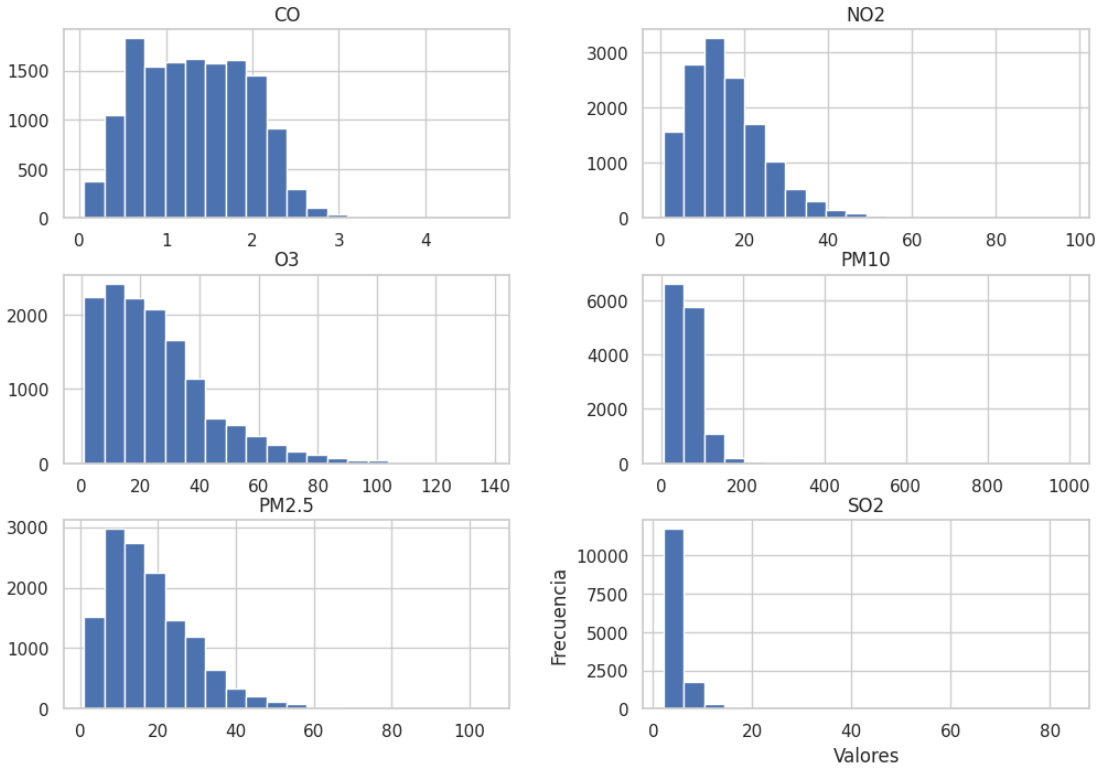


Figura 2: Histogramas de variables de interés [10]

Al observar los gráficos se puede apreciar claramente que este es el caso para las variables de los compuestos Dióxido de Azufre, Material Particulado menor a 10 micrómetros, Material Particulado menor a 2.5 micrómetros y Dióxido de Nitrógeno.

La variable que representa la concentración de Ozono también tiene un sesgo a la izquierda, aunque presenta algunos datos que alcanzan niveles mas altos del compuesto.

La concentración de Monóxido de Carbono es la que presenta mayor cantidad de datos en un rango elevado para el compuesto, lo que indica mayor presencia en la atmósfera.

En la Figura 3 se muestra el mapa de calor de la matriz de correlación de las variables de nuestro interés. Esta correlación se obtuvo con el método '.corr()', que es una función de la biblioteca Pandas en Python. Se utiliza para calcular la matriz de correlación entre las columnas numéricas de un DataFrame. La matriz de correlación es una representación tabular que muestra cómo las variables numéricas en un conjunto de datos están relacionadas entre sí.

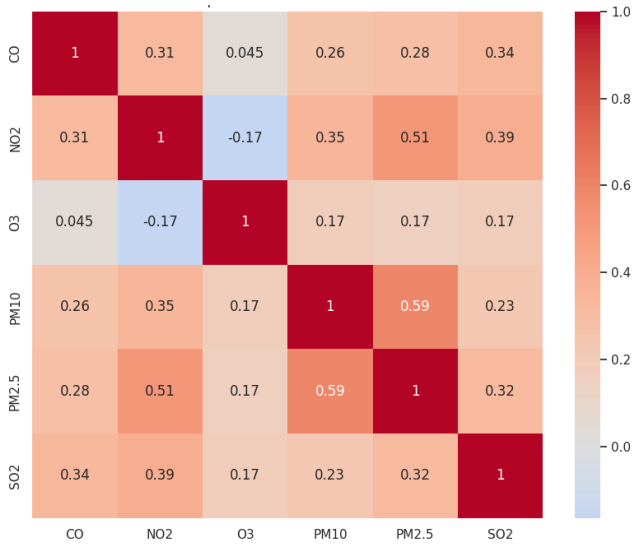


Figura 3: Correlación entre variables de interés

De igual forma se analizaron las variables con mayores correlaciones para determinar si había una dependencia significativa a través del p-value. En la Figura 4 se muestran los resultados, todos los p-values fueron menores al nivel de significancia 0.05, por lo que podemos decir con un 95 % de confianza que hay una relación significativa entre las variables mostradas.

Variables	Correlación	p-value
NO2 vs PM10	0.332535	0.000000e+00
NO2 vs PM2.5	0.509525	0.000000e+00
NO2 vs SO2	0.385133	0.000000e+00
NO2 vs CO	0.301379	6.336621e-270
CO vs SO2	0.331957	0.000000e+00
PM10 vs PM2.5	0.586167	0.000000e+00

Figura 4: Correlación y p-value

Las variables en cuestión no exhiben indicios de correlación significativamente alta entre sí. No obstante, se distingue una relación moderada entre el Material Particulado de tamaño inferior a 10 micrómetros y el Material Particulado de tamaño inferior a 2.5 micrómetros, con un coeficiente cercano a 0.6. Este fenómeno podría fundamentarse en las reacciones químicas que desprenden este tipo de partículas al ambiente, las cuales tienen la capacidad de generar

ambos compuestos.

Al mismo tiempo, existe una correlación moderada entre el Material Particulado de tamaño inferior a 10 micrómetros y el Dióxido de Nitrógeno. No obstante, aún no contamos con evidencia que indique las causas de esta relación. Cabe destacar que el resto de las variables presentan una correlación débil, lo cual es esperado ya que son compuestos muy distintos entre sí.

3.2. Preparación de los Datos

3.2.1. Selección conjunto de datos

Tal como se ha expuesto en la sección introductoria de este documento, conforme a los resultados de nuestra investigación, se identifican las variables que ejercen un impacto directo sobre la calidad del aire y que potencialmente pueden acarrear efectos perjudiciales para la salud. Estas variables son las siguientes:

1. Monóxido de carbono (CO)
2. Dióxido de Nitrógeno (NO2)
3. Dióxido de azufre (SO2)
4. Ozono (O3)
5. Material Particulado menor a 10 micrómetros (PM10)
6. Material Particulado menor a 2.5 micrómetros (PM2.5)

Dichas columnas de la base de datos fueron utilizadas para determinar el vector objetivo: Calidad del aire. Para este propósito, se tomó como referencia los límites de riesgo previamente investigados para cada componente específico. Esta aproximación permitió efectuar una evaluación concreta para discernir la eventual existencia de riesgos en el ambiente debido a la presencia de determinados componentes químicos.

Hasta el presente momento, las demás variables no fueron consideradas debido a la ausencia de evidencia que indique un impacto directo en la calidad del aire con implicaciones adversas para la salud. Cabe señalar que estas columnas no fueron eliminadas totalmente con el propósito de mantener su disponibilidad para análisis futuros; su omisión en el presente estudio reside en su falta de relevancia para nuestro vector objetivo, no obstante, podrían resultar significativas en contextos analíticos posteriores.

3.2.2. Limpieza de datos

Tal como se destacó previamente, varias columnas de la base de datos presentaban valores nulos, lo cual requería un abordaje específico. En esta instancia, se optó por no proceder a la eliminación de filas que contuvieran valores nulos, incluso en casos donde únicamente una columna poseía un valor y el resto se encontraba desprovisto de datos. La razón radica en el potencial aporte de información de estas filas. A modo de ejemplo, si en la columna correspondiente al monóxido de carbono (CO) se registrara un valor de 100 mg/m^3, y en las

demás columnas los valores fueran nulos, el descarte de esta fila resultaría en la pérdida de un dato de gran importancia, dado que dicho valor de CO indica una situación de riesgo en la calidad del aire.

Es precisamente por este motivo que se optó por descartar exclusivamente aquellas filas en las cuales todos los campos presentaran valores nulos. En el transcurso de este proceso, se identificaron un total de 29 filas que cumplían con este criterio y, en consecuencia, fueron eliminadas de manera definitiva de la base de datos. Para el análisis de correlación entre las variables, fue necesario llevar a cabo una eliminación temporal de todas las filas que contenían al menos un valor nulo. Una vez culminado dicho análisis, estas filas previamente excluidas fueron reintegradas al conjunto de datos original.

No fue necesario llevar a cabo correcciones ortográficas ni la conversión de variables a categóricas, dado que nuestra labor se enfocó en variables continuas. Si bien se constató la presencia de un número considerable de valores atípicos, se decidió no proceder a su eliminación. Esta elección se fundamenta en la consideración de que tales valores atípicos revisten un interés sustancial, ya que podrían actuar como desencadenantes de situaciones de riesgo en lo que concierne a la calidad del aire, ya que son los que sobrepasan los límites de riesgo establecidos y son nuestro foco de interés.

3.3. Transformación de Datos

3.3.1. Discretización y normalización

A pesar de que algunas variables contaban con distintas unidades, no decidimos normalizar los datos ya que el vector objetivo se determinó tomando cada variable de forma independiente. Además se trabajo con variables cuantitativas continuas y tampoco fue necesario transformarlas en categorías discretas, ya que su propia naturaleza ya proporciona la información que necesitamos.

Además se buscó un interpretación de los resultados más clara y directa manteniendo las variables en sus unidades originales. Esto nos facilita la comprensión de cómo los cambios en una variable se relacionan con cambios en el vector objetivo.

A pesar de la disparidad de unidades entre algunas variables, se optó por no llevar a cabo la normalización de los datos, dado que la determinación del vector objetivo se efectuó considerando cada variable de manera independiente. Adicionalmente, considerando la naturaleza de las variables involucradas (cuantitativas continuas) tampoco se consideró necesario efectuar su conversión en categorías discretas ya que la propia naturaleza de estas variables ya provee la información requerida para nuestros propósitos.

Es relevante resaltar que se buscó obtener una interpretación de los resultados clara y directa manteniendo las variables en sus unidades originales. Esto facilitó la comprensión de cómo los datos en una variable se relacionan con cambios en el vector objetivo.

3.3.2. Atributos derivados

Con el propósito de realizar un análisis exhaustivo de la base de datos, se introdujeron atributos adicionales. Inicialmente, la columna 'date' fue desglosada en tres columnas suple-

mentarias: 'día', 'mes' y 'año'. Este procedimiento se llevó a cabo con la finalidad de facilitar la generación posterior de gráficos que visualicen los meses con una mayor incidencia de componentes de riesgo, así como para investigar la posible relación entre el mes y los niveles de contaminación del aire.

Dada la importancia de evitar que la presencia de componentes exceda los límites establecidos en términos de horas o días, ya que a medida que se prolonga la exposición a contaminantes también aumenta el riesgo para la salud humana, se introdujo una columna denominada "día único". En esta columna, se asignó el valor '1' a todas las filas correspondientes al primer día, '2' a todos los registros del segundo día, y así sucesivamente hasta alcanzar el día '594', que corresponde al último registrado en la base de datos. El propósito de esta nueva columna es realizar un conteo de las horas por día para cada componente, lo que permitirá determinar si dichos niveles representan un riesgo significativo para la salud de las personas.

El vector objetivo fue creado como una clase binaria, en la que un valor '1' denota la presencia de riesgo en la calidad del aire debido a la existencia de componentes tóxicos perjudiciales para la salud. En contraste, un valor '0' indica la ausencia de riesgo en el aire por la presencia de dichos componentes nocivos.

Primero se estableció un listado de los límites de cada componente, investigados anteriormente, en donde su presencia no representa un riesgo:

1. lim_CO = 9 ppm
2. lim_NO2 = 53 ppb
3. lim_O3 = 65 ppb
4. lim_PM10 = 70μg/m^3
5. lim_PM25 = 41 μg/m^3
6. lim_SO2 = 20 ppb

La representación de los límites a través de la asignación de las variables fue seleccionada con el propósito de brindar flexibilidad para futuras actualizaciones de investigaciones y para permitir ajustes en la rigurosidad o amplitud de los intervalos correspondientes a cada componente. Esta estructura posibilita la modificación sencilla de estos límites, otorgando la capacidad de desplazar estos intervalos de manera eficaz según las necesidades emergentes.

Para establecer la presencia de riesgo, es decir, la clase '1', se llevó a cabo considerando si al menos uno de los siguientes seis escenarios se cumplía:

1. Si el valor en la columna CO es mayor o igual a lim_CO
2. Si el valor en la columna NO2 es mayor o igual a lim_NO2
3. Si el valor en la columna O3 es mayor o igual a lim_O3
4. Si el valor en la columna PM10 es mayor o igual a lim_PM10
5. Si el valor en la columna PM25 es mayor o igual a lim_PM25

6. Si el valor en la columna SO2 es mayor o igual a lim_SO2

En caso contrario la clase asignada era '0', es decir, no hay presencia de riesgo en el aire.

Adicionalmente, se incorporaron seis columnas binarias, cada una de las cuales representaba uno de los seis componentes analizados. En situaciones en las que se determinara mediante nuestro vector objetivo que existe un riesgo en el aire, se procedería a identificar qué componentes rebasaron dicho umbral. Aquellos componentes que fueran identificados se registrarían como '1' en sus respectivas columnas, mientras que aquellos que no excedieran el límite se registrarían como '0'.

3.3.3. Creación de columnas para clasificación de la calidad del aire

Con el fin de visualizar cuales son los contaminantes que influyen en la calidad del aire se hace una análisis de los contaminantes principales CO, NO2, O3, PM10, PM2.5 y SO2 respecto a la clasificación de la calidad del aire.

En el código llamado ClasificacionCalidadAire.Rmd se leen la hojas xsls y se asegura que los datos tengan tipo número. Se leen los datos de la estación deseada y se crean ciclos for para tomar cuenta de que contaminantes caen en una clasificación del aire buen, aceptable, mala, muy mala o extremadamente mala. Para poder visualizar las variables que impactan directa y altamente a la clasificación del calidad del aire se obtuvo la columna 'CalidadAire'. Con los intervalos proporcionados por el SEMARNAT en la norma 172 se crea una columna para cada contaminante y su valor desde 1 (Extremadamente Mala) hasta 5 (Buena). También se tiene el posible valor NULL para los datos no proporcionados.

Para cada uno de los contaminantes se creo la columa CCO, CNO2, CO3, CPM10, CPM2.5 y CSO2. C es acrónimo de Clasificación. Estas clasificación se hacen para la cantidad de observaciones en la estación deseada y se miden por hora de registro. Con esto se tienen 6 nuevas columnas con la calidad del aire por contaminante clasificadas numéricamente y de manera ordinal. Para obtener la calidad del aire en cierta hora según la combinación de los contaminantes, se hace una comparación entre los valores de los contaminantes y el valor más chico se inserta en la columna 'CalidadAire'. Como la calidad del aire es influyente por todos estos contaminantes en conjunto, la presencia de uno en exceso es nocivo para la salud, por lo que se toma el valor más nocivo para definir como está la calidad del aire en esa hora.

Con esto se obtiene la columna 'CalidadAire', la cual consiste de valores del 1-5, indicando la calidad del aire. 1 siendo Extremadamente Mala y 5 Buena. En este caso no es necesario tener los valores de manera numérica porque son clases que no tendrían significancia estadística en algún análisis, sea correlación o de regresión. Por esta razón, los valores de la columna de calidad del aire se transforman de numéricos ordinales a tipo carácter (chr). Entonces finalmente se tienen los valores Extremadamente Mala, Muy Mala, Mala, Aceptable y Buena en la columna.

3.4. Exploración con Análisis Factorial por Mínimo Residuo y Verosimilitud

El análisis factorial es una técnica estadística ampliamente utilizada con diversos propósitos, y sus aplicaciones principales incluyen [17]:

- Comprender la estructura de un conjunto de variables. Permite identificar las relaciones y las dimensiones subyacentes que explican las variaciones observadas en las variables originales.
- Ayuda a identificar patrones y relaciones no evidentes en los datos que pueden ser útiles para la interpretación y la toma de decisiones.
- Reducción de dimensionalidad del conjunto de datos conservando la mayor parte de la información original mientras simplifica el análisis y la interpretación de los datos.

Al aplicar una técnica de análisis de relaciones de interdependencia en el contexto de nuestra investigación, nos interesa como análisis exploratorio implementar esta técnica y disminuir la dimensionalidad del dataset ya que contamos con una gran cantidad de variables que equivalen a cada compuesto químico.

En nuestro contexto, se identifican algunas variables que podrían presentar redundancia en la información y, posiblemente, ejercer un impacto negativo en el rendimiento del modelo. Entre estas variables se incluyen NO, NO2 y NOX. Con el propósito de abordar esta redundancia y mejorar la eficiencia del análisis, se plantea la creación de una nueva variable que represente una simplificación de estas tres a través de un análisis factorial.

3.4.1. Verificación de supuestos

Para poder aplicar el análisis factorial es de suma importancia asegurar que el dataset es apto. Por lo que aplicamos el "Test of Sphericity", esta prueba se utiliza para determinar si las correlaciones entre las variables son significativamente diferentes de cero en conjunto.

Bartlett's test of sphericity suggests that there is sufficient significant correlation in the data for factor analysis ($Chisq(105) = 175319.17,\ p < 0.001$).

Se obtuvo un p-value de 0.001 el cual es menor que el nivel de significancia preestablecido de 0.05. Este resultado nos indica que hay evidencia para rechazar la hipótesis nula y concluir que las variables están relacionadas entre sí y que un análisis factorial puede ser apropiado para explorar la estructura subyacente de los datos.

3.4.2. Matriz de comunalidades

La matriz de comunalidades representa la proporción de la varianza total de una variable que es explicada por los factores extraídos en el modelo. Buscamos obtener valores altos para las variables de NO, NO2 y NOX.

Calculamos los resultados con los dos métodos: análisis factorial de verosimilitud y de mínimo residuo. La elección del método depende de la naturaleza de los datos y las suposiciones que se pueden hacer. Si los datos son aproximadamente normales, el análisis factorial de verosimilitud puede ser una elección adecuada. A continuación mostramos el código donde se evaluaron ambos modelos.

Variable	verosimilitud	mínimo residuo
NOX	0.997854178	0.996980711
NO2	0.995963324	0.729697210
NO	0.995897224	0.718639714
PM2.5	0.290777724	0.543510368
SO2	0.168208542	0.521163282
CO	0.129483261	0.511368424
PM10	0.126448988	0.457184794
WSR	0.102553095	0.430122830
O3	0.083005031	0.350687104
TOUT	0.076002887	0.269337233
WDR	0.032973373	0.207691008
PRS	0.026994435	0.186422322
RH	0.019124563	0.101322126
SR	0.010586939	0.061694633
RAINF	0.001144925	0.002972587

Cuadro 5: Valores de verosimilitud y mínimo residuo para cada variable.

El output de los "scores" en el Cuadro: 5 nos permite identificar que el modelo 1 de máxima verosimilitud es mucho más claro. Esto ya que los valores con son más extremos ya sea para los factores con mayor o menor importancia lo que nos permite identificar más fácilmente cuales tienen una significancia superior para nuestra base de datos.

3.4.3. Número óptimo de factores

Para determinar el número óptimo de factores se realizó el cálculo de la varianza acumulada de los factores y un gráfico de Scree Plot.

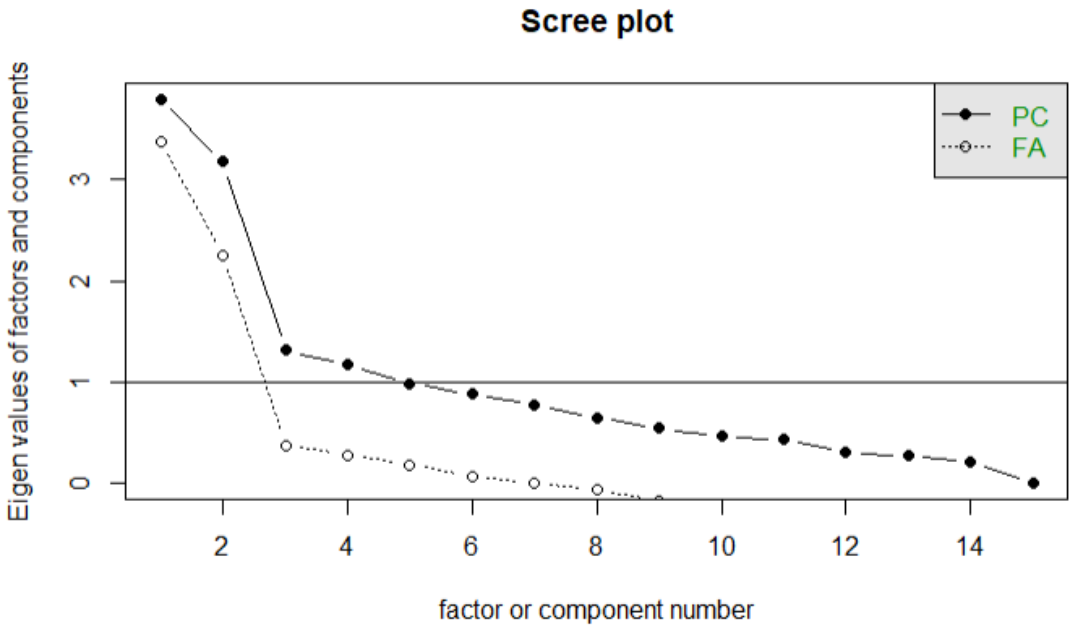


Figura 5: Gráfico Scree para análisis factorial. [19]

Gráficamente observamos que con el factor 1 y el factor 2 se obtiene gran parte de la información de los datos originales. La decisión de tomar solamente el primero es porque su varianza acumulada es de 0.8644652. Además, buscamos la representación de las variables NO, NO2 y NOX.

3.4.4. Rotación de la matriz

La rotación facilita la asignación de un mayor peso a las variables que están mejor representadas por los factores extraídos.

Generamos el modelo con dos factores y rotación "quartimax" ya que proporciona una solución más interpretable al simplificar la estructura de los datos. Se usaron dos factores para proyectar a nuestras variables que buscamos agrupar como vectores. Se puede observar en la Figura 6 que nuestras variables de interés: NO, NO2 y NOX son las que tienen una mayor proyección en el primer factor. Por lo que elegimos el Factor 1 para representarlas.

3.5. Resultados Preliminares

Se puede observar en la Figura 7 que durante las primeras tres horas del día 1 de enero de 2022 se encontró la presencia del componente PM10 por encima de los 70 ($\mu g/m^3$). Como

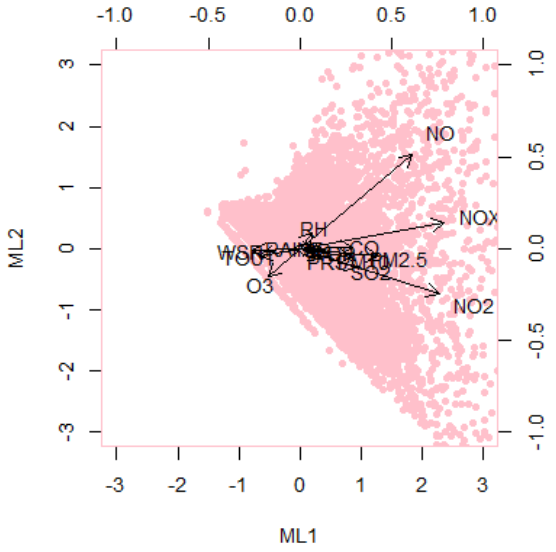


Figura 6: Gráfico análisis factorial.

mencionamos anteriormente, las partículas PM10 no debe sobrepasar los límites en el aire ambiente de 70 ($\mu g/m^3$) en 24 horas [7] porque representan un riesgo para la salud. En este caso no fueron 24 horas, pero es un indicativo de que hay un potencial riesgo. De igual forma con PM2.5, las primeras dos horas se registraron partículas por encima de los 41 ($\mu g/m^3$), lo que de igual forma sobrepasar ese límite representa un riesgo para la salud [7].

	date	CO	NO	NO2	NOX	O3	PM10	PM2.5	PRS	RAINF	...	día	fecha_completa	día_único	CA	CO_	NO2_	O3_	PM10_	PM2.5_	SO2_
0	2022-01-01 00:00:00	1.62	2.5	21.6	24.1	33.0	106.0	53.0	711.4	0.0	...	1	1/1/2022	1	1	0	0	0	1	1	0
1	2022-01-01 01:00:00	1.33	2.5	12.9	15.5	36.0	89.0	49.0	711.3	0.0	...	1	1/1/2022	1	1	0	0	0	1	1	0
2	2022-01-01 02:00:00	1.29	2.6	14.9	17.5	34.0	77.0	39.0	711.3	0.0	...	1	1/1/2022	1	1	0	0	0	1	0	0
3	2022-01-01 03:00:00	1.40	2.8	21.2	24.0	25.0	54.0	35.0	711.1	0.0	...	1	1/1/2022	1	0	0	0	0	0	0	0
4	2022-01-01 04:00:00	1.39	2.8	19.8	22.6	22.0	68.0	35.0	711.1	0.0	...	1	1/1/2022	1	0	0	0	0	0	0	0
5	2022-01-01 05:00:00	1.23	2.7	15.2	17.9	23.0	60.0	27.0	711.1	0.0	...	1	1/1/2022	1	0	0	0	0	0	0	0
6	2022-01-01 06:00:00	1.12	2.4	13.1	15.5	23.0	50.0	22.0	711.1	0.0	...	1	1/1/2022	1	0	0	0	0	0	0	0
7	2022-01-01 07:00:00	1.10	3.1	14.5	17.7	20.0	37.0	13.0	711.3	0.0	...	1	1/1/2022	1	0	0	0	0	0	0	0
8	2022-01-01 08:00:00	0.81	2.6	5.5	8.2	24.0	40.0	7.0	711.4	0.0	...	1	1/1/2022	1	0	0	0	0	0	0	0
9	2022-01-01 09:00:00	0.78	3.0	4.4	7.5	25.0	36.0	7.0	711.6	0.0	...	1	1/1/2022	1	0	0	0	0	0	0	0
10	2022-01-01 10:00:00	0.86	4.7	8.0	12.7	23.0	34.0	0.0	711.9	0.0	...	1	1/1/2022	1	0	0	0	0	0	0	0
11	2022-01-01 11:00:00	0.79	3.3	5.1	8.5	29.0	34.0	3.0	712.0	0.0	...	1	1/1/2022	1	0	0	0	0	0	0	0
12	2022-01-01 12:00:00	0.76	3.4	4.4	7.9	32.0	90.0	7.0	711.9	0.0	...	1	1/1/2022	1	1	0	0	0	1	0	0
13	2022-01-01 13:00:00	0.75	3.3	4.4	7.8	37.0	205.0	7.0	711.5	0.0	...	1	1/1/2022	1	1	0	0	0	1	0	0
14	2022-01-01 14:00:00	0.74	3.2	4.5	7.8	40.0	150.0	8.0	711.3	0.0	...	1	1/1/2022	1	1	0	0	0	1	0	0
15	2022-01-01 15:00:00	0.74	3.3	5.1	8.5	41.0	191.0	5.0	711.1	0.0	...	1	1/1/2022	1	1	0	0	0	1	0	0

Figura 7: Base de datos con atributos agregados

En posteriores análisis buscaremos relaciones entre la presencia de estos componentes y si existen algunos que sobrepasen el tiempo establecido. Además de hacer una representación gráfica por mes de los componentes tóxicos para visualizar si existe alguna dependencia de la época del año con la presencia de estos contaminantes.

Para la clasificación de la calidad del aire se hizo el análisis de variables y sus gráficas entre ellas, asignando el color de la calidad del aire respectivo a cada observación en los renglones de la base de datos de la estación seleccionada.

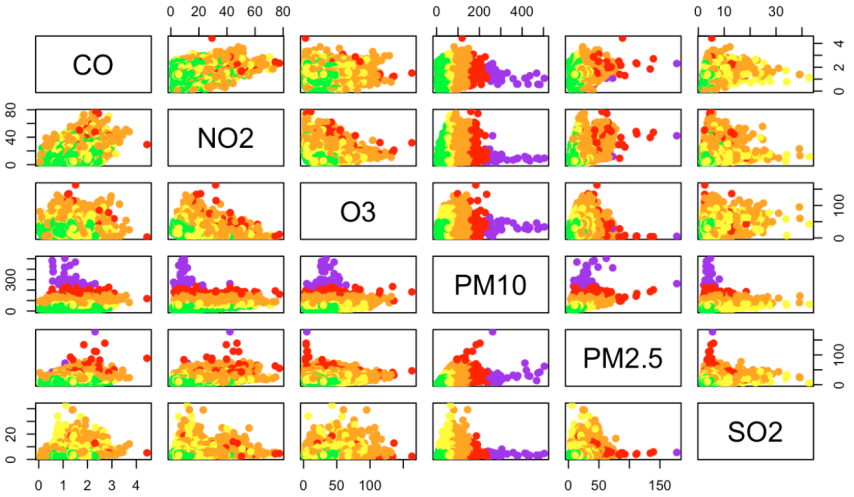


Figura 8: Gráfica de contaminantes y calidad del aire de estación SURESTE

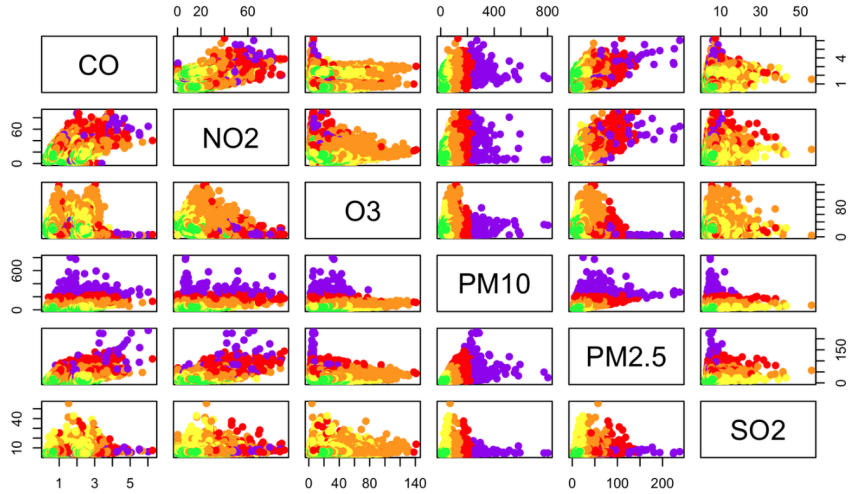


Figura 9: Gráfica de contaminantes y calidad del aire de estación SUROESTE

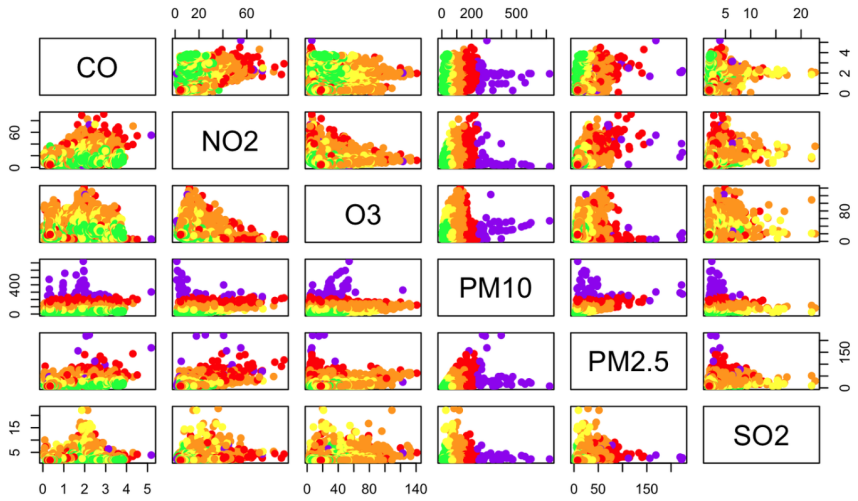


Figura 10: Gráfica de contaminantes y calidad del aire de estación NOROESTE 2

Estas gráficas muestran la calidad del aire para la combinación de todos los contaminantes. Se presentan sólo 3 estaciones al azar par mostrar la presencia de una calidad del aire extremadamente mala para valores altos de PM10 y PM2.5. Esta observación conteste más ampliamente el resultado preliminar de que los contaminantes PM10 y PM2.5 toman valores altos. En estás gráficas se ve que evidentemente sí.

4. Adecuación y Validación del Modelo

4.1. Regresión Multivariada

Un modelo de regresión lineal múltiple es un modelo estadístico versátil para evaluar las relaciones entre un destino continuo y los predictores. Los predictores pueden ser campos continuos, categóricos o derivados, de modo que las relaciones no lineales también estén soportadas. El modelo es lineal porque consiste en términos de aditivos en los que cada término es un predictor que se multiplica por un coeficiente estimado. El término de constante (intercepción) también se añade normalmente al modelo [16].

Para modelo, es necesario que previamente se haga un análisis de las variables y como se relacionan entre ellas, así poder seleccionar las variables para poder tener una predicción asertiva. Este análisis se base del concepto de correlación entre variables, el cual es la medida de la relación estadística entre dos variables.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon$$

Donde:

- y es la variable dependiente que se desea predecir.
- $x_1, x_2, \ldots, x_n$ son las variables independientes.
- $\beta_0, \beta_1, \beta_2, \ldots, \beta_n$ son los coeficientes de regresión que determinan cómo influyen las variables independientes en la variable dependiente.
- ϵ representa el término del error.

El objetivo de la regresión lineal multivariable es encontrar los valores óptimos para los coeficientes $\beta_0, \beta_1, \beta_2, \ldots, \beta_n$ que minimicen la suma de los cuadrados de los errores (el término ϵ) entre los valores observados y los valores predichos por el modelo. Para ajustar un modelo de regresión lineal multivariable, se utilizan métodos estadísticos, como el método de mínimos cuadrados, que busca encontrar los coeficientes que minimizan la suma de los residuos al cuadrado.

El procedimiento de selección de modelos depende de si un predictor categórico está presente o no. Cuando sólo se especifica un predictor continuo, se tienen en cuenta los tres modelos siguientes.

1. Un modelo constante que siempre predice la media general.
2. El modelo lineal con el predictor único se ha añadido a la constante.
3. Modelo cuadrático en el que se añade el predictor cuadrado al modelo lineal [16].

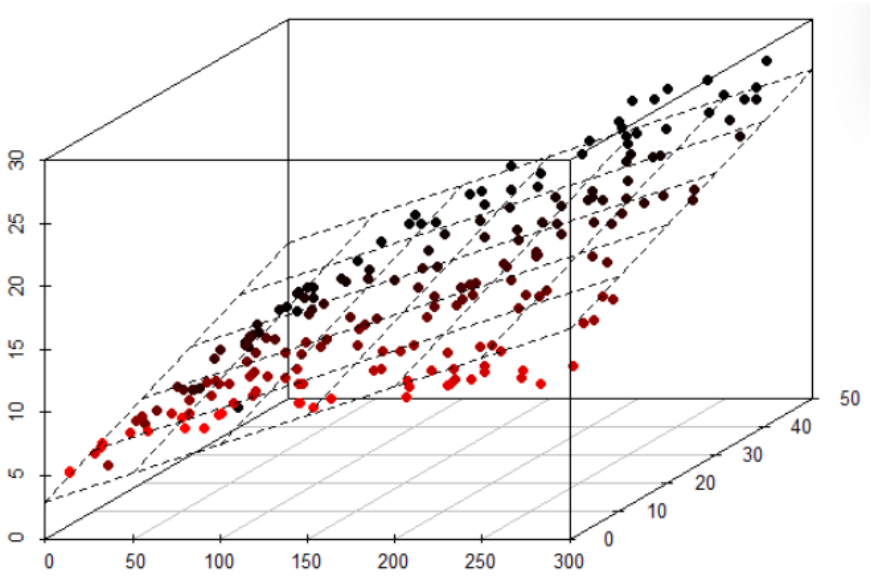


Figura 11: Gráfico de Regresión Lineal Múltiple. [19]

4.1.1. Selección de variables modelo regresión multivariado

En este estudio, nos centraremos en el análisis de la calidad del aire en Monterrey durante los meses de Julio y Agosto. Estos meses son de particular interés debido a las condiciones climáticas y atmosféricas que pueden influir en la concentración de ozono. Utilizaremos un modelo de regresión lineal para investigar la relación entre el ozono como variable dependiente y otras variables que podrían ser consideradas como independientes.

El objetivo principal de este análisis es identificar y cuantificar la influencia de diversas variables ambientales, como la temperatura, la humedad relativa, la velocidad del viento, además de la concentración de otros gases contaminantes, entre otras, en la concentración de ozono en el aire durante los meses de verano. Además, realizaremos un riguroso análisis de independencia entre estas variables para asegurarnos de que el modelo de regresión lineal sea válido y que las suposiciones subyacentes sean satisfechas.

Llevamos a cabo un análisis de covarianzas y correlación para evaluar la relación y la influencia mutua entre la concentración de ozono (O3) y un conjunto de otras variables ambientales clave. Estas variables incluyeron monóxido de carbono (CO), óxido de nitrógeno (NO), dióxido de nitrógeno (NO2), óxidos de nitrógeno totales (NOX), partículas suspendidas PM10 y PM2.5, presión atmosférica (PRS), precipitación (RAINF), humedad relativa (RH), dióxido de azufre (SO2), radiación solar (SR), temperatura del aire exterior (TOUT), velocidad del viento (WSR), y dirección del viento (WDR).

Este análisis permitió identificar patrones de asociación, dependencia o independencia entre estas variables y la concentración de ozono. Los resultados de las covarianzas y correlaciones proporcionaron información valiosa sobre cómo estas variables podrían estar relacionadas entre sí y cómo podrían influir en la concentración de ozono durante los meses de Julio y Agosto en Monterrey. Estos hallazgos contribuyeron a una comprensión más completa de los factores que afectan la calidad del aire en la región y respaldaron las conclusiones de nuestro estudio.

Y a través de nuestro análisis de correlación estos fueron nuestros resultados en torno a determinar las variables predictoras candidatas a nuestro modelo de regresión multivariada:

Cuadro 6: Matriz de Correlación (Parte 1)

	CO	NO	NO2	NOX	O3	PM10	PM2.5	PRS
CO	1.0000	0.3177	0.3324	0.3708	0.2208	0.2747	0.2301	0.3337
NO	0.3177	1.0000	0.5467	0.8380	-0.1899	0.2254	0.2986	0.1906
NO2	0.3324	0.5467	1.0000	0.9138	0.0667	0.4494	0.5348	0.3103
NOX	0.3708	0.8380	0.9138	1.0000	-0.0496	0.4012	0.4913	0.2960
O3	0.2208	-0.1899	0.0667	-0.0496	1.0000	0.3108	0.4309	0.2336
PM10	0.2747	0.2254	0.4494	0.4012	0.3108	1.0000	0.6842	0.3340
PM2.5	0.2301	0.2986	0.5348	0.3103	0.0045	0.3898	0.4081	0.3102

Cuadro 7: Matriz de Correlación (Parte 2)

	RAINF	RH	SO2	SR	TOUT	WSR	WDR
RAINF	0.0124	0.2258	0.5158	0.1817	0.1746	0.0338	-0.2020
RH	-0.0238	0.3172	0.1899	0.1943	0.0169	-0.2676	0.0591
SO2	0.0045	0.3898	0.4081	0.3102	0.1099	-0.2215	0.1403
SR	-0.0088	0.4085	0.3570	0.2949	0.0809	-0.2723	0.1205
TOUT	-0.0055	-0.4035	0.4830	0.6632	0.5529	0.4790	-0.1274
PM10	0.0950	0.4704	0.4366	0.3633	-0.0353	0.0950	0.4704
PM2.5	0.0236	0.4606	0.6463	0.3633	-0.0203	0.0236	0.4606
PRS	0.4993	0.4301	0.1348	0.7732	0.4297	0.2735	0.2549

- **Radiación Solar (SR):** La variable que mostró la correlación más alta con el ozono fue la radiación solar (SR) con un valor de 0.6632. Esto sugiere una relación positiva significativa entre la radiación solar y la concentración de ozono. Es decir, cuando la radiación solar es alta, es más probable que los niveles de ozono también lo sean.
- **Temperatura del Aire Exterior (TOUT):** La temperatura del aire exterior también mostró una fuerte correlación positiva con el ozono, con un valor de 0.5529. Esto indica que a medida que la temperatura del aire exterior aumenta, los niveles de ozono tienden a aumentar en consecuencia.
- **Partículas Suspendidas PM2.5:** Las partículas suspendidas PM2.5 mostraron una correlación positiva significativa con el ozono, con un valor de 0.4309. Esto sugiere que la presencia de partículas finas en el aire (PM2.5) puede estar relacionada con mayores concentraciones de ozono.
- **Humedad Relativa (RH):** La humedad relativa mostró una correlación negativa importante con el ozono, con un valor de -0.4035. Esto significa que cuando la humedad relativa es alta, es menos probable que se registren niveles altos de ozono.

- **Dióxido de Azufre (SO2):** El dióxido de azufre mostró una correlación positiva sólida con el ozono, con un valor de 0.4830. Esto sugiere que la presencia de dióxido de azufre en el aire puede contribuir al aumento de los niveles de ozono.
- **Velocidad del Viento (WSR):** La velocidad del viento también mostró una correlación positiva significativa con el ozono, con un valor de 0.4790. Esto indica que mayores velocidades del viento pueden estar asociadas con mayores niveles de ozono.

Es sumamente notable el aumento en la correlación entre las partículas PM10 y PM2.5, además del gas contaminante SO2 en comparación con la visión del panorama completo de un año. Esto sugiere que, en ciertos períodos o condiciones específicas, las partículas PM10 y PM2.5 pueden estar relacionadas de manera más estrecha. La influencia del ozono en esta correlación ya sea negativa o positiva podría ser significativa, ya que el ozono es un contaminante atmosférico que puede interactuar con otras partículas y gases en la atmósfera, formando compuestos secundarios. Esto podría contribuir a la cohesión entre las partículas PM10 y PM2.5 en ciertos momentos, lo que resulta en una correlación más fuerte.

4.1.2. Análisis de dependencia

A continuación, se realizará una parte del análisis en la que se explorará la dependencia entre las siguientes variables utilizando la prueba de correlación de Pearson (cor.test):

r <dbl>	t <dbl>	valor p <dbl>
-0.276	-16.653	0.000
-0.110	-6.389	0.000
-0.237	-14.101	0.000
0.024	1.370	0.171
0.415	26.410	0.000
0.354	21.920	0.000
0.646	49.059	0.000
0.437	28.111	0.000
0.669	52.120	0.000
0.363	22.588	0.000

Figura 12: Resultados de la prueba de correlación

Donde la única correlación no negada debido al p-value menor a 0.05 es la r15 la cual representa la dependencia de las partículas PM2.5 con un p-value = 0.171 con respecto de esas variables debido al nivel de significancia, por lo que en nuestro modelo la selección de variables estará dada por la selección de las partículas PM2.5, y dado que las demás variables parecen independientes entre si, nuestro modelo se ve reducido a añadir las variable con mayor correlación de las restantes, la cuales son: Radiación Solar (SR) con -0.040. y Humedad Relativa (RH) con 0.66

Por lo que procedemos a la realización de nuestra combinación lineal para nuestro modelo multivariado de la siguiente manera:

$$Y = 29{,}59838 - 0{,}2792167x_1 + 37{,}31402x_2 + 0{,}193375x_3 \tag{1}$$

Donde x_1 es representada por la variable de humedad relativa (RH), x_2 por medio radiación solar (SR), mientras que x_3 por la concentración de partículas PM2.5.

4.1.3. Evaluación y validación del modelo regresión multivariado

Donde x_1 Representa a la variable SR mientras que x_2 representa a la humedad relativa y x_3 la concentración de partículas PM2.5. Con esto formamos nuestro modelo y lo evaluamos en torno a encontrar su coeficiente de determinación.

```
Call:
lm(formula = Y ~ x1 + x2 + x5)

Residuals:
     Min       1Q   Median       3Q      Max
-35.455  -8.877  -1.620   7.945  62.555

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 29.59838    0.81598  36.273  < 2e-16 ***
x1          -0.27922    0.01429 -19.538  < 2e-16 ***
x2          37.31402    1.22038  30.576  < 2e-16 ***
x5           0.19337    0.03443   5.616 2.11e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14 on 3353 degrees of freedom
Multiple R-squared:  0.4971,    Adjusted R-squared:  0.4967
F-statistic:  1105 on 3 and 3353 DF,  p-value: < 2.2e-16
```

Figura 13: Summary modelo de regresión lineal multivariado.

Los signif. codes indican el nivel de significancia estadística de cada coeficiente. En este caso, todas las variables utilizadas son altamente significativos por lo que se etiquetan con (***), lo que significa que son importantes para el modelo.

El coeficiente de determinación (R-cuadrado) es 0.49667, lo que significa que aproximadamente el 49.67 % de la variabilidad en el nivel de ozono (O3) se explica por las variables independientes x1 x2 y x3 en el modelo. Esto indica que el modelo tiene una capacidad moderada para explicar la variación en el ozono. En la Figura 18 se muestra gráficamente la dependencia entre la variable x_2 (Radiación solar) con nuestra variable dependiente y (Ozono) con una pendiente positiva. Esto nos dice que un aumento en la radiación solar impacta generalmente en un aumento en la cantidad de ppb de ozono presente en el aire.

El estadístico F es 1105, y su valor p es prácticamente cero (p-value menor a 2.2e-16). Esto indica que el modelo en su conjunto es altamente significativo, lo que sugiere que al menos una de las variables independientes tiene un efecto significativo en el ozono.

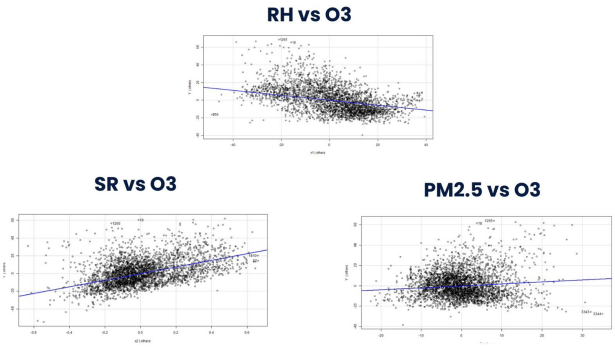


Figura 14: Gráfica de regresión entre x_1 (Radiación Solar), x_2 (Humedad Relativa) x_3(PM2.5) y y (Ozono).

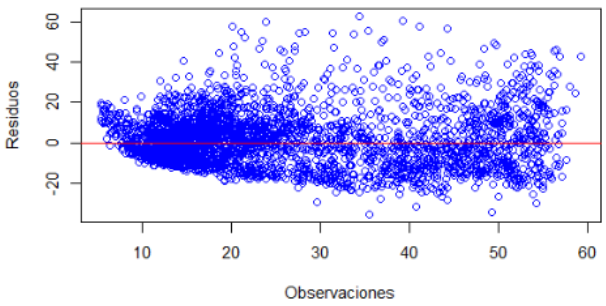
La siguiente prueba para validar el modelo de regresión es el análisis de residuales.

```
        One Sample t-test

data:  reg$residuals
t = -7.6554e-15, df = 3356, p-value = 1
alternative hypothesis: true mean is not equ
95 percent confidence interval:
 -0.4735232  0.4735232
sample estimates:
    mean of x
-1.848855e-15


        Anderson-Darling normality test

data:  residuos
A = 18.869, p-value < 2.2e-16
```

Figura 15: Análisis de residuales

Como resultado obtenemos evidencia para rechazar la hipótesis nula con un p-value menor a 0.05 por lo que los residuos no pertenecen a una distribución normal

Por último se aplica una prueba de homocedasticidad para asegurar una varianza constante.




```
        studentized Breusch-Pagan test

data:  reg
BP = 461.21, df = 3, p-value < 2.2e-16
```

Figura 16: Análisis de homocedasticidad

En nuestro caso hay evidencia suficiente para rechazar la hipótesis nula con un p-value mucho menor a 0.05. Por lo que no se presenta una varianza constante en los residuos.

A pesar de las pruebas estadísticas de normalidad, en las cuales esperábamos resultados deficientes por la naturaleza de los datos llegando al 49.67 % respecto al coeficiente de determinación. El cual muestra una gran significancia y superioridad en comparativa con un análisis general. Tomando en cuenta los coeficientes significativos de las variables, nuestro coeficiente de determinación, y nuestro p value menor a 0.05.

En conclusión la radiación solar (SR), es predictora superior en la presencia de Ozono(O3) y este a su vez tiene una correlación especial en esta época del año con la permanencia de las partículas PM2.5 Y PM10, las cuales son independientes a las otras variables.

4.2. Análisis Discriminante

El análisis discriminante es una técnica estadística que es utilizada en un conjunto con variables independientes para predecir si la variable dependiente es perteneciente a un grupo o categoría específica. Encuentra una combinación lineal de variables independientes que describen a la categoría a predecir para maximizar la separación entre los grupos categóricos.

4.2.1. Aplicación del análisis en el modelo propuesto

Para comprender mejor los factores que influyen en la percepción de la calidad del aire, se incorporó una variable categórica denominada CA (Calidad del Aire) en el conjunto de datos. Esta variable se ha definido como la variable dependiente y divide los días en dos categorías en función de la calidad del aire percibida:

- **0 = Buen día**
- **1 = Mal día**

El objetivo principal de este análisis es identificar y cuantificar la variable que ejerce la mayor influencia en la percepción de la calidad del aire en Monterrey durante el verano y confirmar la relación con las variables predichas anteriormente en el modelo de regresión lineal multivariado. Para alcanzar este objetivo, se aplicó un análisis de discriminante, una técnica estadística poderosa que permite determinar qué variables independientes (características del aire) tienen el mayor poder discriminatorio entre los días calificados como buenos y malos en términos de calidad del aire con las restricciones declaradas anteriormente.

A través de la realización del análisis de discriminante, los resultados fueron los siguientes en torno a la única función discriminante que se encontró:

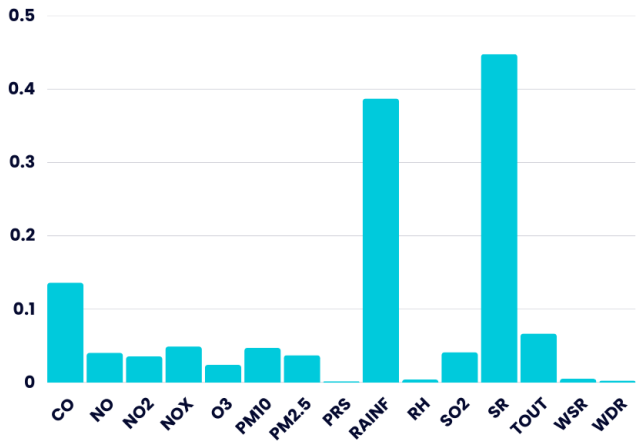


Figura 17: Influencia absoluta de variables sobre la función discriminante

La función discriminante se utiliza para predecir la calidad del aire en función de estas variables independientes. En este contexto, se observa que la radiación solar (SR) tiene una influencia significativamente negativa en la función discriminante, con un coeficiente de -0.447290404. Esto sugiere que un aumento en la radiación solar se asocia fuertemente con

Variable	Valor LD1
CO	-0.135590898
NO	-0.039981126
NO2	-0.035327332
NOX	0.048770277
O3	0.023898846
PM10	0.046894820
PM2.5	0.036930220
PRS	-0.001153171
RAINF	0.386476751
RH	-0.003755444
SO2	0.040917269
SR	-0.447290404
TOUT	-0.066253367
WSR	0.004889659
WDR	0.002338793

Cuadro 8: Valores de los coeficientes de la función discriminante LD1

una disminución en la probabilidad de que la calidad del aire sea buena.

Este resultado plantea cuestiones interesantes sobre la relación entre la radiación solar y la calidad del aire. Aunque pueda parecer contraintuitivo a simple vista, la influencia negativa de la radiación solar podría deberse a una serie de factores, como la mayor formación de contaminantes atmosféricos en condiciones soleadas o la interacción entre la radiación solar y otros contaminantes presentes en la atmósfera.

4.2.2. Predicciones y validación del modelo de la función discriminante

Y dada la influencia de dicha variable así como la de los coeficientes con un mayor peso, se realizó una predicción en torno a observar el poder de predicción así como el accuracy entre otras métricas importantes para el análisis de nuestro modelo.

	0	**1**
1	0.99891727	0.001082729
2	0.99414061	0.005859388
3	0.99781523	0.002184766
4	0.99647919	0.003520811
5	0.03608167	0.963918329
6	0.75436402	0.245635982

Cuadro 9: Predicciones del análisis de discriminante en torno a un buen o mal día

Esto nos da una predicción en torno a conocer que día sera bueno o malo y la calidad de porcentaje de predicción que tenemos en torno a determinarlo con esa exactitud.

Y dadas las observaciones reales:

- 0
- 0
- 0
- 0
- 1
- 0

Podemos confirmar de manera asertiva la predicción de nuestro modelo en torno a que clasificó correctamente todas las observaciones anteriores. Por lo que proseguiremos al calculo de nuestras estadísticas para conocer la efectividad real del modelo de clasificación por análisis de discriminante.

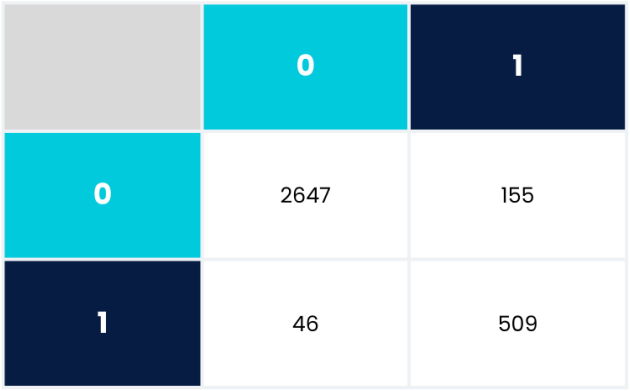


Figura 18: Matriz de confusión de la predicción de nuestro modelo con función discriminante

El análisis discriminante realizado revela resultados significativos en relación con la influencia de las variables predictoras en la calidad del aire, con un enfoque particular en dos de ellas: la radiación solar (SR) y la precipitación (RAINF), considerando los coeficientes proporcionados y los resultados de la matriz de confusión.

En primer lugar, la influencia de la radiación solar (SR) se destaca por su coeficiente negativo significativo de -0.447290404 en el modelo. Esta ponderación negativa sugiere que un aumento en la radiación solar se asocia fuertemente con una disminución en la probabilidad de que se experimente un buen día en términos de calidad del aire. Este hallazgo puede parecer contraintuitivo a simple vista, pero plantea la hipótesis de que la radiación solar podría tener un impacto negativo al promover la formación de contaminantes atmosféricos o interacciones complejas en la atmósfera que afectan negativamente la calidad del aire.

Los resultados de la matriz de confusión son alentadores, con una alta precisión del 76.6 %, un alto recall del 91.7 % y un sólido F1-Score del 83.5 %. Pero sobretodo con un accuracy equivalente a un 94.2 %. Esto sugiere que el modelo de análisis discriminante es capaz de identificar de manera efectiva los días de buena y mala calidad del aire en función de las

variables predictoras, incluyendo la influencia significativa de SR y RAINF.

En resumen, este análisis resalta la importancia de considerar la radiación solar y la precipitación al evaluar y predecir la calidad del aire. Además, demuestra que el modelo de análisis discriminante es un valioso recurso para la toma de decisiones en la gestión ambiental y la mejora de la calidad del aire en ciertas estaciones del año en especial, y como afrontar dichas medidas con ello.

Y creando una mancuerna con el análisis de regresión lineal multivariado, observamos que si la radiacion solar tiene influencia sobre el ozono como variable predictora por medio de este modelo, este a su vez lo tiene con respecto de la determinación de la calidad del aire en torno a clasificar con un gran accuracy del 94 % de certeza estos días por medio de esta variable fundamental en las estaciones de verano en especial.

5. Resultados

Regresión Lineal Multivariable

- Las variables independientes seleccionadas fueron la radiación solar (SR), la humedad relativa (RH) y la concentración de partículas suspendidas PM2.5,
- El modelo obtenido fue: $O3 = 29{,}59838 + 37{,}31402 * SR - 0{,}2792167 * RH + 0{,}193375 * PM2{,}5$
- El coeficiente de determinación (R-cuadrado) del modelo fue del 49.67 %, lo que indica que aproximadamente el 50 % de la variabilidad en los niveles de ozono se explica por las variables independientes incluidas en el modelo.
- La radiación solar (SR) mostró una correlación positiva significativa con el ozono, lo que sugiere que un aumento en la radiación solar se relaciona con niveles más altos de ozono.
- La concentración de partículas PM2.5 también mostró una correlación positiva significativa con el ozono, lo que indica que la presencia de estas partículas finas en el aire puede estar relacionada con mayores concentraciones de ozono.
- En períodos de verano se tienen altos niveles de Ozono y PM2.5, provocando mala calidad del aire entre los meses de Julio y Agosto en Monterrey.

Análisis Discriminante

- Se realizó un análisis discriminante para predecir la calidad del aire en Monterrey durante el verano, dividiendo los días en dos categorías: buenos días (0) y malos días (1) en función de la percepción de la calidad del aire.
- La variable que tuvo el mayor poder discriminatorio en la percepción de la calidad del aire fue la radiación solar (SR), con una influencia significativamente negativa. Esto significa que un aumento en la radiación solar se asocia con una disminución en la probabilidad de que la calidad del aire sea buena.

- Este resultado puede ser contraintuitivo, pero sugiere que la radiación solar puede desempeñar un papel en la formación de contaminantes atmosféricos o interactuar con otros contaminantes para influir en la percepción de la calidad del aire.

Contaminantes y Calidad del Aire

- Los contaminantes PM10 y PM2.5 indican altos rasgos de causalidad en la mala calidad del aire a nivel Estatal.
- Las variables de PM10, PM2.5, O3, pero sobretodo SR la cual tiene una relación estrecha con el ozono por medio de los análisis de regresión múltiple y discriminante, contribuyen a la mala calidad del aire específicamente en los periodos de verano entre Junio y Agosto.

Con todos estos resultados se puede concluir que los periodos de verano tienen una presentación particular de PMs, con una alta correlación entre el Ozono y Radiación solar. A través de una comparación con otras épocas del año se encontró que solo en verano se puede ver que estas variables impactan con esta naturaleza alta y directamente a la calidad del aire. En un análisis más profundo y con más tiempo se podría hacer la comparación entre periodos y conseguir más certeza si estas variables causan mala calidad del aire de manera general en todo tiempo y estación de Nuevo León.

6. Discusión y Conclusión

En este reporte, se realizó un análisis exhaustivo de la calidad del aire en Monterrey durante los meses de verano, utilizando técnicas estadísticas como la regresión lineal multivariable y el análisis discriminante. A continuación, se discuten los principales hallazgos y sus implicaciones:

Episodios de Mala Calidad del Aire: La detección de valores de PM10 y PM2.5 por encima de los límites establecidos en las primeras horas del 1 de enero de 2022 sugiere la posibilidad de episodios de mala calidad del aire en la región. Estos eventos pueden estar relacionados con diversos factores, como condiciones climáticas adversas, actividades industriales o cambios en la atmósfera. Se recomienda investigar más a fondo estas relaciones para comprender las causas detrás de estos episodios y tomar medidas preventivas. Lo mismo para el análisis de causalidad de la mala calidad del aire, se tienen altos rasgos de que se debe a las variables PM pero es posible profundizar más.

Regresión Lineal Multivariable: El modelo de regresión lineal multivariable reveló que la radiación solar (SR), la humedad relativa (RH) y la concentración de partículas PM2.5 son factores significativos que influyen en la concentración de ozono (O3) durante el verano en Monterrey. Estos resultados enfatizan la importancia de considerar estas variables en la gestión de la calidad del aire y la toma de decisiones en salud pública. La relación positiva entre la radiación solar y el ozono indica que un aumento en la radiación solar se asocia con mayores niveles de ozono.

Análisis Discriminante: El análisis discriminante reveló un hallazgo contraintuitivo: la radiación solar (SR) tuvo una influencia negativa en la percepción de la calidad del aire. Esto

sugiere que la radiación solar podría estar relacionada con episodios de mala calidad del aire debido a interacciones complejas con otros contaminantes atmosféricos. Esta relación inusual requiere una investigación más profunda para comprender sus implicaciones y tomar medidas adecuadas.

La radiación solar (SR) y la concentración de partículas PM2.5 emergieron como factores clave que influyen en la calidad del aire durante el verano en Monterrey. Estos factores deben ser considerados en la gestión ambiental y la salud pública para mitigar los efectos adversos en la población. La influencia negativa de la radiación solar en la percepción de la calidad del aire plantea preguntas intrigantes y destaca la necesidad de investigar más a fondo las complejas interacciones entre los contaminantes atmosféricos y los factores climáticos en la región.

En resumen, este reporte proporciona información valiosa sobre la calidad del aire en Monterrey durante los meses de verano, identificando factores clave que influyen en los niveles de ozono y la percepción de la calidad del aire. Estos resultados pueden ser fundamentales para la toma de decisiones en gestión ambiental y salud pública en la región de Monterrey. Se enfatiza la importancia de una investigación continua para comprender mejor las causas detrás de los episodios de mala calidad del aire y tomar medidas preventivas adecuadas.

La importancia de analizar la calidad del aire en Nuevo León radica en su impacto directo en la salud de sus habitantes. El descubrimiento de altos niveles de contaminación del aire durante el verano enfatiza la urgencia de esta vigilancia. La exposición prolongada a la contaminación del aire puede tener efectos perjudiciales en el sistema respiratorio y cardiovascular de las personas, aumentando el riesgo de enfermedades respiratorias, cardiovasculares y otros problemas de salud. Por lo tanto, este análisis resulta invaluable al proporcionar información esencial para tomar medidas preventivas y mitigar los riesgos, mejorando la calidad de vida de la población y promoviendo un entorno más saludable en Nuevo León.

7. Anexo de códigos en R y Python

Los archivos de los códigos utilizados para el presente reporte los puede encontrar en el siguiente link:
https://drive.google.com/drive/folders/1QGs0-wlG5TGT1IrKbGZSHKoNnRWerguT?usp=sharing

Referencias

[1] Contaminación del aire ambiental exterior y en la vivienda: Preguntas frecuentes - OPS/OMS — Organización Panamericana de la Salud. (2018). Paho.org. https://www.paho.org/es/temas/calidad-aire-salud/contaminacion-aire-ambiental-exterior-vivienda-preguntas-frecuentes

[2] Endara, A. D. L. M. G., Heinert, M. E. J., Solórzano, H. X. P. (2020). Contaminación del agua y aire por agentes químicos. RECIMUNDO, 4(4), 79-93.

[3] WHO. (2015). Children's health and the environment: a global perspective. A resource guide for the health sector. Geneva. CONTAMINACIÓN DEL AGUA Y AIRE POR AGENTES QUÍMICOS.

[4] Horrillo, C., Matatagui, D., Marín, P., Navarro, E., López-Sánchez, J., Peña, Á. (2022). Sensor químico resistivo para la detección de NO2.

[5] Engelking, P. (2009). Pollution. Microsoft® Encarta® (2009) (DVD).

[6] Mateos, A. C., Amarillo, A. C., Tavera Busso, I., Gonzalez, C. M. (2018). Evaluación espacial y temporal de la contaminación por SO2, NO2, O3 y CO en la ciudad de Córdoba.

[7] DOF - Diario Oficial de la Federación. (2021). Dof.gob.mx. https://www.dof.gob.mx/notadetalle.php?codigo=5633855fecha=27/10/2021gsc.tab=0

[8] DOF - Diario Oficial de la Federación. (2019). Dof.gob.mx. https://www.dof.gob.mx/notadetalle.php?codigo=5568395fecha=20/08/2019gsc.tab=0

[9] DOF - Diario Oficial de la Federación. (2020). Dof.gob.mx. https://www.dof.gob.mx/notadetalle.php?codigo=5601281fecha=25/09/2020gsc.tab=0

[10] DOF - Diario Oficial de la Federación. (2020). Dof.gob.mx. https://www.dof.gob.mx/notadetalle.php?codigo=5601282fecha=25/09/2020gsc.tab=0

[11] DOF - Diario Oficial de la Federación. (2021). Dof.gob.mx. https://dof.gob.mx/notadetalle.php?codigo=5633854fecha=27/10/2021gsc.tab=0

[12] World. (2019, July 30). Contaminación atmosférica. Who.int; World Health Organization: WHO. https://www.who.int/es/health-topics/air-pollutiontab=tab1

[13] Gobierno de Nuevo León. (Agosto 23, 3023). SISTEMA INTEGRAL DE MONITOREO AMBIENTAL [SIMA] NUEVO LEÓN. http://aire.nl.gob.mx/index.html

[14] World. (2022, December 19). Contaminación del aire ambiente (exterior). Who.int; World Health Organization: WHO. https://www.who.int/es/news-room/fact-sheets/detail/ambient-28outdoor29-air-quality-and-health

[15] ¿Cómo se mide la calidad del aire? (2022, March 11). Fundación Aquae. https://www.fundacionaquae.org/wiki/como-se-mide-calidad-aire/

[16] IBM. (Enero 03, 2023). Regresión lineal múltiple. https://www.ibm.com/docs/es/cognos-analytics/11.1.0?topic=tests-multiple-linear-regression

[17] Análisis Factorial. (2023) Aplicación de métodos multivariados en ciencia de datos. Recuperado de https://experiencia21.tec.mx/courses/423027/pages/1-dot-3-analisis-factorial?moduleitemid=24895867

[18] Análisis Factorial (2016). Universidad de Antofagasta. Recuperado de https://intranetua.uantof.cl/facultades/csbasicas/Matematicas/academicos/emartinez/economia/fac

[19] stackoverflow. (Octubre, 2020). Plot linear model in 3d with Matplotlib. https://stackoverflow.com/questions/26431800/plot-linear-model-in-3d-with-matplotlib