Tecnológico de Monterrey

Materia:
Análisis de Métodos Multivariados para Ciencia de Datos

Reto - Etapa 1

Profesores:
Blanca Rosa Ruiz Hernández
Monica Guadalupe Elizondo Amaya

Alumnos:
Gian Marco Innocenti A00834310
Andrés Villarreal González A00833915
Andrea Hernandez A00835225

28 de septiembre de 2023, Monterrey, N.L.

PARTE I. Conociendo el Negocio

Introducción

La calidad del aire es un aspecto crítico para la salud humana y el medio ambiente. La contaminación del aire puede tener efectos devastadores en la salud de las personas y en el equilibrio ecológico. En este contexto, se ha realizado una revisión de la literatura para comprender la importancia del análisis de la calidad del aire, identificar los principales contaminantes y factores involucrados, y conocer las normativas internacionales que regulan la calidad del aire.

Resumen de la revisión de bibliografía

El análisis de la calidad del aire es esencial debido a su impacto en la salud y el medio ambiente. La contaminación del aire puede causar graves problemas de salud, como enfermedades respiratorias y cardiovasculares, y contribuir al cambio climático y dañar los ecosistemas. En este análisis, se miden diversos contaminantes comunes, como el ozono, el monóxido de carbono, el dióxido de azufre, el dióxido de nitrógeno y las partículas en suspensión, considerando factores como las condiciones meteorológicas y la ubicación de las fuentes de emisión.

La Organización Mundial de la Salud (OMS) establece directrices globales para la calidad del aire, centrándose en contaminantes clave como las partículas finas, el ozono, el dióxido de nitrógeno, el dióxido de azufre y el monóxido de carbono, debido a su daño potencial para la salud. La calidad del aire se mide mediante el Índice de Calidad del Aire (AQI), que utiliza una escala de 0 a 500 y monitores especializados. Para mejorar la calidad del aire, se pueden reducir las emisiones de contaminantes, mejorar la ventilación y utilizar sistemas de limpieza y filtración de aire. La OMS establece normas oficiales para los niveles de calidad del aire en todo el mundo, pero medir la calidad del aire puede ser un desafío en países en desarrollo debido a la falta de acceso a tecnología avanzada y sistemas de alerta.

Descripción del problema específico (preguntas de investigación u objetivo del proyecto)

El problema específico que abordaremos se refiere a la calidad del aire en Monterrey durante los años 2020 y 2021, con un enfoque en comprender si hubo una diferencia significativa en la calidad del aire entre estos dos años, particularmente debido a la pandemia de COVID-19 que afectó a 2020. Para resolver este problema, se plantean las siguientes preguntas de investigación:

- ¿Hubo una variación significativa en la concentración de los contaminantes clave en el aire de Monterrey entre 2020 y 2021?
- ¿Cuáles fueron las condiciones meteorológicas y las fuentes de emisión relevantes en ambos años y cómo pueden haber influido en la calidad del aire?
-

Objetivos (se definen con base en las preguntas de investigación u objetivo del proyecto)

- Evaluar y comparar las concentraciones de contaminantes clave en el aire de Monterrey durante los años 2020 y 2021.
- Analizar las condiciones meteorológicas y las fuentes de emisión en ambos años para comprender su influencia en la calidad del aire.
- Identificar cualquier cambio significativo en la calidad del aire que pueda estar relacionado con la pandemia de COVID-19 y otras variables ambientales.

Justificación de los objetivos

Estos objetivos son fundamentales debido a la importancia de la calidad del aire para la salud pública y el medio ambiente. Comprender si hubo diferencias significativas en la calidad del aire entre 2020 y 2021 puede ayudar a evaluar el impacto de la pandemia y las políticas ambientales en la región. Además, identificar las condiciones meteorológicas y las fuentes de emisión relevantes es esencial para tomar medidas efectivas para mejorar la calidad del aire en el futuro.

Descripción de las fuentes de información (datos)

Los datos utilizados en este proyecto provienen de mediciones de calidad del aire en la Ciudad de Monterrey durante los años 2020 y 2021. Estos datos se presentan en una tabla con múltiples columnas que registran diversas variables ambientales. A continuación, se proporciona una descripción de las columnas clave en la base de datos, proporcionada por la OSF

- date: Esta columna indica la fecha y la hora de cada medición. Los datos se registran por hora y están en el formato "dd/mm/aaaa hh:mm".
- CO: Representa la concentración de monóxido de carbono (CO) en el aire.
- NO: Indica la concentración de óxido de nitrógeno (NO).
- NO2: Refleja la concentración de dióxido de nitrógeno (NO2) en el aire.
- NOX: Representa la concentración total de óxidos de nitrógeno (NOX) en el aire, que incluye tanto NO como NO2.
- O3: Indica la concentración de ozono (O3) en el aire.
- PM10: Representa la concentración de partículas suspendidas en el aire con un diámetro de 10 micrómetros o menos (PM10), medido en µg/m³.
- PM2.5: Refleja la concentración de partículas finas en el aire con un diámetro de 2.5 micrómetros o menos (PM2.5), también medido en µg/m³.
- PRS: Indica la presión atmosférica en la ubicación de la medición.
- RAINF: Representa la cantidad de lluvia registrada en la ubicación durante la hora de la medición.
- RH: Refleja la humedad relativa del aire en porcentaje (%).
- SO2: Indica la concentración de dióxido de azufre (SO2) en el aire.
- SR: Representa la radiación solar en la ubicación de la medición.
- TOUT: Indica la temperatura ambiente en la ubicación de la medición.
- WSR: Representa la velocidad del viento en la ubicación de la medición.
- WDR: Indica la dirección del viento en grados.

Impacto social principal

Este proyecto se enfoca en analizar y comparar la calidad del aire en Monterrey durante los años 2020 y 2021, con el objetivo de comprender mejor los factores que influyen en la calidad del aire y su impacto en la salud pública y el medio ambiente.

PARTE 2. COMPRENSIÓN Y PREPARACIÓN DE LOS DATOS

Para nuestra solución propondremos un estudio donde se verá si la calidad del aire mejoró en el año 2020 donde se experimentó la pandemia de COVID 19 donde la movilidad de las personas se redujo drásticamente en comparación a un año donde la actividad se normalizo tanto en industria como en movilidad humana en el año 2021. Para esto se tomarán registros de 3 diferentes estaciones de recolección de datos en 2020 y 2021. Las estaciones escogidas fueron la Estación Centro, la Estación Noreste2, y la Estación Sureste3. También durante este análisis solo se tomarán en cuenta los contaminantes que se incluyen en el Índice de Aire y Salud y las variables meteorológicas proporcionadas por la estación de recolección.

Indica el número de variables y el tamaño del dataset

Para esta problemática utilizaremos un dataset con 17 variables y 52,462 registros. Por lo que el tamaño del dataset es de 52,462 x 17.

Describe las variables que consideramos importantes y su relevancia para cumplir con el objetivo del proyecto.

Para nuestra solución utilizaremos las siguiente variables:

- **date: Variable Categórica**
- **Año: Variable Categórica**
- **Estacion: Variable Categórica**
  - Esta variable representa la estación donde fue recolectada la información.
- **CO: Variable Numérica**
  - Esta variable representa el nivel del contaminante CO en el índice de calidad del aire y salud medido por el promedio móvil ponderado medido en la estación durante una hora. Es importante porque queremos descubrir el impacto de este contaminante en la calidad del aire.
- **NO2: Variable Numérica**
  - Esta variable representa el nivel del contaminante NO2 en el índice de calidad del aire y salud medido por el promedio móvil ponderado medido en la estación durante una hora. Es importante porque queremos descubrir el impacto de este contaminante en la calidad del aire.
- **O3: Variable Numérica**
  - Esta variable representa el nivel del contaminante O3 en el índice de calidad del aire y salud medido por el promedio móvil ponderado medido en la estación durante una hora. Es importante porque queremos descubrir el impacto de este contaminante en la calidad del aire.
- **PM10: Variable Numérica**
  - Esta variable representa el nivel del contaminante PM10 en el índice de calidad del aire y salud medido por el promedio móvil ponderado medido en la estación durante una hora. Es importante porque queremos descubrir el impacto de este contaminante en la calidad del aire.
- **PM2.5: Variable Numérica**
  - Esta variable representa el nivel del contaminante PM2.5 en el índice de calidad del aire y salud medido por el promedio móvil ponderado medido en la estación durante una hora. Es importante porque queremos descubrir el impacto de este contaminante en la calidad del aire.

- **SO2: Variable Numérica**
  - Esta variable representa el nivel del contaminante SO2 en el índice de calidad del aire y salud medido por el promedio móvil ponderado medido en la estación durante una hora. Es importante porque queremos descubrir el impacto de este contaminante en la calidad del aire.
- **CO3Concentración: Variable Numérica**
  - Esta variable representa el nivel del contaminante NO2 en el índice de calidad del aire y salud medido por la concentración medida en la estación durante una hora. Es importante porque queremos descubrir el impacto de este contaminante en la calidad del aire.
- **PRS: Variable Numérica**
  - Esta variable representa la presión atmosférica en la estación durante una hora. Es importante para el modelo ya que queremos descubrir que tanto este factor meteorológico impactó con la calidad del aire.
- **RAINF: Variable Numérica**
  - Esta variable representa la cantidad de precipitación en la estación durante una hora. Es importante para el modelo ya que queremos descubrir que tanto este factor meteorológico impactó con la calidad del aire.
- **RH: Variable Numérica**
  - Esta variable representa el porcentaje de humedad relativa en la estación durante una hora.Es importante para el modelo ya que queremos descubrir que tanto este factor meteorológico impactó con la calidad del aire.
- **SR: Variable Numérica**
  - .Esta variable representa la radiación solar en la estación durante una hora.Es importante para el modelo ya que queremos descubrir que tanto este factor meteorológico impactó con la calidad del aire.
- **TOUT: Variable Numérica**
  - Esta variable representa la temperatura promedio en la estación durante una hora.Es importante para el modelo ya que queremos descubrir que tanto este factor meteorológico impactó con la calidad del aire.
- **WSR: Variable Numérica**
  - Esta variable representa la velocidad del viento en la estación durante una hora.Es importante para el modelo ya que queremos descubrir que tanto este factor meteorológico impactó con la calidad del aire.
- **WDR: Variable Numérica**
  - Esta variable representa la dirección del viento en la estación durante una hora. Es importante para el modelo ya que queremos descubrir que tanto este factor meteorológico impactó con la calidad del aire.
- **Objetivo: Variable Numérica**
  - El valor del Índice de Calidad de Aire más alto medido cada hora en la estación. Es importante porque es la variable que trataremos de predecir.

Los diferentes valores que las siguientes variables pueden tomar para los diferentes años y estaciones se pueden ver en la siguiente Tabla1.

	Dataset	date	CO	NO2	O3	PM10	PM2.5	PRS	RAINF	RH	SO2	SR	TOUT	WSR	WDR
0	Centro_2020	[2020-01-01 00:00:00, 2020-01-01 01:00:00, 202...	[0.34-6.36]	[0.0-35.5]	[1.0-148.0]	[2.0-720.0]	[2.25-112.84]	[689.4-724.8]	[0.0-0.0]	[1.0-92.0]	[0.7-33.5]	[0.0-0.282]	[3.15-37.93]	[0.7-23.4]	[1.0-360.0]
1	Noreste2_2020	[2020-01-01 00:00:00, 2020-01-01 01:00:00, 202...	[0.15-16.93]	[0.0-42.7]	[1.0-125.0]	[2.0-763.0]	[2.09-140.45]	[692.1-715.4]	[0.0-0.48]	[1.0-90.0]	[0.5-192.1]	[0.0-0.878]	[1.89-38.03]	[0.6-29.2]	[1.0-359.0]
2	Sureste3_2020	[2020-01-01 00:00:00, 2020-01-01 01:00:00, 202...	[0.05-17.71]	[0.3-75.3]	[2.0-101.0]	[3.0-711.0]	[2.0-158.0]	[706.9-747.6]	[0.0-4.39]	[1.0-94.0]	[0.5-90.8]	[0.0-0.889]	[2.38-40.2]	[0.6-11.6]	[1.0-360.0]
3	Centro_2021	[2021-01-01 00:00:00, 2021-01-01 01:00:00, 202...	[0.05-14.6]	[0.1-90.4]	[1.0-135.0]	[2.0-634.0]	[2.17-144.66]	[700.7-722.7]	[0.0-0.0]	[1.0-92.0]	[0.7-61.9]	[0.0-1.007]	[-4.75-40.26]	[0.9-21.9]	[1.0-360.0]
4	Noreste2_2021	[2021-01-01 00:00:00, 2021-01-01 01:00:00, 202...	[0.05-22.38]	[0.0-80.1]	[2.0-139.0]	[2.0-718.0]	[2.05-160.57]	[680.0-713.3]	[0.0-22.38]	[1.0-90.0]	[0.5-58.5]	[0.0-0.774]	[-6.3-41.13]	[0.1-62.1]	[1.0-360.0]
5	Sureste3_2021	[2021-01-01 00:00:00, 2021-01-01 01:00:00, 202...	[0.11-4.15]	[0.0-77.9]	[1.0-129.0]	[2.0-613.0]	[2.0-131.0]	[680.0-745.7]	[0.0-2.1]	[3.0-94.0]	[0.5-175.5]	[0.0-0.803]	[-3.22-42.49]	[0.1-42.0]	[1.0-360.0]

Tabla 1. Posibles Valores Para las Variables

Resume el trabajo realizado en la preparación de datos: necesidad de unión de bases, limpieza de los datos, imputación (de haber sido necesaria), porcentaje de datos faltantes, datos atípicos, duplicados, etc

Durante este proceso primero identificamos los datos faltantes para cada estación y su respectiva columna. El desglose de los datos faltantes se puede ver en la siguiente Tabla2.

	Dataset	date	CO	NO2	O3	PM10	PM2.5	PRS	RAINF	RH	SO2	SR	TOUT	WSR	WDR	Total_Estacion	Porcentaje
0	Centro_2020	0	124	967	1450	380	1845	132	123	126	447	54	135	131	133	6047	0.021553
1	Noreste2_2020	0	108	3496	1784	253	1438	116	108	147	1325	33	109	781	1726	11424	0.040719
2	Sureste3_2020	0	918	136	2505	510	1456	128	126	137	1109	15	126	131	141	7438	0.026511
3	Centro_2021	0	654	1294	1064	288	834	208	198	200	1483	131	200	240	206	7000	0.02495
4	Noreste2_2021	0	315	592	517	400	481	899	313	598	1869	74	657	764	1480	8959	0.031933
5	Sureste3_2021	0	481	770	517	693	1311	505	484	727	1271	40	483	491	493	8266	0.029463
6	Total	0	2600	7255	7837	2524	7365	1988	1352	1935	7504	347	1710	2538	4179	49134	0.175128

Tabla 2. Valores Faltantes para cada Estación y sus Respectivas Variables

En esta tabla se puede apreciar que la estación con más datos faltantes es la Estación Noreste 2. También que en total en todo el dataset faltan 49134 datos lo que representa el 17% del dataset.

El siguiente paso de este proceso es identificar datos atípicos. Esto se hizo calculando el valor de Z con respecto a cada estación y su año. En otras palabras cada valor será evaluado sólo con los datos en su columna de su estación y de su año. Si se calculara en respecto a toda la población estaría sesgado por los datos recolectados en un lugar diferente o en otro año. La metodología implementada fue calcular Z.

$$i = columna$$

$$j = estación$$

$$Z = (x - \mu_{ij}) / \sigma_{ij}$$

Posteriormente un dato sería considerado como atípico si su valor Z es mayor a 2. Lo cual significa que ese valor se encuentra a más de dos desviaciones estándar de la media para su respectiva estación y año. El desglose de los datos atípicos se puede ver en la Tabla 3.

	Dataset	CO	NO2	O3	PM10	PM2.5	PRS	RAINF	RH	SO2	SR	TOUT	WSR	WDR	Total_Estacion	Porcentaje
0	Centro_2020	505	587	389	336	396	513	0	199	395	134	329	312	547	4642	0.016545
1	Noreste2_2020	336	80	373	286	380	487	94	247	206	632	347	156	830	4454	0.015875
2	Sureste3_2020	296	545	476	339	388	512	48	246	244	633	365	322	0	4414	0.015733
3	Centro_2021	777	461	391	321	438	471	0	209	304	691	370	336	627	5396	0.019233
4	Noreste2_2021	321	350	463	325	379	480	1	265	763	10	380	181	825	4743	0.016905
5	Sureste3_2021	225	486	331	364	410	346	16	243	288	649	426	350	183	4317	0.015387
6	Total	2460	2509	2423	1971	2391	2809	159	1409	2200	2749	2217	1657	3012	27966	0.099679

Tabla 3. Valores Atípicos para cada Estación y sus Respectivas Variables

En esta tabla se puede apreciar que el 9% de los datos en el dataset son atípicos.

Después verificamos si los valores de cada variable se distribuyen normalmente excepto la variable date. Para esto se utilizó una prueba de shapiro para verificar la normalidad en respecto a su año y estación.En esta prueba se calculó el valor p para la distribución y se hará la siguiente prueba de hipótesis.

$$H_0: Los\ datos\ provienen\ de\ una\ población\ que\ sigue\ distribución\ normal$$

$$H_1: Los\ datos\ provienen\ de\ una\ población\ que\ no\ sigue\ distribución\ normal$$

$$\alpha = 0.05$$

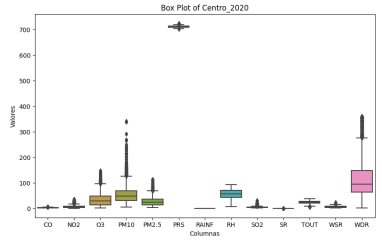
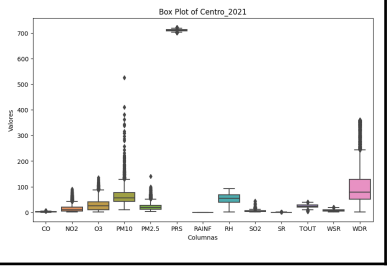
El desglose de la prueba de normalidad para las variables se puede ver en la Tabla 4 Donde se le asigna un valor de 1 cuando no se Rechaza H0 y la población sigue una distribución normal. Se asigna un 0 cuando se rechaza H0 y la población no sigue una distribución normal.

	Dataset	CO	NO2	O3	PM10	PM2.5	PRS	RAINF	RH	SO2	SR	TOUT	WSR	WDR	Total_Estacion
0	Centro_2020	1	0	0	0	0	0	1	0	0	0	0	0	0	2
1	Noreste2_2020	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	Sureste3_2020	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	Centro_2021	0	0	0	0	0	0	1	0	0	0	0	0	0	1
4	Noreste2_2021	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	Sureste3_2021	1	0	0	0	0	0	0	0	0	0	0	0	0	1

Tabla 4. Prueba de Normalidad para cada Estación y sus Respectivas Variables

En esta tabla se puede ver como solo 4 poblaciones siguen una distribución normal.

En la Tabla 5 se expondrán la gráfica de Caja de las variables con respecto a su estación y año .

Estación	Año 2020	Año 2021
Centro		

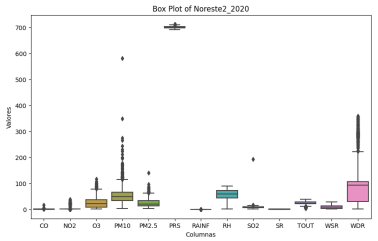
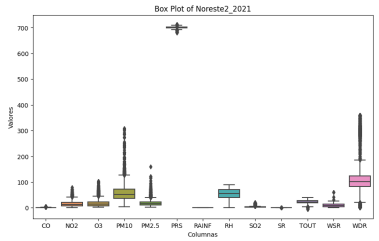
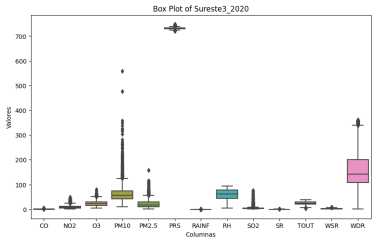
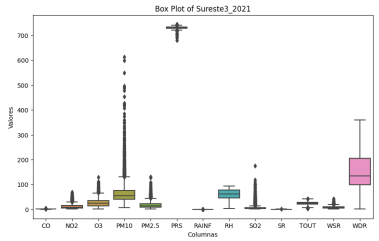
Noreste2		
Sureste3		

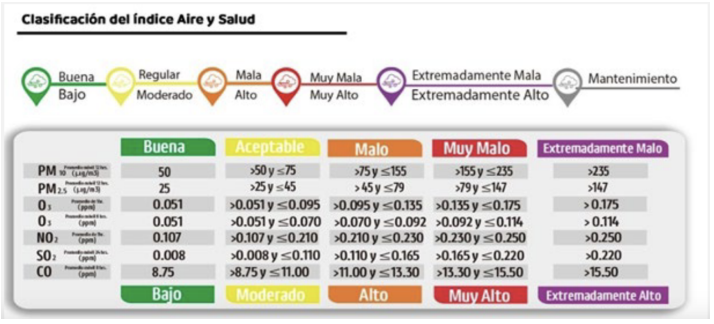
Tabla 5. Boxplot de cada estación respecto a su año

En la Tabla 5 se puede apreciar la diferencia de los valores. Y se puede observar como la alta variación de los mismos significa que hay pocas poblaciones que siguen una distribución normal.

El siguiente paso fue el tratamiento de los datos faltantes. Para esto se utilizó la siguiente metodología. Un valor faltante se reemplazará por el promedio entre el valor previo que tuvo una medición y el siguiente valor que tuvo una medición.

$$x_i = \frac{x_{i-1} + x_{i+1}}{2}$$

El siguiente paso para transformar el dataset fue estandarizar los valores de los contaminantes según su clasificación en el Índice de Aire y Salud(Imagen 1).(Norma Oficial Mexicana 2019)



	Buena	Aceptable	Malo	Muy Malo	Extremadamente Malo
PM 10	50	>50 y ≤75	>75 y ≤155	>155 y ≤235	>235
PM 2.5	25	>25 y ≤45	>45 y ≤79	>79 y ≤147	>147
O3	0.051	>0.051 y ≤0.095	>0.095 y ≤0.135	>0.135 y ≤0.175	>0.175
O3	0.051	>0.051 y ≤0.070	>0.070 y ≤0.092	>0.092 y ≤0.114	>0.114
NO2	0.107	>0.107 y ≤0.210	>0.210 y ≤0.230	>0.230 y ≤0.250	>0.250
SO2	0.008	>0.008 y ≤0.110	>0.110 y ≤0.165	>0.165 y ≤0.220	>0.220
CO	8.75	>8.75 y ≤11.00	>11.00 y ≤13.30	>13.30 y ≤15.50	>15.50
	Bajo	Moderado	Alto	Muy Alto	Extremadamente Alto

Imagen 1. Clasificación del Índice de Aire y Salud (Norma Oficial Mexicana 2019)

Para que esto pueda ser representada numéricamente las Calificaciones serán representadas de la siguiente manera:

- Buena =1
- Aceptable = 2
- Malo = 3
- Muy Malo=4
- Extremadamente Malo =5

Para lograr estos resultados el primer paso fue convertir los valores de la columna O3,SO2 y NO2 de ppb a ppm ya que esa es la unidad utilizada en el Índice de Aire y Salud. Después se calculó el promedio móvil ponderado según la Normativa de la Secretaría de Gobernación de México. Cada contaminante tiene su tratamiento específico en el cual se considera un número de horas previas diferente para el cálculo de su promedio móvil ponderado. El único Contaminante que no recibe ninguna transformación es CO. En la Tabla 6 se presentan las transformaciones necesarias para cada variable.

Contaminante	Concentración base
PM10	Concentración promedio móvil ponderado de 12 horas*
$PM_{2.5}$	
ozono (O3)	Concentración promedio móvil de 8 horas
monóxido de carbono (CO)	
dióxido de nitrógeno (NO2)	Concentración promedio horaria
ozono (O3)	
dióxido de azufre (SO2)	Concentración promedio móvil de 24 horas (como aproximación al promedio de 24 horas)

Tabla 6 Transformación de Datos necesaria para cada Contaminante (Norma Oficial Mexicana 2019)

La metodología implementada para el promedio ponderado móvil se puede ver en la Imagen 2. Cabe resaltar que para nuestra implementación no se tomarán los criterios de datos faltantes ya que estos fueron previamente rellenados con la metodología previamente explicada.

$$\bar{C} = \frac{\sum_{i=1}^{N} C_i W^{i-1}}{\sum_{i=1}^{N} W^{i-1}}$$

Donde:

$$W = \begin{cases} w \text{ si } w > 0.5 \\ 0.5 \text{ si } w \leq 0.5 \end{cases} \text{ y } w = 1 - \frac{C_{max} - C_{min}}{C_{max}}$$

$$\bar{C} = \frac{\sum_{i=1}^{12} (C_i W^{i-1})}{\sum_{i=1}^{12} (W^{i-1})}$$

$\bar{C}$ = Concentración promedio móvil ponderada.

$N = 12$

Σ = Sumatoria de datos.

C_i = Concentración promedio horaria de la hora i.

i = hora consecutiva de medición (la hora más reciente de medición es la hora 1 y la primera hora de medición en el conjunto de datos considerados en el cálculo sería la hora 12).

W = Factor de ponderación.

w = Valor del peso.

C_{max} = Concentración promedio horaria máxima en el periodo de 12 horas.

C_{min} = Concentración promedio horaria mínima en el periodo de 12 horas.

Imagen 2. Metodología para Cálculo de Promedio Móvil Ponderado (Norma Oficial Mexicana 2019)

Al tener los criterios necesarios para cada variable se clasificaron según los criterios mencionados en la Imagen 1.

El último paso en la generación del dataset es crear nuestra variable objetivo. La cual se generó tomando el valor más grande de los contaminantes para cada fila, ya que esta es la normativa implementada por Sima para asignar la calidad de aire que medio la estación en una hora. En la Imagen 3 se puede apreciar un ejemplo de como se ve el dataset que se utilizara en la investigación.

	date	CO	NO2	O3	PM10	PM2.5	PRS	RAINF	RH	SO2	SR	TOUT	WSR	WDR	O3Concentracion	Estacion	Year	Objetivo
0	2020-02-01 00:00:00	1	1.0	1.0	1	1	708.7	0.0	88.0	NaN	0.156	12.30	1.4	263.0	1.0	Centro_2020	2020	1.0
1	2020-02-01 01:00:00	1	1.0	1.0	1	1	708.1	0.0	85.0	NaN	0.155	12.41	3.1	239.0	1.0	Centro_2020	2020	1.0
2	2020-02-01 02:00:00	1	1.0	1.0	1	1	707.2	0.0	81.0	1.0	0.155	12.88	3.2	239.0	1.0	Centro_2020	2020	1.0
3	2020-02-01 03:00:00	1	1.0	1.0	1	1	707.0	0.0	77.0	1.0	0.155	13.22	3.9	258.0	1.0	Centro_2020	2020	1.0
4	2020-02-01 04:00:00	1	1.0	1.0	1	1	706.6	0.0	76.0	1.0	0.155	13.10	2.4	235.0	1.0	Centro_2020	2020	1.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
52457	2021-12-31 19:00:00	1	1.0	1.0	3	2	723.6	0.0	50.0	1.0	0.002	26.92	7.8	127.0	1.0	Sureste3_2021	2021	3.0
52458	2021-12-31 20:00:00	1	1.0	1.0	3	2	723.9	0.0	56.0	1.0	0.001	25.43	5.5	148.0	1.0	Sureste3_2021	2021	3.0
52459	2021-12-31 21:00:00	1	1.0	1.0	3	3	724.1	0.0	58.0	1.0	0.001	24.75	6.5	31.0	1.0	Sureste3_2021	2021	3.0
52460	2021-12-31 22:00:00	1	1.0	1.0	4	4	724.1	0.0	57.0	1.0	0.001	24.70	8.9	47.0	1.0	Sureste3_2021	2021	4.0
52461	2021-12-31 23:00:00	1	1.0	1.0	4	3	724.3	0.0	66.0	1.0	0.001	22.10	4.9	309.0	1.0	Sureste3_2021	2021	4.0

Imagen 3. Ejemplo del Dataset Final

Liga con acceso a base de datos limpia, código de python de tratamiento de datos y datos fuente.

https://drive.google.com/drive/folders/1fuYWXOz1Dj7pQSrnmh4mIJmnz0Udltbs?usp=sharing

Referencias

Clarity. (2023, Marzo 21). The purpose and importance of air quality monitoring. Recuperado de https://www.clarity.io/blog/what-is-air-quality-monitoring-why-is-it-important

UNEP. (2021, Noviembre 2). 5 dangerous pollutants you're breathing in every day. Recuperado de https://www.unep.org/news-and-stories/story/5-dangerous-pollutants-youre-breathing-every-day

NOAA SciJinks. (2023). ¿Cómo se mide la calidad del aire? Recuperado de [https://scijinks.gov/air-quality/]

Organización Mundial de la Salud. (2021, Septiembre 22). Nuevas directrices mundiales de la OMS sobre la calidad del aire tienen como objetivo salvar millones de vidas de la contaminación del aire. Recuperado de https://www.who.int/news/item/22-09-2021-new-who-global-air-quality-guidelines-aim-to-save-millions-of-lives-from-air-pollution

Norma Oficial Mexicana (2019)https://www.dof.gob.mx/nota_detalle.php?codigo=5579387&fecha=20/11/2019#gsc.tab=0