

# Conocimiento de la naturaleza de contaminantes que influyen en la calidad del aire y sus interrelaciones en la región La Pastora en Nuevo León

Medina, F.<sup>1</sup>, Porcayo, E.<sup>2</sup>, González K.<sup>3</sup>, Derbéz, J.<sup>4</sup>, Hernández, J.<sup>5</sup>

<sup>1</sup>A01721441, Federico Medina García Corral

<sup>2</sup>A01423285, Eduardo Emiliano Porcayo Arrieta

<sup>3</sup>A01411597, Karen González Ugalde

<sup>4</sup>A01781313, José Emilio Derbez Safie

<sup>5</sup>A01730548, Javier Hernández Arellano

21/Octubre/2022

## 1. Resumen

La contaminación del aire a través de contaminantes, sobre todo en zonas metropolitanas, como es el caso, representan una amenaza que atenta contra la salud de sus habitantes. Solamente cabe considerar que el 99 % de la población respira un aire que supera los límites de calidad del aire establecidos por la Organización Mundial de la Salud.

A raíz de la base de datos proporcionada por el curso, donde tenemos las emisiones de contaminantes por zona, aunados a sus banderas, se resolvió por utilizar a la región Sureste del estado de Nuevo León, específicamente en la estación de La Pastora, en el municipio de Guadalupe.

Como primer objetivo, se dio a la tarea de explorar la base, donde se rescata el hecho de que variables de contaminantes como el CO o SO<sub>2</sub> contaban de primera instancia con valores atípicos que influían en sus medidas de tendencia central, así que se eliminaron aquellos que no eran de nuestro interés.

En la base de datos se contaba con una gran cantidad de datos nulos, ya sea porque estos no fueron registrados en la base de datos original, o porque las mediciones se consideraban invalidas (a criterio del SIMA). Tomando esto en cuenta, se logró observar que existían 4 columnas que contaban con más del 80 % de datos nulos. Estas fueron excluidas de la base de datos. Además, como detalle de que hubiera errores que no sean claros a la simple vista, se analizó si existían valores repetidos.

El trato que se le decidió dar a los datos nulos fue el siguiente. Para períodos mayores o iguales a 36 horas (1 día y medio) donde solamente se presentan datos nulos, estos no se modifican. Para períodos menores a 36 horas donde existan datos nulos, estos son remplazados por la última observación realizada antes de que inicie la falta de datos.

De igual manera, para poder generar los modelos estadísticos, se transformaron todos los datos de las variables a la misma unidad de medida. En este caso, únicamente la variable CO se cambió a ppb, ya que se encontraba en ppm.

Para la comprensión de datos, se calcularon las medidas de tendencia central, así como el rango y la varianza de todos los datos numéricos. Además se graficaron histogramas, boxplots y gráficas de densidad, los cuales permitieron entender la distribución de las diferentes variables e identificar valores atípicos.

Como primer modelo, se escogió la Regresión Lineal Múltiple con el objetivo de encontrar alguna variable que pueda ser explicada por otras variables. Para esto, se eligió la variable que consideramos que tiene el mayor número de correlaciones "altas" o "medianas" con otras variables, siendo la variable O<sub>3</sub>. Con esto, se utilizó Minitab para encontrar la función de la variable objetivo con las variables PM2, SO<sub>2</sub>, CO, T, RH, SR, RF y WS. Después de encontrar su función, se hizo una prueba de normalidad con los residuales para encontrar si se puede explicar el comportamiento de O<sub>3</sub> con la función. Finalmente, se encontró que los residuales no se comportan de manera normal, por lo que la hipótesis de si O<sub>3</sub> es definida por las otras variables con un nivel de significancia de 0.05 fue rechazada.

Posteriormente se optó por una clasificación a través de análisis discriminante lineal. El principal objetivo de este fue clasificar cualquier hora dada en función del cumplimiento de la normativa relacionada

con el PM10. Se construyó el discriminante en función de todas las variables trabajadas (excepto el PM10), y se obtuvo un modelo de clasificación con una precisión del 75.15 %

Y, por último, se realizó un Análisis de Componentes Principales. Se llevó a cabo con la finalidad de simplificar la complejidad del espacio muestral. Cabe aclarar, que debido a que es una variable cualitativa, se excluyó a "Fecha", además de excluir a los datos nulos ya que de lo contrario no podría llevarse a cabo el análisis.

Se obtuvo que de los 11 componentes principales creados en primera instancia, con 7 PC se explica un 85.07 % de la varianza de los datos, sin embargo con 9 de ellos se obtendría una explicación el 94.07 % de la varianza observada en los datos, por lo que se concluye que con 9 componentes principales se puede tener mucha credibilidad.

Además de estos tres modelos se realizó dentro del proceso un modelo de clasificación llamado "Random Forest" el cual su objetivo era predecir si el contaminante  $PM_{10}$  iba a sobrepasar la normativa o no en cualquier hora dada. Se obtuvieron muy buenas medidas de exactitud, clasificando el 93.38 % de los casos correctamente. Este modelo puede ser utilizado en caso de que exista una falla en el sensor meteorológico o esté deshabilitado por largos períodos de tiempo.

**Keywords:** SIMA, Contaminantes, Calidad del Aire, Modelo, Clasificación, Normativa

## 2. Introducción

Cada año, dado el crecimiento de la civilización y el aumento de emisiones sucias por parte de distintas industrias, automóviles y actividades cotidianas, se deterioran críticamente las condiciones atmosféricas. Entre distintos tipos de contaminantes, como el agua, suelo, térmicos y ruido, la contaminación en el aire es una de las más peligrosas, provocando cambios climáticos y dando origen a distintas enfermedades que atentan contra la vida y la salud de los seres vivos en el planeta. De acuerdo a la Organización Mundial de la Salud (OMS), 90 % de la población mundial respira aire contaminado, mismo que causa alrededor de 7 millones de muertes al año [1].

Los efectos contra la salud en la contaminación del aire incluyen derrames cerebrales, cáncer de pulmón, y enfermedades al corazón. Además, la contaminación en el aire ocasiona problemas ambientales como el agotamiento de ozono y cambios climáticos drásticos. Debido a estas razones, un monitoreo constante y manejo adecuado de la calidad del aire en las ciudades es un tema de creciente interés y preocupación. Según datos de la OMS, los principales aspectos que se deben considerar para analizar la calidad de aire son las concentraciones de los siguientes contaminantes:

1. Materia particulada ( $PM$ )
2. Ozono ( $O_3$ )
3. Dióxido de nitrógeno ( $NO_2$ )
4. Dióxido de azufre ( $SO_2$ )

Las directrices impuestas para estos contaminantes son las siguientes: [2]

- **Materia particulada fina ( $PM_{2.5}$ )**
  - $5 \mu g/m^3$  de media anual
  - $15 \mu g/m^3$  de media diaria
- **Materia particulada gruesa ( $PM_{10}$ )**
  - $15 \mu g/m^3$  de media anual
  - $45 \mu g/m^3$  de media diaria
- **Ozono ( $O_3$ )**
  - $100 \mu g/m^3$  máximo diario, promedio móvil de 8 horas
  - $60 \mu g/m^3$  promedio móvil de 8 horas, durante temporada máxima
- **Dióxido de Nitrógeno ( $NO_2$ )**
  - $10 \mu g/m^3$  de media anual
  - $25 \mu g/m^3$  de media diaria
- **Dióxido de azufre ( $SO_2$ )**
  - $40 \mu g/m^3$  de media diaria

Entre estos contaminantes, los más dañinos para la salud son el dióxido de nitrógeno, el ozono, y las materias particuladas finas. El  $NO_2$  aumenta los síntomas de bronquitis en niños asmáticos y disminuye el desarrollo de la capacidad pulmonar. El ozono causa efectos similares, como problemas respiratorios, asma y reducción de la función pulmonar. Por último, la materia particulada fina, al ser de un tamaño tan pequeño, puede atravesar la barrera pulmonar y entrar directamente al torrente sanguíneo. La exposición continua a estos compuestos aumenta el riesgo de enfermedades cardiovasculares, así como al cáncer de pulmón. [3]

Con esto, se arroja la pregunta sobre cómo se obtienen dichas mediciones y qué variables intervienen para la mejora de la calidad del aire. Para encontrar las variables que intervendrán en la mejora de la calidad de aire, se debe de conocer las herramientas que se utilizan para identificar los contaminantes, la cantidad, la frecuencia y las principales razones por las cuales están presentes en un momento dado. Estas herramientas son:

- Inventario de emisiones de contaminantes atmosféricos
- Monitoreo Atmosférico
- Índice de la Calidad del Aire
- Programas de Gestión para Mejorar la Calidad de Aire

El inventario de emisiones de contaminantes atmosféricos proporciona información sobre la cantidad de emisiones que son liberadas al aire. El monitoreo atmosférico permite que se conozca el estado de la calidad del aire en una zona determinada en tiempo real. Este sirve para que la población pueda tomar decisiones con base en la calidad del aire en dicha zona. Para medir la calidad del aire, se implementa un sistema que utiliza como medición el ICA (Índice de la Calidad del Aire) [4] [5]. Esta medición toma en cuenta las distintas partículas que se presentan en el medio ambiente tales como las Partículas de Suspensión ( $PM_{10}$ ), Ozono Troposférico ( $O_3$ ), el dióxido de nitrógeno ( $NO_2$ ), el dióxido de azufre ( $SO_2$ ), etc. Usando este índice, se clasifica el valor de la calidad del aire entre 0 y 500, siendo entre más alto el valor, mayor es la contaminación de aire en la zona donde se detectó. La manera en que se clasifican los valores de ICA son:

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



Figura 1: Escala de colores de ICA

Por otro lado, los Programas de Gestión para Mejorar la Calidad de Aire (ProAire) establecen acciones que la población debe tomar para la mejora de la calidad de aire [6]. Estas están direccionaladas tanto al control como a la disminución de contaminantes emitidos para que la salud no se vea afectada, y el ambiente sane.

El programa ProAire, entre muchos otros, busca mejorar la salud de todos los individuos con base en los efectos que tiene la contaminación en el cuerpo. La manera en que estas organizaciones buscan cumplir con sus objetivos es mediante una serie de normas establecidas de los distintos contaminantes para la protección de la salud de la población. Estas tienen como nombre Normas Oficiales Mexicanas de la Calidad del Aire Ambiente, conocidas como NOMs, donde, de acuerdo al Gobierno de México, se encuentran vigentes las siguientes:

- NOM-025-SSA1-2014:
  - Referente a las Partículas ( $PM_{10}$ ), utilizando una base promedio de 24 horas con exposición aguda, su Valor límite Indicador con el que se evalúa es  $75\mu g/m^3$  Máximo. En el caso de una exposición crónica,  $40\mu g/m^3$  Promedio anual.
  - Referente a las Partículas ( $PM_{2.5}$ ), utilizando una base promedio de 24 horas con exposición aguda, su Valor límite Indicador con el que se evalúa es  $45\mu g/m^3$  Máximo. En el caso de una exposición crónica,  $12\mu g/m^3$  Promedio anual.
- NOM-020-SSA1-2014:
  - Para el Ozono ( $O_3$ ) El indicador promedio de una hora no debe exceder el valor de 0.095 ppm. El indicador promedio de ocho horas no debe exceder el valor de 0.070 ppm.
- NOM-022-SSA1-2010:
  - Para el Dióxido de azufre ( $SO_2$ ), el indicador promedio diario no debe exceder una vez al año el valor de 0.110 ppm. El indicador promedio anual debe ser menor o igual 0.025 ppm. El indicador promedio de ocho horas no debe exceder una vez al año el valor de 0.200 ppm.
- NOM-023-SSA1-1993:
  - Para el Dióxido de nitrógeno ( $NO_2$ ), el indicador promedio de una hora no debe exceder una vez al año el valor de 0.210 ppm.
- NOM-021-SSA1-1993:

- Para el Monóxido de carbono ( $CO$ ), el indicador promedio de ocho horas no debe exceder una vez al año el valor de 11 ppm.
- NOM-026-SSA1-1993:
  - Para el Plomo ( $Pb$ ), el indicador promedio trimestral debe ser menor o igual a  $1.5\mu g/m^3$ .

Estas normas dan una idea general sobre cómo combatir la contaminación con base en las diferentes mediciones que se presentan de cada elemento. Sin embargo, el simple hecho de hacer una medición de la calidad del aire puede ser más complicado de lo que se supone [7]. Para poder medirla de manera efectiva se requiere de una red de monitoreo extensa. También se debe tener en consideración, que dependiendo de la ubicación de la estación, pueden variar considerablemente los datos que se recolectan. Así mismo, existen factores externos que afectan los datos que se recolectarán, como son las condiciones meteorológicas o conductas sociales. Se debe considerar que los datos suelen seguir características de estacionalidad. Por ejemplo, los contaminantes causados por tráfico vehicular suelen aumentar durante la hora pico. Por esta razón, el trato que se le debe dar a los datos nulos debe tener este tipo de factores en cuenta.

Por último, es importante saber cuál es la organización que monitorea el estado del medio ambiente a quien se le debe reportar los distintos números y cambios encontrados a la hora de hacer una medición. El SIMA es el Sistema Integral de Monitoreo Ambiental que empezó a operar en noviembre de 1992 y su principal propósito es proporcionar información continua y fidedigna de los niveles de contaminación ambiental en el Área Metropolitana de Monterrey. Actualmente cuenta con 10 estaciones de monitoreo ubicadas en diferentes zonas del Área Metropolitana de Monterrey. El SIMA proveé un Índice de Calidad de Aire de Monterrey, actualizado cada hora y desglosado tanto por estaciones como por contaminantes.



Figura 2: Red de monitoreo del SIMA

### 3. Descripción de la Problemática

De acuerdo a la OMS, alrededor de 249 mil muertes prematuras fueron atribuibles a la contaminación del aire exterior y alrededor de 83 mil muertes prematuras fueron atribuibles a la contaminación del aire debido al uso de combustibles sólidos en la vivienda en las Américas en 2016 [8].

La exposición a altos niveles de contaminación del aire puede causar una variedad de resultados adversos a la salud. La contaminación del aire puede aumentar el riesgo de infecciones respiratorias, enfermedades cardíacas, accidentes cerebro vasculares y cáncer de pulmón. Tanto la exposición a corto como a largo plazo a los contaminantes del aire se ha asociado con impactos adversos en la salud. Los impactos más severos afectan a las personas que ya están enfermas. Los niños, los ancianos y los pobres son más susceptibles. Los contaminantes más nocivos para la salud, estrechamente asociados con la mortalidad prematura excesiva, son partículas finas  $PM_{2.5}$  que penetran profundamente en los conductos pulmonares [9].

Concretando esta problemática expuesta previamente, la cuál se enfocará en el peligro que implica tener niveles altos de contaminación del aire y su relación con el clima en un área específica del estado de Nuevo León. Se planea desarrollar las siguientes preguntas de investigación:

### 4. Preguntas de Investigación

- ¿Qué contaminante sobrepasa la normativa con mayor frecuencia en el aire de La Pastora?
- ¿En qué periodo de tiempo al año se observa la peor calidad del aire dentro de la zona?
- ¿Existe una relación lineal múltiple con algún contaminante al relacionarse con otros?
- ¿Se podrá clasificar si alguna variable sobrepasa la normativa utilizando algún modelo de clasificación?

Con estas preguntas, se tiene planeado tener un parteaguas hacia la interpretación de los datos de la región Sureste del estado de Nuevo León, en cuál es la forma en que se manejan, si existe una falta de control respecto a las Normas que establecen los límites, o si existe información concreta a falta de muchos datos para poder generar una conclusión confiable hacia esta problemática.

## 5. Objetivos

Teniendo esto en cuenta, el objetivo de este proyecto es detectar la relación que hay entre los contaminantes del aire y diferentes factores del clima en una estación determinada, la cual será Sureste (SE) llamada "La Pastora" ubicada en el municipio de Guadalupe, Nuevo León. Así mismo se pretende informar y concientizar a la población con datos reales para que puedan darse cuenta de la gravedad de este problema, generar un interés genuino por querer apoyar a la causa del cuidado del medio ambiente y empezar a tomar acción. Para lo anterior, se parte de una base de datos proporcionada por SIMA, con información correspondiente a las concentraciones de distintos contaminantes captadas a distintas horas del día por las estaciones meteorológicas del estado de Nuevo León, con datos que datan a inicios de 2017 hasta mediados del 2021.

## 6. Estación SE "La Pastora"

La estación meteorológica sureste está ubicada dentro de una zona llamada "La Pastora" en el municipio de Guadalupe, Nuevo León. Se encuentra a un lado del Cerro de la Silla, de cuál baja el agua hasta llegar al Río La Silla. Habitán unas 3,000 personas aproximadamente con una edad promedio de 35 años y una escolaridad promedio de 12 años cursados.

Esta zona es muy transitada ya que dentro de ella existen diversos negocios como ferreterías, supermercados, restaurantes, el Estadio BBVA Bancomer, etc. Así mismo tiene muchas avenidas principales que conectan los municipios de Monterrey, Juarez y Valle Alto con Guadalupe.

## 7. Justificación

Con todo lo mencionado anteriormente, se establece que es de suma importancia ya que representa un riesgo para la salud de la población, los animales y el ecosistema. El comprender el proceso que se lleva a cabo para obtener mediciones de distintos elementos en el aire que potencialmente pueden afectar negativamente la salud de la población es de suma importancia ya que así se puede tomar medidas leves, medianas o inclusive severas para comenzar a tomar acciones para el retroceso de la pésima calidad de aire que se respira. Obteniendo efectivamente las medidas necesarias, se podrá comenzar a crear un efecto dominó en donde la población podrá entender, en base a datos reales, la situación de la calidad del aire y así la misma estará convencida de tomar acción para hacer que el aire sea de una buena calidad para todos.

## 8. Comprensión de los datos

Para comenzar a hacer dicha sección, lo primero que se hizo fue generar un tratamiento de las bases datos originales. Fueron a partir de dos donde surgió la nueva, las cuales eran correspondientes a los contaminantes y a las condiciones meteorológicas. Se procedió a obtener únicamente los datos de la zona con la que trabajaremos, la Sureste, las cuales eran las columnas *SE* y *SE<sub>b</sub>* de cada hoja de Excel de las dos bases, para posteriormente unirlas en una misma y colocar cada dato a la fecha correspondiente que le tocaba, para que de esta manera no se corriera el orden de estos, y se tuvieran resultados erróneos.

Posterior a esto, se hizo caso a las banderas que fueron proporcionadas. Para aquellas que eran inválidas, los datos que las acompañaban fueron reemplazados por valores nulos (aún no se reemplazan nuevamente por escalares dependiendo del criterio que utilizamos), y las que eran válidas simplemente se dejaron los datos como estaban.

- Dimensión del dataset: 39394 filas por 17 columnas
- Descripción de variables:

**Fecha:** Es la variable que contiene únicamente las fechas en las cuales se tomaron las mediciones de los contaminantes, siendo cada hora el cambio. En la base de datos, esta variable inicia el primero de enero del 2017 a las 00:00, y termina el 30 de junio del 2021. Cabe aclarar que se utilizará sólo como variable categórica.

**PM10:** Variable referente a pequeñas partículas sólidas o líquidas de polvo menor a 10 micrómetros, la cual es de tipo numérica. Como aún no se asignan valores a los nulos, aún se cuentan con muchos de ellos debido a las banderas, de hecho, este dato se va a repetir para todas las siguientes variables.

**PM2:** Variable con mismas características que la anterior, con la diferencia de que su diámetro es menor a 2.5 micrómetros.

**SO<sub>2</sub>:** El Dióxido de Azufre es una variable de tipo cuantitativa, cuya característica es que se origina durante la combustión de carburantes fósiles que contienen azufre.

**O<sub>3</sub>:** Variable cuantitativa referente al ozono, que puede formarse en grandes concentraciones debido a la reacción química entre otros contaminantes como los óxidos de nitrógeno (NO<sub>x</sub>), que son justamente una de las variables que tiene la base.

**CO:** El Monóxido de Carbono es cuantitativo, causado principalmente por la combustión incompleta del carbono.

**NO:** El Monóxido de Nitrógeno es un contaminante que forma parte de las reacciones atmosféricas causando el “smog”, en esta base tomará valores cuantitativos.

**NO<sub>2</sub>:** El dióxido de nitrógeno es una de las variables cuantitativas causadas principalmente por la combustión de los vehículos a motor, las calefacciones y del transporte marítimo.

**NO<sub>x</sub>:** Término utilizado para referirse a las variables cuantitativas de óxidos nitrosos.

**T:** Variable cuantitativa referente a la temperatura de la fecha y hora en la que se registró un dato, la cual puede influir directamente en el aumento de un contaminante específico debido a las condiciones meteorológicas.

**RH:** Humedad Relativa, variable con datos cuantitativos.

**SR:** Radiación Solar, la cual de igual manera, afecta a que ciertos contaminantes aparezcan con mayor frecuencia en el aire.

**RF:** Precipitación que se tiene al momento de tomar los demás datos de las variables que complementan la base.

**PRS:** Presión Atmosférica, variable cuantitativa.

**WS:** Velocidad del Viento, variable cuantitativa, lo cual hace que ciertos contaminantes (se mostrará posteriormente en la matriz de correlaciones) disminuyan su intensidad.

**WD:** Dirección del Viento, la última de las variables, la cual contiene, como todas exceptuando la Fecha, variables cuantitativas. [10]

■ Medidas estadísticas:

- Variables cuantitativas: **Media:**

PM10	50.227384
PM2	18.452335
SO2	713.056493
O3	27.248761
CO	1.552716
NO	711.001705
NO2	699.464670
NOx	12.411209
T	22.960996
RH	62.091569
SR	0.136696
RF	0.074820
PRS	726.971988
WS	12.192040
WD	84.572546

Figura 3: Promedio de variables cuantitativas

**Mediana:**

PM10	43.715
PM2	15.390
SO2	713.200
O3	24.000
CO	1.520
NO	711.100
NO2	699.700
NOx	9.000
T	23.695
RH	64.000
SR	0.012
RF	0.000
PRS	726.600
WS	11.300
WD	61.000

Figura 4: Mediana de variables cuantitativas

**Moda:**

	PM10	PM2	SO2	O3	CO	NO	NO2	NOx	T	RH	SR	RF	PRS	WS	WD
0	28.0	8.0	713.1	3.0	0.63	713.0	700.4	7.4	23.81	83.0	0.007	0.0	723.6	2.9	352.0
1	Nan	Nan	Nan	Nan	0.67	Nan	Nan	Nan	25.11	Nan	Nan	Nan	Nan	Nan	Nan

Figura 5: Moda de variables cuantitativas

**Rango: máximo - mínimo, varianza y desviación estándar:**

	PM10	PM2	SO2	O3	CO	NO	NO2	NOx	T	RH	SR	RF	PRS	WS	WD
std	32.469541	12.372887	57.079332	18.550203	0.810326	3.274469	6.375705	10.797982	7.219237	20.583161	0.375352	1.025447	6.749582	6.436835	512.296655
var	1054.2321	153.085323	3258.050198	344.112243	0.656629	10.722092	40.846532	116.99474	52.117386	423.666506	0.140289	1.195626	45.556446	41.45602	2624.8733
min	2	2.01	-9999	1	0.05	700.6	400.3	0	-9.62	2	0	0	1.4	-9999	
25%	28	9.99	712.3	12	0.95	708.9	697.4	6.4	18.34	46	0.007	0	723.3	6.8	17
50%	43.715	19.39	713.2	24	1.52	711.1	699.7	9	23.656	64	0.012	0	726.6	11.3	61
75%	64	23.7	714.3	39	2.17	713	701.675	13.8	27.98	79	0.235	0	730.3	16.8	113
max	706.65	349	738.9	139	32	721.9	711.1	86.7	42.12	98	64	17.99	740	64	360

Figura 6: Varianza, desviación estándar y rangos de las variables cuantitativas

- Variables cualitativas: Debido a que la única variable cualitativa es la de "Fecha", y en este caso son valores específicos de la hora y día en que se toman los datos de las variables cuantitativas, sería contraproducente el generar algún histograma o gráfico de frecuencia, al igual que obtener la moda.
- Herramientas de visualización:
  - Variables cuantitativas:

#### Medidas de posición no-central:

Observando los boxplots (Vease Anexos: b), se pueden discernir algunas cosas importantes. El monóxido de carbono presenta un par de valores atípicos, pero especialmente uno bastante alejado de los demás [Fig. 48]. El monóxido de nitrógeno presenta algunos valores atípicos tanto por debajo como sobre los cuartiles extremos [Fig.49]. Como el monóxido de carbono, el dióxido de nitrógeno presenta un valor atípico muy alejado de los demás [Fig.50]. Los óxidos de nitrógeno presentan muchos valores atípicos sobre el último cuartil [Fig.51]. Este es el mismo caso para el ozono [Fig.52]. El SO<sub>2</sub> presenta algunos valores atípicos alejados de la mayoría de datos 53 . La temperatura presenta varios valores atípicos debajo del primer cuartil [Fig.54]. La presión atmosférica presenta un valor atípico muy por debajo de la mayoría de los datos [Fig.55]. La lluvia presenta una gran cantidad de valores atípicos [Fig.56]. La humedad relativa no presenta valores atípicos [Fig.57]. Tanto la dirección como la velocidad del viento presentan algunos valores atípicos sobre el último cuartil [Fig.58-59].

#### Histogramas:

Utilizando estos gráficos se puede analizar la forma en que los datos están distribuidos para cada variable (ver Anexos: a.), entre otros aspectos importantes para comprenderlos mejor. En primer lugar es importante notar una asimetría en los datos para casi todas las variables, a excepción quizás de los correspondientes al Óxido de Nitrógeno (NO) [Fig.36]. Esta asimetría en los datos indica que podrían no presentar una distribución normal.

Asimismo se aprecia la presencia de valores *outliers* nuevamente para casi todas las variables. Los casos más notorios son para el Monóxido de Carbono (CO), Dióxido de Nitrógeno (NO<sub>2</sub>), Dióxido de

Azufre (SO<sub>2</sub>) y la dirección del viento (WD), por mencionar algunos ejemplos [Figs.35, 37, 40, 45]. La aparición de estos datos atípicos puede atribuirse a distintos factores, como errores de captura debido a fallos de los dispositivos de medición, condiciones climatológicas inusuales o algún otro factor externo.

Por último, con ayuda de la función de distribución ajustada en las gráficas es posible darse una idea de la manera en que los datos se ajustan a su distribución. En este caso, el mejor ejemplo de un buen ajuste según las gráficas es representado por la Humedad Relativa (RH), la Temperatura (T) y posiblemente el Monóxido de Carbono (CO) [Figs. 44, 41, 35]. Con esta información se podría decir que para la implementación del futuro modelo vale la pena considerar transformar los datos para normalizarlos y que se ajusten de mejor manera a su distribución.

### Análisis de correlación:



Figura 7: Análisis de correlación de los datos

- Variables cualitativas:

Mismo caso que en las medidas estadísticas.

## 9. Preparación de los datos:

### ▪ Conjunto de datos que se utilizará:

Para poder separar información extra con la información que realmente se utilizará para el análisis, se optó por eliminar ciertas columnas de datos que presentaban más de un 80 % de datos nulos, lo cual se explica en la sección de "Manejo de valores faltantes". Dichos datos eliminados fueron aquellos que no se utilizarán para el análisis gracias a que muestran información que no es pertinente para la solución al igual que se eliminaron otras columnas gracias a la falta de información, tal y como se muestra más adelante.

	Fecha	PM10	PM2	SO2	O3	CO	T	RH	SR	RF	PRS	WS
0	2017-01-01 00:00:00	NaN	NaN	712.5	6.0	3.77	18.71	72.0	0.007	0.0	720.5	8.8
1	2017-01-01 01:00:00	373.0	349.0	712.4	5.0	3.01	19.67	64.0	0.007	0.0	720.1	13.1
2	2017-01-01 02:00:00	373.0	187.0	712.3	4.0	1.80	22.33	51.0	0.007	0.0	719.6	12.9
3	2017-01-01 03:00:00	124.0	101.0	712.1	4.0	2.11	20.51	58.0	0.007	0.0	719.5	9.1
4	2017-01-01 04:00:00	124.0	81.0	711.9	4.0	1.77	20.22	60.0	0.007	0.0	719.1	9.2

Figura 8: DataFrame Final

La nueva matriz de correlación, la cual contiene únicamente los datos que se utilizarán, es la siguiente.



Figura 9: Análisis de correlación de los datos seleccionados

■ **Eliminación de duplicados:**

No se encontraron valores duplicados en la base de datos, por lo que no hizo falta un tratamiento especial.

■ **Corrección de valores erróneos:**

Se consideraron como valores erróneos aquellos que varían en sobre medida sus valores en relación a los demás registros dentro de la misma variable, un ejemplo de esto ocurre para la medición de contaminantes, con valores como -9999 o -9.62. Ante estos casos, la estrategia consistió en transformarlos a valores nulos (NaNs).

De la misma manera, todos aquellos valores que estaban acompañados por una bandera que los declaraba invalidos fueron remplazados por valores nulos (NaNs). Las banderas son valores que acompañan a todas las mediciones. Ciertos valores presentes en estas banderas, indicados por el SIMA, causa una invalidez en la medición.

■ **Manejo de valores faltantes:**

Estos valores nulos (NaNs) se graficaron para poder tener una mejor visualización de cómo estaban distribuidos dentro de las variables:

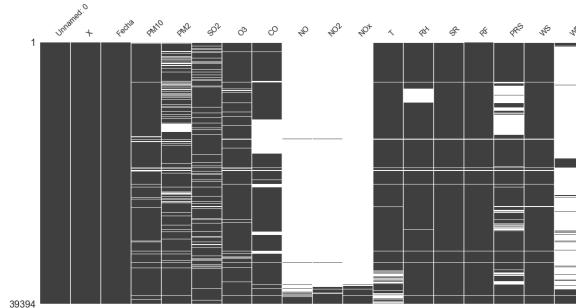


Figura 10: Gráfica de valores nulos

En la gráfica anterior [Fig. 10], los espacios en blanco representan las instancias de datos faltantes. Como se puede observar, existen diversas columnas donde la mayoría de los datos son nulos. Para estos casos se decidió no tomar en cuenta aquellas columnas que presentaran más del 80 % de datos nulos. Utilizando este criterio, se ignoraron las columnas *NO*, *NO<sub>2</sub>*, *NO<sub>x</sub>*, y *WD*.

Posteriormente se tomó la decisión de remplazar los demás valores faltantes con el valor anterior correspondiente utilizando la función "ffill", ya que si se rellenaban con la media u algún otro método iba a crear un sesgo en los datos.

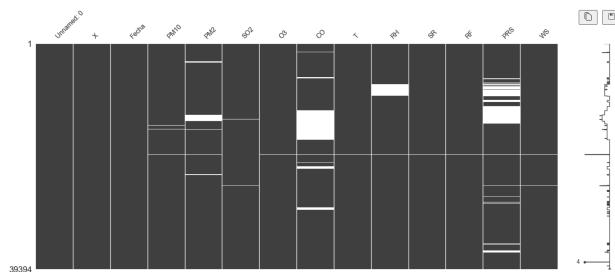


Figura 11: Gráfica reemplazando los valores nulos

Cabe mencionar que después de realizar este proceso, todavía existen valores faltantes debido a que se le indicó a la función "ffill" que tomara como límite un máximo de 36 horas donde existieran NaNs para decidir si los reemplazaba o no. Este parámetro se determinó tomando en cuenta los días de la semana, es decir, es probable que dos días consecutivos tengan la misma cantidad de contaminantes o por lo menos similar, a diferencia de un domingo en la mañana en donde existe poca contaminación a un miércoles por la tarde que es uno de los días con más tráfico en la ciudad.

- **Manejo de datos categóricos:**

Dentro de la base de datos, la única variable con datos categóricos que se toma en cuenta es la Fecha, que como se ha mencionado anteriormente, datan desde inicios de 2017 hasta mediados del 2021. Teniendo esto en cuenta, no se requiere una transformación a otro tipo de formato.

- **Manejo de valores atípicos (outliers):** La mayoría de los histogramas de los contaminantes presentaron valores atípicos, sin embargo se decidió no realizar ningún cambio a estos ya que existe la posibilidad de que si sean reales. Las variables en las que se consideraría remover estos valores outliers serían en las de monóxido de carbono, dióxido de nitrógeno y SO<sub>2</sub> en caso de que representaran un sesgo dentro del modelo. Los únicos valores que si se removieron fueron los valores outliers negativos ya que, como mencionamos anteriormente, son considerados un error.

## 10. Transformación de Datos:

Unas de las consideraciones a tomar sería cambiar la unidad de medida de la columna de CO a ppb (partes por billón) ya que está dada en ppm (partes por millón) cuando todos los demás contaminantes están dados en ppb.

## 11. Reformatea/Reestructura los datos en caso de necesario:

Al momento de analizar la base con los datos limpios, se decidió no juntar variables para crear una nueva columna, es decir, reformatear los datos al combinarlos en una nueva columna. Esto gracias a que no se encontró una forma eficiente en que se pueda utilizar los datos unidos sin que la variabilidad incremente drásticamente al igual que la correlación no decrezca. La única manera que se logró encontrar para unir dos o mas variables es utilizando las columnas NO y PRS ya que tienen una correlación de 0.99 como se muestra en la Figura 6. Sin embargo, dichas variables fueron eliminadas para el análisis posterior, por lo que las columnas restantes no muestran un comportamiento similar entre sí como para unirlas en una nueva columna.

## 12. Límites de Contaminantes

Dentro de esta sección se generaron gráficos representando los niveles de cada contaminante por hora donde se observa que tan frecuentemente sobrepasó el límite de la normativa durante el año 2020.

### 12.1. Materia particulada gruesa ( $PM_{10}$ )

El valor límite indicador con el que se evalúa es  $75 \mu g/m^3$  como máximo:

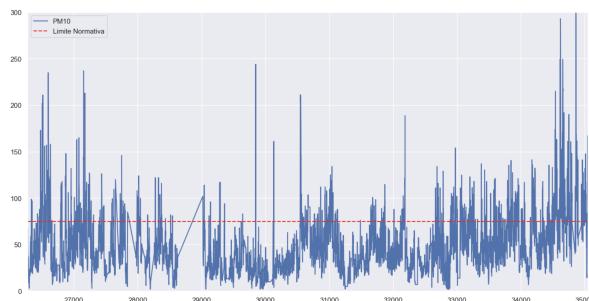


Figura 12: Niveles de  $PM_{10}$  sobre pasando la normativa

Las veces que el contaminante  $PM_{10}$  sobre pasó la normativa fueron 1347 de 8776, lo cual representa un 15.35 % de las horas durante todo el año.

## 12.2. Materia particulada fina ( $PM_{2.5}$ )

El valor límite indicador con el que se evalúa es  $45 \mu\text{g}/\text{m}^3$  como máximo:

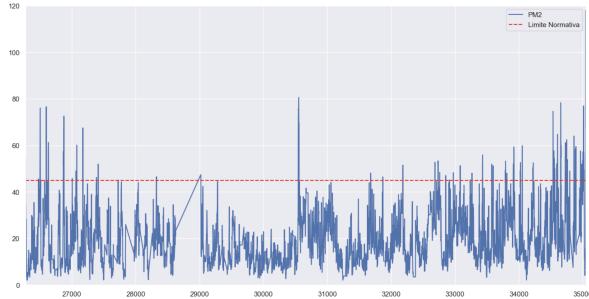


Figura 13: Niveles de  $PM_{2.5}$  sobrepasando la normativa

Las veces que el contaminante  $PM_{2.5}$  sobre pasó la normativa fueron 254 de 8776, lo cual representa un 2.89 % de las horas durante todo el año.

## 12.3. Dióxido de Azufre ( $SO_2$ )

El valor límite indicador con el que se evalúa en una hora es de 0.075 ppm como máximo de un promedio aritmético de 3 años consecutivos de los percentiles 99 anuales, obtenidos de los máximos diarios. Los datos trabajados están dados en ppb, por lo que este valor máximo convertido a ppb es de 75. [11]

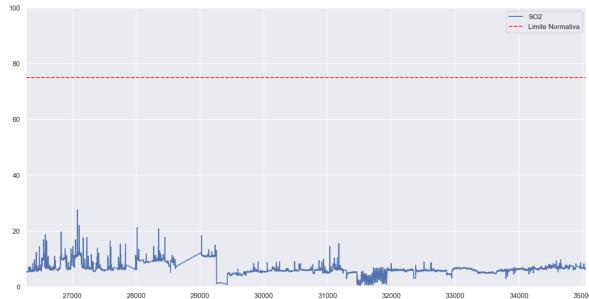


Figura 14: Niveles de  $SO_2$  sobrepasando la normativa

Se puede observar que todas las lecturas de este contaminante están por debajo de la normativa.

## 12.4. Ozono ( $O_3$ )

El valor límite indicador con el que se evalúa es 95 ppb como máximo:

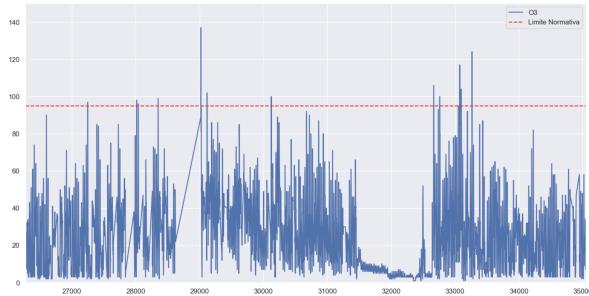


Figura 15: Niveles de  $O_3$  sobrepasando la normativa

Las veces que el contaminante  $O_3$  sobre pasó la normativa fueron 25 de 8776, lo cual representa un 0.27 % de las horas durante todo el año.

## 12.5. Monóxido de Carbono (*CO*)

El valor límite indicador con el que se evalúa es 11 ppm como máximo, lo cual es 11,000 ppb:



Figura 16: Niveles de *CO* sobre pasando la normativa

En ninguna ocasión el contaminante *CO* sobre pasó la normativa.

Analizando todas las veces en las que los contaminantes sobre pasaron la normativa, se pudo generar la siguiente tabla:

Contaminante	Veces excedidas
PM10	15.35 %
PM2.5	2.89 %
O3	0.27 %
SO2	0 %
CO	0 %

Cuadro 1: Porcentaje de incumplimiento de la normativa, por contaminante, para el 2020

En donde se puede observar que el contaminante con niveles más altos durante 2020 fue *PM<sub>10</sub>*. Según datos del Registro Estatal de Emisiones y Fuentes de Contaminantes de España, el 77.9 % de la cantidad total emitida de *PM<sub>10</sub>* procede del polvo suspendido existente en la atmósfera. La industria, la construcción y el comercio suman un 7.6 % y el transporte rodado un 6.5 %. Como fuentes minoritarias están las quemas agrícolas con un 3.7 % y fuentes de origen doméstico que representan el 3.3 %. [12]

Para tener una mejor visualización de las veces que este contaminante sobre pasa la normativa cada mes durante el año 2020, se generó la siguiente gráfica:

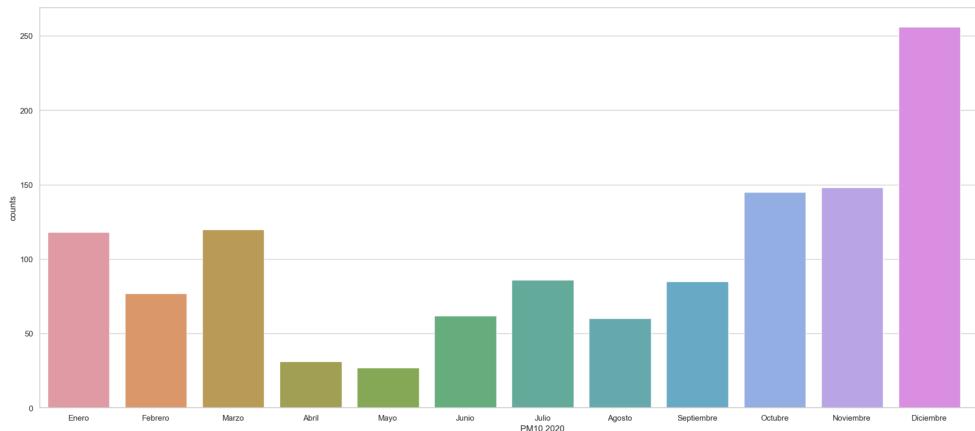


Figura 17: Cantidad de veces que *PM<sub>10</sub>* sobre pasó la normativa en 2020

Cabe recordar la cantidad de registros que se mencionan dentro de este reporte, están dados por hora y no por día. Ahora, teniendo en cuenta que el año que se analizó fue 2020, la pandemia de COVID-19 comenzó en marzo, lo cual fue un factor muy importante para que disminuyeran drásticamente los valores de *PM<sub>10</sub>*. Sin embargo, volvieron a aumentar a partir de junio debido a cuestiones sociales del fin del confinamiento y el hecho de que las

personas daban por terminada la pandemia. Posteriormente en agosto bajó debido a los casos de rebrotes y que el gobierno sugirió seguir tomando medidas de precaución. A partir de septiembre los niveles de este contaminante fueron aumentando cada vez más hasta llegar a diciembre que fue el mes más alto. Este comportamiento se vio afectado debido a que las bajas temperaturas también influyen a que aumenten los niveles de contaminación.

En invierno ocurre un fenómeno llamado inversión térmica, el cual altera el orden de las capas de aire. Cuando la temperatura baja, el suelo se enfriá rápidamente y provoca también una bajada drástica de temperatura en el aire de las capas más bajas (el que respiramos y el que presenta una mayor concentración de contaminantes que salen del tubo de escape de los vehículos). Ese enfriamiento del aire más cercano al suelo altera la dinámica atmosférica. Como las capas de aire que tiene encima están más calientes, la masa de aire contaminado queda atrapado a escasos metros del suelo. [13]

Ahora para tener una visualización más enfocada, se generaron boxplots por cada hora del día durante todo el año 2020 para analizar los valores que sobrepasaron la normativa de este contaminante a determinadas horas del día y la cantidad de datos que se registraron dentro de este rango.

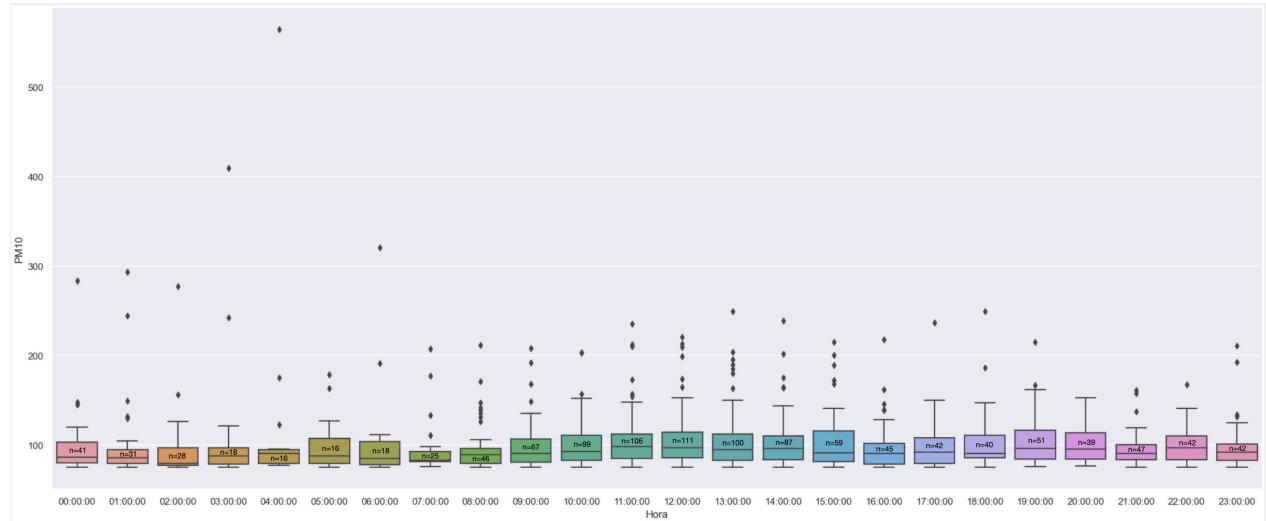


Figura 18: Boxplots por hora de  $PM_{10}$  en 2020

En general, durante el día se observaron concentraciones alrededor de aproximadamente  $100 \mu g/m^3$ , es decir 25 unidades por encima de la normativa y el rango de horas que contienen no sólo los valores más altos en concentración  $PM_{10}$ , sino también una mayor cantidad de registros es de 11:00 a.m. a 15:00 p.m. aproximadamente, siendo al medio día el punto crítico con 111 registros. Este comportamiento coincide justamente con algunas de las horas en las que normalmente hay más tráfico, actividad laboral y movimiento en la ciudad.

Ahora haciendo un mismo análisis pero con el contaminante  $PM_{2.5}$ :

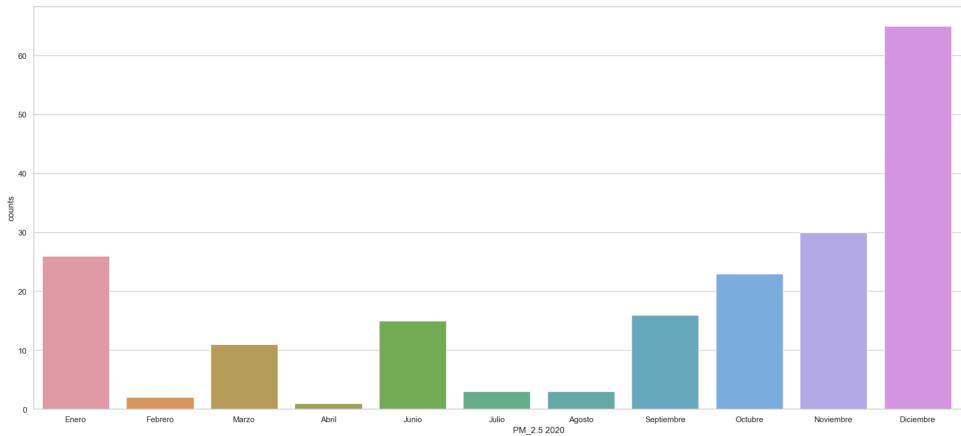


Figura 19: Cantidad de veces que  $PM_{2.5}$  sobrepasó la normativa

De manera similar a lo que ocurre con el  $PM_{10}$ , se observa que durante la primera mitad del año la cantidad de veces que  $PM_{2.5}$  sobrepasa la normativa son menos a comparación del resto del año. Abril fue el mes con menos

incidencias con alrededor de 2 y diciembre fue el mes que más veces sobrepasó el límite, registrando alrededor de 65 ocasiones. Estas fechas también coinciden con los meses críticos de la pandemia donde generalmente había menos actividad.

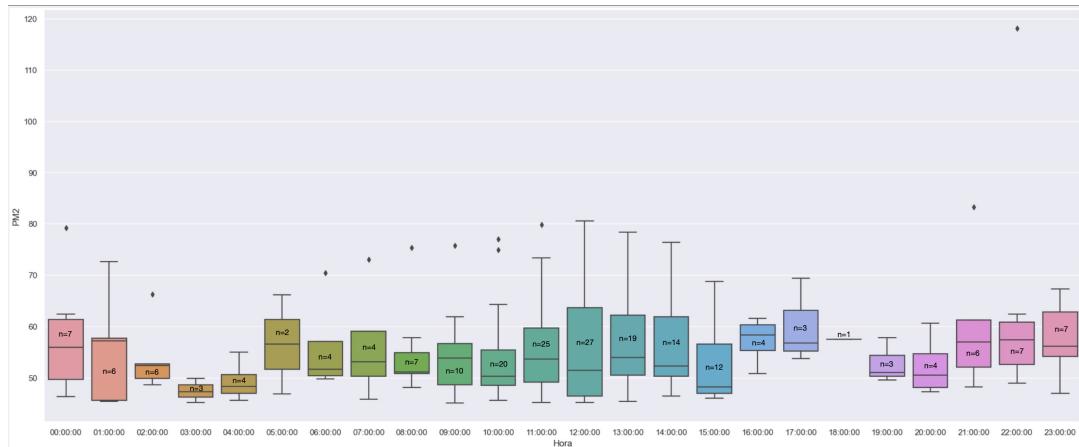


Figura 20: Boxplots por hora de  $PM_{2.5}$

Igualmente las horas del año en las que se observa una mayor concentración y mayor cantidad de datos registrados de  $PM_{2.5}$  coinciden en un rango entre 11:00 y 15:00 hrs aproximadamente. Presentan valores que van entre 50 y  $60 \mu g/m^3$ . En contraste con el caso de las partículas  $PM_{10}$ , para las  $PM_{2.5}$  no hay tantos registros de ocasiones en las que se haya sobrepasado la normativa, donde el mayor caso sucede a medio día, con 27 registros que mayormente se concentran alrededor de  $50 \mu g/m^3$  [Fig. 20]. Este comportamiento bien podría relacionarse, como se menciona anteriormente, a que en estas horas ocurren la mayor cantidad de actividades en la ciudad de todo tipo, cotidianas, industriales y en el hogar.

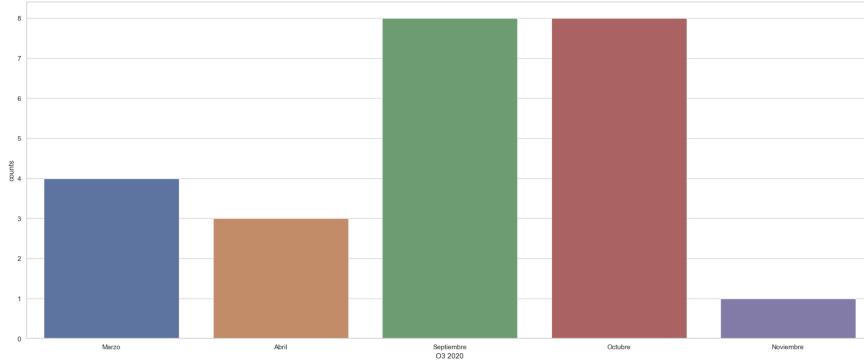


Figura 21: Cantidad de veces que  $O_3$  sobrepasó la normativa

Para el Ozono ( $O_3$ ) se observa un comportamiento totalmente distinto. En primera instancia, los únicos meses cuando se observa un sobreponerse de partículas concentradas según la normativa, corresponden a marzo, abril, septiembre, octubre y noviembre; siendo septiembre y octubre los meses con mayor incidencia de este fenómeno con 8 registros cada uno y Noviembre el menor, con sólo una ocasión. [Fig. 21]

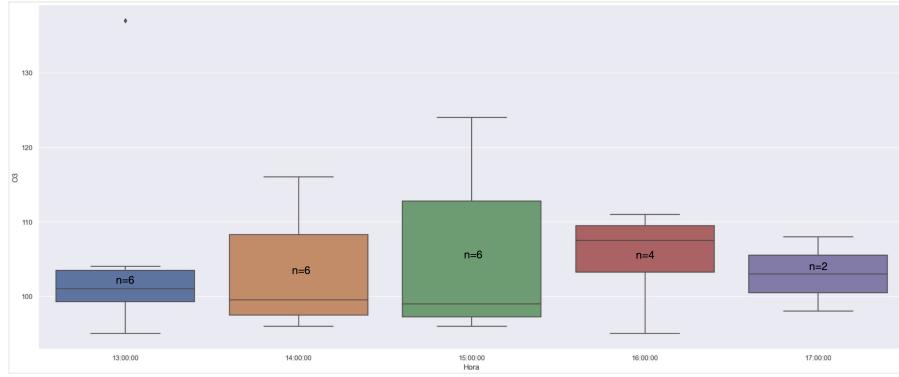


Figura 22: Boxplots por hora de  $O_3$

Asimismo, al año en promedio las únicas horas en las que se observa mayor incidencia, y por lo tanto un sobrepaso en la normativa, para  $O_3$  es en un rango de horas de 13:00 a 17:00 horas. Afortunadamente para este contaminante no son tantas las ocasiones al día en que llegue por encima de la normativa, pues los casos con más registros suceden durante el rango de 13:00 a 15:00 horas con un total 6 registros en cada hora. A las 15:00 hrs se observa el mayor incremento en la concentración de  $O_3$ , llegando poco más arriba de los 120 ppb.

En las ciudades, el  $O_3$  puede formarse en grandes concentraciones debido a la reacción química entre los óxidos de nitrógeno ( $NO_x$ ) y los compuestos orgánicos volátiles (COV) en presencia de luz solar [14]. Este comportamiento coincide justamente con las horas donde el sol se encuentra en su punto más alto durante el día.

## 13. Modelos Escogidos

### 13.1. Regresión Lineal Múltiple

Para la regresión primeramente se tuvo que encontrar qué variables tienen relación lineal con la variable objetivo  $O_3$ . Para esto, no se puede simplemente basarse en la gráfica de correlación, sino que se deben hacer otras pruebas estadísticas para encontrar una relación entre todas las variables independientes, las cuales se hacen pruebas mediante estadísticos multivariados, y la variable objetivo. Para esto, se tuvo que hacer múltiples pruebas y errores.

Primeramente, utilizando Python, se leyeron todos los datos y se utilizó la librería de Scikit-Learn para entrenar aleatoriamente el 70 % de los datos y el otro 30 % se utilizó para hacer una prueba y finalmente comparar los resultados con los datos reales. Haciendo esto y utilizando el modelo de LinearRegressor() de la misma librería, se encontró que el Mean Squared Error (MSE) es igual a 169.61949, lo cual viendo el Boxplot de Ozono dentro de la sección XIIb, la raíz del MSE queda dentro del rango intercuartil, por lo cual el resultado es bueno. Por otro lado, se encontró que Mean Absolute Error (MAE) es igual a 9.85116, lo cual significa que en promedio la magnitud de la diferencia entre lo que se encontró en las pruebas hechas con el 30 % de los datos y los datos reales, lo cual da un resultado bueno para la búsqueda de linealidad entre Ozono y las demás variables.

Luego, para encontrar qué variables son aquellas que mejor describen el comportamiento de la variable objetivo, se evaluaron todas las variables en un análisis de regresión, y, utilizando el p-valor para encontrar si hay linealidad entre las variables y la de Ozono, se fueron descartando variables hasta encontrar cuáles son las que mejor se adaptan a la regresión. Después de varias pruebas, se encontró que las variables que mejor se adaptan son las de  $PM2$ ,  $SO2$ ,  $CO$ ,  $T$ ,  $RH$ ,  $SR$ ,  $RF$  y  $WS$  como se muestra en la Figura 23.

OLS Regression Results						
Dep. Variable:	O3	R-squared:	0.515			
Model:	OLS	Adj. R-squared:	0.515			
Method:	Least Squares	F-statistic:	4010.			
Date:	Thu, 13 Oct 2022	Prob (F-statistic):	0.00			
Time:	22:56:43	Log-Likelihood:	-1.2068e+05			
No. Observations:	30162	AIC:	2.414e+05			
Df Residuals:	30153	BIC:	2.415e+05			
Df Model:	8					
Covariance Type:	nonrobust					
coef	std err	t	P> t	[0.025	0.975]	
const	33.3686	0.554	60.270	0.000	32.283	34.454
PM2	-0.0584	0.006	-9.605	0.000	-0.070	-0.046
SO2	0.0845	0.022	3.764	0.000	0.040	0.128
CO	-0.0019	9.34e-05	-19.894	0.000	-0.002	-0.002
T	0.2171	0.012	17.735	0.000	0.193	0.241
RH	-0.3129	0.004	-71.039	0.000	-0.322	-0.304
SR	30.4434	0.472	64.480	0.000	29.518	31.369
RF	0.2978	0.062	4.819	0.000	0.177	0.419
WS	0.5927	0.013	44.389	0.000	0.567	0.619
Omnibus:	2183.414	Durbin-Watson:	0.201			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	5555.894			
Skew:	0.429	Prob(JB):	0.00			
Kurtosis:	4.920	Cond. No.	1.33e+04			

Figura 23: Análisis de Regresión para  $PM2$ ,  $SO2$ ,  $CO$ ,  $T$ ,  $RH$ ,  $SR$ ,  $RF$  y  $WS$

Como se puede observar en la Figura 23, todos los p-values que se encuentran ahí para las variables asignadas son menores a 0.05, se hecho, son aún más pequeños solo que el modelo no permite observar el valor exacto de estos, sin embargo, con sus cifras significativas son suficientes para rechazar la hipótesis nula, con lo que se puede observar que evidentemente tienen una relación lineal con  $O_3$ . Igualmente, en la tabla se puede observar el valor del coeficiente de determinación el cual es 0.515, lo cual no es ideal pero sigue teniendo un impacto significativo en la variable objetivo.

Igualmente, observando la Figura 23, encontramos que hay una columna llamada "coef", la cual representa los coeficientes que se tienen para cada una de las variables al momento de que se formule una función para  $O_3$ . Una vez representadas en la función objetivo, esta sería:

$$O_3 = 33.3686 - 0.0584PM2 + 0.0845SO2 - 0.0019CO + 0.2171T - 0.3129RH + 30.4434SR + 0.2978RF + 0.5927WS$$

Ahora, una vez teniendo la función establecida, se debe de hacer un análisis de residuos para verificar que realmente se comporta de manera lineal la función. Primeramente, se verificó si los residuos se comportan de forma normal. Para esto, se hizo un análisis de Anderson-Darling para obtener un gráfico sobre el comportamiento de los mismos, el cual se muestra en la Figura 24.

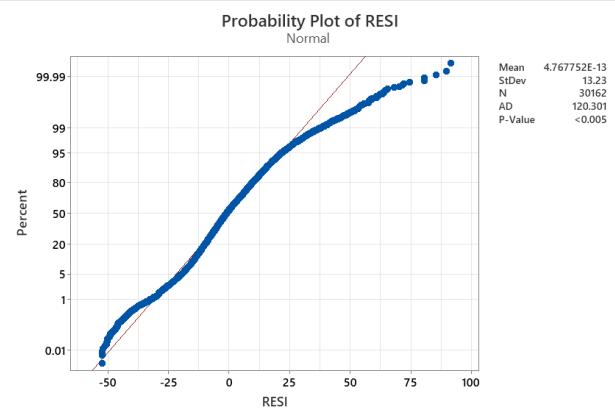


Figura 24: Análisis de Anderson-Darling para Residuos de la Regresión lineal de  $O_3$

Observando la Figura 24, se puede apreciar cómo es que los datos no siguen la linea diagonal mostrada, la cual entre más alto sea el porcentaje que los residuos que están cercano a dicha recta, más normal se comportan estos. Con esto, se puede ver que la gran mayoría de los datos recaen fuera de dicha recta, al igual que el p-valor mostrado en la columna derecha tiene un valor menor a 0.05, por lo que la hipótesis sobre la normalidad de los residuos es rechazada, significando que los residuos se no comportan de manera Normal.

Para poder observar lo encontrado anteriormente de una manera más visual, a continuación en la Figura 25 se muestra el histograma de los residuos de la regresión lineal.

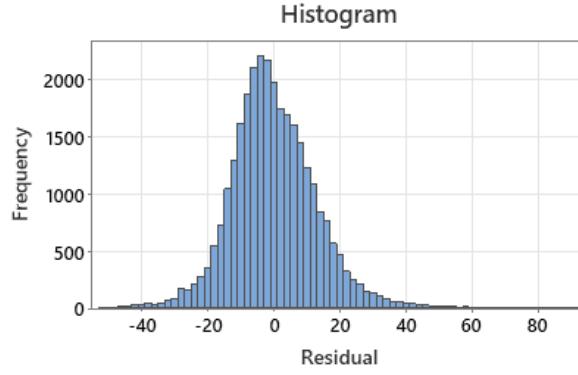


Figura 25: Histograma de los residuos de la regresión lineal

Con todo lo mostrado anteriormente, tanto estadísticos como las pruebas y los análisis, se puede demostrar que el nivel de Ozono en el ambiente no puede ser descrito mediante una regresión lineal múltiple utilizando las variables  $PM2$ ,  $SO2$ ,  $CO$ ,  $T$ ,  $RH$ ,  $SR$ ,  $RF$  y  $WS$  para dicha función, todo esto con un nivel de significancia de 0.05.

### 13.2. Análisis de Discriminante

Como tercer modelo se eligió utilizar el análisis de discriminante, el cual busca clasificar las distintas variables objetivos en diferentes grupos, por lo que es un análisis categórico. Para este, se optó por un análisis discriminante lineal.

Para este modelo se definió como objetivo clasificar si un contaminante sobrepasa o no la normativa impuesta por el Gobierno de México en cualquier hora dada. El contaminante escogido fue el PM10. La normativa "NOM-025-SSA1-2014" indica que el límite para un periodo de 24 horas debe ser inferior a los  $75 \mu\text{g}/\text{m}^3$ , lo que equivale a 75 ppb [7]. Para poder hacer el cálculo sobre el límite, se debe conocer el promedio móvil de las últimas 24 horas. Cabe recalcar que la normativa indica que para poder tener un cálculo válido, se debe contar con las mediciones de, mínimo, el 75 % de las horas anteriores (i.e. 18 horas). Tomando esto en consideración, utilizando los datos disponibles, se creó una nueva columna en la base de datos, la cual indica el promedio móvil de las últimas 24 horas. Tomando esta nueva columna como condición, se agregó otra columna más, la cual simplemente corresponde a si el contaminante está por debajo, es decir, cumpliendo la normativa en esa hora (representado por la clase 0) o si el límite ha sido rebasado (representado por la clase 1).

Realizando este proceso, se puede contar el número de instancias en las cuales sí se cumplió la normativa, así como en las cuales no:

Clase	Porcentaje
Cumplimiento de la normativa (0)	85.08 %
Incumplimiento de la normativa (1)	14.92 %

Cuadro 2: Porcentaje de instancias, por clase

Usando el lenguaje de programación  $R$ , se creó el modelo de clasificación y se obtuvo la siguiente ecuación para calcular el discriminante:

$$D = 0.63 \cdot PM2 - 0.0015 \cdot SO2 - 0.0005 \cdot O3 + 0.0001 \cdot CO - 0.4983 \cdot T - 0.325 \cdot RH - 1.4306 \cdot SR - 0.0079 \cdot RF - 0.0373 \cdot PRS - 0.0083 \cdot WD$$

Usando la ecuación anterior, el centroide del discriminante se encuentra en: -28.915.

Así mismo, se calculó la exactitud del modelo y fue de un 75.11 %. No obstante, la exhaustividad para la clase 1 (incumplimiento) fue de 29.79 %. En la siguiente gráfica se pueden observar los datos que fueron categorizados como "Cumplimiento" e "Incumplimiento" por la función discriminante, graficados con respecto a su dsicriminante, distancia al centroide, y probabilidad.

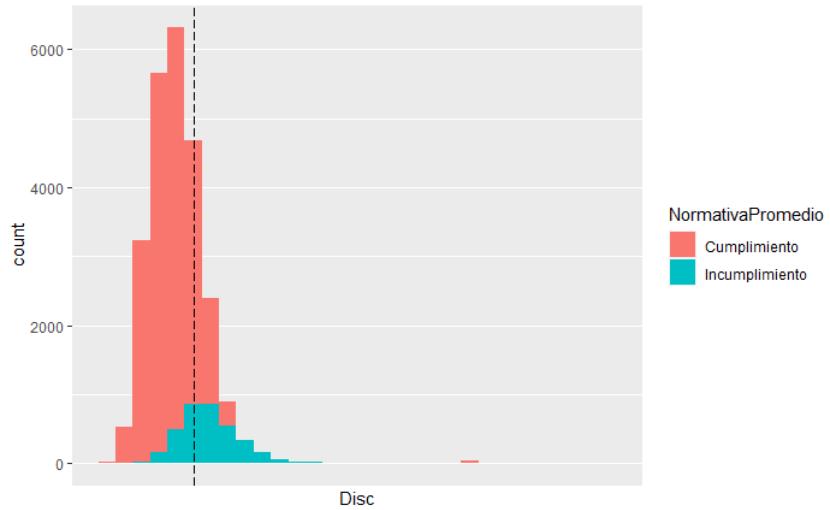


Figura 26: Histograma de los datos según el valor del discriminante, divididos por clase. La línea punteada vertical representa el centroide del discriminante

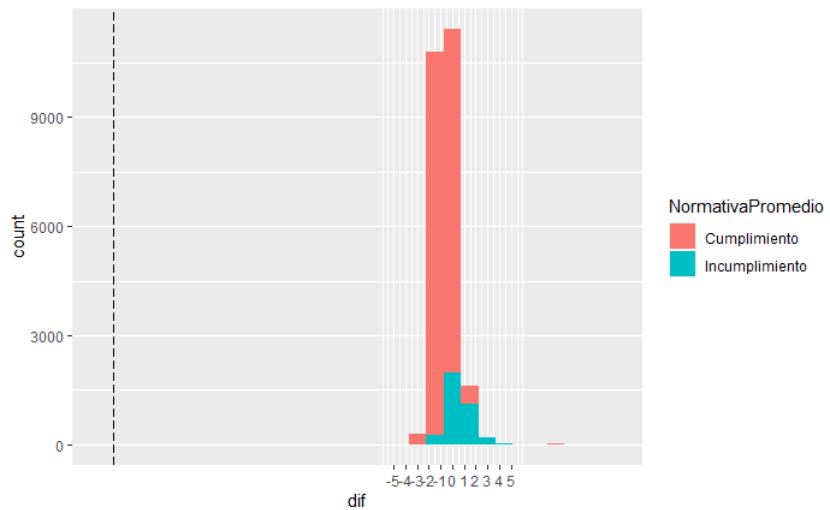


Figura 27: Histograma de la distancia entre el discriminante y el centroide, dividido por clase.

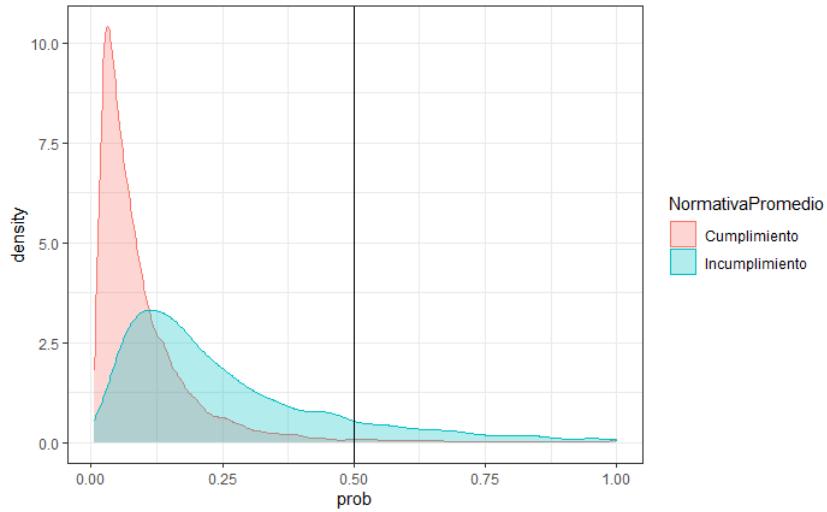


Figura 28: prob

### 13.3. Análisis de Componentes Principales

El modelo PCA, se utilizará con el fin de simplificar la complejidad del espacio muestral, tomando en cuenta la base de datos en cuestión. Debido a que, al final de realizar el modelo se busca obtener con cuántas componentes se puede explicar el mayor porcentaje de la varianza observada de los datos, se excluyó a la variable "Fecha", ya que con una cualitativa no iba a ser posible realizar el análisis.

De primera instancia, se excluyeron los NA que quedaban en la base después de depurarla de la manera más honesta, sin colocar medias en todos los datos nulos ni sustituirlos por valores de horas que no fueran similares. A partir de ese punto, se obtuvieron la media y varianza para conocer cómo se comportaba la base.

Se observa que, en el caso de CO, hay muchas unidades que sobresalen de los demás, pero no se puede inferir que hay una correlación decente con todos, de hecho, ya se realizó una matriz de correlación, por lo que compararlas al azar no tiene sentido. Adentrando más en saber si la concentración de este contaminante es dañina, nos encontramos en que no es de alarmarse, ya que el límite permitido para el CO es de 11,000 ppb, por lo que la media está dentro del rango permitido. Las varianzas obtenidas difieren mucho entre cada variable.

```
> apply(X = df1, MARGIN = 2, FUN = mean)
   PM10      PM2      SO2       O3        CO        T      RH
50.0461558 18.3033986 6.6127949 27.1527274 1509.6153984 22.7219492 61.3081122
   SR       RF      PRS      WS
0.1351673 0.1099404 728.0547779 12.2695587
> apply(X = df1, MARGIN = 2, FUN = var)
   PM10      PM2      SO2       O3        CO        T      RH
1.102466e+03 1.653888e+02 1.260902e+02 3.605926e+02 8.214992e+05 5.648181e+01 4.368067e+02
   PRS      WS
3.534495e-02 1.670393e+00 3.404973e+01 4.247935e+01
```

Figura 29: Medias y varianzas de la base de datos sin NA

Se estandarizaron las variables para que tengan media cero y desviación estándar 1 antes de realizar el estudio PCA, ya que de lo contrario dominará la variable CO en la mayoría de componentes principales.

Se decidió generar 11 componentes principales, considerando las 11 variables que se tienen, de esta manera, además de comprobar cómo se conforma cada componente, se podrá saber hasta cuál de ellos se tiene un porcentaje considerable que explique sus varianzas.

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
PM10	0.201	-0.571	-0.209	-0.006	-0.048	0.200	0.010	-0.270	0.076	0.430
PM2	0.068	-0.567	-0.355	0.064	0.099	0.288	0.148	0.143	0.059	-0.466
SO2	0.124	-0.208	0.443	-0.012	0.828	-0.124	-0.002	-0.163	-0.121	0.006
O3	0.469	0.062	0.130	-0.099	-0.197	-0.051	0.011	-0.255	-0.265	-0.655
CO	-0.172	-0.452	0.283	-0.140	-0.201	-0.304	-0.522	0.478	-0.103	-0.087
T	0.415	0.170	-0.171	0.019	0.249	-0.159	0.035	0.506	0.591	-0.067
RH	-0.432	0.080	-0.197	0.019	0.257	0.168	0.328	0.379	-0.416	-0.130
SR	0.396	-0.120	0.129	-0.188	-0.201	-0.247	0.563	0.308	-0.348	0.313
RF	-0.050	0.081	-0.180	-0.955	0.121	0.121	-0.088	-0.058	0.051	0.006
PRS	-0.213	-0.102	0.627	-0.114	-0.196	0.377	0.384	0.044	0.428	-0.116
WS	0.355	0.185	0.162	0.065	0.027	0.704	-0.352	0.295	-0.257	0.171
Variable	PC11									
PM10	-0.532									
PM2	0.434									
SO2	0.045									
O3	-0.379									
CO	-0.123									
T	-0.266									
RH	-0.482									
SR	0.213									
RF	0.050									
PRS	-0.103									
WS	0.077									

Figura 30: Componentes principales

Se atiende que, del primer componente principal sobresale el contaminante O3, y los parámetros meteorológicos de temperatura y radiación solar, lo cual debe a que a mayor temperatura, mayor concentración de ozono se tendrá en el área del aire en cuestión. Esto que es consecuencia de las radiaciones ultravioleta e infrarroja [15].

Tomando el valor absoluto de los componentes principales, en el segundo, el PM10 y PM2 son los que más sobresalen, los cuales siempre están correlacionados. Esto se observa de mejor manera con la siguiente representación bidimensional de las dos primeras componentes.

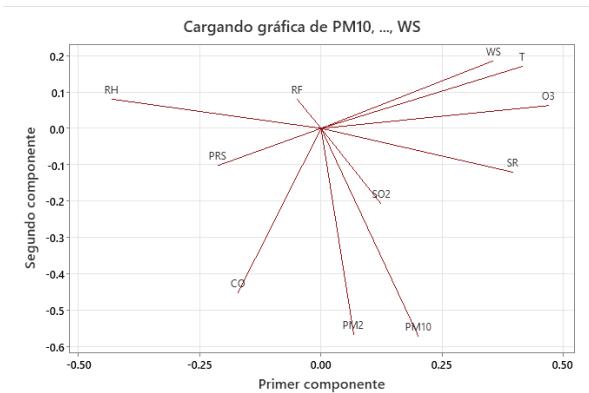


Figura 31: Comparación de los primeros dos componentes principales

Una vez obtenidos los componentes principales correspondientes, se puso como objetivo calcular la proporcionalidad de varianza de cada uno de los componentes, sacar su acumulado.

```
> prop_varianza <- pca$sdev^2 / sum(pca$sdev^2)
> prop_varianza
[1] 0.27285442 0.16829879 0.11601395 0.09031971 0.07832530 0.06355152 0.06135143 0.04630822 0.04364300
[10] 0.03330350 0.02603017
> prop_varianza_acum <- cumsum(prop_varianza)
> prop_varianza_acum
[1] 0.2728544 0.4411532 0.5571672 0.6474869 0.7258122 0.7893637 0.8507151 0.8970233 0.9406666
[10] 0.9739698 1.0000000
```

Figura 32: Varianza explicada por componente principal

Se observa que, dentro del séptimo y noveno componente se tiene una acumulación considerablemente buena, ya que para el PC7 se explica el 85.07 % de la varianza observada en los datos, y para el PC9 el 94.07 %. Debido a que únicamente se sumarían dos componentes, y a que significaría un mejor modelo, se concluye que con 9 componentes principales se puede tener mucha credibilidad. De igual manera, se anexa de manera gráfica la proporcionalidad de varianza explicada del modelo.

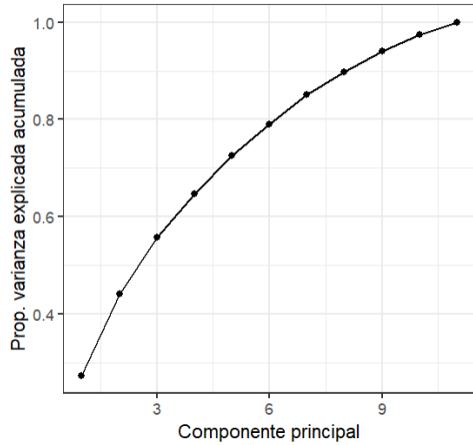


Figura 33: Proporcionalidad de varianza explicada acumulada por componente principal

## 14. Discusión y Conclusiones

Una vez analizando los modelos que se utilizaron para encontrar relaciones multivariables, se puede observar cómo es que algunas variables sí pueden ser de cierta manera "explicadas" por otras, sin embargo, no todas pueden. En el caso de regresión, se observó que a pesar de que todas las variables elegidas tuvieran alta correlación con  $O_3$ , no se puede afirmar que hay manera de explicarse utilizando regresión lineal múltiple utilizando las variables  $PM2$ ,  $SO2$ ,  $CO$ ,  $T$ ,  $RH$ ,  $SR$ ,  $RF$  y  $WS$  gracias al análisis de residuos que se hizo después de encontrar la función de  $O_3$  respecto a las otras variables mencionadas. En este, se encontró que los residuos no se comportan con base en una distribución normal debido a que se rechaza la hipótesis nula por el p-value encontrado en la prueba de normalidad con un nivel de significancia de 0.05.

Por otro lado, al momento de hacer un análisis de discriminante para clasificar cuándo la variable  $PM10$  sobrepasa la normativa NOM-025-S1-2014, la cual dice que el nivel de concentración de  $PM10$  debe ser menor que  $75 \mu g/m^3$ . En este, utilizando los datos disponibles, se creó una nueva columna en la base de datos que representa el promedio móvil de las últimas 24 horas registradas para poder condicionar otra columna nueva que se creó que indica si los datos sobrepasaron la normativa o no. Esta nueva columna es con la que se buscó clasificar los datos. Con esto, se encontró que el 85.08 % de los datos registrados cumplen con la normativa, mientras que el resto no lo hace. Después, se buscó encontrar el centroide para que se pueda clasificar correctamente los datos, el cual se encuentra en el valor de -28.915. Así mismo se encontró que la exactitud del modelo de clasificación fue de 75.11 %. Este es un valor alto, sin embargo, no es ideal para usarse debido a que la clase de interés no está siendo clasificada de manera correcta, presentando una exhaustividad solamente del 29.78 %. El modelo es bueno para clasificar el cumplimiento de la norma, pero bastante deficiente para clasificar el incumplimiento de esta.

En el caso del PCA, existen un par de cosas que se pueden destacar. La primera de ellas es que se necesita de bastantes componentes principales para lograr explicar una buena proporción de la varianza (7 componentes para el 85.07 % de la varianza, y 9 para el 94.07 %). Esto puede sugerir que, en general, los datos que se están manejando son, por naturaleza, muy variados. No obstante, realizando y aplicando un PCA se logaría reducir las dimensiones del dataset. Por otra parte, observando los coeficientes de los componentes principales sobresalen algunas cuestiones. Para el primero de estos, es interesante notar que las variables meteorológicas correspondientes a la temperatura, la radiación solar, y la velocidad del viento se juntan con el contaminante del ozono, y este grupo está contrapuesto a la humedad relativa. Para el segundo componente principal cabe recalcar que, además de que se encuentren presentes el PM2.5 y el PM10, el CO también juega un papel. Esto puede sugerir una relación entre la materia particulada y el monóxido de carbono la cual no se esté tomando en consideración. Observando el tercer componente principal, los coeficientes que sobresalen son los de la presión ambiental y el dióxido de azufre, por una aparte, y por otra el PM2.5.

Es de importancia recalcar que para regresión lineal, no se encontró este tipo de relación al analizar los residuos, ya que no cumplen con la normalidad. Sin embargo, no podemos rechazar que haya un comportamiento lineal con otra variable objetivo en la base de datos, ya que no se realizó el modelo para cada una de ellas. En el caso del análisis discriminantes realizado, se buscó clasificar cuando la variable  $PM10$  sobrepasa la normativa. Los resultados obtenidos presentaron una precisión del 75.11 %, sin embargo, la exhaustividad para el incumplimiento de la normativa, la clase de interés presenta solamente un 29.78 %. Esto quiere decir que, de todas las instancias donde se incumplía la normativa, solamente se logró identificar el 29.78 % de los casos. Por ello, se usó random forest como método de apoyo, el cual presentó un 93.38 % de precisión, y más importante, una exhaustividad del 89.08 %. Es por ello que se concluye que es mejor modelo para clasificar la normativa de  $PM10$ . Finalmente para PCA, lo anteriormente revisado deja abierta la pregunta sobre ¿Qué relación tienen el PM2.5, el PM10, y el CO para la estación ubicada en la pastora? Sería interesante investigar en un futuro si existe una relación entre la

presión atmosférica y la presencia de SO<sub>2</sub> en la atmósfera, además de que, para futuras prácticas sería de ayuda aplicar la regresión lineal múltiple para más variables.

## 15. Anexos

### 15.1. Clasificación con Random Forest

De entre todos los modelos de aprendizaje supervisado para clasificación, se optó por utilizar *Random Forest* por su capacidad para manejar grandes bases de datos de manera eficiente, su habilidad de ejecutarse con incluso miles de variables de entrada y de lidiar con *outliers* y ruido en los datos, y por ser computacionalmente más rápido que otros modelos. [16]

El Random Forest es un conjunto de árboles de decisión, otro modelo de clasificación, donde cada árbol contribuye con un solo peso para la asignación de la clase más frecuente a los datos de entrada. Utiliza un subconjunto aleatorio de características en la división de cada nodo, en lugar de seleccionar directamente las mejores variables, lo que reduce el error de generalización. [16]

Para este modelo se definió como objetivo clasificar si un contaminante sobrepasa o no la normativa impuesta por el Gobierno de México en cualquier hora dada. El contaminante escogido fue el PM10. La normativa "NOM-025-SSA1-2014" indica que el límite para un periodo de 24 horas debe ser inferior a los 75  $\mu\text{g}/\text{m}^3$ , lo que equivale a 75 ppb [7]. Para poder hacer el cálculo sobre el límite, se debe conocer el promedio móvil de las últimas 24 horas. Cabe recalcar que la normativa indica que para poder tener un cálculo válido, se debe contar con las mediciones de, mínimo, el 75 % de las horas anteriores (i.e. 18 horas). Tomando esto en consideración, utilizando los datos disponibles, se creó una nueva columna en la base de datos, la cual indica el promedio móvil de las últimas 24 horas. Tomando esta nueva columna como condición, se agregó otra columna más, la cual simplemente corresponde a si se está cumpliendo la normativa en esa hora (representado por la clase 0) o si el límite ha sido rebasado (representado por la clase 1).

Realizando este proceso, se puede contar el número de instancias en las cuales sí se cumplió la normativa, así como en las cuales no:

Clase	Porcentaje
Cumplimiento de la normativa (0)	85.08 %
Incumplimiento de la normativa (1)	14.92 %

Cuadro 3: Porcentaje de instancias, por clase

Como se puede observar en la tabla 3, existe un gran desequilibrio entre las dos clases a clasificar (14.92 % de la clase 1 contra un 85.08 % de la clase 0). Esto puede causar problemas al momento de entrenar el modelo y su capacidad para clasificar correctamente las clases. Para empezar, al tener una menor cantidad de datos en una clase, es posible que el modelo no se pueda ajustar de manera adecuada para poder clasificar las instancias de la clase minoritaria correctamente. El segundo problema es que se puede tener un modelo preciso, pero no adecuado. Para esto se debe tomar en cuenta otras métricas de desempeño, como la exhaustividad y el valor-f.

Para tratar el problema de desequilibrio de clases, se hará un remuestreo (*resampling*) de los datos. Específicamente, se hará un sobre-muestreo (*oversampling*) para la clase 1. El sobre-muestreo creará nuevas instancias de la clase minoritaria para poder crear un mejor modelo. Para este trabajo, se utilizó la técnica de sobre-muestreo sintético minoritario (SMOTE, por sus siglas en inglés). [17]. La SMOTE crea nuevas instancias de la clase minoritaria basándose en los atributos de las instancias reales presentes en los datos. De esta manera, se crea nuevos datos sintéticos, los cuales pertenecen a la clase minoritaria y comparten las mismas características. Esto es importante ya que no se altera ningún parámetro general de la clase minoritaria, mientras que se logra el objetivo de crear nuevas instancias. Por ejemplo, la distribución de los diferentes datos en base a su clase se mantiene igual. Esto se puede observar en las gráficas de densidad, en el Anexo c (Fig. 60); el área bajo la curva aumenta, pero la forma de la función permanece similar.

El remuestreo se realizó en *Python*, utilizando la librería *imblearn*. Se buscó crear una relación de 2 a 1 entre la clase 0 y la clase 1.

Clase	Porcentaje
Cumplimiento de la normativa (0)	66.67 %
Incumplimiento de la normativa (1)	33.33 %

Cuadro 4: Porcentaje de instancias, por clase, en los datos remuestreados

Utilizando los datos remuestreados, se creó un modelo de clasificación *Random Forest* en *Python* utilizando la librería de *scikit-learn*. Los parámetros utilizados en el modelo *Random Forest* son los siguientes:

- Crietrlion: Gini
- n\_estimators: 100
- min\_samples\_split: 2

Las métricas que se utilizarán para validar el modelo son las siguientes: precisión, exhaustividad, valor-f, y matriz de confusión.

Se dividió los datos en dos grupos, uno para el entrenamiento del modelo y otro para la validación de este, con una razón de 75 % y 25 % respectivamente. Una vez entrenado el modelo, se obtuvieron los siguientes resultados:

Métrica	Clase 0	Clase 1	Promedio Ponderado
Precisión	94.57 %	91.02 %	93.38 %
Exhaustividad	95.59 %	89.08 %	93.41 %
Valor-f	95.07 %	90.04 %	93.38 %

Cuadro 5: Resultado de las diferentes métricas

Observando la tabla 5, se puede apreciar que, en general, el modelo logra producir muy buenas métricas. La clase 1 presenta puntajes ligeramente peores que la clase 0. Es importante notar esto ya que la clase que es de interés es precisamente la clase 1. No obstante, el modelo en general logra clasificar correctamente el 93.38 % de los casos. También, presenta una exhaustividad general de 93.41 %, aunque solamente un 89.08 % para la clase 1.

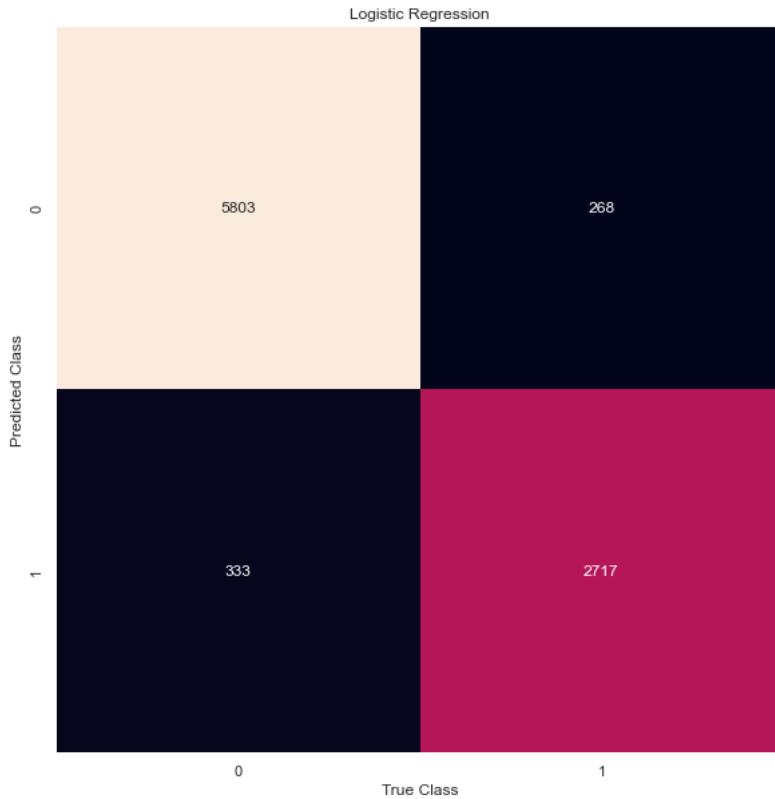


Figura 34: Matriz de confusión para el modelo de clasificación

La matriz de confusión refuerza los resultados obtenidos en la tabla 5.

El modelo de clasificación propuesto puede ser utilizado cuando exista una falla en el sensor de PM10, o este esté deshabilitado por largos períodos de tiempo. También, si se cuentan con los datos, se puede intentar crear una clasificación para años anteriores a 2017

## 15.2. Gráficas de histogramas.

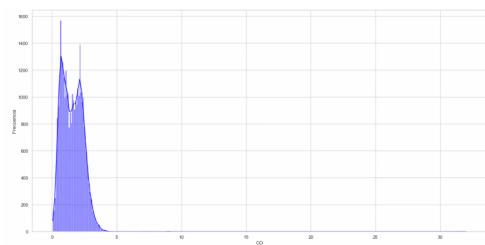


Figura 35: Histograma Monóxido de Carbono

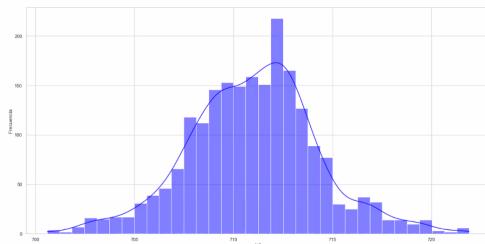


Figura 36: Histograma Monóxido de Nitrógeno

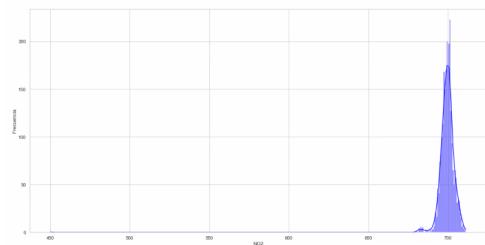


Figura 37: Histograma Dióxido de Nitrógeno

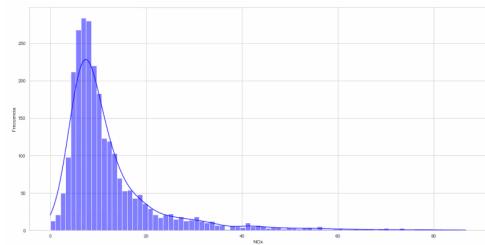


Figura 38: Histograma Óxidos de Nitrógeno

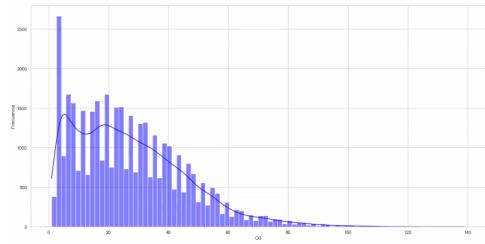


Figura 39: Histograma Ozono

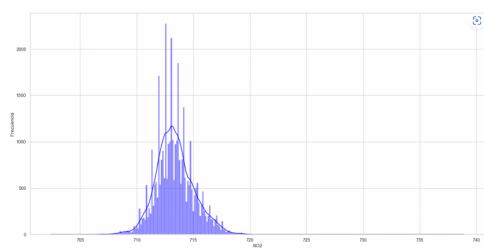


Figura 40: Histograma SO2

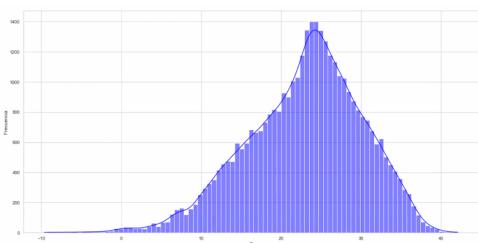


Figura 41: Histograma Temperatura

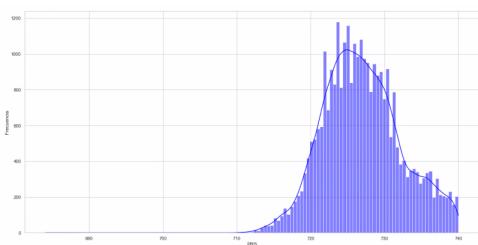


Figura 42: Histograma Presión Atmosférica

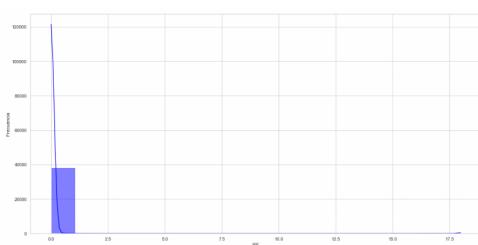


Figura 43: Histograma de Lluvia acumulada en mm

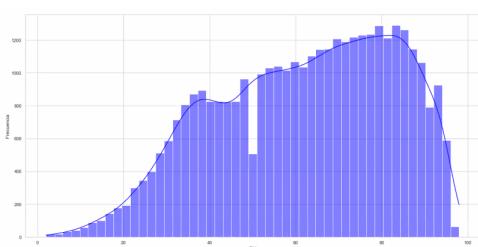


Figura 44: Histograma de Humedad Relativa

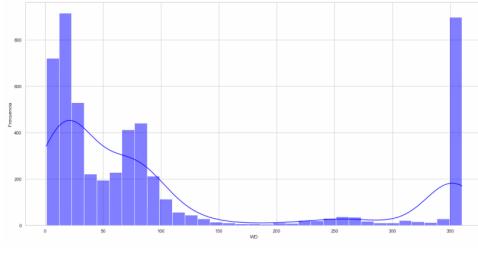


Figura 45: Histograma Dirección del Viento

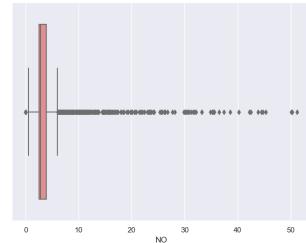


Figura 49: Boxplot Monóxido de Nitrógeno

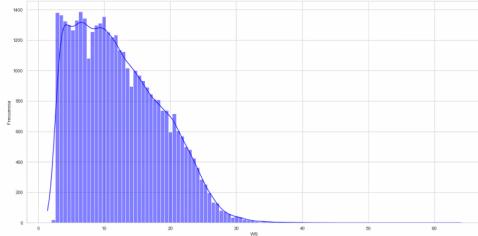


Figura 46: Histograma Velocidad del Viento

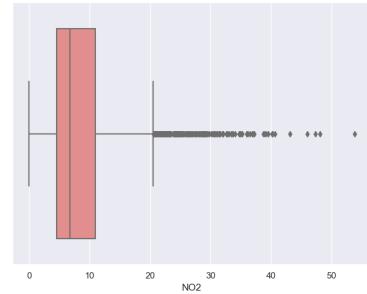


Figura 50: Boxplot Dióxido de Nitrógeno

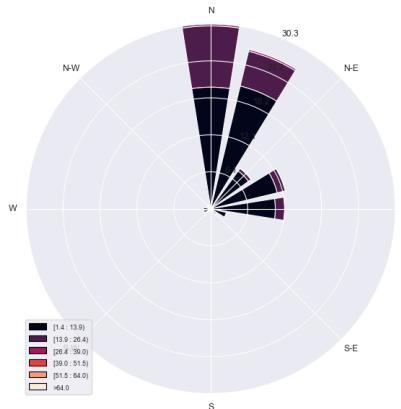


Figura 47: Histograma polar de la dirección del viento. El color indica la velocidad.

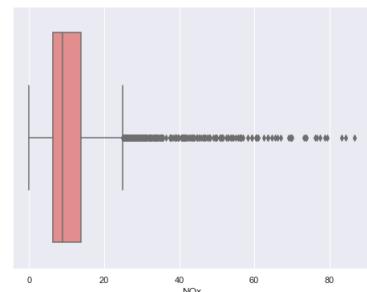


Figura 51: Boxplot Óxidos de Nitrógeno

### 15.3. Boxplots

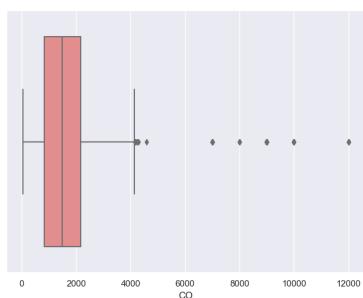


Figura 48: Boxplot Monóxido de Carbono

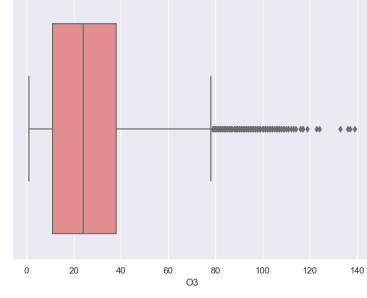


Figura 52: Boxplot Ozono

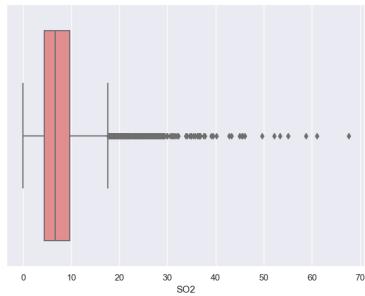


Figura 53: Boxplot SO<sub>2</sub>

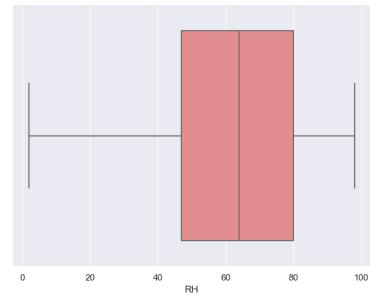


Figura 57: Boxplot Humedad Relativa

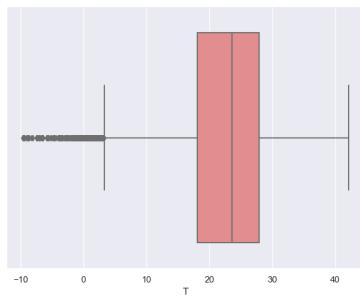


Figura 54: Boxplot Temperatura

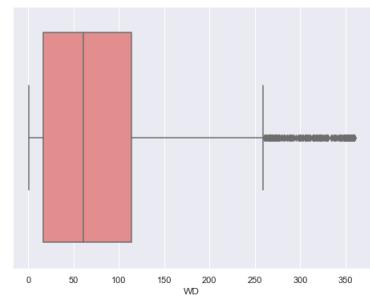


Figura 58: Boxplot Dirección del Viento

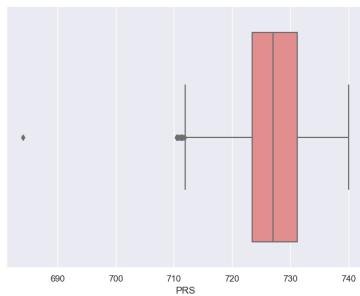


Figura 55: Boxplot Presión Atmosférica

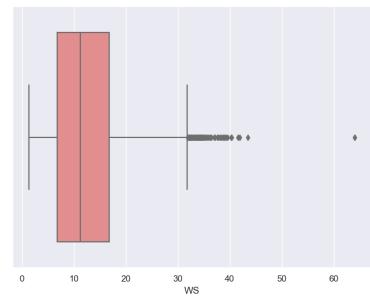


Figura 59: Boxplot Velocidad del Viento

#### 15.4. Gráficas de densidad

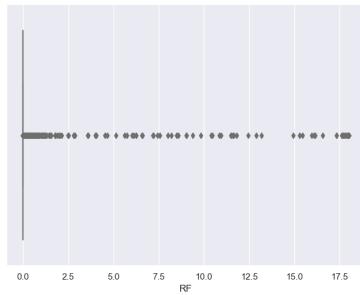


Figura 56: Boxplot de Lluvia acumulada en mm

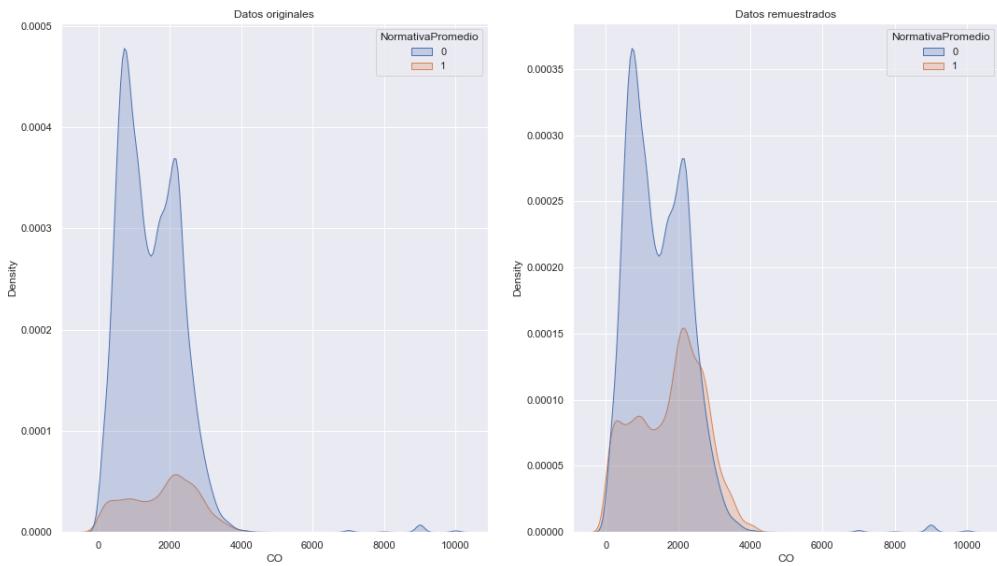


Figura 60: Gráfica de densidad de CO antes y después del remuestreo.

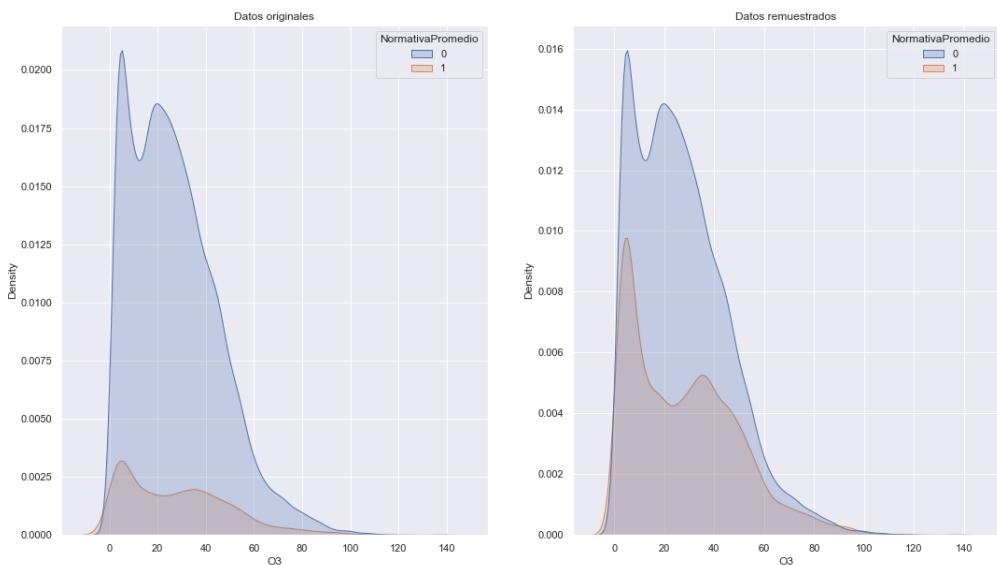


Figura 61: Gráfica de densidad de O<sub>3</sub> antes y después del remuestreo.

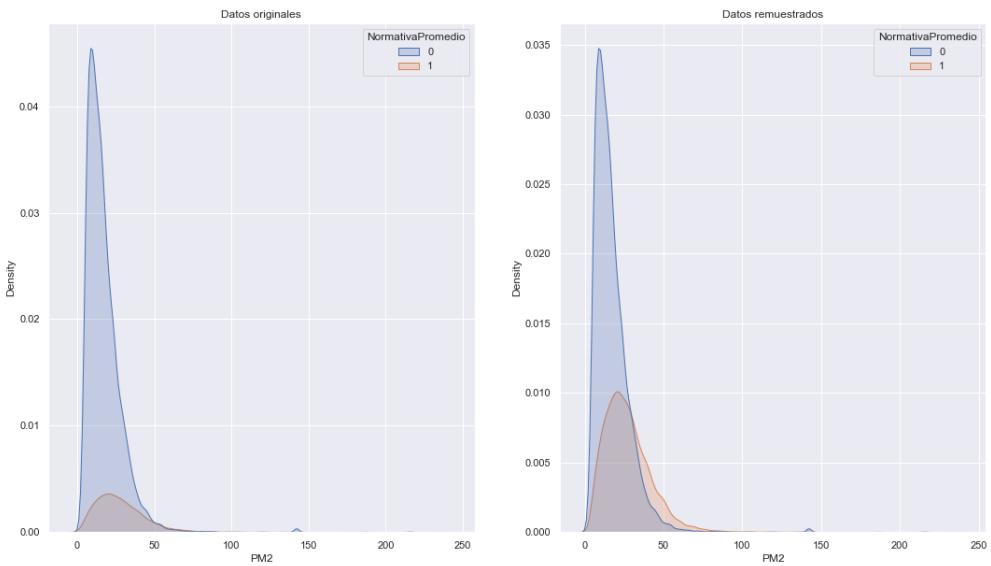


Figura 62: Gráfica de densidad de PM2 antes y después del remuestreo.

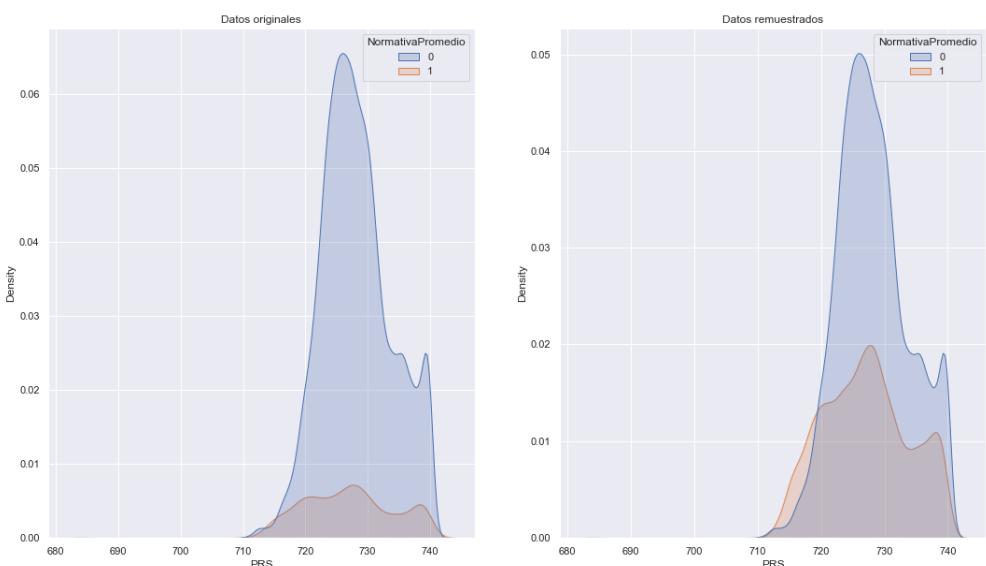


Figura 63: Gráfica de densidad de PRS antes y después del remuestreo.

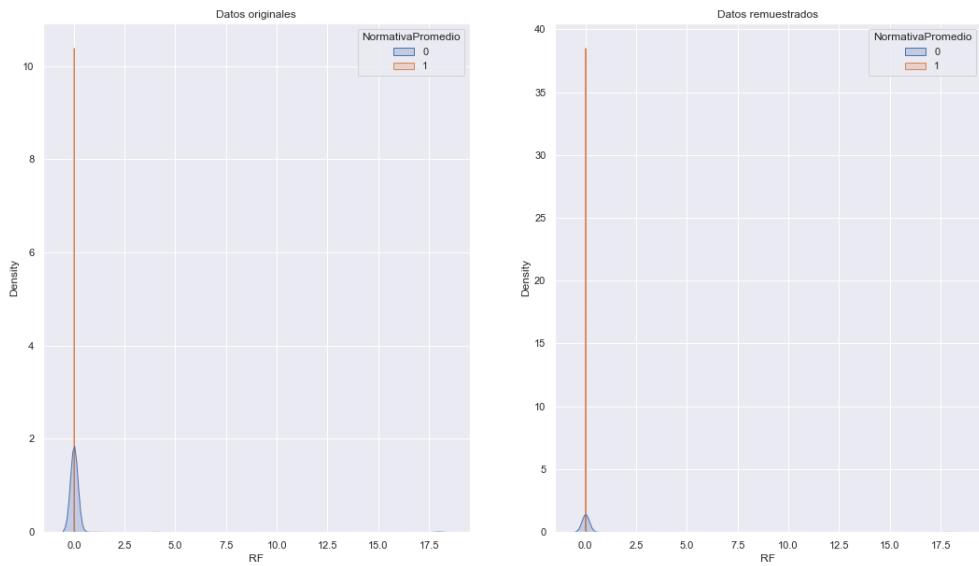


Figura 64: Gráfica de densidad de RF antes y después del remuestreo.

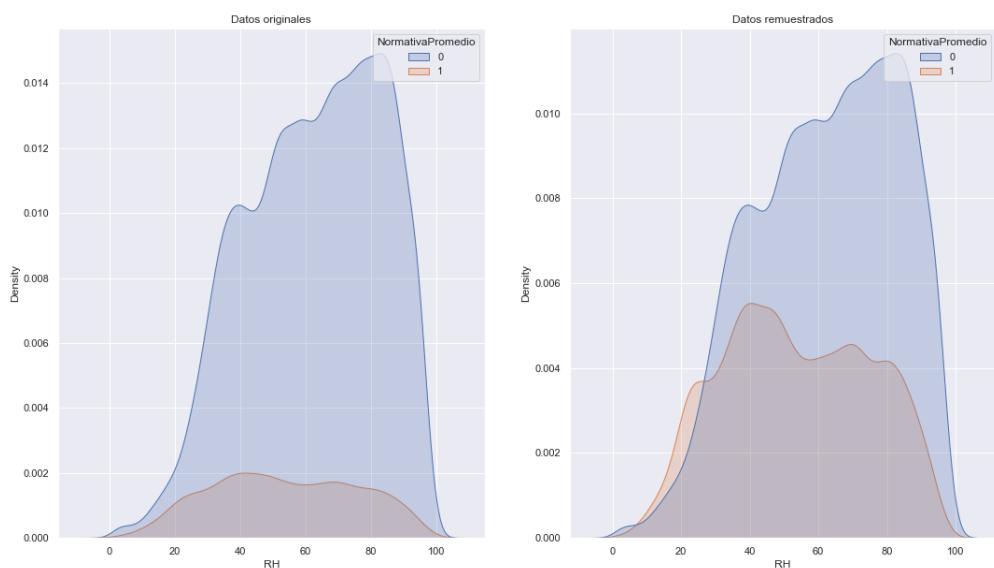


Figura 65: Gráfica de densidad de RH antes y después del remuestreo.

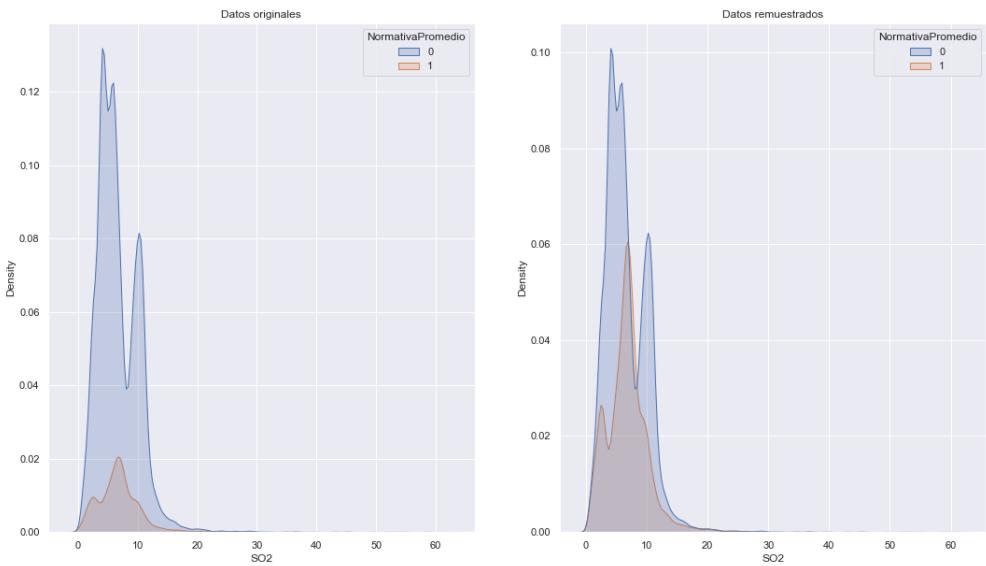


Figura 66: Gráfica de densidad de SO<sub>2</sub> antes y después del remuestreo.

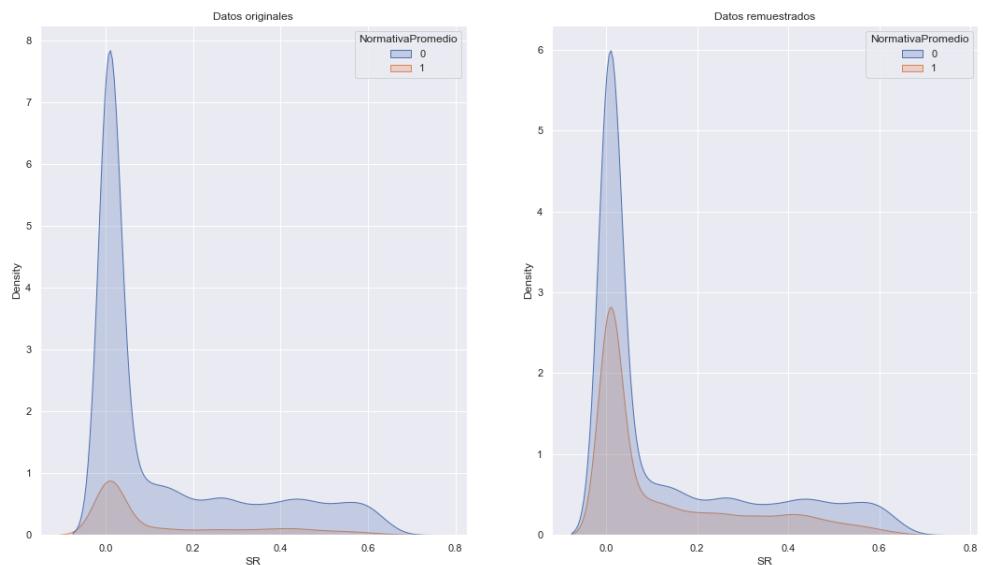


Figura 67: Gráfica de densidad de SR antes y después del remuestreo.

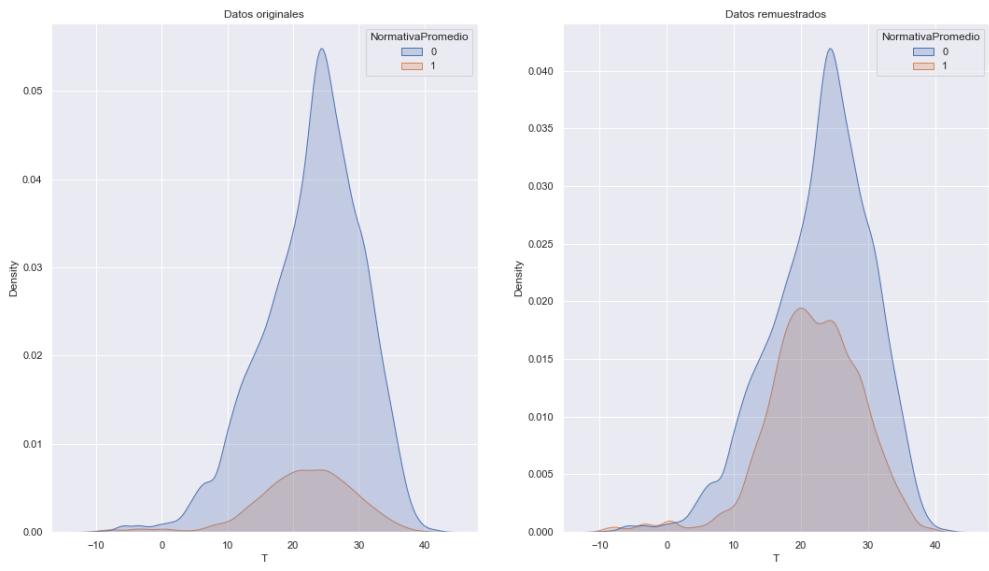


Figura 68: Gráfica de densidad de  $T$  antes y después del remuestreo.

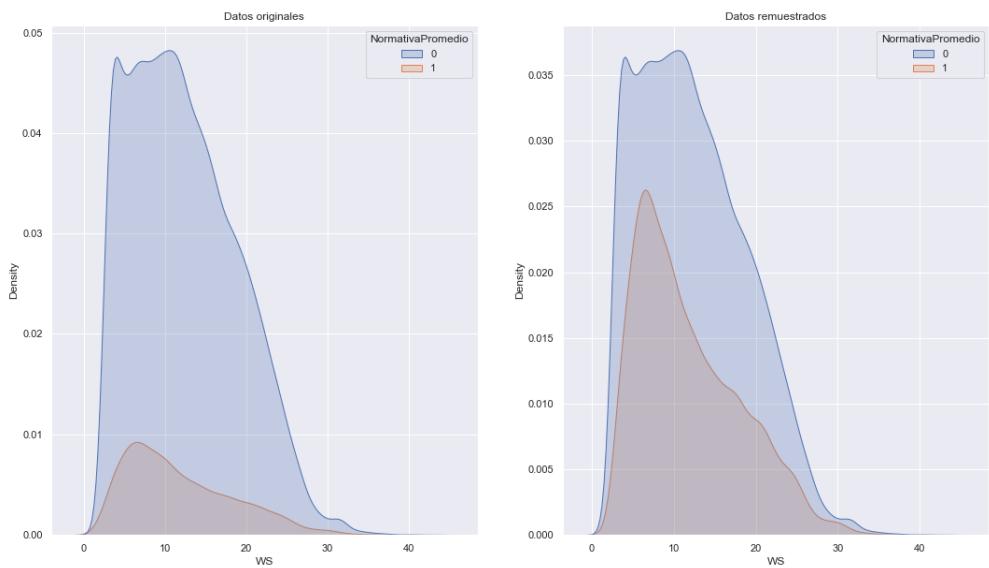


Figura 69: Gráfica de densidad de WS antes y después del remuestreo.

## Referencias

- [1] OMS, “Contaminación del aire ambiente (exterior),” 9 2021. [Online]. Available: [https://www.who.int/es/news-room/fact-sheets/detail/ambient-\(outdoor\)-air-quality-and-health](https://www.who.int/es/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health)
- [2] “Calidad del aire ambiente (exterior),” 9 2021. [Online]. Available: [https://www.who.int/es/news-room/fact-sheets/detail/ambient-\(outdoor\)-air-quality-and-health](https://www.who.int/es/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health)
- [3] SEMARNART, “Calidad de Aire: una práctica de vida,” *SEMARNART*, no. 1, 2013. [Online]. Available: <https://biblioteca.semarnat.gob.mx/janium/Documentos/Ciga/Libros2013/CD001593.pdf>
- [4] “¿Cómo se mide la calidad del aire?” 3 2022. [Online]. Available: <https://www.fundacionaqua.org/wiki/como-se-mide-calidad-aire/>
- [5] OPS/OMS, “Calidad del aire,” 2022. [Online]. Available: <https://www.paho.org/es/temas/calidad-aire#:~:text=La%20exposici%C3%B3n%20a%20altos%20niveles,vulnerable%2C%20ni%C3%B1os%2C%20adultos%20mayores%20y>
- [6] “Manual 1. Principios de Medición de la Calidad del Aire. Instituto Nacional de Ecología,” 10 2022. [Online]. Available: <https://sinaica.inecc.gob.mx/archivo/guias/1-%20Principios%20de%20Medici%C3%B3n%20de%20la%20Calidad%20del%20Aire.pdf>
- [7] C. F. para la Protección contra Riesgos Sanitarios, “Normas Oficiales Mexicanas (NOM) de Calidad del Aire Ambiente,” 12 2017. [Online]. Available: <https://www.gob.mx/cofepris/acciones-y-programas/4-normas-oficiales-mexicanas-nom-de-calidad-del-aire-ambiente>
- [8] OPS/OMS, “Contaminación del aire ambiental exterior y en la vivienda: Preguntas frecuentes,” 2018. [Online]. Available: <https://www.paho.org/es/temas/calidad-aire-salud/contaminacion-aire-ambiental-exterior-vivienda-preguntas-frecuentes#:~:text=La%20exposici%C3%B3n%20a%20altos%20niveles,cerebrovascular%C2%80y%C2%80c%C3%A1ncer%20de%20pulm%C3%B3n>
- [9] R. W. P. . J. M. K. . G. K. . G. H. . Y. A. . S. Terry, “Opportunities and challenges for filling the air quality data gap in low- and middle-income countries,” 8 2019. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1352231019304236?via%3Dihub>
- [10] B. J. . J. Kim, “Development of an IoT-Based Indoor Air Quality Monitoring Platform,” 1 2020. [Online]. Available: <https://www.hindawi.com/journals/js/2020/8749764/>
- [11] J. Novelo, “NORMA Oficial Mexicana NOM-022-SSA1-2019, Salud ambiental. Criterio para evaluar la calidad del aire ambiente, con respecto al dióxido de azufre (SO<sub>2</sub>). Valores normados para la concentración de dióxido de azufre (SO<sub>2</sub>) en el aire ambiente, como medida de protección a la salud de la población.” 8 2019. [Online]. Available: <http://www.aire.cdmx.gob.mx/descargas/monitoreo/normatividad/NOM-022-SSA1-2019.pdf>
- [12] “Partículas PM10.” [Online]. Available: <https://prtr-es.es/particulas-pm10,15673,11,2007.html>
- [13] J. Aunión, “El frío y las escasas lluvias, detrás de los picos de polución,” 12 2016. [Online]. Available: [https://elpais.com/politica/2016/12/28/actualidad/1482961128\\_687402.html](https://elpais.com/politica/2016/12/28/actualidad/1482961128_687402.html)
- [14] C. N. de Prevención de Desastres, “El ozono como contaminante del aire y riesgo para la salud — centro nacional de prevención de desastres — gobierno — gob.mx,” 5 2019. [Online]. Available: <https://www.gob.mx/cenapred/articulos/el-ozono-como-contaminante-del-aire-y-riesgo-para-la-salud>
- [15] J. Allen, “Tango in the Atmosphere: Ozone and Climate Change,” 2 2004. [Online]. Available: [https://www.giss.nasa.gov/research/features/200402\\_tango/](https://www.giss.nasa.gov/research/features/200402_tango/)
- [16] V. F. Rodriguez-Galiano, M. Chica-Olmo, F. Abarca-Hernandez, P. M. Atkinson, and C. Jegannathan, “Random forest classification of mediterranean land cover using multi-seasonal imagery and multi-seasonal texture,” *Remote Sensing of Environment*, vol. 121, pp. 93–107, 6 2012.
- [17] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “Smote: Synthetic minority over-sampling technique,” *J. Artif. Int. Res.*, vol. 16, no. 1, p. 321–357, jun 2002.