Instituto Tecnológico y de Estudios Superiores de Monterrey

Escuela de Ingeniería y Ciencias

Ingeniería en Ciencias de Datos y Matemáticas

Conocimiento de la naturaleza de contaminantes que influyen en la calidad del aire y sus interrelaciones con el medio

Aplicación de métodos multivariados en ciencia de datos - MA2003B.301

Ángel Azahel Ramírez Cabello	A01383328
Annette Pamela Ruiz Abreu	A01423595
Luis Angel López Chávez	A01571000
Jorge Raúl Rocha López	A01740816
Franco Mendoza Muraira	A01383399

Profesoras:

Blanca Rosa Ruiz Hernández

Mónica Guadalupe Elizondo Amaya

Socio Formador: SIMA

Monterrey, Nuevo León

03 de diciembre de 2023

CONTENIDOS

1. INTRODUCCIÓN

La calidad del aire es un tema de relevancia incontestable en la vida de la ciudadanía y, en realidad, para el bienestar general de la humanidad. El análisis de la calidad del aire es un aspecto crucial en la gestión de la contaminación atmosférica, dado que la exposición constante a contaminantes del aire puede tener graves consecuencias para la salud humana y para el entorno ambiental en el que vivimos. En este contexto, la ciudadanía debe estar informada sobre por qué es tan importante evaluar y controlar la calidad del aire, cuáles son los aspectos que se consideran al realizar este análisis, cuáles son los contaminantes más dañinos para la salud, cómo se mide la calidad del aire, qué variables pueden intervenir en la mejora de la calidad del aire, cuáles son las normas oficiales establecidas para proteger la salud y qué dificultades se enfrentan en la medición de la calidad del aire. La importancia de analizar la calidad del aire radica en el impacto directo que tiene sobre la salud y el bienestar de las personas. La exposición a contaminantes atmosféricos puede dar lugar a una amplia variedad de problemas de salud, incluyendo enfermedades respiratorias, cardiovasculares, neurológicas e incluso cáncer. Estos efectos adversos se producen a medida que los contaminantes del aire ingresan al sistema respiratorio y, a través de la inhalación, pueden llegar a afectar otros sistemas y órganos del cuerpo humano. Según el Programa de las Naciones Unidas para el Medio Ambiente, aproximadamente siete millones de personas mueren cada año por enfermedades respiratorias causadas por la contaminación en el aire (UNEP 2017).

A pesar de la importancia del análisis de la calidad del aire y de las leyes que lo promueven, se enfrentan varias dificultades en su medición y control. Las estaciones de monitoreo deben estar distribuidas de manera adecuada para reflejar con precisión la calidad del aire en una región, lo que puede ser un desafío logístico. Además, factores meteorológicos, como la velocidad y dirección del viento, la temperatura y la humedad, pueden influir en la dispersión de los contaminantes, lo que dificulta la interpretación de los datos y la toma de decisiones efectivas. Asimismo, la detección de ciertos contaminantes, como los compuestos orgánicos volátiles, puede requerir equipos de monitoreo más especializados y costosos. Estas dificultades hacen que la gestión de la calidad del aire sea un desafío continuo para las autoridades y la sociedad en general. El Sistema Integral de Monitoreo Ambiental (SIMA) es una plataforma tecnológica diseñada para la recopilación, procesamiento y análisis de datos relacionados con el medio ambiente. Este sistema integra una variedad de sensores, estaciones de monitoreo y dispositivos de medición para evaluar factores ambientales como la calidad del aire, la calidad del agua, la radiación, el ruido y otros parámetros importantes. SIMA permite a las autoridades, científicos y la sociedad en general acceder a información en tiempo real sobre la calidad del ambiente en una determinada región, lo que es esencial para la toma de decisiones informadas, la gestión de recursos naturales y la protección de la salud pública. Además, este sistema puede desempeñar un papel crucial en la identificación de tendencias, la detección temprana de eventos adversos y la formulación de políticas y regulaciones ambientales efectivas. (Belousova, Kuznetsova y Kuznetsov 2009), es por ello que despierta la pregunta: ¿Cómo se diferencían los impactos de las variables en la concentración de PM2.5 según la localización en la Zona metropolitana de Monterrey?

1.1. Problema

La ciudad de Monterrey enfrenta una seria problemática relacionada con la contaminación del aire, situándose entre las más contaminadas de América Latina. Particularmente, la presencia continua de partículas $PM10$ en la atmósfera supera los límites establecidos por la Organización Mundial de la Salud (OMS). De acuerdo con datos de la OMS, los límites recomendados por esta organización son de no más de 10 microgramos por metro cúbico ($\mu g/m^3$) como promedio anual para partículas $PM2{,}5$, y de 20 $\mu g/m^3$ como promedio anual para partículas PM10, sin embargo, en Monterrey estos valores se ven superados de manera constante. Esta situación es alarmante, ya que la mala calidad del aire conlleva graves riesgos para la salud de la población. Este problema de contaminación no se limita únicamente a Monterrey, ya que el Observatorio Ciudadano de la Calidad del Aire de Monterrey (OCCAMM) señala que seis municipios de Nuevo León figuran en la lista de las ciudades más contaminadas de México, incluyendo a Escobedo, Guadalupe, Santa Catarina, San Pedro, San Nicolás y Monterrey, con algunos de los peores índices de calidad del aire a nivel nacional (González Manrique 2023). Por lo que se busca a través de la información recopilada por las diferentes estaciones Metereológicas crear un modelo el cual pueda analizar las relaciones que existen entre las variables que pueden alterar las concentraciones de PM2.5 y con ello poder predecir los días sobre la norma de salud para el próximo mes de diciembre, dado que es uno de los periodos donde tiempo que se detecta una mala calidad del aire durante mayor cantidad de días en la zona metropolitana.

1.2. Justificación

La necesidad de abordar la problemática de la contaminación del aire en Monterrey es evidente debido a sus graves implicaciones para la salud pública y el bienestar de la población. Las cifras muestran que, de acuerdo con datos del Sistema Integral de Monitoreo Ambiental (SIMA) de Nuevo León, la calidad del aire ha reportado valores de hasta 25 microgramos por metro cúbico ($\mu g/m^3$) como promedio anual para partículas $PM2{,}5$, lo que excede considerablemente los límites recomendados por la OMS. Además, según la normativa mexicana, como la NOM-025-SSA1-2021, la mala calidad del aire ha predominado en un 61 % de los días en 2021 y un 71 % en 2022, superando ampliamente las normas oficiales. A pesar de los esfuerzos del gobierno estatal por reducir la contaminación, las cifras demuestran un aumento continuo de los días con mala calidad del aire. Incluso las alertas ambientales han ido en aumento, pasando de 10 en 2020 a 14 en 2022. Además, este problema no solo tiene un impacto negativo en la salud de los habitantes de Monterrey, sino que también tiene implicaciones económicas significativas. De acuerdo con estimaciones del Observatorio Ciudadano de la Calidad del Aire del Área Metropolitana de Monterrey (OCCAMM), la contaminación del aire en la región provoca la pérdida estimada de hasta mil 400 vidas y más de mil 300 millones de dólares al año. Por lo tanto, es crucial abordar de manera efectiva la contaminación del aire en Monterrey para salvaguardar la salud de la población y el bienestar económico de la región (González Manrique 2023). Los meses fríos pueden significar un mayor riesgo para la calidad de aire, esto debido a que el aire frío atrapa más contaminación. Una de las propiedades que tiene el aire frío es que suele a descender (es más denso), mientras que el aire caliente suele a ascender. Esto facilita que el aire que se encuentra cercano a la superficie de la tierra se eleve. Además como

la luz solar es más débil en esa estación del año, el aire que se encuentra cerca de la superficie puede acabar siendo más frío que el aire superior, lo que provoca que el aire superior actué como una tapa y atrape el aire más frió y contaminado que se encuentra. En ciudades contaminadas, esto se puede ver en forma de smog invernal. También puede ocurrir inversiones térmicas lo que puede ser más perjudicial para la salud porque te obliga a respirar contaminantes. (Bannister 2022).

1.3. Objetivos

1. Implementar un modelo de regresión logística , el cual busque predecir la concentración de PM2.5 para los meses de diciembre (Esto porque la contaminación suele ser peor), contemplando las variables de los contaminantes y de las condiciones metereológicas. Se busca comprobar si es que las mismas variables pueden ser usadas para hacer un modelo en las diferentes estaciones y en el caso de que no, se busca poder brindar una posible explicación del porque.

2. Crear un modelo de una serie de tiempo que pueda predecir la calidad del aire en función de la autocorrelación pertinente de las mediciones de PM2.5 a lo largo del tiempo.

1.4. Estaciones utilizadas y contexto

Las estaciones que se escogieron para hacer el análisis fueron **Centro** la cual se ubica en las instalaciones de Agua y Drenaje, en el municipio de Monterrey. La altura de la toma de muestra cumple con los criterios EPA; libre de obstáculos al libre flujo de aire, sin una fuente de emisión fija o móvil que impacte de forma directa sobre la estación. Se ubica viento a favor de muchas fuentes de emisión de sectores económicos diversos lo que la hace susceptible de caracterizar niveles de exposición de la población asociados de diversas fuentes sin estar dominados por una fuente en particular (Barrera et al. 2021). La razón por la cual se seleccionó esa estación es porque es de las que tenía menos datos faltantes.

Se utilizó la estación **Norte2** la cual se encuentra en el estacionamiento del edificio de Graduados en Contaduría Pública y de Administración de la Universidad Autónoma de Nuevo León. Cerca de sus alrededores se encuentran dos vialidades Fidel Velásquez y Gustavo Adolfo Bécquer, que por su cercanía y tránsito podrían estar impactando las mediciones de la calidad de aire (Barrera et al. 2021). Se seleccionó esa estación, ya que es una de las principales zonas industriales del estado.

También se utilizó la estación de **Sureste3** la cual se encuentra en el centro de Rehabilitación Integral del municipio de Cadereyta. La altura de toma de muestra cumple con los criterios EPA, se encuentra libre de obstáculos, se ubica viento a favor de numerosas fuentes de emisión pertenecientes a diversos sectores de actividad destacando Química y Petróleo y Petroquímica. (Barrera et al. 2021) El porqué se seleccionó esta última estación se debe a que se quiere ver el impacto que tiene la refinería de petróleo que se encuentra en Cadereyta, ya que de acuerdo con algunas fuentes (Badillo 2020) esa es la principal fuente de dióxido de azufre (SO_2) y además la quema de combustibles también es un factor importante en la producción de partículas PM2.5. Se investigó que una de las principales fuentes de contaminación del aire en el caso de

PM2.5 son las pedreras, se cuentan alrededor de 33 activas en el estado actualmente (*Las Pedreras en Nuevo León* 2023).

1.5. Impacto social

- Relación causal potencial: Determinar las posibles variables que tengan un mayor impacto en las concentraciones de PM2.5 permite entender de mejor manera, de que forma se relacionan y cualés componentes son los que se deben de procurar reducir para no sobrepasar los niveles permitidos del contaminante PM2.5.

- Planificación y gestión: Con un modelo preciso, las autoridades locales y los planificadores ambientales pueden anticipar y gestionar mejor la calidad del aire en función las variables metereológicas y contaminantes. Esto permite implementar medidas preventivas o correctivas en áreas propensas a altas concentraciones de PM2.5.

- Alertas tempranas: Un modelo de regresión bien ajustado podría proporcionar alertas tempranas sobre posibles aumentos en la concentración de PM2.5 en función de la dirección y el ángulo del viento, permitiendo a las comunidades tomar medidas preventivas.

- Apoyo a la toma de decisiones: Proporcionaría información valiosa para la toma de decisiones en la gestión ambiental, lo que podría llevar a estrategias más efectivas para reducir la contaminación del aire. En el caso de que algún sensor falle o que se tenga una lectura no válida, se podría tener un valor aproximado de la concentración de PM2.5.

2. ANTECEDENTES

2.1. Contaminantes

Para llevar a cabo un análisis adecuado de la calidad del aire, es esencial considerar varios aspectos. En primer lugar, es crucial tener en cuenta la presencia y concentración de contaminantes atmosféricos. Entre los contaminantes más dañinos para la salud se encuentran las partículas finas ($PM2{,}5$), el dióxido de nitrógeno (NO_2), el dióxido de azufre (SO_2), el monóxido de carbono (CO) y el ozono troposférico (O_3) Comisión Federal para la Protección contra Riesgos Sanitarios (COFEPRIS) n.d. Cada uno de estos contaminantes tiene sus propias características y efectos en la salud, por lo que su monitoreo es fundamental. La medición de la calidad del aire se lleva a cabo mediante estaciones de monitoreo que registran la concentración de contaminantes en el aire. Estas estaciones se distribuyen estratégicamente en áreas urbanas y rurales para recopilar datos sobre las concentraciones de los contaminantes a lo largo del tiempo. Los datos recopilados son esenciales para evaluar la calidad del aire en una región específica y para tomar decisiones informadas en cuanto a la gestión de la contaminación atmosférica (Envira n.d.). La mejora de la calidad del aire es un objetivo de suma importancia, y diversas variables pueden intervenir en este proceso. La reducción de emisiones provenientes de fuentes industriales, vehículos y otras actividades humanas, así como la promoción del uso de tecnologías limpias y energías renovables, son algunas de las medidas clave para mejorar la calidad del aire. Además, factores meteorológicos, como la dispersión de los contaminantes, también pueden influir en la calidad del aire y deben ser considerados en cualquier estrategia de mejora.

2.2. La contaminación del aire

Las condiciones meteorológicas están intrínsecamente relacionadas con la contaminación del aire, una de las maneras en las que se relacionan es en la disolución y dispensamiento de los contaminantes que existen en el aire. De acuerdo con el reporte del Consejo Nacional de los Estados Unidos, se estima que alrededor de 1.8 a 3.1 años de vida se pierden, por las personas que viven en las ciudades más contaminadas y debido a la exposición crónica a partículas. Además, cada año alrededor de 4,000 muertes prematuras son ocasionadas por las altas concentraciones superficiales de ozono en los Estados Unidos.

Una persona en promedio requiere alrededor de 13.5 kilogramos de aire cada día, una cantidad demasiada grande comparado con la cantidad de comida o agua que requiere una persona al día (1.2 y 2 kg respectivamente), es por ello que deberíamos de importarnos de la misma manera la calidad del aire como en la comida o el agua que ingerimos.

Los contaminantes atmosféricos son partículas y gases que son llevados por el aire en concentraciones que ponen en peligro la salud y el bienestar de las personas o que perturban el funcionamiento del medio ambiente. Estos contaminantes se suelen clasificar en dos categorías, **los primarios** y **secundarios**. Los contaminantes **primarios** son los que contaminan el aire inmediatamente después de ser emitidos y los **secundarios** producen determinadas reacciones químicas en la atmósfera entre los contaminantes primarios. Algunos ejemplos de algunos compuestos que son contaminantes primarios y el cómo son emitidos son, el

dióxido de azufre que es producido en su mayoría para la generación de electricidad, el monóxido de carbono que es emitido por los vehículos.

Un término que se utiliza a menudo cuando se habla de la contaminación del aire, la materia particulada (PM) el cual es un término general que se utiliza para designar una mezcla de partículas sólidas y gotas líquidas que se encuentran en el aire. El tamaño de estas partículas está relacionado con el daño potencial que puede causar en la salud de una persona (Las partículas con 10 micrómetros de diámetro o menor pueden entrar a los pulmones). Estas partículas se clasifican de acuerdo a su tamaño, las partículas finas tienen menos de 2.5 micrómetros y las partículas gruesas tienen un diámetro mayor a 2.5 micrómetros. Las partículas finas suelen ser emitidas por la combustión de combustibles. Las partículas gruesas usualmente suelen ser emitidas por vehículos que circulan por carreteras sin asfaltar, la manipulación de materiales y las operaciones de trituración. Sin embargo, se puede dar el caso que algunos gases interactúan con otros compuestos del aire para formar partículas finas. Las partículas finas están más relacionadas con problemas de salud y están relacionadas con aumento en la admisión de hospitales, problemas cardiovasculares, enfermedades pulmonares y un aumento en enfermedades respiratorias.

El dióxido de azufre es un gas que se origina de la combustión de combustibles que contienen azufre, como el carbón y aceite. Grandes concentraciones de este gas pueden provocar trastornos respiratorios temporales en niños y adultos que realizan actividades al aire libre. La exposición elevada en personas asmáticas y en personas que hacen deporte puede provocar una reducción en la función pulmonar.

Los óxidos de nitrógeno son gases que se forman en la combustión de alta temperatura de combustibles, cuando el nitrógeno en el combustible reacciona con el oxígeno del aire. Cuando se encuentra en grandes concentraciones, forma gran parte del esmog y contribuye a los problemas de pulmón y corazón. Cuando el aire es húmedo, puede reaccionar con el vapor de agua para formar ácido nítrico, que es una substancia corrosiva que contribuye a la lluvia ácida.

El monóxido de carbono es un gas venenoso que es producido por la combustión incompleta de combustibles como el carbón, petróleo y la madera. Es el contaminante primario más abundante. Puede entrar en la sangre a través de los pulmones y reduce la llegada de oxígeno a los órganos y tejidos del cuerpo. En pequeñas concentraciones puede llegar a provocar sueño, empeora los reflejos y afecta el juicio. En concentraciones altas puede llegar a provocar la muerte.

Los episodios de contaminación atmosférica no suelen producirse por un aumento repentino de la producción de contaminantes, sino más bien por cambios en las condiciones atmosféricas específicas. Los factores clave que influyen en la dispersión de los contaminantes son la fuerza del viento y la estabilidad atmosférica. La velocidad del viento afecta directamente a la concentración de contaminantes, mientras que la estabilidad atmosférica determina cómo los movimientos verticales mezclan la contaminación con el aire más limpio. La profundidad de mezcla, la distancia vertical entre la superficie de la Tierra y la altura de los movimientos convectivos, desempeña un papel crucial. Una mayor profundidad de mezcla conduce generalmente a una mejor calidad del aire. Sin embargo, las inversiones térmicas, caracterizadas por atmósferas estables y una profundidad de mezcla restringida, pueden dar lugar a elevadas concentraciones de contaminación cuando se combinan con vientos suaves, inhibiendo la difusión en zonas con fuentes de contaminación.

(Lutgens y Tarbuck 1979)

2.3. Leyes y Normas

La cooperación del gobierno y la implementación de leyes son esenciales para mejorar la calidad del aire, pues desempeñan un papel fundamental en la protección de la salud pública. Para empezar, la NORMA Oficial Mexicana NOM-172-SEMARNAT-2019 establece una serie de lineamientos para la obtención y comunicación del Índice de Calidad del Aire y Riesgos a la Salud en México. La norma establece los procedimientos y equipos necesarios para llevar a cabo el monitoreo de la calidad del aire. Las estaciones de monitoreo deben ubicarse estratégicamente en todo el país, y se deben medir varios contaminantes clave, como partículas finas ($PM10$ y $PM2{,}5$), dióxido de azufre (SO_2), dióxido de nitrógeno (NO_2), monóxido de carbono (CO) y ozono (O_3). El ICA se calcula tomando en cuenta las concentraciones de los contaminantes mencionados. Se asigna un valor numérico y un color que indica la calidad del aire, que va desde 'Buena' (color verde) hasta 'Muy Mala' (color morado). La ciudadanía puede consultar el ICA en tiempo real a través de diversas plataformas. Cuando el ICA alcanza niveles perjudiciales para la salud, se emiten alertas a la ciudadanía para tomar precauciones, especialmente en grupos vulnerables como niños, personas mayores y aquellos con problemas de salud preexistentes (Comisión Federal para la Protección contra Riesgos Sanitarios (COFEPRIS) n.d.). Por otro lado, el Reglamento de Protección Ambiental e Imagen Urbana de Monterrey establece una serie de restricciones y prohibiciones clave para controlar y prevenir la contaminación del aire en la región. Estas medidas incluyen la prohibición de la quema de residuos a cielo abierto, restricciones en las emisiones industriales, la prohibición de emisiones de compuestos orgánicos volátiles, el control de emisiones vehiculares, restricciones en la quema de neumáticos y materiales tóxicos, así como la regulación de fuentes fijas de contaminación y la reducción de partículas suspendidas $PM10$ y $PM2{,}5$ (Ayuntamiento de Monterrey 2019). Finalmente, la Ley General del Equilibrio Ecológico y la Protección al Ambiente (LGEEPA) es una normativa fundamental en México que tiene como objetivo principal la protección del medio ambiente y la preservación del equilibrio ecológico. Esta ley establece restricciones y regulaciones para prevenir y controlar la contaminación del aire y otros aspectos ambientales, incluyendo la conservación de recursos naturales y la gestión de residuos peligrosos (Cámara de Diputados del H. Congreso de la Unión 2023).

2.4. Límites establecidos para la investigación

- $PM2{,}5$: De acuerdo con lo establecido por la reciente NORMA oficial mexicana NOM-025-SSA1-2021 de Salud ambiental, publicada en el Diario Oficial de la Federación, el límite permisible de $PM2{,}5$ será establecido como una media arímetica de 24 horas con un valor de 33 $\mu g/m^3$ a partir del tercer año la entrada en vigor de la Norma, esto debido a los recientes valores elevados presentes en las zonas metropolitanas del país, lo cual difiere con la NOM-025-SSA1-2014 que marca como mala calidad del aire a partir de un promedio diario de 45 $\mu g/m^3$ y con la comparativa de la normativa de la Organización Mundial de la Salud que establece un valor todavía más bajo como límite permisible (Daño 2021).

(Salud s.f.)

3. DESARROLLO

3.1. Comprensión de los Datos

De acuerdo con los datos proporcionados por la organización socio formadora se analizaron más de quince estaciones meteorológicas que miden la concentración de distintos contaminantes cada cierta cantidad de segundos para hacer un reporte por hora de estos valores, junto con datos meteorológicos del sitio donde está la estación, se nos dio acceso a los registros desde 2020 hasta la fecha, lo cual nos permite analizar un periodo de tiempo reciente que podría ayudar a predecir la lectura de los niveles de los contaminantes en meses próximos, no obstante, para acotar el análisis se utilizarán solamente la información de las estaciones NORTE2, SURESTE3 y CENTRO, puesto que son de interés por sus ubicaciones geográficas, ya que la primera se encuentra en una de las zonas más pobladas de la región, la segunda mide las emisiones de la refinería de PEMEX Ing. Héctor R. Lara Sosa ubicada en el municipio de Cadereyta y finalmente CENTRO es la estación que se encuentra cerca del Obispado.

El conjunto de datos final que se utilizará, consta de un archivo CSV, en los cuales se tienen 16 columnas, la fecha del registro de cada una de las observaciones, las variables que son las diferentes concentraciones de los contaminantes y las variables que son factores metereológicos.

Ya que en las mediciones había muchos datos vacíos, se recurrió a una técnica de imputación. A continuación se presentan las variables que consideramos más importantes y su función para el proyecto:

- **PM10** (numérico): Material Particulado menor a 10 micrómetros, contaminante de alto riesgo en la zona que constantemente levanta alertas en el sistema, se mide en microgramos por metro cúbico. El dominio que puede tomar esta variable de los datos elegidos es de entre 2 a 767 microgramos por metro cúbico. En total, teniendo 1907 datos nulos de 64791, correspondiendo a un 2.94 %.

- **PM2.5** (numérico): Material Particulado menor a 2.5 micrómetros, contaminante de alto riesgo en la zona que constantemente levanta alertas en el sistema, se mide en microgramos por metro cúbico. La variable que se busca predecir. El dominio que puede tomar esta variable de los datos elegidos es de entre 0 a 370 microgramos por metro cúbico. En total, teniendo 5013 datos nulos de 64791, correspondiendo a un 7.74 %.

- **SO2** (numérico): Dióxido de azufre, es un contaminante que se produce por la combustión de combustibles fósiles que tienen azufre como el carbón, el petróleo o el gas natural. Incluso en concentraciones bajas puede traer problemas en la salud. Se mide en microgramos por metro cúbico.

- **NO2** (numérico): Se refiere a la concentración de Dióxido de Nitrógeno que existe en el aire. Se mide en microgramos por metro cúbico.

- **WSR** (numérico): Velocidad del Viento, con esta variable se podrá visualizar la capacidad del ambiente de disipar los gases contaminantes, se mide en Km/hr. El dominio que puede tomar esta variable de los datos elegidos es de entre 0.1 a 62.1. En total, teniendo 1620 datos nulos de 64791, correspondiendo a un 2.50 %.

- **WDR** (numérico): Dirección del viento, con esta variable se puede comprender mejor hacia qué sentido está corriendo el viento, se mide como un ángulo en grados. El dominio que puede tomar esta variable de los datos elegidos es de entre 1° a 360°. En total teniendo 1854 datos nulos de 64791, correspondiendo a 2.86 %.

- **RH** (numérico): Se refiere a la humedad relativa. Con esta variable se mide la relación del vapor de agua en el aire en relación de la máxima cantidad de vapor de agua que este puede contener. Se expresa como un porcentaje, siendo 0 para identificar un aire muy seco, y 100 un aire completamente saturado.

3.2. Preparación de los datos

Para la limpieza de datos, principalmente se encargó del manejo de datos faltantes. Para este caso primero se tomaron en cuenta las horas que no se tuvo registro alguno en cada hora en el periodo de tiempo elegido. Asimismo, se aseguró de que no exista la presencia de registros duplicados para ninguna de las zonas elegidas en el periodo ya establecido.

Una vez haciendo esta limpieza preliminar, se optó por manejar los datos faltantes mediante una imputación. Debido a que ya se tiene un conjunto de datos con una relativa pequeña cantidad de datos nulos, se puede realizar una técnica de imputación con más seguridad. Debido a que visualmente los valores de las variables no sigue una tendencia de una lineal simple y tiene gran variabilidad, se optó por realizar una imputación mediante el promedio de la medición anterior con la siguiente. En este método los valores nulos se igualan a la suma de la anterior observación válida con la siguiente observación válida entre 2, considerando que se encuentran ordenados según el criterio temporal de la fecha de registro.

Asimismo, debido a el propósito de nuestro análisis, optamos por hacer un agrupamiento de promedio aritmético diario. Esto fue para proceder a observar como actúan los datos comparados con las normas establecidas en nuestro país y en el mundo con respecto a nuestra variable de interés, PM2.5. Posteriormente, se hizo una clasificación binaria, separando los días de PM2.5 alto, igual o mayor a 30, asignandoles un 1, estos siendo los días de interés debido a la actualización de la norma para los límites permisibles de PM2.5 en 2014, siendo tomado en cuenta como un límite de promedio de 24 horas para el futuro (Gobernación 2014). Actualmente se tiene un límite de 45 cuando es tomado como malo, pero para la visión que se tiene de llegar a ser un país primer mundista se nos hace de alto interés usar un límite más bajo que se adecúe a normas internacionales, y/o de países más progresivos, esto para la protección de la salud y el bienestar de los habitantes del país de un contaminante tan dañino como lo es el PM2.5; a los demás días se les asigna un 0 como clase negativa.

Finalmente para poder realizar un análisis y pronósticos confiables estadísticamente se optó por usar solamente los datos pertenencientes a los tres diciembres presentes en la base de datos proporcionados por la OSF, en específico se seleccionó este periodo de tiempo debido a la alta cantidad de días sobre la norma de salud existentes a comparación de los demás periodos del año.

Por otro lado, cabe mencionar que también se trabajó con los datos por hora provenientes de las bases de datos originales (ya limpios y con valores nulos imputados) para analizar la estacionalidad de las mediciones

de PM2.5 en cada semana de registros.

3.3. Análisis exploratorio

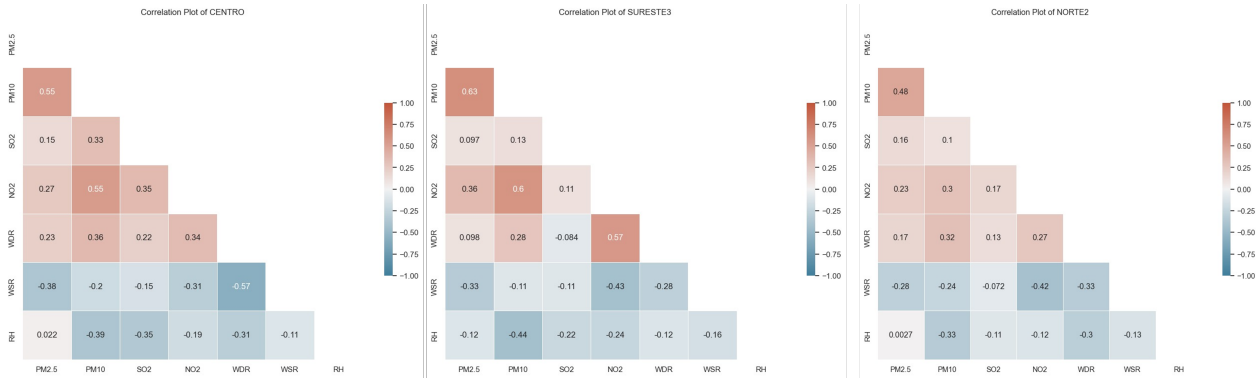

Figura 1: Mapa de calor de las correlaciones de las variables importantes para cada estación utilizada

En la figura 1 se puede observar un mapa de calor con toda las correlaciones con las variables que consieramos de interés. En todas las estaciones el $PM10$ tiene una alta correlación con el $PM2{,}5$, pero con distintos valores de acuerdo con la zona, ya que esta relación depende de la cantidad de combustiones en el ambiente, es por ello que la en estación de Sureste3 en donde tiene un mayor valor, por las actividades constantes de la refinería. Se puede ver que con NO_2 también se tiene un valor relativamente alto para las estaciones de Centro y Sureste3, pero no con la de Norte2. Además se puede observar que con la variable WSR todas parecen tener una correlación negativa, lo cual tiene sentido porque sabemos que el viento es el principal factor en la dispersión de los contaminantes.

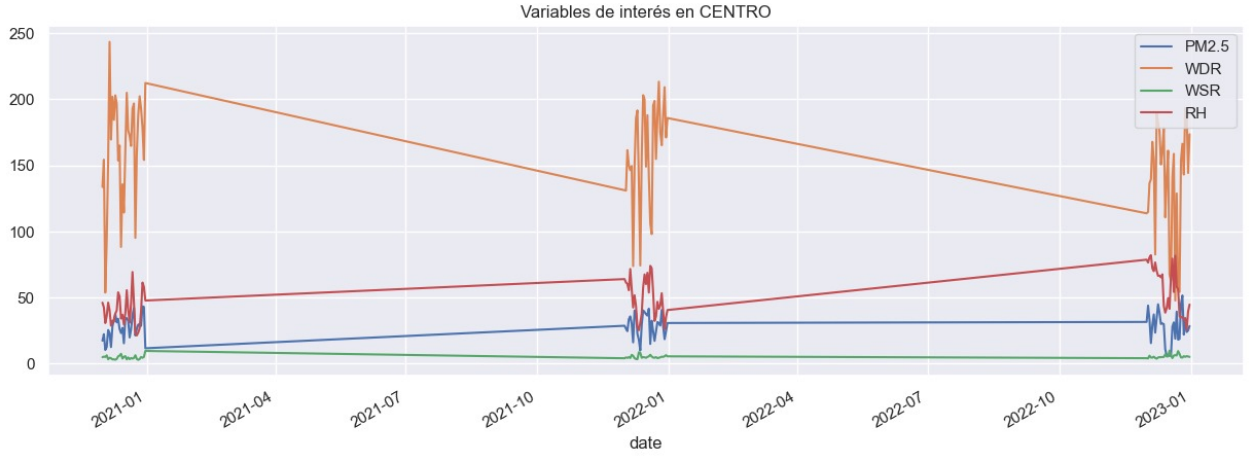

Figura 2: Concentración de PM2.5 en la estación Centro (datos utilizados de cada diciembre de cada año)

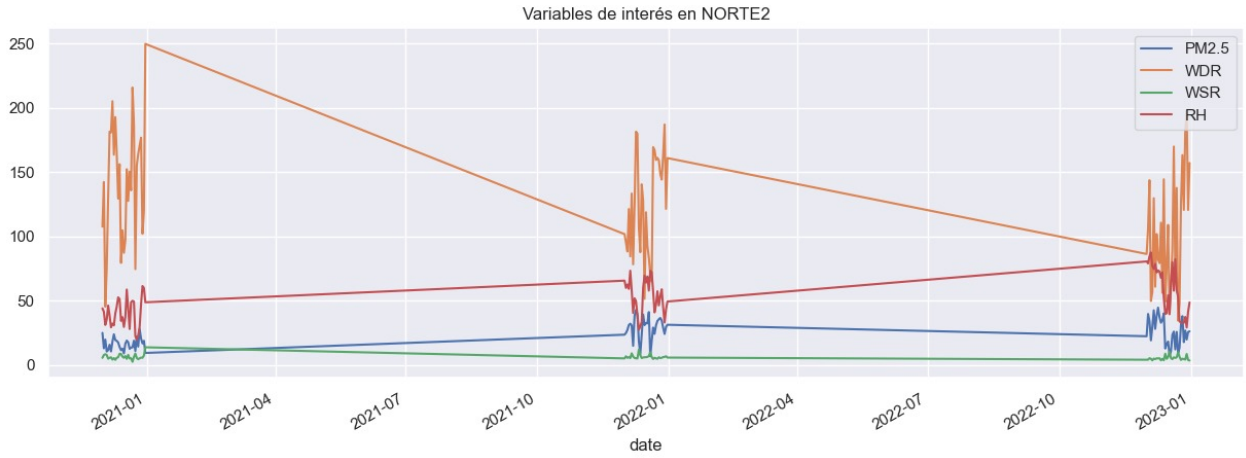

Figura 3: Concentración de PM2.5 en la estación Norte2 (datos utilizados de cada diciembre de cada año)

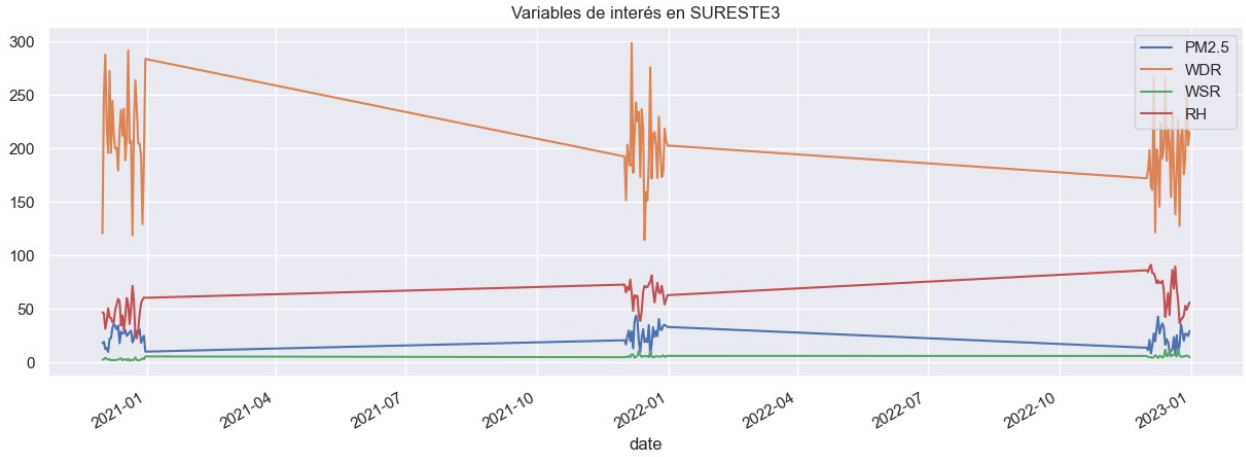

Figura 4: Concentración de PM2.5 en la estación Sureste3 (datos utilizados de cada diciembre de cada año)

Lo que se puede observar en las figuras 2, 3, 4, es como varía a través del tiempo las concentraciones de $PM2{,}5$ y de algunos factores climatológicos. Pareciera ser que al terminar diciembre sería una línea recta, pero esto es porque sólo tomamos los datos de los meses de diciembre. De estas gráficas podemos observar que de las tres estaciones la que tiene una mayor contaminación por $PM2{,}5$ es la estación de Sureste3, después la estación de Centro y finalmente la de Norte2.

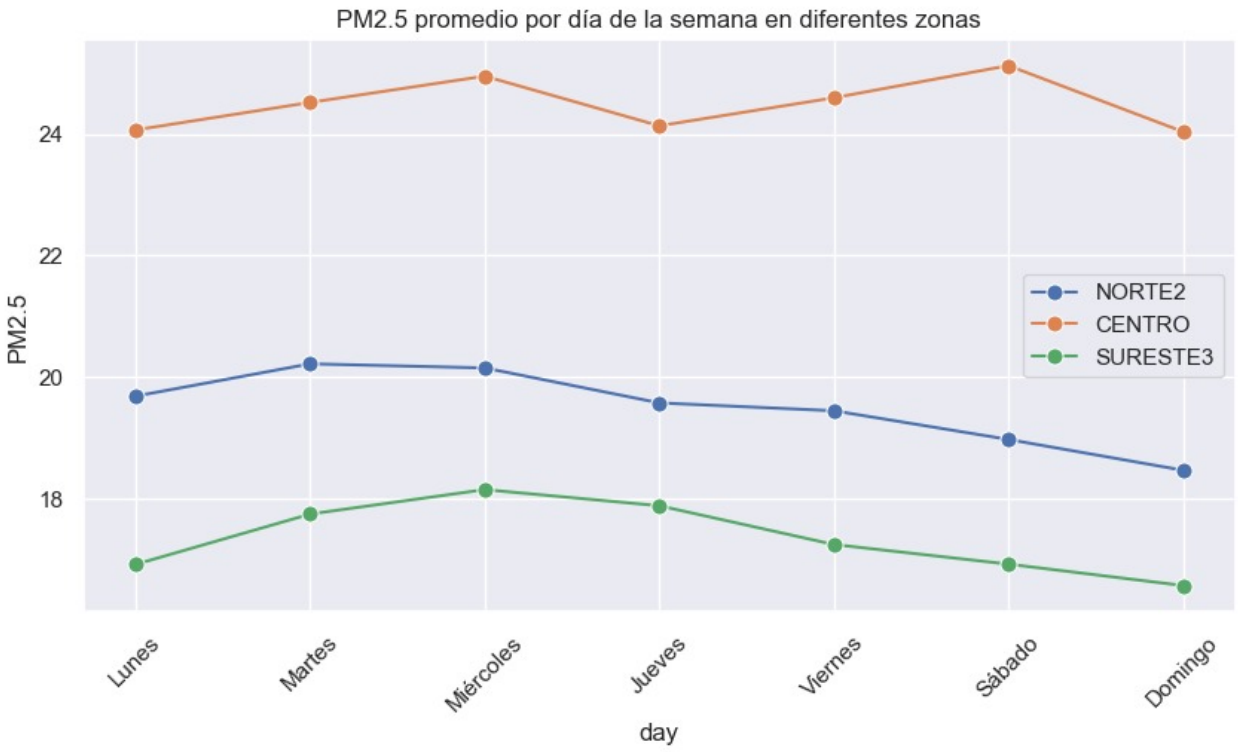

Figura 5: Promedio de concentración de PM2.5 por día de la semana, tomando todos los días (utilizando las 3 estaciones)

En la figura 5 se puede ver la tendencia que tiene la concentraciones de $PM2{,}5$ promedio tomando en cuenta todos los días que se nos fueron proporcionados, se puede apreciar como en casi todas las estaciones los primeros días de la semana, existe una tendencia a la alta; mientras que el período de sábado a domingo tiende a la baja.

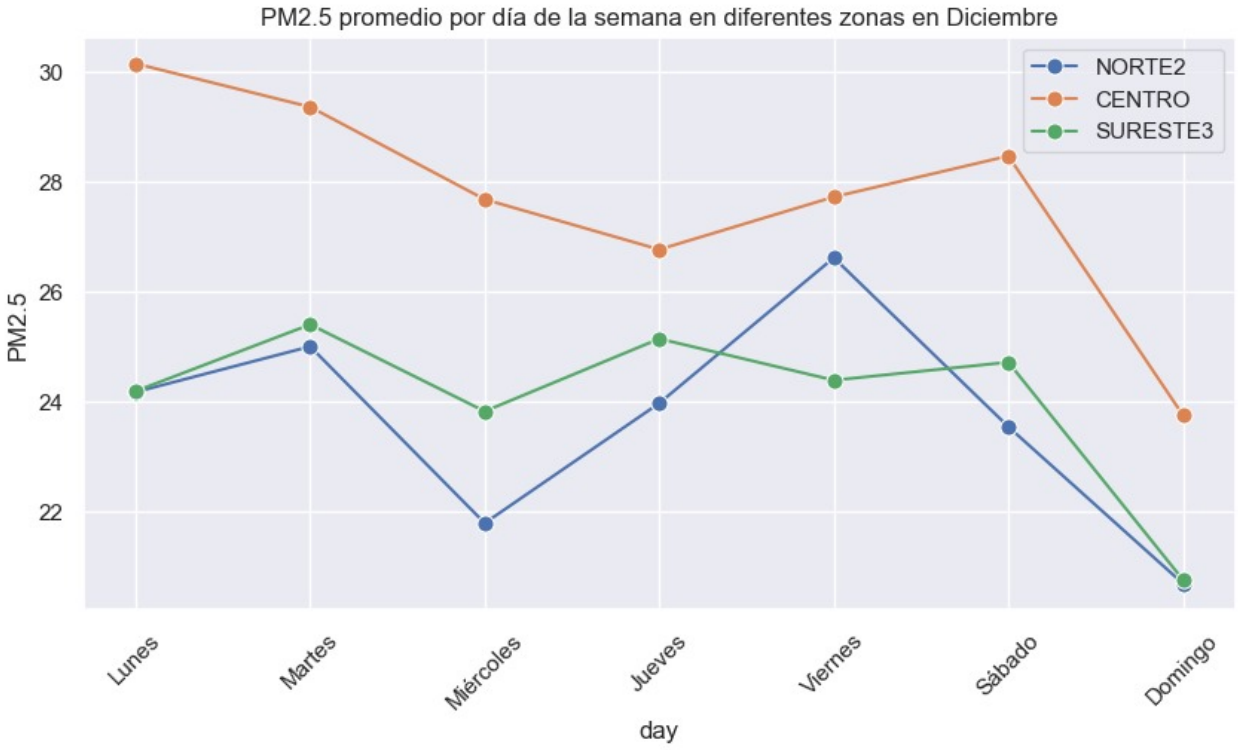

Figura 6: Promedio de concentración de PM2.5 por día de la semana en diciembre

En la figura 6 se puede ver el promedio de los días de la semana para Diciembre, se puede ver algo parecido que en la anterior, es decir que hay una tendencia de que suba las concentraciones a partir del Lunes y que para el Domingo está tiende a bajar. La única estación que pareciera que no se comporta de esa forma, es la estación centro, en la cual las concentraciones bajan y para el día jueves empiezan a subir.

En el siguiente enlace se encuentran los conjuntos de datos finales, junto con los scripts de Python utilizados para la manipulación de la información: Enlace para Datos Finales

4. Modelación y validación

Una vez separados los datos se comenzó por intentar ajustar un modelo de regresión logística binomial utilizando una suma de las variables en cada una de las estaciones de interés, sin embargo, como se pudo encontrar en los script de R (RegLog"Nombre de la estación".Rmd) la significancia de cada una de las variables difería considerablemente para cada caso, por lo que se tuvo que hacer un análisis exhaustivo, usando una división de los datos entre entrenamiento y prueba de forma aleatoria (60 % y 40 % respectivamente), para identificar cuáles combinaciones de variables eran las que podrían presentar un modelo relevante estadísticamente y también para encontrar cuál era el mejor valor de corte a la probabilidad de predecir un día bueno o malo (umbral) usando el comando de R $optCutoff$, por otra parte, también se tomó en cuenta que la regresión logística propuesta para predecir la calidad del aire en un día en función de las partículas $PM2{,}5$ fuera confiable, es decir se buscó encontrar los parámetros indicados para que se cumplieran los cinco supuestos de un método estadístico de esta índole: independencia entre las observaciones, ausencia de multicolinealidad entre las variables predictoras, linealidad entre las probabilidades del logaritmo de la variable respuesta y las variables predictoras, significancia de los coeficientes de las variables predictoras y finalmente un tamaño de muestra significativo (mayor a 10 para el caso menos común). Todo este procedimiento fue llevado a cabo en el mismo script de R mencionado anteriormente (RegLog"Nombre de la estación".Rmd) y con ello se obtuvieron los siguientes resultados.

Ecuación de probabilidad de un día malo (clase 1) en la estación CENTRO, con umbral de 0.4468017:

$$P(\text{día malo}) = \frac{e^{-22{,}05361+0{,}10439\cdot PM_{10}+0{,}04034\cdot WDR+0{,}13722\cdot RH}}{1+e^{-22{,}05361+0{,}10439\cdot PM_{10}+0{,}04034\cdot WDR+0{,}13722\cdot RH}} \tag{1}$$

Ecuación de probabilidad de un día malo (clase 1) en la estación SURESTE3, con umbral de 0.06139201:

$$P(\text{día malo}) = \frac{e^{-13{,}80759+0{,}37629\cdot SO_2+0{,}02458\cdot WDR+0{,}08431\cdot RH}}{1+e^{-13{,}80759+0{,}37629\cdot SO_2+0{,}02458\cdot WDR+0{,}08431\cdot RH}} \tag{2}$$

Ecuación de probabilidad de un día malo (clase 1) en la estación NORTE2, con umbral de 0.07151421:

$$P(\text{día malo}) = \frac{e^{-36{,}78805+0{,}15595\cdot PM_{10}+0{,}32263\cdot NO_2+0{,}15924\cdot RH}}{1+e^{-36{,}78805+0{,}15595\cdot PM_{10}+0{,}32263\cdot NO_2+0{,}15924\cdot RH}} \tag{3}$$

Como se puede ver en cada una de las tres ecuaciones resultantes (1, 2 3) de la regresión logística binomial para analizar la variabilidad de la medición de particulas $PM2{,}5$ en las tres estaciones de interés, se encontraron diferentes variables variables signficativas para su modelo respectivo a partir del conjunto inicial de variables propuestas, destacando que la variable de velocidad de viento no aparece en ninguna de las tres, la humedad resulta ser un factor preponderante en los tres casos y que el coeficiente perteneciente a $PM10$ varía considerablemente entre cada estación, los elementos externos que alteran estos valores serán revisados en la sección de discusión y conclusiones, pero es importante remarcar que cada uno de ellos afecta de forma exponencial a la probabilidad de que el modelo decida que un día es malo o bueno en el conjunto de prueba.

4.1. Supuestos de regresión logística

- Independencia

- Multicolinealidad
- Linealidad entre las variables explicativas y el logit de la variable predictora
- El tamaño de la muestra es suficientemente grande
- Significancia de los coeficientes

Independencia: Se refiere a que las observaciones del conjunto de datos son independientes entre sí. Esto se refiere a que las observaciones no deben de provenir de mediciones repetidas del mismo individuo ni estar relacionadas entre sí de ninguna manera. Una de las maneras más sencillas de comprobar si es que hay independencia es hacer un gráfico con los residuos del modelo contra el tiempo y observar si es que existe una nube de puntos aleatoria (que no haya un patrón aparente). Sin embargo también existe un comando en rstudio **durbinWatsonTest**, en el cual la hipotésis nula es que no existe una correlación entre los residuos y la hipotésis alternativa es que los residuos están autocorrelacionados.

Multicolinealidad: Se refiere a que no haya una alta correlación entre dos o más variables explicativas (independientes). El que dos variables o más tengan una alta correlación puede traer problemas al modelo, ya que no brindan información única o independiente. Una de las maneras más habituales para comprobar si es que este supuesto se cumple es con los coeficientes VIF (Factor de Inflación de la Varianza), el cual mide la fuerza de correlación de las variables independientes de un modelo de regresión, en rstudio se utiliza la función **vif()** para poder calcular estos coeficientes y si es que tiene un valor mayor a 10, entonces no se cumple con este supuesto.

Linealidad variables explicativas y el logit: La regresión logística asume que hay una relación lineal entre cada variable explicativa y el logit de la variable que se quiere predecir. Una de las maneras de comprobar si es que se cumple este supuesto es hacer una prueba Box-Tidwell, en rstudio se utiliza la función **boxTidwell()**, en esta prueba la hipotésis nula es que existe una relación lineal entre la variable predictora y el logit, mientras que la hipotésis alternativa es que no hay una relación lineal.

Tamaño de la muestra: La regresión logística asume que del conjunto de datos es lo suficientemente grande como para obtener conclusiones válidas del modelo de regresión. Existe una regla general, que por lo menos se debe de tener un mínimo de 10 casos con el resultado menos frecuente para cada variable que se quiere predecir.

Significancia de los coeficientes: Se refiere a una prueba de evaluación si los coeficientes de las variables independientes son estadísticamente diferentes de cero. La hipotésis nula es que los coeficiente de la variable independiente es igual a cero, lo que significa que no hay un efecto significativo de esa variable; mientras que la hipotésis alternativa es que es diferente de cero y por lo tanto el coeficiente de esa variable sí es significativo para el modelo.

4.2. Revisión de supuestos en el modelo propuesto

Después de haber generado los tres modelos con las variables correspondientes se llevaron a cabo las pruebas pertinentes para asegurar confianza estadística en el modelo, cuyos valores p resultantes fueron

agrupados en el Cuadro 1 y debido a estos resultados favorables, es posible afirmar que todos los supuestos de los modelos de regresión logística son cumplidos, por lo que los modelos resultan ser confiables (tomando en cuenta que todos los valores p son mayores a un valor de significancia de 0.05 lo cual no rechaza las hipótesis nulas, a excepción de la significancia de los coeficientes de la regresión que requieren rechazar la hipótesis nula para confirmar que son relevantes)

Estación	Significancia	Independencia	Multicolinealidad (VIF)	Colinealidad	Tamaño muestral
NORTE2	PM10 = 0.032 RH = 0.021 NO2 = 0.027	0.3174	PM10 = 1.808 RH = 2.941 NO2 = 2.278	PM10 = 0.2933 RH = 0.09 NO2 = 0.6722	>10 en ambas clases
CENTRO	PM10 = 0.001 RH = 0.002 WDR = 0.006	0.9622	PM10 = 2.177 RH = 2.774 WDR = 1.99	PM10 = 0.3118 RH = 0.1443 WDR = 0.1141	>10 en ambas clases
SURESTE3	SO2 = 0.03 WDR = 0.04 RH = 0.012	0.1751	SO2 = 1.697 WDR = 1.704 RH = 2.587	SO2 = 0.756 WDR = 0.389 RH = 0.799	>10 en ambas clases

Cuadro 1: Valores p de significancia resultantes de cada prueba estadística respectiva con los supuestos

Sin embargo, hay que tomar en cuenta cierta sospecha con respecto a la independencia de NO2 con respecto a las observaciones en la estación NORTE2, debido a que como se puede ver en la Figura 7, los valores de esta variable fueron afectadas por el proceso seguido en el proceso de imputación (donde se tomaron valores promedio en los intervalos de datos nulos), por lo que no se puede asegurar independencia a un 95 % de confianza en el modelo de NORTE2.

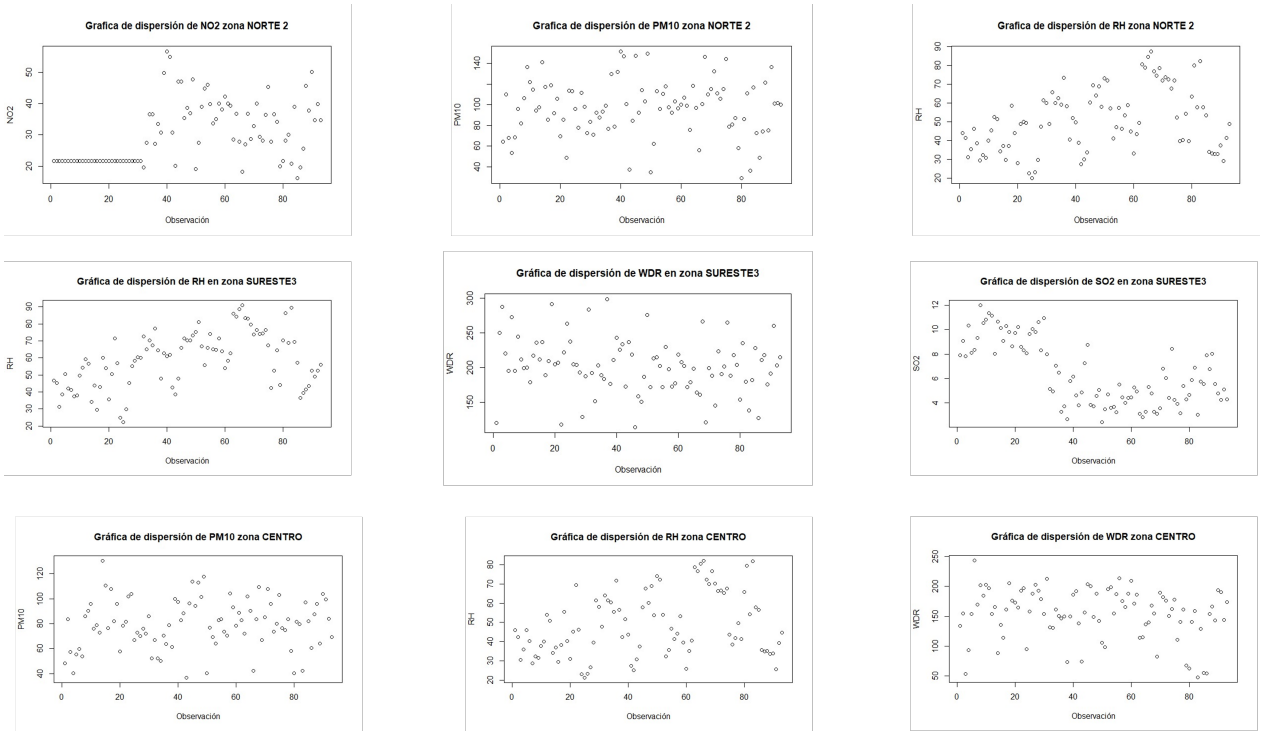

Figura 7: Prueba gráfica de independencia entre observaciones para las tres estaciones

4.3. Supuesto de una serie de tiempo no estacionaria

Por otro lado, para asegurar la efectividad en la predicción de una serie de tiempo estacional, se debe primero asegurar que el periodo de tiempo analizado debe tener un comportamiento no estacionario y para ello se puede aplicar la prueba de Dickey-Fuller que precisamente aborda esta cuestión usando como $H_0 = 0$ La serie no es estacionaria y la hipótesis alternativa como $H_1 \neq 0$ la serie es estacionaria, bajo este supuesto se realizaron pruebas a cada una de las semanas de las tres estaciones de interés, de donde se obtuvo que en los tres casos había menos de 20 de 178 semanas que se trataban de series estacionarias (se pueden consultar los archivos de R SeriesdeTiempo"Nombre de la estación".Rmd para encontrar como fue realizada la prueba en todas las semanas), por lo que los resultados no van a ser totalmente confiables, ya que, hay casos específicos donde este supuesto no se cumple, por otra parte como se va a proponer una regresión lineal de la variable consigo misma como método de predicción, entonces se debería revisar también si los supuestos de esta técnica estadística se cumplen, sin embargo, esto no fue revisado, dado que al tener series estacionarias ya la metodología no es tan confiable, pero todavía funciona como análsis del comportamiento de las lecturas del contaminante $PM2{,}5$ a lo largo del tiempo y de como este factor altera su variabilidad.

5. RESULTADOS

5.1. Regresión Logística zona CENTRO

En la zona CENTRO, después de encontrar como la combinación de variables significativas las de PM10, WDR, y RH, también escogiendolas debido a la zona, y a sus diferencias con las demás, pudimos conseguir la siguiente matriz de confusión, representando las clasificaciones verdaderas de los días buenos y malos, comparado con las clasficiaciones predichas de estos días.

		Clasificaciones Verdaderas	
		Día Bueno	Día Malo
Clasificaciones Predichas	Día Bueno	21	2
	Día Malo	3	11

Teniendo estos resultados de la regresión logística en la zona CENTRO pudimos obtener las siguientes métricas que encontramos ser de mayor interés para evaluar nuestro modelo:

- Se obtuvo una **exactitud**, predicciones correctas, del 86.49 %, solo habiendo obtenido 5 predicciones incorrectas en el modelo. Este modelo nos dió una gran exactitud, para dar más confianza en su uso.
- La **sensibilidad** del modelo fue de 84.62 %, esta métrica clasificando los verdaeros positivos, y teniendo nuestra clase positiva como la de días malos, nos muestra un gran resultado debido al mayor interés que se debe tener de no clasificar días malos como buenos, ya que puede traer peligro el mal informar en esas condiciones.
- El valor de **specificidad** fue de 87.5 %, también siendo un resultado muy bueno del que no hay que quejarse, pero debido a que este valor nos dice la cantidad de días buenos que fueron clasificados correctamente, se debe tomar en cuenta que el valor de **sensibilidad** debería buscar ser mayor al de **specificidad**.

5.2. Regresión Logística zona SURESTE 3

En la zona SURESTE 3, se encontró como la mejor combinación de variables significativas las de SO2, WDR y RH, estas variables siendo importantes en la zona debido al contaminante seleccionado (SO2) siendo el que la caracteriza debido a la refinería encontrada en Cadereyta.

		Clasificaciones Verdaderas	
		Día Bueno	Día Malo
Clasificaciones Predichas	Día Bueno	4	1
	Día Malo	21	11

Teniendo estos resultados de la regresión logística en la zona SURESTE 3 pudimos obtener las siguientes métricas que encontramos ser de mayor interés para evaluar nuestro modelo:

- La **exactitud** del 40.54 % indica que el modelo acertó en sus predicciones solo en un poco más del 40 % de los casos. Aunque este valor es bajo en comparación con el modelo anterior, aún puede proporcionar cierto nivel de confianza en las predicciones realizadas.
- La **sensibilidad** del 91.67 % es bastante alta, lo que indica que el modelo está identificando correctamente la mayoría de los casos de la clase positiva (en este caso, los días malos). Este es un resultado alentador, ya que minimiza los falsos negativos, es decir, los días malos que se clasifican incorrectamente como buenos.
- La **especificidad** del 16 % muestra que el modelo está teniendo dificultades para identificar correctamente la clase negativa (los días buenos). Esta baja especificidad significa que hay una alta tasa de falsos positivos, es decir, días buenos clasificados incorrectamente como malos.

5.3. Regresión Logística zona NORTE 2

En la zona NORTE 2, se escogieron las variables PM10, RH, y NO2, debido a la significancia que se encontró en el modelo, y la alta presencia de NO2 en el área debido al alto tráfico vehicular.

		Clasificaciones Verdaderas	
		Día Bueno	Día Malo
Clasificaciones Predichas	Día Bueno	17	3
	Día Malo	6	11

Teniendo estos resultados de la regresión logística en la zona NORTE 2 pudimos obtener las siguientes métricas que encontramos ser de mayor interés para evaluar nuestro modelo:

- La **exactitud** del 75.68 % muestra que el modelo está acertando en sus predicciones en aproximadamente el 75.68 % de los casos. Esta es una mejora significativa en comparación con el primer conjunto de datos y sugiere una capacidad razonable para predecir con precisión en general.
- La **sensibilidad** del 78.57 % indica que el modelo está identificando correctamente alrededor del 78.57 % de los casos de la clase positiva (por ejemplo, días malos). Esta es una tasa aceptable y muestra una buena capacidad del modelo para detectar los casos de interés (días malos) sin dejar muchos falsos negativos.
- La **especificidad** del 73.91 % muestra que el modelo está identificando correctamente aproximadamente el 73.91 % de los casos de la clase negativa (días buenos). Esta es una mejora con respecto al

segundo conjunto de datos y muestra que el modelo está siendo más preciso al clasificar los días buenos, reduciendo la cantidad de falsos positivos.

5.4. Predicción de serie de tiempo no estacionaria

Una vez aclarada la parte de los supuestos para poder predecir una serie de tiempo que se encuentra como no estacionaria, se procedió a tratar de pronósticar la calidad del aire de acuerdo con la misma clasificación de la regresión logística sobre el contaminante $PM2{,}5$ para los siguientes cuatro días a partir de los registros por hora de una semana, cuya representación visual se puede ver en la Figura 8, donde se tomó una muestra aleatoria de la estación NORTE2 para demostrar como se puede desestacionalizar la serie de tiempo de la semana anterior (puntos negros) para generar un modelo de regresión lineal capaz de predecir la tendencia de la serie en los siguientes registros (como se mencionó en la parte de modelación este modelo de regresión no se encuentra confiable ya que no logra cumplir todos los supuestos estadísticos).

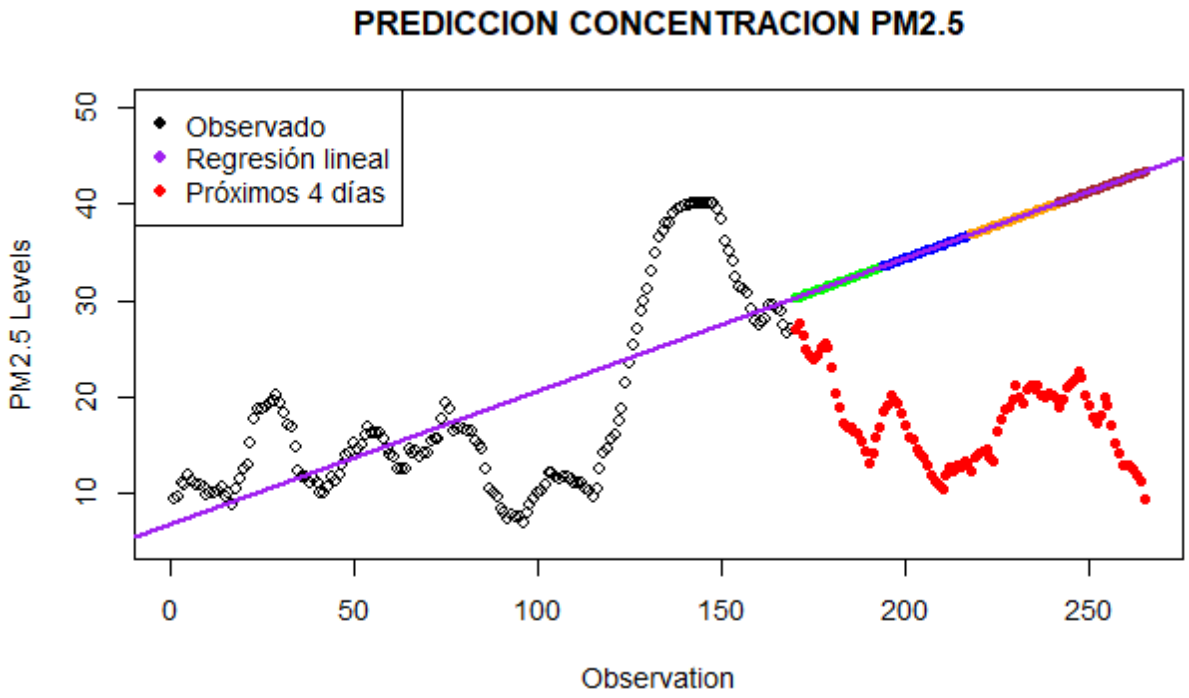

Figura 8: Predicción de serie de tiempo de niveles de PM2.5

Con el modelo se trató de clasificar la calidad del aire como 0 y 1 en los siguientes cuatro días, los resultados promedio de aplicar el modelo en todas las semanas de las tres estaciones se resumen en el Cuadro 2, donde los números del 1 al 4 representan la predicción de cada uno de los días consecutivos después de entrenar el modelo con una semana de información. En el Cuadro 2, MSE se refiere al error cuadrático medio de las predicciones con respecto a los cuatro días de prueba, DIF se refiere a la diferencia entre los promedios por 24 horas propuestos por el pronóstico para el día n y los verdaderos promedios de los siguientes días,

con Exactitud se refiere al porcentaje de acierto para responder correctamente la calidad del aire (promedio arímetico 24 h) en un día pronosticado, CME se refiere al promedio de los cuadrados de los errores entre la regresión lineal y los datos de la semana de entrenamiento y EPAM quiere decir promedio de los errores porcentuales entre el modelo y los datos de la semana de entrenamiento.

Estación	MSE	Diferencia promedios	Exactitud	CME	EPAM
NORTE2	MSE 1 = 106.322	DIF 1 = 6.396			
	MSE 2 = 157.909	DIF 2 = 8.965	85.41 %	45.142	34.86 %
	MSE 3 = 181.036	DIF 3 = 9.501			
	MSE 4 = 231.779	DIF 4 = 10.648			
CENTRO	MSE 1 = 117.735	DIF 1 = 7.006			
	MSE 2 = 187.899	DIF 2 = 9.908	76.82 %	63.132	29.08 %
	MSE 3 = 284.97	DIF 3 = 11.870			
	MSE 4 = 328.711	DIF 1 = 12.999			
SURESTE3	MSE 1 = 95.67	DIF 1 = 6.656			
	MSE 2 = 138.522	DIF 2 = 8.638	88.67 %	40.1844	37.5 %
	MSE 3 = 179.242	DIF 3 = 10.05			
	MSE 4 = 238.087	DIF 4 = 11.624			

Cuadro 2: Valores- p de significancia resultantes de cada prueba estadística respectiva con los supuestos

Como se puede observar en el Cuadro 2, las medidas de error con respecto a la predicción de los cuatro días fueron aumentando conforme se intentaba predecir más hacia el futuro, lo cual tiene mucho sentido con lo propuesto con este modelo tan simple, ya que solamente se toma en cuenta los factores de estacionalidad de una semana y al tratar de predecir al cuarto día no se sabe con certidumbre si esos factores han sido alterados severamente por algún factor externo al tiempo, no obstante, pareciera ser que las predicciones son relativamente buenas para predecir la calidad del aire promedio en las siguientes 24 horas, aunque cabe aclarar que no son del todo confiables al no cumplir los supuestos de una regresión lineal.

6. DISCUSION Y CONCLUSIONES

Pudimos concluir que en la zona sureste 3 donde está Cadereyta, una variable significativa fue el azufre (SO2) un día de contaminación malo, mientras que en las demás zonas se centraba más en PM10 o NO2.

En la zona Norte 2, debido a su alto tránsito, se vio como la combustión de los autos (NO2) impactó significativamente a la predicción de malos días de PM2.5, lo cual tiene sentido debido a la cercanía que tiene la zona de avenidas y vialidades.

También se pudo observar que la humedad (RH) fue una variable siempre presente a través de las zonas, indicando su importancia para predecir los niveles de contaminación independientemente de la zona, esto se debe principalmente a que en ambientes húmedos las partículas de aire y otros contaminantes se agrupen y se depositen más rápido.

Finalmente, se encontró que la variabilidad de registros del contaminante PM2.5 depende de distintos factores al medirse en distintas estaciones y que el tiempo es un elemento muy relevante para pronosticar su comportamiento, aunque este no siempre afecta de la misma forma durante todas las estaciones del año.

Entre algunas áreas de oportunidad que nos hubiera gustado poder hacer, es que debido a que los datos no seguían una distribución normal, no se optó por hacer diferentes modelos. Se intentó hacer que los datos tuvieran una distribución normal a partir de diferentes tranformaciones (Box-cox, YeoJhonson) sin embargo no se pudo hacer que se cumpliera ese supuesto. Algo que se podría haber considerado era separar los datos en diferentes períodos de tiempo, (semanas, meses) para ver si es que en algún un período de tiempo se podría haber hecho otro tipos de modelos. Otra cosa que se pudo hacer, sería una mejor técnica de imputación de los datos, ya que esto pudo haber metido ruido en los modelos. Se pudo haber cambiado la variable dependiente; es decir en lugar de predecir la concentración de $PM2{,}5$ se pudo haber creado otro modelo con otras variables y buscar diferentes relaciones que puedan haber tenido. Hubiera sido bastante útil normalizar (mantenerlos en un rango de 0 a 1) los datos de las variables independientes, esto debido a que las diferentes variables tenían diferentes rangos entre ellas, lo que pueda sesgar a los modelos.

REFERENCIAS

Ayuntamiento de Monterrey (2019). *Reglamento de Protección Ambiental e Imagen Urbana de Monterrey*. URL: https://portal.monterrey.gob.mx/pdf/reglamentos/Reg_proteccion_ambiental.pdf.

Badillo, Diego (ago. de 2020). "Cadereyta, su refinería y cuatro décadas de pasivos ambientales". En: URL: https://www.eleconomista.com.mx/politica/Cadereyta-su-refineria-y-cuatro-decadas-de-pasivos-ambientales-20200809-0001.html.

Bannister, Marie (2022). *Air pollution in winter: why it's worse and what you can do about it*. URL: https://www.airthings.com/resources/air-pollution-winter.

Barrera, Hugo et al. (2021). *Estudio para el rediseño de la red de monitoreo de la calidad del aire de Monterrey*. Inf. téc.

Belousova, N., T. Kuznetsova y V. Kuznetsov (2009). "El concepto metodológico del sistema de monitoreo ambiental". En: *Revista Internacional de Contaminación Ambiental* 25.3, págs. 135-144. URL: https://www.scielo.org.mx/scielo.php?pid=S0186-72102009000300513&script=sci_arttext.

Cámara de Diputados del H. Congreso de la Unión (2023). *Ley General del Equilibrio Ecológico y la Protección al Ambiente*. URL: https://www.diputados.gob.mx/LeyesBiblio/pdf/LGEEPA.pdf.

Comisión Federal para la Protección contra Riesgos Sanitarios (COFEPRIS) (n.d.). *Efectos a la salud por la contaminación del aire ambiente*. URL: https://www.gob.mx/cofepris/acciones-y-programas/3-efectos-a-la-salud-por-la-contaminacion-del-aire-ambiente.

Daño, Salud Sin (2021). *Guías actualizadas de la OMS sobre la calidad del aire y sus implicancias para los países latinoamericanos*. URL: https://saludsindanio.org/sites/default/files/documents-files/6892/Gu%C3%ADa%20actualizada%20de%20la%20OMS%20y%20sus%20implicancias%20en%20AL.pdf.

Envira (n.d.). *Estaciones de medición de la calidad del aire*. URL: https://enviraiot.es/estaciones-medicion-calidad/.

Gobernación, Secretaría de (2014). *NORMA Oficial Mexicana NOM-025-SSA1-2014, Salud ambiental. Valores límite permisibles para la concentración de partículas suspendidas PM10 y PM2.5 en el aire ambiente y criterios para su evaluación*. URL: https://www.dof.gob.mx/nota_detalle.php?codigo=5357042&fecha=20/08/2014#gsc.tab=0.

González Manrique, E. (feb. de 2023). *Contaminación en Nuevo León: versiones encontradas*. URL: https://verificado.com.mx/contaminacion-nuevo-leon-versiones-encontradas/.

Las Pedreras en Nuevo León (feb. de 2023). URL: https://uvleones.com/las-pedreras-en-nuevo-leon/.

Lutgens, Frederick y Edward Tarbuck (1979). *The atmosphere. An introduction to meteorology*. 12th edition. Vol. Capítulo 13. Pearson.

Salud, Organización Panamericana de la (s.f.). *Calidad del Aire Ambiente*. URL: https://www.paho.org/es/temas/calidad-aire/calidad-aire-ambiente.

UNEP (2017). *La importancia del aire*. URL: https://www.unep.org/es/explore-topics/air/la-importancia-del-aire.

ANEXO

Enlace a los códigos utilizados para este proyecto: Enlace a Google Drive

Nota, los códigos de R de interés se encuentran dentro de la siguiente dirección: $codigos > reto > R > entrega_4$

Nota, los códigos de python de interés (limpieza de la base de datos) se encuentran dentro de la siguiente dirección: $codigos > reto > python$ y la carpeta con el número de entrega indica el orden