PREDICCIÓN DE LA CONTAMINACIÓN ATMOSFÉRICA POR PARTÍCULAS ($PM_{2.5}$ Y SO_2) MEDIANTE UN MODELO DE REGRESIÓN DE CRESTAS

Prediction of Particulate Air Pollution ($PM_{2.5}$ and SO_2) Using a Ridge Regression Model

Laura Julieth Contento Patiño
Universidad Libre, Bogotá, Colombia,
lauraj-contentop@unilibre.edu.co

Camila Andrea Ortiz González
Universidad Libre, Bogotá, Colombia,
camilaa-ortizg@unilibre.edu.co

RESUMEN

La contaminación atmosférica repercute una de las mayores problemáticas ambientales en la actualidad, debido a los graves efectos que genera no solo para el ambiente sino también para la salud humana como consecuencia del acelerado crecimiento socioeconómico, que proviene principalmente de la actividad industrial, lo cual ha provocado un considerable aumento en la concentración de material particulado como el PM 2.5, O3, NO2, SO2 y CO, que son algunas de las variables más incidentes en el monitoreo de la calidad del aire, según el registro elaborado por la ***RMCA*** *(Red de Monitoreo de Calidad del Aire) para la ciudad de Bogotá. Sin embargo, la posibilidad de detectar agentes contaminantes con anticipación, podría contribuir al planteamiento de alternativas y planes de manejo para el control y mitigación que eviten la propagación de estas variables, especialmente en sectores con mayor amenaza ambiental. Por lo cual, el presente artículo tiene como objetivo analizar la eficiencia del algoritmo "Ridge Regression" para el estudio y la predicción en cuanto a la fluctuación media de gases contaminantes como el PM 2.5 y SO2 principalmente, teniendo como zona de referencia la localidad de Kennedy.*

Palabras clave: Algoritmo, contaminación, calidad del aire, gases contaminantes, regresión de cresta, predicciones.

ABSTRACT

Atmospheric pollution is one of the biggest environmental problems today, due to the serious effects it generates not only for the environment but also for human health as a consequence of accelerated socio-economic growth, which mainly denotes industrial activity, which has caused a considerable increase in the concentration of particulate matter such as PM 2.5, O3, NO2, SO2 and CO, which are some of the most incident variables in air quality monitoring, according to the registry prepared by the RMCA (The Air Quality Monitoring Network) for the city of Bogotá. However,

the possibility of detecting polluting agents in advance could contribute to the formulation of alternatives and management plans for control and mitigation that prevent the spread of these variables, especially in sectors with greater environmental threat. Therefore, the objective of this article is to analyze the efficiency of the "ridge regression" algorithm for the study and prediction regarding the average fluctuation of polluting gases such as PM 2.5 and SO2 mainly, taking as a reference area the town of Kennedy.

Keywords: About five key words or phrases in alphabetical order are included, separated by commas, for example, Biomedicine, Degrees of freedom, In-situ, Tracking.

1. INTRODUCCIÓN

En la actualidad, la contaminación atmosférica es una de las principales problemáticas globales, y es definida como la conglomeración de sustancias presentes en la atmósfera, emitidas por factores naturales, como la actividad volcánica, oceánica y antropogénica, relacionada principalmente con la quema de combustibles fósiles para suplir procesos industriales vinculados con el sector automotor y de urbanización (Benavides C; Gaviria G. 2011).

Las partículas dispersas en la atmósfera son nocivas debido a que por sus características y elevadas concentraciones afectan negativamente la salud de los seres vivos.

En el año 2016, la Organización Mundial de la Salud consideraba que aproximadamente el 91 % de la población mundial respiraba un aire sumamente contaminado, lo cual se encuentra relacionado con la muerte de aproximadamente 7 millones de personas en el mundo debido a enfermedades pulmonares y cancerígenas asociadas a dicha problemática, en la que el material particulado (PM) y el dióxido de Azufre (SO2) son dos de los contaminantes más comunes en el aire (Franco D. 2020).

El material particulado (PM) está fuertemente relacionado con el cambio climático, la reducción de la visibilidad y efectos nocivos en la salud humana y es definido como un grupo de partículas que se clasifican principalmente en dos amplios grupos, gruesas (PM_{10}) y finas ($PM_{2.5}$) con diámetros inferiores a los 10 y 2.5µm, respectivamente.

El $PM_{2.5}$ es mucho más nocivo para la salud humana debido a que por su insignificante tamaño puede penetrar fácilmente el sistema respiratorio para depositarse en la tráquea, bronquios e incluso en algunos casos en los alveolos; cabe resaltar que estos pueden presentar un origen primario o secundario: en el primer caso son emitidos directamente al ambiente y en el segundo, se forman en la atmósfera debido a transformaciones químicas (Franco D. 2020).

Por su parte, el dióxido de azufre (SO_2) es un gas incoloro muy soluble en agua presente en la atmósfera cuya principal fuente antropogénica es la combustión de carburantes fósiles; dicho compuesto es tóxico para los seres vivos ya que puede ser absorbido por las mucosas de la nariz y vías respiratorias superiores, debido a su solubilidad en el medio acuoso (González G, Robles S. 2003).

En Bogotá, las principales fuentes de contaminación atmosférica son el transporte, la industria, la suspensión de polvo provocada por vías sin pavimentar y un sinnúmero de actividades industriales entre las cuales destaca los restaurantes y vendedores ambulantes y, del mismo modo, las vías de la ciudad se encuentran muy congestionadas (Farrow A, Chen Y, et al. 2022). Pues en una encuesta realizada por la Secretaría de Movilidad en el año 2019 para Bogotá y 18 municipios aledaños, se demostró que diariamente se realizan más de 18 millones de viajes, de los cuales 15 millones se dan dentro de la ciudad (Secretaría Distrital de Movilidad. 2019), de modo que, en el informe mensual de calidad del aire de Bogotá para febrero del año 2020 realizado por la Secretaría Distrital de Ambiente se obtuvieron 16 excedencias diarias en las concentraciones relacionadas con $PM_{2.5}$; en cambio, para SO_2 no se registraron excedencias, de acuerdo con la Resolución 2254 de 2017 del Ministerio de Ambiente y Desarrollo Sostenible (MADS) con respecto a los niveles máximos permisibles (Secretaría Distrital de Ambiente. 2022).

2. REVISIÓN BIBLIOGRÁFICA

2.1 Calidad del aire en Bogotá

La Red de Monitoreo de Calidad del Aire de Bogotá - RMCAB que viene operando desde 1997, ha registrado las concentraciones de material particulado en la atmósfera hasta la fecha, donde se estima que el PM10 es el contaminante con mayor índice de excedencias, según la norma de calidad del aire, el cual presentó un considerable aumento a partir de 2003, debido a la alta demanda de la actividad industrial. Sin embargo, actualmente la influencia de este contaminante se evidencia particularmente en las emisiones procedentes del transporte público impulsado por motores diésel, en tanto que el CO, viene dado principalmente por la incidencia de vehículos particulares que operan con gasolina (Rojas N.2007).

Por otro lado, se encontró que las zonas occidentales de la ciudad, específicamente las localidades de Puente Aranda, Kennedy y Fontibón, generan los mayores índices de material particulado en el aire. No obstante, aunque los resultados obtenidos por la red de monitoreo, estipulan que la concentración de contaminantes suspendidos en la atmósfera en algunas estaciones de la ciudad se encuentra por debajo de los niveles máximos permisibles establecidos por las normas colombianas y, del mismo modo, resultan inaceptables, de acuerdo con los estándares sugeridos por la Organización Mundial de la Salud – OMS (Rojas N.2007).

2.2 Modelo regresión de cresta (Ridge Regression)

La regresión de cresta es un algoritmo implementado para valorar los coeficientes de los modelos de regresión múltiple cuando las variables independientes están altamente correlacionadas. De modo que los coeficientes no se estiman mediante mínimos cuadrados ordinarios (MCO), sino mediante un estimador de crestas, que tiene una varianza menor que el estimador de MCO. Por lo cual, este método es aplicado para mejorar la precisión en cuanto a la predicción y la interpretabilidad de las estadísticas resultantes, gracias a su alto grado de desplazamiento, el cual permite una mayor eficiencia en los modelos aplicados, al ejercer un mayor número de suposiciones, pues capta las diferentes relaciones presentes en la nube de datos aplicada (Msigwa. O, 2023).

Es así, como la regresión de cresta proporciona una posible solución a la imprecisión de los modelos de regresión lineal con un alto grado de multicolinealidad, debido a su baja varianza y desplazamiento, en comparación con las proyecciones de los mínimos cuadrados, en tanto que la estimación simple por el método de cresta está dada por:

$$\hat{\beta}_R = (\mathbf{X}^\intercal \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\intercal \mathbf{y}$$

Donde:

- ***y*** es la matriz de variables independientes
- ***X*** es la matriz de diseño
- ***I*** es la matriz identidad
- ***λ*** es el parámetro de cresta igual o mayor que cero.

A partir de lo cual, es posible determinar que la regresión de cresta funciona con un conjunto de datos estandarizados, los cuales se dividen en matrices x, y respectivamente, para establecer así una matriz de identidad (Msigwa. O, 2023).

No obstante, este algoritmo reduce la complejidad de los modelos aplicados, lo que evita el sobreajuste cuando se emplea un gran número de variables dentro del conjunto de datos, teniendo en cuenta que, la regresión de cresta no es un modelo en sí mismo, sino simplemente una estimación de los coeficientes, que posteriormente se dispondrán en el modelo de regresión lineal:

$$y = X\beta + \varepsilon$$

donde:

Y es el vector de observaciones de la variable dependiente; N x 1

X es la N x K matriz de regresores (hay K regresores);

β es el K x 1 vector de coeficientes de regresión;

ε es el N x 1 vector de errores.

(Taboga y Marco, 2021)

De hecho, en 2016 se realizó un estudio liderado por la Universidad Politécnica del Madrid, el cual buscaba reducir las emisiones de gases contaminantes liberados principalmente por los motores de combustión interna alternativa, aplicando modelos predictivos de regresión, con el objetivo de extraer combinaciones lineales como características derivadas y posteriormente modelar el sistema como una función no lineal acorde con esas particularidades. De este modo, el modelo posibilitaría la predicción del caudal de las emisiones en cada instante en función de los datos tomados por las variables empleadas (Nova A. 2018).

3. ÁREA DE ESTUDIO Y METODOLOGÍA

3.1. Conjunto de datos de descripción

Bogotá se encuentra ubicada en una altiplanicie de la cordillera oriental de los Andes, con coordenadas de 4°35'56" latitud N y 74°04'51" latitud O a 2.640 msnm. Su extensión aproximada es de 33 Km de norte a sur y 16Km de oriente a occidente, con una temperatura promedio de 13.1°C. Según el IDEAM (Ramírez V. 2017), en la ciudad existen 20 estaciones de monitoreo de la calidad del aire que se encuentran ubicadas en puntos estratégicos, que permiten realizar una vigilancia estricta a las concentraciones de contaminantes y analizar distintas variables meteorológicas como la temperatura, humedad y presión, de modo que diariamente se almacenan aproximadamente 5.000 datos relacionados con la calidad del aire (Secretaría de Ambiente. 2022). Los datos extraídos para realizar la predicción mediante un modelo de regresión de crestas fueron material particulado 2.5 (PM2.5) y dióxido de azufre (SO2) correspondientes a la localidad de Kennedy en de Bogotá desde el día 30 de mayo a la 1:00 hasta el día 24 de octubre a las 00:00 del año 2023.

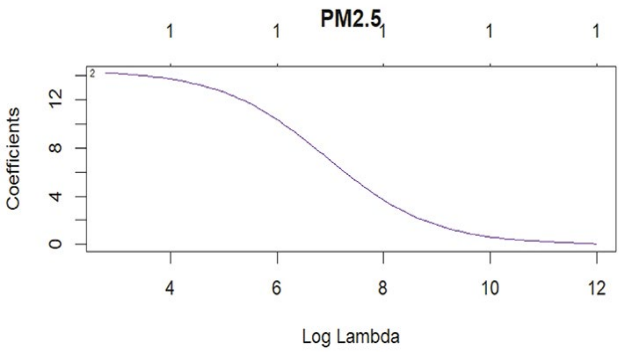
ISSN 2322-8415 / Año 2024 / No. 23 - Vol. 13 / Páginas 65-74 Universidad Libre - Facultad de Ingeniería

4. MODELO

Para realizar la predicción de material particulado (PM2.5) y dióxido de azufre (SO2) se usó un modelo de regresión de crestas que consiste en una extensión de la regresión lineal en el cual la función de pérdida se modifica para minimizar la complejidad del modelo. Los principales argumentos usados fueron nlambda que determina el número de parámetros de regularización que se usaran, lambda que determina los valores lambda que se van a probar y Alpha que depende de la ponderación que se use, cabe resaltar que en el caso de la regresión de crestas este valor será siempre cero.

Para comenzar con la realización del modelo se descargaron los datos de PM2.5 y SO2 de la estación meteorológica de la localidad de Kennedy correspondientes al período de tiempo comprendido entre mayo y octubre del 2023. A continuación, se usó el método de imputación de datos "norm.predict" con el objetivo de completar los datos faltantes en la data. Seguidamente, se crean las variables predictoras y dependientes para aplicar el modelo de regresión de crestas.

Figura I. Gráfica de modelo "Ridge Regression" PM2.5

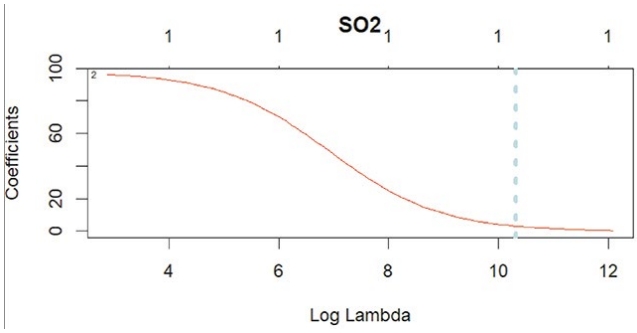


Posteriormente, se analizaron modelos en particular con el objetivo de obtener datos relacionados con el logaritmo de la lambda y sus coeficientes. Además de un gráfico en el que se presenta el modelo por analizar.

Figura 2. Análisis del Modelo 20 para la variable SO2

```
> ridge.model$lambda[20]
[1] 30350.36
> log(ridge.model$lambda[20])
[1] 10.32056
> coef(ridge.model)[,20]
(Intercept) (Intercept)         SO2
1755.950868    0.000000    3.177279
```

Figura 3. Gráfica de modelo 20 "Ridge Regression" para SO2



Posteriormente, se calculó y graficó el mejor Lambda de SO_2 y $PM_{2.5}$, con el objetivo de realizar una predicción con esta variable. Finalmente, se halló el coeficiente de determinación (R^2) y el error cuadrático medio (RMSE), con el objetivo de obtener la eficiencia del método.

5. CRITERIOS DE RENDIMIENTO

Los criterios de rendimiento del modelo de regresión ridge están estipulados por el uso de métricas como el error cuadrático medio (MSE), el modelo de determinación (R2) y la validación cruzada para determinar la calidad del ajuste del modelo y la elección óptima de λ, los cuales están determinados por las siguientes fórmulas:

$$MSE = \frac{1}{m}\sum_{j=1}^{m}(x_i - \hat{x}_i)^2$$

ISSN 2322-8415 / Año 2024 / No. 23 - Vol. 13 / Páginas 65-74

Donde Xi es el pronóstico, m es el número de veces diferentes para las cuales se pronostica la variable, y Xi es el valor real de la cantidad que se pronostica.

$$R^2 = \left[\frac{1}{M} \frac{\sum_{j=1}^{M} \left[(A_j - \bar{A})(B_j - \bar{B}) \right]}{\sigma_a \sigma_b} \right]^2$$

Donde M es el número de observaciones, σb es la desviación estándar de la observación, σa es la desviación estándar del cálculo, Bj es el parámetro observado, B es la media del parámetro observado, Aj es el parámetro calculado, y A es la media del parámetro calculado.

Es así, como la regresión ridge funciona como una técnica de regularización empleada generalmente para mejorar la precisión y el rendimiento de los modelos de regresión lineal, debido a que controla la multicolinealidad y el sobreajuste de las variables.

6. RESULTADOS Y DEBATE

Inicialmente, se visualizó la cantidad de datos faltantes en cada una de las variables mediante las gráficas mostradas en las figuras 4, 5 y 6, en las que se demostró que la variable con más datos faltantes era material particulado (PM2.5).

Figura 4. Gráfico de barras y bloques relacionado con los datos faltantes

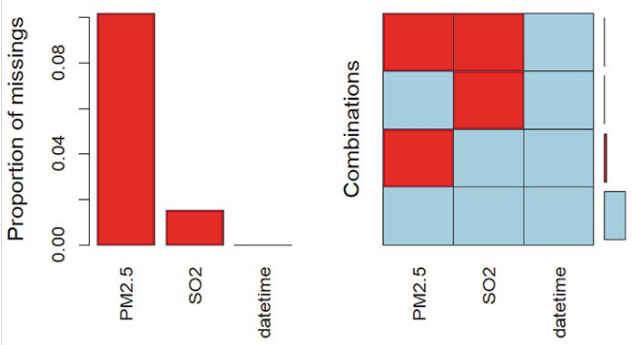


Figura 5. Gráfico date_time vs PM2.5 sin imputación de datos

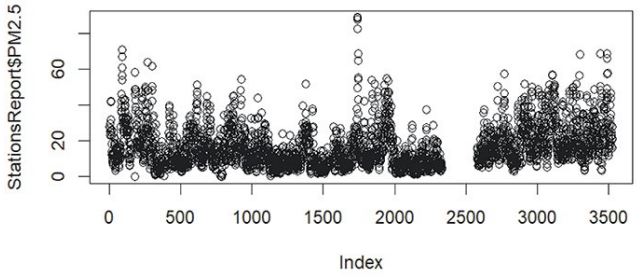
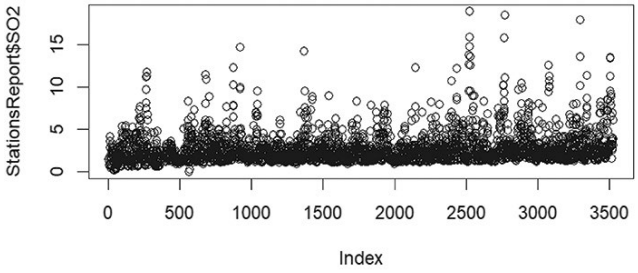
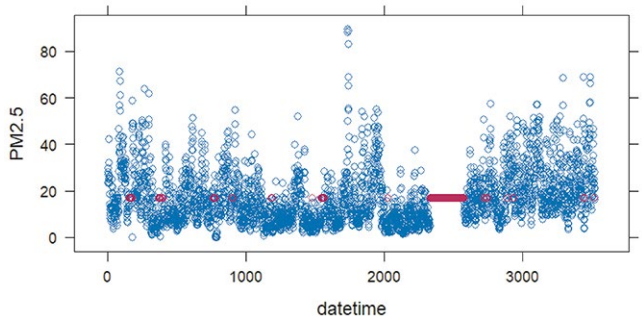


Figura 6. Gráfico date_time vs SO2 sin imputación de datos



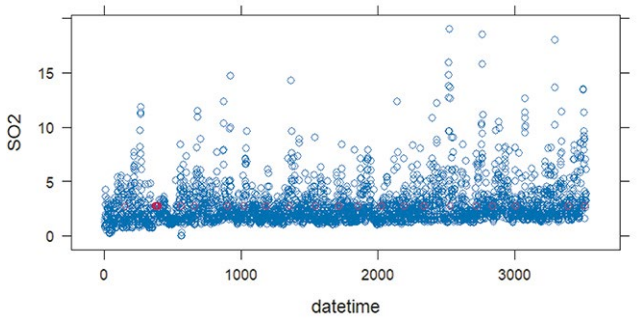
En un principio, para completar la data se usó el método de imputación de datos "mean", que consiste en remplazar los datos que faltan con el valor de la media; sin embargo, en este caso no se obtuvieron resultados óptimos, pues al graficar la información después de ejecutar el modelo "mean", no se observó concordancia entre los datos reales y los obtenidos por imputación como se muestra en las figuras 7 y 8. Cabe resaltar que los círculos azules representan los datos reales y los rojos lo obtenidos por imputación.

Figura 7. Gráfico date_time vs PM2.5 con imputación de datos mediante el modelo "mean"



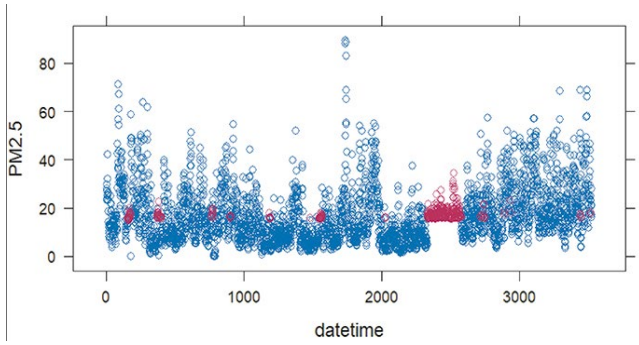
ISSN 2322-8415 / Año 2024 / No. 23 - Vol. 13 / Páginas 65-74 Universidad Libre - Facultad de Ingeniería

Figura 8. Gráfico date_time vs SO2 con imputación de datos mediante el modelo.

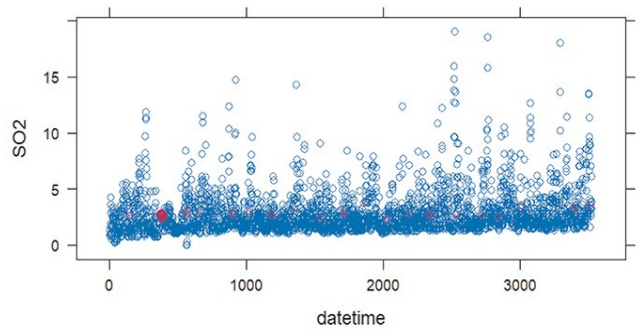


Posteriormente, se usó el modelo de imputación de datos "norm.predict" que presentó resultados óptimos al ejecutar el gráfico, pues se evidenció mayor concordancia entre los datos reales y los obtenidos mediante la imputación, como se muestra en las figuras 9 y 10.

Figura 9. Gráfico date_time vs PM2.5 con imputación de datos mediante el modelo "norm.predict"



Figuras 10. Gráfico date_time vs SO2 con imputación de datos mediante el modelo "norm.predict"



Después de realizar el proceso de imputación para obtener una data completa y poder aplicar el modelo "Ridge regression" se obtuvieron los siguientes resultados.

Para $PM_{2.5}$

La gráfica del modelo "Ridge regression" para $PM_{2.5}$, presenta la información relacionada con el "logaritmo de Lambda" en el eje X y los "coeficientes" en el eje Y; producto de esta regresión se obtiene una función decreciente, en la cual la Lambda más pequeña tiene un valor de 16.16719 y la más grande un valor de 161671.94416 (figura 11); del mismo modo, con base en la gráfica, es posible analizar un modelo en particular, por ejemplo, al ejecutar el modelo 10 de la variable $PM_{2.5}$ se obtienen los valores relacionados con el modelo, el valor del logaritmo y los coeficientes, en este caso 69983.93, 11.15602 y 0.2072377 respectivamente, además de la gráfica, como se muestra en la figura 12.

Figura 11. Gráfica de modelo "Ridge Regression" PM2.5

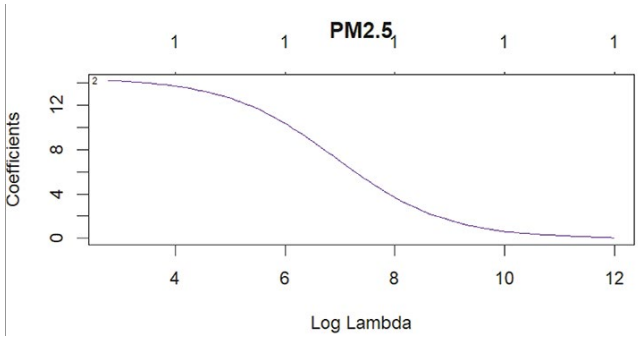
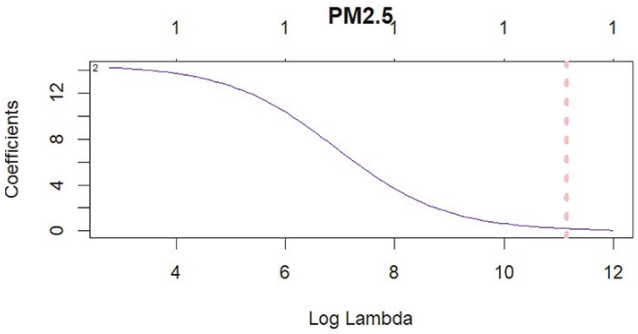
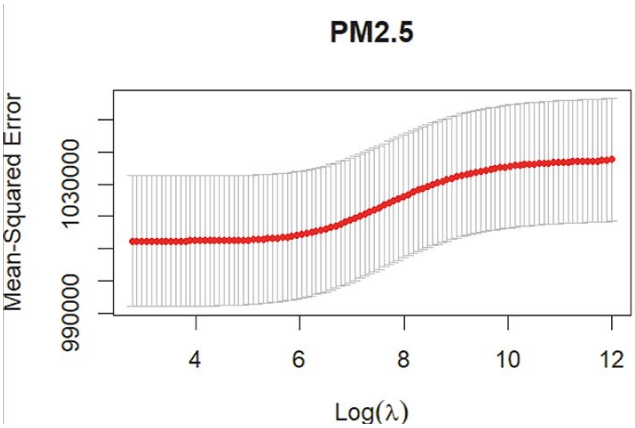


Figura 12. Gráfica de modelo 20 "Ridge Regression" para SO2



A continuación, se halló el valor de la mejor lambda que para la variable PM2.5 fue 16.14674 con un logaritmo de 2.781718, como se muestra en la gráfica representada en la figura 13.

Fig. 13: Gráfica de mejor lambda "Ridge Regression" para PM2.5



Para SO_2

Finalmente, los valores del coeficiente de determinación (r2) y error cuadrático medio (RMSE) fueron 1755.974, 0.03189163, para PM2.5 y 3324.062, 0.03189163 para SO2, respectivamente. Lo cual demuestra que el método "Ridge Regression" no funcionó de una manera óptima para el análisis y predicción de los datos correspondientes a las partículas (PM2.5) y (SO2) en la localidad de Kennedy, puesto que RMSE es una métrica comúnmente utilizada para evaluar la precisión de un modelo de regresión, debido a que compara las predicciones generadas por el modelo con los valores reales del conjunto de datos; generalmente, este valor debe ser cercano a cero y en este caso fue de 1755.974.

Con base en lo anterior, se estima que, aunque el algoritmo de regresión ridge se emplea generalmente para diseñar modelos más simples y con mayor precisión, debido a que busca solucionar los problemas de variabilidad en caso de que algunos predictores se encuentren correlacionados o sean menores que el número de datos, se observa que para este tipo de proyecciones el método es ineficaz, debido a a que el algoritmo cuenta con el input "alpha", que es básicamente un parámetro que a medida que incrementa su valor, se limitará a los valores de las β, con lo que disminuye la varianza total y, por ende, reduce los efectos de correlación entre variables; se minimiza así el riesgo de sufrir "overfitting" o sobreajuste; por esta razón, cuando "alpha" se establece en 0, el modelo se comporta como una regresión lineal.

En consecuencia, para encontrar el valor óptimo, se deben considerar valores entre 1 y 800, con el objetivo de minimizar la raíz cuadrada del error cuadrático medio (RMSE), sin embargo, en el presente estudio el valor obtenido fue de 1755.974, lo cual demuestra un amplio margen de error, inconsecuente con los resultados esperados para el uso de esta técnica.

7. CONCLUSIONES

- En el análisis realizado con base en los datos de la localidad de Kennedy, de acuerdo con la variable de material particulado (PM2.5) y dióxido de azufre (SO2), se identificó que PM2.5 presentaba la mayor cantidad de datos faltantes, de modo que, en un intento por completar la información, se empleó el método de imputación de datos; sin embargo, los resultados no fueron satisfactorios, pues mostraron una falta de concordancia entre los datos reales y los imputados.

- Se recurrió al modelo de imputación "norm.predict", el cual proporcionó resultados más precisos, lo cual evidenció una mayor coherencia entre los datos reales y los imputados para PM2.5. La aplicación del modelo de

ISSN 2322-8415 / Año 2024 / No. 23 - Vol. 13 / Páginas 65-74 Universidad Libre - Facultad de Ingeniería

regresión “Ridge regression” generó gráficas que representan la relación entre el logaritmo de Lambda y los coeficientes. En el caso de PM2.5, se observó una función decreciente con Lambdas que variaban desde 16.16719 hasta 161671.94416.

- Al evaluar el rendimiento del modelo a través de los coeficientes de determinación (r2) y el error cuadrático medio (RMSE), se encontraron valores desfavorables, con un RMSE de 1755.974 para PM2.5. Esto indica que el método “Ridge Regression” no fue eficaz en el análisis y predicción de los datos de PM2.5 y SO2 en la localidad de Kennedy.

- El método “Ridge Regression” puede no ser la elección más adecuada para este tipo de proyecciones en la localidad de Kennedy, por lo cual se sugiere explorar otras técnicas de modelado y ajuste de parámetros para mejorar la precisión y la eficacia en la predicción de los niveles de PM2.5 y SO2 en esta área específica.

REFERENCIAS BIBLIOGRÁFICAS

[1] *Benavides C; Gaviria G; et al. (2011).* "Contaminación por material particulado (pm2,5 y pm10) y consultas por enfermedades respiratorias en Medellín (2008-2009)". Rev. Nac. Salud Pública 2011; 29(3); 241-250.

[2] *Franco D. (2020).* "Análisis y Caracterización del Material Particulado Pm10 y Pm2.5 en la Ciudad de Manizales". Universidad Nacional de Colombia. Facultad de Ingeniería. Manizales.

[3] *González G, Robles S. (2003).* " Mortalidad Asociada con la Contaminación Atmosférica por SO2. A propósito de un caso de autopsia médico legal tras un Episodio de Polución Atmosférica". Madrid.

[4] *Farrow A, Chen Y, et al. (2022).* " La carga de la contaminación del aire en Bogotá, Colombia 2021".

[5] *Secretaría Distrital de Movilidad. (2019).* "Administración Distrital revela resultados de la Encuesta de Movilidad 2019 para Bogotá y 18 municipios vecinos". Alcaldía Mayor de Bogotá D.C., Bogotá.

[6] *Secretaría Distrital de Ambiente. (2022).* "Informe mensual de la calidad del aire de Bogotá para febrero de 2022". Alcaldía Mayor de Bogotá D.C., Bogotá.

[7] *Ramírez V. (2017).* "Ubicación de la Ciudad de Bogotá". https://bogota.gov.co/ubicacion-de-bogota-sitios-turisticos-vias-y-alrededores-de-bogota#:~:text=Ubicada%20en%20el%20Centro%20del,74%C2%B004'51''.

[8] *Secretaría Distrital de Ambiente. (2022).* "Así funcionan las estaciones de monitoreo de calidad del aire en Bogotá". Alcaldía Mayor de Bogotá D.C., Bogotá.

[9] *Taboga, Marco (2021).* "Regresión de crestas", Conferencias sobre teoría de la probabilidad y estadística matemática.

[10] *Msigwa, O. (2023).* Ridge Regression. Data Science and Machine Learnig (part 10).