

Emery Olarte

Modelo predictivo con machine learning para variables ambientales y PM2.5 en Huachac Junín – Perú (1

 My Files My Files Universidad Peruana Union

Detalles del documento

Identificador de la entrega

trn:oid:::29566:419052199

Fecha de entrega

30 dic 2024, 1:22 p.m. GMT-5

Fecha de descarga

30 dic 2024, 1:38 p.m. GMT-5

Nombre de archivo

Modelo predictivo con machine learning para variables ambientales y PM2.5 en Huachac Juníndocx

Tamaño de archivo

1.5 MB

23 Páginas

7,840 Palabras

42,982 Caracteres

16% Similitud general

El total combinado de todas las coincidencias, incluidas las fuentes superpuestas, para ca...

Filtrado desde el informe

- Bibliografía
- Texto citado
- Texto mencionado
- Coincidencias menores (menos de 8 palabras)
- Trabajos entregados

Fuentes principales

- 14%  Fuentes de Internet
- 7%  Publicaciones
- 0%  Trabajos entregados (trabajos del estudiante)

Marcas de integridad

N.º de alertas de integridad para revisión

No se han detectado manipulaciones de texto sospechosas.

Los algoritmos de nuestro sistema analizan un documento en profundidad para buscar inconsistencias que permitirían distinguirlo de una entrega normal. Si advertimos algo extraño, lo marcamos como una alerta para que pueda revisarlo.

Una marca de alerta no es necesariamente un indicador de problemas. Sin embargo, recomendamos que preste atención y la revise.

Fuentes principales

- 14% Fuentes de Internet
- 7% Publicaciones
- 0% Trabajos entregados (trabajos del estudiante)

Fuentes principales

Las fuentes con el mayor número de coincidencias dentro de la entrega. Las fuentes superpuestas no se mostrarán.

1	Internet	
zadoco.site		1%
2	Publicación	
Qian Han, Xinge Nan, Han Wang, Yanjun Hu, Zhiyi Bao, Hai Yan. "Optimizing the S...		1%
3	Publicación	
Ines Aberlí Nieves-Vilchez, Hiram Joshua Castillo-Infante, Josué Isac Carillo-Espino...		1%
4	Internet	
docplayer.es		1%
5	Internet	
sro.sussex.ac.uk		1%
6	Publicación	
Tinoco Ramos José, Yupanqui Arellano Jhoset, Juan J. Soria, Nemias Saboya. "Chap...		0%
7	Internet	
www.preprints.org		0%
8	Internet	
institutodeanalistas.com		0%
9	Internet	
www.researchgate.net		0%
10	Internet	
repositorio.unfv.edu.pe		0%
11	Internet	
www.aicad.es		0%

12	Internet	cubacoopera.uccm.sld.cu	0%
13	Internet	hdl.handle.net	0%
14	Publicación	Cristina Martínez González, María Jesús Cruz Carmona. "Efectos de la contaminaci...	0%
15	Publicación	Vicente Salvador Gallego Blasco. "Análisis de la incidencia de factores causales en...	0%
16	Publicación	Fushuang Jiang, Shanshan Wen, Miaoni Gao, Aiping Zhu. "Assessment of NEX-GD...	0%
17	Publicación	Mohammed Achite, Okan Mert Katipoğlu, Muhammad Jehanzaib, Nehal Elshabou...	0%
18	Internet	alicia.concytec.gob.pe	0%
19	Internet	www.coursehero.com	0%
20	Internet	www.hup.es	0%
21	Internet	ecuciencia.utc.edu.ec	0%
22	Internet	www.webpages.uidaho.edu	0%
23	Internet	eprints.soton.ac.uk	0%
24	Internet	riunet.upv.es	0%
25	Internet	ciencia.lasalle.edu.co	0%

26	Internet	mx.boell.org	0%
27	Internet	www.noticiadesalud.com	0%
28	Publicación	Abul Amir Khan, Kalpana Garsa, Prakhar Jindal, Panuganti C. S. Devara, Shubhans...	0%
29	Publicación	LÍNEA VERDE AC S.A.C.. "Plan de Recuperación de Áreas Degradadas por Residuos...	0%
30	Internet	es.unionpedia.org	0%
31	Internet	nardus.mpn.gov.rs	0%
32	Internet	worldwidescience.org	0%
33	Internet	creativecommons.org	0%
34	Internet	repositorio.upla.edu.pe	0%
35	Internet	www.dnic.unal.edu.co	0%
36	Internet	ia600908.us.archive.org	0%
37	Internet	waqi.info	0%
38	Internet	www.elektrosol.net	0%
39	Internet	www.flacsoandes.edu.ec	0%

40	Internet	www.mujeresdeltercermilenio.hpg.ig.com.br	0%
41	Internet	www.unisdr.org	0%
42	Internet	bibdigital.epn.edu.ec	0%
43	Internet	es.scribd.com	0%
44	Internet	helvia.uco.es	0%
45	Internet	repository.unimilitar.edu.co	0%
46	Internet	www.buzzfeed.com	0%
47	Internet	www.ecofield.com.ar	0%
48	Internet	www.ofita.com	0%
49	Internet	www.revistaeyaacademy.mx	0%
50	Internet	dokumen.pub	0%
51	Internet	iris.unipa.it	0%
52	Internet	livrosdeamor.com.br	0%
53	Internet	perso.univ-lyon2.fr	0%

54	Internet	repositorio.lamolina.edu.pe	0%
55	Internet	repositorio.unesp.br	0%
56	Internet	repository.unipiloto.edu.co	0%
57	Internet	rraae.cedia.edu.ec	0%
58	Internet	www.ingentaconnect.com	0%
59	Internet	www.news-medical.net	0%
60	Internet	www.slideshare.net	0%



Type of the Paper (Article, Review, Communication, etc.)

Modelo predictivo con machine learning para variables ambientales y PM2.5 en Huachac Junín – Perú

Emery Olarte¹, Jhonatan Gutierrez¹, Gwayne Roque¹, Juan J. Soria², Hugo Fernandez¹, Jackson Edgardo Perez Carpio¹, Orlando Poma Porras^{4*}

¹ Escuela profesional de Ingeniería Ambiental, Universidad Peruana Unión, Lima Perú; elsao-olarte@upeu.edu.pe

² Escuela profesional de Ingeniería Ambiental, Universidad Peruana Unión, Lima Perú; jhona-tangg@upeu.edu.pe

³ Escuela profesional de Ingeniería Ambiental, Universidad Peruana Unión, Lima Perú; gwaynero-que@upeu.edu.pe

⁴ Escuela de Posgrado, Universidad Peruana Unión, Lima Perú; jesussoria@upeu.edu.pe

⁵ Escuela profesional de Ingeniería Ambiental, Universidad Peruana Unión, Lima Perú; hugof@upeu.edu.pe

⁶ Escuela profesional de Ingeniería Ambiental, Universidad Peruana Unión, Lima Perú; jacksonperez@upeu.edu.pe

⁷ Escuela profesional de Ingeniería Ambiental, Universidad Peruana Unión, Lima Perú; opoma@upeu.edu.pe

* Correspondence: opoma@upeu.edu.pe; Tel.: +51969332640

Abstract: (1) **Antecedentes:** La contaminación por PM2.5 está en aumento provocando problemas de salud, con el objetivo de modelar el comportamiento del PM2.5AQI (índice de calidad de aire) mediante machine learning se aplicó modelos predictivos de regresión Lineal, Lasso, Ridge y Elastic Net considerando 16, 543 registros de la zona de Huachac-Junin en Perú con regresores de humedad en g/m³ y la temperatura en °C. El punto focal del estudio son el material particulado y variables ambientales, donde se identificó un nuevo modelo de Machine Learning ML capaz de predecir el PM2.5 de la zona de Huachac con considerable precisión, basado en variables regresoras estrechamente relacionadas con el material particulado. (2) **Métodos:** Se aplicó el Análisis Exploratorio de Datos (EDA), el cual permitió una visión clara de los datos, realizando un análisis descriptivo, un ajuste de los tipos de variables, la detección y tratamiento de datos faltantes, la identificación de datos atípicos y la correlación de variables, permitiendo tener una data sólida en la predicción de los modelos predictivos de machine Learning aplicados; (3) **Resultados:** El índice de la calidad del aire tiene tendencia a ser más alta en invierno y primavera con medias de 52.6, 36.9 PM2.5AQI respectivamente y con valores más bajos en verano además presenta valor máximo en septiembre y presenta valores mínimos en febrero. El uso de los modelos de regresión arrojó métricas precisas para elegir el mejor modelo para la predicción del PM2.5AQI. La comparación con otras investigaciones resalta la solidez de los modelos de ML elegidos, subrayando el potencial del ML en la ingeniería ambiental. (4) **Conclusiones:** El modelo predictivo encontrado fue $\alpha=0.111111$ y un valor de $\Lambda=0.150025$ representado por $PM2.5AQI = 83.0846522 - 10.3022000(Humedad) - 0.1268124(Temperatura)$ con un R² ajustado de 0.1483206 y un RMSE de 25.36203.

Keywords: Lasso; Ridge; Elastic Net; PM2.5AQI; Temperatura, Humedad, Aprendizaje automático

Citation: To be added by editorial staff during production.

Academic Editor: Firstname Last-name

Received: date

Revised: date

Accepted: date

Published: date



Copyright: © 2023 by the authors.

Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introducción

Uno de los problemas de carácter global es la contaminación atmosférica, siendo uno de los principales riesgos para la salud pública, situándose como la causa más importante de decesos a nivel mundial [1]. Este es un problema global que afecta tanto a los países desarrollados, especialmente en sus grandes ciudades, como a los países en vías de desarrollo, así la contaminación atmosférica puede tener diversos efectos adversos en el entorno natural, destacándose en primer lugar, la lluvia ácida, smog, pérdida de biodiversidad, cambio climático y agotamiento de la capa de ozono [2].

La contaminación atmosférica es un problema de índole mundial, donde los principales contaminantes que afectan la atmósfera están los gases (CO, NO₂, SO₂, Pb.) y el material particulado PM₁₀ y PM_{2.5} [3]. Por lo general estos son liberados por fuentes antrópicas y naturales, en donde los estudios de morbilidad respiratorias están implicadas para ciertos tipos de contaminantes del aire conforme a la Organización Mundial de la Salud (OMS), siendo el principal la materia particulada, considerándose el PM_{2.5} [4]. Las partículas PM₁₀ y PM_{2.5} son las que generan un efecto negativo en la salud, así como en el medio ambiente y en los ecosistemas, estas partículas peligrosas suspendidas en el aire están compuestas por partículas sólidas y líquidas [5].

Según la OMS [6] muestran que casi toda la totalidad de la población mundial respira aire que excede los límites recomendados por la Organización Mundial de la Salud, que contienen grandes cantidades de contaminantes, además según estimaciones del 2021 la contaminación del aire en ciudades y zonas rurales de todo el mundo provoca 2 millones de muertes prematuras al año.

La contaminación del aire urbano en las últimas décadas y su impacto en la salud humana es solo un ejemplo de la degradación ecológica en curso y sus vínculos con un determinado tipo de sociedad [7]. A nivel urbano, los efectos de la contaminación del aire especialmente en los productos de combustibles fósiles tanto en la producción de energía como en la quema de automóviles y otra materia orgánica muestran como los gases óxidos de azufre (SO_x) y los óxidos de nitrógeno (NO_x) están directamente relacionados con las vías respiratorias, enfermedad y hospitalización; en forma similar sucede con el dióxido de azufre (SO₂) [8].

Los contaminantes atmosféricos peligrosos son sustancias que se sabe que causan efectos graves en la salud, y también se han estudiado los efectos de estos contaminantes junto con el cambio climático, hay muchas fuentes de estos contaminantes, pero las emisiones industriales atraen el interés porque se consideran fuentes importantes de partículas PM₁₀ y PM_{2.5}, gases y metales pesados. Los residentes que viven alrededor de los complejos industriales están expuestos a niveles más altos de contaminación, que los que viven lejos de estas áreas [9].

En la protección ecológica global y el desarrollo urbano, la contaminación del aire se ha convertido en un problema crucial que afecta la salud humana y el medio ambiente, junto con el cambio climático, lo que ha surgido una atención cada vez mayor de académicos y administradores gubernamentales [10]. En América Latina, por lo menos 100 millones de personas están expuestas a niveles de contaminación del aire por encima de los recomendados por la Organización Mundial de la Salud [11]. Los grupos más vulnerables a los efectos dañinos de una mala calidad de aire incluyen a los niños, adultos mayores, personas con problemas de salud y población de bajos estratos socioeconómicos.

Según Cordova [12] realizó un estudio reciente señalando que Lima Metropolitana (LIM) es propensa a altas concentraciones de PM₁₀ y PM_{2.5} debido a la industrialización acelerada y el crecimiento económico además que alberga el 29% de los 34 105 243 de la población total peruana.

Las estaciones meteorológicas y otros dispositivos de medición ambiental generan una gran cantidad de datos difíciles de procesar manualmente, por ello los algoritmos de aprendizaje automático pueden procesar de manera eficiente grandes cantidades de datos e identificar patrones y relaciones entre el PM2.5 y las variables meteorológicas [14]. Los algoritmos de aprendizaje automático pueden proporcionar predicciones precisas y confiables al instante. Adaptabilidad al cambio: las variables en el ambiente del aire pueden cambiar rápidamente en función de factores externos como cambios en las condiciones climáticas o la actividad humana [15].

En el Perú, por el aumento de la industrialización y por el amplio uso de hidrocarburos se ha visualizado un aumento de concentración de partículas PM2.5 en el aire [13], lo cual genera una problemática en términos de salud pública y medio ambiente, por lo que la investigación tiene como objetivo optimizar modelos predictivos de machine Learning con data de la estación astronómica Huachac, Chupaca, Junín, para analizar y predecir la concentración de PM2.5.

2. Materiales y Métodos

2.1 Área de Estudio

El estudio se realizó en el distrito de Huachac, provincia de Chupaca, departamento de Junín – Perú, ubicado a 3275 msnm y a 290 km de la capital del país (Lima), a 22 km al noroeste de Huancayo, limitando por el Norte con el Distrito de Manzanares, por el Sur con la ciudad de Chupaca, por el Este con el Distrito de Sicaya y por el Oeste con el Distrito de Chambara en la provincia de Concepción como se muestra en la Figura 1. Los datos fueron obtenidos de la estación de monitoreo HUAYAO, que está ubicado en distrito de Huachac, provincia de Chupaca, departamento de Junín – Perú.

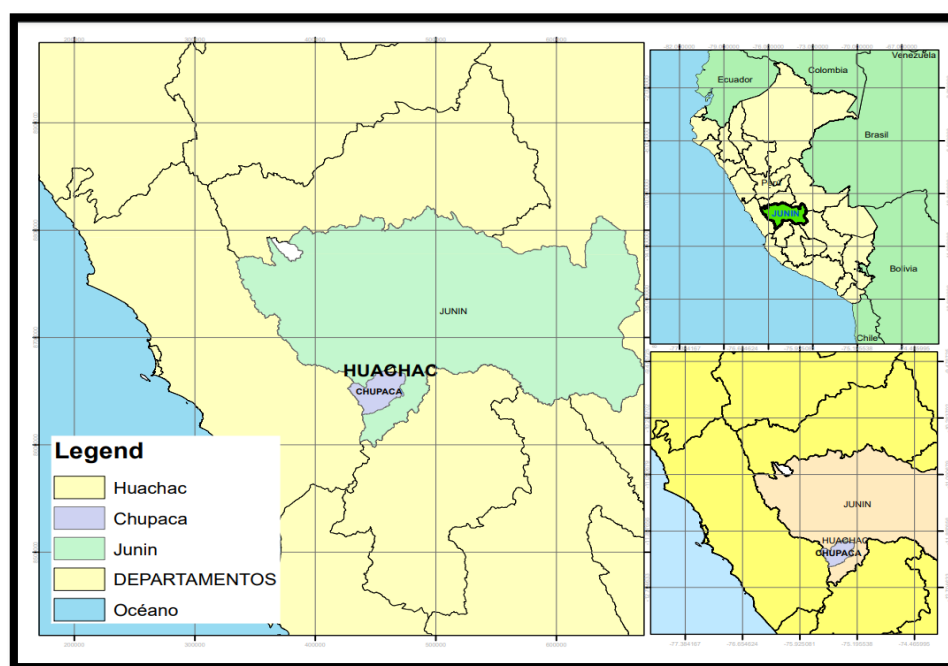


Figure 1. Ubicación de la estación meteorológica Huayao-Junin

2.2. Método de la Investigación

El método de investigación se presenta en la figura 2, en la que se recogió la data de la estación Huayao-Junin-Perú, luego se hizo un análisis exploratorio de los datos (EDA) recogidos, imputando outliers y limpiando datos justificando los prerequisites de normalidad, linealidad y homocedasticidad[16].

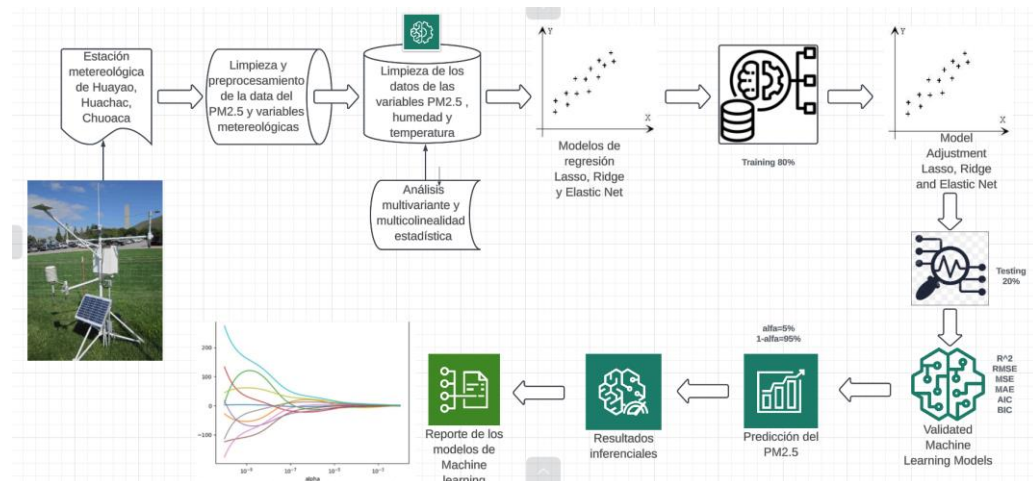


Figure 2. Método de la Investigación

2.2 Modelos de Machine Learning

El aprendizaje automático, es una rama de la inteligencia artificial que se enfoca en desarrollar algoritmos y modelos que permiten a las computadoras aprender y tomar decisiones basadas en datos, sin ser programadas explícitamente cuyo objetivo es permitir a las máquinas aprender de manera autónoma y precisa [17]. El Machine Learning (ML) y la estadística están relacionados, ya que el ML se basa en muchos principios estadísticos[18]

La regresión lineal es una generalización del modelo de regresión poblacional de k variables (FRP) con la variable dependiente Y y $k-1$ variables explicativas $X_2, X_3, X_4, X_5, \dots, X_k$ [19] y se escribe como se muestra en la ecuación 1.

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \dots + \beta_k X_{ki} + u_i \quad ; \quad i = 1, \dots, n \quad (1)$$

Donde β_1 es el intercepto, β_2 a β_k son los coeficientes parciales de pendientes, u es el término de perturbación estocástica e i es la i -ésima observación, con n como tamaño de la población. La ecuación (1) es una expresión abreviada para el conjunto de n ecuaciones simultáneas [20] mostradas en la ecuación 2 y 3.

$$\left. \begin{aligned} Y_1 &= \beta_1 + \beta_2 X_{21} + \beta_3 X_{31} + \beta_4 X_{41} + \dots + \beta_k X_{k1} + u_1 \\ Y_2 &= \beta_1 + \beta_2 X_{22} + \beta_3 X_{32} + \beta_4 X_{42} + \dots + \beta_k X_{k2} + u_2 \\ Y_3 &= \beta_1 + \beta_2 X_{23} + \beta_3 X_{33} + \beta_4 X_{43} + \dots + \beta_k X_{k3} + u_3 \\ &\dots \\ Y_n &= \beta_1 + \beta_2 X_{2n} + \beta_3 X_{3n} + \beta_4 X_{4n} + \dots + \beta_k X_{kn} + u_n \end{aligned} \right\} \quad (2)$$

El sistema de ecuaciones 2 se escribe en forma matricial como se muestra en la ecuación 3 y 4.

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{21} & x_{31} & x_{41} & \cdots & x_{k1} \\ 1 & x_{22} & x_{32} & x_{42} & \cdots & x_{k2} \\ 1 & x_{23} & x_{33} & x_{43} & \cdots & x_{k3} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{2n} & x_{3n} & x_{4n} & \cdots & x_{kn} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ \vdots \\ u_n \end{bmatrix} \quad (3)$$

$$Y_{nx1} = X_{nxk} \cdot \beta_{kx1} + u_{nx1} \quad (4)$$

Donde Y es el vector columna de nx1 de observaciones sobre la variable dependiente Y, la variable X es la matriz nxk, con n observaciones sobre k-1 variables X₂ a X_k y la primera columna de números 1 representa el término del intercepto, el valor de β es el vector columna kx1 de los parámetros desconocidos $\beta_1, \beta_2, \dots, \beta_k$ y el valor de u es el vector columna nx1 de n perturbaciones u_i

La regresión LASSO (Least Absolute Shrinkage and Selection Operator) es un método de regresión utilizando para la selección de variables y la regularización en métodos de regresión lineal. Su objetivo principal es reducir la complejidad del modelo al penalizar los coeficientes de las variables menos importantes, llevándolos a cero.[21]

Según Aliaj [22] la regresión LASSO (Least Absolute Shrinkage and Selection Operator) es una técnica de regularización utilizada en el aprendizaje automático y la estadística para realizar la selección de variables y la reducción de la complejidad de modelo. Por otra parte, existe también el concepto de que la regresión LASSO es una extensión de la regresión lineal que incorpora una penalización para reducir la complejidad del modelo y evitar el sobreajuste [23]. Se utiliza comúnmente en problemas de regresión cuando hay un gran número de características (variables) y se busca realizar la selección de variables al mismo tiempo que se ajusta un modelo predictivo [24]. La función de pérdida de la regresión LASSO se compone de la función de pérdida de la regresión lineal (como el error cuadrático medio) más el valor absoluto de los coeficientes de las características multiplicados por un hiperparámetro lambda λ [25].

La regresión LASSO es un método restringido, dado que selecciona las variables influyentes en una variable llamada respuesta, resultando con unos coeficientes en cero. “Mediante este procedimiento se busca un conjunto de instrumentos fuerte, que proporcione la mayor cantidad de información sobre las variables endógenas en la primera etapa de la estimación [22].

El modelo de regresión LASSO está representado matemáticamente por ecuación 5, cuyo objetivo es minimizar los parámetros $\alpha; \beta$ en los errores tomando la penalización de los coeficientes del modelo [26].

$$\text{Minimize}_{\alpha, \beta} \left\{ \frac{1}{2n} \sum_{i=1}^N (y_i - \alpha - \sum_{j=1}^p x_{ij} \beta_j)^2 \right\} \quad (5)$$

Por lo tanto, el Modelo LASSO considerando las variables anteriormente mencionadas, queda representado por la ecuación 6 y 7.

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \end{bmatrix} \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} \quad (6)$$

$$Y_i = X_{k \times 3} \cdot \beta_{3 \times 1} + u_i ; i = \overline{1,3} \quad (7)$$

El modelo de regresión lineal se muestra en la ecuación 8 y 9.

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + u_i \quad (8)$$

$$u_i = Y_i - \alpha - \beta_1 X_{i1} - \beta_2 X_{i2} \quad (9)$$

Donde u_i es el error que se quiere minimizar. La técnica de mínimos cuadrados realiza la minimización de los errores y su representación es:

$$D = \sum_{i=1}^n u_i^2$$

Al realizar la regresión Lasso, agregamos un factor de penalización a los mínimos cuadrados, que reduce la función de pérdida S a un valor mínimo, representada por la ecuación 10 y 11.

$$S = \underset{\alpha, \beta_1, \beta_2}{Min} [u_i^2 + \lambda(|\beta_1| + |\beta_2|)] \quad (10)$$

$$S = \underset{\alpha, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5}{Min} [(Y_i - \alpha - \beta_1 X_{i1} - \beta_2 X_{i2})^2 + \lambda(|\beta_1| + |\beta_2|)] \quad (11)$$

Las ecuaciones normales por la teoría de mínimos cuadrados para la regresión múltiple Lasso se muestran en la ecuación 12.

$$\begin{cases} \alpha n + \beta_1 \sum_{i=1}^n X_{i1} + \beta_2 \sum_{i=1}^n X_{i2} = \sum_{i=1}^n y_i \\ \alpha \sum_{i=1}^n X_{i1} + \beta_1 \sum_{i=1}^n X_{i1}^2 + \beta_2 \sum_{i=1}^n X_{i1} X_{i2} = \sum_{i=1}^n X_{i1} y_i - \lambda n \\ \alpha \sum_{i=1}^n X_{i2} + \beta_1 \sum_{i=1}^n X_{i1} X_{i2} + \beta_2 \sum_{i=1}^n X_{i2}^2 = \sum_{i=1}^n X_{i2} y_i - \lambda n \end{cases} \quad (12)$$

El modelo de regresión Ridge a diferencia de la regresión LASSO reduce los coeficientes de los predictores correlacionados entre sí, lo que les permite tomar prestada la fuerza de los demás. Asimismo, desde la perspectiva bayesiano, la penalización del modelo RIDGE es apropiado en el caso de haber varios predictores y todos tienen coeficientes distintos de cero, es decir, son extraídos de una distribución gaussiana [27]. Además, es de prioridad considerar en el análisis las propiedades del error cuadrático medio de la regresión de Ridge como la varianza y sesgo del estimador, el teorema sobre la función del cuadrado medio y los comentarios hechos sobre la función del error cuadrático medio [28]. La representación matemática del modelo de regresión Ridge se representa en la ecuación 13.

$$\beta^{ridge} = \underset{\beta \in \mathbb{R}}{argmin} \|Y - X\beta\|_2^2 + \lambda \|\beta\|_2^2 \quad (13)$$

Al realizar la regresión Ridge, agregamos un factor de penalización a los mínimos cuadrados, que reduce la función de pérdida S a un valor mínimo, representada en la ecuación 14 y 15

$$S = \underset{\alpha, \beta_1, \beta_2}{\text{Min}} \left[\frac{1}{2n} \sum_{i=1}^N u_i^2 + \lambda(\beta_1^2 + \beta_2^2) \right] \quad (14)$$

$$S = \underset{\alpha, \beta_1, \beta_2}{\text{Min}} \left[\frac{1}{2n} (Y_i - \alpha - \beta_1 X_{i1} - \beta_2 X_{i2})^2 + \lambda(\beta_1^2 + \beta_2^2) \right] \quad (15)$$

Las ecuaciones normales de la teoría de mínimos cuadrados para la regresión múltiple Ridge se representa en la 16.

$$\begin{cases} \alpha n + \beta_1 \sum_{i=1}^n X_{i1} + \beta_2 \sum_{i=1}^n X_{i2} = \sum_{i=1}^n y_i \\ \alpha \sum_{i=1}^n X_{i1} + \beta_1 \sum_{i=1}^n X_{i1}^2 + \beta_2 \sum_{i=1}^n X_{i1} X_{i2} = \sum_{i=1}^n X_{i1} y_i - 2\lambda n \beta_1 \\ \alpha \sum_{i=1}^n X_{i2} + \beta_1 \sum_{i=1}^n X_{i2} X_{i1} + \beta_2 \sum_{i=1}^n X_{i2}^2 = \sum_{i=1}^n X_{i2} y_i - 2\lambda n \beta_2 \end{cases} \quad (16)$$

La penalización de la regresión Elastic Net es una adaptación de los mínimos cuadrados y permite abordar el problema de estimación produciendo un estimador β sesgado pero con varianzas pequeñas [21]. La representación matemática del modelo de regresión Elastic Net se representa en la ecuación 17 se define:

$$\beta^{EN} = \underset{\beta \in \mathbb{R}}{\text{argmin}} \frac{1}{2n} \|Y - X\beta\|_2^2 + \lambda[(1 - \alpha)\|\beta\|_2^2 + \alpha\|\beta\|_1] \quad (17)$$

Al realizar la regresión Elastic Net, agregamos un factor de penalización a los mínimos cuadrados, que reduce la función de pérdida S a un valor mínimo, representada en la ecuación 18 y 19.

$$S = \underset{\alpha, \beta_1, \beta_2}{\text{Min}} \left[\frac{1}{2n} \sum_{i=1}^N u_i^2 + \lambda(\beta_1^2 + \beta_2^2) \right] \quad (18)$$

$$S = \underset{\alpha, \beta_1, \beta_2}{\text{Min}} \left\{ \frac{1}{2n} (Y_i - \alpha - \beta_1 X_{i1} - \beta_2 X_{i2})^2 + \lambda[(1 - \alpha)(\beta_1^2 + \beta_2^2)] + \alpha[|\beta_1| + |\beta_2|] \right\} \quad (19)$$

Luego las ecuaciones normales de la teoría de mínimos cuadrados para la regresión Elastic net se representa en la ecuación 20.

$$\begin{cases} \alpha n + \beta_1 \sum_{i=1}^n X_{i1} + \beta_2 \sum_{i=1}^n X_{i2} = n\lambda[(\beta_1^2 + \beta_2^2)] - [|\beta_1| + |\beta_2|] + \sum_{i=1}^n y_i \\ \alpha \sum_{i=1}^n X_{i1} + \beta_1 \sum_{i=1}^n X_{i1}^2 + \beta_2 \sum_{i=1}^n X_{i1} X_{i2} = \sum_{i=1}^n X_{i1} y_i + \lambda n[2\beta_1(\alpha - 1) - \alpha] \\ \alpha \sum_{i=1}^n X_{i2} + \beta_1 \sum_{i=1}^n X_{i2} X_{i1} + \beta_2 \sum_{i=1}^n X_{i2}^2 = \sum_{i=1}^n X_{i2} y_i + \lambda n[2\beta_2(\alpha - 1) - \alpha] \end{cases} \quad (20)$$

2.3. Métricas de validación de los modelos predictivos

Las métricas usadas en la investigación dan la validez de los modelos predictivos [22] que son el MAE (Error Absoluto Medio), que midió la magnitud promedio de los errores en las predicciones, sin considerar su dirección. El MAE es el promedio en la muestra, de las diferencias absolutas entre la predicción y la observación real, donde todas las diferencias individuales tienen el mismo peso y se muestra en la ecuación 21.

$$\text{Mean Absolute Error: } MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (21)$$

Así mismo el RMSE (raíz del error cuadrático Medio) es la raíz cuadrada del promedio de las diferencias al cuadrado entre la predicción y la observación real (SSE/n) y está definida por la ecuación 22.

$$\text{Root Mean Square Error: } RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (22)$$

El coeficiente de determinación R^2 es otra métrica que usó los modelos predictivos, el cual midió la varianza en los datos explicado por el modelo y es determinado como uno menos la relación entre la suma de cuadrados de los errores (SSE) y el total (SST) el cual se muestra en la ecuación 23.

$$\text{Coefficient of determination: } R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (23)$$

2.4. Normalización de los datos

En una investigación utilizó 108317 registros de datos de variables escalares, donde Shanker et al [29] sugiere que es necesario normalizar las características de las variables, ya que producen mejores resultados en general. Además los requisitos de los algoritmos usados requieren normalización de los datos donde Atlas et al [30] menciona que se debe usar en los casos de las variables regresoras y predictoras, por el cual se procedió a normalizar los datos mediante la técnica de normalización de min-max para garantizar la homogeneidad de las variables concentradas en un intervalo continuo de [0, 1] menciona M-Dawam [31] como se muestra en la ecuación 24.

$$\hat{X}[:, i] = \frac{X[:, i] - \min(X[:, i])}{\max(X[:, i]) - \min(X[:, i])} \quad (24)$$

2.5. Análisis exploratorio de datos

El Análisis Exploratorio de Datos (EDA), permite una visión clara de los datos, realizando un análisis descriptivo, un ajuste de los tipos de variables, la detección y tratamiento de datos faltantes, la identificación de datos atípicos y la correlación de variables [32]. En el análisis descriptivo, se sumerge en los datos para entender su estructura y naturaleza. Esto incluye la generación de estadísticas resumidas, como media, mediana, desviación estándar, y la creación de gráficos que revelan la distribución de los datos. Luego se ajusta los tipos de variables en que los datos pueden presentarse en diferentes formatos y tipos (numéricos, categóricos, fechas, etc.) [33]. En esta fase, se ajusta y transforma las variables según sea necesario para asegurarnos de que se puedan utilizar eficazmente en el análisis. Además de detecta y se hace un tratamiento de datos faltantes, donde los datos incompletos pueden ser un desafío. Durante el EDA, se identifica las observaciones con datos faltantes y se considera opciones como la imputación o la eliminación de filas o columnas para garantizar la integridad de los resultados, así mismo se identificó los datos atípicos,

conocidos como valores extremos. El EDA nos ayudó a identificar y entender estos valores para determinar si deben ser tratados o si son legítimos y significativos. Finalmente se hizo la correlación de variables para explorar las relaciones entre las variables, utilizando técnicas estadísticas y visualizaciones para descubrir conexiones y patrones que pueden revelar información valiosa sobre el problema del material particulado analizado en la investigación [34]

3. 3. Resultados

3.1. Análisis Descriptivo de las variables

Se analizó 16543 registros para las variables en estudio como se muestra en la Tabla 1, donde el PM2.5AQI, presenta una densidad de los datos con tendencia a la izquierda y con cola a la derecha por los valores atípicos que tiene algunos días del año, pero por lo general presenta una media de 33.1, con una desviación estándar de 2.5 y un valor máximo de 208 lo cual representa un valor atípico y conlleva un riesgo para la salud de la población. En la humedad absoluta mayormente se presentó una distribución normal dado a que tuvo una media de 4.72 g/m³ y con una mediana de 4.8 y una desviación estándar de 0.987, indicándonos una humedad relativamente estable en el tiempo, así mismo, la temperatura presenta un promedio de 10.8 °C con una mediana de 9 y una desviación estándar de 0.23 teniendo una temperatura que oscila entre -6 y 28.

Table 1. Descripción de las variables de estudio.

Title 1	Media	DE
PM2.5AQI	33.1	2.75
Humedad	4.72	0.987 ¹
Temperatura	10.8	0.23

¹ Tables may have a footer

En la figura 3, se observa el PM2.5AQI encontrándose que en Invierno se obtuvo la mayor concentración con un promedio de 52.6 g/m³, en segundo lugar en primavera con un promedio de 36.9 g/m³, en tercer lugar otoño con un promedio de 27.4 g/m³ y finalmente en Verano con un promedio de 15.6 g/m³.

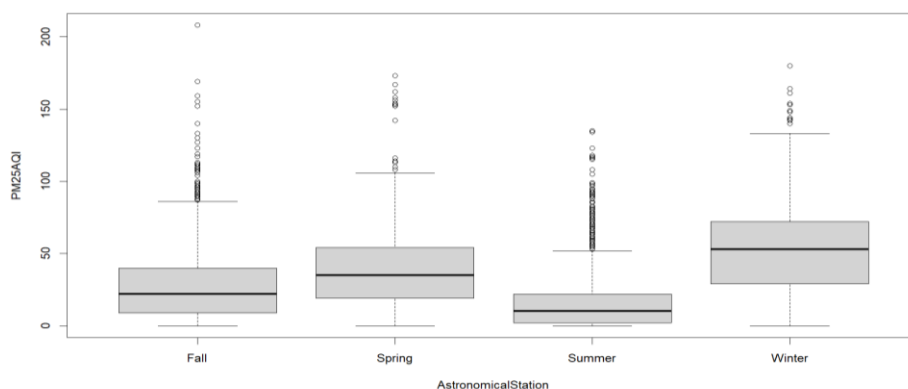


Figure 3. PM2.5AQI por estaciones del año

Los datos muestran que el índice de la calidad del aire (PM2.5AQI) tiene la tendencia a ser más alta en invierno y primavera con medias de 52.6 y 36.9, además de los valores más bajos en verano.

En la figura 4, observamos que el PM2.5AQI por meses en el año analizado, en el mes de setiembre hubo un índice de 64.9, en Segundo lugar en Agosto con 51.7, en tercer lugar julio con 46.0, en cuarto lugar en noviembre con 43.10 y los demás meses con un índice menor a los 38.7.

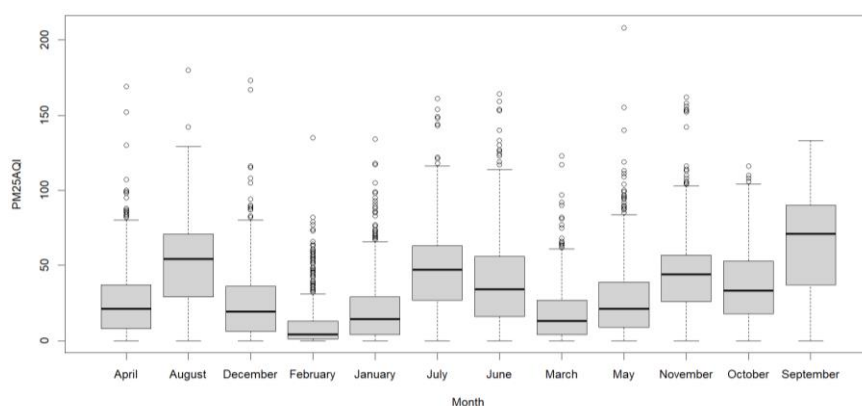


Figure 4. PM2.5AQI por meses

El PM2.5AQI presenta valor máximo en septiembre debido a que es el mes donde los agricultores queman sus restos como parte de la pre siembra y, presenta valores mínimos en febrero porque la lluvia presentes en la sierra central arrastra el material particulado.

En la figura 5, observamos que en verano el 50% tiene una humedad menor a 5.5 y el otro 50% tiene una humedad mayor a 5.5, en segundo lugar en otoño el 50% tiene una humedad menor a 4.9 y el otro 50% tiene una humedad mayor a 4.9, en tercer lugar en primavera el 50% tiene una humedad menor a 4.5 y el otro 50% tiene una humedad mayor a 4.5 y finalmente en invierno el 50% tiene una humedad menor a 4.2 y el otro 50% tiene una humedad mayor a 4.2

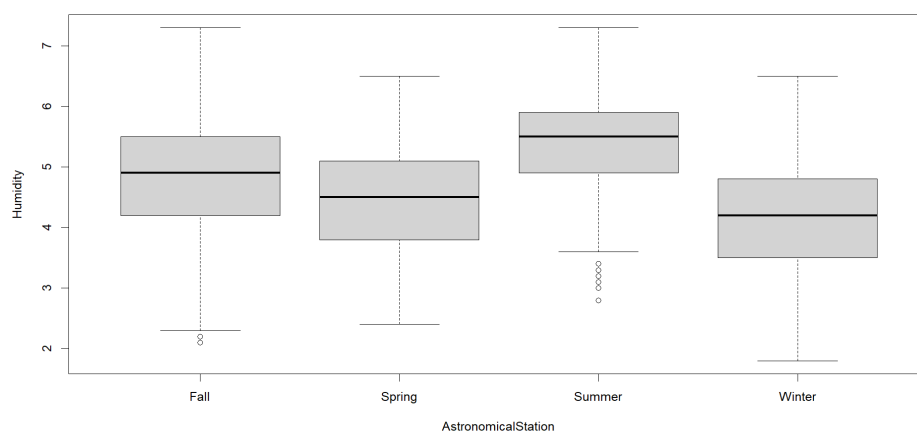


Figure 5. Humedad por estaciones del año

La humedad muestra su valor más alto en enero debido a las Fuertes lluvias y su valor más bajo en julio debido a la temporada seca del año y temporada de heladas. La hume-

dad absoluta muestra que es más alta en Verano seguido de otoño debido al inicio y finalización de las lluvias y en la figura 6, observamos que en febrero y marzo el 50% tiene una humedad menor a 5.5 y el otro 50% tiene una humedad mayor a 5.5, en segundo lugar en enero y abril el 50% tiene una humedad menor a 5.2 y el otro 50% tiene una humedad mayor a 5.2, en tercer lugar en mayo y diciembre el 50% tiene una humedad menor a 4.8 y el otro 50% tiene una humedad mayor a 4.8, en octubre el 50% tiene una humedad menor a 4.7 y el otro 50% tiene una humedad mayor a 4.7 y finalmente en los demás meses el 50% tiene una humedad menor a 4.6 y el otro 50% tiene una humedad mayor a 4.6.

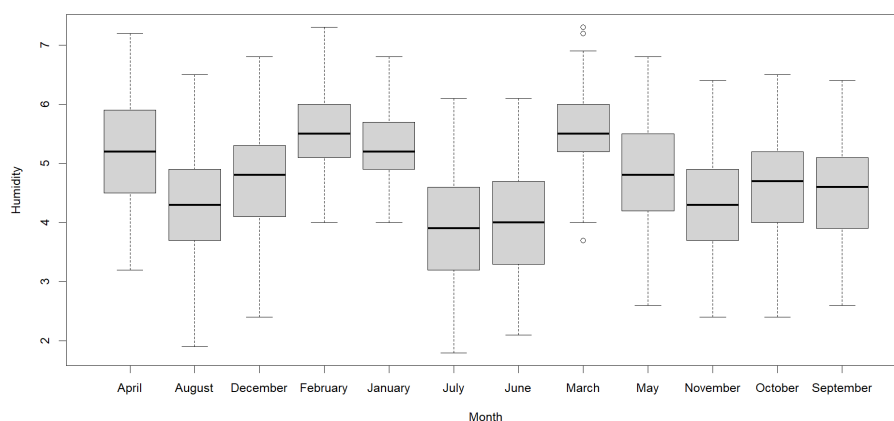


Figure 6. Humedad por meses

En la figura 7, observamos que en primavera el 50% tiene una temperatura menor a 11 °C y el otro 50% tiene una temperatura mayor a 11 °C, y existe un empate de temperatura en otoño, invierno y verano con el 50% de temperatura menor a 9°C y el otro 50% tiene una temperatura mayor a 9°C.

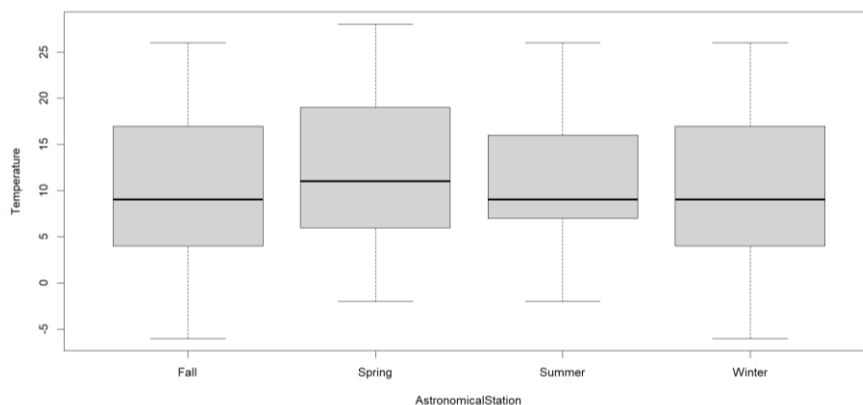


Figure 7. Temperatura por estaciones del año

En la figura 8, observamos que en octubre y Noviembre el 50% tiene una temperatura menor a 12°C y el otro 50% tiene una humedad mayor a 12°C, en segundo lugar en diciembre el 50% tiene una temperatura menor a 11°C y el otro 50% tiene una temperatura mayor a 11°C, en tercer lugar en setiembre el 50% tiene una temperatura menor a 10°C y el otro

50% tiene una temperatura mayor a 10°C, en octubre y finalmente en los demás meses el 50% tiene una temperatura menor a 9°C y el otro 50% tiene una temperatura mayor a 9°C.

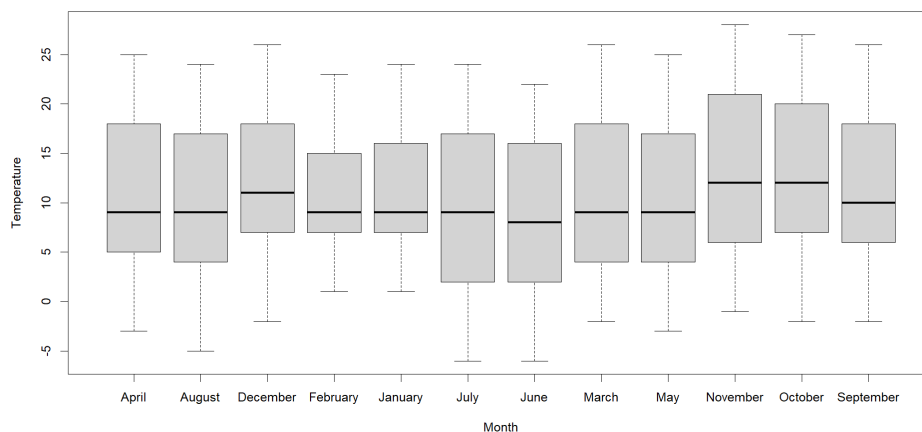


Figure 8. Temperatura por meses del año

La temperatura alcanza su valor más bajo en junio por la presencia de las heladas matutinas y su punto más alto en noviembre por la temporada soleada. Estos valores del índice de la calidad del aire, la humedad y la temperatura, son cruciales para la planificación y la toma de decisiones relacionadas con la salud pública y el medio ambiente ya que por ejemplo el mayor AQI nos indica que el humo de todo el valle del mantaro donde se ubica el centro astronomico de Huachac debería hacer campañas educativas de la no quema de los despojos agrícolas. La temperatura media más alta fue en primavera debido a la ausencia de lluvias y vientos en la zona y la más baja presenta en invierno por la presencia de heladas matutinas. La desviación estándar indica que la variabilidad es mayor en invierno en comparación con las otras estaciones debido a que en estos meses hay la presencia de bajas temperaturas como las más altas. Los valores mínimos y máximos varían según la estación siendo los más altos y bajos asociados con el invierno y la primavera, aunque el más alto se presenta en el otoño con un AQI de 208, esto debido a que los agricultores preparan su terreno para la siembra y queman los rastrojos de las cosechas.

En la investigación de las 16543 observaciones se tomó el 80% (13234) de las observaciones para el entrenamiento y un 20%(3309) para el testeo, encontrando en el entrenamiento una validación cruzada de remuestreo con 3 repeticiones un RMSE=25.36203, un Rsquared=0.1483206 y un MAE=20.51858.

El análisis de la regresión lineal multivariado con 2 variables regresoras analizadas que son la humedad y la temperatura, así como la variable predictora PM2.5AQI cumpliendo los prerequisites estadísticos arrojó los resultados en el software R Studio que se muestra en la figura 9.

```

> summary(lm)

Call:
lm(formula = .outcome ~ ., data = dat)

Residuals:
    Min       1Q   Median       3Q      Max
-53.087 -19.664  -5.102  16.336 177.352

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  83.40932    1.07213   77.798 < 2e-16 ***
Humidity     -10.37024    0.23201  -44.698 < 2e-16 ***
Temperature  -0.12717    0.03169   -4.013 6.04e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 25.37 on 13269 degrees of freedom
Multiple R-squared:  0.1478,    Adjusted R-squared:  0.1477
F-statistic: 1151 on 2 and 13269 DF,  p-value: < 2.2e-16

```

Figure 9. Resultado del modelo Lineal con R Studio

Estos resultados mostrados en la figura 9, donde la significancia para las variables regresoras fueron altamente significativas con p-value menor a 2×10^{-16} permitieron plantear el modelo de regresión múltiple del PM2.5 mostrado en la ecuación 25.

$$PM2.5AQI = 83.40932 - 10.37024(Humidity) - 0.12717(Temperature) \quad (25)$$

Esto indica que tanto la humedad y la temperatura son inversamente proporcionales con el PM2.5AQI por cada unidad que aumenta las variables predictoras, además, los valores del p-value en cada uno de estas variables son menores al nivel de significancia $\alpha=0.05$ los cuales indican que tienen efecto significativo y es corroborado al aplicar un análisis de varianza del modelo, resultando también sus p-valores menores a $\alpha=0.05$ para ambas variables predictoras. Además, la prueba de Durbin Watson nos da una autocorrelación de 0.881 y un DW de 0.238 con un p-value menor a $\alpha=0.05$ lo cual nos indica que existe autocorrelación casi nula lo cual estamos ante la independencia de las variables predictoras, y todo esto lo corrobora la VIF de 1.08 para ambas variables lo cual indica óptimo por estar muy cercano a 1, ambas variables del modelo son independientes y no existe colinealidad entre ellos.

Por lo tanto, el modelo se encuentra muy robusto de manera predictiva con un AIC de y BIC bastante altos y con un RMSE bajo de 25.5 para los 16544 datos observados en el modelo con un p-value menor a 0.05.

Se obtuvo correlaciones de la variable predicha PM2.5AQI con las variables predictoras humedad y la temperatura. En primer lugar, la correlación del PM2.5AQI con la Humedad es de -0.378 con lo cual nos muestra una dependencia lineal y además y es de manera inversamente proporcional, es decir que mientras el PM2.5AQI aumenta la humedad absoluta disminuye y viceversa, lo que significa que la presencia de humedad influye en la disminución o aumento del PM2.5AQI presente en el aire. Asimismo, podemos apreciar que la correlación entre PM2.5AQI y la temperatura tiene una correlación de -0.134, lo cual indica que también se trata de una correlación del tipo inversamente proporcional, lo que quiere decir que mientras el PM2.5AQI aumenta la temperatura disminuye, ello nos da la idea que si aumenta la temperatura disminuye el PM2.5AQI e igualmente de manera inversa, aunque este índice nos muestra que es menor al índice que presenta la humedad, por lo tanto nos indica que es menor su influencia lineal en el PM2.5AQI. Por otro lado, el p-value para la humedad y la temperatura es menor a 0.05, lo cual confirma

la hipótesis de que en ambos casos existe dependencia lineal entre la variable predicha PM2.5AQI y las variables predictoras de la humedad y la temperatura.

Los gráficos mostrados en la figura 10, señalan los residuos normalizados y valores ajustados, con cuantiles residuales y la raíz del error cuadrático medio, por el cual el modelo de regresión lineal cumple con todas las especificaciones y prerrequisitos para la predicción.

Gráfico de perfomance del modelo de regresión lineal.

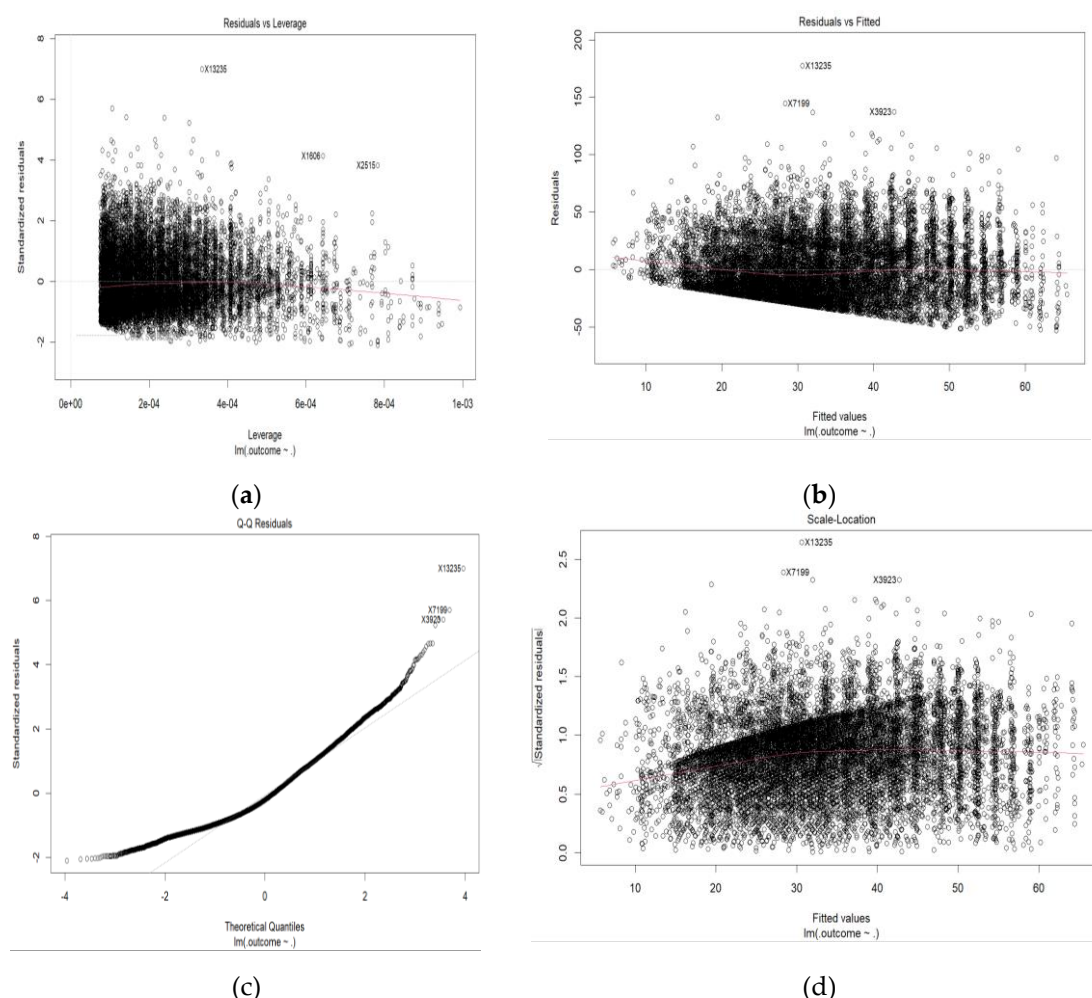


Figure 10. Gráficas de perfomance del modelo de regresión lineal: (a) Describe los residuos normalizados; (b) Describe los residuales de los datos (c) Describe los residuales normalizados; (d) se describe la raíz cuadrada de los residuales.

Los resultados para el modelo de regresión Ridge, con un entrenamiento de 13234 observaciones y 3309 observaciones para el testeo, se encontró que las métricas del remuestreo a través de los parámetros de ajuste son las que se muestran en la table 2.

Table 2. Métricas del performance del modelo de regresión de Ridge.

Lambda	RMSE	Rsquared	MAE
0.000100 *	25.36471	0.1483078	20.53783
0.250075	25.36471	0.1483078	20.53783
0.500050	25.36471	0.1483078	20.53783
0.750025	25.36471	0.1483078	20.53783
1.000000	25.36471	0.1483078	20.53783

* Valores de penalización λ del modelo Ridge.

En la figura 6, se muestra los gráficos para el modelo de regresión Ridge con la validación cruzada y los logaritmos de los valores de penalización Lambdas, así como los valores de los coeficientes de los modelos y la importancia de las variables regresoras.

Gráfico de performance del modelo de regresión Ridge.

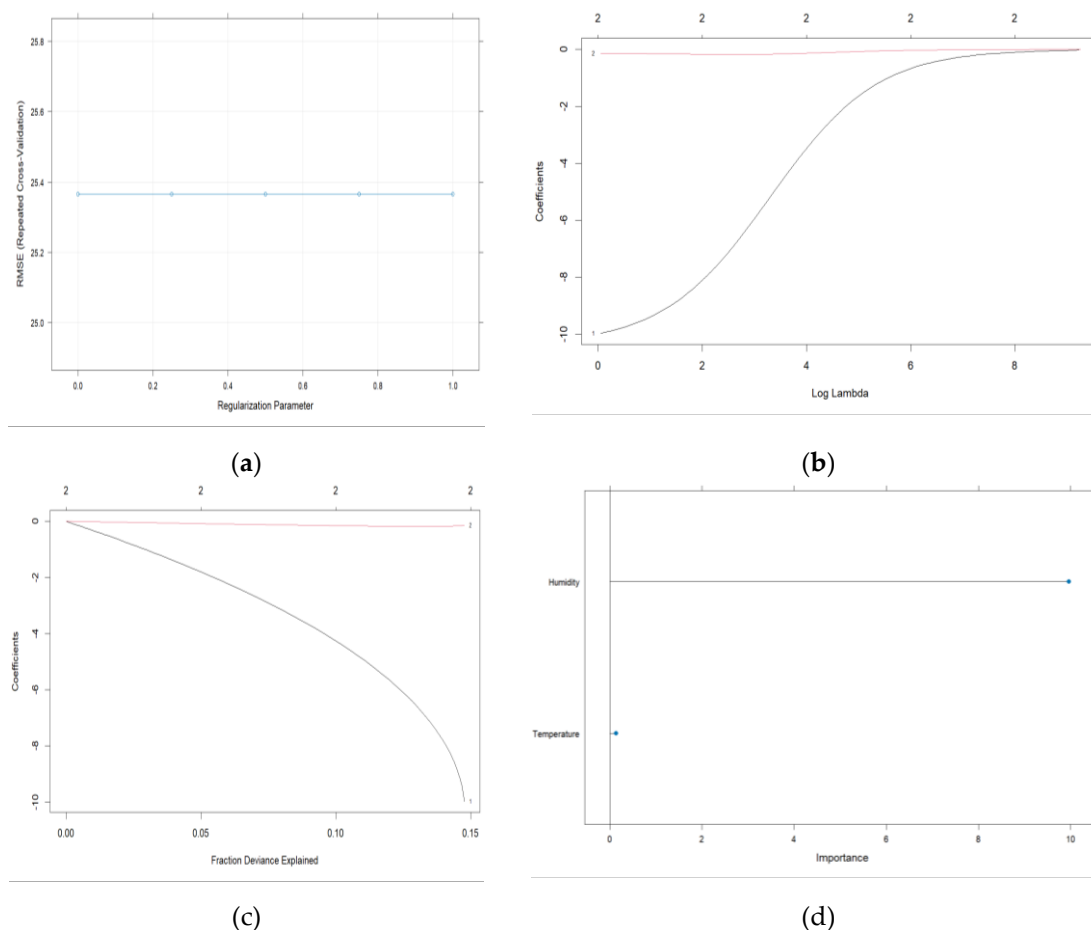


Figure 11. Gráficas de performance del modelo de regresión Ridge: (a) Describe el RMSE de la validación cruzada; (b) Describe los coeficientes de penalización Ridge logarítmica (c) Describe los coeficientes de ridge; (d) Describe la importancia de las variables.

Los resultados para el modelo de regresión lasso, con un entrenamiento de 86654 observaciones y 21663 observaciones para el testeo, se encontró que las métricas del remuestreo a través de los parámetros de ajuste son las que se muestran en la table 3.

Table 3. Métricas del performance del modelo de regresión de Lasso.

Lambda	RMSE	Rsquared	MAE
0.000100 *	25.36206	0.1483177	20.51777
0.050075	25.36206	0.1483177	20.51777
0.100050	25.36229	0.1483129	20.52004
0.150025	25.36268	0.1483044	20.52282
0.200000	25.36323	0.1482928	20.52574

* Valores de penalización λ del modelo Lasso.

En la figura 7, se muestra los gráficos para el modelo de regresión Lasso con la validación cruzada y los logaritmos de los valores de penalización Lambdas, así como los valores de los coeficientes del modelos y la importancia de las variables regresoras.

Gráfico del performance del modelo de la regresión Lasso.

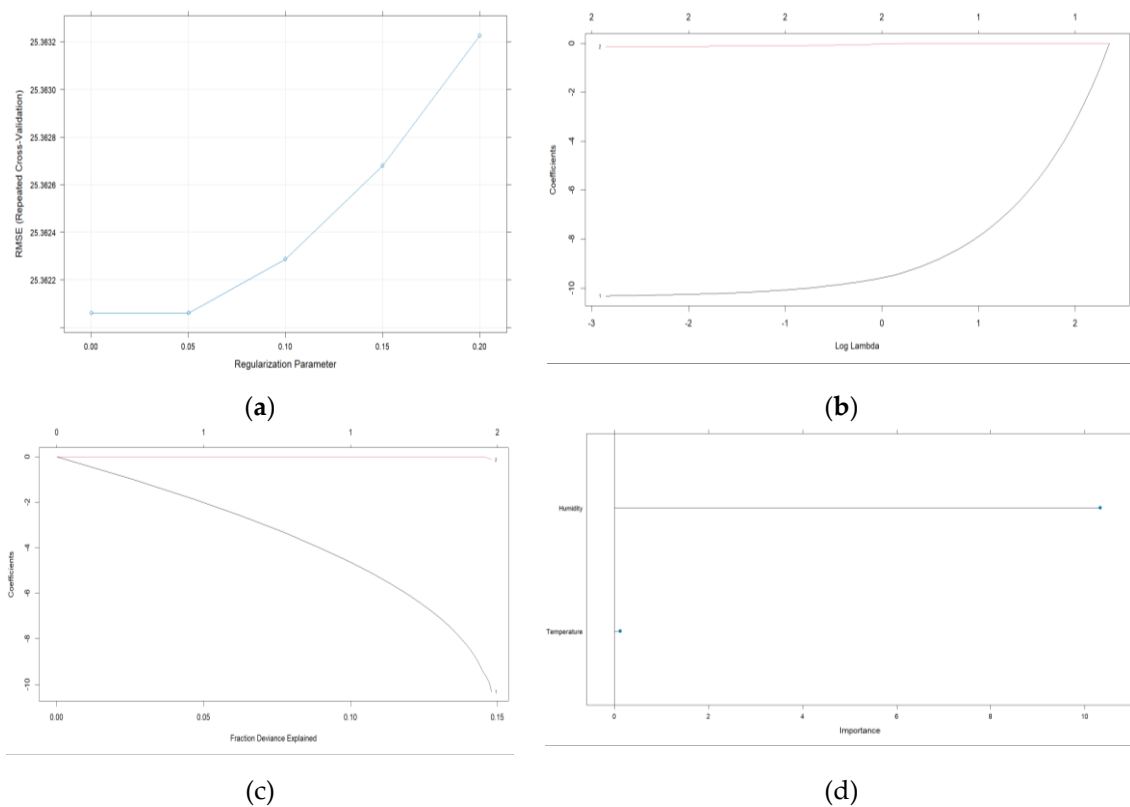


Figure 12. Gráficas del performance del modelo de regresión Lasso: (a) Describe el RMSE de la validación cruzada; (b) Describe los coeficientes de penalización Ridge logarítmica (c) Describe los coeficientes de ridge; (d) Describe la importancia de las variables.

Los resultados para el modelo de regresión Elastic Net, con un entrenamiento de 13234 observaciones y 3309 observaciones para el testeo, se encontró que las métricas del re-muestreo a través de los parámetros de ajuste son las que se muestran en las tablas 4, 5 y 6.

Table 4. Métricas del perfomance del modelo de regresión Elastic Net.

MAE	Min	Mean	Max
LinearModel *	19.72584	20.51489	21.24642
Ridge *	19.75967	20.53783	21.24997
Lasso *	19.73003	20.51777	21.24672
ElasticNet *	19.73139	20.51858	21.24663

* Métrica MAE de los modelos predictivos de penalización

Table 5. Métricas del perfomance del modelo de regresión Elastic Net.

RMSE	Min	Mean	Max
LinearModel *	24.33021	25.36196	26.30641
Ridge *	24.33905	25.36471	26.28795
Lasso *	24.33268	25.36206	26.30250
ElasticNet *	24.33179	25.36203	26.30244

* Métrica RMSE de los modelos predictivos de penalización

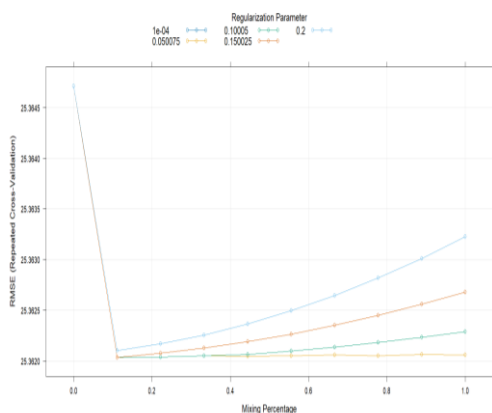
Table 6. Métricas de perfomance del modelo de regresión Elastic Net.

Rquared	Min	Mean	Max
LinearModel *	0.1035190	0.1483207	0.1794977
Ridge *	0.1034498	0.1483078	0.1796028
Lasso *	0.1035378	0.1483177	0.1794510
ElasticNet *	0.1035171	0.1483206	0.1795012

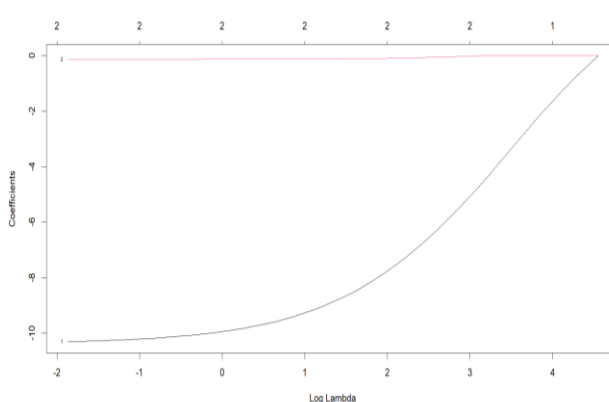
* Métrica Rsquared de los modelos predictivos de penalización

En la figura 8, se muestra los gráficos para el modelo de regresión Lasso con la validación cruzada y los logaritmos de los valores de penalización Lambdas, así como los valores de los coeficientes de los modelos y la importancia de las variables regresoras.

Gráfico de perfomance del modelo de la regresión Elastic Net.



(a)



(b)

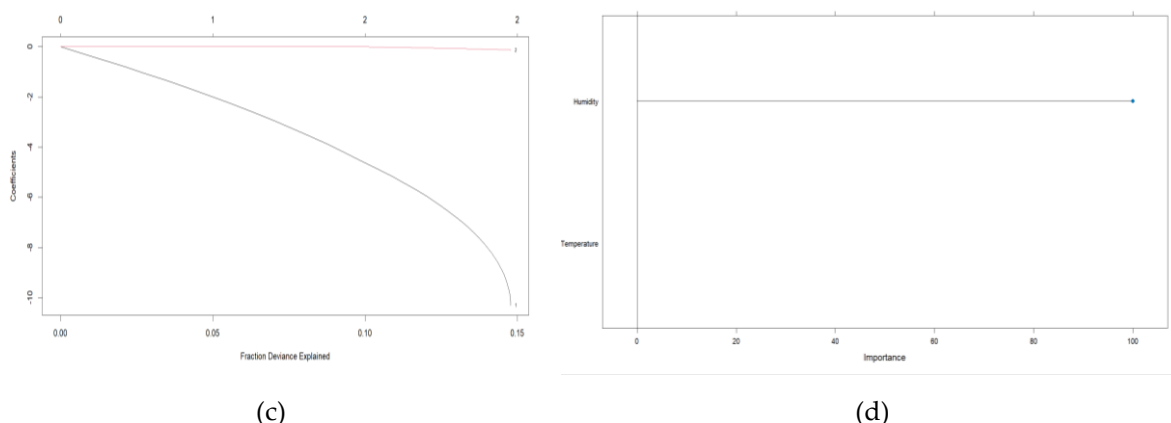


Figure 13. Gráficas de performance del modelo de regresión Elastic Net: (a) Describe el RMSE de la validación cruzada; (b) Describe los coeficientes de penalización Ridge logarítmica (c) Describe los coeficientes de ridge; (d) Describe la importancia de las variables

El mejor modelo que permitio predecir el PM2.5AQI es el décimo tercero con $\alpha=0.1111111$ y un Lambda $\lambda=0.150025$ cuyo modelo se muestra en la ecuación 26.

$$PM2.5AQI = 83.0846522 - 10.3022000(Humidity) - 0.1268124(Temperature) \quad (26)$$

Esto significa que, por cada año, la el PM2.5AQI disminuye en 10.3022000, así mismo para la temperatura el PM2.5AQI disminuye en 0.1268124. En la predicción para error cuadrático medio del entrenamiento se obtuvo un valor de 25.36255 y para el testeo un valor de 25.84308 lo que hace una diferencia de medias de 0.48053, lo que representa el 98.1406% de la eficiencia de la predicción respecto al testeo.

4. Discusión

El método LASSO (Least Absolute Shrinkage and Selection Operator) aplica la regresión con procedimientos de contracción de parámetros para seleccionar las variables predictoras que importan significativamente y minimizan algunas variables que poco o nada afectan al modelo, esto ajustando para que los coeficiente sean máximos y mínimos [24]. En el modelado de MLP y SVM de Qian, encontraron un RMSE de 52.74 y 35.88 pero en nuestra investigación nos presenta un RMSE de 25.5, lo cual es bastante óptimo el modelo encontrado ya que es mucho menor en valor de ajuste con los valores reales registrados.

En una investigación se halló un modelo de regresión lineal $y = 60.219 - 1.7493x$ con un coeficiente de determinación $R^2 = 0.2836$ sobre la concentración de PM2.5 y PM10 en Lima y Callao en la que utilizaron datos del SENAMHI [35], frente al encontrado en esta investigación de $R^2 = 0.1478$.

Geraldo-Campos et al [36] halló modelos de regularización Lasso y Ridge para la predicción del riesgo crediticio en reactiva Perú con 501 298 empresas analizadas, del sector económico, entidad otorgante, monto cubierto y departamento como regresoras y el nivel de riesgo como predictor, determinando un modelo de regresión lasso con $\lambda_{60} = 0.00038$ y un RMSE=0.3573685 ; así como una regresión Ridge con $\lambda_{100} = 0.00910$ y un RMSE=0.3573812, representados en la ecuación 27.

$$Y_{Lasso} = 0.51487 + 0.05878(Economic\ sector) - 0.19292(Credit\ granting\ entity) + 1.29671(Amount\ covered) + 0.03115(Departament) \quad (27)$$

En esta investigación se encontró que los modelos lasso (RMSE=25.36206), Ridge (RMSE=25.36471) y Elastic net (RMSE=25.36203), donde el mejor modelo hallado con un $\alpha=0.1111111$ y un valor de Lambda $\lambda=0.150025$ se representa en la ecuación 28.

$$Y = 83.0846522 - 10.3022000(Humidity) - 0.1268124(Temperature) \quad (28)$$

En general, los datos revelan patrones interesantes sobre las estaciones del año y su impacto en la calidad del aire, la humedad y la temperatura. Estos hallazgos podrían ser útiles para comprender mejor cómo estas variables cambian a lo largo del año y cómo podrían afectar a las personas que son bastante vulnerables al medio ambiente que los rodea porque debido a tener un aire contaminado puede hasta ser mortal en algunos casos.[37]. La investigación encontró el aumento de la calidad del aire en verano, podría tener implicaciones positivas para la salud pública de la zona sierra central, mientras que la mayor humedad en primavera y verano podría influir en la comodidad y el bienestar de las personas ya que son épocas donde mayormente se genera el turismo vivencial en esta zona. La variabilidad estacional en la temperatura también podría tener consecuencias en diversas áreas, desde la agricultura hasta la generación de la Economía [38].

El análisis de los datos proporcionados revela diferencias significativas entre las estaciones del año en términos de calidad del aire, humedad y temperatura. Estos hallazgos podrían ser importantes para comprender mejor los patrones estacionales en estas variables y sus posibles impactos en la sociedad y el medio ambiente. El documento destaca la importancia de considerar las variaciones estacionales al abordar cuestiones relacionadas con la calidad del aire y el clima, y sugiere una mayor investigación para comprender completamente las implicaciones de estos hallazgos.

Los valores de PM25AQI muestran variaciones significativas a lo largo del año, con un pico en septiembre debido a las constantes quemadas de los rastrojos [39] y desechos agrícolas y que lo hacen los agricultores como previo a la siembra anual de sus productos en desmedro de la Calidad del aire, esto amerita un llamado a las autoridades de las poblaciones para realizar campañas educativas medioambientales porque la presencia de PM2.5 en el aire tiene consecuencias muchas veces mortales especialmente en la población vulnerable. [37] La humedad y la temperatura también muestran fluctuaciones interesantes durante los meses de año, en cuanto a la humedad presenta valores más altos en enero debido a la presencia de las lluvias fuertes e intensas que presenta la sierra central del Perú, [40] esto hace que la humedad absoluta sea más concentrado en esta época. Mientras que la temperatura más baja registrada en junio de $-6\text{ }^{\circ}\text{C}$ y un máximo de $22\text{ }^{\circ}\text{C}$ nos da una gran variabilidad de temperatura por la presencia de las heladas matutinas acorde a lo encontrado por Saavedra et. al. [41]. Los datos presentados están dentro de los rangos encontrados por los otros investigadores en cuanto a las variaciones mensuales en la calidad del aire, la humedad y la temperatura.

5. Conclusions

El modelo predictivo encontrado es bastante robusto dado a que cumplen con los indicadores a favor del modelo y de acuerdo al análisis de regresión LASSO con un AIC y BIC altos y con un RMSE bajo para los 16543 datos analizados.

La estación del año con el mayor índice PM25AQI fue el invierno, con un valor de 52.6 con una variabilidad de 29.4 lo cual es menor a 100 que nos indica que es “Moderado” ó puede haber problemas de salud moderado para las personas que son inusualmente sensibles a la contaminación del aire.

El PM25AQI muestra un pico en el mes de septiembre por el humo de la quema de rastrojos por los agricultores, mientras que la humedad alcanza su punto máximo en enero debido a las lluvias de esa temporada y la temperatura mínima en junio por la presencia de heladas.

Estos datos son vitales para comprender las fluctuaciones estacionales en estos parámetros ambientales, lo que tiene implicaciones importantes para la salud pública y su planificación Ambiental. Los datos encontrados muestran que en general la mayor parte del año presenta Buena Calidad del aire, mostrando algunos picos de índice insalobre AQI (208) como discordantes o anómalos.

Supplementary Materials: The following supporting information can be downloaded at: www.mdpi.com/xxx/s1, Figure S1: title; Table S1: title; Video S1: title.

Author Contributions: Conceptualización, EO., J.G y G.R.; metodología, EO., J.G, G.R, JS; software, H.F., J.S.; validación, H.F., J.S. y J.P.; análisis formal, O.P.; investigación, EO., J.G y G.R.; recursos, EO., J.G y G.R.; curación de datos, EO., J.G y G.R.; redacción: preparación del borrador original, EO., J.G y G.R.; redacción: revisión y edición, H.F., J.S. , J.P. y O.P.; visualización, EO., J.G y G.R.; supervisión, O.P.; administración de proyectos, EO., J.G y G.R.; adquisición de financiación, EO., J.G y G.R. Todos los autores han leído y aceptado la versión publicada del manuscrito.

Funding: Please add: Esta investigación no recibió financiación externa. La APC fue financiada por la Universidad Peruana Unión..

Data Availability Statement: Los datos estan disponibles en mendeley data. <https://data.mendeley.com/>

Acknowledgments: Expresamos nuestra gratitud al Ingeniero Luis Flores, director de la estación Director del Observatorio de Huayao, Huachac y al Ingeniero Luis Fernando Suarez por su valiosa contribución, en el acceso a la data y apoyo incondicional en esta investigación. Asimismo a Erick R. Moreno Alvarado de la Universidad Nacional Intercultural Selva Central Juan Santos Atahualpa, por sus aportes y contribución en la discusión y comentarios pertinentes.

Conflicts of Interest: Los autores declaramos no tener conflict de intereses, Los financiadores no tuvieron ningún papel en el diseño del estudio; en la recopilación, análisis o interpretación de datos; en la redacción del manuscrito; o en la decisión de publicar los resultados. Este trabajo no conto con financiamiento externo, fue autofinanciado.

Referencias

- [1] Santillan P., Rodriguez M., Orozco J., Rios I., and Bayas K., “Evaluación de la concentración y distribución espacial de material particulado en los campus de la UNACH - Riobamba,” *NOVASINERGIA REVISTA DIGITAL DE CIENCIA, INGENIERÍA Y TECNOLOGÍA*, vol. 4, no. 2, pp. 111–126, Dec. 2021, doi: 10.37135/ns.01.08.07.
- [2] D. Álvarez-Tolentino and L. Suárez-Salas, “Apportionment of emission sources of PM10 and PM2.5 at urban sites of Mantaro Valley, Peru,” *Revista Internacional de Contaminacion Ambiental*, vol. 36, no. 4, pp. 875–892, 2020, doi: 10.20937/RICA.53473.

- [3] M. Elbayoumi, N. A. Ramli, and N. F. F. M. Yusof, "Development and comparison of regression models and feedforward backpropagation neural network models to predict seasonal indoor PM_{2.5}–10 and PM_{2.5} concentrations in naturally ventilated schools," *Atmos Pollut Res*, vol. 6, no. 6, pp. 1013–1023, 2015, doi: 10.1016/j.apr.2015.09.001.
- [4] M. Czwajdzinska, M. Terpińska, A. Kuźniarski, S. Płaczowska, and A. Piwowar, "Exposure to PM_{2.5} and PM₁₀ and COVID-19 infection rates and mortality: A one-year observational study in Poland," *Biomed J*, vol. 44, no. 6, pp. S25–S36, 2021, doi: 10.1016/j.bj.2021.11.006.
- [5] M. Ponzano *et al.*, "The impact of PM_{2.5}, PM₁₀ and NO₂ on Covid-19 severity in a sample of patients with multiple sclerosis: A case-control study," *Mult Scler Relat Disord*, vol. 68, Dec. 2022, doi: 10.1016/j.msard.2022.104243.
- [6] OMS, "Calidad del Aire Ambiente." Accessed: Feb. 09, 2023. [Online]. Available: <https://www.paho.org/es/temas/calidad-aire-salud/calidad-aire-ambiente>
- [7] D. Álvarez-Tolentino and L. Suárez-Salas, "Apportionment of emission sources of PM₁₀ and PM_{2.5} at urban sites of Mantaro Valley, Peru," *Revista Internacional de Contaminacion Ambiental*, vol. 36, no. 4, pp. 875–892, 2020, doi: 10.20937/RICA.53473.
- [8] M. Załuska and K. Gładyszewska-Fiedoruk, "Regression model of PM_{2.5} Concentration in a single-family house," *Sustainability (Switzerland)*, vol. 12, no. 15, 2020, doi: 10.3390/su12155952.
- [9] R. Rojano, Y. Mendoza, H. Arregoces, and G. Restrepo, "Dispersión de Contaminantes del Aire (PM₁₀, NO₂, CO, COV y HAP) emitidos desde una estación modular de compresión, tratamiento y medición de gas natural," in *Informacion Tecnologica*, Centro de Informacion Tecnologica, 2016, pp. 99–110. doi: 10.4067/S0718-07642016000500012.
- [10] T. Tao, Y. Shi, K. Gilbert, and X. Liu, "Spatiotemporal variations of air pollutants based on ground observation and emission sources over 19 Chinese urban agglomerations during 2015–2019," *Sci Rep*, vol. 12, no. 1, Dec. 2022, doi: 10.1038/s41598-022-08377-9.
- [11] J. Green and S. Sánchez, "La calidad del aire en América Latina: Una visión panorámica," 2012.
- [12] C. Cordova *et al.*, "Air quality assessment and pollution forecasting using artificial neural networks in Metropolitan Lima-Peru," *Sci Rep*, vol. 11, no. 1, Dec. 2021, doi: 10.1038/s41598-021-03650-9.
- [13] H. J. Chu, B. Huang, and C. Y. Lin, "Modeling the spatio-temporal heterogeneity in the PM₁₀-PM_{2.5} relationship," *Atmos Environ*, vol. 102, no. 1, pp. 176–182, Feb. 2015, doi: 10.1016/j.atmosenv.2014.11.062.
- [14] E. Alpaydin, *Introduction to Machine Learning*, vol. 13, no. 1. 1959.
- [15] A. Géron and Rudolph. Russell, "Machine learning step-by-step guide to implement machine learning algorithms with Python," *O'Reilly Media, Inc*, p. 106, 2019.
- [16] S. R. M-Dawam and K. R. Ku-Mahamud, "Reservoir water level forecasting using normalization and multiple regression," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 14, no. 1, pp. 443–449, Apr. 2019, doi: 10.11591/ijeecs.v14.i1.pp443-449.
- [17] "Statistical and Machine-Learning Data Mining Techniques for Better Predictive Modeling and Analysis of Big Data Third Edition."
- [18] A. Géron and Rudolph. Russell, "Machine learning step-by-step guide to implement machine learning algorithms with Python," *O'Reilly Media, Inc*, p. 106, 2019.
- [19] X. Yan and X. Gang Su, "Linear Regression Analysis Theory and Computing."
- [20] Gujarati D., *Econometría*, Quinta., vol. 5, no. 1. McGRAW-HILL, 2010.

- [21] D. Gabauer, R. Gupta, H. A. Marfatia, and S. M. Miller, "Estimating U.S. housing price network connectedness: Evidence from dynamic Elastic Net, Lasso, and ridge vector autoregressive models," *International Review of Economics and Finance*, vol. 89, pp. 349–362, Jan. 2024, doi: 10.1016/j.iref.2023.10.013.
- [22] T. Aliaj, M. Ciganovic, and M. Tancioni, "Nowcasting inflation with Lasso-regularized vector autoregressions and mixed frequency data," *J Forecast*, vol. 42, no. 3, pp. 464–480, Apr. 2023, doi: 10.1002/for.2944.
- [23] F. Emmert-Streib and M. Dehmer, "High-Dimensional LASSO-Based Computational Regression Models: Regularization, Shrinkage, and Selection," *Machine Learning and Knowledge Extraction*, vol. 1, no. 1. MDPI, pp. 359–383, Dec. 01, 2019. doi: 10.3390/make1010021.
- [24] P. Qian, H. Chen, K. Lu, and Z. Liu, "Optimal Prediction Model for PM2.5 Using Lasso Regression," *Proceedings - 2023 2nd International Conference on Data Analytics, Computing and Artificial Intelligence, ICDACAI 2023*, pp. 309–314, Jan. 2023, doi: 10.1109/ICDACAI59742.2023.00065.
- [25] L. A. Geraldo-Campos, J. J. Soria, and T. Pando-Ezcurra, "Machine Learning for Credit Risk in the Reactive Peru Program: A Comparison of the Lasso and Ridge Regression Models," *Economies*, vol. 10, no. 8, Aug. 2022, doi: 10.3390/economies10080188.
- [26] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*. 2014. [Online]. Available: <http://www.springer.com/series/417>
- [27] F. Emmert-Streib and M. Dehmer, "High-Dimensional LASSO-Based Computational Regression Models: Regularization, Shrinkage, and Selection," *Machine Learning and Knowledge Extraction*, vol. 1, no. 1. MDPI, pp. 359–383, Dec. 01, 2019. doi: 10.3390/make1010021.
- [28] K. Das, N. Das Chatterjee, D. Jana, and R. K. Bhattacharya, "Application of land-use regression model with regularization algorithm to assess PM2.5 and PM10 concentration and health risk in Kolkata Metropolitan," *Urban Clim*, vol. 49, May 2023, doi: 10.1016/j.uclim.2023.101473.
- [29] M. Shanker, "Effect of Data Standardization on Neural Network Training," *Int. J. Mgmt Sci*, vol. 24, no. 4, pp. 385–397, 1996.
- [30] L. Atlas *et al.*, "Performance Comparisons Between Backpropagation Networks and Classification Trees on Three Real-World Applications."
- [31] S. R. M-Dawam and K. R. Ku-Mahamud, "Reservoir water level forecasting using normalization and multiple regression," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 14, no. 1, pp. 443–449, Apr. 2019, doi: 10.11591/ijeecs.v14.i1.pp443-449.
- [32] M. Bounessah and B. P. Atkin, "An application of exploratory data analysis (EDA) as a robust non-parametric technique for geochemical mapping in a semi-arid climate," *Applied Geochemistry*, vol. 18, no. 8, pp. 1185–1195, Aug. 2003, doi: 10.1016/S0883-2927(02)00247-0.
- [33] E. D. Marques *et al.*, "Geochemical mapping by stream sediments of the NW portion of Quadrilátero Ferrífero, Brazil: Application of the exploratory data analysis (EDA) and a proposal for generation of new gold targets in Pitangui gold district," *J Geochem Explor*, vol. 250, Jul. 2023, doi: 10.1016/j.gexplo.2023.107232.
- [34] D. Chowdhury, S. Hovda, and B. Lund, "Analysis of cuttings concentration experimental data using exploratory data analysis," *J Pet Sci Eng*, vol. 221, Feb. 2023, doi: 10.1016/j.petro.2022.111254.
- [35] S. A. Pacsi Valdivia, "ANÁLISIS TEMPORAL Y ESPACIAL DE LA CALIDAD DEL AIRE DETERMINADO POR MATERIAL PARTICULADO PM10 Y PM2,5 EN LIMA METROPOLITANA," *Anales Científicos*, vol. 77, no. 2, p. 273, Dec. 2016, doi: 10.21704/ac.v77i2.699.

- [36] L. A. Geraldo-Campos, J. J. Soria, and T. Pando-Ezcurra, "Machine Learning for Credit Risk in the Reactive Peru Program: A Comparison of the Lasso and Ridge Regression Models," *Economies*, vol. 10, no. 8, Aug. 2022, doi: 10.3390/economies10080188.
- [37] Q. Di *et al.*, "Air pollution and mortality in the medicare population," *New England Journal of Medicine*, vol. 376, no. 26, pp. 2513–2522, Jun. 2017, doi: 10.1056/NEJMoa1702747.
- [38] M. Carreras-Sospedra *et al.*, "Air quality and health impacts of the 2020 wildfires in California," *Fire Ecology*, vol. 20, no. 1, p. 6, Dec. 2024, doi: 10.1186/s42408-023-00234-y.
- [39] O. Adegboye, "Field Burning Fallout: Quantifying PM2:5 Emissions from Sugarcane Fires," *Environ Health Perspect*, vol. 130, no. 8, p. 084003, Aug. 2022, doi: 10.1289/EHP11533.
- [40] S. Kumar *et al.*, "Rainfall characteristics in the mantaro basin over tropical andes from a vertically pointed profile rain radar and in-situ field campaign," *Atmosphere (Basel)*, vol. 11, no. 3, p. 248, Mar. 2020, doi: 10.3390/atmos11030248.
- [41] M. Saavedra, C. Junquas, J. C. Espinoza, and Y. Silva, "Impacts of topography and land use changes on the air surface temperature and precipitation over the central Peruvian Andes," *Atmos Res*, vol. 234, p. 104711, Apr. 2020, doi: 10.1016/j.atmosres.2019.104711.