

UNIVERSIDAD NACIONAL INTERCULTURAL DE LA SELVA CENTRAL JUAN SANTOS ATAHUALPA



FACULTAD DE INGENIERÍA

ESCUELA PROFESIONAL DE INGENIERÍA AMBIENTAL

Estimación de concentraciones de Material Particulado fino $PM_{2.5}$
mediante modelos de Machine Learning en la ciudad de Concepción
durante el periodo 2023-2025

TESIS

Para optar el título profesional de Ingeniero Ambiental

AUTORAS

Juliana Neri Pardo Aiquipa

Estefani Nayeli Soto Onofre

ASESOR

Dr. Daniel Martin Álvarez Tolentino

Chanchamayo, Perú

2026

AGRADECIMIENTO

Expresamos nuestro más sincero agradecimiento a todas las personas que hicieron posible la realización de esta investigación.

En primer lugar, agradecemos a nuestro asesor de tesis, por su orientación, apoyo constante y valiosos conocimientos brindados a lo largo de este proceso, los cuales fueron fundamentales para el desarrollo y culminación de este trabajo.

Asimismo, agradecemos a nuestros docentes, quienes a lo largo de nuestra formación académica compartieron sus conocimientos y experiencias, contribuyendo significativamente a nuestro crecimiento profesional.

De manera especial, agradecemos a nuestras familias, por su apoyo incondicional, comprensión y motivación permanente, que nos impulsó a seguir adelante y culminar con éxito esta etapa.

También extendemos nuestro agradecimiento a nuestros compañeros y amigos, por su colaboración, apoyo y ánimo durante todo el proceso de investigación.

Finalmente, agradecemos a todas aquellas personas que, de una u otra forma, contribuyeron en la realización de este trabajo.

DEDICATORIA

Dedicamos el presente trabajo, en primer lugar, a nuestras familias, por ser el pilar fundamental en nuestras vidas, por su amor, apoyo incondicional y constante motivación a lo largo de nuestra formación académica; a nuestros padres, por los valores inculcados, el esfuerzo y sacrificio que hicieron posible el cumplimiento de nuestras metas; a nuestros seres queridos, por su comprensión, paciencia y palabras de aliento en los momentos más desafiantes; y, finalmente, a nosotros mismos, por la perseverancia, compromiso y dedicación que nos permitieron culminar satisfactoriamente esta importante etapa de nuestras vidas.

RESUMEN

La presente investigación evalúa el desempeño de modelos de Machine Learning para estimar las concentraciones de material particulado fino $PM_{2.5}$ en la ciudad de Concepción, Junín, Perú, durante el periodo 2023–2025. Se analizaron las condiciones meteorológicas predominantes, la variabilidad temporal del contaminante y su relación con variables climáticas y temporales, considerando la influencia de la autocorrelación mediante variables con retardo (lags). La base de datos fue sometida a un proceso de preprocesamiento, validación e imputación de datos faltantes, alcanzando un nivel de completitud del 98.28%, garantizando la confiabilidad del análisis. Se implementaron tres modelos predictivos: Regresión Lineal, Random Forest y Redes Neuronales Artificiales (ANN), evaluados mediante métricas como MAE, RMSE, MAPE y coeficiente de determinación (R^2), además de la prueba estadística no paramétrica de Wilcoxon. Los resultados evidencian que las concentraciones de $PM_{2.5}$ presentan alta variabilidad temporal, con incrementos críticos durante los meses secos de junio a septiembre y episodios que superan frecuentemente las guías de la OMS de $25 \mu g/m^3$, especialmente en septiembre de 2024 donde el 73.33% de los días superaron dicho límite. La ANN presentó el mejor desempeño predictivo ($R^2 = 0.611$; RMSE = 5.417; MAE = 2.566), superando significativamente a los otros modelos ($p < 0.001$), contrario a la hipótesis inicial que establecía al Random Forest como el modelo de mejor desempeño. Asimismo, se identificó que la variable más influyente es el $PM_{2.5}$ rezagado ($pm25_lag1$, $\rho = 0.917$), confirmando la fuerte dependencia temporal del contaminante. Las condiciones meteorológicas, especialmente la baja velocidad del viento inferior a 1 m/s y la marcada estacionalidad de la precipitación, favorecen la acumulación de partículas incrementando el riesgo ambiental en periodos secos. Se concluye que las redes neuronales artificiales constituyen una herramienta eficaz para la predicción de la calidad del aire en zonas andinas, con potencial aplicación en sistemas de alerta temprana y gestión ambiental.

Palabras clave: $PM_{2.5}$, Machine Learning, redes neuronales, calidad del aire, variables meteorológicas.

ABSTRACT

This research evaluates the performance of machine learning models for estimating PM_{2.5} fine particulate matter concentrations in the city of Concepción, Junín, Peru, during the period 2023–2025. Prevailing meteorological conditions, the temporal variability of the pollutant, and its relationship with climatic and temporal variables were analyzed, considering the influence of autocorrelation through lagged variables. The database underwent preprocessing, validation, and imputation of missing data, achieving a completeness level of 98.28%, thus ensuring the reliability of the analysis. Three predictive models were implemented: Linear Regression, Random Forest, and Artificial Neural Networks (ANN), evaluated using metrics such as MAE, RMSE, MAPE, and the coefficient of determination (R^2), in addition to the Wilcoxon signed-rank test. The results show that PM_{2.5} concentrations exhibit high temporal variability, with critical increases during the dry months from June to September and episodes that frequently exceed the WHO guidelines of 25 $\mu\text{g}/\text{m}^3$, especially in September 2024, when 73.33% of the days exceeded this limit. The ANN model showed the best predictive performance ($R^2 = 0.611$; RMSE = 5.417; MAE = 2.566), significantly outperforming the other models ($p < 0.001$), contrary to the initial hypothesis, which proposed Random Forest as the best-performing model. Furthermore, the most influential variable was identified as lagged PM_{2.5} (pm25_lag1, $\rho = 0.917$), confirming the strong temporal dependence of the pollutant. Meteorological conditions, especially low wind speeds below 1 m/s and the marked seasonality of precipitation, favor the accumulation of particles, increasing the environmental risk during dry periods. It is concluded that artificial neural networks constitute an effective tool for predicting air quality in Andean regions, with potential applications in early warning systems and environmental management.

Keywords: PM_{2.5}, Machine Learning, Artificial Neural Networks, Random Forest, air quality, Andean region, temporal variability.

ÍNDICE

AGRADECIMIENTO	ii
DEDICATORIA	iii
RESUMEN	iv
ABSTRACT	v
ÍNDICE DE TABLAS	ix
ÍNDICE DE FIGURAS	x
INTRODUCCIÓN	12
CAPÍTULO I	14
PLANTEAMIENTO DEL PROBLEMA	14
1.1. Fundamentación del problema	14
1.2. Formulación del problema	16
1.2.1 Problema general	16
1.2.2. Problemas específicos	16
1.3. Objetivos	16
1.3.1 Objetivo general	16
1.3.2. Objetivos específicos	16
1.4. Definición y operacionalización de variables	17
1.5. Hipótesis	17
1.5.1. Hipótesis general	17
1.6. Justificación	17
1.6.1 Teórico	17
1.6.2 Practica	18
1.6.3 Ambiental	18
1.6.4 Metodológico	18
1.6.5 Social	18

1.6.6 Economía	18
CAPÍTULO II.....	20
MARCO TEÓRICO	20
2.1. Antecedentes	20
2.1.1 Internacionales.....	20
2.1.2 Nacionales	22
2.1.3. Local	24
2.2. Bases teóricas.....	25
2.2.1. Material particulado fino (PM _{2.5}).....	25
2.2.2. Impactos del PM _{2.5}	26
2.2.3. Factores meteorológicos que influyen en las concentraciones de PM _{2.5}	27
2.2.4. Monitoreo y medición de PM _{2.5}	27
2.2.5. Monitoreo y medición de datos meteorológicos.....	29
2.2.6. Modelos relevantes para series temporales ambientales	30
CAPÍTULO III	32
METODOLOGÍA DE LA INVESTIGACIÓN.....	32
3.1. Ámbito de estudio	32
3.2. Nivel y diseño de investigación	32
3.3. Población y muestra.....	33
3.4. Procedimientos, técnicas e instrumentos de recolección de datos	34
3.4.1. Procedimientos	34
3.4.2. Técnicas	36
3.4.3. Instrumentos	36
3.5. Análisis de datos	36
3.6. Consideraciones éticas	38
CAPÍTULO IV	39
RESULTADOS Y DISCUSIONES	39

4.1. Presentación de Resultados y Discusiones	39
4.1.1. Variabilidad de las condiciones meteorológicas predominantes en la ciudad Concepción durante el periodo 2023-2025.	39
4.1.2. Variabilidad de las concentraciones del PM _{2.5} en la ciudad de Concepción, 2023- 2025.	48
4.1.3. Análisis de Correlación entre variables meteorológicas y concentraciones de PM _{2.5}	53
4.1.4. Evaluación y Comparación de Modelos de Machine Learning.....	55
4.2. Prueba de Hipótesis.....	61
4.2.1. Hipótesis general	61
4.2.2. Nivel de significancia	61
4.2.3. Prueba estadística aplicada	61
4.2.4. Resultados de la prueba	61
4.2.5. Decisión estadística	62
4.2.6. Interpretación.....	62
CONCLUSIONES.....	63
RECOMENDACIONES	65
REFERENCIAS BIBLIOGRÁFICAS	66
ANEXOS	74

ÍNDICE DE TABLAS

Tabla 1 Operacionalización de variables	17
Tabla 2 Estadísticos descriptivos y comportamiento interanual de los parámetros meteorológicos en la estación de Concepción 2023-2025.	39
Tabla 3 Estadísticos descriptivos y Análisis Comparativo Anual y Total de $\text{PM}_{2.5}$ $\mu\text{g}/\text{m}^3$ en la estación de Concepción 2023-2025.....	48
Tabla 4 Matriz de correlación de Spearman entre variables meteorológicas y concentraciones de $\text{PM}_{2.5}$ en Concepción	53
Tabla 5 Comparación del desempeño estadístico de los modelos predictivos de $\text{PM}_{2.5}$	55
Tabla 6 Comparación de pesos de variables en los modelos evaluados.....	57
Tabla 7 Comparación estadística del desempeño de modelos (Prueba de Wilcoxon)	58

ÍNDICE DE FIGURAS

Figura 1 Material Particulado	25
Figura 2 PurpleAir PA-II / PA-II-SD	28
Figura 3 Davis Vantage Pro2	28
Figura 4 Ubicación de lugar de estudio.....	32
Figura 5 Flujograma de procesos	34
Figura 6 Variación horaria, semanal y mensual de la temperatura en Concepción durante el periodo 2023-2025.	41
Figura 7 Variación horaria, semanal y mensual de la humedad en Concepción durante el periodo 2023-2025.	42
Figura 8 <i>Variación horaria, semanal y mensual de la velocidad de viento en Concepción durante el periodo 2023-2025.</i>	45
Figura 9 <i>Rosa de vientos y distribución de frecuencias de velocidad y dirección en la estación de Concepción (2023-2025).</i>	46
Figura 10 Variación horaria, semanal y mensual de la precipitación en Concepción durante el periodo 2023-2025.....	47
Figura 11 Variación horaria, semanal y mensual del PM _{2.5} en Concepción 2023-2025....	49
Figura 12 Valores diarios de concentración de PM _{2.5} en comparación con los Estándares de Calidad Ambiental (ECA-Aire Perú) y Guías OMS en la estación de Concepción 2023-2025.	51
Figura 13 Comparación visual de valores observados vs. valores modelados de PM _{2.5} para las tres técnicas de Machine Learning.....	60
Figura 14 Estación meteorológica empleada para el registro de variables ambientales	77
Figura 15 Sensor PurpleAir utilizado para el monitoreo continuo de material particulado PM _{2.5} en la zona de estudio	77
Figura 16 Procedimiento de descarga de registros meteorológicos y de PM _{2.5}	78
Figura 17 Visualización y verificación de la cantidad de datos recolectados durante el periodo de monitoreo.....	78
Figura 18 Generación de gráficos para el análisis y procesamiento de datos ambientales	79
Figura 19 Estación de monitoreo meteorológico con consola de adquisición de datos y computadora portátil para visualización y análisis en tiempo real.....	79
Figura 20 Infraestructura donde se encuentran instalados los dispositivos Larson Davis y PurpleAir	80

Figura 21 Juliana Pardo Aiquipa y Estefani Soto Onofre en el punto de instalación de los equipos Larson Davis y PurpleAir	80
--	----

INTRODUCCIÓN

La contaminación atmosférica es uno de los principales problemas ambientales a nivel global debido a sus efectos adversos en la salud humana y el entorno. Entre los contaminantes, el material particulado fino ($PM_{2.5}$) destaca por su alta peligrosidad, ya que puede penetrar profundamente en el sistema respiratorio y provocar enfermedades cardiovasculares y pulmonares, tal como señalan Thangavel et al. (2022) y Hayes et al. (2020). Asimismo, su concentración en la atmósfera está fuertemente influenciada por las condiciones meteorológicas, especialmente en ciudades con baja ventilación y alta estabilidad atmosférica, como evidenciaron Nguyen et al. (2024) y Özüpak et al. (2025).

En el contexto nacional, la contaminación del aire por $PM_{2.5}$ representa una problemática ambiental importante en diversas ciudades del Perú. De acuerdo con Vu et al. (2019), en Lima Metropolitana las concentraciones de este contaminante presentan una estrecha relación con variables meteorológicas y condiciones atmosféricas. Del mismo modo, Olarte et al. (2025) demostraron que, en ciudades de la región Junín y Lima, factores como el crecimiento urbano, el incremento del parque automotor y las condiciones climáticas propias de cada región influyen en la variabilidad del contaminante, generando la necesidad de fortalecer el monitoreo y la gestión de la calidad del aire.

A nivel local, la ciudad de Concepción presenta condiciones meteorológicas y geográficas que favorecen la acumulación de contaminantes atmosféricos durante determinados periodos del año. Factores como la baja velocidad del viento, la estacionalidad de la precipitación y la ocurrencia de inversiones térmicas contribuyen a reducir la dispersión atmosférica, favoreciendo la acumulación de $PM_{2.5}$ y generando episodios críticos de calidad del aire. En ese sentido, Álvarez-Tolentino y Suárez-Salas (2020) evidenciaron que en la ciudad de Concepción predominan vientos de baja velocidad y existen fuentes locales de emisión asociadas a la quema agrícola, el uso de biomasa y las emisiones vehiculares, factores que incrementan la presencia de partículas contaminantes. Asimismo, la ciudad de Concepción forma parte de la dinámica atmosférica del Valle del Mantaro, área vinculada a la Zona de Atención Prioritaria (ZAP) de Huancayo establecida por el Ministerio del Ambiente, debido a la necesidad de fortalecer acciones de vigilancia, monitoreo y gestión de la calidad del aire en esta cuenca atmosférica (MINAM, 2018; MINAM, s. f.). En este escenario, la predicción del $PM_{2.5}$ constituye una herramienta importante para la gestión ambiental y la prevención de riesgos en la salud pública.

La presente investigación tuvo como objetivo evaluar el desempeño de modelos de Machine Learning en la estimación de las concentraciones de $PM_{2.5}$ durante el periodo 2023–2025. Para ello, se analizaron las condiciones meteorológicas, la variabilidad temporal del contaminante y su relación con variables climáticas y temporales. Se emplearon modelos como Regresión Lineal, Random Forest y Redes Neuronales Artificiales, incorporando variables con retardo (lags) para capturar la dependencia temporal del fenómeno, siguiendo enfoques similares desarrollados por Nguyen et al. (2024), Kunjir et al. (2025) y Makhdoomi et al. (2025), quienes demostraron que los modelos de aprendizaje automático permiten mejorar la precisión en la estimación de contaminantes atmosféricos.

Los resultados obtenidos permitieron identificar el modelo con mayor precisión predictiva y aportan evidencia para el desarrollo de sistemas de alerta temprana, contribuyendo a la toma de decisiones orientadas a mejorar la calidad del aire y proteger la salud de la población.

CAPÍTULO I

PLANTEAMIENTO DEL PROBLEMA

1.1. Fundamentación del problema

El material particulado fino (de $PM_{2.5}$) es un contaminante atmosférico compuesto por partículas menores o iguales a 2.5 micras de diámetro (Campaña & Pérez, 2025), capaces de permanecer en suspensión por largos periodos (Joshi et al., 2025). Este contaminante degrada la calidad del aire, reduce la visibilidad, altera los procesos atmosféricos naturales y contribuye al cambio climático al influir en el balance radiativo terrestre (Sánchez, 2022). Su origen puede ser natural, como el polvo o las cenizas volcánicas, o antrópico, como las emisiones vehiculares, la quema de biomasa, actividades industriales y construcción (Campaña & Pérez, 2025), el aumento de sus concentraciones en la atmósfera representa un desafío creciente para la gestión ambiental, especialmente en ciudades donde no existen sistemas adecuados de monitoreo, control ni predicción de este contaminante.

En el mundo, la contaminación debido a la polución del de $PM_{2.5}$ debe ir es una amenaza para la calidad ambiental, ya que, estas partículas ultrafinas pueden permanecer en la atmósfera durante largos periodos, afectar la visibilidad y alterar el equilibrio natural de los ecosistemas (Chowdhury et al., 2023; Jia et al., 2023; Shu et al., 2024), además es un peligro para la salud humana (Thangavel et al., 2022), habiendo encontrado evidencia de que genera afectaciones cardiovasculares (Hayes et al., 2020) y otros. En china, Lin et al. (2018); Mozaffar et al. (2026); Zhang et al. (2022), encontraron que la acumulación de de $PM_{2.5}$, genera la disminución de la visibilidad atmosférica, altera la radiación solar incidente y la temperatura del aire urbano, además se evidencia la reducción de la evapotranspiración. En Estados Unidos, Yin et al. (2025) evidencio que el aumento del de $PM_{2.5}$ se favorece por la dispersión a través del humo de incendios forestales, lo que modifica el balance radiativo al dispersar o absorber la luz solar, generándose una vulnerabilidad de los ecosistemas y el aumento del riesgo de futuros incendios.

En América Latina, los efectos ambientales del de $PM_{2.5}$ son especialmente visibles en áreas urbanas densamente pobladas de algunos países como: En México, según Zamorategui et al. (2024), la contaminación por de $PM_{2.5}$, perturba la radiación solar, altera la formación de nubes, genera reacciones fotoquímicas y reduce la visibilidad ambiental, lo

que conlleva a la variación del clima afectando así los ecosistemas locales. En Colombia, según Correa-Ochoa et al. (2025) la emisión de $PM_{2.5}$, altera procesos atmosféricos evitando la formación de nubes y generando lluvias ácidas, además de la bioacumulación de contaminantes en ecosistemas terrestres y acuáticos que genera pérdida. En Chile, Basoa et al. (2025) menciona que la presencia de $PM_{2.5}$ influye negativamente ya que genera lluvias ácidas y es partícipe de la generación de compuestos secundarios, además provoca la bioacumulación de contaminantes como son Co, Cr, As, V, Fe, Mn y Ni, en plantas, suelo y agua.

En el Perú, si bien las investigaciones ambientales se han centrado principalmente en Lima Metropolitana, diversas ciudades del interior del país presentan condiciones similares de vulnerabilidad ecológica. En Huancayo, Angeles et al. (2024), evidencio que las concentraciones de $PM_{2.5}$ reducen la visibilidad y alteran el balance radiativo al dispersar y absorber la radiación solar, además de influir en la formación de nubes y la precipitación, lo que incrementa la vulnerabilidad climática de la zona; asimismo, estas partículas interfieren con la fotosíntesis y dañan la vegetación, modifican la fracción de luz difusa y, en consecuencia, pueden disminuir la eficiencia del uso de la luz y el rendimiento de los cultivos locales. En Lima, Moretti-Villegas et al. (2023) evidenciaron que la contaminación por $PM_{2.5}$ causan daños al medio ambiente o dañando bienes materiales debido a que pueden existir en el ambiente durante mucho tiempo.

En Concepción, si bien se ha realizado un estudio que muestra el comportamiento del $PM_{2.5}$ en el periodo de 2007 a 2008 por Álvarez-Tolentino & Suárez-Salas (2020) que evidencio los vientos en esta ciudad son predominantemente del noreste y presentan velocidades reducidas entre 1 a 3 m/s, lo que limita la dispersión de las partículas contaminantes, a ello se suma la influencia de fuentes locales, como el uso de leña en la cocción de alimentos y la quema agrícola entre ellos la quema de pastizales y rastrojos en las zonas de cultivo. Pero aún no se han aplicado predicciones mediante modelos predictivos actuales basados en IA, lo que pone en vilo el conocer que herramienta utilizar para obtener datos precisos que ayuden a la toma de decisiones frente a futuros riesgos ambientales, es por ello por lo que se plantea el siguiente problema de investigación:

1.2. Formulación del problema

1.2.1 Problema general

¿Cuál de los modelos de Machine Learning, permite estimar con mayor precisión las concentraciones del material particulado fino de $PM_{2.5}$ en la ciudad de Concepción, durante el periodo 2023-2025?

1.2.2. Problemas específicos

¿Cuál es la variabilidad de las condiciones meteorológicas predominantes en la ciudad de Concepción durante los años 2023-2025?

¿Cómo varían las concentraciones del material particulado fino $PM_{2.5}$ en la ciudad de Concepción durante los años 2023-2025?

¿Cómo validar el desempeño predictivo de las técnicas de Machine Learning en la estimación de las concentraciones de $PM_{2.5}$ en la ciudad de Concepción durante el periodo 2023–2025?

¿Cuál de los modelos de Machine Learning (Random Forest, Regresión y Red Neuronal Artificial) logra un mejor desempeño predictivo en la estimación de las concentraciones de $PM_{2.5}$ en la ciudad de Concepción durante el periodo 2023-2025?

1.3. Objetivos

1.3.1 Objetivo general

Estimar las concentraciones de Material Particulado fino $PM_{2.5}$ mediante modelos de Machine Learning en Concepción durante el periodo 2023-2025.

1.3.2. Objetivos específicos

Determinar la variabilidad de las condiciones meteorológicas predominantes en la ciudad Concepción durante el periodo 2023-2025.

Determinar la variabilidad de las concentraciones del $PM_{2.5}$ en la ciudad de Concepción durante el periodo 2023-2025.

Validar el desempeño predictivo de las Técnicas de Machine Learning en la estimación de las concentraciones de $PM_{2.5}$ en la ciudad de Concepción durante el periodo de 2023- 2025.

Identificar el mejor desempeño predictivo de las Técnicas de Machine Learning en la estimación de las concentraciones de PM_{2.5} en la ciudad de Concepción durante el periodo de 2023- 2025.

1.4. Definición y operacionalización de variables

Tabla 1

Operacionalización de variables

Variables	Definición conceptual	Dimensiones	Indicadores	Unidad de medida
Independiente: Parámetros meteorológicos	Condiciones atmosféricas que favorecen o dificultan la formación, dispersión y transporte del PM _{2.5} .	Temperatura del aire	Promedio horario de temperatura	(°C)
		Velocidad del viento	Promedio horario de velocidad del viento	(m s ⁻¹)
		Dirección del viento	Dirección dominante horaria	° (grados)
		Precipitación	Suma horaria de precipitación	(mm)
		Presión atmosférica	Promedio horario de presión	mmHg
		Humedad Relativa	Promedio horario de humedad relativa	%
Dependiente: PM _{2.5}	Masa de partículas con diámetro aerodinámico ≤ 2.5 µm suspendidas en el aire de Concepción.	Nivel horario	Media horaria de PM _{2.5}	µg m ⁻³
		Nivel diario	Media diaria de PM _{2.5}	
		Nivel mensual	Media mensual de PM _{2.5}	

1.5. Hipótesis

1.5.1. Hipótesis general

El modelo Random Forest presentará el mejor desempeño predictivo en la estimación de las concentraciones de PM_{2.5} en la ciudad de Concepción durante el periodo 2023–2025.

1.6. Justificación

1.6.1 Teórico

Se justifica teóricamente porque aporta a la generación de conocimiento científico sobre el uso de modelos de Machine Learning para la predicción del material particulado fino (PM_{2.5}), un contaminante poco abordado desde enfoques predictivos en zonas altoandinas. La contrastación de modelos como Random Forest, Regresión y Red Neuronal Artificial permitirá fortalecer el sustento técnico para futuras investigaciones en gestión de calidad del aire.

1.6.2 Practica

Se justifica de manera práctica porque los resultados permitirán brindar a las autoridades locales de Concepción una información y una herramienta útil para anticipar escenarios de contaminación atmosférica. Esto facilitará la implementación de medidas preventivas y respuestas oportunas ante incrementos críticos de $PM_{2.5}$, mejorando así la gestión ambiental urbana.

1.6.3 Ambiental

Se justifica ambientalmente porque el estudio propone una solución tecnológica ante un problema real que afecta la calidad del aire en la ciudad de Concepción. Además, al formar parte de una Zona de Atención Prioritaria (ZAP) por contaminación atmosférica, se evidencia la necesidad de fortalecer el monitoreo y la gestión de la calidad del aire. En ese sentido, la predicción de $PM_{2.5}$ mediante modelos de predicción contribuyen a mitigar los impactos ambientales negativos derivados de fuentes antrópicas y condiciones meteorológicas desfavorables.

1.6.4 Metodológico

Se justifica metodológicamente al aplicar y comparar modelos de aprendizaje automático en un contexto con escasa infraestructura de monitoreo. Esta propuesta permitirá validar la eficacia de estas técnicas en zonas con características geográficas y climáticas complejas, y sentar las bases para su réplica en otras ciudades interandinas del país.

1.6.5 Social

Se justifica socialmente porque la población de Concepción se ve directamente afectada por los niveles elevados de $PM_{2.5}$, un contaminante que impacta la salud respiratoria y cardiovascular, especialmente en niños, adultos mayores y personas con enfermedades preexistentes. El estudio busca contribuir a la protección de la salud pública al ofrecer herramientas de predicción que permitan a la ciudadanía acceder a información anticipada sobre la calidad del aire. Esto fortalece la conciencia ambiental, promueve hábitos de prevención y genera una cultura de corresponsabilidad en el cuidado del entorno.

1.6.6 Economia

Se justifica económicamente porque la exposición a altos niveles de $PM_{2.5}$ genera costos significativos en atención médica, ausentismo laboral y pérdida de productividad. Contar con modelos predictivos permitirá a las autoridades y a los sectores productivos anticipar episodios críticos de contaminación, optimizando la asignación de recursos para la

prevención y mitigación. Asimismo, una gestión más eficiente de la calidad del aire puede reducir gastos públicos en salud y contribuir al desarrollo sostenible de la región, equilibrando crecimiento económico y bienestar poblacional.

CAPÍTULO II

MARCO TEÓRICO

2.1. Antecedentes

2.1.1 Internacionales

Nguyen et al. (2024) en su artículo titulado «*Development of Machine Learning and Deep Learning Prediction Models for PM_{2.5} in Ho Chi Minh City, Vietnam*». El estudio buscó desarrollar y comparar seis algoritmos (random forest, XGBoost, SVR, artificial neural network – ANN, general regression neural network – GRNN y convolutional neural network – CNN) para predecir las concentraciones diarias de PM_{2.5} en Ciudad Ho Chi Minh. Con enfoque cuantitativo y un diseño no experimental correlacional-predictivo, se procesaron 911 observaciones (1-ene-2021 → 30-jun-2023) de PM_{2.5} de la estación del Consulado de EE. UU., junto con temperatura, humedad, viento, lluvia, horas de sol y evaporación registradas en Tan Son Hoa; los datos se depuraron, se evaluó la importancia de los predictores y se dividieron en 80 % entrenamiento y 20 % prueba, aplicando validación interna adicional para ANN y CNN. La ANN obtuvo el mejor desempeño (índice de acuerdo = 0,736; RMSE = 7,98 $\mu\text{g m}^{-3}$), por lo que se considera una herramienta prometedora para respaldar la gestión de la calidad del aire en la ciudad.

Makhdoomi et al. (2025) en su artículo titulado «*PM_{2.5} concentration prediction using machine learning algorithms: an approach to virtual monitoring stations*». El estudio comparó LightGBM, Gradient Boosting Regressor (GBR), XGBoost Regressor y Random Forest para predecir las concentraciones diarias de PM_{2.5} registradas entre 2016 y 2022, con el fin de habilitar estaciones virtuales en Mashhad (Irán). Con enfoque cuantitativo y diseño no experimental correlacional-predictivo, se depuraron 2 203 registros combinados de PM_{2.5}/PM₁₀ y variables meteorológicas; los atípicos se eliminaron mediante el criterio IQR. El conjunto se dividió 80 % entrenamiento y 20 % prueba, y los hiperparámetros se optimizaron por grid-search con validación cruzada de diez pliegues. El desempeño se evaluó mediante MAE, MSE, RMSE, MAPE y R², contrastando los modelos con las pruebas de Friedman y Wilcoxon. El GBR obtuvo los mejores MSE (5,33) y RMSE (2,31), mientras que Random Forest presentó el menor MAE (0,47) y MAPE (1,5); más del 95 % de las predicciones quedaron dentro del rango observado. Con una media anual de 32,84 $\mu\text{g m}^{-3}$ y picos de 160,25 $\mu\text{g m}^{-3}$, los resultados confirman que las técnicas de aprendizaje automático

permiten implementar estaciones virtuales útiles para mejorar la evaluación epidemiológica y la gestión de la calidad del aire.

Kunjir et al. (2025) en su artículo titulado "*Assessing particulate matter ($PM_{2.5}$) concentrations and variability across Maharashtra using satellite data and machine learning techniques*", cuyo objetivo fue estimar y analizar las concentraciones de $PM_{2.5}$ en Maharashtra, India, durante 2023, aplicando modelos de aprendizaje automático para mejorar la precisión espacial y temporal e identificar variaciones estacionales y zonas críticas de contaminación. El estudio adoptó un enfoque cuantitativo, con diseño no experimental, descriptivo y correlacional, empleando cuatro modelos de regresión (Random Forest, regresión múltiple, regresión lineal y Lasso) a partir de datos de 90 estaciones de la Central Pollution Control Board, mediciones satelitales de Fine Aerosol Optical Depth (FAOD) del sensor MODIS y parámetros meteorológicos como humedad relativa, temperatura y componentes del viento. El modelo Random Forest presentó el mejor desempeño ($R^2=0.87$, $RMSE=12.57 \mu g/m^3$, $MAE=6.96 \mu g/m^3$). Las concentraciones anuales promedio alcanzaron $46.22 \mu g/m^3$, superando el límite nacional, con picos en invierno ($75.02 \mu g/m^3$) y mínimos en monzón ($22.38 \mu g/m^3$). Se concluyó que combinar FAOD y datos meteorológicos optimiza la predicción y que las áreas urbanas requieren acciones específicas para reducir la exposición y proteger la salud pública.

Shakya et al. (2023) en su artículo titulado " *$PM_{2.5}$ air pollution prediction through deep learning using meteorological, vehicular, and emission data: A case study of New Delhi, India*", cuyo objetivo fue diseñar y evaluar modelos de aprendizaje profundo para predecir las concentraciones de $PM_{2.5}$ en Delhi, India, con el propósito de mejorar la evaluación de la calidad del aire y orientar estrategias de gestión ambiental. La investigación adoptó un enfoque cuantitativo, con diseño no experimental y alcance predictivo-explicativo. Se emplearon datos históricos de $PM_{2.5}$ de la Central Pollution Control Board, variables meteorológicas (temperatura, humedad relativa, velocidad y dirección del viento) y factores temporales (hora, día y mes), usando como muestra registros horarios de varios años que permitieron capturar patrones diarios y estacionales. Se probaron redes neuronales LSTM, GRU y CNN-LSTM, siendo esta última la de mejor rendimiento ($R^2=0.96$, $RMSE=6.57 \mu g/m^3$, $MAE=4.34 \mu g/m^3$). Los hallazgos demostraron que la inclusión de variables meteorológicas y temporales mejora la precisión predictiva. Se concluyó que las arquitecturas híbridas de aprendizaje profundo son herramientas eficaces para la predicción

de contaminantes y la gestión proactiva de la calidad del aire en contextos urbanos altamente contaminados.

Rad et al. (2024) en su artículo titulado *"Machine learning models for predicting interactions between air pollutants in Tehran Megacity, Iran"* tuvieron como objetivo analizar las interacciones entre seis contaminantes atmosféricos (CO, O₃, NO₂, SO₂, PM₁₀ y PM_{2.5}) en Teherán a lo largo de diez años, simulando escenarios de incremento o reducción del 50 % en cada uno, sin considerar factores ambientales externos. El estudio, de enfoque cuantitativo, diseño no experimental y alcance correlacional, recopiló datos de 16 estaciones de monitoreo entre 2013 y 2022, aplicando correlaciones de Pearson y Spearman, así como modelos de aprendizaje automático Random Forest (RF), XGBoost (XGB) y LightGBM (LGBM). Se utilizaron equipos especializados como BAM-1020 y analizadores Ecotech, obteniendo 3 653 registros diarios por contaminante. Los hallazgos revelaron correlaciones como la negativa entre CO y O₃ ($r = -0.20$; $p < 0.001$) y la alta asociación positiva entre PM₁₀ y PM_{2.5} ($r = 0.85$; $p < 0.001$). El modelo RF mostró el mejor desempeño para contaminantes gaseosos (CO: $R^2 = 0.64$), LGBM para PM_{2.5} ($R^2 = 0.88$) y XGB para PM₁₀ ($R^2 = 0.88$). Las simulaciones evidenciaron que la variación de un contaminante suele afectar a los demás, concluyendo que las estrategias de control dirigidas a uno solo pueden influir de forma significativa en la calidad del aire en general, lo que refuerza la importancia de considerar las interacciones químicas en la gestión ambiental.

2.1.2 Nacionales

Olarte et al. (2025) en su artículo titulado *«Predictive Model with Machine Learning for Environmental Variables and PM_{2.5} in Huachac, Junín, Perú»*. El estudio tuvo como objetivo modelar el índice de calidad del aire para PM_{2.5} a partir de temperatura y humedad, comparando regresión lineal, ridge, lasso y elastic-net. Con un enfoque cuantitativo y un diseño no experimental correlacional-predictivo, se procesaron 16 543 mediciones de la estación HUAYAO, divididas 80 % entrenamiento y 20 % prueba; los datos provinieron de sensores de PM_{2.5}, temperatura (°C) y humedad (%), evaluados con validación cruzada. El mejor ajuste lasso-elastic-net con $\alpha \approx 0,11$ y $\lambda \approx 0,15$ alcanzó un RMSE de 25,4 $\mu\text{g} \cdot \text{m}^{-3}$ y un R^2 ajustado de 0,15; los mayores valores de PM_{2.5} se registraron en invierno ($\approx 52,6 \mu\text{g} \cdot \text{m}^{-3}$). Se concluyó que temperatura y humedad se relacionan inversamente con PM_{2.5} y que la ecuación resultante constituye una herramienta útil para la gestión ambiental del valle del Mantaro.

Oblitas & Escobedo (2024) en su artículo titulado «*Prediction of $PM_{2.5}$ and PM_{10} Concentrations Using XGBoost and LightGBM Algorithms: A Case Study in Lima, Peru*». Su objetivo fue comparar ambos algoritmos para predecir concentraciones horarias de $PM_{2.5}$ y PM_{10} usando dos conjuntos de variables. Investigación cuantitativa, no experimental correlacional-predictiva; empleó 2 810 registros de contaminantes y 2 841 meteorológicos (enero-abril 2024, estación SENAMHI-San Borja), depurados con LOF y normalizados por z-score, con partición 80 / 20 y ajuste de hiperparámetros vía RandomizedSearchCV. Con todas las variables, XGBoost y LightGBM alcanzaron R^2 de 0,75 ($PM_{2.5}$) y 0,84 (PM_{10}); al excluir la variable cruzada descendieron a 0,54 y 0,69. Solo para PM_{10} en el segundo escenario hubo diferencia significativa ($t = -3,04$; $p = 0,0025$). Se concluye que incluir simultáneamente $PM_{2.5}$ y PM_{10} mejora la exactitud, con ligera ventaja de XGBoost como herramienta para la gestión de la calidad del aire limeña.

Vu et al. (2019) en su artículo titulado «*Developing an Advanced $PM_{2.5}$ Exposure Model in Lima, Peru*». El propósito fue estimar diariamente $PM_{2.5}$ a 1 km² durante 2010-2016 mediante un modelo Random Forest que integra AOD-MAIAC, meteorología, salidas de WRF-Chem y variables de uso del suelo. Con enfoque cuantitativo y diseño no experimental correlacional-predictivo, se usaron 8 470 observaciones de 16 estaciones (10 SENAMHI y 6 JHU) para entrenar un bosque de 1 000 árboles, evaluado con validación cruzada de diez pliegues. El modelo obtuvo $R^2 = 0,70$ y $RMSE = 5,97 \mu g \cdot m^{-3}$, con sesgo medio de $-0,09 \mu g \cdot m^{-3}$ y 94,5 % de predicciones dentro de ± 2 DE. Radiación solar, temperatura, humedad relativa, altura de la capa límite y AOD resultaron predictores clave; se observaron menores concentraciones en la costa ($\sim 14,6 \mu g \cdot m^{-3}$) y mayores en la vertiente andina ($\sim 48,5 \mu g \cdot m^{-3}$). El modelo proporciona insumos valiosos para estudios epidemiológicos y políticas de mitigación en Lima.

Choquetico & Silva (2025) en su tesis titulada «*Modelo de regresión funcional para la predicción de $PM_{2.5}$ en función de PM_{10} en Lima – Perú*». Su objetivo fue estimar niveles diarios de $PM_{2.5}$ (2021-2023) mediante regresión funcional concurrente basada en curvas B-Spline de PM_{10} y otros contaminantes. Investigación cuantitativa, no experimental y correlacional-predictiva; se analizaron series de diez estaciones SENAMHI-WAQI, depurando 348 variables y reteniendo cinco predictoras claves. El modelo $PM_{2.5}(t) = \beta_0(t) + \beta_1(t) PM_{10}(t) + \varepsilon(t)$ se ajustó tras transformar los 28 primeros días de cada mes; $\beta_1(t)$ varió entre 0 y $>1,5$, indicando asociación fluctuante. En 2023 los promedios de $PM_{2.5}$ fueron $\approx 84 \mu g \cdot m^{-3}$ (Carabayllo) y $79 \mu g \cdot m^{-3}$ (SJL), mientras PM_{10}

promedió $\approx 44 \mu\text{g}\cdot\text{m}^{-3}$ en San Borja; el intercepto funcional superó $100 \mu\text{g}\cdot\text{m}^{-3}$ a fin de mes, sugiriendo fuentes externas. Se concluye que $\text{PM}_{2.5}$ no depende sólo de PM_{10} y que la regresión funcional permite identificar influencias externas críticas para la salud pública.

2.1.3. Local

Álvarez-Tolentino y Suárez-Salas (2020), en su artículo “Aporte cuantitativo de las fuentes de PM_{10} y $\text{PM}_{2.5}$ en sitios urbanos del Valle del Mantaro, Perú”, tuvieron como objetivo determinar y cuantificar las principales fuentes emisoras de PM_{10} y $\text{PM}_{2.5}$ en Jauja, Concepción y Huancayo. El estudio fue de enfoque cuantitativo, diseño no experimental y alcance descriptivo-correlacional, desarrollado entre julio de 2007 y octubre de 2008 mediante monitoreo mensual con un muestreador Partisol FRM 2000 y análisis químico por fluorescencia de rayos X. Los resultados mostraron que Huancayo registró las mayores concentraciones de PM_{10} y $\text{PM}_{2.5}$, superando en algunos periodos los estándares de calidad ambiental. Asimismo, se identificaron como principales fuentes el polvo del suelo, la quema de biomasa, las emisiones vehiculares, combustible-aceite y fundición. Se concluyó que la presencia de PM_{10} y $\text{PM}_{2.5}$ representa un riesgo para la salud y está asociada principalmente a actividades antrópicas y rurales, recomendándose medidas de control para mejorar la calidad del aire.

2.2. Bases teóricas

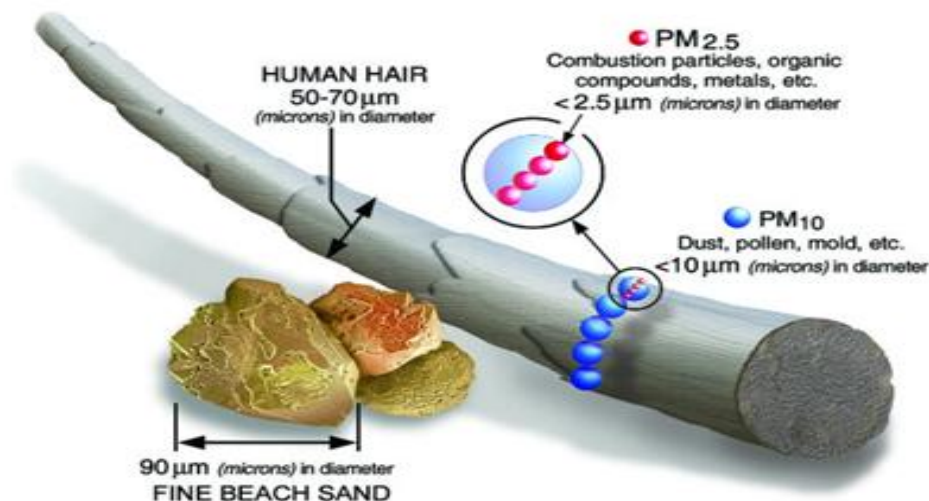
2.2.1. Material particulado fino (PM_{2.5})

Según Carranza (2004) el material particulado (PM) es una mezcla de partículas sólidas y líquidas suspendidas en el aire.

Entre sus características están (Quijano et al., 2010):

- Su reducido diámetro aerodinámico ($\leq 2,5 \mu\text{m}$), lo que le permite permanecer suspendido en la atmósfera por largos periodos y ser respirable en un 100 %.
- Forma parte de la fracción más pequeña y agresiva del material particulado, compuesta generalmente por partículas ácidas, y puede presentar propiedades higroscópicas, es decir, capacidad para absorber y retener humedad del aire.

Figura 1
Material Particulado



Fuente: (U.S. Environmental Protection Agency [EPA], 2023)

El PM_{2.5} contiene iones como sulfatos (SO_4^{2-}), nitratos (NO_3^-) y amonio (NH_4^+), además de metales como plomo (Pb), cadmio (Cd) y zinc (Zn). Estos provienen de procesos industriales, emisiones vehiculares y fuentes naturales, y afectan las propiedades químicas y ópticas del aerosol (Chowdhury et al., 2023).

En cuanto a los compuestos orgánicos, incluye hidrocarburos aromáticos policíclicos y carbono orgánico (OC), junto con carbono elemental (EC) o negro de carbono, generado por combustión incompleta (Chowdhury et al., 2023).

Las fuentes naturales, comprenden fenómenos y procesos que liberan partículas finas a la atmósfera sin intervención directa humana. Entre ellas destacan el polvo mineral arrastrado por el viento, las emisiones de cenizas volcánicas, las tormentas de polvo, la sal marina proveniente del rompimiento de olas y la quema natural de biomasa durante incendios forestales. (Bimenyimana et al., 2025; Correa-Ochoa et al., 2025; Jia et al., 2023). Estos procesos aportan material particulado con composición y tamaño variables, influyendo en la química atmosférica y el balance radiativo terrestre.

Las fuentes antrópicas, derivan de actividades humanas que generan emisiones directas o promueven la formación secundaria de partículas. Entre las más relevantes se incluyen la combustión de combustibles fósiles en centrales termoeléctricas, industrias y vehículos; procesos industriales y de construcción; quema agrícola y forestal controlada; así como el uso de biomasa para calefacción y cocción. (Campaña Tene & Pérez Fernández, 2025; Correa-Ochoa et al., 2025; Jia et al., 2023). Estas fuentes son responsables de una fracción significativa de la contaminación atmosférica urbana y rural, asociada a impactos en la salud y el ambiente.

2.2.2. Impactos del $PM_{2.5}$

El material particulado fino ($PM_{2.5}$) representa un riesgo significativo para la salud debido a su capacidad de penetrar profundamente en el sistema respiratorio y atravesar la barrera alveolar, alcanzando el torrente sanguíneo (Rad et al., 2024). Una vez en el organismo, estas partículas pueden provocar inflamación sistémica y estrés oxidativo, procesos que se han vinculado con enfermedades respiratorias como asma y bronquitis crónica, así como con trastornos cardiovasculares como hipertensión y cardiopatía isquémica (Shu et al., 2024). La evidencia epidemiológica confirma que la exposición prolongada al $PM_{2.5}$ incrementa el riesgo de mortalidad prematura y agrava la carga global de enfermedad, afectando tanto a poblaciones urbanas como rurales (Zhang et al., 2022).

En el medio ambiente, el $PM_{2.5}$ afecta el balance radiativo terrestre al dispersar y absorber radiación solar, modificando los patrones de calentamiento y enfriamiento atmosférico (Rad et al., 2024), también su comportamiento óptico reduce la visibilidad y favorece la formación de neblina fotoquímica, afectando de manera directa la calidad visual del paisaje (Zhang et al., 2022), además, la deposición de partículas sobre las plantas disminuye la eficiencia fotosintética, mientras que en ecosistemas acuáticos altera el pH y la disponibilidad de nutrientes, generando desequilibrios que comprometen la salud de los ecosistemas (Shu et al., 2024).

2.2.3. Factores meteorológicos que influyen en las concentraciones de $PM_{2.5}$

La temperatura y la humedad relativa (HR) influyen directamente en las concentraciones de $PM_{2.5}$. Un aumento de la HR puede favorecer el crecimiento higroscópico de las partículas, lo que incrementa su tamaño y promueve su sedimentación, contribuyendo así a la reducción de contaminantes en el aire. En contraste, temperaturas bajas, especialmente en invierno, pueden favorecer la formación de inversiones térmicas que limitan la mezcla vertical y favorecen la acumulación de contaminantes en las capas bajas de la atmósfera (Nakyai et al., 2025).

La velocidad y dirección del viento determinan la dispersión, transporte y acumulación de $PM_{2.5}$. Vientos predominantes en áreas industriales facilitan la dispersión de contaminantes hacia zonas menos densamente pobladas, mientras que en entornos urbanos con atmósferas más estables puede ocurrir una acumulación de partículas. Velocidades de viento más altas incrementan la capacidad de mezcla y dispersión atmosférica, reduciendo las concentraciones de $PM_{2.5}$, aunque su efectividad depende de la estabilidad atmosférica local (Nakyai et al., 2025).

La precipitación actúa como un mecanismo de lavado atmosférico (“washout”), eliminando $PM_{2.5}$ y otros contaminantes a través de la captación y arrastre en las gotas de lluvia. Por otro lado, la presión atmosférica influye en la estabilidad del aire: presiones bajas favorecen la mezcla y dispersión, mientras que presiones altas están asociadas con acumulación de contaminantes, en gran parte debido a la formación de inversiones térmicas. En zonas industriales, se ha identificado una correlación positiva entre la presión atmosférica y las concentraciones de $PM_{2.5}$ (Nakyai et al., 2025).

2.2.4. Monitoreo y medición de $PM_{2.5}$

Los métodos directos de medición de $PM_{2.5}$ se basan en el uso de equipos de referencia que ofrecen alta precisión y confiabilidad en la cuantificación de partículas en suspensión. Entre los más utilizados se encuentran el Beta Attenuation Monitor (BAM), el Tapered Element Oscillating Microbalance (TEOM) y el sistema Partisol, los cuales operan mediante la recolección y medición directa del material particulado bajo principios físicos estandarizados (Bonilla-Granados et al., 2023).

En los últimos años, se ha incrementado el uso de sensores de bajo costo para ampliar la cobertura del monitoreo atmosférico. Estos dispositivos operan bajo principios ópticos de dispersión de luz, en los que un haz incide sobre las partículas y el patrón de dispersión se

convierte en una estimación de la concentración de $PM_{2.5}$ (Awos et al., 2024). Entre estos, destaca el PurpleAir, que será empleado en esta investigación junto con una estación meteorológica Davis Vantage Pro instalada en la ciudad de Concepción. Este sensor ofrece mediciones continuas de alta resolución temporal, lo que permite generar series de datos consistentes para el análisis y modelado predictivo, aunque su precisión puede verse afectada por factores ambientales como humedad relativa elevada o cambios bruscos de temperatura, por lo que es recomendable realizar procesos de calibración y validación.

Figura 2

PurpleAir PA-II / PA-II-SD



Fuente: EOAS, Florida State University

Figura 3

Davis Vantage Pro2



Fuente: (Ubuy, 2025)

Los métodos indirectos de estimación de $PM_{2.5}$ se basan en la utilización de variables y mediciones asociadas obtenidas de manera remota o a partir de datos auxiliares, que posteriormente son transformadas mediante modelos para inferir la concentración de partículas en la superficie.

Entre las técnicas más empleadas destacan las observaciones satelitales de la profundidad óptica de aerosoles (Aerosol Optical Depth, AOD), las cuales permiten estimar la cantidad de partículas presentes en una columna atmosférica (Awos et al., 2024).. Productos como MODIS (Moderate Resolution Imaging Spectroradiometer) y el algoritmo MAIAC (Multi-Angle Implementation of Atmospheric Correction) facilitan la derivación de concentraciones superficiales de $PM_{2.5}$ al combinar datos de AOD con información de estaciones terrestres.

Además de las mediciones satelitales, los métodos indirectos incluyen el uso de modelos de transporte y dispersión atmosférica, que integran variables meteorológicas (temperatura, humedad relativa, velocidad y dirección del viento, presión atmosférica) y datos de emisiones para estimar la distribución espacial y temporal de contaminantes.

La precisión de estas estimaciones puede mejorarse mediante la integración de datos satelitales con parámetros de uso y cobertura del suelo, variables topográficas y registros meteorológicos locales, lo que permite capturar variaciones estacionales y micro climáticas. Esta combinación ha demostrado aumentar la resolución espacial y la exactitud de los valores estimados de $PM_{2.5}$, proporcionando una representación más fiel de las concentraciones reales y facilitando el desarrollo de modelos predictivos para la gestión de la calidad del aire (Bonilla-Granados et al., 2023).

2.2.5. Monitoreo y medición de datos meteorológicos

El monitoreo de variables meteorológicas constituye un elemento esencial para comprender y predecir las dinámicas de dispersión, transformación y acumulación de contaminantes atmosféricos. La medición continua de parámetros como temperatura, humedad relativa, velocidad y dirección del viento, presión atmosférica y precipitación permite establecer relaciones directas e indirectas con las concentraciones de $PM_{2.5}$ y otros contaminantes.

En esta investigación, las variables meteorológicas se registrarán mediante una estación Davis Vantage Pro instalada en la ciudad de Concepción. Este sistema,

ampliamente utilizado en estudios de calidad del aire, proporciona mediciones de alta resolución temporal y precisión gracias a sus sensores especializados:

- **Temperatura y humedad relativa:** mediante un sensor termo-higrómetro protegido contra radiación solar.
- **Velocidad y dirección del viento:** a través de un anemómetro y veleta de alta sensibilidad.
- **Presión atmosférica:** captada mediante un barómetro digital integrado.
- **Precipitación:** medida con un pluviómetro de balancín.

La estación almacena y transmite los datos en intervalos configurables, exportables en formatos como CSV o Excel, lo que facilita su procesamiento estadístico y modelado predictivo. Esta infraestructura garantiza la generación de series temporales consistentes y comparables, optimizando la calibración y validación de los modelos de Machine Learning propuestos.

2.2.6. Modelos relevantes para series temporales ambientales

El modelo Random Forest es un algoritmo de aprendizaje supervisado que combina múltiples árboles de decisión para generar predicciones más precisas, utilizando el voto mayoritario en clasificación o el promedio en regresión. Su principal fortaleza radica en la capacidad de manejar grandes volúmenes de datos, tolerar variables correlacionadas y minimizar el riesgo de sobreajuste, lo que lo hace adecuado para problemas complejos como el modelado de series temporales ambientales. No obstante, su uso presenta limitaciones, entre ellas una menor interpretabilidad en comparación con modelos más simples y un mayor tiempo de entrenamiento cuando se aplican grandes bases de datos (Vásquez, 2022).

La regresión es una técnica estadística utilizada para modelar la relación que hay entre una variable dependiente y un conjunto de variables independientes, permitiendo estimar y predecir valores futuros a partir de datos históricos. En lo que se refiere a regresión, encontramos una variedad según lo que se desea analizar y encontrar, en particular, la regresión de cresta (*Ridge Regression*) se aplica cuando las variables predictoras presentan alta correlación, ya que incorpora un parámetro de regularización (λ) que penaliza el tamaño de los coeficientes, reduciendo la varianza y controlando la multicolinealidad y el sobreajuste. Este método no estima los coeficientes mediante mínimos cuadrados ordinarios, sino a través de un estimador de cresta que mejora la estabilidad del modelo y su capacidad predictiva, pero se debe tener en cuenta que su rendimiento depende de una adecuada

calibración del parámetro λ , pues valores inapropiados pueden disminuir la precisión del modelo, y al no realizar selección de variables, todas permanecen en la ecuación, lo que puede limitar su interpretabilidad (Contento & Ortiz, 2024).

Las Redes Neuronales Artificiales (ANN, por sus siglas en inglés) son modelos computacionales que funcionan de manera similar que un cerebro humano, capaz de aprender patrones complejos a partir de datos, el cual esta compuesta por capas de nodos interconectados neuronas que transforman las entradas en salidas mediante conexiones ponderadas y funciones de activación, ajustando estos pesos durante el entrenamiento para minimizar el error de predicción, comúnmente a través de algoritmos como *backpropagation* y descenso de gradiente. Su capacidad para modelar relaciones no lineales, han sido empleadas en la predicción de la calidad del aire, integrando variables meteorológicas, datos de emisiones y concentraciones de contaminantes. No obstante, su eficacia depende de contar con grandes volúmenes de datos y de una correcta regularización para evitar el sobreajuste, así como de afrontar la limitación de su carácter de “caja negra”, que dificulta la interpretación de los resultados (Zhou et al., 2024).

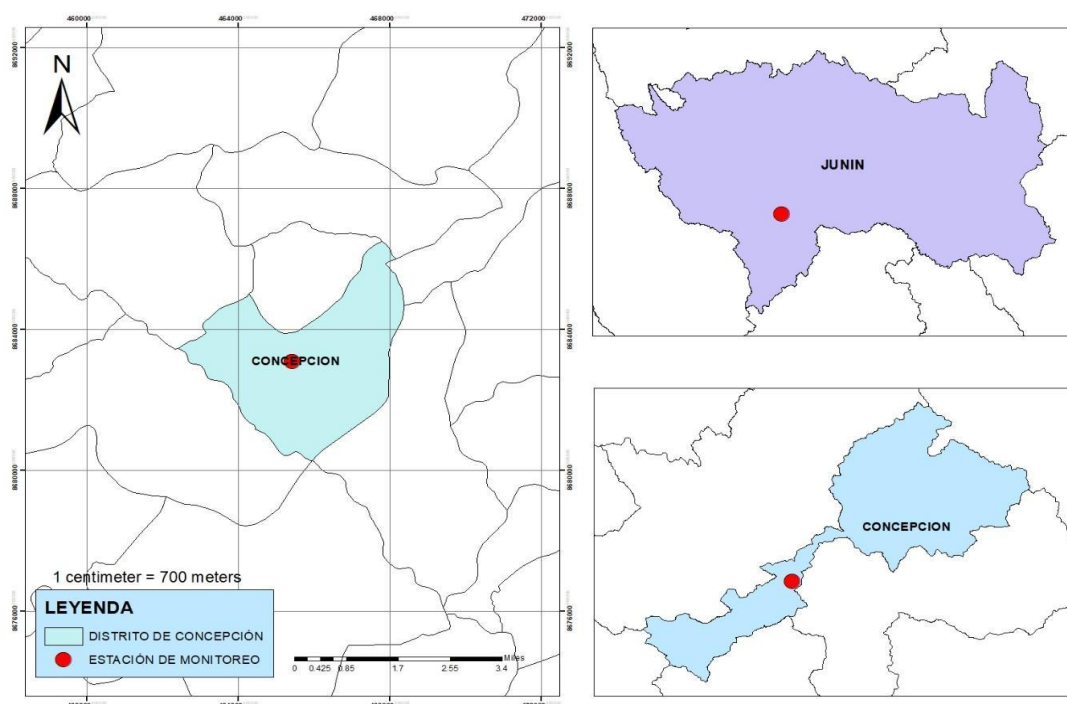
CAPÍTULO III

METODOLOGÍA DE LA INVESTIGACIÓN

3.1. Ámbito de estudio

El estudio se llevó a cabo en la ciudad de Concepción, la cual se encuentra ubicada en la región Junín, Perú, a una altitud aproximada de 3 300 m s. n. m., la ciudad se caracterizó por tener un clima templado subhúmedo de montaña. Asimismo, este lugar fue seleccionado debido al aumento de emisión de $PM_{2.5}$, productos de la quema de biomasa y actividades agrícolas.

Figura 4
Ubicación de lugar de estudio.



3.2. Nivel y diseño de investigación

La presente investigación fue de nivel explicativo-predictivo, ya que buscó identificar relaciones entre variables meteorológicas y las concentraciones de $PM_{2.5}$, y describir el comportamiento mediante modelos de predicción basados en inteligencia artificial. De acuerdo con Hernández et al. (2014), el nivel explicativo tuvo como propósito identificar las causas de los fenómenos y a través de ello establecer relaciones de causalidad,

mientras que el enfoque predictivo se orienta a anticipar resultados futuros a partir de patrones.

El tipo de investigación fue aplicada, debido a que buscó generar soluciones prácticas mediante el uso de modelos de Machine Learning para estimar los niveles de contaminación del aire y así contribuir a la toma de decisiones ambientales en la ciudad de Concepción. Según Hernández et al. (2014), las investigaciones de tipo aplicada tuvieron como finalidad resolver problemas concretos y generar conocimiento útil para la práctica.

En cuanto al diseño fue no experimental, debido a que no se manipularon directamente las variables, sino que se analizaron los registros existentes; además, presento un carácter longitudinal- predictivo, ya que se emplearon datos recopilados a lo largo de más de 2 años, correspondiente a 2023- 2025 para construir y validar modelos como Random Forest, Regresión y Redes Neuronales Artificiales(ANN), lo cuales permitieron estimar y predecir los niveles de $PM_{2.5}$ con base en parámetros atmosféricos. Según Hernández et al. (2014), el diseño no experimental implicó observar los fenómenos tal como ocurrieron en su contexto natural, sin intervención del investigador, mientras que el diseño longitudinal permitió analizar los cambios o comportamientos a lo largo del tiempo.

3.3. Población y muestra

La población de estudio estuvo conformada por los datos del material particulado fino ($PM_{2.5}$) de la ciudad de Concepción, así como por las variables meteorológicas asociadas (temperatura, humedad, velocidad y dirección del viento, presión atmosférica y precipitación) de los periodos comprendido entre 2023 - 2025.

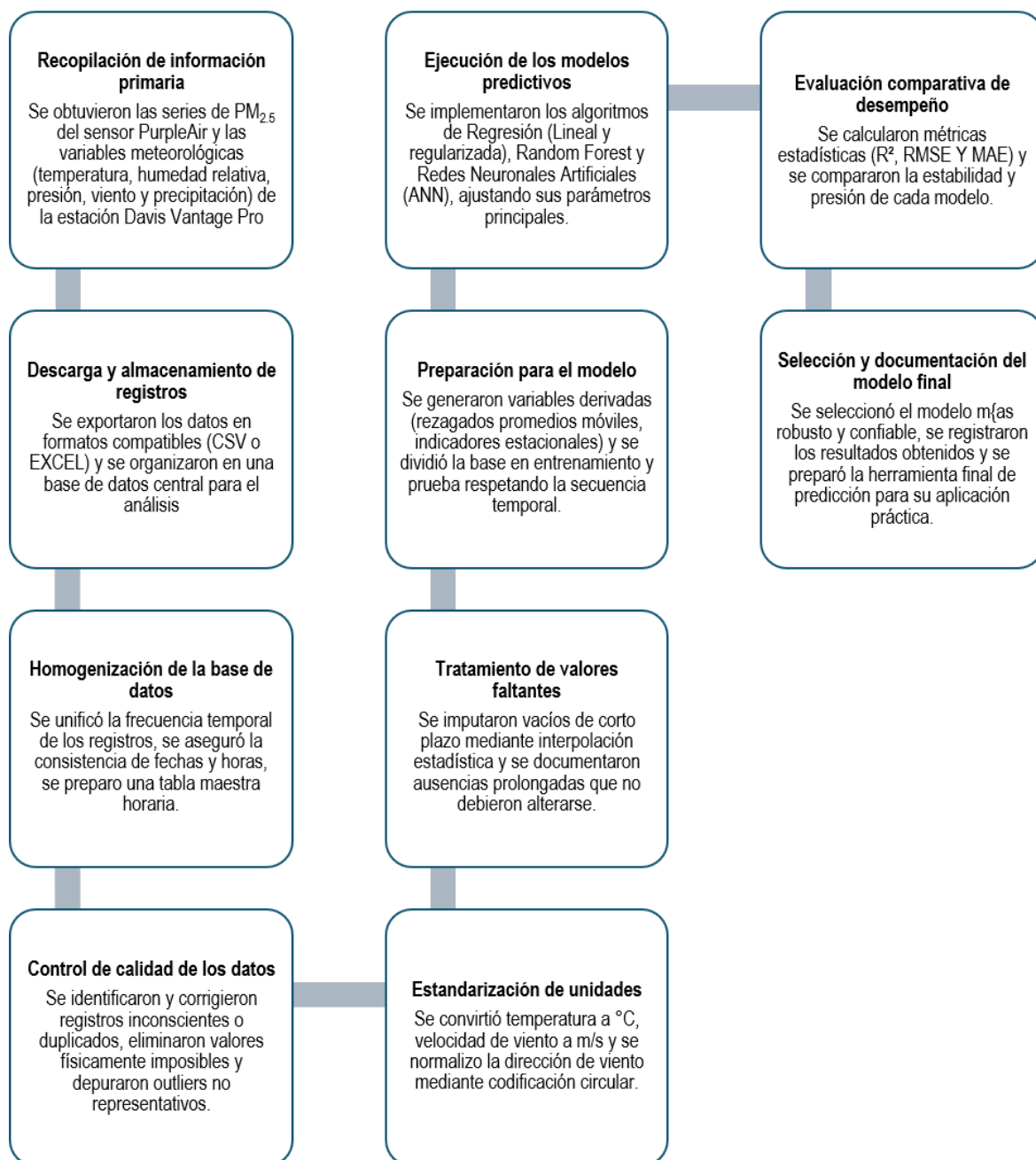
La muestra se estuvo constituida por los datos horarios, diarios y mensuales de las concentraciones de $PM_{2.5}$ y de las variables meteorológicas recopiladas por una estación meteorológica instalada en la ciudad de Concepción, equipada con un sistema Davis Vantage Pro y un sensor PurpleAir para la medición del material particulado fino. La selección de la muestra fue de tipo no probabilística e intencionada, considerando únicamente los días y meses que presentaron información continua, válida y completa, necesaria para el entrenamiento, evaluación y comparación de los modelos de Machine Learning propuestos en la investigación.

3.4. Procedimientos, técnicas e instrumentos de recolección de datos

3.4.1. Procedimientos

Figura 5

Flujograma de procesos



Fuente: Elaboración propia (2025)

En la primera fase, se llevó a cabo la recopilación de datos, la cual consistió en recopilar series temporales horarias y diarias de concentraciones de $PM_{2.5}$ del sensor PurpleAir, también las variables meteorológicas asociadas (temperatura, humedad relativa, velocidad y dirección del viento, presión atmosférica y precipitación) brindadas por la estación meteorológica instalada en la ciudad de Concepción, equipada con un sistema Davis Vantage Pro. Posteriormente, los registros obtenidos fueron descargados y almacenados en formatos compatibles, como CSV y Excel, con la finalidad de facilitar su organización, procesamiento y análisis dentro de la investigación.

En la segunda fase, se llevó a cabo la limpieza y validación de los datos. Este proceso incluyó la identificación y eliminación de valores atípicos y registros incompletos, así como la conversión de unidades para unificar criterios; por ejemplo, la temperatura fue expresada en °C y la velocidad del viento en m/s. Asimismo, se realizó la imputación de datos faltantes mediante tres métodos: la imputación simple (Simple Value Imputation), MICE (Multivariate Imputation by Chained Equations) y MissForest, con el fin de mantener la continuidad y coherencia de las series temporales.

En la tercera fase se llevó a cabo la ejecución de los modelados, la cual abarcó la implementación de tres algoritmos de predicción:

- **Random Forest (RF):** los datos fueron divididos en un conjunto de entrenamiento (70%) y prueba (30%), ajustando los hiperparámetros como el número de árboles (*n_estimators*), la profundidad máxima (*max_depth*) y el número de variables por división (*max_features*). Asimismo, se aplicó validación cruzada *k*-fold (*k*=10) para evaluar la estabilidad y robustez del modelo.
- **Regresión:** se evaluó mediante enfoques (lineal y regularizada), considerando la multicolinealidad entre las variables predictoras y el ajuste global del modelo. Además, los parámetros fueron optimizados mediante validación cruzada para mejorar la capacidad predictiva y evitar el sobreajuste.
- **Red Neuronal Artificial (ANN):** Se configuró una arquitectura especializada para series temporales, defendiéndose el número de capas y neuronas, la función de activación, el tamaño de la ventana temporal y la tasa de dropout, con la finalidad de prevenir el sobreajuste y mejorar el desempeño predictivo del modelo.

Por último, en la cuarta fase, se desarrolló la evaluación de desempeño de cada modelo mediante métricas estadísticas como el coeficiente de determinación (R^2), la raíz del error cuadrático medio (RMSE) y el error absoluto medio (MAE). Con base en los resultados de los indicadores, se seleccionó el modelo que presente el mejor balance entre precisión, estabilidad y capacidad predictiva, siendo considerado como la herramienta principal para la estimación de las concentraciones futuras de $PM_{2.5}$ en el área de estudio.

3.4.2. Técnicas

La técnica empleada fue la observación indirecta, la cual consistió en la recopilación y análisis de datos previamente obtenidos de una estación meteorológica ubicada en la ciudad de Concepción, el cual cuenta con un sistema Davis Vantage Pro y un sensor PurpleAir. Esto permitió acceder a información temporal de concentraciones de $PM_{2.5}$ y las variables meteorológicas sin intervenir de forma directa en el fenómeno de estudio, lo que resulta adecuado para investigaciones longitudinales y predictivas.

Asimismo, se realizó un análisis de datos mediante el procesamiento estadístico y el modelado con algoritmos de Machine Learning (Random Forest, Regresión y Red Neuronal Artificial), lo que permitieron identificar de patrones, relaciones, además de las tendencias en el comportamiento de la contaminación atmosférica y también la evaluación comparativa del desempeño predictivo de cada modelo.

3.4.3. Instrumentos

El instrumento de recolección de datos empleados fue una ficha de registro de datos diseñada específicamente para organizar los datos obtenidos, esto permitió registrar información horaria, diaria y mensual de $PM_{2.5}$ y los parámetros meteorológicos obtenidos de la estación meteorológica instalada en la ciudad de Concepción. La ficha incluyó campos para la fecha, hora, valores medidos, unidades de medida y observaciones adicionales. Posteriormente, esta ficha se digitalizará en hojas de cálculo (Excel o CSV) para su procesamiento estadístico y modelado computacional.

3.5. Análisis de datos

El análisis de los datos se llevó a cabo siguiendo una secuencia estructurada que asegure la calidad, coherencia y validez de la información utilizada en la investigación:

Primero se realizó la preparación y depuración de los datos obtenidos, integrando las series temporales horarias y diarias de concentraciones de $PM_{2.5}$ y variables meteorológicas (temperatura, humedad relativa, velocidad y dirección del viento, presión atmosférica, y

precipitación) registradas por la estación meteorológica instalada en la ciudad de Concepción, equipada con un sistema Davis Vantage Pro y un sensor PurpleAir. Los registros fueron homologados en sus unidades de medida, se eliminaron duplicados y valores atípicos no representativos, y se realizó la imputación de datos faltantes con el fin de reducir el sesgo y evitar resultados no fiables. Para ello se evaluaron tres métodos: la imputación simple (Simple Value Imputation), el paquete MICE (Multivariate Imputation by Chained Equations) y MissForest. La selección del método óptimo se determinó mediante las métricas RMSE, MAE y R^2 .

Posteriormente, se desarrolló un análisis exploratorio de datos, el cual incluyó estadística descriptiva, identificación de patrones estacionales y evaluación de correlaciones entre variables predictoras y la concentración de $PM_{2.5}$. Esta etapa permitió detectar tendencias, variabilidad y posibles interacciones entre factores meteorológicos y niveles de contaminación, así como preparar las variables para el modelado.

En la etapa de modelado predictivo, los datos fueron divididos en conjuntos de entrenamiento (70%) y prueba (30%), preservando la secuencia temporal para evitar fugas de información. Se implementaron tres modelos de Machine Learning: Random Forest (RF), ajustando parámetros como el número de árboles ($n_estimators$) y la profundidad máxima (max_depth); Regresión, evaluando enfoques lineales y regularizados, considerando la multicolinealidad y optimizando parámetros mediante validación cruzada; y Red Neuronal Artificial (ANN), configurando una arquitectura especializada para series temporales, definiendo el número de capas y neuronas, la función de activación, el tamaño de la ventana temporal y la tasa de dropout para evitar sobreajuste. Cada modelo fue validado mediante validación cruzada de tipo temporal (TimeSeriesSplit) para garantizar la estabilidad y la robustez de los resultados.

Por último, se evaluó el desempeño mediante el cálculo de métricas como el coeficiente de determinación (R^2), raíz del error cuadrático medio (RMSE), error cuadrático medio (MSE), el error absoluto medio (MAE) y el error porcentual absoluto medio (MAPE), cada una de estas métricas permitieron cuantificar la precisión y consistencia de las predicciones, facilitando la comparación entre los tres enfoques para seleccionar el modelo con mejor rendimiento. Además, se analizó la importancia relativa de las variables predictoras en el modelo seleccionado, con el fin de aportar evidencia sobre los factores meteorológicos que más influyen en las concentraciones de $PM_{2.5}$ en la ciudad de Concepción.

3.6. Consideraciones éticas

La investigación se desarrolló bajo principios éticos orientados a garantizar la integridad científica y la veracidad y validez de los resultados:

En primer lugar, toda la información utilizada fue obtenida de la estación meteorológica instalada en la ciudad de Concepción, equipada con un sistema Davis Vantage Pro y un sensor PurpleAir para la medición de PM_{2.5}, evitándose así el uso de datos obtenidos de forma ilícita o sin consentimiento. Al ser registros ambientales y meteorológicos, no se involucró información personal ni sensible, por lo que no existe riesgo de vulnerar la privacidad de personas u organizaciones.

Asimismo, se mantuvo transparencia durante el procesamiento y análisis de datos, generando evidencia documentaria que muestre cada uno de los procedimientos como son: la depuración, modelado y validación para garantizar la reproducibilidad del estudio. Se evitó cualquier manipulación de los datos que puedan sesgar los resultados, y se redactó lo hallado de manera objetiva siendo favorables o no favorables los modelos aplicados.

Por último, en la investigación se citó las partes que correspondan a información de otros autores respetándose así la propiedad intelectual, reconociendo adecuadamente las fuentes bibliográficas y las metodológicas utilizadas.

CAPÍTULO IV

RESULTADOS Y DISCUSIONES

4.1. Presentación de Resultados y Discusiones

4.1.1. Variabilidad de las condiciones meteorológicas predominantes en la ciudad Concepción durante el periodo 2023-2025.

La caracterización climática presentada en la Tabla 2 permite entender el "escenario de dispersión" en el que se encuentra la ciudad de Concepción. La temperatura promedio se mantuvo en 12.97 ± 4.84 °C con valores mínimos negativos en 2024 (-1.10 °C) y 2025 (-2.50 °C), lo que sugiere episodios de heladas durante esos años. La humedad relativa presentó una media general de $66.16 \pm 18.89\%$, con una tendencia decreciente a lo largo del periodo, pasando de 69.40% en 2023 a 65.39% en 2025. Un hallazgo crítico es la baja velocidad del viento registrada (promedio inferior a 1 m/s), la cual muestra una tendencia decreciente hacia el año 2025, en cuanto presión atmosférica, se observa una media general de 676.08 ± 9.48 mmHg, manteniéndose relativamente estable entre los tres años.

Tabla 2

Estadísticos descriptivos y comportamiento interanual de los parámetros meteorológicos en la estación de Concepción 2023-2025.

Variable	Unidad	Media $\pm$ DE	Mediana	[Min - Max]	[Q1-Q3]
2023					
Presión atmosférica	mmHg	676.22 a $\pm$ 25.64	674.50 a	[663.90 - 1022.50] a	[672.40 - 676.50] a
Humedad Relativa	%	69.40 a $\pm$ 16.90	72.00 a	[10.00 - 95.00] a	[57.00 - 84.00] a
Temperatura	°C	13.71 a $\pm$ 4.12	12.90 a	[4.10 - 26.80] a	[10.40 - 16.70] a
Velocidad de viento	m/s	0.73 a $\pm$ 0.89	0.40 a	[0.00 - 6.70] a	[0.00 - 1.30] a
2024					
Presión atmosférica	mmHg	676.17 a $\pm$ 3.57	676.20 a	[663.70 - 689.40] a	[673.80 - 678.50] a
Humedad Relativa	%	66.05 b $\pm$ 19.03	68.22 b	[4.00 - 95.00] b	[52.00 - 82.00] b
Temperatura	°C	13.04 b $\pm$ 4.94	12.30 b	[-1.10 - 26.70] b	[9.60 - 16.70] b
Velocidad de viento	m/s	0.73 a $\pm$ 0.84	0.40 a	[0.00 - 7.60] a	[0.00 - 1.30] a
2025					
Presión atmosférica	mmHg	675.94 b $\pm$ 3.82	676.00 b	[662.90 - 690.50] b	[673.40 - 678.40] b
Humedad Relativa	%	65.39 c $\pm$ 19.18	68.00 c	[9.00 - 95.00] c	[52.00 - 81.00] c
Temperatura	°C	12.70 c $\pm$ 4.89	11.90 c	[-2.50 - 26.70] c	[9.10 - 16.50] c
Velocidad de viento	m/s	0.46 b $\pm$ 0.77	0.00 b	[0.00 - 6.70] b	[0.00 - 0.90] b

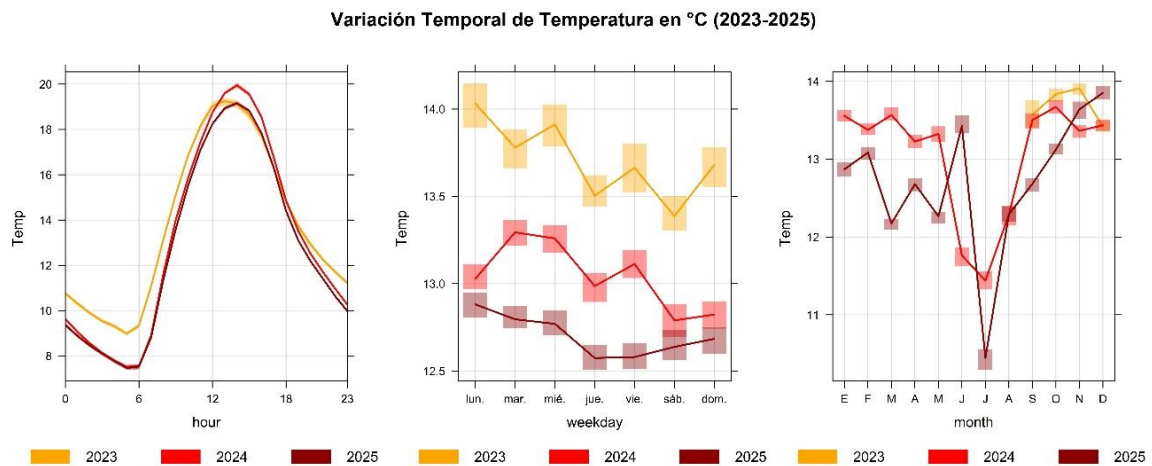
<i>2023-2025</i>					
Presión atmosférica	mmHg	676.08 ± 9.48	675.9	[662.90 - 1022.50]	[673.46 - 678.20]
Humedad Relativa	%	66.16 ± 18.89	69	[4.00 - 95.00]	[53.00 - 82.00]
Temperatura	°C	12.97 ± 4.84	12.2	[-2.50 - 26.80]	[9.50 - 16.60]
Velocidad de viento	m/s	0.61 ± 0.83	0.4	[0.00 - 7.60]	[0.00 - 0.90]

Nota: DE: Desviación estándar, Q1: Primer cuartil, Q3: Tercer cuartil, Min: Valor mínimo, Max: Valor máximo

La Figura 6 ilustra la dinámica térmica de la ciudad en tres escalas de tiempo, permitiendo identificar patrones cíclicos que influyen en la estabilidad de la capa límite atmosférica. A nivel horario, se registraron las temperaturas más bajas entre las 05:00 y 06:00 horas, asociadas al enfriamiento radiativo nocturno y a condiciones de estabilidad atmosférica que favorecen la acumulación de contaminantes; posteriormente, la temperatura aumentó progresivamente hasta alcanzar sus valores máximos entre las 13:00 y 15:00 horas, promoviendo procesos convectivos que facilitan la dispersión de las partículas en suspensión. En la escala semanal, se observó una escasa variabilidad térmica entre los días de la semana, evidenciando que la dinámica de la temperatura está influenciada principalmente por factores meteorológicos regionales más que por actividades antrópicas. Asimismo, a escala mensual, se identificó una marcada estacionalidad, con temperaturas más bajas durante los meses de invierno (junio-agosto) y más elevadas hacia finales de la primavera y comienzos del verano. Estos resultados sugieren que las madrugadas y los meses de invierno constituyen periodos críticos para la acumulación de PM_{2.5}, debido a la mayor estabilidad atmosférica y la limitada dispersión de contaminantes, mientras que las temperaturas más altas favorecen una mayor mezcla atmosférica y capacidad de autodepuración del aire.

Figura 6

Variación horaria, semanal y mensual de la temperatura en Concepción durante el periodo 2023-2025.



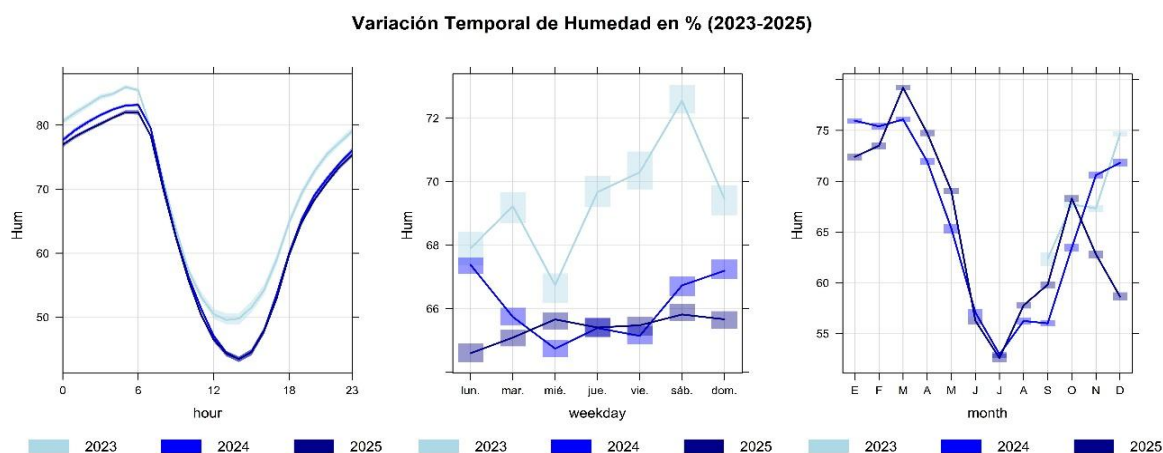
La distribución temporal de la temperatura observada en Concepción demuestra que esta variable meteorológica desempeña un papel importante en la dinámica de las concentraciones de $PM_{2.5}$. Las bajas temperaturas registradas durante las primeras horas de la mañana favorecen una atmósfera más estable, reduciendo los procesos de mezcla vertical y facilitando la acumulación de contaminantes cerca de la superficie. De acuerdo con Prado Fasanando (2021), las condiciones de estabilidad atmosférica asociadas al enfriamiento nocturno limitan la dispersión de los contaminantes y contribuyen al incremento de sus concentraciones. Por el contrario, el aumento de la temperatura durante las horas diurnas genera movimientos convectivos que incrementan la altura de la capa límite atmosférica y favorecen la dispersión de las partículas suspendidas, reduciendo las concentraciones de $PM_{2.5}$, tal como indican Vidal-Daza y Pérez-Vidal (2018). Asimismo, la reducida variabilidad térmica semanal observada sugiere que la temperatura está condicionada principalmente por factores meteorológicos regionales más que por fluctuaciones de la actividad urbana, coincidiendo con lo reportado por la U.S. Environmental Protection Agency (2026). A nivel estacional, las menores temperaturas registradas durante el invierno constituyen un factor que favorece la acumulación de contaminantes debido a la persistencia de condiciones atmosféricas estables, mientras que las temperaturas más elevadas de primavera y verano incrementan la ventilación y la capacidad de autodepuración atmosférica. Estos resultados confirman que la temperatura desempeña un papel fundamental en la dinámica temporal del $PM_{2.5}$ y explican por qué las madrugadas y los

meses invernales representan los periodos de mayor riesgo de acumulación de este contaminante en la ciudad de Concepción.

La Figura 7 presenta la variación temporal de la humedad relativa registrada durante el periodo 2023-2025. En cuanto a la variabilidad horaria, se observa un comportamiento inversamente proporcional al de la temperatura, con valores máximos que oscilan entre 80% y 85%, alcanzados durante las primeras horas de la madrugada (04:00 – 06:00 horas), y valores mínimos cercanos al 45%–50% registrados entre las 12:00 y 14:00 horas, coincidiendo con el periodo de mayor calentamiento superficial. Respecto a la variabilidad semanal, la humedad relativa muestra una tendencia relativamente estable a lo largo de los días de la semana, con fluctuaciones leves que oscilan entre 65% y 70%, sin diferencias marcadas entre días laborales y fines de semana. En cuanto a la variabilidad mensual, se observa un descenso pronunciado entre los meses de junio y agosto, marcando el pico de la temporada seca con valores mínimos cercanos al 53%–55%; por el contrario, durante los meses de diciembre a marzo la humedad se mantiene elevada, superando el 70%–75%, coincidiendo con el periodo de lluvias y el consecuente mayor lavado atmosférico. Este patrón es consistente con la circulación estacional propia de los Andes tropicales.

Figura 7

Variación horaria, semanal y mensual de la humedad en Concepción durante el periodo 2023-2025.



Estos resultados son consistentes con lo reportado por Zhang et al. (2017), quienes demostraron que la humedad relativa no actúa de manera aislada sobre la concentración del $PM_{2.5}$, sino que su efecto depende de la variabilidad temporal y de las condiciones atmosféricas acompañantes. En concordancia con los resultados del presente estudio, dichos autores señalan que bajo condiciones de atmósfera estable, la humedad relativa favorece la

acumulación progresiva de $PM_{2.5}$; asimismo, enfatizan que cuando la humedad relativa se mantiene en el rango de 45% a 84%, se registra un incremento acumulativo en la concentración de $PM_{2.5}$. En esta misma línea, Zender-Świercz et al. (2024) destacan que la humedad relativa constituye un factor determinante en la dinámica del $PM_{2.5}$, dado que promueve el crecimiento higroscópico de las partículas, este proceso incrementa tanto el tamaño como la masa de las partículas, modificando su comportamiento en la atmósfera; en consecuencia, aquellas partículas que alcanzan un tamaño suficiente tienden a depositarse por gravedad, lo que puede traducirse en una reducción temporal de la concentración de $PM_{2.5}$ en el aire. Las fluctuaciones semanales de humedad relativa en Concepción están gobernadas principalmente por factores de macroescala, como el ingreso de masas de aire regionales, y no por microclimas generados por la actividad urbana local, sin diferencias significativas entre días laborables y fines de semana. Este comportamiento es consistente con los patrones climáticos de larga duración reportados para el Valle del Mantaro (Arroyo-Aliaga, 2011), y concuerda con lo señalado por Zender-Świercz et al. (2024), quienes indican que la humedad relativa responde fundamentalmente a procesos atmosféricos de mayor escala, siendo poco sensible a variaciones de origen urbano en ciudades de tamaño intermedio. Las épocas de estiaje favorecen resuspensión de polvo y la permanencia de partículas finas en suspensión por más tiempo al no haber procesos de condensación (Martínez-Soto et al., 2025). Este comportamiento es consistente con lo reportado por Villalobos-Puma et al. (2024), quienes documentaron que durante la temporada seca, cuando la humedad relativa alcanza sus valores más bajos, se registran las mayores concentraciones de carbono negro (julio-noviembre), siendo las contribuciones de quema de biomasa las predominantes tanto en el valle del Mantaro como en la zona de alta montaña de Huaytapallana. La variabilidad de la humedad en Concepción define dos escenarios para el $PM_{2.5}$. 1) Un escenario nocturno/matutino de alta humedad que favorece el aumento de tamaño de las partículas por hidratación, y 2) Un escenario estacional seco (invierno) donde la baja humedad relativa, sumada a la falta de lluvias, crea las condiciones ideales para que las partículas finas permanezcan más tiempo en el aire, incrementando el riesgo de exposición para la población.

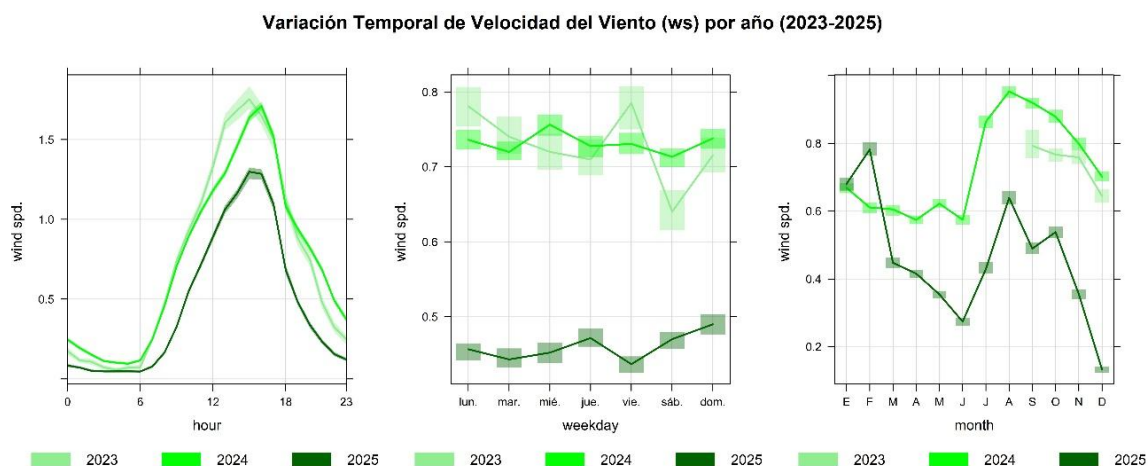
La Figura 8 detalla la dinámica eólica de la ciudad en distintas escalas temporales, revelando un régimen de vientos predominantemente débil que condiciona la calidad del aire local, en la variabilidad horaria se identificó un patrón claro de incremento de velocidad de viento en horas de mayor insolación (14:00 – 16:00), los valores mínimos (cerca de 0.2 m/s) ocurren durante la madrugada y primeras horas de la mañana (00:00 – 08:00); este comportamiento es consistente con lo reportado por Ccasani-Guillén (2019), quienes

documentaron que la máxima velocidad del viento en el Valle del Mantaro ocurre alrededor de las 14:00 horas, forzada por la variación diurna de la radiación solar y las brisas de montaña-valle características de la topografía andina, mientras que las condiciones de calma predominan durante las primeras horas de la mañana, favoreciendo la acumulación de contaminantes en la capa baja de la atmósfera, con respecto a la variabilidad semanal se observa que en los años (2023-2024) presentan una uniformidad notable muy diferente a lo que se presneto en el años 2025 que presento velocidad de vientos menores a 0.5 m/s, pero no se observan variaciones significativas entre los días laborales y el fin de semana, lo que confirma que el régimen de vientos en Concepción está gobernado por la topografía del valle y factores meteorológicos regionales, y no por microclimas inducidos por la actividad urbana. En concordancia con lo señalado por Prado (2021), quien indica que el Valle del Mantaro constituye una zona de topografía compleja que condiciona el comportamiento de la capa límite atmosférica y los patrones de viento local. La variación mensual nos muestra que hay una tendencia decreciente de la velocidad media conforme avanza el año, así mismo los meses de junio a agosto (temporada seca/invierno) presentan velocidades ligeramente menores o más inestables en comparación con el inicio del año. Esta reducción de la velocidad del viento en los meses secos configura el escenario de mayor riesgo ambiental para la ciudad, ya que se reduce significativamente la capacidad de limpieza atmosférica por dispersión eólica del $PM_{2.5}$ (Martínez-Soto et al., 2025).

La ciudad de Concepción se caracteriza por un régimen de vientos extremadamente débiles (calmas), con promedios horarios que rara vez superan los 1.0 m/s. Los resultados sugieren que el estancamiento de masas de aire es la condición predominante, obligando a que la gestión de emisiones de $PM_{2.5}$ sea mucho más rigurosa, dado que la atmósfera local no tiene la capacidad de "limpiarse" por sí sola mediante el transporte eólico.

Figura 8

Variación horaria, semanal y mensual de la velocidad de viento en Concepción durante el periodo 2023-2025.

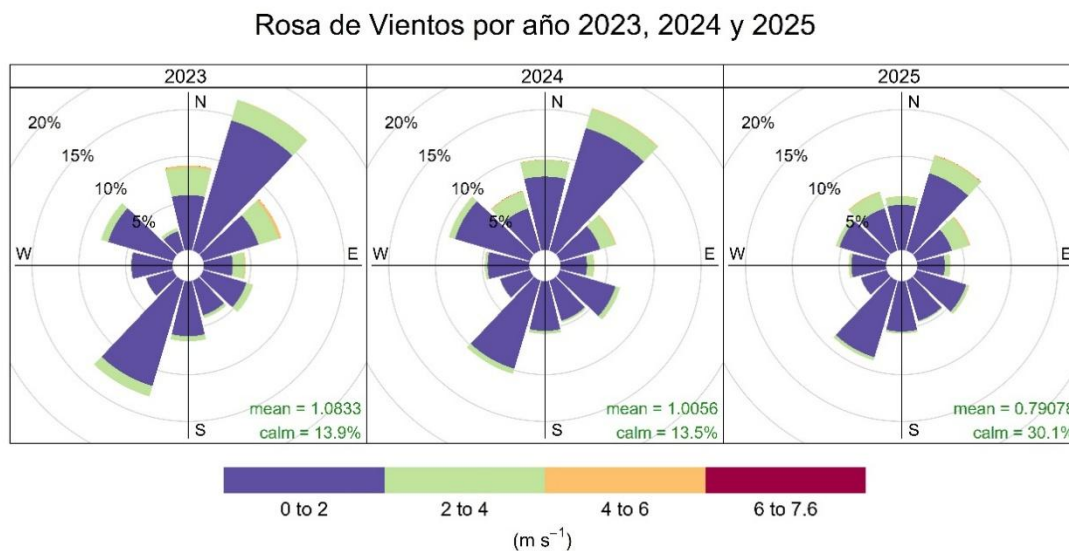


Con respecto a la dirección del viento para el periodo 2023-2025 se visualiza (Figura 9) un patrón bimodal con flujos significativos provenientes del Noreste (N) y del Sur-Sureste (SSE), con velocidades que oscilan principalmente entre 0 y 2 m/s. Durante 2023, la velocidad media fue de 1.08 m/s con un porcentaje de calmas del 13.9%, mientras que en 2024 la velocidad media fue similar (1.01 m/s) y las calmas alcanzaron el 13.5%, evidenciando una notable consistencia entre ambos años. En contraste, el año 2025 registró una reducción considerable de la velocidad media (0.79 m/s) y un incremento significativo del porcentaje de calmas (30.1%), lo que indica condiciones de menor dinámica eólica respecto a los años anteriores. El patrón bimodal NE-SSE observado en la rosa de vientos de Concepción es consistente con lo reportado por Álvarez-Tolentino & Suárez-Salas (2020), quienes documentaron que los vientos en Concepción provienen predominantemente del noreste con velocidades de 1 a 3 m/s. Asimismo, Flores-Rojas et al. (2019) señalan que la orientación noroeste-sureste del Valle del Mantaro favorece la formación de vientos locales por circulación térmica montaña-valle, lo cual explica la presencia de dos sistemas de viento identificados por Ccasani-Guillén (2019): las brisas de montaña-valle de origen térmico y los vientos sinópticos provenientes del SSE-SE asociados al campo de presión regional. La ciudad de Concepción presenta un régimen de vientos de muy baja intensidad, lo que condiciona una capacidad de ventilación limitada. El predominio de velocidades menores a 1 m/s actúa como un factor de riesgo ambiental, ya que facilita que las concentraciones de material particulado fino se eleven rápidamente ante

cualquier incremento en la actividad antrópica, validando la necesidad de correlacionar estas "calmas" con los picos de contaminación observados en el Objetivo 2.

Figura 9

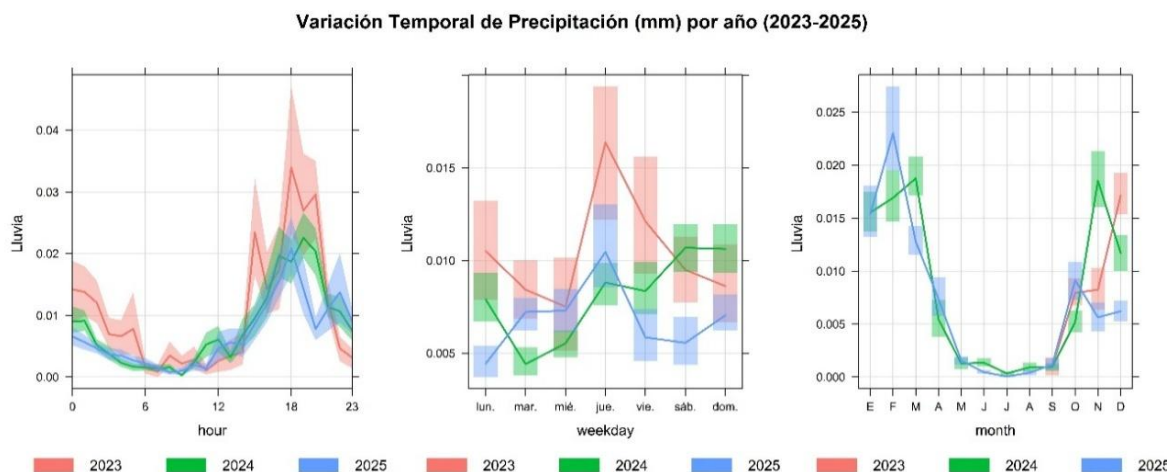
Rosa de vientos y distribución de frecuencias de velocidad y dirección en la estación de Concepción (2023-2025).



La Figura 10 muestra la variación temporal de las lluvias, factor crítico que actúa como un regulador natural de las concentraciones de material particulado fino a escala horaria, semanal y mensual. A nivel horario, las precipitaciones presentaron una mayor intensidad durante las horas de la tarde y noche, registrándose los valores más elevados entre las 16:00 y 21:00 horas, con máximos cercanos a las 18:00 y 19:00 horas. En la escala semanal, la precipitación mostró ligeras fluctuaciones entre los días de la semana, sin evidenciar un patrón claramente definido. Por otro lado, la variación mensual reveló una marcada estacionalidad, con mayores precipitaciones durante los meses de enero a marzo y valores mínimos entre junio y agosto, correspondientes al periodo de estiaje.

Figura 10

Variación horaria, semanal y mensual de la precipitación en Concepción durante el periodo 2023-2025.



El patrón horario observado es característico de las regiones andinas tropicales, donde el calentamiento de la superficie terrestre durante el día favorece el desarrollo de procesos convectivos que generan precipitaciones durante la tarde y primeras horas de la noche (Villalobos et al., 2019; IGP, 2020). Desde el punto de vista ambiental, este comportamiento resulta beneficioso para la calidad del aire, ya que las lluvias contribuyen al lavado atmosférico y a la deposición húmeda de partículas suspendidas, reduciendo las concentraciones de $PM_{2.5}$ acumuladas durante el día. Respecto a la variabilidad semanal, la ausencia de un patrón definido indica que la ocurrencia de precipitaciones depende principalmente de la dinámica atmosférica regional y del ingreso de masas de aire, más que de factores antrópicos locales (Arroyo-Aliaga, 2011; IGP, 2005). Asimismo, la marcada estacionalidad observada a escala mensual coincide con el régimen climático característico del Valle del Mantaro, donde la temporada lluviosa se concentra entre diciembre y marzo, mientras que los meses de invierno presentan precipitaciones escasas o nulas. Esta reducción de las lluvias durante el periodo seco limita los procesos de remoción atmosférica y favorece una mayor permanencia de partículas finas en suspensión, incrementando el riesgo de acumulación de $PM_{2.5}$ (Zhang et al., 2018, citado en Martínez-Soto et al., 2025). En consecuencia, la precipitación constituye uno de los principales mecanismos naturales de autodepuración atmosférica en Concepción, generando condiciones más favorables para la dispersión y eliminación de contaminantes durante la temporada húmeda y una mayor vulnerabilidad a episodios de contaminación durante el estiaje.

4.1.2. Variabilidad de las concentraciones del $PM_{2.5}$ en la ciudad de Concepción, 2023-2025.

Los datos de la Tabla 3 evidencian que las concentraciones de $PM_{2.5}$ en la estación de Concepción presentaron una marcada variabilidad durante el periodo 2023–2025. La media total fue de $14.08 \mu\text{g}/\text{m}^3$, mientras que la desviación estándar alcanzó $14.73 \mu\text{g}/\text{m}^3$, valor similar al promedio, lo que indica una alta dispersión de los datos y la ocurrencia de episodios de contaminación de gran magnitud. Asimismo, la mediana total ($10.8 \mu\text{g}/\text{m}^3$) fue inferior a la media, sugiriendo una distribución asimétrica influenciada por valores extremos. Los registros máximos superaron los $270 \mu\text{g}/\text{m}^3$ en los tres años evaluados, alcanzando $304.60 \mu\text{g}/\text{m}^3$ en 2023, lo que refleja la presencia de eventos puntuales de contaminación severa. En el análisis anual, 2024 presentó la mayor concentración promedio ($16.72 \pm 16.52 \mu\text{g}/\text{m}^3$), seguido de 2023 ($15.98 \pm 13.84 \mu\text{g}/\text{m}^3$), mientras que 2025 registró la menor media ($10.79 \pm 12.24 \mu\text{g}/\text{m}^3$). Del mismo modo, los valores de mediana y rango intercuartílico muestran una reducción de las concentraciones en 2025 respecto a los años anteriores, evidenciando una posible mejora en la calidad del aire durante ese periodo.

Tabla 3

Estadísticos descriptivos y Análisis Comparativo Anual y Total de $PM_{2.5} \mu\text{g}/\text{m}^3$ en la estación de Concepción 2023-2025

Year	Unidad	Media $\pm$ DE	Mediana	[Min - Max]	[Q1-Q3]
2023	$\mu\text{g}/\text{m}^3$	$15.98 \text{ b} \pm 13.84$	13.10 b	[0.40 - 304.60] b	[8.30 - 19.50] b
2024	$\mu\text{g}/\text{m}^3$	$16.72 \text{ a} \pm 16.52$	13.10 a	[0.1 - 271.70] a	[6.90 - 20.70] a
2025	$\mu\text{g}/\text{m}^3$	$10.79 \text{ c} \pm 12.24$	7.60 c	[0.1 - 271.40] c	[3.70 - 14.70] c
Total (2023-2025)	$\mu\text{g}/\text{m}^3$	14.08 ± 14.73	10.8	[0.1 - 304.60]	[5.20 - 18.10]

Nota: DE: Desviación estándar, Q1: Primer cuartil, Q3: Tercer cuartil, Min: Valor mínimo, Max: Valor máximo.

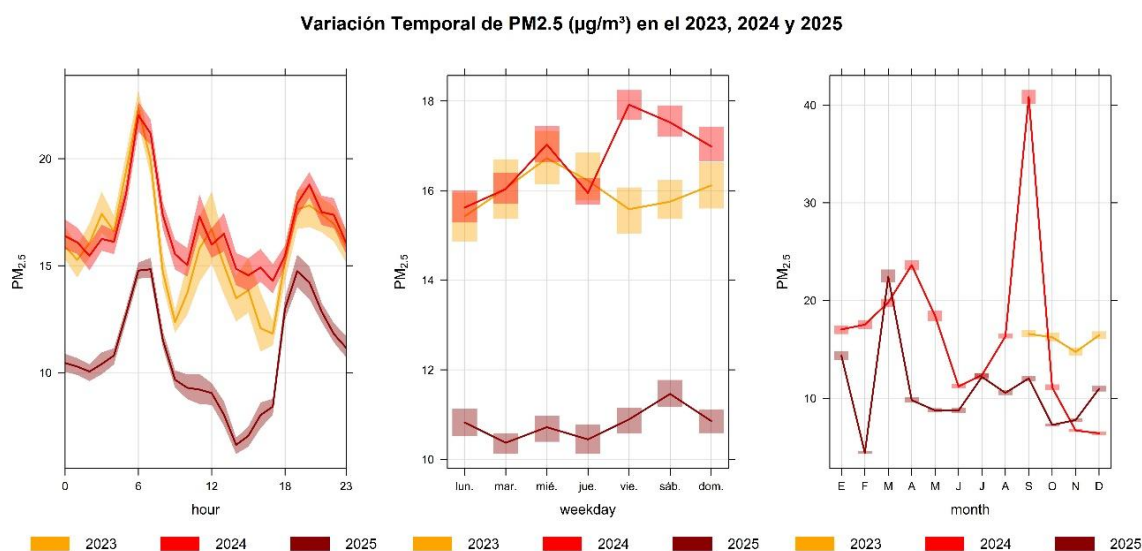
Estos resultados sugieren que el comportamiento del $PM_{2.5}$ en Concepción está caracterizado por una concentración de fondo moderada, interrumpida por episodios extremos que incrementan significativamente los niveles de contaminación atmosférica. La elevada variabilidad observada podría estar asociada a factores meteorológicos que favorecen la acumulación temporal de partículas, así como a fuentes emisoras de origen antropogénico y natural. En este sentido, Romero et al. (2020) señalaron que las bajas velocidades del viento y las condiciones atmosféricas estables favorecen el incremento de las concentraciones de $PM_{2.5}$ en Lima Metropolitana, debido a la reducción de los procesos

de dispersión. De manera similar, Warthon et al. (2023) demostraron que variables como la precipitación, la humedad relativa y la velocidad del viento influyen significativamente en la dispersión del material particulado en ciudades andinas del Perú. Asimismo, Cabello-Torres et al. (2024) identificaron que los periodos de estabilidad atmosférica favorecen la acumulación de partículas finas y generan incrementos abruptos en las concentraciones urbanas. Por tanto, la mayor frecuencia de concentraciones elevadas observada en 2023 y 2024, así como la reducción registrada en 2025, podría estar relacionada con variaciones en las condiciones meteorológicas locales que modifican la capacidad de dispersión y remoción de los contaminantes atmosféricos.

La Figura 11 permite analizar la dinámica temporal de las concentraciones de material particulado fino, identificando los periodos de mayor exposición y los factores que impulsan la contaminación local en las escalas horaria, semanal y mensual. A nivel horario, los mayores valores se registran durante las primeras horas de la mañana y en la noche, mientras que las concentraciones disminuyen durante la tarde. En la escala semanal, las diferencias entre días son moderadas, observándose valores ligeramente más altos hacia el final de la semana, especialmente en 2024. Por otro lado, la variación mensual muestra una marcada estacionalidad, con concentraciones más elevadas durante algunos meses específicos del año, destacando septiembre de 2024 como el periodo de mayor contaminación. En conjunto, los resultados indican que las concentraciones de $PM_{2.5}$ no se distribuyen de manera uniforme en el tiempo, sino que responden a patrones recurrentes que reflejan la influencia de factores temporales sobre la calidad del aire.

Figura 11

Variación horaria, semanal y mensual del $PM_{2.5}$ en Concepción 2023-2025.



Los patrones temporales observados en las concentraciones de $PM_{2.5}$ evidencian la influencia conjunta de las actividades antrópicas y las condiciones meteorológicas sobre la calidad del aire en Concepción. Las mayores concentraciones registradas durante las primeras horas de la mañana y la noche sugieren que las emisiones asociadas al tránsito vehicular y otras actividades urbanas coinciden con condiciones atmosféricas estables que favorecen la acumulación de contaminantes cerca de la superficie. Asimismo, la relativa persistencia de los niveles de $PM_{2.5}$ a lo largo de la semana coincide con lo reportado por Álvarez-Tolentino y Suárez-Salas (2020) en el Valle del Mantaro, quienes señalaron que la contaminación atmosférica regional es resultado de la contribución simultánea de fuentes urbanas y rurales.

La marcada elevación de las concentraciones durante algunos meses del año, especialmente en septiembre del 2024 tiene una confluencia de factores normativos y dinámicos atmosféricos. Villalobos-Puma et al. (2024) y Flores-Rojas et al. (2019) documentaron que existe una estrecha relación entre la cuenca amazónica y los Andes Centrales del Perú, donde el South American Low Level Jet (SALLJ) transporta masas de aire cargadas de aerosoles desde la Amazonía hacia el valle del Mantaro, canalizadas a través de los pasos de la cordillera oriental. Desde el punto de vista normativo, el año 2024 constituyó un hito en la gestión forestal peruana con la promulgación de la Ley 31973 (denominada "Ley Antiforestal"), que eliminó la obligatoriedad de los estudios de clasificación de suelos para el cambio de uso de tierras en la Amazonía. Esto generó un incremento descontrolado de la quema de biomasa forestal en la cuenca amazónica durante la temporada seca de 2024. La combustión incompleta de biomasa produce material particulado fino ($PM_{2.5}$), transportado por el SALLJ y canalizado orográficamente a través de los pasos de la cordillera oriental, alcanzó el valle del Mantaro durante el mes de septiembre el año 2024. En consecuencia, el dato anómalo registrado no responde a fuentes locales de emisión, sino a un evento de transporte de larga distancia de material particulado fino ($PM_{2.5}$) de origen pirógeno, cuya magnitud excepcional está directamente vinculada al descontrol en la gestión forestal propiciado por el marco normativo vigente de ese año.

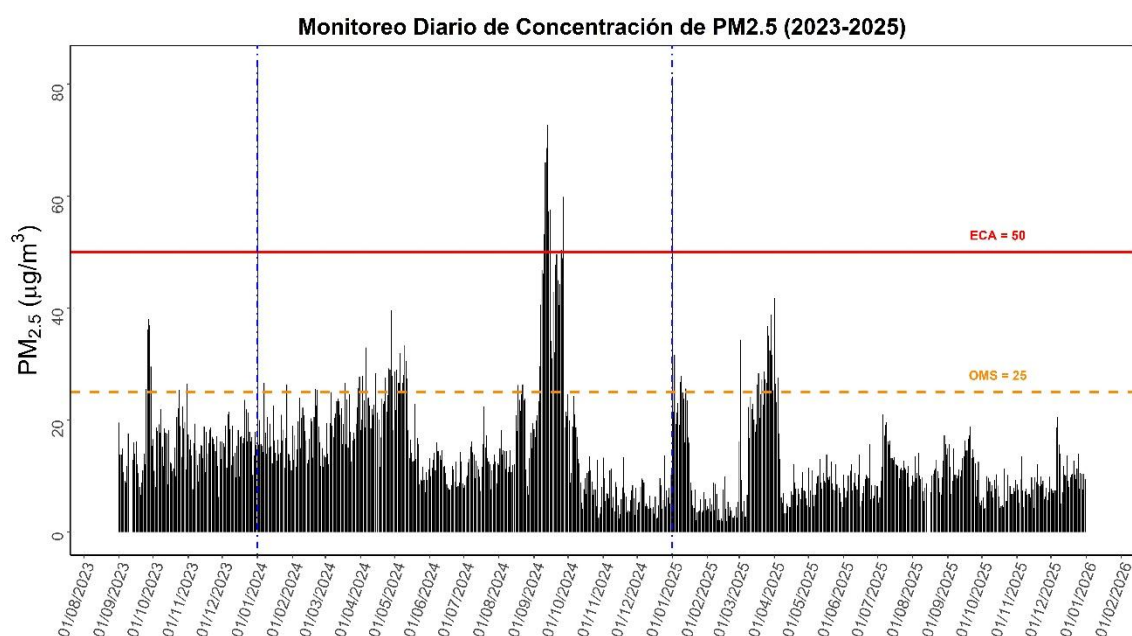
Por otro lado, la marcada elevación de las concentraciones durante determinados meses, especialmente en septiembre de 2024, refleja la influencia de la estacionalidad meteorológica, ya que la disminución de las precipitaciones y la mayor estabilidad atmosférica favorecen la permanencia de partículas en el aire; resultados similares fueron reportados por Warthon et al. (2023) en ciudades andinas del Perú. En conjunto, estos

hallazgos demuestran que las concentraciones de $PM_{2.5}$ presentan patrones temporales recurrentes que pueden ser aprovechados en la modelación predictiva. En este sentido, Olarte et al. (2025) señalan que la incorporación de variables meteorológicas y temporales permite mejorar el desempeño de los modelos de Machine Learning para la predicción de $PM_{2.5}$, al capturar la variabilidad espacial y temporal de los contaminantes atmosféricos.

La Figura 12 muestra la evolución diaria de las concentraciones de $PM_{2.5}$ registradas en la estación de monitoreo de Concepción durante el periodo 2023–2025, comparándolas con el Estándar de Calidad Ambiental (ECA) del Perú para 24 horas ($50 \mu\text{g}/\text{m}^3$) y la guía diaria de la Organización Mundial de la Salud (OMS) ($25 \mu\text{g}/\text{m}^3$). Se observa que la mayoría de los registros diarios se mantienen por debajo del límite establecido por la normativa nacional; sin embargo, existen episodios específicos en los que las concentraciones superan dicho estándar, especialmente durante septiembre de 2024, cuando se registran los valores más elevados de toda la serie, alcanzando aproximadamente $72.73 \mu\text{g}/\text{m}^3$. Asimismo, se evidencia una frecuencia considerable de días que exceden la guía de la OMS, principalmente entre abril y septiembre de 2024 y durante marzo de 2025. Estos resultados indican que, aunque el cumplimiento del ECA peruano es predominante, las concentraciones de $PM_{2.5}$ superan recurrentemente los niveles recomendados internacionalmente para la protección de la salud humana.

Figura 12

Valores diarios de concentración de $PM_{2.5}$ en comparación con los Estándares de Calidad Ambiental (ECA-Aire Perú) y Guías OMS en la estación de Concepción 2023-2025.



Los resultados revelan que la evaluación de la calidad del aire varía significativamente según el estándar de referencia utilizado. Mientras que la mayoría de los registros cumplen con el ECA peruano de $50 \mu\text{g}/\text{m}^3$, la frecuencia de superaciones respecto a la guía de la OMS revela una exposición recurrente de la población a concentraciones potencialmente perjudiciales para la salud. Esta diferencia entre los estándares ha sido ampliamente discutida por la World Health Organization. (2021), la cual sostiene que incluso exposiciones de corta duración a concentraciones moderadas de $\text{PM}_{2.5}$ pueden incrementar el riesgo de enfermedades respiratorias y cardiovasculares, especialmente en grupos vulnerables.

El marcado incremento observado durante septiembre de 2024 sugiere la influencia de condiciones meteorológicas desfavorables para la dispersión de contaminantes. La ocurrencia de episodios de estabilidad atmosférica, bajas velocidades de viento y ausencia de precipitaciones favorece la acumulación de partículas finas cerca de la superficie. Estos hallazgos coinciden con lo reportado por Romero et al. (2020) en Lima Metropolitana, quienes identificaron que las condiciones atmosféricas estables contribuyen significativamente al aumento de las concentraciones de $\text{PM}_{2.5}$ y a la ocurrencia de episodios de contaminación prolongados.

Asimismo, la concentración de excedencias en determinados meses del año evidencia un comportamiento estacional del contaminante. Estudios realizados en ciudades andinas del Perú por Warthon et al. (2023) demostraron que variables meteorológicas como la humedad relativa, la precipitación y la estabilidad atmosférica influyen directamente en la permanencia del material particulado fino en suspensión, favoreciendo incrementos temporales de las concentraciones ambientales. Este comportamiento es consistente con los picos observados en la serie temporal de Concepción.

Por otro lado, la persistencia de superaciones de la guía de la OMS incluso durante 2025, año que presenta menores concentraciones promedio, indica que la problemática de contaminación por $\text{PM}_{2.5}$ mantiene un carácter estructural. Resultados similares fueron reportados por Cabello-Torres et al. (2024), quienes encontraron episodios recurrentes de incremento de $\text{PM}_{2.5}$ asociados a periodos de estabilidad atmosférica en ciudades peruanas, aun cuando los promedios anuales se mantenían dentro de los límites normativos nacionales.

Desde la perspectiva de la gestión ambiental, estos resultados respaldan la necesidad de implementar herramientas predictivas que permitan anticipar episodios críticos de

contaminación. En este sentido, Olarte et al. (2025) destacan que los modelos de Machine Learning constituyen herramientas eficaces para la predicción de PM_{2.5}, ya que permiten integrar variables meteorológicas y ambientales para mejorar la precisión de los pronósticos y apoyar la toma de decisiones orientadas a la protección de la salud pública.

4.1.3. Análisis de Correlación entre variables meteorológicas y concentraciones de PM_{2.5}

La Tabla 4 expone la matriz de correlación de rangos de Spearman y las variables meteorológicas. Este análisis es muy importante ya que nos permite comprender los procesos atmosféricos fundamentales que controlan la dinámica de la calidad de aire en la ciudad de Concepción. Todos los coeficientes resultaron estadísticamente significativos ($p < 0.001$), lo que indica una confianza superior al 99.9% en las relaciones observadas.

Respecto a la relación con el PM_{2.5}, la presión atmosférica mostró la correlación positiva más alta ($r_s = 0.145$), sugiriendo que condiciones de alta presión favorecen la acumulación de contaminantes al reducir la dispersión vertical. La humedad relativa presentó una correlación positiva débil ($r_s = 0.116$), mientras que la temperatura ($r_s = -0.172$) y la velocidad del viento ($r_s = -0.129$) mostraron correlaciones negativas, indicando que tanto el aumento de temperatura como una mayor velocidad del viento contribuyen a la dispersión del PM_{2.5}. La precipitación tuvo la correlación más débil ($r_s = -0.044$), aunque igualmente significativa.

Tabla 4

Matriz de correlación de Spearman entre variables meteorológicas y concentraciones de PM_{2.5} en Concepción

PM _{2.5} _interp	Temp.Out	Out.Hum	Wind.Speed	Rain	Bar
1NA	-0.172***	0.116***	-0.129***	-0.044***	0.145***
-0.172***	1NA	-0.773***	0.639***	-0.08***	-0.534***
0.116***	-0.773***	1NA	-0.562***	0.197***	0.264***
-0.129***	0.639***	-0.562***	1NA	0.034***	-0.308***
-0.044***	-0.08***	0.197***	0.034***	1NA	-0.026***
0.145***	-0.534***	0.264***	-0.308***	-0.026***	1NA

Nota: Los coeficientes representan la correlación de rangos de Spearman (ρ). Los asteriscos (***) indican una significancia estadística de $p < 0.001$, lo que denota una confianza superior al 99.9% en la relación observada. Los valores cercanos a 1.0 indican una relación directa fuerte, mientras que valores cercanos a -1.0 indican una relación inversa.

La temperatura mostró una correlación negativa con $PM_{2.5}$, se observa que cuando la temperatura es más elevada, las concentraciones del material particulado tienden a disminuir ligeramente, posiblemente por un aumento de dispersión atmosférica. Este resultado concuerda con los hallazgos de Olarte et al. (2025), quienes encontraron que temperaturas más altas se asocian con menores niveles de $PM_{2.5}$ debido a una mayor ventilación y mezcla atmosférica.

La humedad relativa muestra una correlación positiva débil, lo cual da entender que en Concepción se da el proceso de higroscopicidad de las partículas y por la coincidencia horaria entre los máximos de humedad nocturna y las condiciones de mayor estabilidad atmosférica, que limitan simultáneamente la dispersión de contaminantes. La correlación positiva entre humedad relativa y $PM_{2.5}$ ($r = 0.116$) encontrada en Concepción es consistente con lo reportado por Zender-Świercz et al. (2024), quienes mediante el coeficiente de Spearman obtuvieron una correlación positiva moderada ($r_s = 0.5$) entre ambas variables, explicada por el proceso de crecimiento higroscópico de las partículas: al combinarse con la humedad atmosférica, las partículas aumentan su densidad y masa, incrementando su concentración superficial medible. Este comportamiento es igualmente respaldado por Zhang et al. (2017), quienes demostraron que la fluctuación de la humedad favorece el incremento de la concentración de masa del $PM_{2.5}$, especialmente bajo condiciones de estabilidad atmosférica.

La velocidad de viento presenta una correlación negativa débil, lo cual es consecuente con el efecto de dispersión del viento, el cual ayuda a reducir la concentración de contaminantes. Este patrón es consistente con lo reportado por Bhatta et al. (2026) en Bangkok, quienes obtuvieron una correlación negativa débil similar ($r = -0.112$), atribuyendo su baja magnitud a los patrones complejos de circulación urbana que reducen la efectividad de la dispersión eólica. En Concepción, esta limitación se acentúa por los vientos predominantemente débiles (1–3 m/s) documentados por Álvarez-Tolentino y Suárez-Salas (2020).

La precipitación tiene una correlación negativa muy débil, lo que señala un efecto de lavado atmosférico poco marcado con el conjunto de los datos analizados. Este comportamiento es consistente con lo reportado por Bhatta et al. (2026) en Bangkok, quienes obtuvieron una correlación negativa similar ($r = -0.260$), confirmando el papel de la precipitación en el arrastre y eliminación del $PM_{2.5}$ de la atmósfera. La menor magnitud encontrada en Concepción se explica por la marcada estacionalidad de las lluvias,

concentradas entre enero y marzo, lo que limita su efecto depurador durante los meses secos de mayor acumulación de PM_{2.5} (junio–septiembre).

La presión atmosférica muestra una correlación positiva más alta dentro del grupo, aunque sigue siendo débil, lo que podría señalar condiciones de estabilidad atmosférica asociadas a acumulación de contaminantes. Este comportamiento es consistente con lo reportado por Bhatta et al. (2026) en Bangkok, quienes obtuvieron la correlación positiva más fuerte entre todas las variables meteorológicas evaluadas ($r = 0.473$), confirmando que los sistemas de alta presión inducen subsidencia atmosférica que suprime la capa límite y reduce la eficiencia de dispersión de contaminantes. La menor magnitud encontrada en Concepción refleja la influencia dominante de la topografía del valle interandino sobre los patrones sinópticos de presión.

4.1.4. Evaluación y Comparación de Modelos de Machine Learning

La Tabla 5 resume las métricas de error y precisión utilizadas para evaluar el desempeño de los tres algoritmos seleccionados: Regresión Lineal (LM), Random Forest (RF) y Redes Neuronales Artificiales (ANN). Esta comparativa permite determinar cuál de los modelos posee la mayor capacidad para estimar las concentraciones de PM_{2.5} en la ciudad de Concepción.

Tabla 5

Comparación del desempeño estadístico de los modelos predictivos de PM_{2.5}

Modelo	MAE	MSE	RMSE	MAPE	R ²
Regresión Lineal	3.349	29.588	5.44	74.488	0.608
Random Forest	3.055	31.436	5.607	55.914	0.583
ANN	2.566	29.349	5.417	36.069	0.611

Los resultados posicionan a la ANN como el modelo más preciso, alcanzando el mayor coeficiente de determinación ($R^2 = 0.611$) y los errores más bajos en casi todas las dimensiones (MAE = 2.566; RMSE = 5.417). Un hallazgo crítico es la reducción drástica del MAPE, donde la ANN logra un 36.06% frente al 74.48% de la Regresión Lineal, lo que indica un ajuste superior en términos porcentuales.

La superioridad observada de la ANN respecto a la Regresión Lineal y Random Forest sugiere que la dinámica del PM_{2.5} en Concepción está influenciada por relaciones no

lineales complejas entre las variables meteorológicas y las concentraciones del contaminante. Este comportamiento coincide con lo reportado por Zhou et al. (2020), quienes encontraron que los modelos basados en redes neuronales poseen una mayor capacidad para representar procesos atmosféricos complejos y generar predicciones más precisas que los enfoques convencionales.

Aunque el modelo Random Forest mostró un desempeño competitivo, presentó errores superiores respecto a la ANN, particularmente en el MAPE. Estudios desarrollados por Kamińska (2020) indican que los modelos basados en árboles de decisión pueden verse limitados ante series temporales altamente dinámicas, especialmente cuando existen cambios abruptos o patrones no lineales complejos en las concentraciones de $PM_{2.5}$.

Por otro lado, la Regresión Lineal presentó el desempeño más limitado, lo que sugiere que las relaciones entre el $PM_{2.5}$ y las variables meteorológicas en Concepción no responden únicamente a comportamientos lineales. Este resultado es consistente con investigaciones realizadas por Romero et al. (2020) en Perú, donde se identificó que la dinámica del material particulado está influenciada por interacciones meteorológicas complejas y procesos atmosféricos no lineales.

En la comparación por métricas, la ANN presenta el mejor desempeño global, al obtener los menores valores en MAE (2.566), MSE (29.349), RMSE (5.417) y MAPE (36.069%), lo que indica una menor desviación absoluta y relativa frente a los otros modelos. En contraste, la Regresión Lineal muestra los mayores errores en la mayoría de las métricas, especialmente en el MAPE (74.488%), mientras que el Random Forest presenta un desempeño intermedio, con errores superiores a la ANN en casi todos los casos. Finalmente, en el R^2 , la ANN alcanza el valor más alto (0.611), lo que confirma una mejor capacidad para explicar la variabilidad del $PM_{2.5}$ en comparación con los otros enfoques. Este comportamiento se alinea con lo señalado por Olarte et al. (2025), quienes destacan que los modelos con menores errores y mayores valores de R^2 presentan una mayor capacidad predictiva y utilidad en la toma de decisiones ambientales.

La Tabla 6 detalla la influencia que cada variable independiente ejerce sobre la predicción del $PM_{2.5}$ según la arquitectura del modelo. Permite observar cómo cada algoritmo jerarquiza la información meteorológica y temporal para generar su estimación final.

Tabla 6*Comparación de pesos de variables en los modelos evaluados*

Variable	Peso_LM	Peso_RF	Peso_ANN
Temp.Out	0.583	0.047	0.158
Out.Hum	4.456	0.057	0.009
Wind.Speed	-0.093	0.077	0.004
Lluvia	-0.101	0.021	0.01
Bar	0.145	0.027	0.002
temp_lag1	-0.806	0.06	0.268
hum_lag1	-5.011	0.055	0.083
wind_lag1	-0.463	0.065	0.077
bar_lag1	-0.152	0.028	0.014
pm25_lag1	13.093	0.564	0.376

En todos los modelos, la variable $PM_{2.5_lag1}$ es la que ostenta el mayor peso (13.093 en LM; 0.564 en RF; 0.376 en ANN). Este comportamiento confirma que la persistencia temporal del $PM_{2.5}$ constituye el principal predictor de las concentraciones futuras, y investigaciones como la de Bhatta et al. (2026), quienes identificaron que la variable $PM_{2.5}$ con retardo de 1 día ($PM_{2.5_lag1}$) constituyó el segundo predictor más importante en sus modelos de Random Forest (20.63%) y Gradient Boosting (17.58%), lo confirman. Sin embargo, se observa una diferencia cualitativa en el uso de la temperatura: mientras que en el RF tiene un peso bajo (0.047), en la ANN la temperatura actual y rezagada suman un peso combinado significativo (0.426), superando incluso la importancia del lag del propio contaminante en términos de activación de red. Este hallazgo es consistente con lo reportado por Zghaid et al. (2025), quienes demostraron que la temperatura ejerce efectos retardados significativos sobre las concentraciones de $PM_{2.5}$, con impactos que se extienden más allá del día inmediato. En esta misma línea, Yang et al. (2021), en un estudio de 284 ciudades chinas, evidenciaron que el $PM_{2.5}$ presenta patrones de efecto rezagado diversos según la variable meteorológica, siendo la temperatura uno de los factores con mayor influencia no lineal y retardada, lo que explica por qué modelos con mayor capacidad para capturar estas relaciones complejas, como la ANN, aprovechan mejor esta información temporal en comparación con el Random Forest.

Aunque el "pasado" es el mejor predictor general, los modelos interpretan el clima de forma distinta. El porqué de la superioridad de la ANN podría explicarse por su mayor atención a la variable térmica (temperatura). En ingeniería ambiental, esto sugiere que la red neuronal detecta mejor los patrones de estabilidad o inversión térmica de Concepción que la regresión o el bosque aleatorio. Resultados similares fueron descritos por Feng et al. (2015), quienes señalaron que las redes neuronales artificiales son capaces de representar dependencias complejas entre las condiciones meteorológicas y la contaminación atmosférica.

La ANN no solo se apoya en la inercia del contaminante, sino que extrae más información del ciclo térmico para ajustar su predicción. Desde la física atmosférica, la temperatura influye directamente sobre la estabilidad de la capa de mezcla y los procesos de dispersión vertical del contaminante. Temperaturas bajas favorecen condiciones de estabilidad atmosférica e inversión térmica, promoviendo la acumulación de $PM_{2.5}$ cerca de la superficie, mientras que temperaturas más elevadas facilitan la convección y dispersión de partículas. Este comportamiento también fue reportado por Tai et al. (2010), quienes encontraron relaciones significativas entre temperatura y concentraciones de partículas finas.

La Tabla 7 presenta los resultados de la prueba de rangos con signo de Wilcoxon, aplicada para determinar si las diferencias de precisión observadas en la Tabla 6 son estadísticamente significativas. Este paso es el rigor final que valida si un modelo es verdaderamente mejor que otro o si el resultado fue producto del azar.² Descripción de los resultados: Todas las comparaciones realizadas (RF vs LM, RF vs ANN, y LM vs ANN) arrojaron un p-valor < 0.001 en todas las métricas evaluadas (MAE, MSE, RMSE y R^2). Esto indica que existe una diferencia estadística contundente entre los tres modelos, sin margen de ambigüedad.

Tabla 7

Comparación estadística del desempeño de modelos (Prueba de Wilcoxon)

Nro.	Comparación	p_value	MAE	MSE	RMSE	R^2
1	RF vs LM	<0.001	<0.001	<0.001	<0.001	<0.001
2	RF vs ANN	<0.001	<0.001	<0.001	<0.001	<0.001
3	LM vs ANN	<0.001	<0.001	<0.001	<0.001	<0.001

Estos hallazgos coinciden con investigaciones desarrolladas en Perú sobre modelamiento predictivo de $PM_{2.5}$. Romero et al. (2020) señalaron que las relaciones entre variables meteorológicas y contaminantes atmosféricos en ciudades peruanas presentan comportamientos altamente no lineales, lo que incrementa la complejidad de su modelamiento y favorece el uso de técnicas capaces de capturar patrones complejos. En este contexto, las redes neuronales artificiales poseen una mayor capacidad para representar estas interacciones, lo que puede explicar el mejor desempeño observado en la predicción de las concentraciones de $PM_{2.5}$.

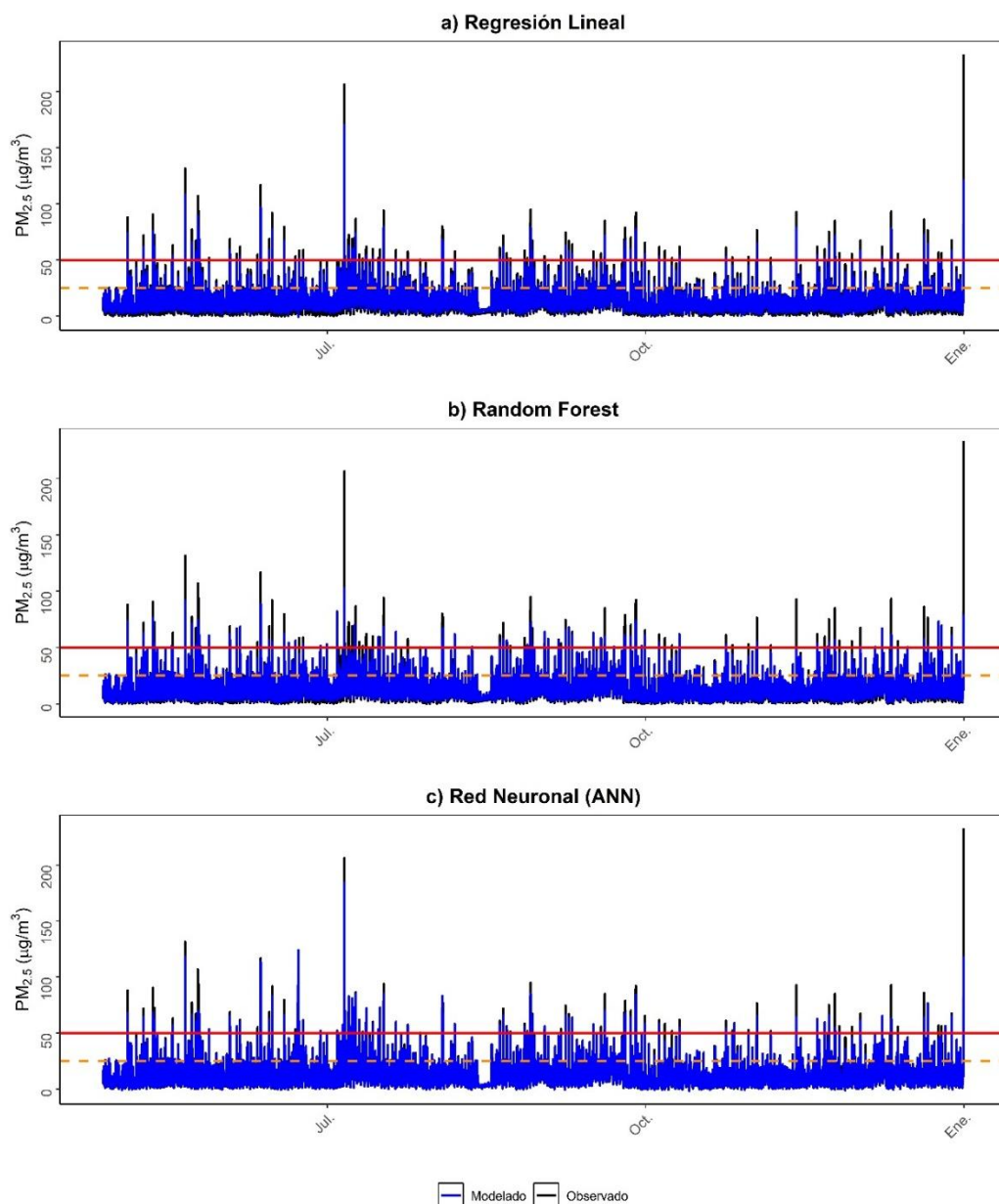
El porqué de estos valores tan bajos (altamente significativos) se debe a la consistencia de los errores de predicción a lo largo del periodo 2023-2025. Al haber utilizado una base de datos robusta, la prueba detecta que las estructuras de los algoritmos son fundamentalmente diferentes en su capacidad de respuesta. Esto permite afirmar con seguridad científica que la Red Neuronal no es solo numéricamente mejor, sino que es un sistema predictivo cualitativamente superior para la realidad atmosférica de Concepción.

Desde una perspectiva de ingeniería ambiental, estos resultados validan el uso de modelos basados en redes neuronales artificiales como herramientas de apoyo para sistemas de alerta temprana y gestión de calidad del aire. Estudios recientes desarrollados en Perú resaltan que el uso de técnicas de Machine Learning permite mejorar la predicción de episodios críticos de contaminación y optimizar estrategias de vigilancia ambiental en ciudades urbanas y andinas. En este contexto, Olarte et al. (2025) aplicaron modelos de regresión lasso, ridge y elastic net para predecir $PM_{2.5}$ en Huachac, Junín ciudad andina a 3,275 msnm obteniendo un R^2 de 0.148 al utilizar únicamente temperatura y humedad como variables predictoras. Este limitado poder explicativa evidencia que dos variables meteorológicas contemporáneas son insuficientes para capturar la dinámica del contaminante, limitación que la presente investigación supera mediante la incorporación de variables meteorológicas adicionales con retardo temporal, mejorando sustancialmente la capacidad predictiva del modelo.

La Figura 13 presenta la validación visual del desempeño de los tres modelos desarrollados: Regresión Lineal (LM), Random Forest (RF) y Red Neuronal Artificial (ANN). El gráfico superpone las concentraciones de $PM_{2.5}$ observadas (reales) frente a las estimadas por cada algoritmo, permitiendo evaluar intuitivamente la precisión de las predicciones a lo largo del periodo de prueba y la capacidad de cada técnica para seguir la tendencia de los datos.

Figura 13

Comparación visual de valores observados vs. valores modelados de $PM_{2.5}$ para las tres técnicas de Machine Learning.



En el gráfico se observa que la línea correspondiente a la Red Neuronal (ANN) es la que sigue con mayor fidelidad la trayectoria de los datos observados, capturando con mayor precisión tanto los valores base como las fluctuaciones repentinas. Por el contrario, la Regresión Lineal muestra una tendencia a sobreestimar los valores bajos y subestimar los picos de contaminación, evidenciando una "rigidez" en su ajuste. El Random Forest presenta

un comportamiento intermedio, logrando un buen ajuste general, pero mostrando mayor dispersión en los valores extremos en comparación con la ANN.

Esta comparación visual es la confirmación empírica de que la ANN posee la mayor fidelidad de respuesta ante la dinámica del aire en Concepción. El porqué de este resultado reside en la capacidad de la red neuronal para modelar la complejidad no lineal del fenómeno. Mientras que la regresión lineal intenta forzar una línea recta que no logra adaptarse a los cambios bruscos de la meteorología y la inercia del contaminante, la ANN, gracias a sus capas ocultas y funciones de activación, logra "aprender" y replicar las interacciones complejas, resultando en un modelo mucho más plástico y preciso que se ajusta a la realidad observada.

4.2. Prueba de Hipótesis

4.2.1. Hipótesis general

H_0 (Hipótesis nula): No existen diferencias significativas en el desempeño predictivo entre los modelos de Machine Learning (Regresión Lineal, Random Forest y Redes Neuronales Artificiales) en la estimación de las concentraciones de $PM_{2.5}$ en la ciudad de Concepción durante el periodo 2023–2025.

H_1 (Hipótesis alterna): Existen diferencias significativas en el desempeño predictivo entre los modelos de Machine Learning, siendo uno de ellos superior en la estimación de las concentraciones de $PM_{2.5}$.

4.2.2. Nivel de significancia

Se adoptó un nivel de significancia de $\alpha = 0.05$.

4.2.3. Prueba estadística aplicada

Para contrastar la hipótesis se utilizó la prueba no paramétrica de rangos con signo de Wilcoxon, debido a que permite comparar el desempeño de modelos predictivos sin asumir normalidad en los errores.

4.2.4. Resultados de la prueba

De acuerdo con la Tabla 11, todas las comparaciones entre modelos (Regresión Lineal vs Random Forest, Random Forest vs ANN y Regresión Lineal vs ANN) presentan valores de $p < 0.001$ en todas las métricas evaluadas (MAE, MSE, RMSE y R^2).

4.2.5. Decisión estadística

Dado que los valores de p son menores que el nivel de significancia ($p < 0.05$), se rechaza la hipótesis nula (H_0).

4.2.6. Interpretación

Los resultados evidencian que existen diferencias estadísticamente significativas entre los modelos evaluados. En particular, la Red Neuronal Artificial (ANN) presenta el mejor desempeño predictivo, al registrar los menores errores ($MAE = 2.566$; $RMSE = 5.417$) y el mayor coeficiente de determinación ($R^2 = 0.611$), superando tanto a la Regresión Lineal como al Random Forest.

Por lo tanto, se rechaza la hipótesis planteada inicialmente, la cual proponía que el modelo Random Forest sería el de mejor desempeño, y se concluye que el modelo más adecuado para estimar las concentraciones de $PM_{2.5}$ en la ciudad de Concepción es la Red Neuronal Artificial.

CONCLUSIONES

En relación con el primer objetivo específico, se determinó que las condiciones meteorológicas en Concepción durante el periodo 2023–2025 se caracterizan por una alta estabilidad atmosférica, con velocidades de viento predominantemente inferiores a 1 m/s, más del 70% de los registros en el rango de 0.2 a 0.5 m/s, y una marcada estacionalidad de la precipitación concentrada entre enero y marzo. Esta combinación de calmas eólicas y ausencia de lluvias durante los meses de junio a agosto configura un entorno intrínsecamente desfavorable para la dispersión de contaminantes, convirtiendo a la ciudad en una cuenca receptora con baja capacidad de autodepuración eólica e hídrica.

Respecto al segundo objetivo específico, se evidenció que las concentraciones de $PM_{2.5}$ presentan una alta variabilidad temporal, con una media multianual de $14.08 \mu\text{g}/\text{m}^3$ y una desviación estándar casi igual a la media ($14.73 \mu\text{g}/\text{m}^3$), lo que refleja la presencia de episodios extremos que superan ampliamente el valor promedio, llegando hasta $304.60 \mu\text{g}/\text{m}^3$. Los incrementos más significativos ocurren durante los meses secos de junio a septiembre, con un pico crítico en septiembre de 2024 donde el 73.33% de los días superaron la guía de la OMS de $25 \mu\text{g}/\text{m}^3$. Asimismo, se identificó un patrón horario bimodal, con picos entre las 07:00–09:00 horas y concentraciones críticas durante la noche, asociados a las emisiones del parque automotor y a la compresión de la capa de mezcla atmosférica nocturna.

En cumplimiento del tercer objetivo específico, se validó el desempeño de los tres modelos de Machine Learning mediante métricas estadísticas (MAE, MSE, RMSE, MAPE y R^2). La prueba no paramétrica de Wilcoxon confirmó con un $p < 0.001$, muy por debajo del nivel de significancia $\alpha = 0.05$, que las diferencias entre los modelos son estadísticamente significativas con un 99.9% de confianza, descartando que los resultados sean producto del azar y garantizando la confiabilidad científica de las comparaciones realizadas.

En relación con el cuarto objetivo específico, se identificó que la Red Neuronal Artificial (ANN) presenta el mejor desempeño predictivo en la estimación de las concentraciones de $PM_{2.5}$, superando a la Regresión Lineal y al Random Forest en todas las métricas evaluadas (MAE = 2.566; RMSE = 5.417; MAPE = 36.07%; $R^2 = 0.611$). Este resultado se atribuye a su capacidad para modelar relaciones no lineales y complejas entre las variables meteorológicas y el contaminante, distribuyendo su atención entre múltiples

variables, a diferencia del Random Forest que concentró el 56.4% de su importancia en una sola variable. Sin embargo, el $R^2 = 0.611$ indica que el 39% de la variabilidad del $PM_{2.5}$ no es explicada por las variables meteorológicas consideradas, sugiriendo la influencia de factores adicionales como tráfico vehicular, quema de biomasa y actividades agrícolas.

En función del objetivo general, se rechaza la hipótesis inicial que establecía al Random Forest como modelo de mejor desempeño, siendo la Red Neuronal Artificial la herramienta cualitativamente superior para representar la realidad atmosférica de Concepción ($R^2 = 0.611$; RMSE = 5.417; MAE = 2.566), validada estadísticamente mediante la prueba de Wilcoxon ($p < 0.001$), confirmando que existen diferencias significativas entre los modelos evaluados.

Los resultados evidencian que la calidad del aire en Concepción está fuertemente gobernada por la inercia atmosférica, siendo la variable `pm25_lag1` el predictor más influyente ($\rho = 0.917$, $p < 0.001$), confirmando que las partículas $PM_{2.5}$ permanecen suspendidas formando una nube persistente que condiciona el estado atmosférico siguiente. La combinación de variables meteorológicas con retardo temporal y modelos de aprendizaje automático constituye una herramienta robusta para la predicción de la calidad del aire en zonas andinas.

RECOMENDACIONES

- Gestionar ante las autoridades ambientales la continuidad y mejora en la calidad del registro de datos de PM_{2.5} y variables meteorológicas.
- Incluir variables con retardo temporal, especialmente pm25_lag1, en futuros estudios y sistemas de predicción de calidad del aire.
- Ampliar el conjunto de variables de entrada en futuras investigaciones, integrando emisiones vehiculares, quemas agrícolas y variables topográficas, para mejorar la precisión de los modelos en episodios extremos.
- Fortalecer las estrategias de control de emisiones durante los periodos de mayor vulnerabilidad ambiental, mediante la regulación de quemas agrícolas, control del parque automotor y promoción de tecnologías limpias.
- Evaluar modelos de aprendizaje profundo como LSTM en futuras investigaciones, con el fin de mejorar la captura de la dependencia temporal del PM_{2.5} especialmente en episodios críticos de contaminación.
- Promover en la comunidad académica de la Escuela Profesional de Ingeniería Ambiental de la UNISCJSA la incorporación de herramientas de inteligencia artificial y aprendizaje automático en investigaciones ambientales.

REFERENCIAS BIBLIOGRÁFICAS

- Álvarez-Tolentino, D., & Suárez-Salas, L. (2020). Aporte cuantitativo de las fuentes de PM₁₀ y PM_{2.5} en sitios urbanos del Valle del Mantaro, Perú. *Revista Internacional de Contaminación Ambiental*. <https://doi.org/10.20937/RICA.53473>
- Angeles Suazo, J. M., Angeles Vasquez, R., Chavarría Márquez, E. Y., Lavado Meza, C., Huamani Astocaza, L. L., & Cusiche Huamani, M. L. (2024). Efecto del material particulado y humedad de suelo en Huancayo y Pampas/Perú. *Revista Alfa*, 8(24), 1001-1012. <https://doi.org/10.33996/revistaalfa.v8i24.319>
- Arroyo-Aliaga, J. (2011). Tendencias y cambio del clima del valle del Mantaro mediante los análisis de índices de Precipitación Efectiva y Temperatura Eficiente. *Apuntes de Ciencia & Sociedad*, 1(1), 45–54. <https://doi.org/10.18259/acs.2011009>
- Awos, A., Thompson, S., Adedeji, O., Zvomuya, F., & Zhang, Q. (2024). Monitoring of Air Quality for Particulate Matter (PM_{2.5}, PM₁₀) and Heavy Metals Proximate to a Cement Factory in Ewekoro, Nigeria. *Journal of Geoscience and Environment Protection*, 12(10), 152-180. <https://doi.org/10.4236/gep.2024.1210009>
- Basoa, K., Fleming, Z. L., Leiva, M. A., Concha, C., & Menares, C. (2025). Current Status, Trends, and Future Directions in Chilean Air Quality: A Data-Driven Perspective. *Atmosphere*, 16(6), 733. <https://doi.org/10.3390/atmos16060733>
- Bhatt, J., Acharya, S. R., & Yang, K. M. (2026). Enhanced air quality forecasting and trend analysis using machine learning: A comprehensive evaluation of five years of PM_{2.5} concentrations in Bangkok, Thailand. *Environmental Challenges*, 22, 101442. <https://doi.org/10.1016/j.envc.2026.101442>
- Bimenyimana, E., Sciare, J., Oikonomou, K., Iakovides, M., Pikridas, M., Vasiliadou, E., Savvides, C., & Mihalopoulos, N. (2025). Cross-validation of methods for quantifying the contribution of local (urban) and regional sources to PM_{2.5} pollution: Application in the Eastern Mediterranean (Cyprus). *Atmospheric Environment*, 343, 120975. <https://doi.org/10.1016/j.atmosenv.2024.120975>
- Bonilla-Granados, C. A., Sánchez-Delgado, A. Y., Rubio-Gómez, Y., & Cortéz-Huerta, M. (2023). Niveles de concentración por PM_{2.5} mediante sensores de bajo costo. Caso de estudio: Pamplona, Colombia. *Revista UIS Ingenierías*, 22(3). <https://doi.org/10.18273/revuin.v22n3-2023003>

- Cabello-Torres, R., Reyes, F., y colaboradores. (2024). An exploratory analysis of $PM_{2.5}/PM_{10}$ ratio during spring 2016–2018 in Metropolitan Lima. *Scientific Reports*, 14, 10234. <https://doi.org/10.1038/s41598-024-59831-9>
- Campaña Tene, R. C., & Pérez Fernández, E. (2025). Análisis del impacto ambiental del material particulado PM_{10} y $PM_{2.5}$ en zonas urbanas e industriales: Propuesta de tecnologías para su monitoreo y control. *Código Científico Revista de Investigación*, 6(E1), 576-595. <https://doi.org/10.55813/gaea/ccri/v6/nE1/703>
- Carranza Noriega, R. (2004). Gestión de la calidad del aire Causas, Efectos y soluciones. Instituto Investigación Ingeniería Industrial. https://sisbib.unmsm.edu.pe/bibvirtualdata/libros/mambiente/gest_cal/pdf/t_completo.pdf
- Ccasani Guillén, J. C. (2019). Ciclo diurno y ciclo anual de los vientos superficiales observados en la zona central oeste del Valle del Mantaro, Perú [Tesis de licenciatura, Universidad Nacional Mayor de San Marcos]. Repositorio Cybertesis UNMSM. <https://hdl.handle.net/20.500.12672/10814>
- Choquetico Apaza, H. A., & Silva Rubio, E. (2025). Modelo de regresión funcional para la predicción de $PM_{2.5}$ en función de PM_{10} en Lima—Perú. *Revista Científica: BIOTECH AND ENGINEERING*, 4(3). <https://doi.org/10.52248/eb.Vol4Iss3.135>
- Chowdhury, S., Pillarisetti, A., Oberholzer, A., Jetter, J., Mitchell, J., Cappuccilli, E., Aamaas, B., Aunan, K., Pozzer, A., & Alexander, D. (2023). A global review of the state of the evidence of household air pollution's contribution to ambient fine particulate matter and their related health impacts. *Environment International*, 173, 107835. <https://doi.org/10.1016/j.envint.2023.107835>
- Contento Patiño, L. J., & Ortiz González, C. A. (2024). Predicción de la contaminación atmosférica por partículas ($PM_{2.5}$ y SO_2) mediante un modelo de regresión de cretas. *Universidad Libre - Facultad de ingeniería*, 13(23), 65-74.
- Correa-Ochoa, M. A., Gómez-Marín, M., Patiño-López, K. V., Aguiar, D., & Franco, S. A. (2025). Characteristic Chemical Profile of Particulate Matter ($PM_{2.5}$)—A Comparative Study Between Two Periods, Case Study in Medellín, Colombia. *Sustainability*, 17(12), 5380. <https://doi.org/10.3390/su17125380>

- Feng, X., Li, Q., Zhu, Y., Hou, J., Jin, L., & Wang, J. (2015). Artificial neural networks forecasting of PM_{2.5} pollution using air mass trajectory based geographic model and wavelet transformation. *Atmospheric Environment*, 107, 118–128. <https://doi.org/10.1016/j.atmosenv.2015.02.030>
- Flores-Rojas, J. L., Moya-Álvarez, A. S., Kumar, S., Martínez-Castro, D., Villalobos-Puma, E., & Silva-Vidal, Y. (2019). Análisis de posibles mecanismos desencadenantes de tormentas severas en los Andes centrales tropicales del Perú, Valle del Mantaro. *Atmosphere*, 10(6), 301. <https://doi.org/10.3390/atmos10060301>
- Hayes, R. B., Lim, C., Zhang, Y., Cromar, K., Shao, Y., Reynolds, H. R., Silverman, D. T., Jones, R. R., Park, Y., Jerrett, M., Ahn, J., & Thurston, G. D. (2020). PM_{2.5} air pollution and cause-specific cardiovascular disease mortality. *International Journal of Epidemiology*, 49(1), 25-35. <https://doi.org/10.1093/ije/dyz114>
- Hernández Sampieri, R., Fernandez-Collado, C. F., & Baptista Lucio, P. (2014). Metodología de la investigación (Sexta edición). McGraw-Hill Education. https://apiperiodico.jalisco.gob.mx/api/sites/periodicooficial.jalisco.gob.mx/files/metodologia_de_la_investigacion_-_roberto_hernandez_sampieri.pdf
- Instituto Geofísico del Perú. (2020). Dispersión de material particulado 2.5 emitido por pollerías usando modelos de dispersión de contaminantes atmosféricos en Huancayo, Junín. Repositorio IGP.
- Jia, N., Li, Y., Chen, R., & Yang, H. (2023). A Review of Global PM_{2.5} Exposure Research Trends from 1992 to 2022. *Sustainability*, 15(13), 10509. <https://doi.org/10.3390/su151310509>
- Joshi, D. C., Negi, P., Devi, S., Lohani, H., Kumar, R., Gupta, M., & Ming, L. C. (2025). Fine particulate matter (PM_{2.5}, PM₁₀): A silent catalyst for chronic lung diseases in India; a comprehensive review. *Environmental Challenges*, 20, 101215. <https://doi.org/10.1016/j.envc.2025.101215>
- Kunjir, G. M., Tikle, S., Das, S., Karim, M., Roy, S. K., & Chatterjee, U. (2025). Assessing particulate matter (PM_{2.5}) concentrations and variability across Maharashtra using satellite data and machine learning techniques. *Discover Sustainability*, 6(1), 238. <https://doi.org/10.1007/s43621-025-01082-3>

- Kamińska, J. A., Sciavicco, G., Lucena-Sánchez, E., & Jiménez, F. (2020). Lag variables in air pollution modeling based on traffic flow and meteorological factors. *Proceedings*, 51(1), 1. <https://doi.org/10.3390/proceedings2020051001>
- Lin, Y., Zou, J., Yang, W., & Li, C.-Q. (2018). A Review of Recent Advances in Research on PM_{2.5} in China. *International Journal of Environmental Research and Public Health*, 15(3), 438. <https://doi.org/10.3390/ijerph15030438>
- Makhdoomi, A., Sarkhosh, M., & Ziaei, S. (2025). PM_{2.5} concentration prediction using machine learning algorithms: An approach to virtual monitoring stations. *Scientific Reports*, 15(1), 8076. <https://doi.org/10.1038/s41598-025-92019-3>
- Martínez-Soto, A., Jentsch, M. F., Martínez-Gallegos, V., Duchêne, J., y Zipf, A. (2025). Influence of meteorological variables on PM_{2.5} concentrations in cities heated with solid fuels: A case study from Temuco, Chile. *Atmospheric Environment: X*, 25, 100401. <https://www.sciencedirect.com/science/article/pii/S2590162125000917>
- Ministerio del Ambiente (MINAM). (2018, 22 de enero). Resolución Ministerial N.º 020-2018-MINAM: Aprueban los Lineamientos para la determinación de las Zonas de Atención Prioritaria. Gobierno del Perú. <https://www.gob.pe/institucion/minam/normas-legales/4347-20-2018-minam>
- Ministerio del Ambiente (MINAM). (s. f.). Zonas priorizadas de calidad del aire. Dirección General de Calidad Ambiental. <https://www.minam.gob.pe/calidadambiental/zonas-priorizadas-de-calidad-del-aire/>
- Moretti-Villegas, L. F., Tafur-Anzualdo, V. I., & Valiente-Saldaña, Y. M. (2023). Contaminación del aire en la ciudad de Lima, Perú. *Revista Arbitrada Interdisciplinaria Koinonía*, 8(2), 822-830. <https://doi.org/10.35381/r.k.v8i2.2975>
- Mozaffar, A., Zhang, Y.-L., Hong, Y., Fan, M.-Y., & Lin, Y.-C. (2026). A machine learning and box modeling approach to comparing the atmospheric chemistry of high- and low-ozone and PM_{2.5} episodes. *Atmospheric Research*, 327, 108373. <https://doi.org/10.1016/j.atmosres.2025.108373>
- Nakyai, T., Santasnachok, M., Thetkathuek, A., & Phatrabuddha, N. (2025). Influence of meteorological factors on air pollution and health risks: A comparative analysis of industrial and urban areas in Chonburi Province, Thailand. *Environmental Advances*, 19, 100608. <https://doi.org/10.1016/j.envadv.2024.100608>

- Nguyen, P. H., Dao, N. K., & Nguyen, L. S. P. (2024). Development of Machine Learning and Deep Learning Prediction Models for PM_{2.5} in Ho Chi Minh City, Vietnam. *Atmosphere*, 15(10), 1163. <https://doi.org/10.3390/atmos15101163>
- Oblitas Mantilla, J. A., & Escobedo Cárdenas, E. J. (2024). Prediction of PM_{2.5} and PM₁₀ Concentrations Using XGBoost and LightGBM Algorithms: A Case Study in Lima, Peru. *Interfases*, 020, 183-206. <https://doi.org/10.26439/interfases2024.n020.7417>
- Olarte, E., Gutierrez, J., Roque, G., Soria, J. J., Fernandez, H., Carpio, J. E. P., & Poma, O. (2025). Predictive Model with Machine Learning for Environmental Variables and PM_{2.5} in Huachac, Junín, Perú. *Atmosphere*, 16(3), 323. <https://doi.org/10.3390/atmos16030323>
- Özüpak, Y., Alpsalaz, F., & Aslan, E. (2025). Air Quality Forecasting Using Machine Learning: Comparative Analysis and Ensemble Strategies for Enhanced Prediction. *Water, Air, & Soil Pollution*, 236(7), 464. <https://doi.org/10.1007/s11270-025-08122-8>
- Prado Fasanando, J. J. (2021). Caracterización de la capa límite atmosférica en el Valle del Mantaro empleando radiosondas y sensores remotos [Tesis de pregrado, Universidad Nacional José Faustino Sánchez Carrión]. Repositorio Institucional
- Quijano Parra, A., Quijano Vargas, M. J., & Henao Martínez, J. A. (2010). Caracterización fisicoquímica del material particulado fracción respirable PM_{2.5} en Pamplona-Norte de Santander-Colombia. *Bistua: Revista de la Facultad de Ciencias Básicas*, 8(1), 1-20.
- Rad, A. K., Razmi, S.-O., Nematollahi, M. J., Naghipour, A., Golkar, F., & Mahmoudi, M. (2024). Machine learning models for predicting interactions between air pollutants in Tehran Megacity, Iran. *Alexandria Engineering Journal*, 104, 464-479. <https://doi.org/10.1016/j.aej.2024.08.023>
- Romero, J., Moya-Álvarez, A., Ccoyllo, O., Silva, Y., y Martínez-Castro, D. (2020). Temporal and spatial analysis of traffic-related pollutant under the influence of the seasonality and meteorological variables over an urban city in Peru. *Heliyon*, 6(6), e04029. <https://doi.org/10.1016/j.heliyon.2020.e04029>
- Sánchez Meza, A. L. (2022). Influencia de características antrópicas de tres zonas de la provincia de Coronel Portillo en la concentración de material particulado PM₁₀ y

- PM_{2.5}, Ucayali, Perú [Tesis Pregrado, Universidad Nacional de Ucayali].
<https://apirepositorio.unu.edu.pe/server/api/core/bitstreams/e7d76b67-e0f7-4452-b1a6-9d2bf120374e/content>
- Shakya, D., Deshpande, V., Goyal, M. K., & Agarwal, M. (2023). PM_{2.5} air pollution prediction through deep learning using meteorological, vehicular, and emission data: A case study of New Delhi, India. *Journal of Cleaner Production*, 427, 139278. <https://doi.org/10.1016/j.jclepro.2023.139278>
- Shu, X., Cao, J., Liu, Q., Wang, Y., Jiang, F., Wu, C., & Shu, J. (2024). Global Trends and Hotspots in the Research of the Effects of PM_{2.5} on Asthma: A Bibliometric and Visualized Analysis. *Journal of Epidemiology and Global Health*, 14(4), 1720-1736. <https://doi.org/10.1007/s44197-024-00331-4>
- Tai, A. P. K., Mickley, L. J., & Jacob, D. J. (2010). Correlation between fine particulate matter (PM_{2.5}) and meteorological variables in the United States: Implications for the sensitivity of PM_{2.5} to climate change. *Atmospheric Environment*, 44(32), 3976–3984. <https://doi.org/10.1016/j.atmosenv.2010.06.060>
- Thangavel, P., Park, D., & Lee, Y.-C. (2022). Recent Insights into Particulate Matter (PM_{2.5})-Mediated Toxicity in Humans: An Overview. *International Journal of Environmental Research and Public Health*, 19(12), 7511. <https://doi.org/10.3390/ijerph19127511>
- U.S. Environmental Protection Agency. (2026). *Las islas de calor*. <https://espanol.epa.gov/la-energia-y-el-medioambiente/las-islas-de-calor>
- Vásquez Arenas, J. P. (2022). Beneficios de los modelos basados en algoritmos de machine learning para la predicción de concentraciones de PM_{2.5} en el Valle de Aburrá [Monografía, Universidad de Antioquia]. <https://bibliotecadigital.udea.edu.co/server/api/core/bitstreams/e31a1faf-b2d7-4798-a5f3-c33a649b89f8/content>
- Vu, B. N., Sánchez, O., Bi, J., Xiao, Q., Hansel, N. N., Checkley, W., Gonzales, G. F., Steenland, K., & Liu, Y. (2019). Developing an Advanced PM_{2.5} Exposure Model in Lima, Peru. *Remote Sensing*, 11(6), 641. <https://doi.org/10.3390/rs11060641>
- Vidal-Daza, O. A., & Pérez-Vidal, A. (2018). Estimación de la dispersión de contaminantes atmosféricos emitidos por una industria papelera mediante el modelo AERMOD.

- *Ingeniería*, 23(1), 31-47.
https://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0121-750X2018000100031
- Villalobos, E. E., Martinez-Castro, D., Kumar, S., Silva, Y., & Fashe, O. (2019). Estudio de tormentas convectivas sobre los Andes Centrales del Perú usando los radares PR-TRMM y KuPR-GPM. *Revista Cubana de Meteorología*, 25(1), 59-75.
- Villalobos-Puma, E., Suarez, L., Gillardoni, S., Zubieta, R., Martinez-Castro, D., Miranda-Corzo, A., Bonasoni, P., & Silva, Y. (2024). Atmospheric black carbon observations and its valley-mountain dynamics: Eastern cordillera of the central Andes of Peru. *Environmental Pollution*, 355, 124089.
<https://doi.org/10.1016/j.envpol.2024.124089>
- Warthon, E., Zúñiga, L., y colaboradores. (2023). Correlación de material particulado PM_{2.5} en función a la humedad relativa y precipitación usando Python, en la ciudad del Cusco. *arXiv*. <https://doi.org/10.48550/arXiv.2303.00526>
- World Health Organization. (2021). WHO global air quality guidelines: Particulate matter (PM_{2.5} and PM₁₀), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. World Health Organization.
<https://www.who.int/publications/i/item/9789240034228>
- Yang, Z., Yang, J., Li, M., Chen, J., & Ou, C.-Q. (2021). Non-linear and lagged meteorological effects on daily PM_{2.5} and O₃ levels in the ambient air: Evidence from 284 Chinese cities. *Journal of Cleaner Production*, 278, 123931.
<https://doi.org/10.1016/j.jclepro.2020.123931>
- Yin, L., Bai, B., Zhang, B., Zhu, Q., Di, Q., Requía, W. J., Schwartz, J. D., Shi, L., & Liu, P. (2025). Regional-specific trends of PM_{2.5} and O₃ temperature sensitivity in the United States. *Npj Climate and Atmospheric Science*, 8(1), 12.
<https://doi.org/10.1038/s41612-024-00862-4>
- Zamorategui Molina, A., Delgado Galván, X., Villalobos Pérez, S., Alcántara Concepción, P. C., Esquivel Longoria, M. I., & De La Peña Torre, A. (2024). Effects of meteorological variables and dry seasonal months on PM_{2.5} concentration in non-industrial cities: The case of Guanajuato city. *Acta Universitaria*, 34, 1-22.
<https://doi.org/10.15174/au.2024.4044>

- Zender-Świercz, E., Galiszewska, B., Telejko, M., & Starzomska, M. (2024). The effect of temperature and humidity of air on the concentration of particulate matter - PM_{2.5} and PM₁₀. *Atmospheric Research*, 312, 107733. <https://doi.org/10.1016/j.atmosres.2024.107733>
- Zghaid, M., Benchrif, A., Tahri, M., Arfaoui, A., Elouardi, M., Derdaki, M., Quyou, A., & Ouahidi, M. L. (2025). Assessment of air pollution and lagged meteorological effects in an urban residential area of Kenitra City, Morocco. *Atmosphere*, 16(1), 96. <https://doi.org/10.3390/atmos16010096>
- Zhang, L., Cheng, Y., Zhang, Y., He, Y., Gu, Z., & Yu, C. (2017). Impact of air humidity fluctuation on the rise of PM mass concentration based on the high-resolution monitoring data. *Aerosol and Air Quality Research*, 17(2), 543-552.
- Zhang, Q., Meng, X., Shi, S., Kan, L., Chen, R., & Kan, H. (2022). Overview of particulate air pollution and human health in China: Evidence, challenges, and opportunities. *The Innovation*, 3(6), 100312. <https://doi.org/10.1016/j.xinn.2022.100312>
- Zhou, Y., Chang, L.-C., & Chang, F.-J. (2020). Explore a multivariate Bayesian uncertainty processor driven by artificial neural networks for probabilistic PM_{2.5} forecasting. *Science of the Total Environment*, 711, 134792. <https://doi.org/10.1016/j.scitotenv.2019.134792>
- Zhou, S., Wang, W., Zhu, L., Qiao, Q., & Kang, Y. (2024). Deep-learning architecture for PM_{2.5} concentration prediction: A review. *Environmental Science and Ecotechnology*, 21, 100400. <https://doi.org/10.1016/j.es.2024.100400>

ANEXOS

Anexo 1. Matriz de consistencia

FORMULACIÓN DEL PROBLEMA	OBJETIVOS	HIPÓTESIS	VARIABLES, DIMENSIONES E INDICADORES	METODOLOGÍA DE LA INVESTIGACIÓN
Problema General	Objetivos General	Hipótesis General	VARIABLE INDEPENDIENTE: Parámetros meteorológicos Dimensiones: - Temperatura del aire - Humedad Relativa - Velocidad del viento - Dirección del viento - Precipitación - Presión atmosférica Indicadores: (°C) (%) (m s⁻¹) ° (grados) (mm) (mmHg)	a) Tipo, nivel y diseño de investigación: Investigación aplicada, nivel explicativo-predictivo, diseño no experimental, longitudinal y predictivo. b) Población y muestra: Registros meteorológicos horarios y diarios de 2023 a 2025. c) Técnicas e instrumentos: Recolección de datos mediante descarga en formato CSV/Excel desde una estación meteorológica instalada en la ciudad de Concepción, equipada con un sistema Davis Vantage Pro y un sensor
¿Cuál de los modelos de Machine Learning, permite estimar con mayor precisión las concentraciones del material particulado fino (PM _{2.5}) en la ciudad de Concepción, durante el periodo 2023-2025?	Estimar las concentraciones de Material Particulado fino (PM _{2.5}) mediante modelos de Machine Learning en concepción durante el periodo 2023-2025.	El modelo Random Forest presentará el mejor desempeño predictivo en la estimación de las concentraciones de (PM _{2.5}) en la ciudad de Concepción durante el periodo 2023–2025.		
Problemas Específicos	Objetivos Específicos			
¿Cuál es la variabilidad de las condiciones meteorológicas predominantes en la ciudad de Concepción durante los años 2023-2025?	Determinar la variabilidad de las condiciones meteorológicas predominantes en la ciudad Concepción durante el periodo 2023-2025.			
¿Cómo varían las concentraciones del material particulado fino (PM _{2.5}) en la	Determinar la variabilidad de las concentraciones del (PM _{2.5}) en la ciudad de Concepción, 2023- 2025.			

ciudad de Concepción durante los años 2023-2025?			VARIABLE DEPENDIENTE: PM_{2.5} Dimensiones: - Nivel horario - Nivel diario - Nivel mensual Indicadores: - Media horaria de PM_{2.5} - Media diaria de PM_{2.5} - Media mensual de PM_{2.5}	PurpleAir para la medición de PM _{2.5} , registrando la información en una ficha diseñada para el estudio. d) Análisis de datos: Procesamiento estadístico y modelado predictivo con algoritmos de Machine Learning: Random Forest (RF), Regresión y Red Neuronal Artificial (ANN).
¿Cómo validar el desempeño predictivo de las técnicas de Machine Learning en la estimación de las concentraciones de (PM _{2.5}) en la ciudad de Concepción durante el periodo 2023–2025?	Validar el desempeño predictivo de las Técnicas de Machine Learning en la estimación de las concentraciones de (PM _{2.5}) en la ciudad de Concepción durante el periodo de 2023-2025.			
¿Cuál de los modelos de Machine Learning (Random Forest, Regresión y Red Neuronal Artificial) logra un mejor desempeño predictivo en la estimación de las concentraciones de (PM _{2.5}) en la ciudad de Concepción durante el periodo 2023-2025?	Identificar el mejor desempeño predictivo de las Técnicas de Machine Learning en la estimación de las concentraciones de (PM _{2.5}) en la ciudad de Concepción durante el periodo de 2023-2025			

Anexo 2. *Instrumentos de recolección de datos*

FICHA DE RECOLECCIÓN DE DATOS – VARIABLES METEOROLÓGICAS

Fecha	Hora	Temperatura (°C)	Humedad Relativa (%)	Velocidad del Viento (m/s)	Dirección del Viento (°)	Presión Atmosférica (mmHg)	Precipitación (mm)	Observaciones

Nota: La tabla se adaptará a un Excel

Anexo 3. Panel fotográfico

Figura 14

Estación meteorológica empleada para el registro de variables ambientales



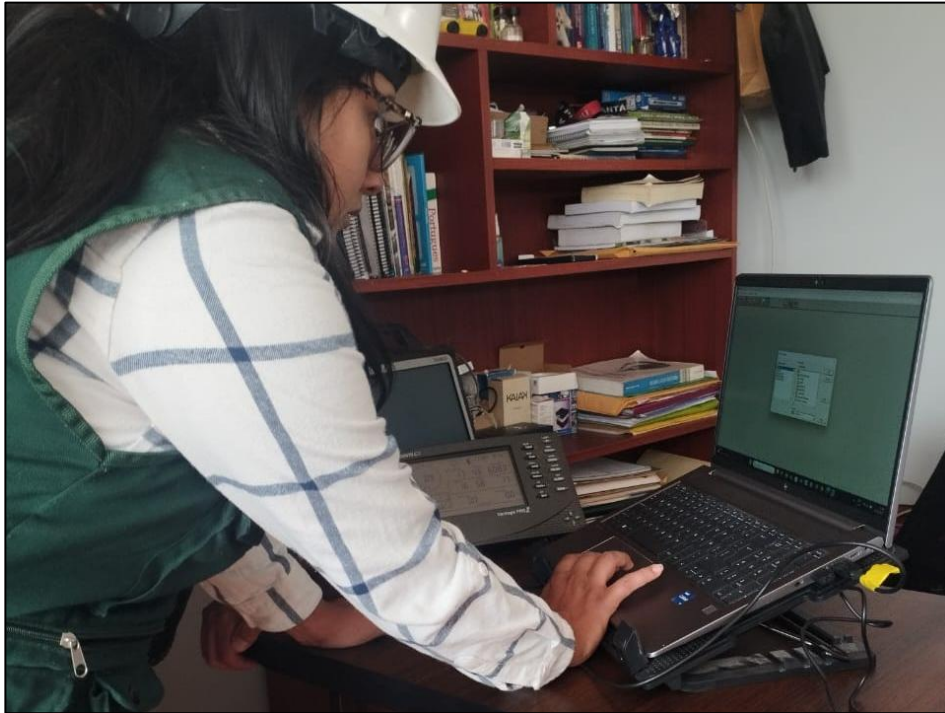
Figura 15

Sensor PurpleAir utilizado para el monitoreo continuo de material particulado $PM_{2.5}$ en la zona de estudio



Figura 16

Procedimiento de descarga de registros meteorológicos y de PM_{2.5}

**Figura 17**

Visualización y verificación de la cantidad de datos recolectados durante el periodo de monitoreo



Figura 18

Generación de gráficos para el análisis y procesamiento de datos ambientales

**Figura 19**

Estación de monitoreo meteorológico con consola de adquisición de datos y computadora portátil para visualización y análisis en tiempo real

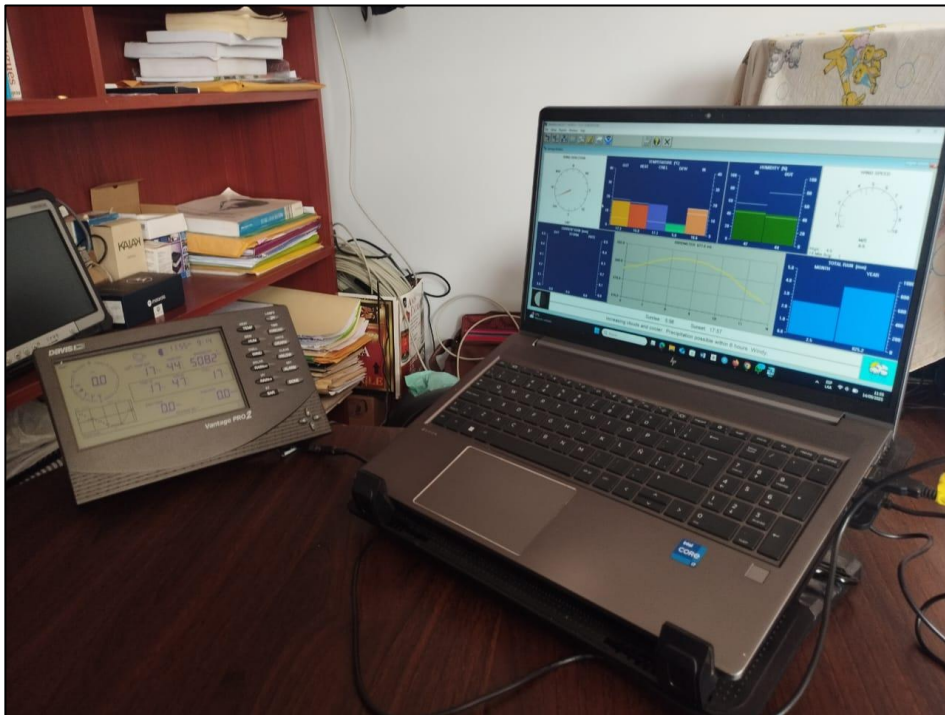


Figura 20

Infraestructura donde se encuentran instalados los dispositivos Larson Davis y PurpleAir

**Figura 21**

Juliana Pardo Aiquipa y Estefani Soto Onofre en el punto de instalación de los equipos Larson Davis y PurpleAir



Anexo 4. Script en R para el procesamiento, análisis estadístico y modelamiento predictivo de PM_{2.5}

```
# =====
# 1. LIBRERÍAS
# =====

library(tidyverse)
library(lubridate)
library(Metrics)
library(sjPlot)
library(randomForest)

setwd("E:/TRABAJOS DE INVESTIGACION/YUYU")
dat <- read.csv("data_unida.csv", header=TRUE, sep = ",", stringsAsFactors = FALSE)

# 1. Corregir formato de Fecha y Hora (Esto recupera el 00:00:00 omitido)
dat$datetime <- parse_date_time(dat$datetime, orders = c("ymd HMS", "ymd"))
str(dat)

dat$datetime <- as.POSIXct(dat$datetime, format="%Y-%m-%d %H:%M:%S")
datmach<-dat[,-c(8:11)]
head(datmach)

# =====
# 2. PREPARACIÓN DE DATOS
# =====

data <- datmach
data$datetime <- as.POSIXct(data$datetime)
vars <- c("Temp.Out", "Out.Hum", "Wind.Speed", "Rain", "Bar")
data_model <- data %>%
  select(datetime, PM25_interp, all_of(vars)) %>%
  drop_na()

# =====
# 1. INGENIERÍA DE VARIABLES (Lags) Y PREPARACIÓN
# =====

# Creamos retrasos temporales: el clima de hace 1 hora influye en el PM2.5 de ahora.
data_model <- data_model %>%
  arrange(datetime) %>%
  mutate(
    temp_lag1 = lag(Temp.Out, 1),
    hum_lag1 = lag(Out.Hum, 1),
    wind_lag1 = lag(Wind.Speed, 1),
    bar_lag1 = lag(Bar, 1),
    pm25_lag1 = lag(PM25_interp, 1) # Auto-regresivo (Clave para R2 positivo)
  ) %>%
  drop_na()

# Volvemos a dividir con los nuevos datos con Lags
train_index <- floor(0.7 * nrow(data_model))
train <- data_model[1:train_index, ]
test <- data_model[(train_index + 1):nrow(data_model), ]

# Estandarización: Fundamental para Regresión Lineal y ANN
# Escalamos basándonos en TRAIN para no contaminar con datos de TEST
train_scaled <- train
test_scaled <- test

# Identificamos columnas numéricas predictoras (excluyendo datetime y el target original)
cols_to_scale <- names(train)[!names(train) %in% c("datetime", "PM25_interp")]

# Escalado manual consistente
train_mean <- apply(train[, cols_to_scale], 2, mean)
train_sd <- apply(train[, cols_to_scale], 2, sd)
train_scaled[, cols_to_scale] <- scale(train[, cols_to_scale], center = train_mean, scale = train_sd)
```

```

test_scaled[, cols_to_scale] <- scale(test[, cols_to_scale], center = train_mean, scale = train_sd)

# =====

# 2. MODELOS OPTIMIZADOS
# =====

# --- Modelo 1: Regresión Lineal ---
# Ahora con datos escalados y lags, el R2 debería dejar de ser negativo
model_lm <- lm(PM25_interp ~ ., data = train_scaled[, -1])
pred_lm <- predict(model_lm, newdata = test_scaled)
# --- Modelo 2: Random Forest ---
set.seed(123)
model_rf <- randomForest(
  PM25_interp ~ .,
  data = train_scaled[, -1],
  ntree = 800,    # Aprovechando tus 6 hilos lógicos
  mtry = 3,
  nodesize = 10,  # Evita el sobreajuste (overfitting)
  importance = TRUE
)
pred_rf <- predict(model_rf, newdata = test_scaled)
# --- Modelo 3: Red Neuronal (ANN) ---
library(nnet)
set.seed(123)
model_ann <- nnet(
  PM25_interp ~ .,
  data = train_scaled[, -1],
  size = 12,      # Tamaño intermedio para evitar ruido
  linout = TRUE,
  maxit = 1000,
  decay = 0.3,    # Regularización para estabilizar el modelo
  trace = FALSE
)
pred_ann <- predict(model_ann, newdata = test_scaled)

# =====

# MÉTRICAS COMPLETAS (ROBUSTAS)
# =====

calc_metrics_full <- function(obs, pred){
  # =====

  # FILTRAR CEROS PARA MAPE
  # =====

  idx <- obs > 0
  # =====

  # R2 REAL
  # =====

  ss_res <- sum((obs - pred)^2)
  ss_tot <- sum((obs - mean(obs))^2)
  data.frame(
    MAE = mae(obs, pred),
    MSE = mean((obs - pred)^2),
    RMSE = rmse(obs, pred),
    MAPE = mean(abs((obs[idx] - pred[idx]) / obs[idx])) * 100,
    R2 = 1 - (ss_res / ss_tot)
  )
}

# =====

# MÉTRICAS TRAIN
# =====

metrics_lm_tr <- calc_metrics_full(train$PM25_interp, predict(model_lm, train))

```

```

metrics_rf_tr <- calc_metrics_full(train$PM25_interp, predict(model_rf, train))
metrics_ann_tr <- calc_metrics_full(train$PM25_interp, predict(model_ann, train_scaled))
# =====
# MÉTRICAS TEST
# =====
metrics_lm_te <- calc_metrics_full(test$PM25_interp, pred_lm)
metrics_rf_te <- calc_metrics_full(test$PM25_interp, pred_rf)
metrics_ann_te <- calc_metrics_full(test$PM25_interp, pred_ann)
# =====
# TABLA FINAL
# =====
df_results <- data.frame(
  Modelo = c("Regresión Lineal", "Random Forest", "ANN"),
  # TRAIN
  MAE_train = c(metrics_lm_tr$MAE, metrics_rf_tr$MAE, metrics_ann_tr$MAE),
  MSE_train = c(metrics_lm_tr$MSE, metrics_rf_tr$MSE, metrics_ann_tr$MSE),
  RMSE_train = c(metrics_lm_tr$RMSE, metrics_rf_tr$RMSE, metrics_ann_tr$RMSE),
  MAPE_train = c(metrics_lm_tr$MAPE, metrics_rf_tr$MAPE, metrics_ann_tr$MAPE),
  R2_train = c(metrics_lm_tr$R2, metrics_rf_tr$R2, metrics_ann_tr$R2),
  # TEST
  MAE_test = c(metrics_lm_te$MAE, metrics_rf_te$MAE, metrics_ann_te$MAE),
  MSE_test = c(metrics_lm_te$MSE, metrics_rf_te$MSE, metrics_ann_te$MSE),
  RMSE_test = c(metrics_lm_te$RMSE, metrics_rf_te$RMSE, metrics_ann_te$RMSE),
  MAPE_test = c(metrics_lm_te$MAPE, metrics_rf_te$MAPE, metrics_ann_te$MAPE),
  R2_test = c(metrics_lm_te$R2, metrics_rf_te$R2, metrics_ann_te$R2)
)
library(sjPlot)
tab_df(
  df_results,
  title = "Tabla X. Evaluación comparativa de modelos (Train vs Test)",
  footnote = "Nota. El MAPE fue calculado excluyendo valores de PM2.5 iguales a cero.",
  digits = 3,
  show.type = FALSE,
  alternate.rows = FALSE
)
# =====
# OBJETIVO ESPECÍFICO 4
# COMPARACIÓN Y SELECCIÓN DEL MEJOR MODELO
# =====
library(sjPlot)
library(dplyr)
library(neuralnet)
# =====
# 1. PESOS MODELO LINEAL
# =====
coef_lm <- coef(model_lm)

df_lm <- data.frame(
  Variable = names(coef_lm)[-1],
  Peso_LM = coef_lm[-1]
)
# =====
# 2. PESOS RANDOM FOREST
# =====
imp <- importance(model_rf)
df_rf <- data.frame(
  Variable = rownames(imp),

```

```

Peso_RF = imp[,1]
)
# Normalizar RF
df_rf$Peso_RF <- df_rf$Peso_RF / sum(df_rf$Peso_RF)
# =====

# 3. PESOS RED NEURONAL (APROXIMACIÓN)
# =====

# Extraer pesos ANN
pesos_ann <- model_ann$wts
# Aproximación: usar promedio de pesos por variable
vars <- names(train)[-c(1,2)] # excluir datetime y target

df_ann <- data.frame(
  Variable = vars,
  Peso_ANN = abs(pesos_ann[1:length(vars)])
)
# Normalizar ANN
df_ann$Peso_ANN <- df_ann$Peso_ANN / sum(df_ann$Peso_ANN)
# =====

# 4. UNIR PESOS DE LOS 3 MODELOS
# =====

df_pesos <- df_lm %>%
  full_join(df_rf, by="Variable") %>%
  full_join(df_ann, by="Variable")
# =====

# 5. MÉTRICAS (YA CALCULADAS ANTES)
# =====

df_metrics <- data.frame(
  Modelo = c("Regresión Lineal", "Random Forest", "ANN"),
  MAE = c(metrics_lm_te$MAE, metrics_rf_te$MAE, metrics_ann_te$MAE),
  MSE = c(metrics_lm_te$MSE, metrics_rf_te$MSE, metrics_ann_te$MSE),
  RMSE = c(metrics_lm_te$RMSE, metrics_rf_te$RMSE, metrics_ann_te$RMSE),
  MAPE = c(metrics_lm_te$MAPE, metrics_rf_te$MAPE, metrics_ann_te$MAPE),
  R2 = c(metrics_lm_te$R2, metrics_rf_te$R2, metrics_ann_te$R2)
)
# =====

# 6. TABLA FINAL UNIFICADA (APA)
# =====

tab_df(
  df_metrics,
  title = "Tabla X. Comparación de desempeño de modelos predictivos",
  footnote = "Nota. Se presentan métricas de error y ajuste para cada modelo evaluado.",
  show.type = FALSE,
  alternate.rows = FALSE,
  digits = 3,
  CSS = list(
    css.table = "border-top: 2px solid black; border-bottom: 2px solid black;",
    css.thead = "border-bottom: 1px solid black;"
  )
)
# =====

# 7. TABLA DE PESOS (LOS 3 MODELOS)
# =====

tab_df(
  df_pesos,
  title = "Tabla X. Comparación de pesos de variables en los modelos evaluados",
  footnote = "Nota. Los pesos del modelo Random Forest y ANN fueron normalizados.",

```

```

show.type = FALSE,
alternate.rows = FALSE,
digits = 3,
CSS = list(
  css.table = "border-top: 2px solid black; border-bottom: 2px solid black;",
  css.thead = "border-bottom: 1px solid black;"
)
)
# =====
# 8. ECUACIONES DE LOS MODELOS
# =====
# ECUACIÓN REGRESIÓN LINEAL
ecuacion_lm <- paste0(
  "PM2.5 = ",
  round(coef_lm[1], 3), " + ",
  paste(
    paste0(round(coef_lm[-1], 3), "**", names(coef_lm)[-1]),
    collapse = " + "
  )
)
)
cat("\n=====\\n")
cat(" MODELO REGRESIÓN LINEAL:\\n")
cat(ecuacion_lm)
# ECUACIÓN RANDOM FOREST (APROXIMADA)
ecuacion_rf <- paste0(
  "PM2.5 ≈ ",
  paste(
    paste0(round(df_rf$Peso_RF, 3), "**", df_rf$Variable),
    collapse = " + "
  )
)
)
cat("\n\\n=====\\n")
cat("MODELO RANDOM FOREST (APROX.):\\n")
cat(ecuacion_rf)

# ECUACIÓN ANN (APROXIMADA)
ecuacion_ann <- paste0(
  "PM2.5 ≈ f(",
  paste(
    paste0(round(df_ann$Peso_ANN, 3), "**", df_ann$Variable),
    collapse = " + "
  ),
  ")",
  "\\n"
)
)
cat("\n\\n=====\\n")
cat("MODELO RED NEURONAL (APROX.):\\n")
cat(ecuacion_ann)
cat("\n=====\\n")
# =====
# 9. MEJOR MODELO
# =====
best_model <- df_metrics %>%
  arrange(desc(R2), RMSE, MAE)
print("Mejor modelo:")
print(best_model[1, ])
# =====
# CORRELACIÓN SPEARMAN

```

```

# =====
library(Hmisc)
library(dplyr)
library(sjPlot)
# =====

# MATRIZ DE CORRELACIÓN (Spearman)
# =====

corr_data <- data_model %>%
  select(PM25_interp, all_of(vars))
corr <- rcorr(as.matrix(corr_data), type = "spearman")
corr_r <- corr$r      # coeficientes
corr_p <- corr$p      # p-values
# =====

# FUNCIÓN PARA SIGNIFICANCIA
# =====

get_stars <- function(p){
  ifelse(p < 0.001, "***",
    ifelse(p < 0.01, "**",
      ifelse(p < 0.05, "*", "")))
}
# =====

# COMBINAR VALOR + SIGNIFICANCIA
# =====

corr_final <- matrix("", nrow = nrow(corr_r), ncol = ncol(corr_r))
for(i in 1:nrow(corr_r)){
  for(j in 1:ncol(corr_r)){
    valor <- round(corr_r[i,j], 3)
    stars <- get_stars(corr_p[i,j])
    corr_final[i,j] <- paste0(valor, stars)
  }
}

corr_final <- as.data.frame(corr_final)
colnames(corr_final) <- colnames(corr_r)
rownames(corr_final) <- rownames(corr_r)
# =====

# TABLA APA (CON SIGNIFICANCIA)
# =====

tab_df(
  corr_final,
  title = "Tabla X. Correlación de Spearman entre variables meteorológicas y PM2.5",
  footnote = "Nota. *** p < 0.001; ** p < 0.01; * p < 0.05.",
  show.type = FALSE,
  alternate.rows = FALSE,
  CSS = list(
    css.table = "border-top: 2px solid black; border-bottom: 2px solid black;",
    css.thead = "border-bottom: 1px solid black;"
  )
)
# =====

# CORRELACIÓN VISUAL (TIPO PAPER)
# =====

# Seleccionar variables
base <- data_model %>%
  select(PM25_interp, all_of(vars))
library(corrplot)
# =====

# 1. MATRIZ DE CORRELACIÓN (Spearman)

```

```

# =====
correlacion <- cor(base, method = "spearman", use = "complete.obs")
# =====

# 1. MATRIZ DE CORRELACIÓN (Spearman) - CORREGIDO
# =====

correlacion <- cor(base, method = "spearman", use = "complete.obs")
# Creamos el vector con los 11 nombres (incluyendo los rezagos/lags)
nuevos_nombres <- c(
  "PM2.5",
  "Temperatura",
  "Humedad",
  "Vel. Viento",
  "Lluvia",
  "Presión",
  "Temp (Lag1)",
  "Hum (Lag1)",
  "Viento (Lag1)",
  "Presión (Lag1)",
  "PM2.5 (Lag1)"
)
# Ahora sí, asignamos (11 nombres para 11 columnas/filas)
colnames(correlacion) <- nuevos_nombres
rownames(correlacion) <- nuevos_nombres
# Verificamos
print(correlacion[1:5, 1:5]) # Muestra solo una parte para confirmar
# =====

# 2. GRÁFICO
# =====

jpeg(
  "Correlacion FINAL.jpeg",
  width = 9.5, height = 6, # 📏 reduce tamaño (antes 13.5 x 6.5)
  units = "in",
  res = 600, # 📏 baja de 600 → 400 (suficiente para tesis)
  quality = 100 # máxima calidad JPEG
)

corrplot(
  correlacion,
  method = "number", # muestra valores
  type = "upper", # triángulo superior
  tl.col = "black", # nombres variables
  tl.srt = 45 # rotación etiquetas
)
dev.off()
# =====

# CARGA DE LIBRERÍAS
# =====

library(sjPlot)
library(ggplot2)
library(patchwork)
# =====

# 1. DEFINICIÓN DE LA FUNCIÓN DE COMPARACIÓN (Ejecutar esto primero)
# =====

compare_models <- function(p1, p2, obs){
  # Cálculo de errores
  err1 <- obs - p1
  err2 <- obs - p2

```

```

mae1 <- abs(err1)
mae2 <- abs(err2)
mse1 <- err1^2
mse2 <- err2^2
# Pruebas de Wilcoxon (Comparación no paramétrica)
p_mae <- wilcox.test(mae1, mae2)$p.value
p_mse <- wilcox.test(mse1, mse2)$p.value
p_rmse <- p_mse # El RMSE deriva del MSE
# Cálculo de R2 para la comparación
ss_tot <- sum((obs - mean(obs))^2)
r2_1 <- 1 - sum(err1^2)/ss_tot
r2_2 <- 1 - sum(err2^2)/ss_tot
# Comparación de residuos al cuadrado para R2
p_r2 <- wilcox.test(err1^2, err2^2)$p.value
# P-value global (Promedio de métricas)
p_global <- mean(c(p_mae, p_mse, p_r2))
return(data.frame(
  p_value = p_global,
  MAE = p_mae,
  MSE = p_mse,
  RMSE = p_rmse,
  R2 = p_r2
))
}
# =====
# 2. EJECUCIÓN DE COMPARACIONES ESTADÍSTICAS
# =====
# Asegúrate que pred_rf, pred_lm, pred_ann y test$PM25_interp existan en tu ambiente
comp1 <- compare_models(pred_rf, pred_lm, test$PM25_interp)
comp2 <- compare_models(pred_rf, pred_ann, test$PM25_interp)
comp3 <- compare_models(pred_lm, pred_ann, test$PM25_interp)
# Unir resultados en una tabla
df_stats <- rbind(
  cbind('S. No` = 1, Comparison="RF vs LM", comp1),
  cbind('S. No` = 2, Comparison="RF vs ANN", comp2),
  cbind('S. No` = 3, Comparison="LM vs ANN", comp3)
)
# Función para formatear P-values (APA style)
format_p <- function(x){
  ifelse(x < 0.001, "<0.001", round(as.numeric(x), 4))
}
df_stats[,3:7] <- lapply(df_stats[,3:7], format_p)
# Mostrar Tabla tipo APA en la consola/viewer
tab_df(
  df_stats,
  title = "Tabla X. Comparación estadística del desempeño de modelos (Prueba de Wilcoxon)",
  footnote = "Nota. Los valores p inferiores a 0.05 indican diferencias estadísticamente significativas entre los modelos comparados.",
  show.type = FALSE,
  alternate.rows = FALSE,
  CSS = list(
    css.table = "border-top: 2px solid black; border-bottom: 2px solid black;",
    css.thead = "border-bottom: 1px solid black;"
  )
)
# =====
# 3. GENERACIÓN DEL MULTIGRÁFICO (Objetivos 3 y 4)
# =====

```

```

# Redefinir la función de plot para asegurar el formato PM2.5
plot_model <- function(real, pred, titulo){
  df <- data.frame(real = real, pred = pred, t = test$date)

  ggplot(df, aes(x = t)) +
    # Cambiamos 'size' por 'linewidth' para evitar el aviso
    geom_line(aes(y = real, color = "Observado"), linewidth = 0.8) +
    geom_line(aes(y = pred, color = "Modelado"), linewidth = 0.8) +
    # Líneas de referencia internacional (también con linewidth)
    geom_hline(yintercept = 50, col = "red", linetype = "solid", linewidth = 0.8) +
    geom_hline(yintercept = 25, col = "darkorange", linetype = "dashed", linewidth = 0.8) +
    # Etiquetas con subíndice PM2.5
    labs(
      title = titulo,
      x = "",
      y = expression(paste(PM[2.5], " (" , mu, "g/", "m^3, ")"))
    ) +
    scale_color_manual(values = c("Observado" = "black", "Modelado" = "blue")) +
    # Estilo de Tesis
    theme(
      panel.background = element_rect(fill = 'white', colour = 'black'),
      panel.grid.major = element_blank(),
      panel.grid.minor = element_blank(),
      axis.line = element_line(colour = "black"),
      plot.title = element_text(hjust = 0.5, size = 14, face = "bold"),
      axis.title.y = element_text(size = 12),
      axis.text.x = element_text(size = 10, angle = 45, hjust = 1),
      axis.text.y = element_text(size = 10, angle = 90, hjust = 1),
      legend.title = element_blank(),
      legend.position = "top"
    )
  }

# Ahora vuelve a ejecutar la creación de los objetos:
p1 <- plot_model(test$PM25_interp, pred_lm, "a) Regresión Lineal")
p2 <- plot_model(test$PM25_interp, pred_rf, "b) Random Forest")
p3 <- plot_model(test$PM25_interp, pred_ann, "c) Red Neuronal (ANN)")
# =====

# 4. GUARDAR EN JPEG (Alta Resolución)
# =====

jpeg(
  "Multigrafico_PM25_Final corregido.jpeg",
  width = 10, height = 12,
  units = "in",
  res = 600, # Calidad máxima para impresión de tesis
  quality = 100
)

# Unir gráficos con patchwork y recolectar leyenda
(p1 / p2 / p3) + plot_layout(guides = "collect") & theme(legend.position = 'bottom')
dev.off()

print("Proceso finalizado. El gráfico se ha guardado en tu carpeta de trabajo.")
#####

library(ggplot2)
library(patchwork)

# FUNCIÓN PARA GRÁFICO DE DISPERSIÓN (CORRELACIÓN)
plot_correlacion <- function(real, pred, titulo) {
  df <- data.frame(real = real, pred = pred)

  # Calcular R2 para mostrarlo en el gráfico
  r2_val <- round(cor(real, pred)^2, 3)

```

```

ggplot(df, aes(x = real, y = pred)) +
  # Puntos con transparencia para evitar solapamiento (overplotting)
  geom_point(alpha = 0.4, color = "black", size = 1.5) +
  # Línea de identidad (45 grados) - El ideal
  geom_abline(intercept = 0, slope = 1, linetype = "dashed", color = "red", linewidth = 0.8) +
  # Línea de regresión ajustada
  geom_smooth(method = "lm", color = "blue", se = FALSE, linewidth = 1) +
  # Etiquetas con subíndices
  labs(
    title = titulo,
    subtitle = paste("Coeficiente de determinación R² =", r2_val),
    x = expression(paste("Observado ", PM[2.5], " (" , mu, "g/", m^3, ")")),
    y = expression(paste("Modelado ", PM[2.5], " (" , mu, "g/", m^3, ")"))
  ) +
  # Estilo igual a tus otros gráficos
  theme_minimal() +
  theme(
    panel.background = element_rect(fill = 'white', colour = 'black'),
    panel.grid.major = element_line(color = "grey95"),
    panel.grid.minor = element_blank(),
    plot.title = element_text(hjust = 0.5, size = 12, face = "bold"),
    plot.subtitle = element_text(hjust = 0.5, size = 10, face = "italic"),
    axis.title = element_text(size = 11),
    axis.text = element_text(size = 10)
  )
}

# 2. Generar los 3 gráficos de correlación
c1 <- plot_correlacion(test$PM25_interp, pred_lm, "a) Regresión Lineal")
c2 <- plot_correlacion(test$PM25_interp, pred_rf, "b) Random Forest")
c3 <- plot_correlacion(test$PM25_interp, pred_ann, "c) Red Neuronal (ANN)")

# 3. Guardar en JPEG (Organizados en una sola fila para comparar mejor)
jpeg(
  "Multigrafico_Correlacion_Final.jpeg",
  width = 12, height = 6, # Formato alargado para 3 gráficos lado a lado
  units = "in",
  res = 700,
  quality = 100
)

# Unir los 3 gráficos horizontalmente
print(c1 | c2 | c3)
dev.off()

```

Anexo 5. *Script en R para la comparación estadística y validación gráfica de modelos predictivos de PM_{2.5} mediante pruebas de Wilcoxon*

```

# =====
# COMPARACIÓN ESTADÍSTICA CORREGIDA (TIPO PAPER)
# =====

library(sjPlot)

# =====
# FUNCIÓN CORRECTA
# =====

compare_models <- function(p1, p2, obs){
  err1 <- obs - p1
  err2 <- obs - p2
  mae1 <- abs(err1)

```

```

mae2 <- abs(err2)
mse1 <- err1^2
mse2 <- err2^2
# Wilcoxon tests
p_mae <- wilcox.test(mae1, mae2)$p.value
p_mse <- wilcox.test(mse1, mse2)$p.value
p_rmse <- p_mse # RMSE depende de MSE → mismo contraste
# pseudo comparación R2 (más correcto)
ss_tot <- sum((obs - mean(obs))^2)
r2_1 <- 1 - sum(err1^2)/ss_tot
r2_2 <- 1 - sum(err2^2)/ss_tot
p_r2 <- wilcox.test(err1^2, err2^2)$p.value
# p-value global (promedio como en papers)
p_global <- mean(c(p_mae, p_mse, p_r2))
return(data.frame(
  p_value = p_global,
  MAE = p_mae,
  MSE = p_mse,
  RMSE = p_rmse,
  R2 = p_r2
))
}
# =====
# COMPARACIONES
# =====
comp1 <- compare_models(pred_rf, pred_lm, test$PM25_interp)
comp2 <- compare_models(pred_rf, pred_ann, test$PM25_interp)
comp3 <- compare_models(pred_lm, pred_ann, test$PM25_interp)
# =====
# TABLA FINAL
# =====
df_stats <- rbind(
  cbind('S. No` = 1, Comparison="RF vs LM", comp1),
  cbind('S. No` = 2, Comparison="RF vs ANN", comp2),
  cbind('S. No` = 3, Comparison="LM vs ANN", comp3)
)
# Evitar 0.0000
format_p <- function(x){
  ifelse(x < 0.001, "<0.001", round(x, 4))
}
df_stats[,3:7] <- lapply(df_stats[,3:7], format_p)
# =====
# TABLA APA
# =====
tab_df(
  df_stats,
  title = "Tabla X. Comparación estadística del desempeño de modelos",
  footnote = "Nota. Valores p obtenidos mediante la prueba de Wilcoxon para comparación de errores entre modelos.",
  show.type = FALSE,
  alternate.rows = FALSE,
  CSS = list(
    css.table = "border-top: 2px solid black; border-bottom: 2px solid black;",
    css.thead = "border-bottom: 1px solid black;"
  )
)
# =====
# GRÁFICO COMPARATIVO (OBJETIVOS 3 Y 4)

```

```

# =====
library(ggplot2)
library(patchwork)

# 1. Función actualizada con el estilo de tu segundo script
plot_model <- function(real, pred, titulo){
  df <- data.frame(real = real, pred = pred, t = test$date)
  ggplot(df, aes(x = t)) +
    # Líneas de los modelos
    geom_line(aes(y = real, color = "Observado"), size = 0.8) +
    geom_line(aes(y = pred, color = "Modelado"), size = 0.8) +
    # Línea ECA Perú (50) y OMS (25)
    geom_hline(yintercept = 50, col = "red", linetype = "solid", size = 0.8) +
    geom_hline(yintercept = 25, col = "darkorange", linetype = "dashed", size = 0.8) +
    # Eje Y con Subíndice y Unidades
    labs(
      title = titulo,
      x = "",
      y = expression(paste(PM[2.5], " (" , mu, "g/", m^3, ")"))
    ) +
    # Colores manuales para las líneas
    scale_color_manual(values = c("Observado" = "black", "Modelado" = "blue")) +
    # Configuración del Tema igual a tu ejemplo
    theme(
      panel.background = element_rect(fill = 'white', colour = 'black'),
      panel.grid.major = element_blank(),
      panel.grid.minor = element_blank(),
      axis.line = element_line(colour = "black"),
      plot.title = element_text(hjust = 0.5, size = 14, face = "bold"),
      axis.title.y = element_text(size = 12),
      axis.text.x = element_text(size = 10, angle = 45, hjust = 1),
      axis.text.y = element_text(size = 10, angle = 90, hjust = 1),
      legend.title = element_blank(),
      legend.position = "top"
    )
}

# 2. Generar los gráficos (se asume que pred_lm, pred_rf y pred_ann ya existen)
p1 <- plot_model(test$PM25_interp, pred_lm, "a) Regresión Lineal")
p2 <- plot_model(test$PM25_interp, pred_rf, "b) Random Forest")
p3 <- plot_model(test$PM25_interp, pred_ann, "c) RNN")

# 3. Guardar en JPEG
jpeg(
  "Multigrafico_PM251.jpeg",
  width = 10, height = 11, # Un poco más alto para que luzcan los ejes
  units = "in",
  res = 700,
  quality = 100
)

# Unir y recolectar leyendas para que no se repitan
(p1 / p2 / p3) + plot_layout(guides = "collect") & theme(legend.position = 'bottom')
dev.off()

```