**Logotipo

Descripción generada automáticamente**

**Análisis de la Correlación entre Variables Atmosféricas y Contaminantes mediante Técnicas de Machine Learning para la Predicción de la Calidad del Aire en el Oriente Antioqueño**

Stefania Nieto Téllez

Diego Alejandro Mora Suárez

Monografía presentada para optar al título de Especialista en Analítica y Ciencia de Datos

Asesor  
Nombres completos, Título académico más alto

Universidad de Antioquia  
Facultad de Ingeniería

Especialización en Analítica y Ciencia de Datos

Medellín, Antioquia, Colombia

2024

|  |  |
| --- | --- |
| **Cita** | (Mora Suarez & Nieto Tellez, 2024) |
| **Referencia**  **Estilo APA 7 (2020)** | Mora Suárez, D.A., & Nieto Tellez, S. (2024). *Análisis de la Correlación entre Variables Atmosféricas y Contaminantes mediante Técnicas de Machine Learning para la Predicción de la Calidad del Aire en el Oriente Antioqueño* Trabajo de grado especialización]. Universidad de Antioquia, Medellín, Colombia. |

**** 

Especialización en Analítica y Ciencia de Datos, CohorteVIII.

Centro de Investigación Ambientales y de Ingeniería (CIA).

|  |  |
| --- | --- |
|  | Diagrama  Descripción generada automáticamente con confianza media |

Centro de Documentación Ingeniería (CENDOI)

**Repositorio Institucional:** http://bibliotecadigital.udea.edu.co

Universidad de Antioquia - www.udea.edu.co

Rector: John Jairo Arboleda Céspedes.

Decano: Julio Cesar Saldarriaga Molina

Jefe departamento: Diego José Luis Botia Valderrama

El contenido de esta obra corresponde al derecho de expresión de los autores y no compromete el pensamiento institucional de la Universidad de Antioquia ni desata su responsabilidad frente a terceros. Los autores asumen la responsabilidad por los derechos de autor y conexos.

**Dedicatoria**

Texto de dedicatoria centrado.

**Agradecimientos**

Texto de agradecimientos centrado.

**Tabla de contenido**

[Resumen 8](#_Toc183095120)

[Abstract 9](#_Toc183095121)

[1. Descripción del problema 10](#_Toc183095122)

[1.1. Problema de negocio 10](#_Toc183095123)

[1.2. Aproximación desde la analítica de datos 12](#_Toc183095124)

[1.3. Origen de los datos 12](#_Toc183095125)

[1.4. Métricas de desempeño 12](#_Toc183095126)

[2. Objetivos 14](#_Toc183095127)

[2.1. Objetivo general 14](#_Toc183095128)

[2.2. Objetivos específicos 14](#_Toc183095129)

[3. Datos 15](#_Toc183095130)

[3.1. Datos originales 15](#_Toc183095131)

[3.2. Datsets 16](#_Toc183095132)

[3.3. Analítica descriptiva 16](#_Toc183095133)

[18](#_Toc183095134)

[Referencias 19](#_Toc183095135)

[Anexos 20](#_Toc183095136)

**Lista de tablas**

[**Tabla 1**](#_heading=h.2u6wntf) Descripción de los campos almacenados en los archivos CSV.16-17

**Lista de figuras**

[**Figura 1**](#_heading=h.2r0uhxc) Distribución de los datos para cada una de las variables numéricas 18

[**Figura 2**](#_heading=h.1664s55) Matriz de correlación entre los contaminantes y los factores meteorológicos 19

**Siglas, acrónimos y abreviaturas**

**APA** American Psychological Association

**Cms.** Centímetros

**ERIC** Education Resources Information Center

**Esp.** Especialista

**MP** Magistrado Ponente

**MSc** Magister Scientiae

**Párr.** Párrafo

**PhD** Philosophiae Doctor

**PBQ-SF** Personality Belief Questionnaire Short Form

**PostDoc** PostDoctor

**UdeA** Universidad de Antioquia

**CORNARE** Corporación Aútonoma Regional de las Cuencas de los Ríos Negro y Nare

# Resumen

La calidad del aire en el Oriente Antioqueño se ha convertido en un desafío crítico debido al crecimiento poblacional e industrial de la región, lo que ha incrementado los niveles de contaminantes atmosféricos criterio como PM10, PM2.5, O3, SO2, NO2 y CO. Además, las condiciones meteorológicas adversas, como altas temperaturas y bajas humedades, han exacerbado el impacto de estos contaminantes. Este proyecto busca implementar herramientas basadas en ciencia de datos para analizar y modelar estas relaciones, proporcionando una base técnica para apoyar a las autoridades ambientales en la toma de decisiones.

Para ello, se utilizaron datos históricos de calidad del aire y meteorología recolectados por CORNARE entre 2019 y 2024 en la estación San Antonio de Pereira. Tras un riguroso proceso de limpieza y pre procesamiento de los datos, se identificaron patrones clave mediante análisis estadísticos y exploratorios. También se desarrollaron modelos predictivos con técnicas de machine learning, capaces de evaluar escenarios críticos y generar alertas tempranas.

Los resultados preliminares muestran correlaciones significativas entre variables meteorológicas como la temperatura y concentraciones de ozono troposférico, confirmando estudios previos. Estas herramientas no solo permiten identificar tendencias, sino también generar estrategias para mitigar el impacto ambiental y proteger la salud pública.

El código del proyecto se encuentra disponible en el repositorio de GitHub asociado (<https://acortar.link/HDNCHO> ).

*Palabras clave: Calidad del aire, análisis de datos, machine learning, variables meteorológicas, contaminantes atmosféricos criterio.*

# Abstract

Air quality in the Oriente Antioqueño region has become a critical challenge due to the population and industrial growth of the area, which has increased the levels of criteria air pollutants such as PM10, PM2.5, O3, SO2, NO2, and CO. Additionally, adverse meteorological conditions, such as high temperatures and low humidity levels, have exacerbated the impact of these pollutants. This project aims to implement data science-based tools to analyze and model these relationships, providing a technical foundation to support environmental authorities in decision-making.

To achieve this, historical air quality and meteorological data collected by CORNARE between 2019 and 2024 at the San Antonio de Pereira station were used. After a rigorous data cleaning and preprocessing phase, key patterns were identified through statistical and exploratory analyses. Predictive models were also developed using machine learning techniques, capable of evaluating critical scenarios and generating early warnings.

Preliminary results show significant correlations between meteorological variables such as temperature and tropospheric ozone concentrations, confirming previous studies. These tools not only enable the identification of trends but also the development of strategies to mitigate environmental impacts and protect public health.

The code of this project is available in the associated GitHub repository (<https://acortar.link/HDNCHO> ).

*Keywords****:*** *Air Quality, Data Analysis, Machine Learning, Meteorological Variables, Criteria Air Pollutants*

# Descripción del problema

Según la Organización Mundial de la Salud (OMS), cada año la contaminación ambiental contribuye a la muerte de alrededor de 7 millones de personas en el mundo, al agravar enfermedades respiratorias y cardiovasculares (OMS, 2021). En el Oriente antioqueño, el rápido crecimiento industrial y urbano ha generado preocupaciones sobre el impacto de los contaminantes atmosféricos en la salud pública y el ambiente (Departamento Administrativo Nacional de Estadística -en adelante DANE-, censo 2018 – actualizado 2019). Por ello, las entidades responsables y las comunidades deben implementar planes de gestión de calidad del aire que permitan monitorear y reducir los niveles de contaminación.

Tradicionalmente, la detección de contaminantes requiere la toma de muestras y su análisis en estaciones de monitoreo fijo, un proceso que consume tiempo y limita las acciones preventivas inmediatas. Ante esta situación, surge la necesidad de crear herramientas predictivas que, mediante modelos de machine learning con datos históricos de mediciones, anticipen episodios de contaminación atmosférica, permitiendo tomar medidas oportunas y proteger la salud de la población y la calidad ambiental de la región. Las métricas para

## Problema de negocio

Según la Organización Mundial de la Salud (OMS), la contaminación atmosférica se cobra trece vidas en el mundo cada minuto (OMS, 2022). Más del 32 % de las muertes se deben a la contaminación atmosférica, que afecta negativamente al bienestar general, al clima y a la economía (Wu et al., 2021). La Agencia de Protección del Medio Ambiente de Estados Unidos (EPA) define la contaminación atmosférica como sustancias, agentes físicos o regulares que degradan el aire y dañan sus propiedades (EPA, 2023). Entre los contaminantes atmosféricos criterio se encuentran el Material Particulado (MP), el Ozono (O3) troposférico, el dióxido de nitrógeno (NO2) y el dióxido de azufre (SO2) proceden principalmente de actividades humanas y procesos naturales (OMS, 2021).

Los factores meteorológicos como la temperatura, la precipitación o la humedad, tienen un gran efecto el cual puede ser negativo o positivo, en las proporciones de los contaminantes presentes en el aire. Los cambios en las condiciones meteorológicas pueden amplificar el impacto del cambio climático al influir en la formación, el transporte y la dispersión de los contaminantes (Ramanathan, 2020). Un estudio anterior ha determinado que los casos de olas de calor e inversiones de temperatura contribuyeron a un aumento de las concentraciones de O3 en un 28 % y un 14 %, respectivamente, sobre la región (Joshi et al., 2020). Un estudio realizado en China reveló que un aumento de 10 μg/m3 en el O3 provocaba un incremento del 0,55 % en la mortalidad asociada a enfermedades respiratorias (Y. Zhang et al., 2021).

La calidad y la contaminación ha sido un problema cada vez más preocupante debido a los efectos pueden variar desde enfermedades respiratorias hasta problemas cardiacos y derrames cerebrales, en donde la población vulnerable, como niños y personas de tercera edad, son los más propensos a sufrir las consecuencias de una mala calidad el aire en la zona donde viven. (OMS,2021). El registro, estudio y análisis de la información nos permite ser capaces de tomar acciones en el presente y ser proyectado.

En Colombia mediante la Resolución 2254 de 2017 expedida por el Ministerio de Medio ambiente, estableció los niveles máximos permisibles de contaminantes criterio mencionados anteriormente, estos y datos meteorológicos son monitoreados por las Autoridades Ambientales (Corporaciones Autónomas Regionales, Corporaciones de Desarrollo Sostenible y Autoridades Ambientales Urbanas) mediante Sistemas de Vigilancia de la Calidad del Aire – SVCA, diseñados y operados conforme a los criterios que el Ministerio de Ambiente y Desarrollo Sostenible ha definido. La información de los contaminantes atmosféricos es reportada por las Corporaciones Autónomas Regionales y las Autoridades Ambientales de los grandes centros urbanos en el Subsistema de Información sobre Calidad del Aire – SISAIRE (Minambiente, 2021).

El Oriente Antioqueño es una de las 9 subregiones que componen el departamento de Antioquia. La subregión aportó en el 2022 el 10.4% del PIB de Antioquia, siendo la segunda subregión de mayor importancia económica en el departamento después del Valle de Aburrá. Según la Cámara de Comercio del Oriente Antioqueño, el producto interno bruto (PIB) de la subregión se duplicó en los últimos 10 años, pasando de COP$5.75 billones en el 2009 a COP$11.06 billones en el 2018. A medida que las capacidades territoriales del área metropolitana del Valle de Aburrá se agotan, las empresas y los habitantes han buscado nuevas posibilidades de expansión en el Oriente Antioqueño. Lo anterior ha llevado a un crecimiento del 59,5% en la cantidad total de empresas ubicadas en la subregión del 2015 al 2022, y un crecimiento de la población del 11% en estos mismos años (Cámara de Comercio, 2023). Este crecimiento exponencial llevó a que en el 2021 el oriente contará con una población de 706.477 habitantes, que representan el 10.4% de Antioquia y el 1.4% de Colombia (Departamento Administrativo Nacional de Estadística -en adelante DANE-, censo 2018 – actualizado 2019).

## Aproximación desde la analítica de datos

Se desarrollaron modelos de regresión de machine learning para predecir cantidades de PM10, PM2.5, O3, SO2, NO2 y CO en el oriente antioqueño según las cantidades de las variables meteorológicas, con esto poder anticipar eventos de contaminación críticos para la región.

## Origen de los datos

El conjunto de datos se obtuvo a través de la plataforma SISAIRE, mediante una búsqueda específica por departamento. En este caso, los archivos fueron subidos por CORNARE desde la estación de monitoreo continuo ubicada en San Antonio de Pereira. Los datos, en formato .csv y con valores promedios diarios, abarcan el periodo entre 2020 y 2024, aunque presentan intervalos de tiempo discontinuos debido a vacíos temporales entre los registros. Este conjunto de datos, de acceso público y disponible gratuitamente en SISAIRE, no tiene restricciones ni condiciones de uso especificadas. En el contexto de este proyecto, se mantendrán sus condiciones originales sin modificaciones, combinaciones, ni usos ajenos a los objetivos planteados.

## Métricas de desempeño

Para evaluar el desempeño de los modelos de aprendizaje automático en la predicción de episodios de contaminación atmosférica, la precisión (accuracy) es una de las métricas más utilizadas, ya que indica el porcentaje de predicciones correctas en relación con el total de predicciones. En este caso, se espera alcanzar una precisión de al menos un 90 %, comparable con la de los métodos tradicionales de monitoreo, que suelen oscilar entre el 85 % y el 95 % (este valor puede variar entre estaciones y debe notificarse junto con su incertidumbre). Además, se evaluará la sensibilidad (recall) del modelo, que muestra el porcentaje de episodios de contaminación correctamente identificados sobre el total de episodios relevantes. Para una detección eficaz de episodios de alta contaminación, una tolerancia de hasta un 10 % en falsos negativos es razonable, lo que implica un recall mínimo del 90 % para asegurar alertas tempranas precisas que ayuden a mitigar riesgos significativos para la salud pública y el ambiente.

Se espera que el modelo predictivo influya en la atención temprana de episodios de contaminación atmosférica, un indicador clave para comprender las tendencias de calidad del aire. El pronóstico, expresado en términos de precisión porcentual, refleja la efectividad de las alertas tempranas y facilita la toma de decisiones con un margen de anticipación significativo. Las concentraciones elevadas de ciertos contaminantes durante periodos prolongados incrementan el riesgo de enfermedades respiratorias y pueden intensificar la degradación ambiental. Con un mayor volumen de datos disponibles, se prevé que el modelo reduzca los tiempos de respuesta ante eventos de contaminación, mejorando la exactitud en el pronóstico de estos episodios.

# Objetivos

## Objetivo general

Desarrollar herramientas de análisis de datos, que permitan a las Autoridades Ambientales y/o entes que les compete identificar patrones en las variables atmosféricas relacionadas con la calidad del aire, contribuyendo de manera efectiva a la prevención de problemas de salud y la mitigación de los efectos negativos de los contaminantes en el Oriente Antioqueño, apoyando la toma de decisiones y la implementación de estrategias de control ambiental.

## Objetivos específicos

* Procesar y analizar datos históricos de calidad del aire y meteorología para identificar patrones relevantes.
* Implementar técnicas estadísticas y de aprendizaje automático para modelar relaciones entre contaminantes y variables meteorológicas.
* Diseñar herramientas visuales que permitan evaluar tendencias y predecir escenarios críticos.
* Evaluar el impacto de las medidas de control ambiental mediante análisis comparativos.
* Implementar un modelo de Machine Learning que logre pronosticar episodios de contaminación atmosférica en el Oriente Antioqueño.

# Datos

Este capítulo presenta la estructura de los datos de la estación utilizados en el estudio, comenzando con un análisis detallado de los datos originales y sus características. A continuación, se describe la preparación de los conjuntos de datos específicos para el modelado, seguido de una analítica descriptiva que examina tendencias en el dataset, proporcionando una base para los análisis posteriores.

## Datos originales

Los datos utilizados provienen de la estación San Antonio de Pereira, proporcionados por CORNARE para los años 2019 a 2024. Estos incluyen mediciones de contaminantes atmosféricos criterio y variables meteorológicas en formato CSV. Cada registro contiene información como presión atmosférica, concentración de contaminantes (CO, NO2, O3, PM10, PM2.5, SO2), velocidad y dirección del viento, temperatura, y precipitación líquida. En la tabla 1 se describen las características de cada una de las columnas de los datos.

**Tabla 1**  *Descripción de los campos almacenados en los archivos CSV.*

|  |  |  |
| --- | --- | --- |
| **Campo** | **Descripción** | **Ejemplo de Dato** |
| Estación | Estación donde se realiza la medición | San Antonio de Pereira |
| Fecha inicial | Fecha y hora de inicio de registro | 01/01/2024 01:00 |
| Fecha inicial | Fecha y hora de fin de registro | 01/01/2024 01:59 |
| P | Presión atmosférica medida | 591.8 |
| CO | Concentración de monóxido de carbono (CO) | 619.8 |
| NO | Concentración de dióxido de nitrógeno (NO₂) | 6.4 |
| O3 | Concentración de ozono (O₃) | 40.7 |
| PM2.5 | Partículas en suspensión de tamaño menor a 2.5 micras (PM2.5) | 8.3 |
| PM10 | Partículas en suspensión de tamaño menor a 10 micras (PM10) | 8.3 |
| DViento | Dirección del viento. | 202.5 |
| HAire10 | HAire10: Humedad del aire (a 10 metros de altura) | 91.0 |
| PLiquida | Precipitación líquida o cantidad de lluvia | 0.25 |
| RGolbal | Radiación solar global | 1.8 |
| TAire10 | Temperatura del aire (a 10 metros de altura) | 17.3 |
| Vviento | Velocidad del viento | 0.9 |
| SO2 | Concentración de dióxido de azufre (SO₂) | 0.9 |

## Datsets

***Limpieza de Datos***

El tratamiento de los datos consistió en la depuración de columnas irrelevantes, estandarización de formatos y unificación de datos por años en un solo conjunto bajo una granularidad diaria. Posteriormente, se imputaron valores nulos utilizando el método KNN Imputer para garantizar la integridad del análisis estadístico y la preparación de los modelos predictivos, el proceso se de limpieza e imputación se puede consultar en el archivo “ME03 - G06 - [1128437044] - [1017239733].ipynb”, alojado en el repositorio de GitHub (<https://acortar.link/HDNCHO>) .

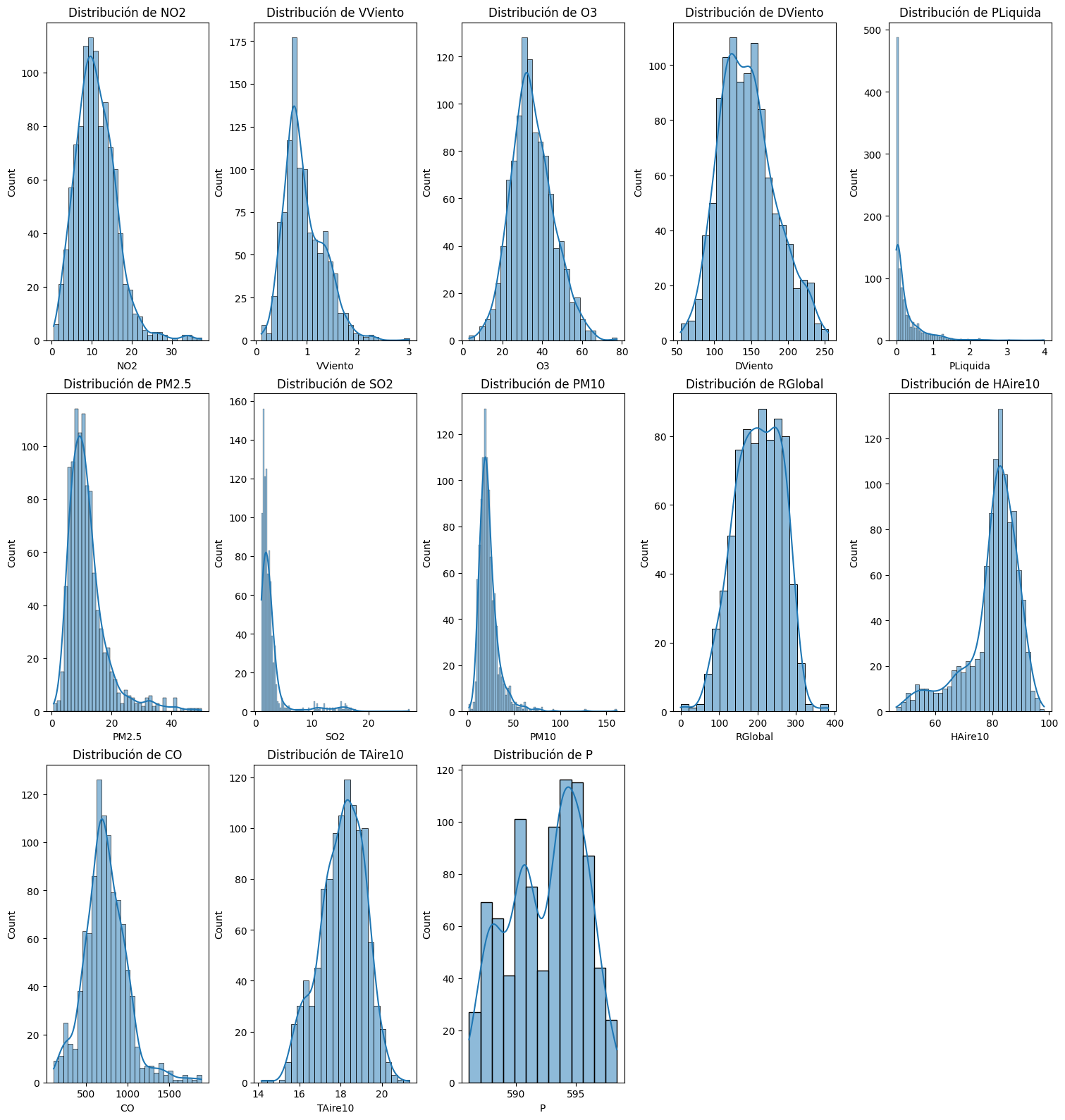
## Analítica descriptiva

El análisis exploratorio de los datos se detalla en el archivo “ME03 - G06 - [1128437044] - [1017239733].ipynb” del repositorio de GitHub asociado (<https://acortar.link/HDNCHO> ). En el análisis de los datos meteorológicos y de calidad del aire para la estación de San Antonio de Pereira (2019-2024), se realizaron las siguientes actividades:

***Distribución de las variables***:

**Figura 1**

*Distribución de los datos para cada una de las variables numéricas*



Nota. Elaboración propia, código disponible en el repositorio asociado al proyecto

Se generaron histogramas para cada variable para identificar patrones de distribución y posibles sesgos. Ejemplo: en la Figura 1 el PM2.5 muestra una distribución sesgada a la derecha, lo que indica episodios de contaminación puntuales.

**Análisis estadístico descriptivo:**

Se calculó la media, mediana, desviación estándar y percentiles para las variables clave. Ejemplo: La media anual de PM10 fue de 45.3 µg/m³, superando el estándar nacional en varias ocasiones.

**Visualización:**

* **Histogramas y diagramas Q-Q:** Evaluaron la normalidad de las distribuciones de las variables. Identificaron distribuciones normales en algunas variables como la presión atmosférica y distribución no normal en otras como el PM2.5.
* **Diagramas de caja y bigotes:** Identificaron valores atípicos en PM10 y PM2.5, relacionados con eventos de contaminación puntuales. Resaltaron valores atípicos en variables como el PM2.5 durante picos de contaminación.

**Matriz de Correlación**

Se calculó una matriz de correlación para explorar relaciones significativas entre los contaminantes y los factores meteorológicos:

**Figura 2**

Matriz de correlación entre los contaminantes y los factores meteorológicos

# C:\Users\snt\Desktop\descarga.png

El análisis descriptivo identificó en la **Figura 2** tendencias críticas, como el aumento de ozono en días cálidos y la dispersión de material particulado en presencia de vientos fuertes. Estos resultados son esenciales para la construcción de modelos predictivos y para la implementación de estrategias de gestión ambiental en el Oriente Antioqueño.

# 

# 

# Referencias

[Palmeira, A., Pereira, É., Ferreira, P., Diele-Viegas, L. M., & Moreira, D. M. (2022). Long-Term Correlations and Cross-Correlations in Meteorological Variables and Air Pollution in a Coastal Urban Region. Sustainability, 14(21), 14470.

Statheropoulos, M., Vassiliadis, N., & Pappa, A. (1998). Principal component and canonical correlation analysis for examining air pollution and meteorological data. Atmospheric environment, 32(6), 1087-1095.

Rad, A. K., Shamshiri, R. R., Naghipour, A., Razmi, S. O., Shariati, M., Golkar, F., & Balasundram, S. K. (2022). Machine learning for determining interactions between air pollutants and environmental parameters in three cities of Iran. Sustainability, 14(13), 8027.

Zapata González, E. A. (2017). Estimación de la calidad del aire en la ciudad de Talca utilizando algoritmos de aprendizaje automático (Tesis de maestría, Universidad de Talca). Facultad de Ingeniería, Universidad de Talca

AlOmar, M. K., Khaleel, F., AlSaadi, A. A., Hameed, M. M., AlSaadi, M. A., & Al-Ansari, N. (2022). The Influence of Data Length on the Performance of Artificial Intelligence Models in Predicting Air Pollution. Advances in Meteorology, 2022, Article ID 5346647. <https://doi.org/10.1155/2022/5346647>

Vásquez Arenas, J. P. (2022). Beneficios de los modelos basados en algoritmos de machine learning para la predicción de concentraciones de PM2.5 en el Valle de Aburrá [Trabajo de grado especialización, Universidad de Antioquia]. Universidad de Antioquia, Medellín, Colombia. ]

# Anexos