



VERANO DE INVESTIGACIÓN DELFIN

PREDICCIÓN DE LA CALIDAD DEL AIRE EN
ZONAS URBANAS DE MÉXICO MEDIANTE
APRENDIZAJE AUTOMÁTICO, UN ENFOQUE
COMPARATIVO INTERNACIONAL ANTE LA
ESCASEZ DE DATOS LOCALES

P R E S E N T A:

Br. Andrés Antonio Chan Cach

Marcos Vallejo Leyva

ASESORA DE INVESTIGACIÓN

Mtra. Edurnet Jhaquelin Luna Becerril

TEXCOCO, EDO. MÉX.

JULIO 2025

Resumen.

La contaminación del aire constituye uno de los desafíos ambientales más urgentes a nivel mundial, con impactos significativos en la salud humana, el cambio climático y la calidad de vida en las zonas urbanas. Este proyecto, enmarcado en el Programa Delfín 2025, propone el uso de técnicas de aprendizaje automático para analizar y predecir la calidad del aire (AQI) en distintas regiones, con especial énfasis en el contexto mexicano.

Para ello, se trabajó con tres conjuntos de datos que incluyen variables ambientales y meteorológicas como temperatura, humedad, presión atmosférica, velocidad del viento y niveles de contaminantes (PM2.5, PM10, O₃, NO₂, CO y SO₂). A través de un riguroso proceso de limpieza, transformación y análisis exploratorio, se desarrollaron modelos de regresión y clasificación con algoritmos como Random Forest, Regresión Lineal y MLPRegressor.

Los resultados muestran que las variables meteorológicas tienen una fuerte relación con el AQI y que el modelo MLPRegressor ofrece el mejor rendimiento predictivo. Asimismo, se definió un índice de contaminantes que permitió comparar la calidad del aire entre países, ubicando a México en una posición intermedia en el ranking internacional.

Este estudio demuestra que el aprendizaje automático es una herramienta poderosa para complementar los sistemas tradicionales de monitoreo ambiental, facilitando la identificación de zonas críticas, la generación de alertas tempranas y el diseño de políticas públicas basadas en datos.

Palabras clave: calidad del aire, AQI, aprendizaje automático, regresión, clasificación, contaminación atmosférica, México.

Índice general

Resumen	II
1. Marco de Referencia	1
1.1. Introducción	1
1.2. Objetivo General	2
1.3. Objetivos Específicos	2
1.4. Alcance	3
1.5. Limitaciones	3
1.6. Problemática	4
1.7. Justificación	5
1.8. Antecedentes	5
2. Marco Teórico	7
2.1. Contaminación atmosférica en contextos urbanos	7
2.2. Sistemas de monitoreo ambiental y disponibilidad de datos en México	8
2.2.1. Disponibilidad y aprovechamiento de datos abiertos para modelado predictivo	9
2.3. Enfoques internacionales y adaptabilidad en contextos con escasez de datos .	9
3. Desarrollo Metodológico	11
3.1. Herramientas utilizadas	11
3.2. Fase 1 Recolección de datos	12
3.3. Fase 2 Limpieza y preprocesamiento	12

3.4. Fase 3 Análisis exploratorio de datos (EDA)	13
3.5. Fase 4 Modelado predictivo	14
3.5.1. Regresión	16
3.5.2. Clasificación	18
3.6. Fase 5 Evaluación del modelo	18
4. Resultados	19
4.1. Features y algoritmos con mayor FoM	19
4.2. Análisis de contaminantes en México y otros países	20
4.2.1. Evolución temporal de valores de AQI de contaminantes	20
4.2.2. Distribución de contaminantes por país	20
4.2.3. Contaminantes predominantes en México	20
4.2.4. Definición del índice de contaminantes	20
4.2.5. Ranking de países según índice de contaminantes	22
4.3. Conclusiones	22

Índice de figuras

1.1. Calidad del Aire en México.	2
3.1. Distribución de valores de AQI por dataset	13
3.2. Conteo de entradas por estado para el dataset 1.	14
3.3. Matriz de correlación de las variables del Dataset 1 transformado.	15
3.4. Valores de AQI medido, mínimo, máximo y promedio para el Dataset 1 transformado.	15
4.1. Distribución de los seis contaminantes principales en tres estados de México en función de la hora del día.	21
4.2. Distribución de contaminantes por país	22
4.3. Ranking de países por índice de contaminantes (IC)	23

Capítulo 1

Marco de Referencia

1.1. Introducción

La contaminación del aire es un problema que afecta cada vez más a las ciudades del mundo. Sus consecuencias no solo se reflejan en el medio ambiente, sino también en la salud de las personas, especialmente en quienes viven en zonas urbanas densamente pobladas. Diversos estudios han demostrado que una mala calidad del aire está directamente relacionada con enfermedades respiratorias, cardiovasculares y un aumento en la mortalidad prematura. En particular, el índice de calidad del aire (AQI, por sus siglas en inglés) es una métrica ampliamente utilizada para evaluar el nivel de contaminación en función de la concentración de distintos contaminantes como el material particulado (PM2.5, PM10), ozono (O_3), dióxido de nitrógeno (NO_2), dióxido de azufre (SO_2) y monóxido de carbono (CO).

En México, existen redes de monitoreo como el SINAICA, sin embargo la información disponible es limitada o fragmentada, lo cual dificulta una evaluación precisa de la calidad del aire en muchas regiones. Esta falta de datos se convierte en un obstáculo para diseñar políticas públicas efectivas o para anticipar los riesgos que representa la exposición constante a contaminantes como las partículas finas, el ozono o el dióxido de nitrógeno.

Frente a este panorama, los avances en ciencia de datos y aprendizaje automático ofrecen una alternativa prometedora. Estas herramientas permiten procesar grandes volúmenes de información, identificar patrones y generar modelos predictivos incluso en contextos con datos parciales. Este proyecto, desarrollado en el marco del Programa Delfín 2025, tiene

como objetivo aplicar modelos de regresión y clasificación sobre distintos conjuntos de datos de calidad del aire provenientes de estaciones distribuidas en México y otros países.

En la Figura 1.1, se presenta una ilustración de las variables clave relacionadas con la calidad del aire en México.



Figura 1.1: Calidad del Aire en México.

1.2. Objetivo General

Diseñar e implementar un modelo predictivo, basado en técnicas de aprendizaje automático, que permita anticipar los niveles de calidad del aire en zonas urbanas de México. Dicho modelo integra datos sobre contaminantes atmosféricos, condiciones meteorológicas y fuentes geoespaciales, para facilitar la toma de decisiones orientadas a la protección del medio ambiente y la salud pública.

1.3. Objetivos Específicos

1. Recolectar y organizar conjuntos de datos públicos internacionales y nacionales sobre calidad del aire, variables meteorológicas y factores geoespaciales.
2. Conocer la distribución y niveles de contaminantes por país, especialmente México, e

identificar si se encuentran dentro de los límites aceptables establecidos por normas internacionales como la OMS.

3. Explorar y analizar las principales tendencias, correlaciones y patrones de comportamiento entre los contaminantes atmosféricos y las variables ambientales, utilizando técnicas de análisis exploratorio de datos.
4. Diseñar y entrenar modelos predictivos de aprendizaje automático que permitan estimar los niveles de calidad del aire, comparando diferentes algoritmos y evaluando su desempeño con métricas estándar.
5. Evaluar la adaptabilidad de los modelos entrenados en contextos urbanos mexicanos e internacionales y las posibles estrategias de transferencia o ajuste.
6. Definir un índice que represente el carácter contaminante de un país para realizar análisis comparativos.
7. Desarrollar una visualización clara y accesible de los resultados del modelo, con el fin de facilitar la interpretación de los datos y brindar una mejor toma de decisiones.

1.4. Alcance

Esta investigación tiene un alcance exploratorio y propositivo, centrado en el diseño, implementación y validación preliminar de un modelo predictivo de calidad del aire basado en técnicas de aprendizaje automático. Aunque el estudio emplea datos internacionales debido a la limitada disponibilidad de información local en muchas regiones de México, su enfoque está orientado al contexto nacional, con la intención de generar una herramienta adaptable y replicable en zonas urbanas mexicanas.

1.5. Limitaciones

Debido a la limitada disponibilidad de datos continuos y abiertos sobre calidad del aire en muchas zonas urbanas de México, esta investigación emplea conjuntos de datos interna-

cionales como punto de partida. No obstante, el modelo desarrollado busca ser replicable y adaptable al contexto nacional, una vez que se fortalezcan los sistemas de monitoreo ambiental en el país.

1.6. Problemática

A pesar del reconocimiento generalizado sobre los efectos perjudiciales de la contaminación del aire en la salud humana, México aún no cuenta con una plataforma nacional sólida, integrada y estandarizada que permita predecir y analizar de forma eficiente la calidad del aire. Si bien existen redes como SINAICA o RAMA, estas presentan limitaciones importantes: su cobertura es desigual, la densidad de monitoreo varía según la región y, en muchos casos, los datos disponibles son incompletos, poco actualizados o difíciles de analizar en formatos accesibles para investigadores y autoridades. En las zonas urbanas, esta situación se ve agravada por el crecimiento acelerado del parque vehicular, la expansión urbana descontrolada y la escasa regulación sobre emisiones. Como resultado, se ha observado un incremento sostenido en contaminantes como partículas finas (PM10 y PM2.5), óxidos de nitrógeno (NO_x), monóxido de carbono (CO) y compuestos orgánicos volátiles (COV). Lo preocupante es que estos contaminantes están presentes no solo en las grandes ciudades metropolitanas, sino también en ciudades intermedias que no cuentan con políticas de mitigación ni sistemas adecuados de monitoreo [Movilidad2023]. Aunque algunas investigaciones aisladas, como las realizadas en Hidalgo o la Ciudad de México, han comenzado a utilizar técnicas de aprendizaje automático para anticipar la calidad del aire, estas iniciativas no han sido replicadas de manera sistemática en otras regiones con necesidades similares. La ausencia de una política pública que fomente el uso de modelos predictivos, sensores de bajo costo y acceso a datos abiertos impide democratizar el conocimiento ambiental y limita la posibilidad de implementar medidas preventivas basadas en evidencia. Ante este escenario, se vuelve prioritario diseñar modelos predictivos que no solo integren datos históricos sobre calidad del aire, sino que también consideren variables meteorológicas, factores socioespaciales y tecnologías de monitoreo accesibles. Una posible solución sería desarrollar un repositorio nacional de datos abiertos, acompañado por una infraestructura tecnológica que permita aplicar algoritmos de

inteligencia artificial para generar alertas tempranas, estimar riesgos poblacionales y apoyar la toma de decisiones en salud y medio ambiente. En definitiva, la creación de un modelo replicable y escalable para predecir la calidad del aire en zonas urbanas mexicanas podría convertirse en una herramienta clave. Esto permitiría cerrar brechas en materia de monitoreo ambiental, reducir riesgos sanitarios y avanzar hacia una gestión más eficiente y equitativa de los compromisos ambientales del país.

1.7. Justificación

La predicción de la calidad del aire mediante modelos de aprendizaje automático ofrece un enfoque moderno, automatizado y escalable para complementar las estrategias actuales de vigilancia ambiental. Esta investigación tiene justificación en tres dimensiones principales:

- **Salud pública:** Una estimación precisa del AQI permite alertar a la población vulnerable (niños, adultos mayores y personas con enfermedades respiratorias) ante condiciones ambientales desfavorables.
- **Toma de decisiones:** Los gobiernos y organismos ambientales podrán utilizar los modelos propuestos para establecer políticas de reducción de emisiones, definir restricciones temporales y ubicar estaciones de monitoreo estratégicamente.
- **Aplicabilidad tecnológica:** El enfoque propuesto demuestra cómo los datos abiertos pueden ser aprovechados con técnicas de inteligencia artificial, impulsando el desarrollo de soluciones digitales sostenibles.

Además, al centrarse en el contexto mexicano y regional, esta propuesta aporta evidencia localizada sobre los patrones de contaminación, su estacionalidad y su relación con las condiciones climáticas.

1.8. Antecedentes

La calidad del aire en las ciudades es una preocupación creciente a nivel global. La contaminación atmosférica no solo impacta directamente en la salud de millones de personas,

sino que también representa un desafío ambiental complejo y persistente. La Organización Mundial de la Salud (OMS) señala que más del 92 % de la población mundial respira aire contaminado a diario.

En América Latina, y particularmente en México, esta situación se agrava por factores como la limitada cobertura de los sistemas de monitoreo, la desigualdad en la distribución de tecnologías y una urbanización acelerada. El crecimiento del parque vehicular, la falta de planeación urbana y la expansión desordenada de las ciudades han contribuido al deterioro de la calidad del aire.

Estudios internacionales han demostrado el potencial del aprendizaje automático para abordar este problema. Por ejemplo, Kumar et al. (2020) emplearon Random Forest para estimar PM2.5 en India; Li y Zhao (2021) aplicaron redes neuronales para predecir el AQI en Beijing; y González y Méndez (2022) usaron modelos de clasificación en Bogotá y Santiago de Chile.

En México, investigaciones recientes en San Luis Potosí y la Ciudad de México han utilizado sensores móviles y algoritmos como SVM, C4.5 y redes neuronales para predecir contaminantes con precisión superior al 90 %. No obstante, estos avances aún son aislados y requieren estrategias nacionales que permitan su replicación y escalabilidad.

Capítulo 2

Marco Teórico

2.1. Contaminación atmosférica en contextos urbanos

La contaminación atmosférica es un fenómeno complejo derivado principalmente de actividades humanas en zonas densamente pobladas. Las principales fuentes urbanas incluyen el transporte motorizado, las industrias, las emisiones residenciales, la construcción y, en menor medida, procesos naturales como el polvo y los incendios forestales. Estas fuentes generan contaminantes como material particulado (PM10, PM2.5), óxidos de nitrógeno (NO_x), dióxido de azufre (SO_2), monóxido de carbono (CO) y ozono troposférico (O_3), los cuales afectan tanto la salud humana como el medio ambiente.

En los entornos urbanos, estos contaminantes tienden a concentrarse por el efecto de las condiciones meteorológicas, las barreras arquitectónicas y la densidad vehicular. Particularmente preocupante es la exposición prolongada a niveles elevados de PM2.5, la cual se asocia con enfermedades respiratorias crónicas, cardiovasculares y aumento en la tasa de mortalidad prematura.

Además de los impactos en la salud, la contaminación del aire urbano contribuye al cambio climático a través de emisiones de gases de efecto invernadero (GEI) y afecta la calidad de vida de los habitantes al reducir la visibilidad, deteriorar estructuras y limitar actividades al aire libre.

En países como México, los problemas se agravan en ciudades con crecimiento urbano desordenado, falta de regulación eficaz o deficiencia en el transporte público. Estas caracte-

rísticas hacen indispensable el desarrollo de sistemas predictivos y herramientas inteligentes que permitan una vigilancia proactiva de la calidad del aire en tiempo real.

2.2. Sistemas de monitoreo ambiental y disponibilidad de datos en México

En México, la calidad del aire es evaluada principalmente por el Sistema Nacional de Información de la Calidad del Aire (SINAICA), administrado por la Secretaría de Medio Ambiente y Recursos Naturales (SEMARNAT). SINAICA recopila datos en tiempo real provenientes de más de 120 estaciones de monitoreo automático distribuidas en zonas urbanas e industriales del país.

Sin embargo, este sistema presenta diversas limitaciones:

- Las estaciones están desigualmente distribuidas, con una mayor concentración en zonas metropolitanas como la Ciudad de México, Guadalajara y Monterrey, mientras que regiones del norte y sur del país muestran una cobertura limitada.
- Existen inconsistencias en la periodicidad y calidad de los datos, con periodos sin mediciones o registros incompletos, lo que afecta la continuidad de las series temporales.
- La plataforma SINAICA, aunque pública, tiene una interfaz limitada para consultas masivas o automatizadas, lo que dificulta el acceso sistemático a grandes volúmenes de datos ambientales.

A pesar de estas restricciones, el sistema representa una valiosa fuente de información para la validación de modelos y para estudios históricos de contaminación. En este contexto, el uso de fuentes complementarias como OpenAQ, que recopila datos de calidad del aire a nivel global, y de sensores móviles o redes ciudadanas, puede ampliar la cobertura espacial y temporal del monitoreo.

Dada la naturaleza fragmentada de los datos, el uso de algoritmos robustos para el procesamiento y predicción se vuelve fundamental. Tal como menciona Lin et al. (1999), la

integración de métodos estadísticos y de inteligencia computacional permite compensar la falta de datos completos o ruidos significativos en los registros [lin1999robust].

2.2.1. Disponibilidad y aprovechamiento de datos abiertos para modelado predictivo

El creciente acceso a datos abiertos ambientales ha permitido el desarrollo de soluciones de ciencia de datos aplicadas al monitoreo ambiental. Plataformas como OpenAQ, WAQI (World Air Quality Index) y SINAICA ofrecen registros históricos y en tiempo real que pueden ser combinados con variables meteorológicas obtenidas de servicios como OpenWeather o NOAA.

Estos datos pueden ser procesados mediante herramientas de machine learning para construir modelos predictivos del AQI, detectar anomalías, inferir condiciones futuras y optimizar la colocación de estaciones de monitoreo. No obstante, el éxito de estas aplicaciones depende de una correcta limpieza, transformación y estandarización de los datos, así como de su contextualización geográfica.

En el caso específico de México, los investigadores enfrentan el reto de trabajar con datasets incompletos, pero con un potencial significativo si se integran con datos auxiliares como series climáticas, datos satelitales o información sociodemográfica.

Esto resalta la importancia de generar soluciones locales y de código abierto que permitan automatizar la descarga, análisis y visualización de datos, y faciliten su uso tanto para la comunidad académica como para tomadores de decisiones.

2.3. Enfoques internacionales y adaptabilidad en contextos con escasez de datos

A nivel internacional, diversos países han desarrollado sistemas sofisticados de monitoreo y predicción de la calidad del aire basados en sensores distribuidos, modelado meteorológico y aprendizaje automático. Sin embargo, estos enfoques no siempre son directamente transferibles a regiones con limitaciones técnicas o presupuestales, como es el caso de muchas ciudades en América Latina, África y Asia meridional.

Por ejemplo, en Estados Unidos y la Unión Europea, los modelos de predicción de la calidad del aire combinan datos satelitales, estaciones fijas, modelos meteorológicos tridimensionales y datos sociodemográficos. Sistemas como el Community Multiscale Air Quality (CMAQ) o el Copernicus Atmosphere Monitoring Service (CAMS) son capaces de estimar la concentración de contaminantes en resoluciones espaciales finas y horizontes de tiempo extendidos. No obstante, su implementación requiere recursos computacionales, datos de entrada consistentes y personal capacitado.

En contraste, en países con infraestructuras limitadas, se ha explorado el uso de modelos ligeros que pueden entrenarse con datasets parciales, incluso con datos recolectados por sensores de bajo costo o con fuentes abiertas como OpenAQ. Estas soluciones, aunque menos precisas que los modelos completos, permiten obtener resultados útiles en zonas donde antes no existía monitoreo alguno.

Un enfoque adaptativo consiste en entrenar modelos generalistas con datos internacionales y luego afinarlos con datos locales disponibles mediante técnicas de transferencia de aprendizaje o fine-tuning. También se han utilizado métodos de imputación avanzada para reconstruir valores faltantes en series temporales, incluyendo el uso de redes neuronales recurrentes (RNNs), interpolación bayesiana y k-nearest neighbors (k-NN).

Además, se ha demostrado que incluso modelos simples, como la regresión lineal o los árboles de decisión, pueden ser eficaces si se seleccionan adecuadamente las variables de entrada y se realiza una buena ingeniería de características. Estos modelos son más interpretables y fáciles de implementar, lo cual es especialmente relevante en contextos donde la explicación del modelo es tan importante como su precisión.

El éxito de estos enfoques adaptativos radica en la capacidad de reutilizar conocimientos previos, adaptar soluciones según la disponibilidad local de datos y simplificar arquitecturas sin comprometer totalmente la calidad de las predicciones. Este tipo de soluciones son especialmente valiosas para contextos como el mexicano, donde existen regiones con gran heterogeneidad en la cantidad y calidad de datos ambientales.

Capítulo 3

Desarrollo Metodológico

El desarrollo de esta investigación se llevará a cabo en cinco fases interrelacionadas, diseñadas para abordar de manera ordenada el reto de predecir la calidad del aire en zonas urbanas de México mediante modelos de aprendizaje automático, incluso ante la escasez de datos locales.

3.1. Herramientas utilizadas

Durante el desarrollo de este proyecto se emplearon diversas herramientas de análisis, visualización y modelado, destacando:

- **Python 3.11** como lenguaje principal de programación.
- **Pandas y NumPy** para manipulación y análisis de datos tabulares.
- **Matplotlib y Seaborn** para generación de gráficas exploratorias y estadísticas.
- **Scikit-learn** para implementación de modelos de regresión y clasificación, escalado de variables, validación cruzada y evaluación.
- **Google Colab** como entorno colaborativo para ejecutar notebooks en la nube.

3.2. Fase 1 Recolección de datos

En esta primera etapa se identificaron y recopilaron conjuntos de datos abiertos sobre calidad del aire, meteorología y variables geoespaciales, priorizando fuentes internacionales con cobertura urbana detallada. Aunque el enfoque está en México, la carencia de datos nacionales actualizados y estandarizados justifica el uso de información de otras regiones comparables.

Se obtuvieron tres conjuntos de datos o *datasets* que contenían información correspondiente a diversos países. Se describen las principales características de estos datasets.

El primer dataset o dataset 1 contiene columnas categóricas que indican país, estado y ciudad de la lectura, así como la fecha, el contaminante medido y el contaminante predominante. Sus columnas numéricas indican la latitud y longitud de la estación, los valores de AQI mínimo, máximo, promedio y medido, y valores de temperatura (°C) y humedad (%).

El segundo dataset o dataset 2 tiene como columnas categóricas al país y estado. No se da registro de la fecha de medición. Las columnas numéricas son temperatura (°C), humedad (%) velocidad del viento (km/h), presión (hPa), la hora del día y la medición de AQI. El tercer dataset o dataset 3 es una extensión del dataset 2, obtenido a partir de la inclusión de seis columnas correspondientes a los contaminantes.

3.3. Fase 2 Limpieza y preprocesamiento

Una vez recopilada la información, se aplicaron técnicas de depuración para eliminar duplicados, manejar valores faltantes y homogeneizar unidades. Se ajustaron variables para asegurar consistencia entre las distintas fuentes y se normalizaron los datos para prepararlos para su análisis y modelado.

Originalmente el dataset 1 contenía 4,620 filas, mientras que el 2 y 3 contenían 5000. Después de eliminar valores nulos las filas del Dataset 1 se redujeron a 4,538. Los Datasets 2 y 3 no tenían valores nulos. Para el Dataset 1 se observó que las columnas de latitud y longitud no contenían duplicados, por lo que cada fila en este dataset representa una estación única. Este dataset contiene datos desde el 18 de Junio de 2024 hasta el 20 de Junio de 2025.

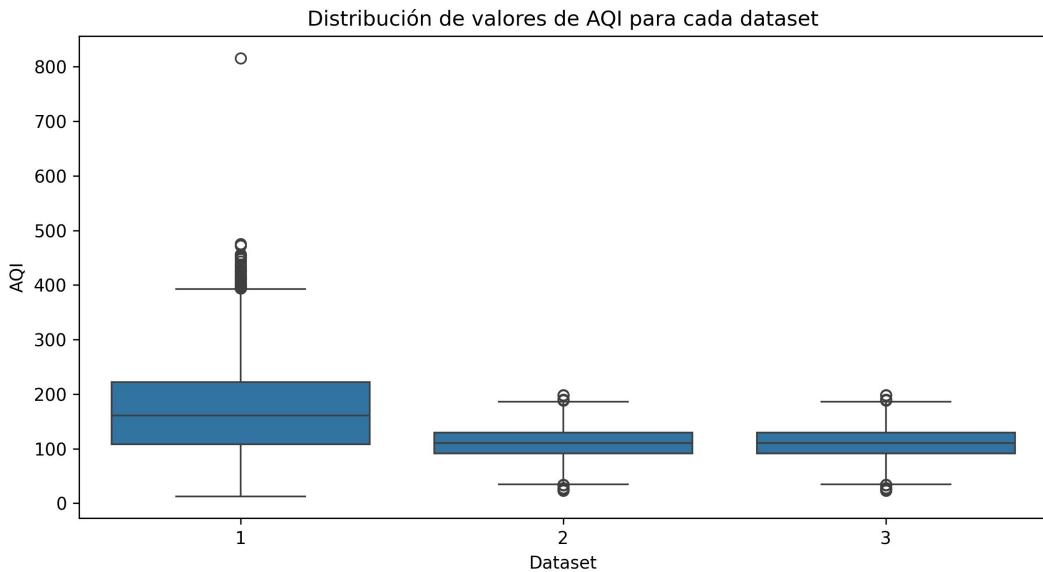


Figura 3.1: Distribución de valores de AQI por dataset

También se encontró un valor inconsistente para el AQI de 800. El valor máximo teórico es de 500, por lo que se elimina esta entrada.

3.4. Fase 3 Análisis exploratorio de datos (EDA)

Esta etapa permitió comprender el comportamiento y las relaciones entre las variables. A través de estadísticas descriptivas y gráficas, se buscaron patrones clave, tendencias temporales y correlaciones entre contaminantes, clima y ubicación geográfica.

El dataset 1 tiene una distribución considerablemente uniforme en cuanto al número de datos por estado. De la Figura 3.2 es posible observar que el valor mínimo de entradas para un estado es de poco menos que 200, y el máximo es cercano a 250.

Previo a los restantes análisis exploratorios se llevó a cabo la transformación de las columnas de los datasets. Para el Dataset 1, el cual incluye datos temporales, se desglosaron los mismos en año, mes, día, hora y minuto para convertir la columna Fecha en cinco columnas numéricas. Posteriormente se aplicó la transformación de las columnas empleando la clase StandardScaler de scikit-learn. Los Datasets 2 y 3 no incluyen datos temporales, por lo que se procedió a transformar sus columnas sin procesos previos. Por otro lado, la codificación de

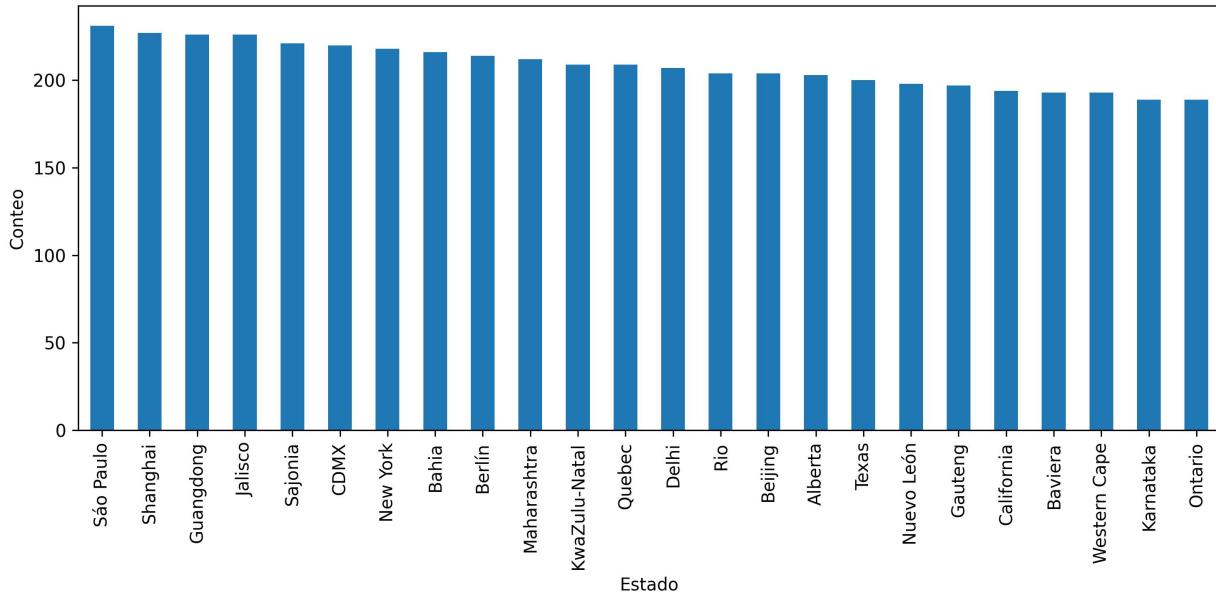


Figura 3.2: Conteo de entradas por estado para el dataset 1.

las columnas categóricos de los tres datasets se llevó a cabo empleando la clase LabelEncoder de scikit-learn.

De la Figura 3.3 se observa que la propiedad AQI tiene una correlación casi de 1 con Avg, así como valores de correlación positiva altos para las propiedades de Min y Max. Se observan otras variables correlacionadas pero no relevantes como Latitud y Estado, o Fecha_anio y Fecha_mes. Además de estas, no hay indicios de otras variables correlacionadas.

3.5. Fase 4 Modelado predictivo

Se entrenaron y compararon distintos modelos de aprendizaje automático como regresión lineal, random forest y redes neuronales profundas. La meta es identificar cuál algoritmo ofrece mejores resultados al predecir niveles de contaminación bajo distintos escenarios urbanos.

A cada dataset se le aplicaron tres modelos de regresión (random forest, regresión lineal y red neuronal) múltiples veces con el propósito de variar los features que se introducían a los modelos para estudiar la influencia de estos en la predicción del AQI. Para la clasificación, se definió un valor de corte (threshold) del AQI, el cual es la mediana de sus valores. Por lo tanto, la clasificación es binaria (arriba o debajo del valor de corte). Solo se empleó un

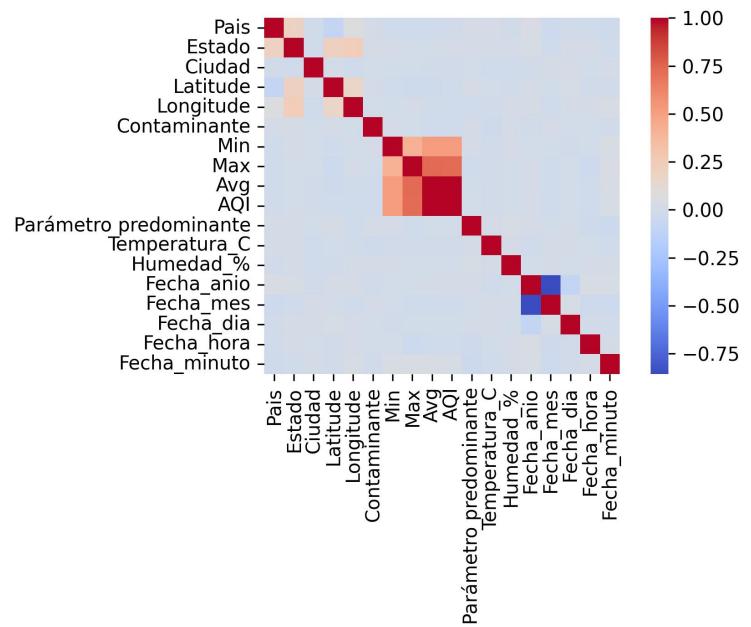


Figura 3.3: Matriz de correlación de las variables del Dataset 1 transformado.

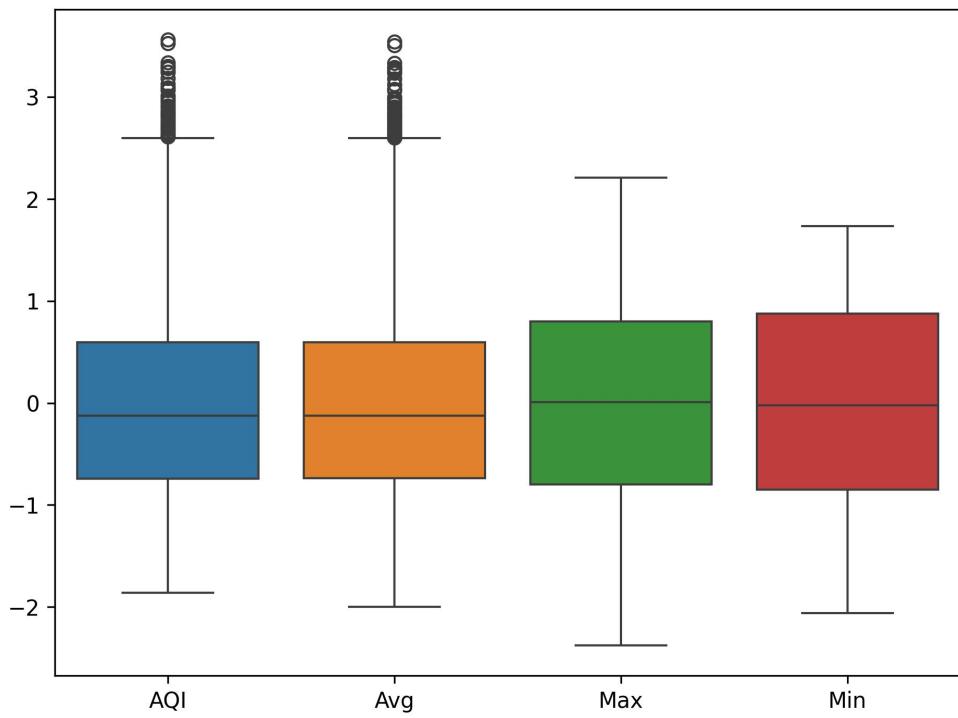


Figura 3.4: Valores de AQI medido, mínimo, máximo y promedio para el Dataset 1 transformado.

algoritmo de clasificación para los tres datasets (RandomForestClassifier).

3.5.1. Regresión

Se ha mencionado que se diseñaron tres modelos de regresión (Random Forest, Regresión Lineal y Red Neuronal) y estos fueron aplicados a los tres datasets múltiples veces variando los features en cada caso. Los pasos son:

1. **Establecer una lista de listas de features.** Por ejemplo:

[[Pais, Estado, Min]

[Temperatura, Humedad, Estado]

[Estado, Min, Max, Contaminante]

...]

Cada lista representa un conjunto de features con los que se quiere probar los tres modelos. Con la lista de ejemplo mencionada, se tendrían $3 \times 3 = 9$ ejecuciones totales entre los 3 modelos.

2. **Ejecutar los tres modelos de regresión para cada lista de features.** Se obtienen tres métricas: MAE, RMSE y R^2 .
3. **Crear un dataframe con los datos de cada ejecución y las métricas obtenidas.**

Esto servirá para obtener una figura mérito y poder obtener una calificación general de cada combinación de features y modelo.

Los siguientes son los parámetros empleados en los modelos.

- Random Forest: Implementado con RandomForestRegressor, con n_estimators=100.
- Regresión lineal: Implementado con LinearRegression.
- Red neuronal: Implementado con MLPRegressor, con hidden_layer_sizes=(50,30) y max_iter=500.

A continuación se detallan las listas de features empleadas para cada dataset.

■ Dataset 1:

1. Pais, Estado
2. Pais, Estado, Contaminante
3. Pais, Estado, Fecha_anio, Fecha_mes, Fecha_dia
4. Min, Max, Avg
5. Contaminante, Parámetro predominante
6. Temperatura_C, Humedad_ %
7. Temperatura_C, Humedad_ %, Parámetro predominante
8. Temperatura_C, Humedad_ %, Contaminante
9. Temperatura_C, Humedad_ %, Fecha_anio, Fecha_mes, Fecha_dia, Fecha_hora

■ Dataset 2:

1. Temperatura_C, Humedad_ %, Velocidad_Viento_kmh, Hora_dia
2. Pais, Estado, Presion_hPa
3. Humedad_ %, Presion_hPa, Estado
4. Pais, Temperatura_C, Velocidad_Viento_kmh
5. Velocidad_Viento_kmh, Hora_dia
6. Humedad_ %, Velocidad_Viento_kmh
7. Temperatura_C, Hora_dia
8. Velocidad_Viento_kmh, Presion_hPa
9. Temperatura_C, Humedad_ %, Velocidad_Viento_kmh

■ Dataset 3:

1. Temperatura_C, Humedad_ %, Velocidad_Viento_kmh, Hora_dia
2. Pais, Estado, PM10],

3. Humedad_ %, PM2.5, Estado],
4. Pais, Temperatura_C, Velocidad_Viento_kmh, PM10],
5. Velocidad_Viento_kmh, Hora_dia, PM2.5],
6. Humedad_ %, Velocidad_Viento_kmh],
7. Temperatura_C, Hora_dia],
8. Velocidad_Viento_kmh, Presion_hPa],
9. Temperatura_C, Humedad_ %, Velocidad_Viento_kmh,],
10. Todas las columnas a excepción de la columna target (AQI)

3.5.2. Clasificación

Se empleó un único algoritmo de clasificación para los tres datasets, siendo este Random Forest. Se empleó la mediana de la columna AQI del dataframe transformado como valor de corte o threshold. Este es un problema de clasificación binaria. La columna objetivo es calidad_aire y determina si el valor de AQI está por debajo o por encima de la mediana.

3.6. Fase 5 Evaluación del modelo

Los modelos se evaluaron con métricas como el error absoluto medio (MAE), el error cuadrático medio (RMSE) y el coeficiente de determinación (R^2). Este análisis permite validar la precisión y robustez del modelo, así como detectar posibles limitaciones o sesgos.

Para determinar las columnas que tienen una mayor influencia en la predicción de AQI se emplean las columnas de MAE, RMSE y R^2 . Primero se calculan los z-scores de los valores de estas columnas empleando StandardScaler. Posteriormente se define una figura de mérito (FoM) como el componente principal obtenido al aplicar PCA a las columnas de z-score. Esta figura de mérito permite ordenar las diferentes combinaciones de features y algoritmos empleados, donde un mayor valor de la FoM es preferible.

Capítulo 4

Resultados

4.1. Features y algoritmos con mayor FoM

Debido a que la columna Avg en el Dataset 1 tiene una correlación prácticamente de 1 con la variable objetivo, se descarta la lista de features que contiene esta columna, la cual es la que tuvo la mayor FoM. Esta lista es [Min, Max, Avg], donde los tres primeros puestos corresponden a esta lista con los tres algoritmos. Por lo tanto, la lista de features y algoritmo que mejor desempeño mostraron fue la que ocupaba la cuarta posición del ranking correspondiente a este dataset, esta es la lista [Pais, Estado, Contaminante] con el algoritmo MLPRegressor. Sus métricas fueron: MAE: 69.17, RMSE: 85.50, R2: 0.003, FoM: -0.46. Esto sugiere que este dataset no permite predecir los valores de AQI de la manera deseada.

Para el Dataset 2 la lista con mejor desempeño fue la lista [Temperatura_C, Humedad_%, Velocidad_Viento_kmh, Hora_dia] con MLPRegressor, con métricas: MAE: 8.11, RMSE: 10.21, R2: 0.86, FoM: 2.635. Esta lista y algoritmo son también los que mejor desempeño mostraron para el Dataset 3, con métricas: MAE: 8.12, RMSE: 10.12, R2: 0.86, FoM: 2.279. Estas métricas permiten observar que los Datasets 2 y 3 proporcionan información suficiente para predecir los valores de AQI con error mínimo. En comparación, el Dataset 1 genera resultados con dimensiones de error de un orden mayor.

4.2. Análisis de contaminantes en México y otros países

4.2.1. Evolución temporal de valores de AQI de contaminantes

La figura 4.1 muestra las series temporales asociadas a los seis principales contaminantes para tres estados de México. Específicamente, las gráficas muestran las variaciones de la mediana de los valores de AQI de estos contaminantes en función de la hora del día. Dichos valores de AQI corresponden a los Datasets 2 y 3.

4.2.2. Distribución de contaminantes por país

Se graficaron los niveles de contaminantes (PM2.5, PM10, NO₂, SO₂, CO y O₃) por país, con especial atención en México. La Figura 4.2 muestra la distribución de cada contaminante para los distintos países del Dataset 3. En general, se observa que México presenta niveles moderados, con algunos picos de PM10 y O₃ en ciudades industriales.

4.2.3. Contaminantes predominantes en México

Analizando el número de apariciones de cada contaminante como “parámetro predominante” dentro del Dataset 1 (columna ‘Parámetro predominante’), se observó que los más comunes son PM2.5 y O₃, con más del 70 % de los registros. Esto indica que estos contaminantes son los principales indicadores de la calidad del aire en zonas urbanas de México.

4.2.4. Definición del índice de contaminantes

Para comparar la contaminación entre países, se propone el siguiente índice de contaminantes (IC):

$$IC_{pais} = \frac{1}{N} \sum_{i=1}^N AQI_i$$

Donde AQI_i representa el índice de calidad del aire de cada estación o ciudad dentro del país, y N es el número total de registros asociados a dicho país. Este índice captura la tendencia general de la contaminación observada en el país.

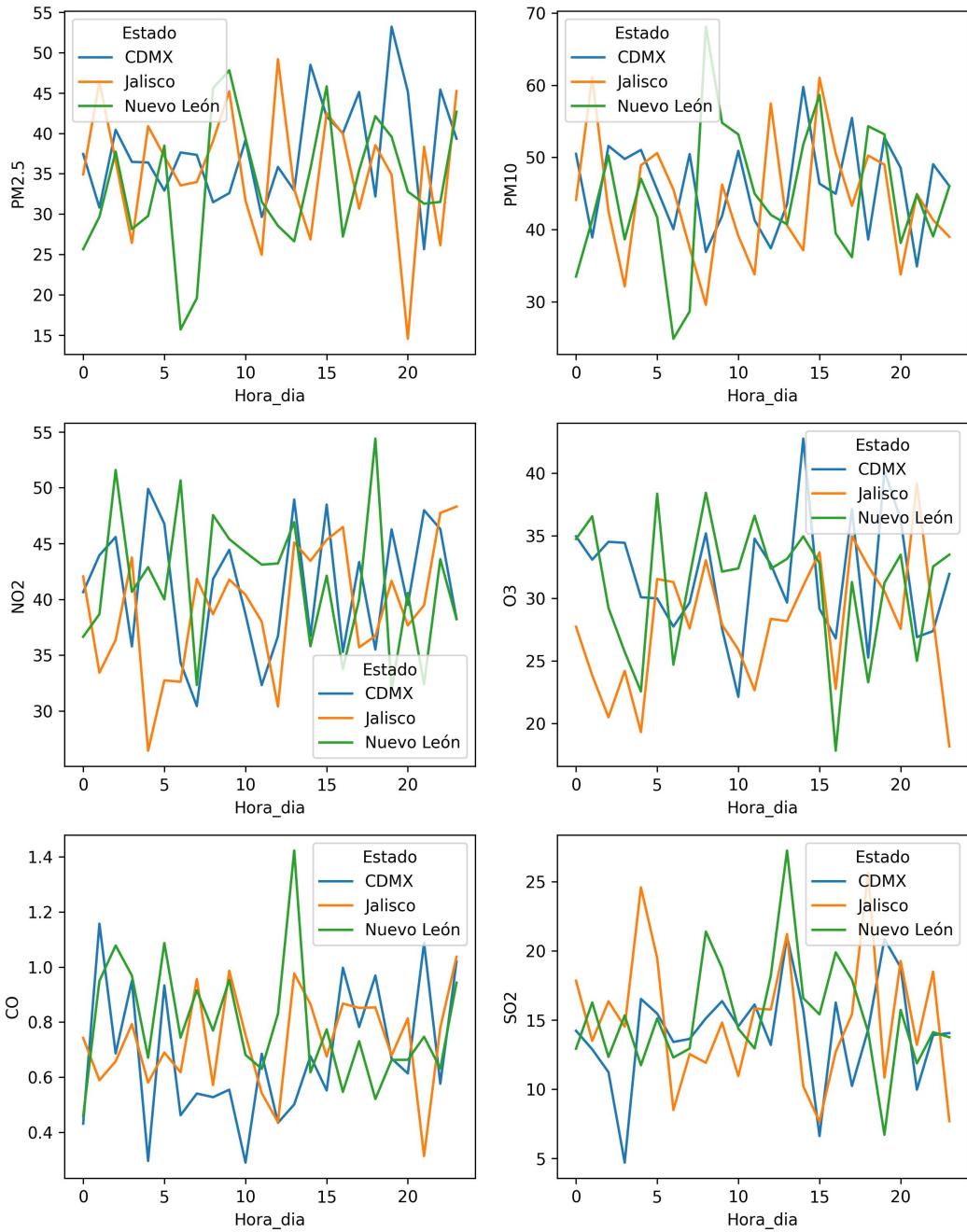


Figura 4.1: Distribución de los seis contaminantes principales en tres estados de México en función de la hora del día.

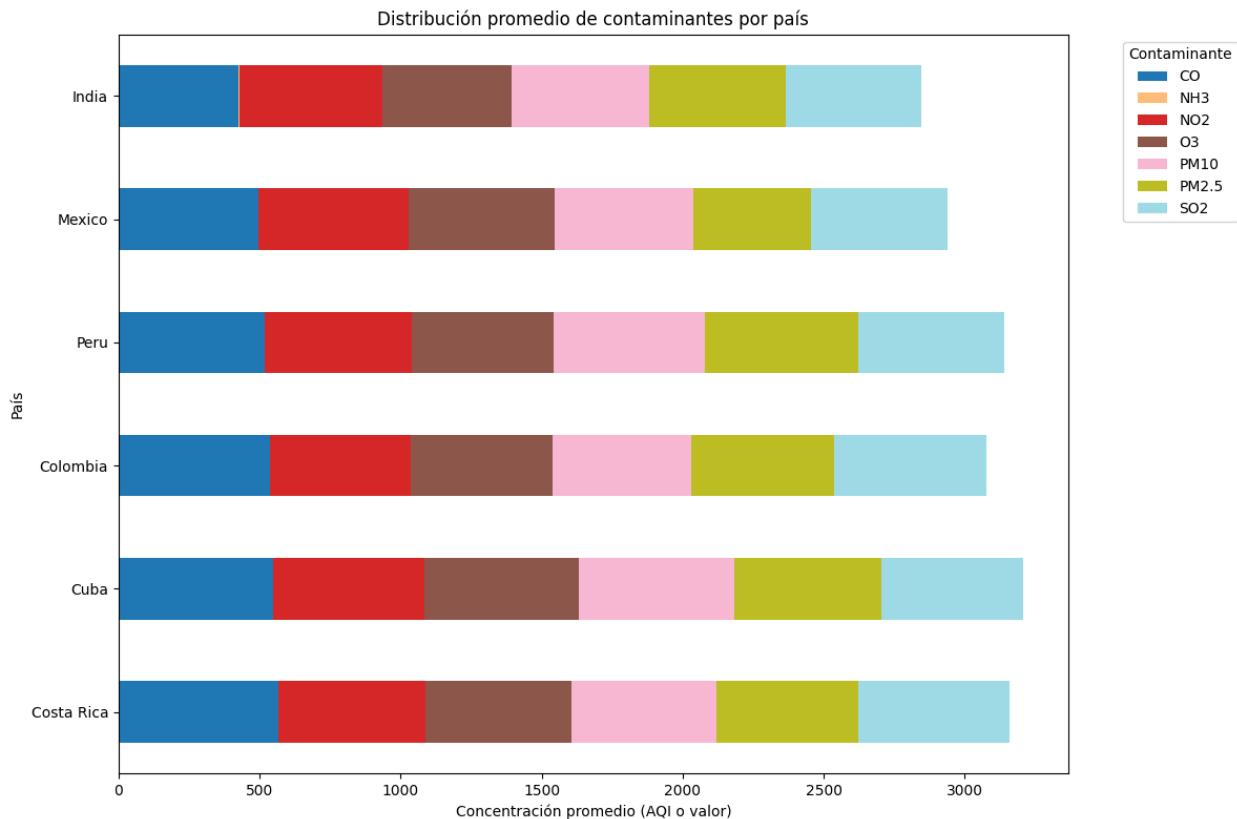


Figura 4.2: Distribución de contaminantes por país

4.2.5. Ranking de países según índice de contaminantes

A partir del índice definido, se obtuvo un ranking de países según su nivel promedio de contaminación (ver Figura 4.3). México ocupa el quinto puesto, donde el cuarto y sexto puestos son ocupados por Colombia e India, respectivamente.

4.3. Conclusiones

Los análisis revelaron que México tiene como principales contaminantes al PM2.5 y al ozono, especialmente en zonas urbanas. En cuanto a la determinación de la influencia de las variables en la predicción de los valores de AQI, las variables meteorológicas como temperatura, humedad y velocidad del viento resultaron ser altamente predictivas en los datasets 2 y 3. Los menores valores de error y, consecuentemente, los mayores valores de la figura de mérito fueron alcanzados con el uso de redes neuronales. Específicamente, en los Datasets

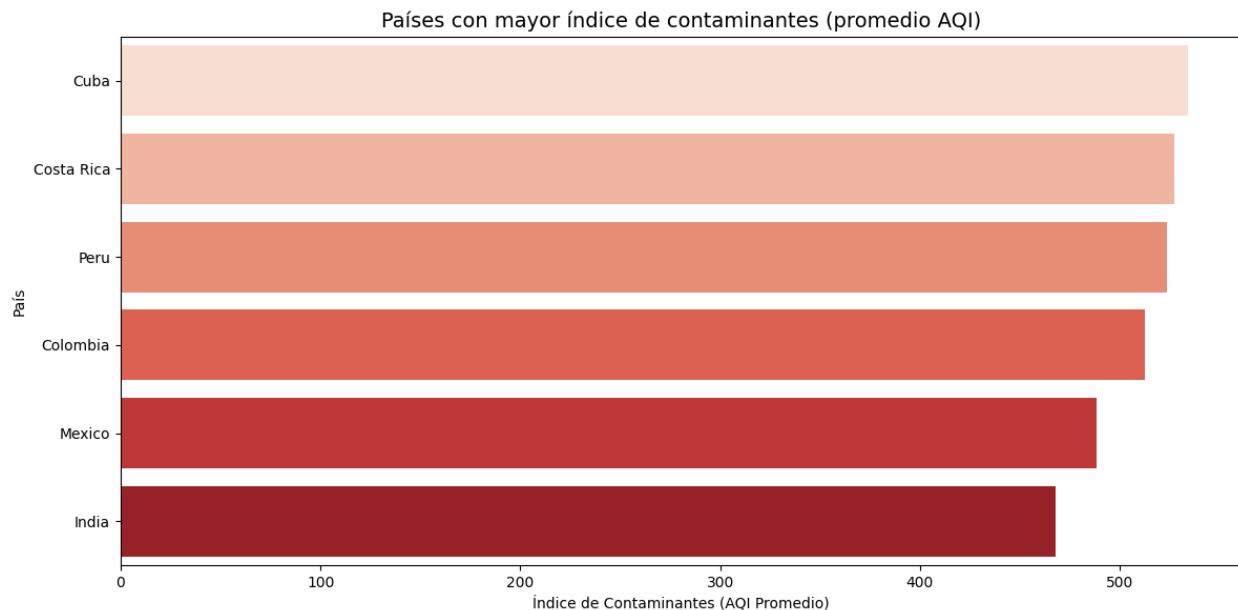


Figura 4.3: Ranking de países por índice de contaminantes (IC)

2 y 3 el modelo que mejor desempeño mostró para predecir el AQI fue el MLPRegressor usando las variables [Temperatura_C, Humedad_%, Velocidad_Viento_kmh, Hora_dia]. En contraste, ninguna combinación de variables del Dataset 1 que no contuviera la variable Avg (la cual presentaba una correlación prácticamente igual a 1 con la variable objetivo) mostró valores de error por debajo de 69 y 85 para el MAE y RMSE, respectivamente.

La creación de un índice de contaminantes permitió establecer un ranking entre países para cuantificar el carácter contaminante de cada uno de ellos. México se encuentra en una posición intermedia, lo que sugiere que aún existe margen para mejorar las condiciones del aire. Los resultados obtenidos pueden servir como base para sistemas predictivos que apoyen decisiones de política ambiental en zonas urbanas.