Tecnológico de Monterrey

Profesor Titular: Dra. Grettel Barceló Alonso
Profesor Tutor: Dra. María de la Paz Rico

TC5035 - Proyecto Integrador (Gpo 10)

Avance 1: Proyecto Integrador

Visualización interactiva de calidad de aire en AR en aplicaciones móviles con análisis y forecasting con AI y ML

Septiembre 28, 2025

Equipo #56
Paulina Escalante Campbell A01191962

Tabla de Contenidos

Aire: Visualización interactiva de calidad de aire en AR en aplicaciones móviles con análisis y forecasting con AI y ML

1.Objetivo de EAD

Para el proyecto de Aire, como primer avance se busca hacer un análisis exploratorio de datos del dataset que se busca usar para las siguientes semanas. Este avance consiste en elegir las características más relevantes para reducir la dimensionalidad y aumentar la capacidad de generalización del modelo y abordar y corregir los problemas identificados en los datos.

Como contexto, la contaminación del aire en entornos urbanos representa uno de los desafíos de salud pública más críticos de este siglo, con la Organización Mundial de la Salud identificando que 14 de las 15 ciudades más contaminadas del mundo se encuentran en India, mientras que regiones como China y otras áreas de Asia mantienen niveles alarmantes de material particulado PM2.5 y otros contaminantes atmosféricos. En este contexto, el proyecto de revitalización de la aplicación móvil "Aire" busca democratizar el acceso a información ambiental compleja a través de visualizaciones interactivas en realidad aumentada, modelos predictivos de inteligencia artificial, y interfaces centradas en el usuario que permitan a las poblaciones urbanas tomar decisiones informadas sobre su exposición a la contaminación atmosférica.

Este primer avance se enfoca en realizar un análisis exploratorio de datos (EDA) exhaustivo para identificar patrones, tendencias y características relevantes en múltiples datasets globales de calidad del aire que abarcan desde 2010 hasta 2024, con énfasis particular en responder la pregunta central: ¿cuáles ciudades tienen históricamente la peor calidad del aire y qué factores contribuyen a esta problemática? A través de técnicas estadísticas y de visualización, este análisis permitirá seleccionar las características más predictivas para reducir la dimensionalidad del modelo, abordar problemas de calidad de datos como valores faltantes y outliers, e identificar las variables más relevantes para la construcción de modelos de machine learning robustos que posteriormente alimentarán las funcionalidades predictivas y de visualización de la aplicación móvil en desarrollo.

2. Liga a github

Se usa github para documentar y mantener el código y avances de manera ordenada y en un repositorio remoto.

```
None
La liga del repositorio es la siguiente:
-https://github.com/pauescalantec/ProyectoIntegrador_Aire
Este avance se encuentra en su respectivo directorio de Avance1:
-https://github.com/pauescalantec/ProyectoIntegrador_Aire/blob/main/Avance1%2356/Avance1_56.ipynb
```

Se busca crear pull requests para cada avance y tener un directorio de avances semanales.

3. Dataset seleccionado

Dado que el enfoque es en ciudades y a través de al menos un año de datos recopilados, se usará este dataset. Abarca 170 países y más de 300 ciudades, proporciona una vista holística de la dinámica global de la calidad del aire. Enfocado en contaminantes importantes como el Monóxido de Carbono, Ozono, Dióxido de Nitrógeno y Material Particulado (PM2.5), sirve como un recurso valioso para científicos ambientales, formuladores de políticas e investigadores. Los insights derivados de este dataset empoderan a los usuarios para analizar tendencias de calidad del aire, formular políticas efectivas y contribuir a fomentar un planeta más saludable.

Global Air Quality Dataset 🌎

Comprehensive Air Quality Measurements from Major Cities Worldwide 🌍

https://www.kaggle.com/datasets/sazidthe1/global-air-pollution-data/data

Columna	Descripción
country_name	Name of the Country
city_name	Name of the City
aqi_value	Overall AQI value of the city
aqi_category	Overall AQI category of the city
co_aqi_value	AQI value of Carbon Monoxide of the city
co_aqi_category	AQI category of Carbon Monoxide of the city
ozone_aqi_value	AQI value of Ozone of the city
ozone_aqi_category	AQI category of Ozone of the city
no2_aqi_value	AQI value of Nitrogen Dioxide of the city
no2_aqi_category	AQI category of Nitrogen Dioxide of the city
pm2.5_aqi_value	AQI value of Particulate Matter with a diameter of 2.5 micrometers or less of the city
pm2.5_aqi_category	AQI category of Particulate Matter with a diameter of 2.5 micrometers or less of the city

Este dataset fue el más completo y de alta calidad ya que incluye categorías y valores numéricos y es reciente, del año 2024. Otros posibles datasets tenían valores interesantes que pueden ser calculados como variables dependientes y calculadas como: proximidad a

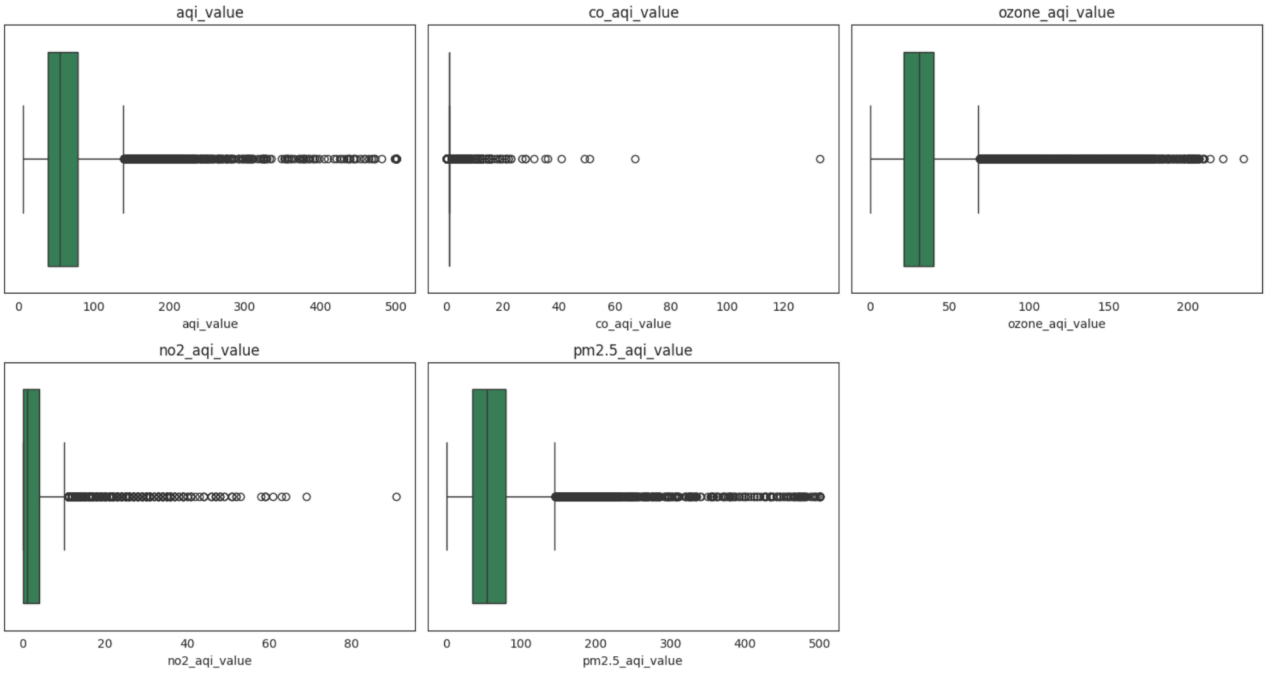
zona industrial (km.) y densidad de población (personas/km2), se considerarán en siguientes avances.

4. Exploración de los datos y observaciones

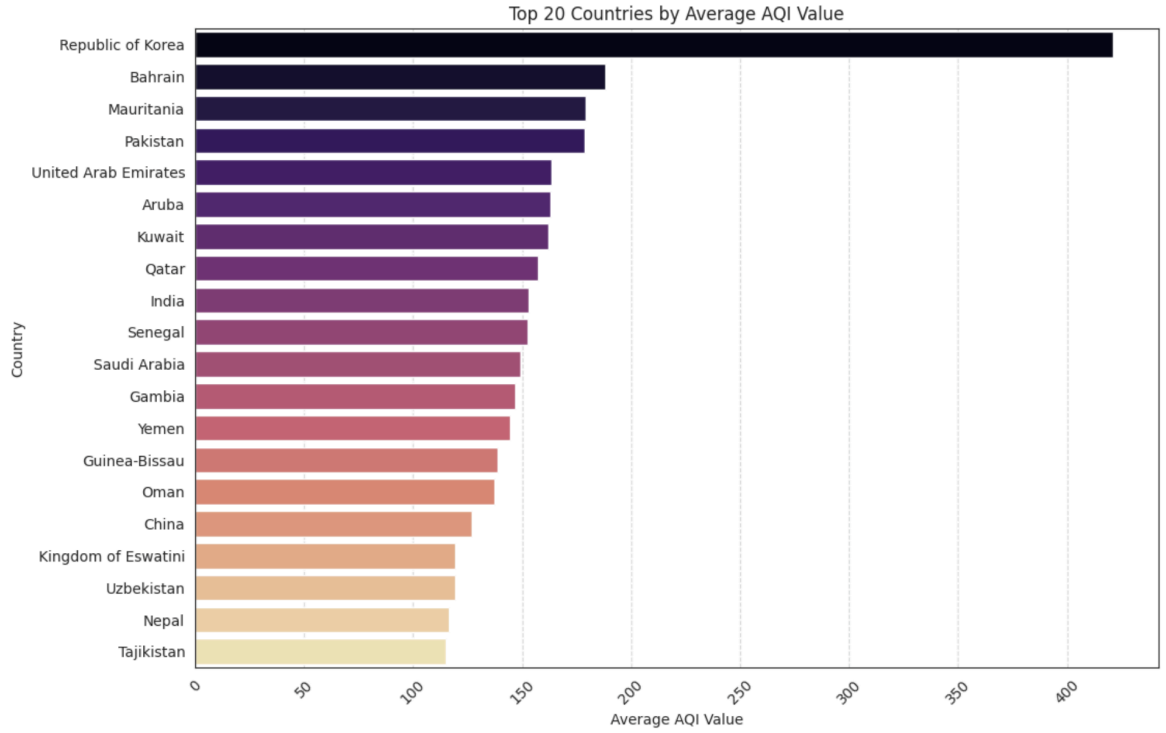
La mayoría de los insights se encuentra en el github en el jupyter notebook:

```
None
https://github.com/pauescalantec/ProyectoIntegrador_Aire/blob/main/Avance1%2356/Avance1_56.ipynb
```

Algunas visualizaciones fueron interesantes cómo la distribución de datos por contaminante, ya que a veces fuegos o volcanes o eventos pueden crear eventos climáticos que empeoran la calidad del aire, es importante considerar los outliers en este caso y no tratar de borrarlos o simplificarlos.



Los países con peor contaminación en 2024 fueron los siguientes:



Muchas observaciones y comentarios se hacen en el jupyter notebook

5. Conclusiones

Este primer avance se centró en comprender el contexto de los datos disponibles sobre calidad del aire y evaluar su estructura para definir una estrategia de análisis y modelado. El propósito general es respaldar el rediseño y extensión de la aplicación móvil Aire, incorporando visualización en realidad aumentada y capacidades de predicción basadas en inteligencia artificial.
En esta fase se trabajó con un conjunto de datos global que reporta medidas de contaminación atmosférica por ciudad, país y tipo de contaminante, enfocándose en las variables numéricas relacionadas con el índice de calidad del aire (AQI) y los principales contaminantes: PM2.5, CO, NO_2 y ozono.

El análisis univariado permitió identificar que todas las variables numéricas relevantes presentan distribuciones sesgadas hacia la derecha y con colas largas. Esto se confirmó mediante el cálculo de coeficientes de skewness y kurtosis, donde la mayoría de las variables superan ampliamente los valores típicos de una distribución normal. En particular, la variable pm2.5_aqi_value se comporta de manera similar al aqi_value, mostrando una correlación muy alta (0.98), lo cual sugiere que el índice general de calidad del aire está fuertemente determinado por los niveles de partículas finas en suspensión. Esta observación refuerza la decisión de considerar pm2.5_aqi_value como la principal variable objetivo para los modelos de predicción en futuras etapas del proyecto.

En el análisis bivariado se examinaron las relaciones entre variables categóricas como aqi_category o country_name y los valores numéricos de los contaminantes. A través de visualizaciones como boxplots y mapas coropléticos, se identificaron diferencias

significativas entre países y categorías. Se destaca que países como India, Pakistán, Irán y Bangladesh concentran altos niveles de PM2.5, y que las categorías de AQI más severas se asocian claramente con valores más elevados de los contaminantes analizados. Además, se generó un mapa interactivo a nivel mundial que muestra el promedio de AQI por país, lo cual permitió visualizar las regiones más afectadas por contaminación atmosférica de forma compacta y comprensible.

Respecto al manejo de calidad de datos, se identificaron valores faltantes en algunas columnas, que fueron tratados mediante imputación con medidas de tendencia central o eliminación en casos con muy baja frecuencia de aparición. En cuanto a los valores atípicos, se utilizaron métodos estadísticos como el rango intercuartílico (IQR) para identificarlos. Sin embargo, **se decidió no eliminarlos, ya que muchos de estos registros reflejan condiciones reales de contaminación extrema**. En su lugar, se propone el uso de modelos robustos como Random Forest y XGBoost, y la aplicación de transformaciones logarítmicas a variables muy sesgadas para mejorar la estabilidad de los modelos sin comprometer la integridad del fenómeno observado.

También se abordó el problema de alta cardinalidad en variables categóricas. Aunque country_name tiene un número manejable de categorías, la variable city_name cuenta con cientos de valores distintos, lo cual podría representar un problema si se codifica como variable categórica tradicional (one-hot). Para este caso, se sugirió emplear técnicas como target encoding o frequency encoding, que permiten reducir la dimensionalidad sin perder información relevante sobre la relación entre las ciudades y la variable objetivo.

El análisis realizado proporciona una comprensión sólida del comportamiento de los datos y establece criterios técnicos claros para su transformación y modelado posterior. Con base en estos hallazgos, se podrán construir modelos predictivos más precisos, escalables y alineados con los objetivos del proyecto, permitiendo así una implementación exitosa de capacidades de inteligencia artificial en la aplicación Aire.

Referencias

INECC. (2020). El Sistema Nacional de Cambio Climático. Calidad del aire. https://cambioclimatico.gob.mx/estadosymunicipios/Aire.html

Instituto Nacional de Ecología y Cambio Climático, (2020). Cumplimiento de la Norma Oficial Mexicana (NOM). Informe Nacional de Calidad del Aire 2017.

Instituto Nacional de Ecología y Cambio Climático, (2020). Estaciones de monitoreo. Sistema de Monitoreo de la Calidad del Aire (SINAICA).

Natalia Garcia Torres, Paulina Escalante Campbell. (2019). Aire: visualize air quality. In ACM SIGGRAPH 2019 Appy Hour (SIGGRAPH '19). Association for Computing Machinery, New York, NY, USA, Article 1, 1–2. https://doi.org/10.1145/3305365.3329869

The World Air Quality Index waqi. (2019). Air Pollution in the World: Frequently Asked Questions. http://aqicn.org/faq/

Governments, C. A. C. of. (n.d.). *Air Central Texas*. English. https://aircentraltexas.org/en/regional-air-quality/aqi

Mateen, M. (2024, December 4). *Air Quality and pollution assessment*. Kaggle. https://www.kaggle.com/datasets/mujtabamatin/air-quality-and-pollution-assessment