

Entrega 1 Proyecto Desarrollo de soluciones

Brayan Sthefen Gomez Salamanca

Juan Sebastian Ordoñez Acuña

Maria Alejandra Rojas Garzon

Hainer Jair Torrenegra Jimenez

Universidad de los Andes

Maestria en Inteligencia Artificial MAIA

Proyecto - Desarrollo de Soluciones

24 de agosto de 2025

Proyección de Calidad de Aire en Risaralda

1. Problema y contexto

La mala calidad del aire es uno de los principales factores del medio ambiente que afectan la salud humana, en los últimos años, ha adquirido una relevancia mayor. Entre los contaminantes más relevantes se encuentran las partículas PM10 (material particulado con diámetro menor o igual a 10 micrómetros) y PM2.5 (diámetro menor o igual a 2.5 micrómetros). Debido a su tamaño microscópico, estas partículas pueden permanecer en el aire durante largos periodos de tiempo y ser inhaladas fácilmente por las personas. Mientras las PM10 tienden a alojarse en las vías respiratorias, las PM2.5, de menor tamaño, pueden incluso ingresar al torrente sanguíneo, lo que las convierte en un contaminante que representa un riesgo para la salud.

La exposición constante partículas PM10 y PM2.5 está asociada a enfermedades a nivel respiratorio, incrementa el riesgo de asma, bronquitis crónica, infecciones respiratorias agudas y reducción de la función pulmonar. A nivel cardiovascular, la exposición sostenida a estas partículas eleva la probabilidad de infartos, hipertensión y enfermedades cerebrovasculares. Asimismo, se ha evidenciado una asociación con cáncer de pulmón, alteraciones en el sistema nervioso central e incluso con disminución cognitiva y riesgo de demencia.

En Colombia, la contaminación atmosférica constituye un desafío para la salud pública. Se estima que una alta proporción de enfermedades respiratorias están asociadas a la exposición a partículas finas. Residuos generados por medios de transporte terrestre, la industria, la quema de biomasa y los incendios forestales son considerados como los principales emisores de este tipo de partículas en el país.

El departamento de Risaralda enfrenta retos significativos en materia de calidad del aire. Pereira y su área metropolitana concentran buena parte de las emisiones. Al ser una región de topografía montañosa, la dispersión de estas partículas contaminantes puede verse disminuida, favoreciendo la acumulación de partículas en condiciones meteorológicas específicas. La acumulación de estas partículas incrementa la probabilidad de enfermedades en sectores vulnerables de la población como niños, adultos mayores y personas con enfermedades respiratorias y cardiovasculares preexistentes.

Frente a esta problemática, la disponibilidad de datos locales de PM10 y PM2.5 en Risaralda constituye una oportunidad para desarrollar una solución que permitan predecir la calidad del aire. Este proyecto basa en el análisis de los datos históricos recolectados entre los años 2007 y 2023 por la Corporación Autónoma Regional de Risaralda (CARDER), datos que brindan una base para identificar

patrones de comportamiento de la contaminación atmosférica del departamento. Este tipo de herramientas puede anticipar episodios críticos de contaminación, orientar estrategias de políticas públicas y, lo más importante, proteger la salud de los ciudadanos.

2. Pregunta de negocio y alcance del proyecto

El proyecto busca responder la siguiente pregunta de negocio: ¿Cómo predecir episodios críticos de contaminación del aire en Risaralda y, a partir de ello, apoyar decisiones en salud pública y gestión ambiental?

Para dar respuesta a este interrogante, el proyecto desarrollara una aplicación para la proyección de la calidad del aire en el departamento de Risaralda, utilizando como principal indicador las concentraciones de material particulado PM10 y PM2.5. Se emplearán los datos históricos de calidad del aire registrados entre 2007 y 2023 por la Corporación Autónoma Regional de Risaralda (CARDER), lo que permitirá identificar patrones y generar estimaciones de concentraciones futuras. El alcance del proyecto contempla las siguientes acciones:

- Procesamiento de datos históricos: limpieza, organización y análisis exploratorio de los datos con el fin de garantizar la calidad de la información utilizada.
- Desarrollo y entrenamiento del modelo de aprendizaje automático: implementación de un algoritmo que permita estimar concentraciones futuras de PM10 y PM2.5 a partir de los patrones históricos identificados.
- Construcción de la aplicación: diseño e implementación de una interfaz que facilite la interacción del usuario, permitiendo seleccionar un rango de fechas futuras y visualizar las proyecciones generadas.
- Visualización de resultados: inclusión de gráficos y reportes que presenten tendencias y posibles escenarios de calidad del aire en el departamento.

3. Conjunto de datos a emplear y descripción

El proyecto utilizará principalmente datos ambientales existentes en formato tabular y de series temporales. Se trabajará con registros de calidad del aire (PM10 y PM2.5) recolectados en estaciones municipales de Risaralda entre 2007 y 2023, disponibles en la página oficial del gobierno colombiano de datos abiertos:

https://www.datos.gov.co/Ambiente-y-Desarrollo-Sostenible/Calidad-del-Aire/53gx-j5pc/about_data

Cada registro incluye una marca de tiempo (Fecha), lo cual es de gran importancia para la construcción de un modelo de predicción de la calidad del aire. Las variables categóricas (Estación, Diámetro aerodinámico y Municipio) requerirán un proceso de codificación antes de su uso en el modelado. Adicionalmente, la columna Fecha ofrece la posibilidad de generar variables derivadas (día, mes, año y día de la semana), lo que permitirá identificar patrones temporales de manera más detallada.

4. Exploración de los datos

El conjunto de datos contiene **5047 registros** con las siguientes columnas:

- Municipio
- Estación
- Fecha
- Diámetro aerodinámico
- Medición ($\mu\text{g}/\text{m}^3$)

Durante la revisión inicial no se identificaron valores faltantes/nulos, pero si 9 registros con valor 0 en la columna de “medicion” lo cual posiblemente se deba a que la medicion esta por debajo del limite de deteccion, pero al ser pocos datos no compromete el análisis.

El resto de variables no presenta valores nulos.

Análisis Descriptivo:

Municipio	count	mean	std	min	25%	50%	75%	max
Dosquebradas	1976.0	30.182	17.178	1.0	15.36	28.79	40.82	116.0
La Virginia	743.0	22.803	10.554	0.0	15.61	21.4	27.87	73.31
Pereira	1507.0	28.598	15.472	0.0	18.94	25.58	34.93	149.59
Santa Rosa de Cabal	817.0	30.568	14.045	3.65	18.95	30.24	40.1	75.16

Tabla 1: Resumen por municipio

- Dosquebradas presenta la mayor variabilidad (desviación estándar = 17.178).
- Se detectan valores atípicos con concentraciones superiores a $100 \mu\text{g}/\text{m}^3$.

Limpieza de datos:

Se realizaron los siguientes ajustes:

- Conversión de `Fecha` a formato datetime.
- Conversión de `Medicion` a valor numérico.
- Creación de variables derivadas de `Fecha`:
`Dia`, `Mes`, `Año` y `DiaSemana` (ordenados de lunes a domingo).
- Se guardó un dataset enriquecido en
`data/processed/Calidad_del_Aire_enriquecido.csv`.

Visualizaciones:

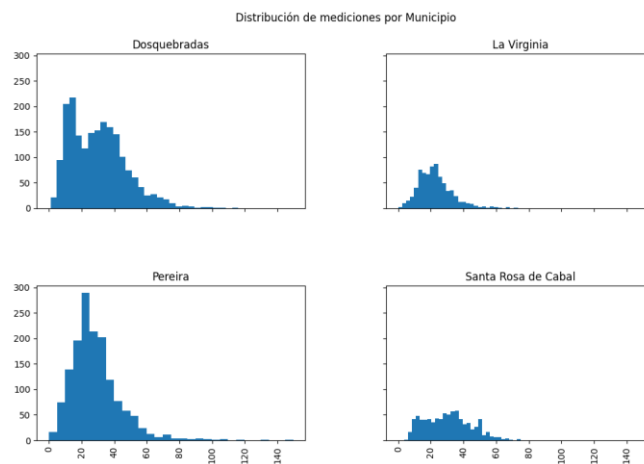


Imagen 1: Distribución mediciones por municipio

Esta gráfica muestra la distribución de valores de medición en cada municipio. Se observa que la mayoría de las mediciones se concentran entre 10 y 40 $\mu\text{g}/\text{m}^3$, aunque Pereira y Dosquebradas presentan colas más largas hacia la derecha, lo que indica la presencia de eventos con altos niveles de contaminación. Esto sugiere que estos municipios podrían tener mayores riesgos de superar los límites recomendados de calidad del aire.

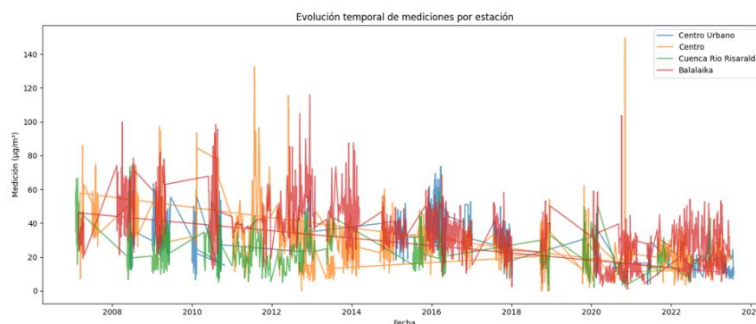


Imagen 2 Serie temporal por estación

La serie temporal permite identificar la evolución de los niveles de mediciones a lo largo de los años. Se evidencian picos de contaminación en Pereira y Dosquebradas, mientras que La Virginia y Santa Rosa presentan valores más estables. Esta gráfica ayuda a detectar episodios críticos y a analizar posibles patrones estacionales o asociados a condiciones locales específicas.

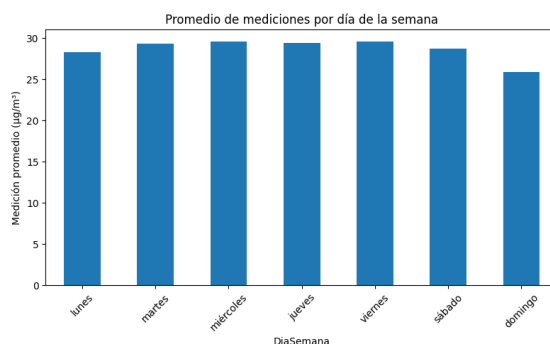


Imagen 3 Promedio de mediciones por día de la semana

Aquí se comparan los niveles promedio de mediciones entre los distintos días de la semana. La variación es ligera, pero se aprecia una tendencia a menores valores en fines de semana, lo que podría estar relacionado con la reducción de la actividad vehicular e industrial durante esos días. Esto refuerza la hipótesis de que parte importante de la contaminación proviene de fuentes móviles y productivas.

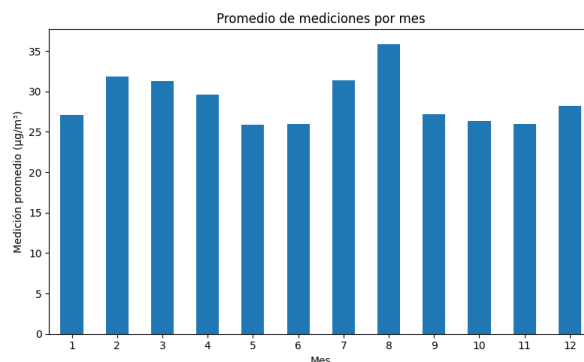


Imagen 4 Promedio de mediciones por mes

Este gráfico permite visualizar la estacionalidad de la contaminación. Se observa que algunos meses presentan picos más altos que otros, lo que podría estar asociado a condiciones climáticas específicas (ej. temporadas secas con mayor re-suspensión de polvo, o meses con menor dispersión atmosférica). Identificar estos patrones es clave para planificar medidas de control en épocas críticas.

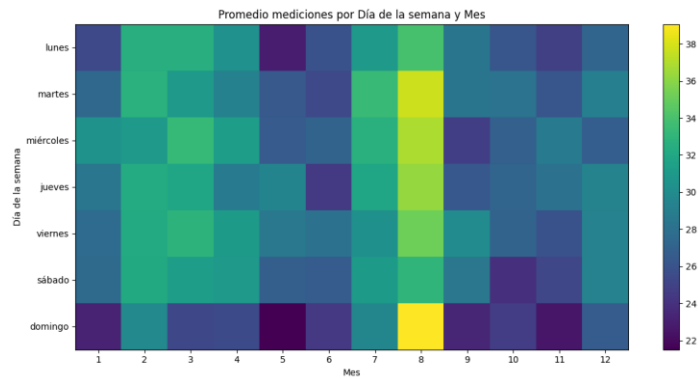


Imagen 5 Heatmap (Mes × Día de la semana)

El mapa de calor combina la dimensión temporal mensual y semanal. Se aprecian combinaciones de ciertos meses y días (ej. días laborales en meses secos) donde las concentraciones son mayores. Esta visualización facilita detectar patrones conjuntos y ayuda a enfocar políticas de mitigación en períodos de mayor riesgo.

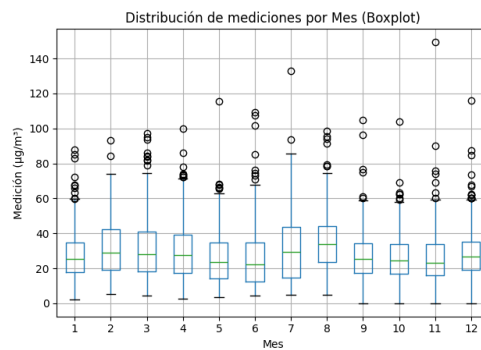


Imagen 6 Boxplot por mes

El grafico de cajas permite analizar la dispersión y los valores atípicos de mediciones a lo largo de los meses. Se evidencia que algunos meses presentan mayor variabilidad y valores extremos, lo que indica la ocurrencia de episodios de contaminación aguda. Estos resultados sugieren la necesidad de estrategias de monitoreo reforzado en dichos periodos.

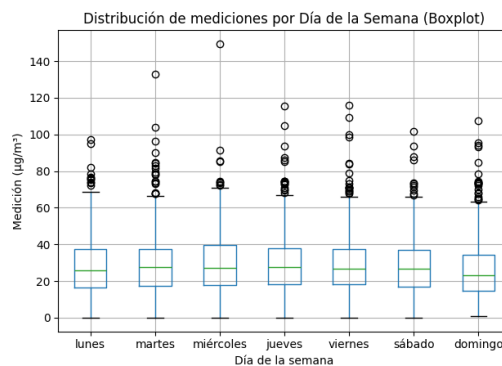


Imagen 7 Boxplot por día de la semana

Este gráfico muestra la variabilidad de los niveles de las mediciones según el día de la semana. Aunque la mediana no cambia de forma drástica, se observan diferencias en la dispersión: algunos días presentan mayor rango y valores atípicos más altos. Esto podría asociarse a actividades específicas de ciertos días, como mayor movilidad en inicios de semana o variaciones en las rutinas industriales.

5. Repositorios

5.1. GitHub

Con el objetivo de garantizar un adecuado manejo de versiones, se creó un repositorio en **GitHub** para el control del código y los entregables, y un repositorio en **DVC** con almacenamiento en **Amazon S3** para la gestión de versiones de los datos.

El repositorio principal del proyecto se encuentra en el siguiente enlace:

<https://github.com/SebastianOrd/Microproyecto-DS>

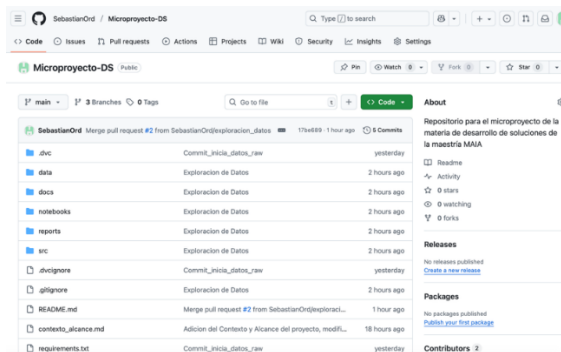


Imagen 8: Repositorio del proyecto

se ha definido una estructura para el repositorio de la siguiente manera:

```
data/
├── raw/           # Datos originales (solo lectura)
├── processed/     # Datos enriquecidos/intermedios
├── notebooks/     # Todos los ipynb que se desarrollen
├── src/           # Código modularizado en python
├── reports/       # Informes y entregables
├── figures/       # Imágenes usadas en los documentos
├── .gitignore
├── requirements.txt # Dependencias requeridas
└── README.md      # resumen del proyecto y su uso
```

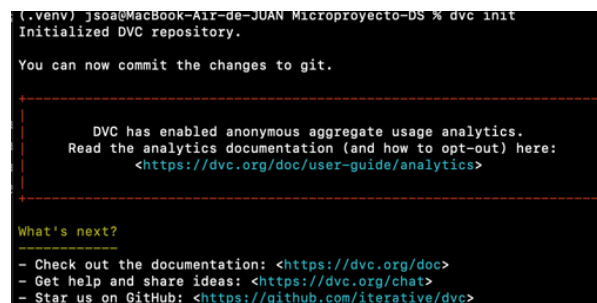
Imagen 9: Estructura del repositorio inicial

La dinámica de trabajo definida consiste en que cada integrante del grupo cree una rama para el desarrollo de las tareas asignadas. Todos los miembros cuentan con permisos para crear y realizar commits en sus propias ramas, pero no en la rama principal (main). El responsable del repositorio será quien revise los cambios propuestos y realice los merges correspondientes hacia main.

Finalmente, se estableció que cada integrante del grupo debe crear un entorno virtual en su máquina local e instalar las dependencias listadas en requirements.txt. De esta manera, se asegura la reproducibilidad del entorno de desarrollo y se facilita el trabajo colaborativo.

5.2. Repositorio de datos DVC

Con el objetivo de asegurar un adecuado manejo de los datos, se creó un **bucket de S3 en AWS** como almacenamiento remoto para el proyecto.

A terminal window showing the output of the 'dvc init' command. The text indicates that a DVC repository has been initialized and linked to a local Git repository. It also mentions that anonymous aggregate usage analytics are enabled and provides a link to the documentation for opting out. Finally, it lists 'What's next?' with links to documentation, chat, and GitHub.

```
(.venv) jsoa@MacBook-Air-de-JUAN Microproyecto-DS % dvc init
Initialized DVC repository.

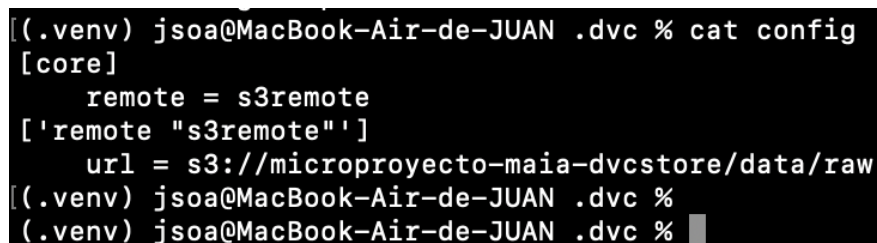
You can now commit the changes to git.

DVC has enabled anonymous aggregate usage analytics.
Read the analytics documentation (and how to opt-out) here:
<https://dvc.org/doc/user-guide/analytics>

What's next?
- Check out the documentation: <https://dvc.org/doc>
- Get help and share ideas: <https://dvc.org/chat>
- Star us on GitHub: <https://github.com/iterative/dvc>
```

Imagen 10 Repositorio DVC

Cada integrante del equipo debe clonar el repositorio de GitHub, el cual contiene la carpeta .dvc y el archivo de configuración .config.

A terminal window showing the output of the 'cat config' command in the .dvc directory. The output shows the configuration for a remote S3 storage, including the remote name 's3remote' and the URL 's3://microproyecto-maia-dvcstore/data/raw'.

```
(.venv) jsoa@MacBook-Air-de-JUAN .dvc % cat config
[core]
  remote = s3remote
['remote "s3remote"']
  url = s3://microproyecto-maia-dvcstore/data/raw
(.venv) jsoa@MacBook-Air-de-JUAN .dvc %
(.venv) jsoa@MacBook-Air-de-JUAN .dvc %
```

Imagen 11 Configuración repositorio remoto

Posteriormente, será necesario configurar el **AWS CLI** en su máquina local para establecer las credenciales de acceso al bucket. Una vez configurado, cada miembro podrá ejecutar un dvc pull para descargar los datos y trabajar con ellos localmente, garantizando así la sincronización y consistencia de la información entre todos los colaboradores.

5.3. Maqueta del prototipo

El prototipo consta de dos secciones principales:

- Carga inicial de la información, filtrado, obtención de métricas principales y visualización de los datos.
- Panel que resalte los días en que se prevean episodios críticos de contaminación, permitiendo a los usuarios tomar medidas preventivas.



Imagen 12 Maqueta sección de identificación de los datos



Imagen 13 Maqueta sección de alertas

5.4. Reporte de trabajo en equipo

Encargado	Rol Principal	Actividades Específicas
Hainer Torrenegra	Contexto y Alcance	<ul style="list-style-type: none">Definición del problema y su contextoElaboración de la pregunta de negocio y el alcance del proyectoRedacción y alineación del marco contextual del análisis de calidad del aire.
Juan Sebastian Ordoñez	Datos y Repositorios	<ul style="list-style-type: none">Descripción de los conjuntos de datos a emplearConfiguración del repositorio en GitHub para el control de versiones Implementación del repositorio de datos con DVC y Amazon S3
Brayan Gomez	Exploración y Análisis de Datos (EDA)	<ul style="list-style-type: none">Realización del análisis exploratorio y descriptivo de los datosGeneración de tablas y gráficos para documentar hallazgos inicialesLimpieza y preparación inicial de los datos
Maria Alejandra Rojas	Prototipo y Coordinación	<ul style="list-style-type: none">Diseño de la maqueta del prototipo de la aplicación.Elaboración del reporte de trabajo en equipo.Verificación final de todos los entregables en los repositorios.

5.5. Evidencia uso git

```
* commit bea7bee671af86a28565a32a93bee497898b21ab (HEAD -> main, origin/main, origin/HEAD)
Merge: 564dabd e81a01f
Author: SebastianOrd <js.ordonezal@uniandes.edu.co>
Date: Sun Aug 24 19:45:34 2025 -0500

    Merge pull request #5 from SebastianOrd/SebastianOrd-patch-1

    Update README.md

* commit e81a01f30d7fed3d0f2938955c0c25cb6f2c85e5
Author: SebastianOrd <js.ordonezal@uniandes.edu.co>
Date: Sun Aug 24 19:44:07 2025 -0500

    Update README.md

* commit 564dabdc91f1def920b1faaec64bd10de1ed22c8
Merge: 531d45b b4325b5
Author: SebastianOrd <js.ordonezal@uniandes.edu.co>
Date: Sun Aug 24 19:44:38 2025 -0500

    Merge pull request #4 from SebastianOrd/SebastianOrd-patch-1-1

    Update README.md

* commit b4325b56a2f7587a2854d1c49d671e78d2d46741
Author: SebastianOrd <js.ordonezal@uniandes.edu.co>
Date: Sun Aug 24 19:44:27 2025 -0500

    Update README.md

* commit 531d45b8093bed5708fa4206d165ad71cc51aef9
Merge: 17be689 d5e73c0
Author: SebastianOrd <js.ordonezal@uniandes.edu.co>
Date: Sun Aug 24 15:54:24 2025 -0500

    Merge pull request #3 from SebastianOrd/datos_repositorio

    Manejo de datos y documento de entrega_1

* commit d5e73c0d2bc5e1ff9f994ecff8a158e0dc5848ab (origin/datos_repositorio)
Author: sebastian <js.ordonezal@uniandes.edu.co>
Date: Sun Aug 24 15:53:19 2025 -0500

    Manejo de datos y documento de entrega_1

* commit 17be6895b0dcc0f97f2818c2b8e5ebfdd6c1d549
Merge: 6761314 74bfbad
Author: SebastianOrd <js.ordonezal@uniandes.edu.co>
Date: Sun Aug 24 12:45:16 2025 -0500

    Merge pull request #2 from SebastianOrd/exploracion_datos

    Exploracion de Datos

* commit 74bfbad602847d0f4bba282821c9145e550214c3 (origin/exploracion_datos)
Author: sebastian <js.ordonezal@uniandes.edu.co>
Date: Sun Aug 24 11:39:40 2025 -0500

    Exploracion de Datos

* commit 67613147fc10057eec9d55a0494b7ae70c6c8482
Merge: 5a7a383 74dff9d
Author: SebastianOrd <js.ordonezal@uniandes.edu.co>
Date: Sun Aug 24 11:41:08 2025 -0500

    Merge pull request #1 from SebastianOrd/contexto_alcance

    Adicion del Contexto y Alcance del proyecto, modificacion de resumen ...

* commit 74dff9dba72029b63fe180d2a17a02874f84032b (origin/contexto_alcance)
Author: hainer torrenegra <h.torrenegra@uniandes.edu.co>
Date: Sat Aug 23 20:00:52 2025 -0500
```