Proyecto Final

Análisis de calidad del aire y su relación con enfermedades respiratorias

Problema y estrategia

Definición del problema

El Valle de Aburrá presenta de forma recurrente concentraciones de contaminantes atmosféricos, especialmente material particulado fino ($PM_{2.5}$ y PM_{10}), que superan los límites recomendados por la Organización Mundial de la Salud. Esta situación provoca un deterioro progresivo en la calidad del aire y genera impactos negativos sobre la salud respiratoria de la población, evidenciados en el aumento de consultas médicas y hospitalizaciones durante los periodos de alta contaminación.

A pesar de contar con sistemas de monitoreo ambiental robustos como el SIATA, la información disponible se emplea principalmente con fines descriptivos y no se explota analíticamente para anticipar riesgos sanitarios. La ausencia de integración entre los datos ambientales, meteorológicos y epidemiológicos impide establecer relaciones causales claras y limita la toma de decisiones basadas en evidencia científica. Por tanto, se requiere una solución que unifique fuentes de datos abiertas y permita modelar, analizar y predecir los impactos de la contaminación sobre la salud respiratoria en el territorio.

Necesidad del negocio

Las instituciones encargadas de la salud y el medio ambiente en Medellín, como el Área Metropolitana del Valle de Aburrá y la Secretaría de Salud, enfrentan el desafío de prevenir episodios críticos de contaminación y sus efectos sobre la población. Actualmente, las estrategias se implementan de manera reactiva, una vez que los niveles de contaminación ya han superado los umbrales de alerta.

Existe, por tanto, la necesidad de desarrollar un sistema analítico de soporte a la decisión que permita anticipar incrementos en los casos de enfermedades respiratorias a partir de los registros históricos de calidad del aire, variables meteorológicas y datos de morbilidad. Esta herramienta facilitaría la planificación preventiva, la asignación eficiente de recursos médicos y la formulación de políticas públicas basadas en evidencia.

Alineación con la estrategia corporativa

El proyecto se alinea con los ejes estratégicos de sostenibilidad, salud pública y transformación digital del Plan de Desarrollo de Medellín y del Área Metropolitana del Valle de

Aburrá, los cuales promueven el uso de datos abiertos y analítica avanzada para la gestión ambiental y sanitaria.

Asimismo, responde a los Objetivos de Desarrollo Sostenible (ODS), especialmente el ODS 3 (Salud y bienestar) y el ODS 11 (Ciudades sostenibles), impulsando el aprovechamiento de tecnologías de datos para reducir la exposición de la población a la contaminación atmosférica.

En el ámbito institucional, la implementación de este sistema fortalecería la inteligencia analítica corporativa, mejoraría la capacidad de respuesta ante emergencias ambientales y consolidaría una cultura de gestión basada en datos dentro de las entidades públicas del territorio

Propuesta inicial del modelado de procesos (BPM)

Descripción general

El proceso propuesto busca integrar y analizar datos ambientales y de salud pública para identificar patrones que relacionen la contaminación atmosférica con la incidencia de enfermedades respiratorias.
Actualmente, las instituciones responsables del monitoreo del aire (como SIATA) y las entidades de salud (como la Secretaría de Salud de Medellín) gestionan la información de forma independiente.
El modelo BPM plantea un flujo que unifica estas fuentes en una secuencia estructurada de **ingesta, procesamiento, análisis y difusión de resultados**, favoreciendo la toma de decisiones preventivas.

Objetivo del proceso

Automatizar la recopilación, limpieza y análisis de datos de calidad del aire y salud respiratoria para generar alertas y reportes predictivos que apoyen la planificación sanitaria y ambiental.

Etapas del proceso

1. **Inicio del proceso**
    - Evento: actualización de datos diarios de calidad del aire o registros de salud.
    - Actor: *Sistema SIATA* / *Secretaría de Salud*.
    - Resultado: activación del flujo de análisis.

2. **Ingesta y consolidación de datos abiertos**
    - Tareas:
        - Conexión a fuentes abiertas (SIATA, IDEAM, MinSalud).

        - Descarga automática de archivos CSV o consulta API.
        - Validación de estructura y campos.
    - Actor: *Data Engineer*.
    - Artefacto: dataset crudo (raw_data).

3. **Limpieza y normalización**
    - Tareas:
        - Eliminación de valores nulos, duplicados o inconsistentes.
        - Homologación de unidades (µg/m^3, días, municipios).
        - Generación de dataset limpio (clean_data).
    - Actor: *Data Analyst*.
    - Herramientas: Python / Pandas / Jupyter.

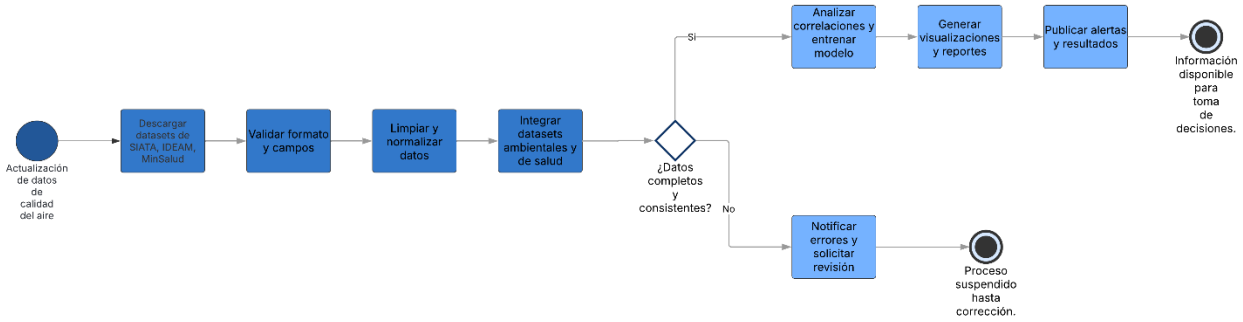
4. **Integración y almacenamiento**
    - Tareas:
        - Unión de datos ambientales, meteorológicos y de salud.
        - Almacenamiento en base analítica o repositorio central (p. ej. S3 o base SQL).
    - Artefacto: *Data Warehouse ambiental-sanitario*.

5. **Análisis exploratorio y modelado predictivo**
    - Tareas:
        - Cálculo de correlaciones entre $PM_{2.5}$ y casos de enfermedades respiratorias.
        - Entrenamiento de modelos de predicción (regresión / red neuronal).
        - Validación de resultados.
    - Actor: *Científico de Datos*.
    - Artefacto: *Modelo predictivo entrenado*.

6. **Visualización y generación de reportes**
    - Tareas:
        - Creación de dashboards (mapas de calor, series temporales).
        - Publicación de reportes semanales de riesgo respiratorio.
    - Actor: *Analista de Políticas / Gestor Ambiental*.
    - Herramienta: Power BI / Plotly Dash.

7. **Toma de decisiones y retroalimentación**
    - Tareas:
        - Comunicación de alertas preventivas a las autoridades de salud.
        - Evaluación de efectividad del modelo.
        - Ajuste de umbrales o variables según desempeño.
    - Actor: *Comité de Gestión Ambiental y Sanitaria*.
    - Resultado: ciclo de mejora continua.

8. **Fin del proceso**



Roles principales

Rol	Responsabilidad
Data Engineer	Automatizar la ingesta y validación de datos.
Data Analyst	Limpiar, transformar y documentar la información.
Científico de Datos	Modelar correlaciones y entrenar modelos predictivos.
Gestor Ambiental / Sanitario	Interpretar resultados y emitir alertas.
Comité de Gobernanza de Datos	Supervisar la calidad y seguridad del proceso.

Búsqueda y selección de datos

1. Aire (Mediciones estaciones calidad del aire) – SIATA/AMVA (por contaminante, JSON)
    a. Qué trae: PM2.5, PM10, PM1, O_3, NO_2, NO, SO_2, CO por estación y tiempo.
    b. Para qué sirve: es la base principal para medir la calidad del aire.
    c. Nivel: estación/fecha (luego se puede resumir por municipio/día).
    d. Enlace: https://datosabiertos.metropol.gov.co/dataset/a78f43fe-6711-4f41-bbbc-521fb8646bde
2. Meteorología (Variables meteorológicas) – SIATA (Vaisala, JSON)
    a. Qué trae: temperatura, humedad, precipitación y viento por estación y tiempo.
    b. Para qué sirve: explica cambios en la contaminación (clima).
    c. Nivel: estación/fecha.
    d. Enlace: https://datosabiertos.metropol.gov.co/dataset/ca7225ca-6754-4687-9cb5-0ffb76872ca8
3. Salud (Cálculo de casos de morbimortalidad asociados a contaminación del aire) – SIVISA/SISPRO (AMVA)
    a. Qué trae: estimaciones de enfermedades y muertes asociadas a contaminación del aire.
    b. Para qué sirve: me da la señal de impacto en salud (agregado, sin PII).

    c. Nivel: municipal / periodo (semanal o anual según recurso).
    d. Enlace: https://datosabiertos.metropol.gov.co/dataset/7281fa8e-c666-4fed-a8bd-a4af7d0ea749
4. Demografía – DANE (población municipal)
    a. Qué trae: población por municipio y año.
    b. Para qué sirve: convertir conteos en tasas por 100.000 hab.
    c. Enlace: https://www.dane.gov.co/.../proyecciones-de-poblacion
5. DIVIPOLA – Códigos de municipios
    a. Qué trae: código oficial de cada municipio.
    b. Para qué sirve: llave única para unir todo sin enredos de nombres.
    c. Enlace: https://www.datos.gov.co/.../gdxc-w37w

Opcionales

- Ciudadanos Científicos (SIATA): útil para mapas y comparación, no como serie principal. https://datosabiertos.metropol.gov.co/dataset/bd2e0f13-6834-4a58-af84-a648adc4b190
- IDEAM (DHIME/geoportal): respaldo si necesito estaciones fuera del Valle.
- Plan Siembra (AMVA): árboles sembrados por municipio/año como variable de contexto. https://datosabiertos.metropol.gov.co/dataset/5d7841e4-5d95-4723-83f3-ebb93c27bb72

Reglas simples de unión

- Tiempo: fecha (aire/meteo, se puede resumir a día) y semana_epidemiológica (salud).
- Espacio: municipio_code (DIVIPOLA).
- Estación: estacion_id (para aire/meteo, luego asigno municipio).

Boceto de arquitectura de datos

1. Traer datos:
    - Descarga de Aire SIATA y Meteo SIATA (JSON) + Salud (archivo abierto) + DANE + DIVIPOLA.
    - Se guarda tal cual en /data/bronze/.
2. Ordenar y limpiar:
    - Establecer nombres claros (ej.: pm25_ugm3, temp_c).
    - Eliminar duplicados y marco vacíos. (Reglas simples: PM ≥ 0; humedad 0–100; viento ≥ 0).
3. Unir:

- Aire y Meteo por estación + fecha → agrego a día.
- Se asigna cada estación a su municipio y agrego por municipio/día.
- Salud se maneja por semana epidemiológica y municipio.
- Con DIVIPOLA estandarizo el municipio; con DANE calculo tasas.

4. Armar tablas para análisis:
   - Tabla diaria por municipio: PM2.5, PM10, clima (día).
   - Tabla semanal por municipio: tasas de salud (semanal).
   - Si se necesita, se crean promedios móviles (7 y 14 días) para comparar con la semana de salud.

Usar y mostrar
Un notebook de exploración (tendencias, mapas simples).
Un gráfico PM2.5 vs. tasas por municipio/semana para ver la relación.