Contaminación de Madrid

Objetivo

Con esta prueba, queremos ver:

- Tu estilo y habilidades de programación
- Cómo analizas y resuelves los problemas
- Cómo presentas resultados o justificas tus decisiones

Alcance

No pretendemos que le dediques un tiempo excesivo. Vamos a plantear un problema abierto, en el que tú decides hasta donde llegas. La idea es que hagas lo básico y comentes posibles puntos de mejora o próximos pasos, pero sin llegar a implementarlos.

Entrega

El formato de salida debería ser algo que permita incluir texto, código y salida. Te proponemos un notebook de Jupyter.

La programación debe ser sobre Python 3. Si no nos indicas otra cosa, usaremos Python 3.8 para correrlo.

Sobre la forma de entrega, puedes:

- O enviarnos un zip con la prueba
- O subirlo a un repositorio al que podamos acceder (p.e. público de Github)

Problema

Queremos almacenar en una base de datos propia los datos de niveles de NO_2 del aire de Madrid durante el año 2018. El objetivo es que otros compañeros del departamento puedan:

- Hacer una visualización de la evolución de la contaminación: gráficas de evolución diaria / mensual, mapas de calor animados, ...
- Hacer análisis de situaciones extraordinarias: días con restricciones, medición de la efectividad de ciertas medidas, ...

Los datos están disponibles en el portal de datos abiertos de Madrid, aquí. Asegúrate de consultar el PDF con la documentación, te proporcionará información de utilidad sobre el dataset. Algunos comentarios:

- Descarga los datos del año 2018
- Solo nos interesan las mediciones sobre el NO_2. Puedes descartar las de otros elementos contaminantes

Objetivo 1: lectura y tratamiento

Lee y trata los datos. Debes dejarlos limpios en el formato que consideres más cómodo para su análisis. Comenta cómo almacenarías esto: fichero, BD, ...

Objetivo 2: arquitectura

Diseña una arquitectura para almacenar, limpiar, explotar y presentar estos datos. El uso de esta plataforma será:

- Dispondrá de información de niveles de NO_2, las estaciones donde se han recogido y datos de temperaturas de Madrid
- Hay dos tipos de cargas de información:
    - Masivas, puntuales: sirven para hacer una carga inicial de datos. También, de forma esporádica, para corregir información capturada en tiempo real.
    - Incrementales. Los orígenes son:
        - Datos en tiempo real (actualizados cada hora) de contaminación, más información aquí
        - Datos de temperaturas. Se capturarán consultando alguna API de servicios metereológicos.
- La explotación será:
    - Vía API: los usuarios de la plataforma pueden consultar y descargar los datos por API.
    - Vía portal web: el equipo de Front diseñará un portal para permitir que los usuarios visualicen mapas de calor, gráficas de evolución, etc.
    - Modelo predictivo: el equipo de Data creará un modelo para predecir la contaminación de los próximos 3 días.

    - Análisis *adhoc*: además, el equipo de Data necesita explorar y crear pequeños análisis *adhoc* que puedan surgir. P.e. prevemos que el Ayuntamiento nos asigne la tarea de medir el efecto de las medidas tomadas en la legislatura pasada: Madrid Central, restricciones los días de aviso de contaminación, etc.
- El volumen de los datos: profundidad histórica desde el año 2001.
- La cantidad de usuarios: 500-1000 usuarios diarios.
- Las características de nuestro equipo: somos una *startup* pequeña y tecnológica, con unos pocos perfiles por disciplina: *front*, *back* y *data*. No tenemos perfiles especializados de sistemas.

No hace falta que programes nada en esta parte. Simplemente describe en la forma que creas más conveniente cómo lo harías (p.e. con un diagrama de arquitectura + comentarios).

Anexo

Puedes ver la información adicional aquí:

- Estaciones de medida
- Temperaturas horarias