Laboratorio 3 - Explorando la calidad del aire con Streamlit

Curso: Programación Científica
cristhian.rabi@ce.ucn.cl

13 de noviembre de 2025

1. Dataset

El dataset contiene 9.358 registros horarios de un dispositivo multisensor químico instalado en una calle muy contaminada de una ciudad italiana. Se encuentra acá:
https://archive.ics.uci.edu/dataset/360/air+quality

1.1. Sensores MOX (óxidos metálicos)

Son 5 sensores que responden a distintos gases:

- PT08.S1 (CO): sensible a CO
- PT08.S2 (NMHC): sensible a hidrocarburos no metánicos
- PT08.S3 (NOx): sensible a NOx
- PT08.S4 (NO2): sensible a NO_2
- PT08.S5 (O3): sensible a ozono

  Estos sensores son crudos, no dan la concentración real, sino una señal eléctrica dependiente de gas, temperatura y humedad.

1.2. Concentraciones reales (Ground Truth)

Las concentraciones reales fueron medidas por un analizador certificado e incluyen:

- CO (mg/m^3)
- NMHC ($\mu g/m^3$)
- C_6H_6 (Benceno)
- NO_x (ppb)
- NO_2 ($\mu g/m^3$)

Estas variables permiten calibrar los sensores MOX comparando la señal registrada con la concentración verdadera.

1.3. Variables ambientales

- Temperatura
- Humedad relativa (RH)
- Humedad absoluta (AH)

Estas variables influyen de manera significativa en las lecturas de los sensores.

1.4. Características del dataset

- Presencia de *sensor drift* (cambios en el tiempo en la respuesta del sensor).
- *Cross-sensitivities*: los sensores responden a varios gases, no solo al contaminante objetivo.
- Valores faltantes indicados con `-200`.

2. Desarrollo

2.1. Preparación de datos

Si bien el dataset no tiene mayores complicaciones, se sugiere hacer lo siguiente:

1. Manejo de missings: Los valores `-200` son missings
2. Agregar variables asociadas a la fecha de medición:
    - Día de la semana
    - Hora
    - Mes
    - Fin de semana/día laboral

Estas operaciones se realizan en un notebook o script de Python. Luego genere un nuevo csv con los datos procesados.

2.2. Análisis exploratorio

En este apartado veremos cómo se comportan los datos. Es relevante conocer cómo se distribuyen las variables y saber cómo se correlacionan. Esta parte también hágala en un notebook/script. Lo que valga la pena destacar o usted crea que es interesante, seleccionélo y llévelo en reporte de Streamlit.

2.3. Modelamiento I

En nuestros datos tenemos variables obtenidas de 2 maneras: un sensor que nos entrega una señal eléctrica y otra que es de una manera certificada y precisa de lo que estamos midiendo. Por ejemplo, `PT08.S1(CO)` es un sensor de óxidos metálicos, particularmente de óxido de estaño. Este sensor es utilizado para medir el CO de manera *indirecta*, mientras que la medición `CO(GT)` es el valor de la concentración de CO obtenido por un analizador certificado, esto vendría a ser el valor *real*. Por lo tanto, lo interesante ahora es calibrar el sensor que nos entrega una señal eléctrica y ver cómo se relaciona con el valor real.

Para realizar modelamientos no utilizaremos algoritmos de Machine Learning. En primer lugar porque lo que necesitamos resolver no lo amerita, y, en segundo lugar, porque podemos resolverlo de una manera más sencilla ajustando una curva con la librería `scipy`.

Se le pide crear 3 modelos univariable, seleccione 3 sensores MOX y relaciónelos con su contaminante real. Además, cree un modelo multivariable para algún contaminante (use varios sensres).

2.4. Modelamiento II

Se dice que los sensores cambian la respuesta a lo largo del tiempo, esto puede ser por desgaste de materiales por ejemplo. Identifique cómo cambia la respuesta en el tiempo y modele el comportamiento. Para ello puede tomar solo 1 sensor y analizarlo, no es necesario verlos todos.

2.5. Reporte a eleccción

Se le presentarán diferentes propuestas para desarrollar, seleccione 1. Debe presentarlo de una manera *creativa* y con algún mensaje/intención clara.

2.5.1. Clasificación

Etiquetaremos los datos en base a algún criterio, luego mostraremos el comportamiento por categoría.

- Calidad del aire: Según un criterio justificado, definir categorías para el aire (Buena o Mala por ejemplo)
- Días normales vs Días contaminados: Agrupe los datos por día y luego clasifíquelos.

2.5.2. Series de tiempo

- Forecasting de alguna variable (GT)
- Detección de anomalías

3. Entregables

La fecha de entrega es el 10 de diciembre al medio día (12:00).

3.1. Código fuente

El código se divide en 2 partes:

- Desarrollo: Todos los gráficos, experimentos y exploración realizada. Incluye todo lo que hizo, hasta lo que no funcionó como esperaba.
- Código para la App Streamlit: Acá está lo que usted seleccionó y quiere mostrar, es la presentación de su trabajo. La App de Streamlit debe incluir lo siguiente:
  - Título y contexto. Puede darle un enfoque especial.
  - Exploración: Gráficos relevantes (2 a 4), debe incluir 1 filtro como un slider u otro a gusto
  - Modelamiento I: 1 gráfico para cada modelo con la curva ajustada
  - Modelamiento II: Un gráfico que muestre el cambio de la respuesta en el tiempo
  - Reporte a elección: Explicar la decisión que tomó, los criterios, mostrar al menos 1 gráfico con sus filtros respectivos (categoría, tiempo, etc...)

3.2. Video

Un video breve mostrando la App de Streamlit, no es necesaria la edición. Simplemente grabar pantalla, mostrar cómo funciona y explicar lo preciso. No explique el contexto.