# ACTIVIDAD DE APRENDIZAJE #1

JOSE DAVID CARDONA MAZO

Proyecto Integrado III - Analítica de Datos (**Sharon Karin Camacho**) - PREICA2502B020071

## UNIVERSIDAD DIGITAL DE ANTIOQUIA – 2025

### Definición del problema

Se necesita evaluar y analizar de manera cuantitativa cual es la influencia de las variables meteorológicas tales como: temperatura, precipitación, velocidad del viento, radiación solar y humedad relativa en la concentración y variabilidad de los contaminantes atmosféricos (PM10, PM2.5, $NO_2$, $SO_2$, CO y $O_3$) que son registrados desde las estaciones de monitoreo de calidad del aire de Colombia.
El propósito esta investigación y análisis es identificar patrones de correlación y cuáles son esas tendencias temporales que nos permitan predecir todas aquellas posibles excedencias de los límites establecidos por la normativa ambiental, asi podríamos contribuir a la toma de decisiones basadas en evidencia en la gestión ambiental y por sobre todo en la protección de la salud pública.

### Hipótesis

Las condiciones meteorológicas influyen de manera significativa en los niveles de contaminación atmosférica.
En particular evaluaremos las siguientes relaciones:

- Incrementos en la **temperatura** y la **radiación solar** se asocian con mayores concentraciones de ozono ($O_3$).
- Bajos niveles de **precipitación** y **humedad relativa** favorecen la acumulación de partículas PM10 y PM2.5.
- Mayores velocidades de **viento** tienden a reducir la concentración de contaminantes debido a la dispersión atmosférica.

Estas relaciones permitirán construir **modelos predictivos medibles** basados en coeficientes de correlación ($R^2$), precisión del modelo (RMSE) y número de días que cuenten con excedencias de los límites de calidad del aire.

### Granularidad

La granularidad del análisis que se realizará **será horaria y por estación de monitoreo**, ya que eso nos va a permitir observar todas esas variaciones temporales y espaciales en las concentraciones de contaminantes.
Este nivel de detalle es relevante desde mi punto de vista porque facilita la identificación de patrones**,** correlaciones y eventos críticos de contaminación por lo que después en un futuro podremos mejorar la precisión de los modelos predictivos y se harán la toma de decisiones ambientales.

### Diccionario de datos

| Columna | Descripción | Tipo de Dato | Ejemplo |
|---|---|---|---|
| **id_estacion** | Identificador único asignado a la estación de monitoreo. | Numérico | 20383 |
| **autoridad_ambiental** | Autoridad ambiental encargada de la estación. | Texto | CORANTIOQUIA |
| **estacion** | Nombre de la estación de monitoreo. | Texto | ALTAVISTA |
| **latitud** | Coordenada geográfica que indica la latitud de la estación. | Numérico | 6.222.584 |
| **longitud** | Coordenada geográfica que indica la longitud de la estación. | Numérico | -75.628.207 |
| **variable** | Contaminante o variable meteorológica medida en la estación (PM10, PM2.5, temperatura, etc.). | Texto | PM10 |
| **unidades** | Unidad de medida para la variable. | Texto | ug/m3 |
| **tiempo_de_exposicion_horas** | Horas de exposición durante el monitoreo. | Numérico | 1 |
| **a_o** | Año en el que se realizaron las mediciones. | Numérico | 2011 |
| **promedio** | Promedio anual de la medición de la variable (contaminante o meteorológica). | Numérico | 75.5 |
| **suma** | Total acumulado anual de la medición de la variable. | Numérico | 329454.0 |
| **no_de_datos** | Número total de datos válidos recopilados en ese año de medición. | Numérico | 4363 |
| **representatividad_temporal** | Porcentaje del tiempo que tiene datos válidos registrados. | Categórico | 50% |
| **excedencias_limite_actual** | Número de veces que el contaminante ha superado el límite permitido en ese año. | Numérico | 0 |
| **porcentaje_excedencias_limite** | Porcentaje de veces que se superó el límite permitido durante el año. | Numérico | 0.0% |
| **mediana** | Mediana de la medición del contaminante durante el año. | Numérico | 64.0 |
| **percentil_98** | Percentil 98 de la medición del contaminante durante el año (valor por debajo del cual está el 98% de los datos). | Numérico | 208.0 |
| **maximo** | Valor máximo de la medición durante el año. | Numérico | 524.0 |
| **fechas_horas_del_maximo** | Fecha y hora del valor máximo registrado durante el año. | Fecha/Tiempo | 27/10/2011 8:00:00 p. m. |
| **minimo** | Valor mínimo de la medición durante el año. | Numérico | 7.0 |

| | | | |
|---|---|---|---|
| **fechas_horas_del_minimo** | Fecha y hora del valor mínimo registrado durante el año. | Fecha/Tiempo | 27/10/2011 8:00:00 p. m. |
| **dias_de_excedencias** | Número de días en que se superó el límite permitido. | Numérico | 0 |
| **codigo_del_departamento** | Código DANE del departamento donde se encuentra la estación. | Numérico | 5 |
| **nombre_del_departamento** | Nombre del departamento. | Texto | ANTIOQUIA |
| **codigo_del_municipio** | Código DANE del municipio donde se encuentra la estación. | Numérico | 5001 |
| **nombre_del_municipio** | Nombre del municipio. | Texto | MEDELLÍN |
| **tipo_de_estacion** | Tipo de estación de monitoreo (fija o móvil). | Texto | Fija |
| **ubicacion** | Información geoespacial de la estación (coordenadas en formato JSON). | Geoespacial | {'type': 'Point', 'coordinates': [-75.628207, 6.222584]} |

## Posibles problemas en cada aspecto analizado en la exploración

- **Datos faltantes (nulos):** Se evidenció la presencia de valores nulos en columnas y variables, esto puede afectar el análisis y la correlación de las variables:
- **Inconsistencia en formatos de fecha y hora:** algunos de los formatos no están adecuados para su agrupación correcta lo que nos puede generar conflictos al momento de comparar.
- **Duplicados:** Existe gran cantidad de duplicados
- **Inconsistencias geográficas:** Algunas estaciones presentan coordenadas nulas o ubicaciones fuera del rango de Colombia, lo que afectará de alguna manera la representación espacial y el mapeo geográfico de los datos.
- **Valores atípicos, raros o extremos:** Existen valores atípicos dentro de los datos lo que nos puede decir que puede deberse a un sensor malo o que surgieron errores en la captura de esos datos

## Conclusiones

- Tenemos una fuente de datos muy útil y que nos permitirá llevar a cabo el análisis correcto, se requiere de una gran limpieza para asegurar confiabilidad.
- Se identifican tendencias y valores claros, con respecto a la contaminación en zonas urbanas o industriales con respecto a la comparativa con zonas rurales.
- Los datos meteorológicos que tenemos en el dataset presentan una variabilidad que desde mi perspectiva es suficiente para construir modelos de correlación y predicción.

**Tareas específicas para la limpieza de datos**

- Se normalizarán formatos de fecha y hora.
- Se eliminarán los valores nulos y duplicados.
- Se filtrará con los valores atípicos para corregirlos usando percentiles.
- Se validan coordenadas de estaciones dentro del rango geográfico de Colombia.

**ENLACE A GITHUB**

https://github.com/josedav-17/analiticaDeDatos3.git

**ENLACE A TRELLO**

https://trello.com/invite/b/6911115e4d2b851c2700433f/ATTI558c5564fca7516329b39ce7a443fbed99E1B4C2/analiticadedatos3

**ENLACE DEL DATASET**

datos.gov.co/resource/kekd-7v7h.json?