**DATA MINING Y MACHINE LEARNING**

**(ICCD623)**

**PROYECTO 2 – 1er Bimestre**

**Integrantes:** Arias Johanna

Narváez Jeferson

**Curso:** GR2CC\_2023-1

**Fecha:** 30 de junio de 2023

**Tema:** Analítica Prescriptiva

# Introducción

La calidad del aire es un aspecto crucial para la salud y el bienestar de las personas, y su monitorización es de suma importancia. En este proyecto, se busca desarrollar un modelo de Machine Learning capaz de predecir el Índice de Calidad del Aire (AQI) o el Nivel de Contaminación del Aire en Beijing, China. Utilizando el conjunto de datos "Beijing Multi-Site Air-Quality Data Data Set", se pretende crear un modelo predictivo que pueda proporcionar información anticipada sobre los niveles de contaminación atmosférica en la ciudad.

# Definición del problema

El objetivo principal es predecir el AQI o el Nivel de Contaminación del Aire en Beijing. Para ello, se utilizarán datos históricos de contaminantes atmosféricos y variables meteorológicas de 12 sitios de monitoreo de calidad del aire en la ciudad. Estos datos permitirán entrenar un modelo de Machine Learning que pueda predecir el nivel de contaminación del aire en función de sus características.

Los datos de calidad del aire provienen del Centro de Monitoreo Ambiental Municipal de Beijing. Los datos meteorológicos de cada sitio de calidad del aire están relacionados con la estación meteorológica más cercana de la Administración Meteorológica de China. El período de tiempo va desde el 1 de marzo de 2013 hasta el 28 de febrero de 2017.

Hace unos años, China estableció el Índice de Calidad del Aire (AQI) basado en el nivel de cinco contaminantes atmosféricos medidos en las estaciones de monitoreo de cada ciudad, a saber:

dióxido de azufre (SO2),

dióxido de nitrógeno (NO2),

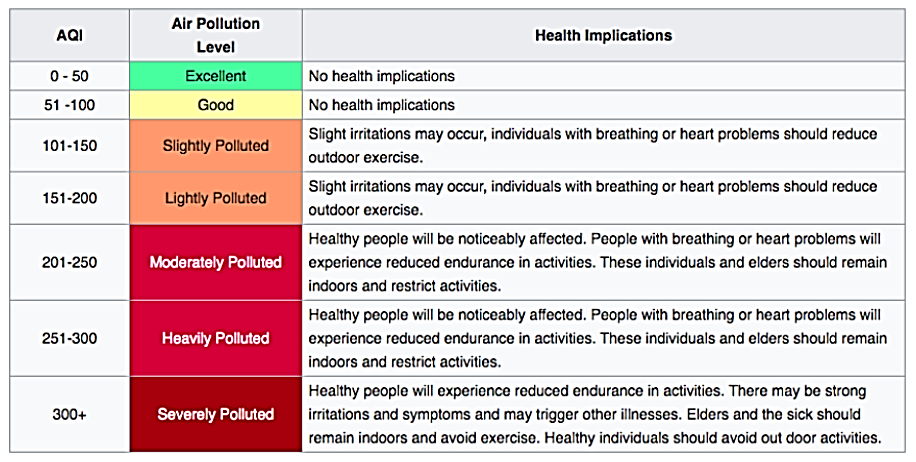
partículas suspendidas (PM10),

monóxido de carbono (CO) y

ozono (O3),

A cada nivel de contaminante se le asigna una puntuación individual, y el AQI final es la puntuación más alta de esos cinco contaminantes.

Los contaminantes pueden medirse de manera bastante diferente. SO2, NO2 y PM10 se miden como un promedio diario. CO y O3 son más dañinos y se miden como un promedio por hora. El valor final del AQI se calcula por día y tiene la interpretación que se muestra en la siguiente tabla.



# Pre-procesamiento de datos

Antes de utilizar los datos para el entrenamiento del modelo, es necesario realizar un preprocesamiento de los mismos. Algunas de las tareas de preprocesamiento incluyen:

Limpieza de datos: Se identifican y manejan los valores faltantes, valores atípicos y otros problemas de calidad de los datos que puedan afectar el rendimiento del modelo.

Normalización de datos: Se asegura que los datos estén en una escala común y comparable. Esto puede implicar la normalización de variables numéricas o la codificación adecuada de variables categóricas.

Selección de características: Se analiza la relevancia de cada característica en relación con la variable objetivo (AQI o Nivel de Contaminación del Aire) y se seleccionan las características más relevantes para el modelo.

División de datos: Se separan los datos en conjuntos de entrenamiento y prueba. El conjunto de entrenamiento se utiliza para entrenar el modelo, mientras que el conjunto de prueba, para evaluar su rendimiento y generalización.

# Análisis exploratorio de datos

Resumir y visualizar los datos de las formas útiles. Los estudiantes deben pensar en preguntas interesantes que se podrían comprobar con los datos disponibles y dar respuestas textuales o mediante visualización de datos.

# Modelado predictivo: configuración experimental y resultados obtenidos

Se debe definir una tarea predictiva que pueda ayudar a predecir la contaminación del aire, a través del valor del AQI o del Nivel de Contaminación del Aire, en función de sus características. Después de definir la tarea, se debe utilizar los datos disponibles para seleccionar y obtener un buen modelo para esta tarea. Se debe justificar el modelo sugerido.

Selección del modelo: De acuerdo a la naturaleza de los datos trabajados, se ha seleccionado el modelo KNN para la tarea de predicción.

Evaluación del modelo: Se utilizarán métricas de evaluación adecuadas, como el error cuadrático medio (MSE), el coeficiente de determinación (R²) o el error absoluto medio (MAE), para medir el rendimiento del modelo en la predicción del AQI o el Nivel de Contaminación del Aire.

Validación cruzada: Se aplicará validación cruzada para evaluar el rendimiento del modelo de manera robusta. Esto implica dividir los datos en múltiples conjuntos de entrenamiento y prueba, y calcular las métricas de evaluación promediadas sobre todas las divisiones.

Ajuste de hiperparámetros: Se utilizarán técnicas de búsqueda de hiperparámetros, como la búsqueda en cuadrícula o la optimización bayesiana, para encontrar la combinación óptima de hiperparámetros del modelo.

Comparación de modelos: Se compararán diferentes modelos en función de su rendimiento y se seleccionará aquel que obtenga los mejores resultados en la tarea de predicción del AQI o el Nivel de Contaminación del Aire.

El proceso experimental se realizará iterativamente, ajustando el preprocesamiento de datos, la selección del modelo y los hiperparámetros, con el fin de obtener un modelo final que proporcione predicciones precisas y confiables de la contaminación del aire en Beijing.

# Conclusiones, limitaciones y trabajos futuros

# Anexos (opcional).

Un cuaderno Jupyter/RMarkdown/Colab dinámico:

<https://www.kaggle.com/code/jefersonnarvaez/proyecto-2>