



## **ESCUELA POLITÉCNICA NACIONAL**

### **Proyecto de Analítica Prescriptiva**

Integrantes: Javier Salazar

Edgar Velasco

### **Introducción**

La calidad del aire es un tema de preocupación global, y recientemente se ha informado que en Quito, Ecuador, los niveles de contaminación atmosférica han superado ampliamente los límites tolerados por la Organización Mundial de la Salud (OMS), según una noticia publicada en el portal Primicias.

Este problema de contaminación del aire plantea serias consecuencias para la salud de la población y el medio ambiente. En este contexto, el presente proyecto tiene como objetivo predecir la contaminación del aire en Beijing, China, utilizando un conjunto de datos llamado "Beijing Multi-Site Air-Quality Data Data Set", el cual está disponible en un enlace proporcionado. El período de tiempo de los datos recogidos va desde el 1 de marzo de 2013 hasta el 28 de febrero de 2017.

### **Definición del Problema**

El problema de la contaminación del aire en Beijing, China, ha sido motivo de preocupación debido a los altos niveles de contaminantes atmosféricos presentes en la ciudad. Este proyecto se propone utilizar el conjunto de datos mencionado para analizar y predecir la calidad del aire en Beijing. El conjunto de datos contiene información sobre contaminantes atmosféricos recolectada por 12 sitios de monitoreo de calidad del aire, supervisados a nivel nacional por el Centro de Monitoreo Ambiental Municipal de Beijing.

China, a establecido hace unos años el Índice de Calidad del Aire (AQI, por sus siglas en inglés), el cual se basa en la medición de cinco contaminantes atmosféricos: dióxido de azufre (SO<sub>2</sub>), dióxido de nitrógeno (NO<sub>2</sub>), partículas suspendidas (PM<sub>10</sub>), monóxido de carbono (CO) y ozono (O<sub>3</sub>). Cada contaminante se evalúa individualmente y se le asigna una puntuación. El AQI final corresponde a la puntuación más alta entre los cinco contaminantes. Es importante destacar que la forma de medición varía según el contaminante, ya que SO<sub>2</sub>, NO<sub>2</sub> y PM<sub>10</sub> se miden como promedios diarios, mientras que CO y O<sub>3</sub>, considerados más perjudiciales, se miden como promedios por hora.

### **Procesamiento de Datos**

El procesamiento de datos es un proceso fundamental en el análisis de información, que implica transformar y manipular conjuntos de datos con el fin de obtener información significativa y útil. En el contexto de este informe, el procesamiento de datos se realizó utilizando el lenguaje de programación R.

En este informe, el procesamiento de datos con R fue crucial para obtener información valiosa a partir de los conjuntos de datos utilizados. Esto incluyó desde la exploración inicial de los archivos .csv de las diferentes estaciones de monitoreo, la detección y manejo de valores atípicos o faltantes como los "NA" y "", hasta la transformación de variables y la generación de nuevos archivos. Asimismo, el procesamiento de datos permitió preparar los datos para su posterior análisis y visualización que se lo realizó usando Python, lo que contribuyó a obtener resultados más precisos y significativos.

(a) Limpiar los registros, guardar en un nuevo archivo .csv

Para el procesamiento de datos se utilizó R

### Análisis exploratorio de datos

### Modelado predictivo: configuración experimental y resultados obtenidos;

La configuración inicial se hizo con kernel lineal

## 5.- Calcular valores de exactitud, precisión, esfuerzo y F1

```
In [217]: from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score

accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred, average='micro')
recall = recall_score(y_test, y_pred, average='micro')
f1 = f1_score(y_test, y_pred, average='micro')

print("Accuracy:", accuracy)
print("Precision:", precision)
print("Recall:", recall)
print("F1 Score:", f1)
## haciendo todo lo posible haciendo un ajuste de datos con la desviación estandar
```

  

```
Accuracy: 0.8243559718969555
Precision: 0.8243559718969555
Recall: 0.8243559718969555
F1 Score: 0.8243559718969555
```

## 8.- Se realiza el Test

```
In [288]: from datetime import datetime
def numero_dias(desde_fecha):
    fecha = datetime.strptime(desde_fecha, '%d/%m/%Y')
    fecha_base = datetime(1970, 1, 1)
    diferencia = fecha - fecha_base
    numero_dias = diferencia.days
    return numero_dias
## 2013-10-27 →19 *63 *136 →36 *6 →136 →16005 →poco contaminado
fecha = '27/10/2013'
valor1 = numero_dias(fecha)
valor2 = 19
valor3 = 63
valor4 = 136
valor5 = 36
valor6 = 6

dato_prediccion = [[valor1, valor2, valor3, valor4, valor5, valor6]]
prediccion = best_model.predict(dato_prediccion)

print("La categoría predicha para la fecha", fecha, "es:", prediccion)
```

La categoría predicha para la fecha 27/10/2013 es: [5]

mal

## 8.- Se realiza el Test

```
In [289]: from datetime import datetime
def numero_dias(desde_fecha):
    fecha = datetime.strptime(desde_fecha, '%d/%m/%Y')
    fecha_base = datetime(1970, 1, 1)
    diferencia = fecha - fecha_base
    numero_dias = diferencia.days
    return numero_dias
## 2013-10-27 →19 *63 *136 →36 *6 →136 →16005 →poco contaminado
## 2013-10-28 →26 *61 *174 →57 *44 *174 →16006 →ligeramente contaminado

fecha = '28/10/2013'
valor1 = numero_dias(fecha)
valor2 = 26
valor3 = 61
valor4 = 174
valor5 = 57
valor6 = 44

dato_prediccion = [[valor1, valor2, valor3, valor4, valor5, valor6]]
prediccion = best_model.predict(dato_prediccion)

print("La categoría predicha para la fecha", fecha, "es:", prediccion)
```

La categoría predicha para la fecha 28/10/2013 es: [3]

bien

## **Conclusiones, limitaciones y trabajos futuros;**

### **Conclusiones**

1. Variable objetivo (o variable dependiente): La variable objetivo es aquella que deseas predecir, estimar o modelar. En este proyecto calculamos la variable de valor considerando los 5 gases que se ha proporcionado en los archivos. Se procesa a calcular la variable AQI como se la denominó, con la cual podremos proceder a realizar los análisis y construir el modelo de predicción.
2. Variables independientes (o variables predictoras): Las variables independientes son aquellas que se utilizan para predecir o explicar la variable objetivo. Son las variables que se consideran como factores que podrían influir en la variable objetivo. Estas variables se utilizan para construir el modelo y estimar su impacto en la variable objetivo.

### **Anexos**

Enlace de procesamiento de datos en línea:

[https://colab.research.google.com/drive/1mPzumY82SHW-6oJCdVWWknkO-lvVf\\_Q1?usp=sharing](https://colab.research.google.com/drive/1mPzumY82SHW-6oJCdVWWknkO-lvVf_Q1?usp=sharing)

Enlace del modelado Predictivo en línea:

[https://colab.research.google.com/drive/1eN\\_hZJ6hDDdxjEQDSYSqAPWFhkzB24JT?usp=sharing](https://colab.research.google.com/drive/1eN_hZJ6hDDdxjEQDSYSqAPWFhkzB24JT?usp=sharing)