ESCUELA POLITÉCNICA NACIONAL

Proyecto de Analítica Prescriptiva

Integrantes: Javier Salazar

Edgar Velazco

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_2), dióxido de nitrógeno (NO_2), partículas suspendidas (PM_{10}), monóxido de carbono (CO) y ozono (O_3). 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_2, NO_2 y PM_{10} se miden como promedios diarios, mientras que CO y O_3, 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 realizo 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 utilizo R

o Preprocesamiento de datos;

o Análisis exploratorio de datos;

o Modelado predictivo: configuración experimental y resultados obtenidos;

La configuración inicial se hizo con kernel linear

5.- Calcular valores de exactitud, presición, 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 y la media, descartando valores que generan
```

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

8.- Se realiza el Test

In [221]:
```
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

fecha = '01/10/2013'
valor1 = numero_dias(fecha)
valor2 = 7
valor3 = 36
valor4 = 85
valor5 = 22
valor6 = 21

prediccion = best_model.predict(dato_prediccion)


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

```
La categoría predicha para la fecha 01/10/2013 es: [0]
```

'firefox/central/

o Conclusiones, limitaciones y trabajos futuros;

o Anexos (opcional).

Conclusiones

1. Variable objetivo (o variable dependiente): La variable objetivo es aquella que deseas predecir, estimar o modelar. Es la variable cuyo valor quieres entender o predecir en función de otras variables. Por ejemplo, si estás estudiando el precio de las viviendas, la variable objetivo podría ser el precio de una vivienda. Es importante tener claridad sobre cuál es la variable objetivo antes de comenzar el modelado.

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. Siguiendo el ejemplo anterior, las variables independientes podrían ser características de la vivienda, como el tamaño, la ubicación, el número de habitaciones, etc.

Al seleccionar las variables independientes, es importante considerar:

- Relevancia teórica: Las variables independientes deben tener una base teórica o lógica que las relacione con la variable objetivo. Deben ser factores que razonablemente puedan influir en la variable objetivo.

- Disponibilidad y calidad de los datos: Debes asegurarte de que las variables independientes estén disponibles en tus datos y que sean de calidad suficiente para su análisis.

- Multicolinealidad: Si varias variables independientes están altamente correlacionadas entre sí, podría ser problemático incluirlas todas en el modelo debido a la multicolinealidad. En esos casos, es posible que debas seleccionar solo algunas de ellas o realizar técnicas de reducción de dimensionalidad.

En resumen, identificar la variable objetivo y las variables independientes implica determinar qué variable deseas predecir y qué variables pueden influir en ella. La elección adecuada de estas variables es fundamental para construir un modelo de regresión efectivo y útil.

Enlace de procesamiento de datos esta en gogle colab:

https://colab.research.google.com/drive/1mPzumY82SHW-6oJCdVWWknkO-IvVf_Q1?usp=sharing

Enlace del modelado Predictivo:
https://colab.research.google.com/drive/1eN_hZJ6hDDDxjEQDSYSqAPWFhkzB24JT?usp=sharing