ESCUELA POLITECNICA NACIONAL

FACULTAD DE SISTEMAS

INGENIERIA EN COMPUTACION

DATA MINING Y MACHINE LEARNING

PROYECTO BIMESTRAL

INTEGRANTES

MORALES STHEVVEN

QUISHPE JORDY

CURSO: GR2CC

FECHA DE ENTREGA:30-06-2023

PROFESOR: IVÁN CARRERA

1. INTRODUCCIÓN

El desarrollo de la industrialización y el avance de la tecnología ha llevado a un aumento en los niveles de contaminantes en el aire. En China, ha llevado a un aumento de condiciones climáticas extremas, como un mayor aumento registrado en la temperatura, neblina, smog y efectos adversos en las condiciones de salud de los habitantes de dichas cuidades. La investigación sobre la concentración de estos contaminantes proporciona información que ayuda a la mejora significativa la salud, las condiciones ambientales y beneficios para la economía al tiempo que reduce los costos de las soluciones para reducir la contaminación del aire. Se realizará un análisis con el conjunto de datos de contaminantes atmosféricos por hora de 12 sitios de monitoreo de calidad del aire controlados a nivel local. 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, a saber, dióxido de azufre (SO2), dióxido de nitrógeno (NO2), partículas suspendidas (PM10), monóxido de carbono (CO) y ozono (O3) medidos en las estaciones de monitoreo de cada ciudad. 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.

2. DEFINICIÓN DEL PROBLEMA

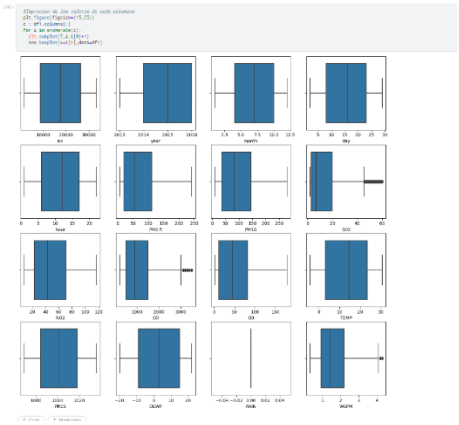
Mediante este proyecto, nuestro objetivo es investigar e incorporar relaciones entre la concentración de contaminantes del aire y las variables meteorológicas a lo largo de un periodo determinado. Crear e implementar un modelo de predicción basado en los niveles de concentración de contaminantes individuales que pueda predecir la calidad del aire. Un sistema que contenga un modelo de predicción que genere advertencias basadas en la calidad del aire son muy necesarios e importantes para un cambio, ya que pueden desempeñar un papel importante en las alertas de salud cuando los niveles de contaminación del aire pueden exceder los niveles especificados. Este proyecto consiste en el desarrollo de un modelo de Machine Learning que sea capaz de predecir el AQI o el Nivel de Contaminación del Aire para un día determinado. Se debe empezar por el conjunto de datos de uno de los sitios de monitoreo y, luego, si es posible, ampliar el estudio a los demás sitios de monitoreo

3. PREPROCESAMIENTO DE DATOS

- **Atípicos**

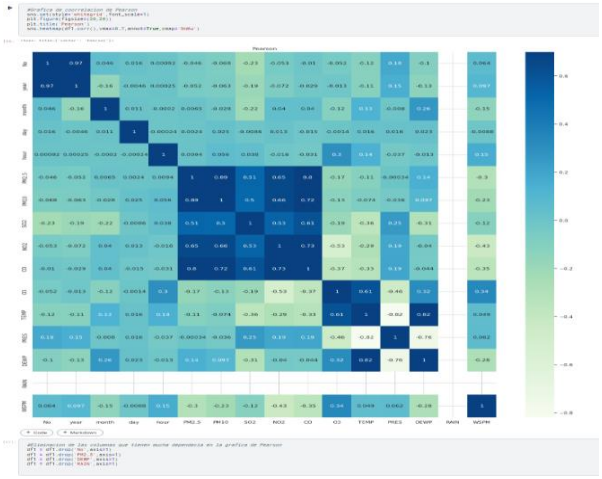
  Se procede a crear una copia del DtaFrame sin que contenga las columnas “wd” y “station” puesto que son de tipo object y no se podría realizar la respectiva grafica. Para esta sección se procedió a usar la lógica de dividir en cuartiles, entonces si se encuentra en valor atípico se lo reemplaza por el promedio al cuartil al que pertenece de esta manera no nos causara mucho ruido a nuestro modelo al momento de entrenar con los datos.

  Grafica después de reemplazar los valores atípicos



- **Pearson**

  Se procede a hacer la gráfica de correlación de Pearson, esto nos dirá si tenemos columnas con mucha dependencia lo cual se procede a eliminar dichas columnas



- **Categóricos**

  Aquí nos centraremos en las 2 columnas de tipo object que son “wd” y “station”, entonces se escogió el método labelencoder en dichas columnas puesto que cada una de estas tiene varios valores dentro, la mejor opción sería usar el método labelencoder.

- **Imputación**

  Se utilizo una imputación simple lo cual hace uso de estas estrategias: Imputación por media, Imputación por mediana, Imputación por moda. Una vez echa la imputación no se tendrían valores nulos

- **Columna AQI**

  Una vez echo el proceso de limpieza se procede a crear la columna AQI la cual consiste en tomar en mayor valor de las columnas 'SO2', 'NO2', 'PM10', 'CO', 'O3'. Después se procede a categorizar dicha columna entre rangos de valores.

4. ANÁLISIS EXPLORATORIO DE DATOS

El conjunto de datos abarca un período de tiempo desde el 1 de marzo de 2013 hasta el 28 de febrero de 2017. Contamos con información sobre 12 sitios de monitoreo de calidad del aire

en Beijing. Las columnas disponibles incluyen: year, month, day, hour, PM10, SO2, NO2, CO, O3, TEMP, PRES, WSPM, wd, station, AQI y AQI Category.

5. MODELADO PREDICTIVO: CONFIGURACIÓN EXPERIMENTAL Y RESULTADOS OBTENIDOS

- **Modelo 1: Support Vector Classifier (SVC)**

  En este modelo, utilizamos un clasificador de vectores de soporte (SVC) para predecir la categoría del Índice de Calidad del Aire (AQI) en Beijing. A continuación, se detalla la configuración experimental y los resultados obtenidos.

  Resultados obtenidos:

```
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)
```

```
Accuracy: 0.9998647401419433
Precision: 0.9998647401419433
Recall: 0.9998647401419433
F1 Score: 0.9998647401419433
```

+ Code + Markdown

- **Modelo 2: cross-validation**

  Este modelo hace uso de una secuencia de dividir los datos en este caso en 5 partes para tomar las partes 1,2,3,4 y validar con la 5, después toma las partes 1,2,3,5 y valida con la 4 y se repite esta secuencia hasta terminar con todas las partes.

  Resultados obtenidos:

[41]:

```
from sklearn.model_selection import cross_val_score

# Example of cross-validation using 5 folds
scores = cross_val_score(model, X, y, cv=5)
print("Cross-Validation Scores:", scores)
print("Average Accuracy:", scores.mean())
```

```
Cross-Validation Scores: [0.99985678 0.99998807 0.99997613 0.99992839 0.9999642 ]
Average Accuracy: 0.9999427135858291
```

- **Modelo 3:**

  En este modelo, utilizamos la técnica de Grid Search para ajustar los hiperparámetros del clasificador de vectores de soporte (SVC) con el objetivo de mejorar su rendimiento en la predicción.

Resultados obtenidos:

[42]:
```
from sklearn.model_selection import GridSearchCV

# Example of hyperparameter tuning using grid search
param_grid = {'C': [0.1, 1, 10], 'kernel': ['linear', 'rbf']}
grid_search = GridSearchCV(model, param_grid, cv=5)
grid_search.fit(X_train, y_train)

best_params = grid_search.best_params_
best_model = grid_search.best_estimator_
print('Best parameters', best_params)
```

```
Best parameters {'C': 1, 'kernel': 'linear'}
```

6. CONCLUSIONES, LIMITACIONES Y TRABAJOS FUTUROS

- El modelo de predicción de la calidad del aire basado en el aprendizaje supervisado lo cual se utiliza para predecir la calidad del aire en Beijing. El objetivo del proyecto era investigar y estudiar los diversos modelos de Machine Learning disponibles para implementar el mejor ajuste para nuestro conjunto de datos. Como equipo, tuvimos la oportunidad de limpiar los datos, observar la importancia de los contaminantes y su impacto en la calidad del aire e implementar algoritmos de ML para predecir el AQI y observar cómo el valor predicho difiere del valor original.

- Este proyecto como tal ayudara en un futuro a que nuestro aprendizaje vaya más allá de solo formulas y comandos, ya que permitirá tener un mejor razonamiento y entendimiento sobre las problemáticas que existe en el mundo real, entonces estos tipos de análisis sirve como un eje fundamental para ayudar a que otras personas puedan entender lo que está pasando en el mundo y así poder contribuir con investigaciones , resultados y análisis para alcanzar un propósito positivo ante la sociedad generando confianza a través de los datos.

- Se concluyo con todos los objetivos que teníamos planteados para este proyecto, en donde era de buscar un modelo de machine learning que sea capaz de predecir el nivel de contaminación en el aire en una parte de China durante un día lo cual los datos que fueron proporcionados para esta análisis fueron de gran ayuda ya que el comportamiento de la contaminación es muy diferente para cada sitio ya que los factores contaminantes varían según que cantidad de sustancias fueron expuestas en el aire en cada uno de los 12 sitios en China. Es bueno contar con algún tipo categorizaciones en donde se pueda establecer agrupaciones con los diferentes resultados en donde se califique con un puntaje de riesgo como alto, medio, bajo entre otras tanto numérico como categórico para así explicar a las personas de cuanto es el daño que están ocasionando con mucha contaminación en el aire y que consecuencias puede traer a futuro para las siguientes generaciones.