ESCUELA POLITÉCNICA NACIONAL
FACULTAD DE INGENIERÍA DE SISTEMAS
Data Mining y Machine Learning

ESTUDIANTE: Daniel Samaniego Zapata.

GRUPO: GR2

INGENIERO: Iván Marcelo Carrera Izurieta

PERÍODO ACADÉMICO: 2023-A

Proyecto 2 – 1er Bimestre

Introducción

El presente proyecto tiene como objetivo predecir la contaminación del aire en Beijing, China, utilizando el conjunto de datos "Beijing Multi-Site Air-Quality Data Data Set".

Este conjunto de datos incluye datos de contaminantes atmosféricos por hora de 12 sitios de monitoreo de calidad del aire controlados a nivel nacional. 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:

AQI	Air Pollution Level	Health Implications
0 - 50	Excellent	No health implications
51 -100	Good	No health implications
101-150	Slightly Polluted	Slight irritations may occur, individuals with breathing or heart problems should reduce outdoor exercise.
151-200	Lightly Polluted	Slight irritations may occur, individuals with breathing or heart problems should reduce outdoor exercise.
201-250	Moderately Polluted	Healthy people will be noticeably affected. People with breathing or heart problems will experience reduced endurance in activities. These individuals and elders should remain indoors and restrict activities.
251-300	Heavily Polluted	Healthy people will be noticeably affected. People with breathing or heart problems will experience reduced endurance in activities. These individuals and elders should remain indoors and restrict activities.
300+	Severely Polluted	Healthy people will experience reduced endurance in activities. There may be strong irritations and symptoms and may trigger other illnesses. Elders and the sick should remain indoors and avoid exercise. Healthy individuals should avoid out door activities.

Ilustración 1Tabla AQI

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.

Definición del problema

El presente proyecto tiene como objetivo predecir la contaminación del aire en Beijing, China, utilizando el conjunto de datos "Beijing Multi-Site Air-Quality Data Data Set", disponible en este [link: http://archive.ics.uci.edu/dataset/501/beijing+multi+site+air+quality+data].

Preprocesamiento de Datos

- Obtener los datos del link mencionado anteriormente.
- Cargar el/los archivos al entorno donde se va a trabajar (Python).
- Leer el/los archivos y proceder a unir los archivos separados.
- Eliminar filas con datos vacíos y columnas que no se utilizarán.
- Comprobamos si existe valores nulos o NA.
- Agrupamos las columnas year, month y day en una sola (date).
- Asignamos un formato adecuado para la columna (date) que será: YYYY-MM-DD, para manejras los datos de una mejor manera.
- Agrupamos los datos por la columna (date) y calculamos la media aritmética para las columnas PM10, SO2, NO2 y el valor máximo para las columnas CO y O3.
- Definimos una función para calcular el AQI de cada gas.
- Para calcular el AQI de cada gas, haciendo uso de la función definida anteriormente.
- Calculamos el AQI Final, que es escogiendo el valor máximo de los 5 gases.
- Según el AQI Final obtenido, asignamos el Nivel de Contaminación del Aire.

Análisis exploratorio de datos

Realizamos un gráfico de dispersión agrupando por “station = Changping” los 20 primeros valores, el tamaño de los puntos esta determinado por el valor de la columna AQI, mientras el valor de AQI aumenta, el tamaño de los puntos también aumenta.

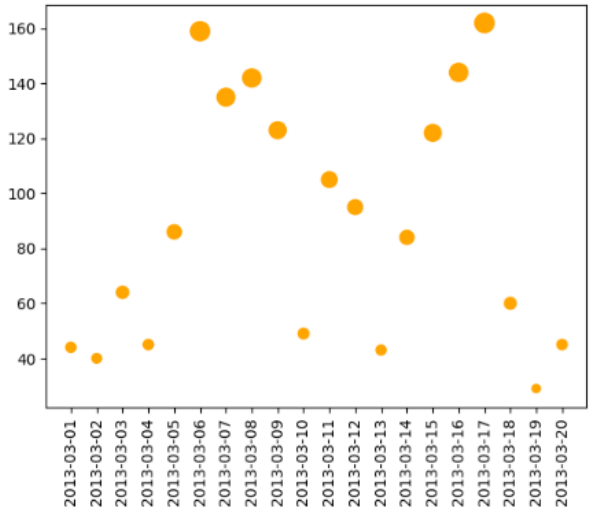

Ilustración 2 Gr. dispersión Station-Changping

Modelado predictivo

Utilizaremos el modelo SVC (Support Vector Classifier).

Definimos la variable **X** como un DataFrame que contiene las columnas de características: 'AQI_PM10', 'AQI_SO2', 'AQI_NO2', 'AQI_CO', 'AQI_O3'. Estas columnas representan las variables predictoras o

características del modelo, definimos la variable **y** como una serie que contiene la columna 'Air_Pollution_Level', esta columna representa nuestra variable objetivo.

Utilizamos la función **train_test_split** para dividir los conjuntos de datos X y y en conjuntos de entrenamiento y prueba. El conjunto de prueba se crea con un tamaño del 30% de los datos originales y se utiliza una semilla aleatoria (random_state=42) para garantizar la reproducibilidad de la división.

Calculamos varias métricas de evaluación como accuracy, precisión, recall y f1 score para evaluar el rendimiento del modelo de clasificación SVM en las predicciones realizadas, obteniendo los siguientes valores.

```
Accuracy: 0.953810623556582
Precision: 0.953810623556582
Recall: 0.953810623556582
F1 Score: 0.953810623556582
```

Mientras más se aproxime a uno quiere decir que es mejor.

Utilizando el método cross_val_score de scikit-learn realizamos una validación cruzada con 5 folds y se muestra los puntajes obtenidos en cada fold y el promedio de los puntajes. Esto ayuda a tener una estimación más precisa del rendimiento del modelo al considerar diferentes particiones del conjunto de datos en el proceso de evaluación. Obteniendo el siguiente resultado:

```
Cross-Validation Scores: [0.96193772 0.9550173  0.95833333
0.95486111 0.95138889]
Average Accuracy: 0.9563076701268743
```

Luego realizamos una búsqueda exhaustiva de hiperparámetros utilizando Grid Search Cross-Validation para encontrar los mejores valores de los hiperparámetros 'C' y 'kernel' para el modelo y mostramos los mejores parámetros encontrados.

```
Best parameters {'C': 10, 'kernel': 'rbf'}
```

Posteriormente realizamos funciones para predecir el nivel de contaminación del aire en función de las variables (PM10, SO2, NO2, CO, O3). Filtramos las filas donde la columna "station" no es igual a 'Changping' y seleccionamos una muestra aleatoria de las filas filtradas. Obteniendo el siguiente resultado:

[122…

	date	station	PM10	SO2	NO2	CO	O3	AQI_PM10	AQI_SO2	AQI_NO2	AQI_CO	AQI_O3	AQI	Air_Pollution_Level
2131	2013-09-07	Shunyi	47.75	1.625	21.458	1900.0	124.0	44	1	11	19	79	79	Good

Por último, utilizando el modelo SVC realizamos la predicción de los datos ingresados manualmente.

```
        PM10     SO2      NO2       CO      O3
2131   47.75   1.625   21.458   1900.0   124.0
       AQI_PM10  AQI_SO2  AQI_NO2  AQI_CO  AQI_O3
2131         44        1       11      19      79
[125… array(['Good'], dtype=object)
```

Conclusiones y trabajos futuros.

- Se llegó a implementar un modelo valido para predecir la contaminación del aire en Beijing, China, utilizando el conjunto de datos "Beijing Multi-Site Air-Quality Data Data Set".
- Para trabajos futuros se puede utilizar el modelo para predecir el nivel de contaminación del aire para otros años, o en su efecto se puede reproducir para predecir el nivel de contaminación del aire en otras ciudades o países.

Anexos

GitHub link:

https://github.com/DanSamaniego/DataMining/tree/65624e51eaf9c52ed5fb1e38c0e270919bed8657/Proyecto%202