ESCUELA POLITÉCNICA NACIONAL

ICCD623 Data Mining y Machine Learning

Nombre: Noemí Uyaguari

Fecha: 30-06-2023

Introducción

La contaminación del aire es un problema ambiental mundial que afecta la calidad de vida de las personas y el equilibrio de los ecosistemas. Las emisiones de gases y partículas contaminantes provenientes de fuentes industriales, vehículos y actividades humanas que pueden tener graves repercusiones en la salud humana, causando enfermedades respiratorias, cardiovasculares y otros problemas de salud.

El objetivo general de este proyecto es desarrollar un modelo predictivo que pueda estimar el nivel de contaminación del aire en la ciudad de Aotizhongxin- China, utilizando datos disponibles sobre diversas características ambientales. El modelo tendrá como objetivo predecir el Índice el Nivel de Contaminación del Aire en función de sus características.

Se espera que este proyecto contribuya a la comprensión de los factores que influyen en la contaminación del aire en Aotizhongxiny proporcione una herramienta útil para pronosticar y monitorear los niveles de contaminación en las demás ciudades e incluso en nuestro país Ecuador. Estos estudios pueden ayudar a las autoridades ambientales a tomar decisiones informadas para implementar medidas de control y mitigación de la contaminación del aire, así como para informar y concienciar a la población sobre los riesgos asociados y la necesidad de adoptar prácticas más sostenibles.

Definición del problema

El problema que se aborda en este proyecto es la predicción de la contaminación del aire, específicamente el Índice de Calidad del Aire (AQI) del componente AQI_PM10, utilizando sus características ambientales y meteorológicas disponibles y se centra en la necesidad de desarrollar un modelo predictivo que pueda predecir la contaminación del aire utilizando características ambientales y meteorológicas. Esto permitirá tomar medidas preventivas, mejorar la calidad del aire y proteger la salud de las personas y el medio ambiente.

Preprocesamiento de datos

- Importar los datos proporcionados de la ciudad de Aotizhongxin desde el archivo CSV.
- Verificar la calidad de los datos, identificar y manejar los valores faltantes, duplicados o inconsistentes.
- Realizar la limpieza de datos necesaria, como eliminar columnas irrelevantes o duplicadas, tratar los valores faltantes.
- Aplicar técnicas de preprocesamiento como la codificación de variables categóricas.

Análisis exploratorio de datos

- Resumen estadístico de los datos

	No	year	month	day	hour	PM2.5	PM10	SO2	NO2	CO	O3	TEMP	PRES	DEWP	RAIN	WSPM
count	25931.000000	25931.000000	25931.000000	25931.000000	25931.000000	25931.000000	25931.000000	25931.000000	25931.000000	25931.000000	25931.000000	25931.000000	25931.000000	25931.000000	25931.000000	25931.000000
mean	18051.203502	2014.727123	6.525279	15.691412	11.453280	60.483838	84.738282	10.704189	51.852759	897.267147	54.339418	14.639564	1011.493979	3.764429	0.081443	1.767884
std	9993.137712	1.166904	3.244493	8.791221	6.960841	55.030635	65.957013	10.498739	30.462278	603.449849	46.694846	10.784589	10.412695	14.088958	0.920036	1.243613
min	1.000000	2013.000000	1.000000	1.000000	0.000000	3.000000	2.000000	0.571200	2.000000	100.000000	0.214200	-16.800000	985.900000	-35.300000	0.000000	0.000000
25%	10010.500000	2014.000000	4.000000	8.000000	6.000000	17.000000	31.000000	3.000000	27.000000	400.000000	12.000000	5.400000	1003.000000	-8.200000	0.000000	0.900000
50%	18449.000000	2015.000000	7.000000	16.000000	11.000000	44.000000	69.000000	7.000000	48.000000	700.000000	47.000000	16.400000	1010.500000	5.700000	0.000000	1.500000
75%	26863.500000	2016.000000	9.000000	23.000000	18.000000	87.000000	123.000000	15.000000	71.000000	1200.000000	82.000000	23.500000	1019.700000	16.400000	0.000000	2.300000
max	35064.000000	2017.000000	12.000000	31.000000	23.000000	366.000000	329.000000	48.000000	160.000000	3000.000000	193.000000	40.100000	1042.000000	28.500000	46.400000	11.200000

- Medidas de tendencia central; aplicado al campo PM10

```
Media de PM10: 84.73828236473 72
Mediana de PM10: 69.0
Moda de PM10: 0    6.0
Name: PM10, dtype: float64
```

- Dispersión de los datos; aplicado al campo PM10

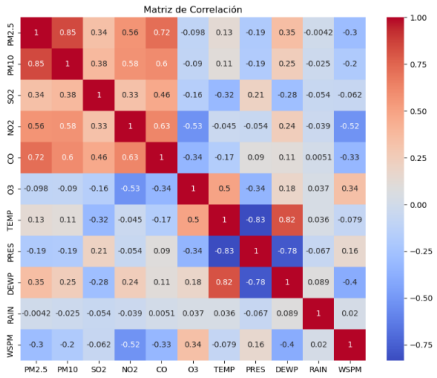
```
Desviación estandar de PM10: 65.95701312735798
Rango: 327.0
Coeficiente de Variancón de PM10: 0.7783614593869227
```

- Correlación entre variables

Se puede observar la intensidad y dirección de las correlaciones entre las variables.

Los colores más oscuros representan una correlación más fuerte (positiva o negativa)

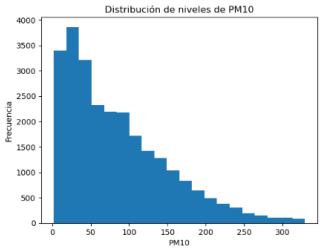
Los colores más claros representan una correlación más débil o cercana a cero.



Preguntas formuladas

¿Cuál es la distribución de las variables relacionadas con la contaminación del aire?

Gráfico de distribución de niveles de los contaminantes. Para PM10

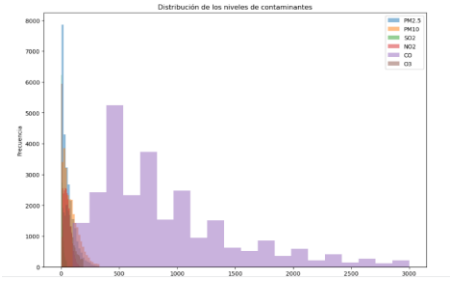


Tendencia de los datos

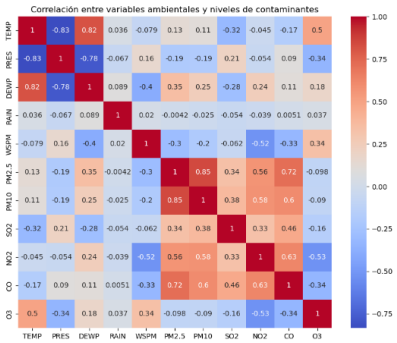
El gráfico muestra un acantilado izquierdo, significa que la distribución de PM10 está sesgada hacia la izquierda, esto implica que hay una concentración de valores más altos en el extremo izquierdo del gráfico y una disminución gradual a medida que avanza hacia la derecha.

Esto puede sugerir que la contaminación del aire en general se mantiene en niveles bajos, pero ocasionalmente puede haber eventos o períodos de tiempo con niveles más altos.

Gráfico de distribución de niveles de los contaminantes combinados



¿Existe alguna relación entre las variables ambientales y los niveles de contaminantes?

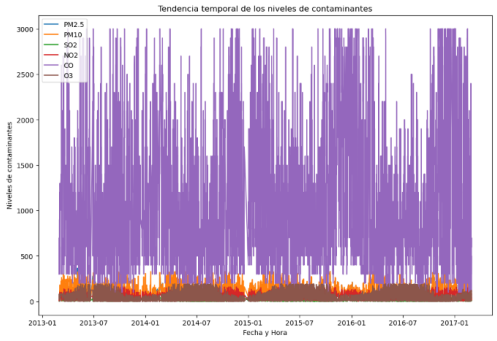


TEMP	PRES	DEWP	RAIN	WSPM	PM2.5	SO2	NO2	CO	O3
0.13	-0.19	0.25	-0.025	-0.2	0.85	0.38	0.58	0.6	-0.09
Correlación positiva moderada	Correlación negativa débil	Correlación positiva débil	No existe una correlación fuerte	Correlación negativa moderada	Correlación positiva fuerte	Correlación positiva moderada	Correlación positiva moderada	Correlación positiva moderada	No existe una correlación fuerte
A medida que la temperatura aumenta, los niveles de PM10 tienden a aumentar	A medida que la presión atmosférica aumenta, los niveles de PM10 tienden a disminuir	A medida que el rocío aumenta, los niveles de PM10 tienden a aumentar	La precipitación no muestra una relación clara con los niveles de PM10	A medida que la velocidad del viento aumenta, los niveles de PM10 tienden a disminuir	Los niveles de PM10 están influenciados en gran medida por los niveles de partículas PM2.5	PM10 tienden a aumentar junto con los niveles de SO2	PM10 tienden a aumentar junto con los niveles de NO2	PM10 tiende a aumentar junto con C0	El ozono no muestra una relación clara con los niveles de PM10

Se observa que los niveles de PM10 están más estrechamente relacionados con los niveles de partículas PM2.5 y los contaminantes gaseosos como SO2, NO2 y CO. La temperatura, la presión atmosférica, la velocidad del viento también tiene influencia.

¿Hay alguna tendencia temporal en los niveles de contaminación?

Gráfico de series de tiempo que muestra cómo los valores de diferentes contaminantes (PM2.5, PM10, SO2, NO2, CO, O3) cambian a lo largo del tiempo.

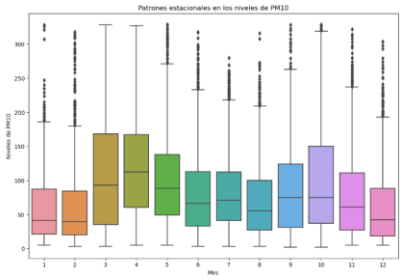


Los niveles de CO son significativamente más altos en comparación con los otros contaminantes, esto sugiere que el CO podría ser uno de los principales contribuyentes a la contaminación del aire en la ciudad de Aotizhongxin.

La agrupación de los demás colores en la parte inferior del gráfico podría significar que los niveles de los otros contaminantes son relativamente bajos, esto podría indicar que los niveles de los otros contaminantes son más estables o menos variables en el tiempo.

¿Existen patrones estacionales en la calidad del aire?

Para PM10



Mediana: No existen cambios consistentes en la posición de la mediana a lo largo del tiempo, por lo tanto, se podría decir que no existen patrones estacionales definitivos. Como la mediana es más alta en ciertos meses o estaciones, indica que los niveles de contaminantes tienden a ser más altos durante esos períodos.

Altura de la caja: La altura de la caja es mayor en ciertos meses o estaciones, lo que indica una mayor variabilidad en los niveles de PM10 durante esos períodos. Esto podría sugerir fluctuaciones estacionales en la calidad del aire.

Longitud de los bigotes: Los bigotes varían en longitud entre diferentes meses o estaciones, indica diferencias en los niveles extremos de PM10 excepcionalmente altas durante esas estaciones.

Valores Atípicos: Presencia de valores fuera de los rangos definidos, que podrías representar casos con excepciones.

Modelado predictivo

- Calcular puntuaciones individuales para cada contaminante, utilizando las fórmulas específicas correspondientes a cada uno.
- Comparar las puntuaciones AQI individuales de cada contaminante y seleccionar la puntuación más alta.

	SO2	AQI_SO2	NO2	AQI_NO2	PM10	AQI_PM10	CO	AQI_CO	O3	\
0	4.0	4.0	7.0	8.750000	4.0	4.0	300.0	-1	77.0	
1	4.0	4.0	7.0	8.750000	8.0	8.0	300.0	-1	77.0	
2	5.0	5.0	10.0	12.500000	7.0	7.0	300.0	-1	73.0	
3	11.0	11.0	11.0	13.750000	6.0	6.0	300.0	-1	72.0	
4	12.0	12.0	12.0	15.000000	3.0	3.0	300.0	-1	72.0	
...	...	...	...	...	...	...	...	...	...	
25926	5.0	5.0	35.0	43.750000	29.0	29.0	400.0	-1	95.0	
25927	7.0	7.0	45.0	56.025641	37.0	37.0	500.0	-1	81.0	
25928	10.0	10.0	66.0	82.410256	37.0	37.0	700.0	-1	58.0	
25929	12.0	12.0	87.0	103.969697	44.0	44.0	700.0	-1	35.0	
25930	10.0	10.0	79.0	98.743590	31.0	31.0	600.0	-1	42.0	

	AQI_O3
0	122.000000
1	122.000000
2	108.000000
3	104.500000
4	104.500000
...	...
25926	174.210526
25927	136.000000
25928	60.800000
25929	32.407407
25930	38.888889

- El valor final del AQI será la puntuación más alta obtenida.

Definir la tarea de predecir el nivel de contaminación del aire (AQI) en función de las características disponibles.

Configuración experimental

El modelo sugerido para el estudio es un modelo de regresión de bosques aleatorios (Random Forest Regressor), que es una elección adecuada para la tarea de predicción de una variable continua como el nivel de contaminación del aire en función de PM10.

Selección de características y variable objetivo

Se seleccionan las características ambientales (year, month, day, hour, PM2.5, PM10, SO2, NO2, CO, O3, TEMP, PRES, DEWP, RAIN, wd, WSPM) como las variables de entrada (X) y la variable objetivo 'AQI_PM10' como la variable que queremos predecir (y).

Dividir los datos en conjuntos de entrenamiento y prueba

Los datos se dividen en conjuntos de entrenamiento (X_train, y_train) y prueba (X_test, y_test) utilizando la función train_test_split de sklearn. El 20% de los datos se utiliza para el conjunto de prueba, mientras que el 80% se utiliza para entrenar el modelo.

Creación y entrenamiento del modelo

Se crea una instancia del modelo RandomForestRegressor con 100 estimadores (árboles) y un estado aleatorio de 42 para reproducibilidad. Luego, el modelo se ajusta utilizando los datos de entrenamiento codificados.

Predicciones y evaluación del rendimiento

Se utilizan los datos de prueba codificados para realizar predicciones sobre los valores del índice de calidad del aire (AQI) para PM10 utilizando el modelo entrenado. Se evalúa el rendimiento del modelo calculando el error cuadrático medio (MSE) y el coeficiente de determinación R^2.

```
Error cuadrático medio (MSE): 0.00019555638792853475
Coeficiente de determinación R^2: 0.9999998656986252
```

Mejorar el rendimiento y la capacidad predictiva del modelo

Se procede a realizar una búsqueda en la cuadrícula de hiperparámetros especificada y encontrar los mejores hiperparámetros para un modelo de regresión de bosques aleatorios.

Justificación de la elección del modelo

El modelo adoptado se considera adecuado para abordar este tipo de problema de predicción debido a su:

- Flexibilidad: Los bosques aleatorios son modelos no paramétricos, lo que significa que no hacen suposiciones sobre la forma funcional de los datos y son flexibles para ajustarse a relaciones complejas y no lineales entre las características y la variable objetivo.
- Manejo de características: Los bosques aleatorios pueden manejar características categóricas y numéricas sin requerir una codificación especial, como se muestra en el código utilizando el OneHotEncoder para codificar la variable categórica 'wd'.
- Reducción de sobreajuste (overfitting): Los bosques aleatorios combinan múltiples árboles de decisión, lo que reduce el riesgo de sobreajuste al promediar las predicciones de cada árbol y limitar la profundidad de los árboles individuales.
- Robustez: Los bosques aleatorios son robustos frente a valores atípicos (outliers) y ruido en los datos, ya que utilizan múltiples árboles y promedian sus resultados.

Resultados obtenidos

- Error cuadrático medio (MSE): es una métrica que mide el promedio de los errores al cuadrado entre las predicciones del modelo y los valores reales.
  El valor de MSE obtenido es 0.00019555638792853475, al ser un valor cercano a cero indica que el modelo tiene un ajuste muy cercano a los datos reales, lo cual es un resultado muy positivo.

- Coeficiente de determinación R^2: representa la proporción de la varianza en la variable objetivo que es explicada por el modelo.
  El valor de R^2 obtenido es 0.9999998656986252, al ser un valor cercano a 1 indica que el modelo explica la gran mayoría de la varianza en los datos, lo cual es un resultado excelente.

En resumen, los resultados obtenidos indican que el modelo de regresión de bosques aleatorios se ajusta muy bien a los datos y es capaz de predecir con gran precisión el nivel de contaminación del aire (AQI_PM10) utilizando las características proporcionadas.

Conclusiones

- En este análisis se utilizó un modelo de regresión de bosques aleatorios para predecir el nivel de contaminación del aire (AQI_PM10) utilizando diversas características relacionadas con la calidad del aire y datos meteorológicos. Los resultados obtenidos mostraron un

ajuste muy bueno del modelo, con un error cuadrático medio (MSE) cercano a cero y un coeficiente de determinación (R^2) cercano a 1. Esto indica que el modelo es capaz de explicar y predecir la variabilidad en los niveles de contaminación del aire con gran precisión.

- Se debe tener en cuenta que el R^2 extremadamente cercano a 1 puede sugerir un riesgo de sobreajuste, y es posible que se necesite una validación adicional del modelo en datos no vistos para asegurar que su rendimiento sea igualmente bueno en nuevos conjuntos de datos.

Limitaciones

- Generalización de resultados de los resultados obtenidos se basan en los datos utilizados en este estudio específico. Es posible que los resultados no sean generalizables a otros conjuntos de datos o a diferentes condiciones ambientales y geográficas.
- Es necesario realizar pruebas y validaciones adicionales en diferentes conjuntos de datos para evaluar la robustez del modelo.
- De la optimización del modelo, se obtuvieron resultados prometedores, pero es posible que exista margen para mejorar y optimizar el modelo. Se pueden explorar diferentes técnicas de preprocesamiento de datos, ajustar los hiperparámetros del modelo y probar otros algoritmos de aprendizaje automático para obtener un mejor rendimiento.

Trabajos futuros

Con base en los resultados y las limitaciones mencionadas, se sugieren puntos de mejora y trabajos futuros como la exploración de diferentes modelos, la optimización del modelo actual, la validación en conjuntos de datos independientes y la incorporación de características adicionales.

Bibliografía

https://github.com/ivan-carrera/handson_2023A/blob/master/notebooks/04_ModelEvaluation.ipynb

https://www.kaggle.com/code/dansbecker/using-categorical-data-with-one-hot-encoding/notebook

https://towardsdatascience.com/random-forest-in-python-24d0893d51c0

https://www.cienciadedatos.net/documentos/py08_random_forest_python