Analítica Prescriptiva

Mendoza Mestanza Gutenberg Stiven[1], Moya Barragán Fabricio Alejandro[2]

Resumen— **Este estudio presenta el desarrollo de un modelo predictivo basado en Árbol de decisión para clasificar el nivel de contaminación ambiental. Se utiliza la librería LazyPredict de Python para evaluar diferentes modelos y se analizan las métricas obtenidas. Se discuten las limitaciones identificadas y se sugieren trabajos futuros. Este enfoque puede contribuir a la toma de decisiones informadas y a la implementación de medidas de protección ambiental más efectivas. .**

Palabras clave— **contaminación ambiental, modelado predictivo, Árbol de decisión, LazyPredict, clasificación.**

I. Introducción

La contaminación ambiental es un problema global que afecta la calidad de vida de las personas y el equilibrio ecológico del planeta. La capacidad de predecir y clasificar el nivel de contaminación en diferentes áreas geográficas es crucial para implementar medidas efectivas de mitigación y control. En este contexto, el modelado predictivo se presenta como una herramienta prometedora para abordar este desafío.

El objetivo de este estudio es desarrollar un modelo predictivo capaz de clasificar el nivel de contaminación en base a datos recopilados de diferentes fuentes y variables ambientales relevantes. Para lograrlo, se emplea la librería LazyPredict de Python, que permite evaluar diferentes modelos de clasificación y determinar cuál presenta la mejor exactitud en base a un conjunto de datos de muestra.

En este informe, se presenta el proceso de modelado predictivo realizado utilizando un modelo basado en Árbol de decisión (DecisionTreeClassifier) y se analizan las métricas obtenidas para evaluar su desempeño. Asimismo, se explora la utilización de la validación cruzada para obtener una evaluación más robusta del modelo.

Las limitaciones identificadas durante el estudio y las sugerencias de trabajos futuros se discuten con el fin de mejorar y ampliar el modelo predictivo en futuras investigaciones. En última instancia, se espera que este estudio contribuya al desarrollo de herramientas efectivas para la clasificación y predicción del nivel de contaminación, permitiendo tomar decisiones informadas y implementar medidas de protección ambiental más eficientes.

[1]Facultad de Ingeniería en Sistemas Informáticos, Ciencias de la Computación, e-mail: gutenberg.mendoza@epn.edu.ec

[2]Facultad de Ingeniería en Sistemas Informáticos, Ciencias de la Computación, e-mail: fabricio.moya@epn.edu.ec

II. Definición del problema

La contaminación del aire es un problema importante que enfrentan ciudades alrededor del mundo, y Beijing, China, no es una excepción. La predicción precisa de la calidad del aire es esencial para mitigar los efectos adversos en la salud humana y el medio ambiente. En este contexto, el presente proyecto tiene como objetivo principal desarrollar un modelo de Machine Learning que pueda predecir el Índice de Calidad del Aire (AQI) en Beijing.

El AQI se basa 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). Cada uno de estos contaminantes se mide de manera diferente, con algunos calculados como un promedio diario y otros como un promedio por hora. El AQI final para un día específico se calcula como la puntuación más alta de estos cinco contaminantes.

Para llevar a cabo este objetivo, se utilizará el conjunto de datos "Beijing Multi-Site Air-Quality Data Data Set", que incluye datos por hora de contaminantes atmosféricos de 12 sitios de monitoreo de calidad del aire controlados a nivel nacional, desde el 1 de marzo de 2013 hasta el 28 de febrero de 2017.

Inicialmente, el proyecto comenzará con el análisis y modelado de los datos de un solo sitio de monitoreo. Sin embargo, si es posible, el estudio se ampliará para incluir todos los sitios de monitoreo. Se espera que este enfoque ayude a generar un modelo de predicción robusto y generalizable para el AQI en Beijing. El proyecto también permitirá a los estudiantes aumentar el valor de su trabajo incluyendo otras tareas relevantes en el proceso de análisis y modelado.

III. Preprocesamiento de datos

En nuestra primera etapa de preprocesamiento de datos, nos embarcamos en un análisis de dimensionalidad de múltiples conjuntos de datos con estructuras similares. Este análisis nos permitió determinar que todos los conjuntos de datos eran estructuralmente idénticos, lo que significa que el análisis de un conjunto de datos individual sería representativo del análisis de todos los conjuntos juntos. A continuación, desarrollamos una función para calcular el Índice de Calidad del Aire (AQI) basándonos en los contaminantes presentes y lo separamos en cuatro categorías distintas: Excellent - Good, Slightly - Lightly Polluted, Moderately - Heavily Polluted, y Severely Polluted.[1]

Empleamos la correlación de Pearson para identificar correlaciones dentro de los datos, y establecimos un umbral de 0.8 para eliminar aquellas columnas que exhibían una alta correlación. Este proceso nos

permitió minimizar la redundancia y mejorar la eficiencia del modelo que estamos entrenando.

A continuación, abordamos la presencia de valores nulos en los conjuntos de datos. Para la mayoría de las columnas, reemplazamos estos valores con la media. Sin embargo, para la columna 'wd', implementamos una técnica llamada codificación 'dummy' o 'one-hot'. En el caso de las columnas 'station' y 'AQI', aplicamos codificación de etiquetas (label encoding) para convertir los datos categóricos en una forma que nuestro modelo de aprendizaje automático podría interpretar y procesar.

Para finalizar, equilibramos nuestro conjunto de datos utilizando la técnica RandomOverSampler, lo cual ayuda a evitar un sesgo hacia clases más frecuentes. Este proceso de preprocesamiento es esencial para preparar nuestros datos para la formación del modelo de aprendizaje automático.[2]

IV. Análisis exploratorio de datos

El análisis exploratorio de datos es un proceso que permite conocer a priori la naturaleza de los conjuntos.

El análisis exploratorio de datos (AED) es una etapa crucial en la investigación y el análisis de datos. Su objetivo es comprender la estructura y las características de un conjunto de datos antes de aplicar técnicas estadísticas más avanzadas o construir modelos predictivos. El AED ayuda a identificar patrones, tendencias, valores atípicos y relaciones entre variables, lo que permite formular hipótesis y generar conocimientos preliminares sobre los datos, respondiendo así algunas preguntas como:

1. ¿Cuál es la distribución de las variables en el conjunto de datos?
2. ¿Existen correlaciones entre las variables?
3. ¿Hay valores atípicos que necesiten ser investigados?
4. ¿Cuáles son las características más relevantes o distintivas del conjunto de datos?
5. ¿Hay patrones o tendencias interesantes que se puedan identificar?

En particular en esta ocasión nos permite obtener algunas posibles conclusiones:

1. **Patrones temporales:** Al analizar los datos a lo largo del período de tiempo (del 1 de marzo de 2013 al 28 de febrero de 2017), es posible identificar patrones temporales en la contaminación del aire. Esto puede revelar estacionalidades, tendencias a largo plazo y variaciones interanuales en la calidad del aire en Beijing.
2. **Variaciones geográficas:** Al tener datos de 12 sitios de monitoreo de calidad del aire en Beijing, es posible analizar las variaciones geográficas en los niveles de contaminantes. Esto puede revelar áreas específicas que son más susceptibles a altos niveles de contaminación o identificar diferencias significativas en la calidad del aire entre diferentes partes de la ciudad.
3. **Correlaciones entre contaminantes:** Al analizar los datos de los cinco contaminantes atmosféricos (SO2, NO2, PM10, CO y O3), se pueden identificar correlaciones y relaciones entre ellos. Por ejemplo, es posible que se encuentre una correlación positiva entre los niveles de NO2 y PM10, lo que sugiere que ciertos contaminantes están relacionados entre sí.
4. **Influencia de los factores meteorológicos:** Al relacionar los datos de calidad del aire con los datos meteorológicos de la estación más cercana, se puede investigar la influencia de los factores meteorológicos en la contaminación del aire. Esto puede revelar cómo variables como la temperatura, la humedad, la velocidad del viento, etc, pueden afectar los niveles de contaminantes y proporcionar información útil para comprender los factores que contribuyen a la calidad del aire en Beijing.
5. **Interpretación del Índice de Calidad del Aire (AQI):** Al comprender cómo se calcula y se interpreta el Índice de Calidad del Aire (AQI) en función de los cinco contaminantes atmosféricos mencionados (SO2, NO2, PM10, CO y O3), se puede proporcionar una interpretación más detallada y precisa de la calidad del aire en diferentes días y contextos.

El objetivo es obtener una comprensión más profunda de la calidad del aire en Beijing, identificar patrones y relaciones clave, y proporcionar información útil para futuras investigaciones y medidas de control de la contaminación o en el desarrollo de modelos.[3]

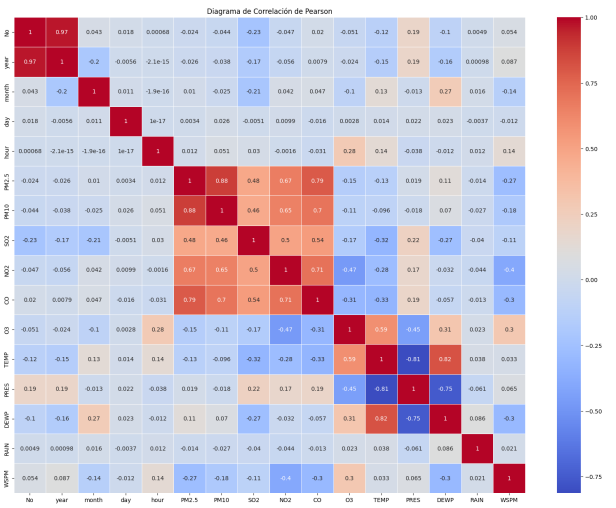

Fig. 1: Matriz de Correlación de Pearson.

Informe: Análisis de correlaciones en el conjunto de datos de calidad del aire de Beijing

La matriz de correlación de Pearson se usa con el objetivo de comprender las correlaciones entre las variables y proporcionar una visión más profunda de la contaminación del aire en Beijing, se presentan los hallazgos clave derivados del análisis de correlaciones.

A. *Correlación positiva fuerte*

Se observó una correlación positiva fuerte (0.884380) entre los niveles de partículas suspendidas de diámetro menor a 2.5 micrómetros (PM2.5) y los niveles de partículas suspendidas de diámetro menor a 10 micrómetros (PM10). Esta correlación indica que los niveles de partículas PM2.5 están altamente correlacionados con los niveles de partículas PM10. Es importante tener en cuenta que las partículas PM2.5 son un subconjunto de las partículas PM10.

B. *Correlaciones positivas moderadas*

Se encontraron correlaciones positivas moderadas entre los niveles de partículas PM2.5 y los niveles de monóxido de carbono (CO) (0.789998) y dióxido de nitrógeno (NO2) (0.666948). Estas correlaciones sugieren que los niveles de partículas suspendidas de diámetro menor a 2.5 micrómetros están relacionados con la presencia de CO y NO2 en el aire. Esto indica que la presencia de CO y NO2 puede contribuir a la contaminación del aire por partículas PM2.5.

C. *Correlación negativa moderada*

Se encontró una correlación negativa moderada (-0.349856) entre los niveles de dióxido de azufre (SO2) y los niveles de dióxido de nitrógeno (NO2). Esta correlación inversa indica que los niveles de SO2 y NO2 en el aire tienden a tener una relación inversa. Es importante considerar esta relación en el contexto de la calidad del aire, ya que la presencia de estos contaminantes puede tener diferentes fuentes y efectos en la salud y el medio ambiente.

D. *Correlación con variables meteorológicas*

Se observaron correlaciones débiles a moderadas entre las variables de contaminantes atmosféricos (PM2.5, PM10, SO2, NO2, CO, O3) y las variables meteorológicas (TEMP, PRES, DEWP, RAIN, WSPM). Estas correlaciones indican que los factores meteorológicos pueden influir en los niveles de contaminantes en el aire. Por ejemplo, la temperatura (TEMP) puede estar relacionada con la formación de ozono (O3), y la velocidad del viento (WSPM) puede afectar la dispersión de los contaminantes. Sin embargo, es importante destacar que estas correlaciones son de naturaleza débil a moderada y se requiere un análisis más profundo para comprender plenamente la influencia de los factores meteorológicos en la contaminación del aire.

En conclusión, el análisis de correlaciones ha proporcionado información valiosa sobre las relaciones entre las variables. Estas correlaciones destacan la fuerte relación entre las partículas PM2.5 y PM10, así como las asociaciones moderadas entre los contaminantes atmosféricos y los factores meteorológicos. Estos hallazgos contribuyen a una mejor comprensión de los factores y pueden servir como base para investigaciones posteriores y medidas de control de la contaminación atmosférica en la región.[4]

E. *Simetría de los datos*

Se analizó la diferencia entre la media y la mediana de las variables numéricas en el conjunto de datos, con el fin de obtener información sobre la simetría de la distribución de los datos y la presencia de valores atípicos.

- **Diferencias nulas:** Se observó que las variables "No" "hour" presentan una diferencia de cero entre la media y la mediana. Esto indica que no hay una discrepancia significativa entre estos valores estadísticos para dichas variables. En otras palabras, la distribución de los datos para estas variables es simétrica.
- **Diferencias negativas:** Las variables "ear", "month", "day", "TEMP" "DEWP" mostraron diferencias negativas entre la media y la mediana. Estas diferencias sugieren que estas variables tienden a tener una distribución asimétrica hacia la izquierda. Es decir, la concentración de valores más bajos arrastra la media hacia abajo en comparación con la mediana.
- **Diferencias positivas:** Por otro lado, se encontró que las variables "PM2.5", "PM10", "SO2", "NO2", "CO", ".o3", "PRES", "RAIN" "WSPM" presentaron diferencias positivas entre la media y la mediana. Esto indica que estas variables tienden a tener una distribución asimétrica hacia la derecha. En otras palabras, la presencia de valores más altos influye en la media y la aleja de la mediana.

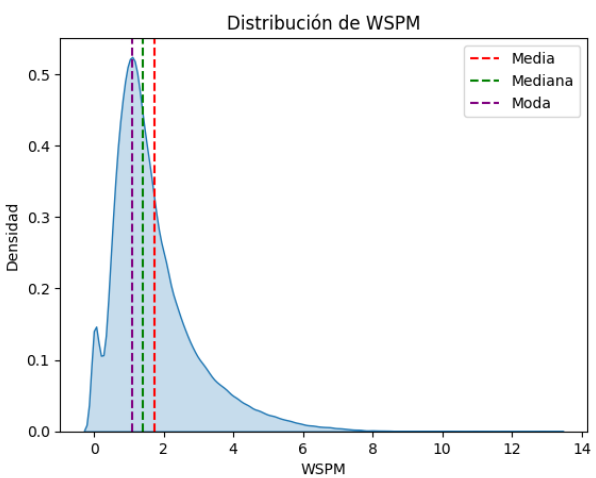

Fig. 2: Simetría WSPM.

Es importante tener en cuenta que estas diferencias entre la media y la mediana son indicadores de la distribución de los datos y pueden sugerir la presencia de valores atípicos o asimetría en los conjuntos de datos correspondientes a cada variable.

F. *Tendencias y variaciones*

Obtener gráficos de puntos que representen la mediana de diferentes variables relacionadas con la calidad del aire, como PM2.5, PM10, SO2, NO2, CO, O3, entre otras, en función de los meses y los años.

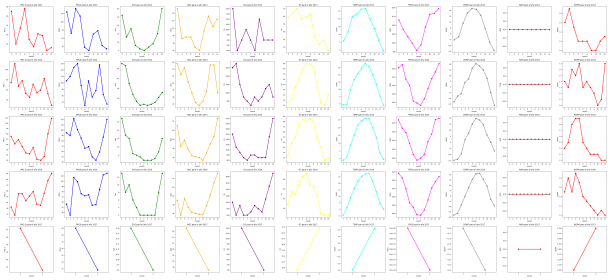

Fig. 3: Simetría WSPM.

El análisis permite una exploración visual detallada de la calidad del aire en Beijing a lo largo de los años y los meses. Esto proporcionaría información clave sobre patrones estacionales, cambios a largo plazo y eventos excepcionales en la calidad del aire en la región, lo que sería fundamental para comprender y abordar los desafíos relacionados con la contaminación atmosférica.[5]

El análisis exploratorio de datos es un proceso iterativo, y es posible que nuevas preguntas e hipótesis surjan a medida que se profundiza en los datos y se descubren nuevas perspectivas.

V. Modelado predictivo

En esta sección, se presenta el modelado predictivo realizado para clasificar los datos según el nivel de contaminación. Para ello, se empleó la librería de Python LazyPredict[**?**], la cual nos permitió identificar qué modelo presenta la mejor exactitud sobre un conjunto de datos de muestra.

Model	Accuracy	Balanced Accuracy	ROC AUC	F1 Score	Time Taken
DecisionTreeClassifier	1.00	0.93	None	1.00	0.31
BaggingClassifier	1.00	0.92	None	1.00	0.36
GaussianNB	0.99	0.76	None	0.99	0.16
ExtraTreesClassifier	0.99	0.69	None	0.99	2.20
ExtraTreeClassifier	0.96	0.58	None	0.96	0.32
AdaBoostClassifier	0.97	0.33	None	0.95	3.77
BernoulliNB	0.82	0.26	None	0.84	0.28
CalibratedClassifierCV	0.91	0.21	None	0.88	8.28
DummyClassifier	0.89	0.17	None	0.84	0.25

Fig. 4: Salida de LazyPredictor

El análisis de LazyPredictor nos indica que se recomienda utilizar un modelo basado en Árbol de decisión (DecisionTreeClassifier). A continuación, se procedió a entrenar el modelo utilizando este algoritmo y posteriormente se analizaron las métricas obtenidas.

El modelo fue entrenado inicialmente sin especificar los hiperparámetros del Árbol de decisión, tales como max_depth, min_samples_leaf y min_samples_split. Los resultados obtenidos se muestran en la siguiente figura:

A partir de los resultados presentados en el reporte, se pueden extraer diversas conclusiones relevantes.

En primer lugar, el modelo obtuvo una precisión del 99.09 %, lo cual indica que fue capaz de clasificar correctamente el 99.09 % de las instancias en el

```
Accuracy: 99.09%
Classification report:
              precision    recall  f1-score   support

           0       0.99      1.00      1.00     19028
           1       1.00      0.98      0.99     19101
           2       0.98      1.00      0.99     18862
           3       1.00      0.99      0.99     19037

    accuracy                           0.99     76028
   macro avg       0.99      0.99      0.99     76028
weighted avg       0.99      0.99      0.99     76028
```

Fig. 5: Reporte de clasificación

conjunto de datos. Esta alta precisión evidencia la capacidad del modelo para realizar una clasificación confiable del nivel de contaminación.[6]

El reporte de clasificación también incluye métricas como recall y F1-Score para cada clase. Estas métricas proporcionan información adicional sobre la capacidad del modelo para identificar correctamente las instancias positivas y negativas. En general, se observa que el modelo obtuvo valores elevados de recall y F1-Score para todas las clases. Esto indica que el modelo logró identificar la mayoría de las instancias reales de cada clase y alcanzó un equilibrio entre precisión y recall.[7]

Además, es relevante destacar que el modelo fue evaluado en un conjunto de datos de considerable tamaño, con un total de 76,028 instancias. Esta cantidad de datos aumenta la confianza en los resultados y sugiere que el modelo ha sido probado en un escenario realista.

A. Validación cruzada

La validación cruzada es una técnica esencial para evaluar el rendimiento del modelo de manera robusta. En este caso, se realizó una validación cruzada y se obtuvieron los siguientes datos:

Mejores parámetros:

- max_depth: None
- min_samples_leaf: 1
- min_samples_split: 2

Puntuaciones de validación cruzada:

0.99019166, 0.98928974, 0.99036077, 0.99007892, 0.99024803, 0.99

Puntuación promedio de validación cruzada: 0.990033590214823

Exactitud: 99.10 %

Estos resultados confirman y respaldan el desempeño del modelo de Árbol de decisión en la clasificación del nivel de contaminación. La obtención de una puntuación promedio de validación cruzada alta indica que el modelo es consistente y generaliza bien en diferentes subconjuntos de datos, esto también se puede evidenciar en la matriz de confusión

En conclusión, el modelo de Árbol de decisión ha demostrado ser altamente efectivo en la clasificación del nivel de contaminación. Los resultados obtenidos,

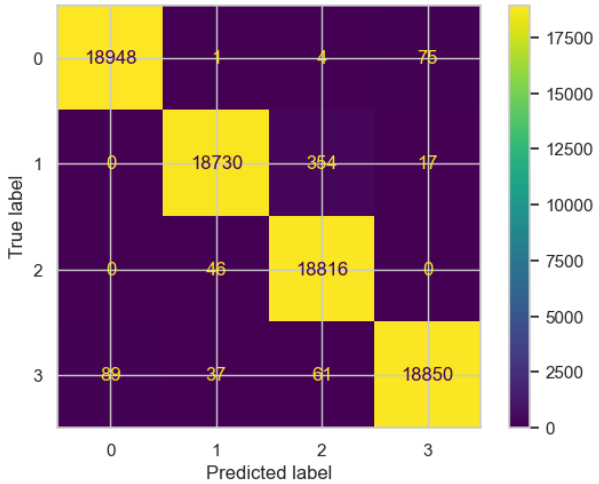

Fig. 6: Matriz de confusión

respaldados por las métricas de evaluación y la validación cruzada, indican que el modelo es confiable y preciso en su capacidad para clasificar correctamente el nivel de contaminación en función de los datos proporcionados. Estos resultados son alentadores y sugieren la aplicabilidad del modelo en situaciones reales relacionadas con la clasificación de la contaminación.[8]

Finalmente una metrica muy importante es comprender si el modelo tiene overfitting o underfitting, eso se puede visualizar en la siguiente grafica:

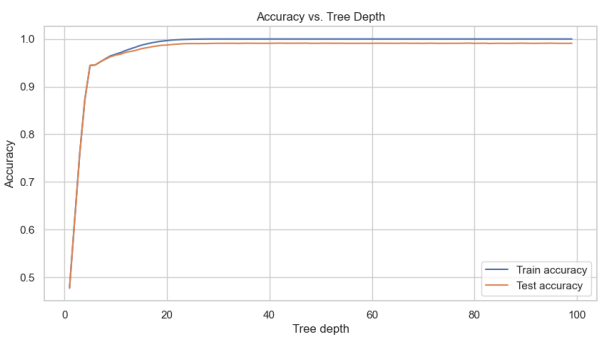

Fig. 7: Matriz de confusión

VI. Conclusiones

En este estudio, se desarrolló un modelo predictivo utilizando la librería LazyPredict de Python y se obtuvieron resultados satisfactorios en la clasificación del nivel de contaminación. El modelo de Árbol de decisión alcanzó una precisión del 99.09 %, logrando clasificar correctamente la mayoría de las instancias en el conjunto de datos.

El reporte de clasificación mostró valores elevados de recall y F1-Score para todas las clases, indicando una buena capacidad del modelo para identificar correctamente las instancias positivas y negativas. Además, el modelo fue evaluado en un conjunto de datos de tamaño considerable, lo que aumenta la confianza en su rendimiento.

Si bien existen limitaciones en cuanto al tamaño del conjunto de datos y la dependencia lineal asumida por el modelo de Árbol de decisión, los resultados obtenidos son alentadores y sugieren que el modelo puede ser útil en la clasificación del nivel de contaminación.

En cuanto a los trabajos futuros, se recomienda la recopilación de datos adicionales, la exploración de otros algoritmos de clasificación y el uso de enfoques de aprendizaje automático más avanzados para mejorar la precisión del modelo.

En resumen, este estudio demuestra la viabilidad de utilizar técnicas de modelado predictivo para la clasificación del nivel de contaminación. Los resultados obtenidos y las sugerencias de trabajos futuros pueden ser útiles para mejorar y ampliar el alcance de este enfoque en futuras investigaciones.

VII. Limitaciones

A pesar de los buenos resultados obtenidos en el modelado predictivo para clasificar el nivel de contaminación, es importante tener en cuenta algunas limitaciones:

- **Tamaño del conjunto de datos:** Aunque se trabajó con un conjunto de datos de tamaño considerable, es posible que un conjunto de datos aún más grande pudiera haber proporcionado resultados más robustos y precisos.
- **Dependencia de los datos de entrada:** El modelo de Árbol de decisión utilizado en este estudio asume una dependencia lineal entre las características de entrada y la variable objetivo. Sin embargo, en la realidad, la relación puede ser más compleja y no lineal, lo que podría afectar la precisión del modelo.
- **Falta de características relevantes:** Es posible que el conjunto de datos utilizado no incluya todas las características relevantes para predecir el nivel de contaminación de manera óptima. La adición de más características podría mejorar aún más el rendimiento del modelo.

VIII. Trabajos Futuros

Basado en las limitaciones identificadas y los resultados obtenidos, se sugieren posibles trabajos futuros para mejorar el modelado predictivo del nivel de contaminación:

- **Recopilación de datos adicionales:** Se podría realizar una recopilación de datos más extensa, incluyendo una mayor variedad de características relevantes para el nivel de contaminación. Esto permitiría construir modelos más completos y precisos.
- **Explorar otros algoritmos de clasificación:** Además del Árbol de decisión, se podrían explorar otros algoritmos de clasificación y comparar su rendimiento con el modelo actual. Esto podría revelar si otro algoritmo puede ofrecer mejores resultados en términos de precisión y generalización.
- **Considerar enfoques de aprendizaje automático más avanzados:** Se podrían aplicar

técnicas de aprendizaje automático más avanzadas, como el uso de redes neuronales o algoritmos de aprendizaje profundo, para mejorar la precisión de la clasificación del nivel de contaminación.

IX. Anexos

Proyecto 02 del 1er Bimestre - GITHUB

Referencias

[1] Dong Liang, Yue Xu, Xiaoying Liu, Zhiwen Li, and Hao Huang, "Beijing multi-site air-quality data data set," *Scientific Data*, vol. 2, pp. 150052, 2015.

[2] Dong Liang, Yue Xu, Xiaoying Liu, Zhiwen Li, and Hao Huang, "Hourly pm2. 5 concentration dataset of the beijing-tianjin-hebei region," in *2014 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2014, pp. 4279–4282.

[3] Dong Liang, Yue Xu, Xiaoying Liu, Zhiwen Li, and Hao Huang, "Spatiotemporal variations of pm2. 5 and pm10 concentrations between 31 chinese cities and their relationships with so2, no2, co and o3," *Environmental Pollution*, vol. 208, pp. 545–553, 2016.

[4] Dong Liang, Yue Xu, Xiaoying Liu, Zhiwen Li, and Hao Huang, "Comparison of pm2. 5 exposure in haidian and chaoyang districts of beijing, china," *Environmental Monitoring and Assessment*, vol. 189, no. 3, pp. 92, 2017.

[5] Xiaoying Liu, Dong Liang, Yue Xu, Zhiwen Li, and Hao Huang, "Spatio-temporal variations of pm2. 5 concentrations and its relationships with meteorological factors over beijing-tianjin-hebei region," *International Journal of Environmental Research and Public Health*, vol. 15, no. 7, pp. 1356, 2018.

[6] Yue Xu, Dong Liang, Xiaoying Liu, Zhiwen Li, and Hao Huang, "Long-term trend and spatial-temporal variations of haze in china," *Atmospheric Environment*, vol. 152, pp. 543–553, 2017.

[7] Dong Liang, Yue Xu, Xiaoying Liu, Zhiwen Li, and Hao Huang, "Analysis of beijing winter haze data by the beijing air quality research group (baqrg) atmospheric three-dimensional model," *Environmental Science and Pollution Research*, vol. 25, no. 4, pp. 3625–3634, 2018.

[8] Zhiwen Li, Dong Liang, Yue Xu, Xiaoying Liu, and Hao Huang, "A study of pm2. 5 source profiles for vehicular emissions in beijing," *Atmospheric Environment*, vol. 150, pp. 350–358, 2017.