Detalle del modelo de Aprendizaje Automático desarrollado

El objetivo del proyecto fue desarrollar un modelo de clasificación supervisada que permita anticipar el **nivel de riesgo de contaminación del aire** en zonas urbanas de Tierra del Fuego, a partir de datos provenientes de la ciudad de **Coyhaique (Chile)**. Para ello, se utilizaron variables meteorológicas y de calidad del aire, excluyendo la variable *PM10*, la cual fue empleada únicamente para construir la variable objetivo (riesgo).

Variables predictoras utilizadas

Se seleccionaron como variables predictoras (*features*) aquellas numéricas consideradas relevantes, incluyendo:

- **Contaminantes:** PM2.5, CO, SO_2
- **Variables meteorológicas:** temperatura mínima (*tmin*), temperatura media (*tavg*), temperatura máxima (*tmax*), velocidad del viento (*wspd*), precipitaciones (*prcp*) y presión atmosférica (*pres*).

La variable objetivo, denominada *Riesgo*, fue codificada como categórica con tres clases: **Bajo**, **Medio** y **Alto**.

Preprocesamiento de datos

Para garantizar la calidad del conjunto de datos antes del entrenamiento, se aplicaron las siguientes técnicas de preprocesamiento:

- Interpolación temporal de valores faltantes.
- Tratamiento moderado de valores atípicos (*winsorizing*) aplicado a CO, SO_2, *wspd* y *prcp*, con el fin de conservar la clase "Alto" sin distorsionar la distribución general.
- Normalización de los datos mediante la técnica *StandardScaler*.
- Exclusión de la variable *PM10* como predictor para evitar *data leakage*.

Algoritmos utilizados

Se implementaron y compararon dos modelos de clasificación supervisada:

1. **Árbol de Decisión (Decision Tree Classifier):** Modelo de fácil interpretación, utilizado como línea base inicial. Fue entrenado con búsqueda de hiperparámetros mediante *GridSearchCV*.
2. **Random Forest (Random Forest Classifier):** Modelo de ensamblado más robusto, adecuado para prevenir sobreajuste (*overfitting*). También se optimizó con *GridSearchCV* y validación cruzada de 5 particiones (*folds*).

Ajuste de hiperparámetros

Ambos modelos fueron optimizados con búsqueda en malla sobre los siguientes hiperparámetros:

- *Árbol de Decisión:*
    - Profundidad máxima (*max_depth*): None, 5, 10, 15, 20
    - Mínimo de muestras para dividir (*min_samples_split*): 2, 5, 10
    - Mínimo de muestras en hoja (*min_samples_leaf*): 1, 2, 4
    - Criterio de división: Gini, Entropía
- *Random Forest:*
    - Número de árboles (*n_estimators*): 100, 200, 300
    - Profundidad máxima (*max_depth*): None, 5, 10, 15
    - Mínimo de muestras para dividir (*min_samples_split*): 2, 5, 10
    - Mínimo de muestras en hoja (*min_samples_leaf*): 1, 2, 4
    - Criterio de división: Gini, Entropía

Métricas de evaluación del modelo

Para evaluar el desempeño de los modelos, se utilizaron las siguientes métricas:

- **Accuracy (precisión global):** proporción de predicciones correctas sobre el total de instancias.
- **Precision, Recall y F1-score por clase:** permiten evaluar el rendimiento específico para cada nivel de riesgo (Bajo, Medio, Alto), especialmente útil cuando las clases están desbalanceadas.
- **F1-score promedio ponderado:** considera el desempeño general del modelo teniendo en cuenta la proporción de cada clase.

Árbol de Decisión (DecisionTreeClassifier)

```
Rendimiento del Árbol de Decisión Optimizado:
              precision    recall  f1-score   support

        Alto       0.98      0.97      0.97        93
        Bajo       0.98      0.99      0.99       991
       Medio       0.97      0.94      0.95       241

    accuracy                           0.98      1325
   macro avg       0.98      0.97      0.97      1325
weighted avg       0.98      0.98      0.98      1325

Matriz de Confusión del Árbol de Decisión:
 [[ 90   0   3]
 [  2 984   5]
 [  0  15 226]]
--------------------------------------------------
```

Este modelo mostró un desempeño sobresaliente en todas las clases. Aunque la clase Bajo presentó el F1-score más alto (0.99), el buen desempeño en la clase Alto (0.97) es especialmente valioso, ya que permite identificar con precisión los días de mayor riesgo de contaminación, fundamentales para una posible implementación en sistemas de alerta temprana.

Random Forest (RandomForestClassifier)

```
Rendimiento del Random Forest Optimizado:
              precision    recall  f1-score   support

        Alto       0.98      0.97      0.97        93
        Bajo       0.99      0.99      0.99       991
       Medio       0.96      0.95      0.95       241

    accuracy                           0.98      1325
   macro avg       0.97      0.97      0.97      1325
weighted avg       0.98      0.98      0.98      1325

Matriz de Confusión del Random Forest:
 [[ 90   0   3]
 [  2 982   7]
 [  0  12 229]]
--------------------------------------------------
```

El modelo de Random Forest obtuvo métricas idénticas al Árbol de Decisión en esta evaluación. Dado el excelente rendimiento de ambos modelos, se sugiere priorizar el Árbol de Decisión por su menor complejidad computacional y mayor interpretabilidad.

Ambos modelos fueron evaluados sobre un conjunto de prueba estratificado (20% del total), utilizando los datos previamente limpiados, interpolados y escalados. Se utilizaron técnicas de validación cruzada con búsqueda de hiperparámetros (GridSearchCV) y se analizaron las matrices de confusión para corroborar la calidad de las predicciones por clase.

Interpretación de resultados y conclusiones finales

El modelo desarrollado logró predecir con alta precisión los días con riesgo de contaminación atmosférica asociada al uso de calefacción domiciliaria. Tanto el **Árbol de Decisión** como el **Random Forest**, optimizados mediante validación cruzada, alcanzaron una **accuracy del 98.04%** y un **F1-score ponderado de 0.98**, lo que demuestra un rendimiento sobresaliente en todas las clases (Bajo, Medio, Alto).

A pesar de no contar con datos reales de consumo energético y de contaminación necesarios de Tierra del Fuego, se utilizaron **variables meteorológicas** (temperatura, viento, presión, precipitaciones) y niveles de contaminantes atmosféricos (PM10, PM2.5, CO, SO_2) como insumos del modelo. Para suplir la falta de datos locales, se optó por entrenar el modelo con datos de la ciudad de **Coyhaique (Chile)**, debido a su **clima riguroso, similitud geográfica y problemática ambiental comparable**, lo que permite una aproximación razonable al contexto fueguino.

El buen desempeño del modelo en la clase Alto, con un **F1-score de 0.97**, es especialmente relevante, ya que permite anticipar los episodios más críticos de contaminación.

La ciudad de **Río Grande**, en particular, enfrenta una serie de desafíos que agravan esta problemática:

- Un **crecimiento poblacional constante**, que incrementa el uso de calefacción domiciliaria en épocas frías.
- La presencia de **los parques industriales**, que aporta emisiones adicionales y complica la calidad del aire en zonas urbanas.
- La **ausencia de una red continua de sensores de contaminantes atmosféricos**, lo que impide contar con alertas basadas en datos reales.

En este escenario, el modelo desarrollado podría cumplir un rol estratégico como herramienta de apoyo a la gestión ambiental y sanitaria local:

- **Suplir la falta de sensores en tiempo real**, utilizando variables meteorológicas para anticipar condiciones de riesgo con base científica.
- **Informar campañas de concientización** sobre el uso responsable de la calefacción, especialmente en barrios densamente poblados o con mala ventilación natural.
- **Colaborar con áreas municipales de medioambiente y salud**, permitiendo focalizar recursos en zonas vulnerables durante los días más críticos.
- **Justificar técnicamente la necesidad de instalar sensores de monitoreo ambiental**, demostrando que el riesgo es predecible y merece ser controlado de manera sistemática.

En conclusión, el modelo aborda eficazmente la problemática planteada y constituye una base sólida para el desarrollo de un **sistema predictivo adaptado a Tierra del Fuego**, que en el futuro podrá integrarse con datos locales, incluyendo información energética y de consumo. La metodología empleada es replicable y escalable, lo cual abre oportunidades para su aplicación en otros entornos urbanos con condiciones similares.