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

Ambos modelos fueron evaluados en un conjunto de prueba independiente, utilizando muestreo estratificado. Las métricas de desempeño se presentan en la sección siguiente.