Predicción de una Ola de Frío en Río Grande, Tierra del Fuego

Introducción

El informe documenta el desarrollo de un proyecto de Aprendizaje Automático cuyo objetivo es la predicción de una ola de frío en la ciudad de Río Grande, Tierra del Fuego del cual se utilizaron datos históricos meteorológicos y técnicas de clasificación supervisada.

Las olas de frío son un fenómeno climático extremo que puede tener un impacto significativo en la salud, la infraestructura y la actividad económica de la región, especialmente en zonas con condiciones meteorológicas severas como la provincia de Tierra del Fuego. Poder anticipar estos eventos puede ser clave para una mejor planificación y mitigación de riesgos.
El proyecto fue desarrollado en el marco de la materia Aprendizaje Automático, con carácter académico y está dirigido para todo público con interés en la ciencia de datos. A lo largo del informe se describen en detalle las etapas realizadas en el proyecto, desde la recolección y preparación de los datos hasta la selección, entrenamiento y evaluación de distintos modelos de clasificación, destacando el modelo final elegido y las métricas obtenidas en un conjunto de prueba independiente.

Objetivos

Objetivo General

El proyecto tuvo como objetivo desarrollar un modelo de Aprendizaje Automático capaz de de predecir si un día determinado será parte de una ola de frío en la ciudad de Río Grande, Tierra del Fuego, a partir de variables meteorológicas históricas, con el fin de anticipar eventos climáticos y contribuir a la toma de decisiones preventivas. El desarrollo de un modelo predictivo de este tipo requiere abordar distintos aspectos: desde la definición objetiva de lo que se considera una “ola de frío” hasta el tratamiento de datos desbalanceados, pasando por la selección adecuada de variables y modelos.

En este sentido, el trabajo busca responder preguntas clave como:

- ¿Cómo se puede identificar una ola de frío utilizando datos históricos de temperatura?
- ¿Qué variables meteorológicas son más relevantes para anticipar su ocurrencia?

Objetivos específicos

Se presentan algunos de los objetivos específicos que fueron parte del desarrollo del proyecto para la predicción de una ola de frío;

- Recolectar y unificar datos meteorológicos diarios de Río Grande correspondientes a los años 2021 a 2024.
- Realizar un análisis exploratorio y limpieza de datos, incluyendo el tratamiento de valores atípicos y datos faltantes.
- Definir y construir la variable objetivo (`ola_frio`) en base a criterios climáticos objetivos.
- Entrenar y evaluar diferentes modelos de clasificación, incluyendo Árbol de Decisión y Random Forest.

Descripción de los datos

Se utilizaron datos meteorológicos históricos diarios correspondientes a la ciudad de Río Grande, Tierra del Fuego, abarcando el periodo comprendido entre 1 de enero de 2021 hasta el 31 de diciembre de 2024. Los mismos fueron recolectados de la plataforma **Visual Crossing Weather Data**, lo que permitió acceder a registros diarios de temperatura, humedad, precipitación, viento, entre otras variables relevantes.
Luego de unificar los cuatro dataset, se conformó un dataset que contiene 1461 registros, uno por cada día del periodo mencionado.

Preprocesamiento de Datos (ETL)

En primer lugar se unificaron los archivos correspondientes a los años 2021, 2022, 2023 y 2024 en un único dataframe desde Google Collab. La unificación permitió trabajar con un conjunto consolidado y homogéneo de los datos.
Se realizó una limpieza de columnas que no iban a ser necesarias para el proyecto:
Se eliminaron aquellas columnas que contenían valores nulos en más del 50% de sus registros (preciptype y severerisk), o que no aportaban valor predictivo (como stations, description, icon, etc.).

Se borraron las columnas: preciptype y severerisk

```
columnas_a_eliminar = ['name','sunrise','sunset','conditions','description','icon','stations','preciptype','severerisk']
df.drop(columns=columnas_a_eliminar, inplace=True, errors='ignore')
```

Se renombraron las columnas al español para mejorar la legibilidad del dataset (por ejemplo: tempmin → temp_min, dew → punto_rocio, windspeed → viento_velocidad, entre otras). También se transformó la columna datetime a formato fecha y se generaron variables temporales derivadas: dia_del_anio, mes, dia_semana.

Cambio de nombres al español

```
df.rename(columns={
    'datetime': 'fecha',
    'tempmin': 'temp_min',
    'tempmax': 'temp_max',
    'temp': 'temp_promedio',
    'humidity': 'humedad',
    'windspeed': 'viento_velocidad',
    'windgust': 'rafaga_viento',
    'winddir': 'direccion_viento',
    'sealevelpressure': 'presion_nivel_mar',
    'cloudcover': 'cobertura_nubosa',
    'visibility': 'visibilidad',
    'dew': 'punto_rocio',
    'precip': 'precipitacion',
    'precipprob': 'prob_precipitacion',
    'precipcover': 'cobertura_precipitacion',
    'snow': 'nieve',
    'snowdepth': 'profundidad_nieve',
    'solarradiation': 'radiacion_solar',
    'solarenergy': 'energia_solar',
    'uvindex': 'indice_uv',
    'moonphase': 'fase_lunar'
}, inplace=True)
print(df.columns)
```

```
Index(['fecha', 'temp_max', 'temp_min', 'temp_promedio', 'feelslikemax',
       'feelslikemin', 'feelslike', 'punto_rocio', 'humedad', 'precipitacion',
       'prob_precipitacion', 'cobertura_precipitacion', 'nieve',
       'profundidad_nieve', 'rafaga_viento', 'viento_velocidad',
       'direccion_viento', 'presion_nivel_mar', 'cobertura_nubosa',
       'visibilidad', 'radiacion_solar', 'energia_solar', 'indice_uv',
       'fase_lunar'],
      dtype='object')
```

Luego, se realizó una detección de valores atípicos utilizando método estadístico, encontrando valores extremos principalmente en variables como: Precipitación, Profundidad de nieve, Dirección del Viento y Visibilidad. No obstante, se conservó estos valores atípicos ya que se consideró que son coherentes con el clima de la región y que podían contener información valiosa para el modelo, tratándose de eventos climáticos extremos.

Verificación de outliers.

```python
# Seleccionar solo columnas numéricas
columnas_numericas = df.select_dtypes(include=["float64", "int64"]).columns

# Calcular IQR para cada columna numérica
Q1 = df[columnas_numericas].quantile(0.25)
Q3 = df[columnas_numericas].quantile(0.75)
IQR = Q3 - Q1

# Filtrar valores atípicos (outliers)
outliers = df[((df[columnas_numericas] < (Q1 - 1.5 * IQR)) | (df[columnas_numericas] > (Q3 + 1.5 * IQR))).any(axis=1)]

print("Cantidad total de registros con outliers:", outliers.shape[0])
print("\nOutliers detectados en cada variable:")
for col in columnas_numericas:
    num_outliers = ((df[col] < (Q1[col] - 1.5 * IQR[col])) | (df[col] > (Q3[col] + 1.5 * IQR[col]))).sum()
    print(f"{col}: {num_outliers} outliers")
```

```
Cantidad total de registros con outliers: 614

Outliers detectados en cada variable:
temp_max: 0 outliers
temp_min: 10 outliers
temp_promedio: 3 outliers
feelslikemax: 0 outliers
feelslikemin: 27 outliers
feelslike: 0 outliers
punto_rocio: 15 outliers
humedad: 0 outliers
precipitacion: 328 outliers
prob_precipitacion: 0 outliers
cobertura_precipitacion: 53 outliers
nieve: 31 outliers
profundidad_nieve: 107 outliers
rafaga_viento: 6 outliers
viento_velocidad: 5 outliers
direccion_viento: 181 outliers
presion_nivel_mar: 8 outliers
cobertura_nubosa: 2 outliers
visibilidad: 220 outliers
radiacion_solar: 0 outliers
energia_solar: 0 outliers
indice_uv: 0 outliers
fase_lunar: 0 outliers
```

Desarrollo del Modelo

Fue necesario construir una variable objetivo llamada ola_frio, que indique con un **valor 1** si un determinado día pertenecía a una ola de frío y **0 en caso contrario**. Para eso se calculó el percentil 10 de la temperatura mínima de todo el periodo 2021-2024. Como no existía una columna explícita que identificara estos eventos en el conjunto de datos, se definió la ola en base a criterios estadísticos sobre la temperatura mínima diaria (temp_min). El proceso comienza calculando el percentil 10 de la temperatura mínima de todo el periodo de 2021-2024. Este valor estadístico representa el umbral por debajo del cual se encuentran los días más fríos del año (los 10% más fríos). Luego se creó una columna auxiliar llamada dia_frio que vale 1 si la temperatura mínima del día es menor a ese umbral, y 0 en caso contrario. Finalmente, se definió una ola de frío como un período de al menos tres días

consecutivos con **dia_frio = 1**. Para ello, se utilizó una ventana móvil de 3 días con suma acumulada: si la suma era mayor o igual a 3, se marcó como **ola_frio = 1.**

División de datos: El conjunto de datos fue divididos en tres subconjuntos: Entrenamiento, Validación y Prueba. Para el conjunto de entrenamiento se usó el 70%, para la validación 15% y para prueba 15% también.

Balanceo: Al analizar la variable objetivo se confirmó que el dataset estaba fuertemente desbalanceado, con sólo 3.4% de días marcados como ola de frío. Este desbalance podría afectar negativamente el aprendizaje del modelo teniendo a favorecer a la clase mayoritaria, es por esto que para mitigar este problema, se aplicó la técnica SMOTE (Synthetic Minority Oversampling Technique) al conjunto de entrenamiento. SMOTE genera ejemplos sintéticos nuevos de la clase minoritaria a partir de sus vecinos más cercanos, aumentando así su representación sin duplicar instancias reales. Esta técnica solo al ser aplicada al conjunto de entrenamiento asegura una evaluación realista del modelo sobre datos.

Entrenamiento de modelos: Luego de haber aplicado SMOTE, se entrenó y comparó dos modelos de clasificación: Árbol de Decisión y Random Forest. Ambos modelos fueron implementados con la biblioteca Scikit-learn, utilizando hiperparámetros estándar con leves ajustes. En particular, se estableció una profundidad máxima (max_depth) de 5 a 10 niveles, con el objetivo de evitar el sobreajuste y facilitar la interpretación del modelo.
El árbol de decisión fue entrenado con el conjunto balanceado y evaluado sobre el conjunto de validación, y fue elegido por su simplicidad, facilitando la visualización y capacidad para mostrar claramente las reglas de decisión del modelo.
El Random Forest se seleccionó como modelo más robusto al ser un conjunto de múltiples árboles de decisión entrenados con diferentes subconjuntos de datos y características.
Los modelos fueron evaluados en el conjunto de validación utilizando métricas específicas como; Recall, Precision, F1-score, Curva ROC y AUC.
Se eligió el modelo Random Forest como modelo final, debido a su mejor rendimiento en la detección de la clase minoritaria **ola_frio = 1.**

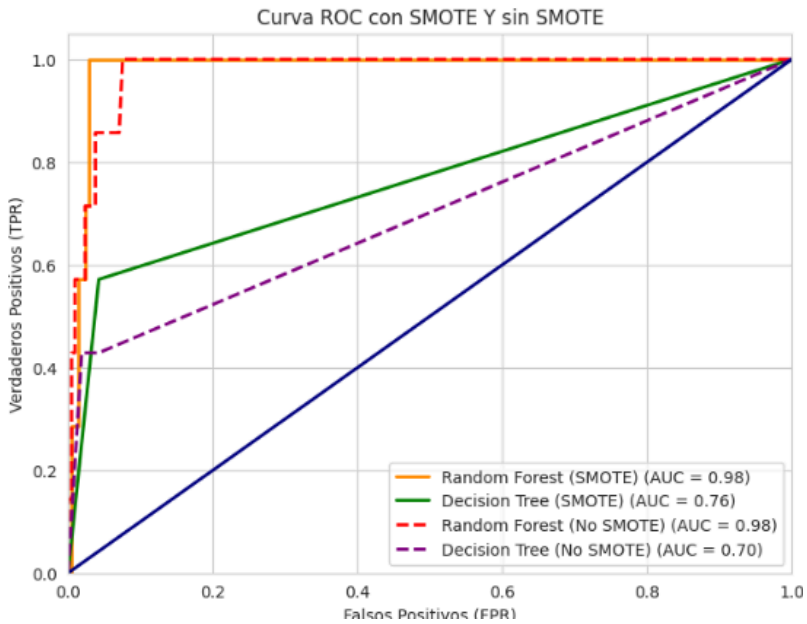
Evaluación: El Árbol de Decisión logró detectar 4 de los 7 días reales con ola de frío (recall = 57%) pero tuvo un bajo nivel de precisión (31%), cometiendo 9 falsos positivos. Aunque tuvo un buen rendimiento general, su capacidad para identificar correctamente las olas de frío fue limitada.
El modelo Random Forest mostró un desempeño superior, con mayor capacidad de detección de eventos (recall = 71%) y mejor equilibrio entre precisión y recuperación (F1-score = 59%). Cometió menos falsos positivos y logró un accuracy global del 97%, lo que confirma su solidez y efectividad.

Es por esto que se seleccionó al modelo Random Forest como el modelo final de este proyecto por ser el que mejor respondió al objetivo: detectar eventos raros de forma confiable.

Comparación de Modelos: Se compararon los desempeños del Árbol de Decisión y del Random Forest utilizando las métricas obtenidas en el conjunto de validación.

	Modelo	Accuracy	Precision	Recall	F1-Score
0	Árbol de Decisión	0.945	0.308	0.571	0.400
1	Random Forest	0.968	0.500	0.714	0.588

A nivel visual, también se compararon ambos modelos mediante la curva ROC, tanto con como sin la técnica de balanceo SMOTE. En todos los casos, el modelo Random Forest mostró un área bajo la curva (AUC) superior al del Árbol de Decisión, y mejor capacidad para distinguir entre días con y sin ola de frío.



Evaluación final en el conjunto de prueba: Una vez seleccionado el modelo Random Forest como el modelo final, se procedió a evaluar su desempeño sobre el conjunto de prueba. El conjunto no fue utilizado para el entrenamiento ni la validación.
El modelo logró identificar correctamente 6 de los 8 días con ola de frío presentes en el conjunto de prueba, con solo 2 falsos negativos y 4 falsos positivos. Estos resultados confirman que el modelo mantiene su capacidad predictiva incluso fuera del entorno de entrenamiento.

Conclusiones

En este proyecto se desarrolló un modelo de clasificación binaria capaz de predecir una ola de frío en Río Grande, Tierra del Fuego, con datos meteorológicos correspondientes al periodo 2021-2024. En él se entrenaron dos algoritmos, un Árbol de decisión y Random Forest del cual este último fue el que mejores resultados proporcionó.

Estos resultados permiten concluir que el modelo Random Forest **es efectivo para anticipar olas de frío en la región analizada**, representando una herramienta útil para la toma de decisiones en contextos de prevención climática.

Como recomendaciones para trabajos futuros sugiero:

- Incluir datos de más años o de otras variables meteorológicas regionales (presión en zonas vecinas, humedad del suelo, etc.)
- Explorar otros algoritmos más avanzados como XGBoost o LightGBM
- Ajustar el umbral de decisión según el tipo de aplicación (conservador o sensible)
- Complementar el modelo con sistemas de alerta temprana o mapas climáticos