Informe Final del Proyecto: Predicción de la Tasa de Ocupación Hotelera y de Plazas en Tierra del Fuego

Proyecto de Aprendizaje Automático

1. Introducción

El presente documento detalla el desarrollo y los resultados de un proyecto de **Aprendizaje Automático** enfocado en la **predicción de la Tasa de Ocupación Hotelera (TOH%)** y la **Tasa de Ocupación de Plazas (TOP%)** en las ciudades de Ushuaia y Río Grande, provincia de Tierra del Fuego, Antártida e Islas del Atlántico Sur.

A diferencia de un análisis estadístico descriptivo, cuyo fin es resumir y caracterizar datos históricos, este proyecto se centra en la **construcción de modelos predictivos**. Estos modelos tienen la capacidad de **aprender patrones complejos a partir de datos históricos** y, posteriormente, **generalizar ese conocimiento para realizar estimaciones precisas sobre datos futuros o no vistos**. El objetivo es proporcionar herramientas robustas para la anticipación de la demanda turística, la optimización de recursos y la toma de decisiones estratégicas en el sector hotelero.

2. Origen y Adquisición de los Datos

Los datos utilizados en este proyecto constituyen una compilación de diversas fuentes oficiales de la provincia de Tierra del Fuego, Argentina, garantizando su fiabilidad y relevancia para el ámbito turístico.

Fuentes de Datos (Instituto Provincial de Estadística y Censos - IPIEC):

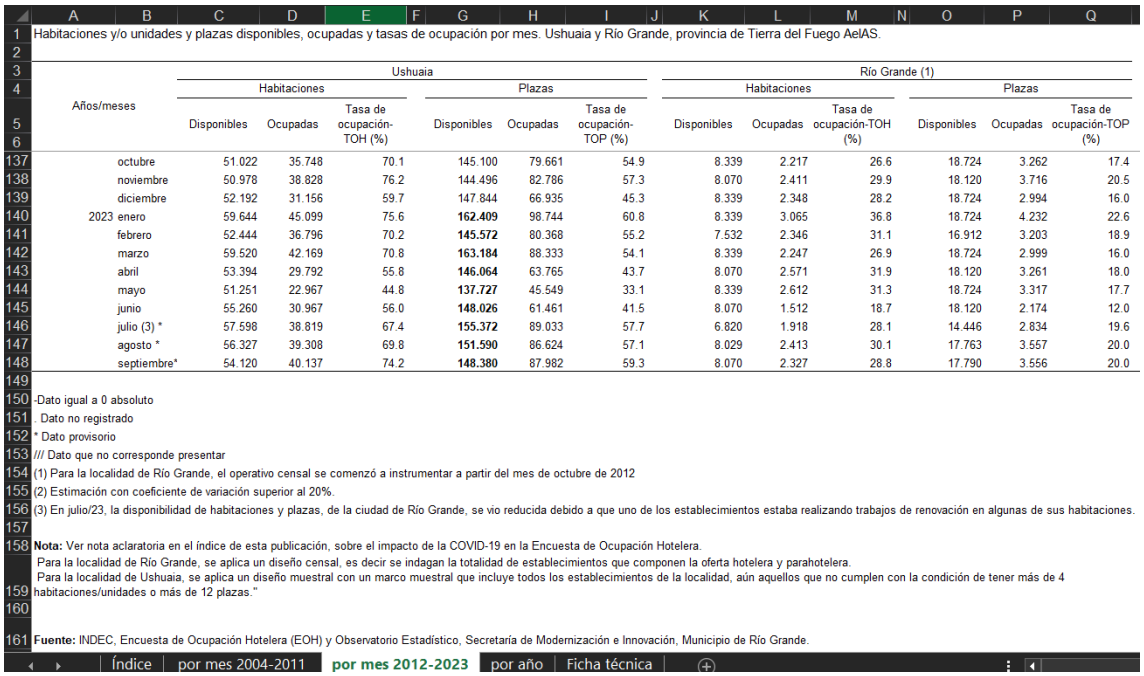
- **Meteorología:**
    - **Fuente:** IPIEC (https://ipiec.tierradelfuego.gob.ar/estadisticas-del-medio-ambiente/)
    - **Archivo Original:** 22_2_01_Meteorologia_Temperatura_Precipitaciones.xlsx
    - **Variables Extraídas (mensuales):**
        - temp_max: Temperatura máxima media (°C)
        - temp_min: Temperatura mínima media (°C)
        - temp_media: Temperatura media promedio (°C)
        - lluvia_mm: Precipitaciones en milímetros
        - dias_nieve: Días con nieve en el mes
- **Turismo** (Variables Objetivo):

- **Fuente:** IPIEC (https://ipiec.tierradelfuego.gob.ar/estadisticas-economicas-2/)
- **Archivo Original:** 16_3_01_Habitaciones_plazas_tasas_ocupacion-1.xlsx
- **Variables Extraídas (mensuales):**
  - toh (Tasa de Ocupación Hotelera %): Porcentaje de habitaciones ocupadas.
  - top (Tasa de Ocupación de Plazas %): Porcentaje de camas o plazas ocupadas.

- **Transporte Aéreo:**
  - **Fuente:** IPIEC (https://ipiec.tierradelfuego.gob.ar/estadisticas-economicas-2/)
  - **Archivo Original:** 14_5_03_Transporte_aereo._Movimiento_de_pasajeros_por_aeropuerto (1).xlsx
  - **Variables Extraídas (mensuales):**
    - aero_rg: Personas desembarcadas en el aeropuerto de Río Grande.
    - aero_ush: Personas desembarcadas en el aeropuerto de Ushuaia.
  - **Nota:** Ambas variables se incluyeron en los datasets finales de cada ciudad, reconociendo que los movimientos de pasajeros en un aeropuerto pueden influir en la ocupación hotelera de la otra ciudad debido a la proximidad geográfica.

- **Transporte Terrestre:**
  - **Fuente:** IPIEC (https://ipiec.tierradelfuego.gob.ar/estadisticas-economicas-2/)
  - **Archivo Original:** 14_5_04_Transporte_terrestre_Ingreso_Egreso_personas_por_San_Sebastian (1).xlsx
  - **Variable Extraída (mensual):**
    - ent_san_sebastian: Personas entradas por el paso fronterizo San Sebastián.

Proceso de Recopilación y Unificación: Los archivos originales, que contenían información de mes y año, fueron unificados manualmente en un archivo Excel (Columnas_unificadasUSHyRG.xlsx) debido a su formato no estándar.



Habitaciones y/o unidades y plazas disponibles, ocupadas y tasas de ocupación por mes. Ushuaia y Río Grande, provincia de Tierra del Fuego AeIAS.

Años/meses	Ushuaia						Río Grande (1)					
	Habitaciones			Plazas			Habitaciones			Plazas		
	Disponibles	Ocupadas	Tasa de ocupación-TOH (%)	Disponibles	Ocupadas	Tasa de ocupación-TOP (%)	Disponibles	Ocupadas	Tasa de ocupación-TOH (%)	Disponibles	Ocupadas	Tasa de ocupación-TOP (%)
octubre	51.022	35.748	70.1	145.100	79.661	54.9	8.339	2.217	26.6	18.724	3.262	17.4
noviembre	50.978	38.828	76.2	144.496	82.786	57.3	8.070	2.411	29.9	18.120	3.716	20.5
diciembre	52.192	31.156	59.7	147.844	66.935	45.3	8.339	2.348	28.2	18.724	2.994	16.0
2023 enero	59.644	45.099	75.6	**162.409**	98.744	60.8	8.339	3.065	36.8	18.724	4.232	22.6
febrero	52.444	36.796	70.2	**145.572**	80.368	55.2	7.532	2.346	31.1	16.912	3.203	18.9
marzo	59.520	42.169	70.8	**163.184**	88.333	54.1	8.339	2.247	26.9	18.724	2.999	16.0
abril	53.394	29.792	55.8	**146.064**	63.765	43.7	8.070	2.571	31.9	18.120	3.261	18.0
mayo	51.251	22.967	44.8	**137.727**	45.549	33.1	8.339	2.612	31.3	18.724	3.317	17.7
junio	55.260	30.967	56.0	**148.026**	61.461	41.5	8.070	1.512	18.7	18.120	2.174	12.0
julio (3) *	57.598	38.819	67.4	**155.372**	89.033	57.7	6.820	1.918	28.1	14.446	2.834	19.6
agosto *	56.327	39.308	69.8	**151.590**	86.624	57.1	8.029	2.413	30.1	17.763	3.557	20.0
septiembre*	54.120	40.137	74.2	**148.380**	87.982	59.3	8.070	2.327	28.8	17.790	3.556	20.0

-Dato igual a 0 absoluto
. Dato no registrado
* Dato provisorio
/// Dato que no corresponde presentar
(1) Para la localidad de Río Grande, el operativo censal se comenzó a instrumentar a partir del mes de octubre de 2012
(2) Estimación con coeficiente de variación superior al 20%.
(3) En julio/23, la disponibilidad de habitaciones y plazas, de la ciudad de Río Grande, se vio reducida debido a que uno de los establecimientos estaba realizando trabajos de renovación en algunas de sus habitaciones.

Nota: Ver nota aclaratoria en el índice de esta publicación, sobre el impacto de la COVID-19 en la Encuesta de Ocupación Hotelera.
Para la localidad de Río Grande, se aplica un diseño censal, es decir se indagan la totalidad de establecimientos que componen la oferta hotelera y parahotelera.
Para la localidad de Ushuaia, se aplica un diseño muestral con un marco muestral que incluye todos los establecimientos de la localidad, aún aquellos que no cumplen con la condición de tener más de 4 habitaciones/unidades o más de 12 plazas."

Fuente: INDEC, Encuesta de Ocupación Hotelera (EOH) y Observatorio Estadístico, Secretaría de Modernización e Innovación, Municipio de Río Grande.

Índice | por mes 2004-2011 | por mes 2012-2023 | por año | Ficha técnica

La falta de una estructura de fila y columna consistente, la presencia de metadatos y notas aclaratorias incrustadas directamente en las hojas, y los encabezados anidados que requerían interpretación manual, hicieron inviable cualquier método de extracción y unificación automatizado. El proceso manual fue esencial para consolidar la información en un formato utilizable para el análisis.

Posteriormente, este archivo se dividió en dos hojas (Columnas_unificadasUSH.csv y Columnas_unificadasRG.csv) para facilitar su procesamiento en herramientas de análisis y Machine Learning.

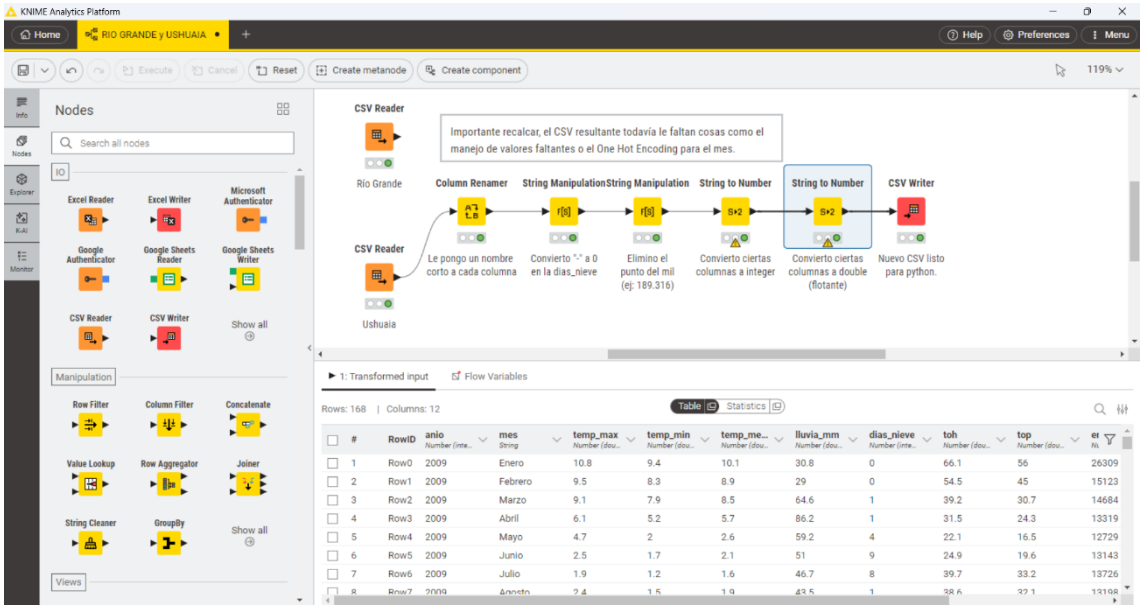
3. Preprocesamiento de Datos para Aprendizaje Automático

El preprocesamiento es una etapa crítica en el desarrollo de modelos de **Aprendizaje Automático**, asegurando que los datos sean de alta calidad, consistentes y estén en el formato adecuado para el entrenamiento de algoritmos.

a. Preprocesamiento Inicial con KNIME

Se utilizó KNIME Analytics Platform para una fase inicial de limpieza y preparación de los datos. Este software de código abierto permitió la automatización de tareas como:

- **Renombrado de Columnas:** Estandarización de los nombres de las características para facilitar su manejo y consistencia.
- **Conversión** de Tipos de **Datos:** Asegurar que cada columna tuviera el tipo de dato correcto (ej., numérico para temperaturas y conteos, categórico para meses).



b. Procesamiento Avanzado con Python (Google Colab)

La fase más intensiva de preprocesamiento, crucial para el modelado de **Aprendizaje Automático**, se llevó a cabo en un entorno de Python (Google Colab).

https://colab.research.google.com/drive/1oTecMKTpre7nw3Uil3ecUkPXoEfmeiwN?usp=sharing]

- **Imputación de Variables Ausentes:**
    - **Tasas de Ocupación (TOH% / TOP% - Años 2009-2010):** Inicialmente, estos valores estaban completamente ausentes. Para su imputación, se exploraron estrategias basadas en la mediana mensual (considerando la estacionalidad). **Posteriormente,** y como parte integral **del enfoque de Aprendizaje Automático, se utilizaron modelos XGBoost entrenados con datos posteriores (2011-2022) para predecir y rellenar estos valores faltantes.** Esta es una aplicación de ML para la imputación de datos, que busca generar estimaciones más precisas que métodos estadísticos simples.
    - **Entradas por San Sebastián (Río Grande):** Se observaron valores ausentes para Río Grande. Para su imputación, se empleó un modelo de **Random Forest**. Este es un algoritmo de **Aprendizaje Automático** robusto, capaz de manejar relaciones no lineales y predecir valores faltantes basándose en otras características del dataset, lo que subraya la aplicación de ML en la fase de preprocesamiento.
- **One-Hot Encoding para la Variable 'Mes':**
    - La característica 'Mes' (ej., "Enero", "Febrero"), inicialmente una cadena de texto, fue transformada utilizando **One-Hot Encoding**. Esta técnica

convierte una variable categórica en múltiples columnas binarias (0 o 1), una por cada categoría única. Esto es fundamental para que los algoritmos de **Aprendizaje Automático**, que operan principalmente con datos numéricos, puedan procesar esta información sin asumir una relación ordinal incorrecta entre los meses.

Tras estas etapas, los datasets de Río Grande y Ushuaia quedaron completos y en el formato adecuado, listos para el entrenamiento y evaluación de los modelos de **Aprendizaje Automático**.

Conclusiones Clave del Análisis Exploratorio:

- **Marcada Estacionalidad:** Se observaron patrones estacionales consistentes en las tasas de ocupación de ambas ciudades, con picos y valles anuales bien definidos, lo que resalta la importancia de las variables de mes.
- **Influencia del Transporte Aéreo:** La cantidad de pasajeros desembarcados en los aeropuertos demostró ser un factor altamente correlacionado con la ocupación hotelera, especialmente aero_ush para Ushuaia.
- **Tendencia Temporal en Río Grande:** El análisis visual sugirió una tendencia de crecimiento o cambio a largo plazo en la ocupación de Río Grande, lo que validó la inclusión de la variable anio.
- **Relación con el Turismo Terrestre:** La entrada por San Sebastián mostró ser un indicador relevante del flujo turístico terrestre.
- **Variabilidad Climática:** Las variables climáticas (temperatura, lluvia, nieve) mostraron patrones estacionales, pero su correlación directa con la ocupación fue menos pronunciada que la de los flujos de pasajeros y la estacionalidad pura.

Estas observaciones del EDA fueron fundamentales para guiar la selección inicial de características para nuestros modelos de **Aprendizaje Automático**.

5. Modelo de Aprendizaje Automático Desarrollado

Para la predicción de la ocupación hotelera, se desarrollaron **cuatro modelos de regresión independientes**, cada uno adaptado a las particularidades de la ciudad y la métrica objetivo:

- **Modelo 1:** Predicción de **TOH% en Ushuaia**.
- **Modelo 2:** Predicción de **TOP% en Ushuaia**.
- **Modelo 3:** Predicción de **TOH%** en Río **Grande**.
- **Modelo 4:** Predicción de **TOP% en Río Grande**.

Arquitectura y Algoritmos Utilizados

El algoritmo central para todos los modelos fue **XGBoost (Extreme Gradient Boosting)**.

- **XGBoost como Algoritmo de Aprendizaje Automático:**
    - **Enfoque Ensemble y Boosting:** XGBoost es un algoritmo de *ensemble learning* basado en *gradient boosting*. Construye secuencialmente una serie de árboles de decisión, donde cada nuevo árbol corrige los errores del árbol anterior. Esta naturaleza iterativa y combinatoria le confiere una **alta capacidad para aprender relaciones complejas y no lineales** en los datos, lo cual es una característica distintiva del **Aprendizaje Automático**.
    - **Rendimiento Superior:** Es ampliamente reconocido por su eficiencia computacional y su capacidad para lograr resultados de vanguardia en problemas de datos tabulares, lo que lo convierte en una elección robusta para la predicción.
    - **Manejo de Datos Diversos:** Se adapta bien a la combinación de características numéricas y categóricas (previamente transformadas a binarias mediante One-Hot Encoding).

Flujo de Desarrollo del Modelo (Metodología de Aprendizaje Automático)

El desarrollo de cada modelo siguió una metodología rigurosa de **Aprendizaje Automático**:

1. **División de Datos (Train-Test Split):**
    - Cada dataset (Ushuaia y Río Grande) se dividió en un **conjunto** de entrenamiento **(80%)** y un **conjunto de prueba (20%)**.
    - Esta división es fundamental en **Aprendizaje Automático** para **evaluar la capacidad de generalización del modelo**. El modelo solo "ve" los datos de entrenamiento durante su fase de aprendizaje, y su rendimiento se mide en el conjunto de prueba (datos no vistos), proporcionando una estimación imparcial de cómo se comportará en escenarios reales. Se utilizó random_state=42 para asegurar la reproducibilidad de esta división.
2. **Pre-procesamiento** de Características para **el Modelo:**
    - Las características seleccionadas (anio, temp_max, temp_min, temp_media, lluvia_mm, dias_nieve, ent_san_sebastian, aero_ush, aero_rg, y las variables mes_ generadas por One-Hot Encoding) fueron preparadas.

- Las columnas booleanas de los meses se convirtieron a enteros (0/1).
- Se eliminaron filas con valores nulos en las características o en las variables objetivo para garantizar la integridad del conjunto de entrenamiento.

3. **Optimización de Hiperparámetros (Tuning) con GridSearchCV:**
    - Para maximizar el rendimiento de cada modelo XGBoost, se realizó una exhaustiva **optimización de hiperparámetros** utilizando **GridSearchCV**. Esta técnica de **Aprendizaje Automático** explora sistemáticamente múltiples combinaciones de configuraciones de modelo.
    - **Hiperparámetros Ajustados:** Se optimizaron parámetros clave como n_estimators (número de árboles), learning_rate (tasa de aprendizaje), max_depth (profundidad máxima de cada árbol), subsample (fracción de muestras utilizadas), colsample_bytree (fracción de características utilizadas) y gamma (reducción mínima de pérdida para una división).
    - **Validación Cruzada (Cross-Validation):** GridSearchCV empleó una estrategia de validación cruzada de 5 pliegues (cv=5). Esto significa que el conjunto de entrenamiento se dividió en 5 subconjuntos, y el modelo se entrenó y validó 5 veces, cada vez utilizando un subconjunto diferente para validación. Esto asegura que los hiperparámetros seleccionados sean robustos y no estén sobreajustados a una partición específica de los datos.
    - **Métrica de Optimización:** La optimización se guio por la minimización del neg_root_mean_squared_error (la negación del RMSE), buscando el modelo que produjera el menor error de predicción.
4. **Selección de Características (Feature Selection Iterativa)::**
    - Tras el entrenamiento inicial y el análisis de importancia de características, se identificaron variables con una contribución muy baja al poder predictivo del modelo.
    - Para los modelos de Río Grande, se realizó una prueba específica eliminando la característica mes_julio. Esta decisión se basó en la importancia de las características del modelo base, donde mes_julio mostró una importancia muy baja para top y una moderada para toh. La re-optimización del modelo sin esta variable permitió evaluar si la simplificación del modelo podría mantener o incluso mejorar el rendimiento. Este proceso iterativo de **selección** de características es una práctica fundamental en Aprendizaje **Automático** para construir modelos más eficientes y robustos.

6. Métricas de Evaluación del Modelo

La evaluación de los modelos se realizó utilizando métricas de regresión estándar, calculadas sobre el **conjunto de prueba (datos no vistos)**. Esto garantiza una estimación imparcial de la capacidad predictiva de cada modelo.

- **RMSE** (Root Mean **Squared Error - Raíz del Error Cuadrático Medio):** Mide la magnitud promedio de los errores de predicción del modelo en las mismas unidades que la variable objetivo. Un RMSE más bajo indica un modelo más preciso.
- **MAPE** (Mean Absolute Percentage Error **- Error Porcentual Absoluto Medio):** Expresa el error promedio de las predicciones como un porcentaje del valor real. Es una métrica intuitiva para comprender la precisión relativa del modelo.
- **R2 Score (Coeficiente de Determinación):** Indica la proporción de la varianza en la variable objetivo que es predecible a partir de las características del modelo. Un valor más cercano a 1 (o 100%) significa que el modelo explica una mayor parte de la variabilidad del target.

Resultados Finales de los Modelos de Aprendizaje Automático (Tuneados y Optimizados):

Ciudad	Variable Objetivo	RMSE	MAPE (%)	R2 Score	Notas
Ushuaia	TOH%	06.09	12.30	0.92	
Ushuaia	TOP%	4.60	13.31	0.94	
Río Grande	TOH%	04.01	9.40	0.92	Mejorado tras eliminación de mes_julio
Río Grande	TOP%	3.18	11.76	0.88	Mejorado tras eliminación de mes_julio

7. Interpretación de Resultados y Conclusiones Finales

Los resultados obtenidos confirman que hemos desarrollado **modelos de Aprendizaje Automático de alta capacidad predictiva** para la ocupación hotelera en Tierra del Fuego.

- **Capacidad Predictiva Demostrada:** Los altos valores de R2 Score (entre 0.88 y 0.94) indican que nuestros modelos son capaces de explicar una proporción muy significativa de la varianza en las tasas de ocupación. Los bajos valores de

RMSE y MAPE (especialmente el MAPE por debajo del 10% para TOH% en Río Grande) demuestran que las predicciones son consistentemente cercanas a los valores reales. Esto es el resultado directo del **aprendizaje de patrones complejos** por parte de los algoritmos de XGBoost.

- **Factores Clave de Influencia (Insights de Machine Learning):**
    - **Impacto del Flujo Aéreo:** La cantidad de pasajeros en el **aeropuerto de Ushuaia (aero_ush)** es la característica más influyente para la ocupación hotelera de Ushuaia, subrayando la importancia de este indicador para la planificación turística.
    - **Tendencia Temporal en Río Grande:** La variable **anio** emerge como el predictor más potente en Río Grande, sugiriendo que factores de crecimiento a largo plazo o cambios estructurales en la ciudad son críticos para su ocupación hotelera.
    - **Estacionalidad Refinada:** Las variables One-Hot Encoding de los meses son fundamentales para capturar la estacionalidad de la demanda. Los modelos tuneados lograron explotar esta información de manera más efectiva que los modelos base.
    - **Optimización por Simplificación (Río Grande):** La prueba de eliminación de mes_julio en los modelos de Río Grande, seguida de un re-tuning, demostró que un modelo más simple puede ser igualmente (o incluso más) preciso. Esto valida la importancia de la **selección de características como una técnica de optimización en Aprendizaje Automático**, permitiendo construir modelos más eficientes sin sacrificar rendimiento.
- **Generalización y Robustez:** La metodología de división de datos en entrenamiento y prueba, junto con la validación cruzada durante la optimización de hiperparámetros, asegura que nuestros modelos no están sobreajustados a los datos históricos. Por lo tanto, tienen una **sólida capacidad para generalizar y realizar predicciones confiables sobre datos futuros o no vistos**, lo cual es la esencia del **Aprendizaje Automático**.

En conclusión, este proyecto va más allá del análisis estadístico descriptivo al **construir y validar modelos predictivos** que pueden ser utilizados activamente para la toma de decisiones. Hemos demostrado cómo los algoritmos de **Aprendizaje Automático** pueden **aprender de datos complejos** y **generar pronósticos precisos**, proporcionando una herramienta valiosa para la gestión estratégica del turismo en Tierra del Fuego.

8. Contenido del Repositorio GIT

https://github.com/SantiagoOroz/AA-SantiagoOroz-PredHOT

El repositorio GIT contiene todo lo generados durante el desarrollo de este proyecto de Aprendizaje Automático, garantizando la reproducibilidad y transparencia del trabajo.

- **Archivos e historial de modificaciones del Dataset:**
- **Notebooks de Python (.ipynb)**
- **Modelos Entrenados (.joblib):**
    - best_model_ush_toh.joblib (Modelo tuneado para TOH% en Ushuaia)
    - best_model_ush_top.joblib (Modelo tuneado para TOP% en Ushuaia)
    - best_model_rg_toh_no_julio.joblib (Modelo tuneado para TOH% en Río Grande, sin mes_julio)
    - best_model_rg_top_no_julio.joblib (Modelo tuneado para TOP% en Río Grande, sin mes_julio)
- **Video de Presentación:** https://drive.google.com/file/d/1oyp6yUB_TXYK7mxjjqIG3RwDrH2tRaWK/view?usp=sharing
- **Documentos:**
    - Descripción del Dataset y Origen para Predicción de Ocupación Hotelera en Tierra del Fuego.pdf
    - Informe_Final_Proyecto_ML.pdf (Este documento, una vez finalizado)