Descripción del Dataset y Origen para Predicción de Ocupación Hotelera en Tierra del Fuego

Santiago Oroz

Aprendizaje automático

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

Este documento detalla el dataset seleccionado y las fases de preprocesamiento aplicadas, fundamentales para el desarrollo de nuestro proyecto de **Aprendizaje Automático**. Se busca proporcionar una comprensión exhaustiva de los datos, su origen y las transformaciones realizadas para prepararlos para el modelado predictivo.

1. Descripción General del Dataset

El dataset principal utilizado en este proyecto es una compilación de diversas fuentes de datos provinciales de Tierra del Fuego, Argentina. Ha sido consolidado y preprocesado para servir como base para la predicción de variables clave en el ámbito turístico.

- **Número de Instancias (Filas):** 170
- **Número de Características (Columnas) Iniciales:** 12
- **Número de Características (Columnas) Finales (post-One-Hot Encoding):** 23
- **Tipos de Datos:** Incluyen valores numéricos (temperaturas, lluvias, personas, porcentajes) y categóricos (meses del año).

2. Origen y Adquisición del Dataset

Los datos provienen del **Instituto Provincial de Estadística y Censos (IPIEC)** de Tierra del Fuego, una fuente pública y oficial que garantiza la fiabilidad de la información.

Se extrajeron datos de las siguientes secciones y URLs:

- **Meteorología:**
    - **Fuente:** https://ipiec.tierradelfuego.gob.ar/estadisticas-del-medio-ambiente/
    - **Archivo Original:** 22_2_01_Meteorologia_Temperatura_Precipitaciones.xlsx
    - **Variables Extraídas (mensuales):**
        - Temperatura máxima promedio
        - Temperatura media promedio

        - Temperatura mínima promedio
        - Lluvia en mililitros
        - Días de nieve

- **Turismo:**
    - **Fuente:** 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 en los hoteles de una ciudad durante un mes específico. Por ejemplo, si un hotel tiene 100 habitaciones y 80 están ocupadas, el TOH% es 80%.
        - **TOP% (Tasa de Ocupación de Plazas):** Porcentaje de camas o plazas ocupadas en los hoteles durante un mes. Esta métrica considera la capacidad total de personas que pueden alojarse, no solo las habitaciones. Por ejemplo, si un hotel tiene 200 plazas y 150 están ocupadas, el TOP% es 75%.
- **Transporte Aéreo:**
    - **Fuente:** 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):**
        - **Personas desembarcadas en el aeropuerto de Río Grande (RG):** Número de pasajeros que desembarcaron en el aeropuerto de Río Grande.
        - **Personas desembarcadas en el aeropuerto de Ushuaia (USH):** Número de pasajeros que desembarcaron en el aeropuerto de Ushuaia.
        - *Nota:* Ambas variables se incluyen en el dataset final, ya que los movimientos de pasajeros en un aeropuerto pueden influir en la ocupación hotelera de la otra ciudad, dada la proximidad geográfica entre Río Grande y Ushuaia.
- **Transporte Terrestre:**

- **Fuente:** https://ipiec.tierradelfuego.gob.ar/estadisticas-economicas-2/
- **Archivo Original:** 14_5_04_Transporte_terrestre_Ingreso_Egreso_personas_por_San_Sebastian (1).xlsx
- **Variables Extraídas (mensuales):**
  - Personas entradas por San Sebastián

Todos los archivos contenían información de mes y año, lo que facilitó su integración temporal.

Proceso de Recopilación y Unificación:

La unificación de estos cuatro datasets originales en una única tabla se realizó manualmente en un archivo Excel, denominado Columnas_unificadasUSHyRG.xlsx. Este proceso fue necesario debido a que los datos no se ofrecían en un formato CSV estándar, sino como informes estructurados con títulos y aclaraciones, lo que impedía una concatenación directa.

Posteriormente, este archivo Excel se dividió en dos hojas, las cuales fueron guardadas individualmente como archivos CSV para facilitar su manejo en KNIME:

- Columnas_unificadasUSH.csv
- Columnas_unificadasRG.csv

El resultado de esta unificación fue una tabla inicial de 12 características para 170 instancias.

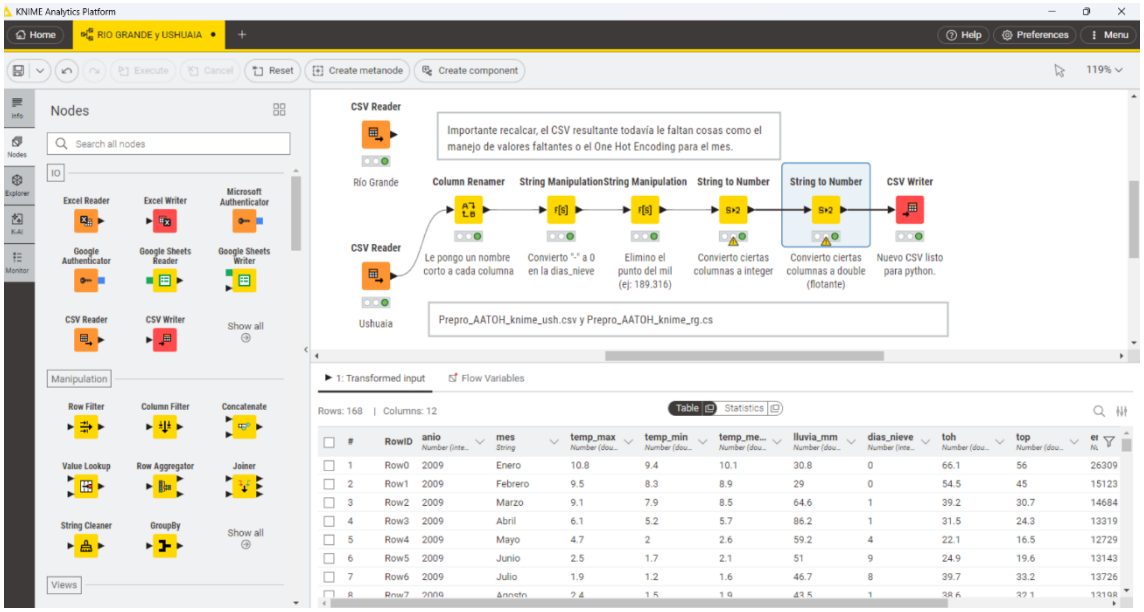
3. Preprocesamiento de Datos para Aprendizaje Automático

El preprocesamiento es una etapa crucial en el Aprendizaje Automático para asegurar la calidad y el formato adecuado de los datos para el entrenamiento de modelos.

a. Preprocesamiento Inicial con KNIME

Se utilizó KNIME para una fase inicial de limpieza y preparación de los datos. Las tareas principales incluyeron:

- **Renombrado de Columnas:** Se estandarizaron los nombres de las características para facilitar su manejo.
- **Conversión de Tipos de Datos:** Se aseguró que cada columna tuviera el tipo de dato correcto (ej., numérico para temperaturas y conteos, etc.).



b. Procesamiento Avanzado con Python (Google Colab)

La fase más intensiva de preprocesamiento se llevó a cabo en un entorno de Python (Google Colab), enfocándose en la imputación de valores ausentes y la codificación de variables categóricas.

Link al colab:

https://colab.research.google.com/drive/1oTecMKTpre7nw3UiI3ecUkPXoEfmeiwN?usp=sharing

Imputación de Variables:

La imputación es vital para manejar datos faltantes y evitar la pérdida de instancias valiosas, permitiendo que los algoritmos de Aprendizaje Automático trabajen con un conjunto de datos completo.

- **Tasas de Ocupación (TOH% / TOP%):**
    - **Problema:** Se identificó una ausencia completa de datos para los años 2009-2010 en estas variables.
    - **Solución:** Para imputar estos valores, se utilizó la **mediana mensual** calculada a partir del período 2011-2022.
    - **Justificación:** Esta estrategia se eligió debido a los fuertes patrones estacionales presentes en el turismo (ej., temporada alta en verano, baja en invierno), asegurando que los valores imputados reflejen el comportamiento típico de cada mes.
- **Entradas por San Sebastián:**

- **Problema:** Se observaron valores ausentes para Río Grande (RG), mientras que los datos para Ushuaia (USH) estaban presentes.
- **Solución:** Se empleó un modelo de **Random Forest** para imputar estos valores.
- **Justificación:** Random Forest es un algoritmo de Aprendizaje Automático robusto que puede manejar relaciones no lineales entre variables, lo que lo hace adecuado para predecir valores faltantes basándose en otras características del dataset.

One-Hot Encoding:

- **Variable 'Mes':** La característica 'Mes' (mes del año), que inicialmente era una cadena de texto (ej., "Enero", "Febrero"), fue transformada utilizando **One-Hot Encoding**.
- **Resultado:** 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 generalmente operan con datos numéricos, puedan procesar esta información sin asumir una relación ordinal incorrecta entre los meses.

4. Conclusión

Tras estas etapas de adquisición y preprocesamiento, los datasets de Río Grande y de Ushuaia está ahora completo y en el formato adecuado, con 23 características, listo para ser utilizado en el entrenamiento y evaluación de modelos de Aprendizaje Automático para el proyecto.