Tarea 2 – Introducción a Data Science
Regresión Lineal

Profesor: Brian Keith Norambuena, brian.keith@ucn.cl.
Ayudante: Matías Salas Villanueva, matias.salas@alumnos.ucn.cl.

Objetivos

1. Preparar el conjunto de datos para optimizar el ajuste aplicando Feature Engineering.
2. Ajustar diferentes modelos de ML con regresión lineal.
3. Reportar el comportamiento de cada modelo.

Entregables

- Un cuaderno interactivo *Jupyter Notebook* con todo el código fuente, resultados, y análisis.
- El cuaderno interactivo debe utilizar apropiadamente celdas de *markdown* y de código.

Fecha de Entrega

26 de septiembre a las 23.59 por Campus Virtual con el formato indicado en la sección de Consideraciones de Evaluación y Administrativas.

Requerimientos Mínimos

Se describen a continuación los requerimientos mínimos para esta tarea.

1. Preprocesado de Datos e Ingeniería de Características [2 puntos]:
   - **Conjunto Baseline:** Cargar el conjunto de datos inicial y realizar únicamente las modificaciones necesarias para que pueda ingresarse al modelo sin errores (por ejemplo, manejo de valores nulos, conversión de variables categóricas, ajuste de tipos de datos). **No se debe aplicar ingeniería de características ni estandarización en esta versión.** Puede generar nuevas columnas de ser necesario. **(+0.5 puntos)**
   - **Conjunto Limpio:** A partir del conjunto baseline, generar un nuevo conjunto de datos aplicando limpieza de columnas sin aporte de información y estandarización de las variables numéricas. Se debe argumentar el descarte de columnas en caso de eliminarlas. **(+0.75 puntos)**
   - **Conjunto con Interacción:** A partir del conjunto limpio, generar un tercer conjunto de datos en el cual se agregue una nueva columna de interacción entre $\boldsymbol{latitude \cdot longitude}$ (es decir, latitude · longitude). Se debe argumentar el descarte de las columnas en caso de eliminarlas. **(+0.75 puntos)**
2. Regresión Lineal [3 puntos]:
   - Dividir el conjunto de datos en datos de **entrenamiento** y **prueba** en un **80%** y **20%** respectivamente. **(+0.5 puntos)**
   - Ajustar 3 modelos de regresión lineal múltiple, uno con cada conjunto de datos. **(+0.5 c/u)**
   - Mostrar en una tabla comparativa los tres modelos y sus respectivos valores para el error cuadrático medio (MSE) y la proporción de la varianza explicada (R^2). **(+1 punto)**
     Para este punto puede utilizar la librería PrettyTable, puede ver su forma de uso en https://www.geeksforgeeks.org/python/creating-tables-with-prettytable-library-python/.

Descripción del conjunto de datos

Se debe hacer uso del conjunto de datos proporcionado housing_Tarea2.csv para la ejecución de esta tarea.

Este conjunto proviene de un scraping de https://www.makaan.com/ de abril de 2024, a continuación, una breve descripción de cada columna:

- **house_type:** Tipo de casa (apartamento, villa, dúplex).
- **house_size:** Tamaño de la casa en pies o metros cuadrados.
- **location:** Zona o barrio específico donde se encuentra la propiedad.
- **city:** Ciudad de la India donde se sitúa la propiedad.

- **latitude:** Coordenadas de latitud geográfica de la ubicación de la propiedad.
- **longitude:** Coordenadas de longitud geográfica de la ubicación de la propiedad.
- **price:** Precio de alquiler de la casa (variable objetivo).
- **currency:** Moneda en la que se denomina el precio (por ejemplo, INR - rupias indias).
- **numBathrooms:** Número total de baños en la casa.
- **numBalconies:** Número total de balcones en la casa.
- **isNegotiable:** Indica si el precio es negociable (Yes/No).
- **verificationDate:** Fecha en la que se verificó la información del alquiler (antigüedad de la publicación).
- **description:** Descripción o detalles adicionales sobre la propiedad.
- **SecurityDeposit:** Monto del depósito de garantía requerido para alquilar la propiedad.
- **Status:** Indica el estado de amueblado de la propiedad (amueblado, sin amueblar, semi-amueblado).

Consideraciones de Evaluación y Administrativas

- La entrega de esta tarea es de carácter individual y cualquier indicio de plagio será motivo de aplicación de nota mínima.
- El **formato de entrega** será el siguiente "NombreApellido1_NombreApellido2.ipynb" (sin tildes) mediante plataforma de campus virtual.
  Ejemplo: **PedroRojas_JosePerez.ipynb**
- A partir de las 00:00 hrs del día siguiente del día de entrega comenzará un descuento de 10 décimas **por cada hora** de retraso, Ejemplo: Si el límite de entrega es el 29 de agosto a las 23:59 hrs y la tarea es entregada el 30 de agosto a las 01:00 hrs, tiene derecho a nota máxima 6.0.
- La escala de evaluación será al 60%.
- Para recibir el puntaje completo debe ser asegurarse de que el cuaderno pueda ser ejecutado de corrido, de lo contrario se aplicará la **mitad del puntaje**.
- Los análisis de resultados y correspondientes conclusiones serán evaluados según la claridad de sus ideas y su presentación (escritura y formato).
- **Si utiliza Inteligencia Artificial generativa** (e.g., ChatGPT) durante su trabajo, por favor incluya todas sus *prompts* al final del cuaderno interactivo como documentación o entregue un anexo con esta información. No habrá penalización por el uso de estas herramientas como apoyo, pero no deberían ser utilizadas para resolver la tarea por ustedes.
- Los Créditos Extras serán puntos extras que se sumarán al final del semestre a tareas que no tengan el puntaje completo, comenzando por la tarea con menor puntaje. Los créditos extra sobrantes se anularán si el estudiante tiene puntaje máximo en todas las tareas.