# Trabajo Práctico 2 : Propiedades en Venta

## Introducción

En este trabajo práctico se propone que cada grupo de alumnos se enfrente a un problema real de ciencia de datos, que trabaje en cada una de las etapas del proceso y que pueda resolverlo aplicando los contenidos que vemos en la materia.
Vamos a utilizar el conjunto de datos provisto por la empresa Properati correspondiente a anuncios de propiedades en venta de la República Argentina publicados durante el año 2021 .
La información fue extraída desde BigQuery (producto de Google Cloud para consultar grandes volúmenes de datos) donde la empresa disponibiliza sus datasets con avisos de propiedades y desarrollos inmobiliarios que están y estuvieron publicados en Properati en todo Latinoamérica desde 2015 hasta dos meses atrás. Los datos se actualizan diariamente para mayor información pueden consultar el siguiente link.
El objetivo principal del trabajo es implementar nuevos modelos predictivos a partir de la experiencia realizada en el TP1.

## Modalidad de entrega

Notebook

El trabajo debe ser realizado en una notebook de python, se espera que la misma contenga **todos** los resultados de la ejecución los cuales siempre deben ser **reproducibles**. La notebook debe respetar la siguiente nomenclatura :

7506R_TP2_GRUPOXX_ENTREGA

En el caso que sea estrictamente necesario entregar más de una notebook las mismas deben contar con una numeración correlativa manteniendo un orden lógico entre ellas (7506R_TP2_GRUPOXX_ENTREGA_**N1**, 7506R_TP2_GRUPOXX_ENTREGA_**N2**, etc )
Las secciones del trabajo deben estar claramente diferenciadas en la notebook utilizando celdas de markdown. Se debe incluir una sección principal con el título del trabajo, el número de grupo y el nombre de todos los integrantes.
Todo análisis realizado debe estar acompañado de su respectiva explicación y toda decisión tomada debe estar debidamente justificada. Cualquier hipótesis que sea considerada en el desarrollo del trabajo práctico debe ser detallada y debe estar informada en la entrega. Cualquier criterio que se utilice basado en fuentes externas (papers, bibliografía, etc.) debe estar correctamente referenciado en el trabajo.

Visualizaciones

Todos los gráficos que se incorporen deben tener su correspondiente título, leyenda, nombres en los ejes, unidades de medidas, y cualquier referencia que se considere necesaria. Es importante que tengan presente que los gráficos son una herramienta que facilita entender el problema, por lo tanto, deben ser comprensibles por quien los vaya a leer.

Preprocesamiento:

A partir de las tareas de preprocesamiento, y de las diferentes estrategias que se planteen, es posible que se generen nuevos datasets sobre los cuales se entrenarán los modelos. Todo conjunto de datos creado debe ser almacenado y debe estar disponible en la entrega para ser utilizado por el equipo docente.

Modelos

Todos los modelos entrenados tanto para clasificación como para regresión deben ser guardados en un archivo (joblib / pickle) y deben estar disponibles en la entrega para ser utilizado por el equipo docente.

Repositorio

Cada grupo deberá crear su propio repositorio en github con la siguiente nomenclatura:

7506R-2C2022-GRUPOXX

En dicho repositorio deberá estar disponible la notebook, los modelos entrenados, los conjuntos de datos utilizados para el entrenamiento y cualquier archivo que sea necesario para la correcta ejecución del trabajo.

# Fechas de entrega

**Entrega: 08/12**

Para esta fecha se espera que todos los grupos hayan resuelto en su totalidad el trabajo práctico cumpliendo con todas las consignas y condiciones de entrega. **Esta fecha es obligatoria.**

# Enunciado

Los conjuntos de datos a utilizar **properati_argentina_2021** y **properati_argentina_2021_decrip** se encuentran disponibles en el siguiente enlace, la descripción de las variables se encuentra disponible aquí . Para este trabajo se plantean los siguientes objetivos generales:

*Procesamiento del Lenguaje Natural*: el objetivo será analizar las descripciones de los avisos para construir nuevos features y reentrenar algunos modelos del TP1.

*Redes Neuronales:* se deberán implementar dos modelos de red neuronal uno para clasificación y otro para regresión.

*Ensamble de Modelos:* el objetivo será construir dos ensambles híbridos uno para clasificación y otro para regresión pudiendo utilizar los modelos del TP1

A continuación se detallan las etapas que deben ser desarrolladas en el trabajo:

**1.Procesamiento del Lenguaje Natural**

**a) Ampliación del dataset**

Utilizar la columna **descripción** para encontrar aspectos de una propiedad. Luego utilizar estos aspectos para crear nuevas columnas, ampliando el *dataset* original.

Se proponen las siguientes técnicas para la detección de aspectos:

- La técnica de *Minqing Hu y Bing Liu* basada en frecuencia y palabras que indican “carga de valor” . Puede ser necesario realizar algunas modificaciones sobre la técnica.
- Regex con algún criterio de automatización teniendo en cuenta frases o valores repetidos. Obs: si se utilizó regex en el TP1 se debe proponer un criterio diferente al utilizado anteriormente.
- Métodos de extracción de conocimiento para la Web (*Open Information Extraction*): ExtrHech, ArgOE, DepOE, ECMes.

Según lo que observen en el texto, se obtendrán los valores posibles de los aspectos, por ejemplo, si encontraron “expensas” como un aspecto posible, seguramente al procesar el texto e intentar extraer los valores encontrarán: “bajas”,”baratas”, etc.

**b) Modelos**

Entrenar un modelo XGBoost para regresión con el nuevo dataset ampliado considerando los siguientes escenarios

- Utilizar los mismos hiperparámetros seleccionados en el TP 1
- Utilizar hiperparametros optimizados con el nuevo dataset ampliado.

Mostrar los resultados obtenidos, según las mismas métricas escogidas en el TP1.
Comparar el desempeño de los modelos, evaluando si mejoró la performance al agregar las columnas nuevas.

**2. Redes Neuronales**

Construir dos modelos de redes neuronales, uno para regresión y otro para clasificación, considerando los datasets creados en el TP1. Mejorar estos modelos de redes neuronales a través de la búsqueda de arquitectura e hiperparámetros adecuados. Se pide:

- Regresión: predecir el precio de la propiedad y utilizar como métrica de evaluación el error cuadrático medio
- Clasificación: predecir el atributo tipo_precio creado en el TP 1 y utilizar como métricas precisión, *recall* y F1-Score

**3. Ensamble de modelos**

Construir dos ensambles de modelos de tipo híbridos :

- Ensamble 1: ensamble tipo Voting para el conjunto de datos de clasificación. Obtener sus métricas y comparar los resultados con los obtenidos en los puntos anteriores.

- Ensamble 2: ensamble de tipo Stacking combinando diferentes modelos de regresión y utilizando un modelo adicional para estimar el valor final. Mostrar las métricas obtenidas y comparar con los resultados obtenidos en los puntos anteriores.

**4. Conclusiones**

Realizar las conclusiones correspondientes al trabajo realizado en su totalidad, destacando principalmente los aspectos que consideren más relevantes. Comentar brevemente qué otras opciones hubiesen explorado y quedaron fuera del alcance de este trabajo.

---

**Nota:** para los problemas de clasificación deberán utilizar el mismo target que en el TP1 . Les recordamos cómo se solicitó construir la variable **tipo_precio:**

Construcción del target

Para esta tarea se debe crear una nueva variable ***tipo_precio*** que tendrá tres categorías: **alto, medio, bajo.** Esta nueva variable será nuestra clase en el problema de clasificación.
Para determinar cuándo el tipo_precio de una propiedad es alto, medio o bajo se deberá analizar el precio por metro cuadrado (pxm2). Se propone evaluar las siguientes alternativas para establecer los límites de cada categoría:

1. Dividir la variable *pxm2* en 3 intervalos con igual cantidad de observaciones.
2. Dividir la variable pxm2 en 3 intervalos, el primero con el 25% de las observaciones, el siguiente con el 50% y el último con el 25% de las observaciones restantes.
3. Trabajar la variable pxm2 relativa a cada tipo de propiedad y luego dividirla como en el punto anterior.

Se pide: Seleccionar una de las alternativas, justificando la misma.

---