# Proyecto 2. Entrega 2
# Árboles de Decisión

## INTRODUCCIÓN:

InmoValor S.A. es una empresa innovadora del sector inmobiliario que ha apostado por la transformación digital para ofrecer valoraciones precisas y objetivas de propiedades. Ante un mercado dinámico y competitivo, la compañía ha adoptado técnicas avanzadas de análisis y modelos de regresión para estimar el valor de inmuebles basándose en un amplio conjunto de datos que recopila información detallada de viviendas. Este dataset incluye variables clave como ubicación, tamaño, calidad constructiva y otros factores determinantes, lo que permite desarrollar modelos predictivos capaces de reflejar con mayor exactitud las condiciones del mercado.

La empresa ha decidido incorporar un equipo de analistas de datos con la finalidad de trabajar con el conjunto de datos "House Prices: Advanced Regression Techniques" para desarrollar modelos predictivos que proyecten de manera precisa el precio de las viviendas. Mediante el análisis de variables clave como la ubicación, el tamaño y la calidad de las propiedades, el equipo utilizará técnicas avanzadas de regresión para mejorar la estimación de valores inmobiliarios y facilitar la toma de decisiones estratégicas en el mercado de bienes raíces.

**Vínculo del conjunto de datos a utilizar**

https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data

**Descripción de la Consultoría:**

La consultoría se dividirá en varias etapas. Cada semana deberá entregar resultados de la aplicación de un algoritmo de predicción y/o clasificación. La conclusión será la elección del mejor algoritmo predictivo para estimar el valor de una vivienda. Le han pedido que cree una variable que agrupe los precios de las casas en 3 categorías: Económicas, Intermedias o Caras, esto teniendo siempre en cuenta, el rango de precios.

**Resultados esperados en la Segunda Entrega Consultoría:**

Se espera la entrega de un informe detallado donde incluya:

- La descripción de la nueva variable categórica, explicando los límites seleccionados para cada categoría y la justificación de la elección de los valores mínimo y máximo para cada una de las categorías.
- Los modelos de árboles de decisión tanto para predicción del precio de las casas como de clasificación usando la variable creada.
- Una comparación para estimar los precios de las viviendas donde determine qué algoritmo funcionó mejor (Regresión Lineal, Árboles de regresión, Random Forest). Compare los resultados del mejor modelo de cada uno de los algoritmos.

**Notas:**

- La consultoría es en grupo, por lo que solo se tendrán en cuenta los grupos conformados por más de un especialista.
- Cada individuo será evaluado de forma individual basado en sus aportes al trabajo grupal, por lo que deben versionar el código para poder revisar las contribuciones de cada uno.

## INSTRUCCIONES

Utilice el data set House Prices: Advanced Regression Techniques que comparte Kaggle. Utilice el análisis exploratorio que hizo en la entrega anterior. Si considera que le faltó algo por explorar y cree que lo necesita, hágalo. Genere un informe con las explicaciones de los pasos que llevó a cabo y los resultados obtenidos. Recuerde que la investigación debe ser reproducible por lo que debe guardar el código que ha utilizado para resolver los ejercicios.

## ACTIVIDADES

1. Use los mismos conjuntos de entrenamiento y prueba que usó para los modelos de regresión lineal en la entrega anterior.
2. Elabore un árbol de regresión para predecir el precio de las casas usando todas las variables.
3. Úselo para predecir y analice el resultado. ¿Qué tal lo hizo?
4. Haga, al menos, 3 modelos más, cambiando el parámetro de la profundidad del árbol. ¿Cuál es el mejor modelo para predecir el precio de las casas?
5. Compare los resultados con el modelo de regresión lineal de la hoja anterior, ¿cuál lo hizo mejor?
6. Dependiendo del análisis exploratorio elaborado cree una variable respuesta que le permita clasificar las casas en Económicas, Intermedias o Caras. Los límites de estas clases deben tener un fundamento en la distribución de los datos de precios, y estar bien explicados
7. Elabore un árbol de clasificación utilizando la variable respuesta que creó en el punto anterior. Explique los resultados a los que llega. Muestre el modelo gráficamente. Recuerde que la nueva variable respuesta es categórica, pero se generó a partir de los precios de las casas, no incluya el precio de venta para entrenar el modelo.
8. Utilice el modelo con el conjunto de prueba y determine la eficiencia del algoritmo para clasificar.
9. Haga un análisis de la eficiencia del algoritmo usando una matriz de confusión para el árbol de clasificación. Tenga en cuenta la efectividad, donde el algoritmo se equivocó más, donde se equivocó menos y la importancia que tienen los errores.
10. Entrene un modelo usando validación cruzada, prediga con él. ¿le fue mejor que al modelo anterior?
11. Haga al menos, 3 modelos más, cambiando la profundidad del árbol. ¿Cuál funcionó mejor?
12. Repita los análisis usando random forest como algoritmo de predicción, explique sus resultados comparando ambos algoritmos.

## EVALUACIÓN

**Nota**: Tiene que poderse comprobar su aporte al trabajo grupal a través de commits. Si no existen al menos 3 commits con su aporte significativo no va a tener nota de la hoja de trabajo. Utilice una herramienta que permita registrar los aportes de cada uno.

- **(10 puntos)** Creación de la variable respuesta para árbol de clasificación. Explicación de los límites de las categorías.
- **(8 puntos)** Generación del modelo de regresión. Análisis de los resultados obtenidos
- **(8 puntos)** Comparación del árbol de regresión con el modelo de regresión lineal de la hoja anterior. Explicación de resultados
- **(8 puntos)** Tuneo del parámetro de la profundidad del árbol. Selección del mejor modelo, todo está explicado claramente.
- **(12 puntos)** Árbol de Clasificación. Representación gráfica del modelo.
- **(12 puntos)** Modelo con validación Cruzada y tuneo de la profundidad del árbol de clasificación.
- **(12 puntos)** Random Forest
- **(30 puntos)** Análisis de resultados de aplicación del algoritmo para predecir o clasificar sobre el conjunto de prueba. Comparación entre algoritmos.

## MATERIAL A ENTREGAR

- Archivo .rmd, .ipynb o Google docs con el informe con lo solicitado en las instrucciones (En caso de que utilicen jupyter notebooks, pueden omitir esta entrega e irlo documentando todo dentro del mismo notebook. Siempre es importante que pueda mostrar evidencias de la evolución de sus avances).
- Script de R o de Python que utilizó debidamente organizado y comentado (Si utilizó rmd debe añadir el html que se genera)
- Link de controlador de versiones utilizado.

## FECHAS DE ENTREGA

- **AVANCE:** Puntos del 1 al 7 de la sección de actividades: viernes 7 de marzo a las 23:59.
- **ENTREGA FINAL:** domingo 9 de marzo a las 23:59

**NOTA:** Solo se calificará el Documento Final si está entregado el avance con todo lo que se pide.