Clase de Talle de Modelado I MCMAI

Tarea 1.

Esta es la primera entrega de 3 tareas que van encaminadas en generar un modelo predictivo para predecir el precio final de cada casa (variable SalePrice) de un conjunto de datos con 79 variables explicativas que describen varios aspectos de las casas residenciales en Ames, Iowa.

Esta serie de tareas debe ser en equipo, de 3 personas. El modelo final se tiene que ajustar o entrenar en los datos Casas.csv (con aproximadamente 1400 registros) y probar su poder predictivo en los datos Casas_Kaggle.csv (con aproximadamente 1400 registros). La descripción de las variables está en el archivo data_description.txt.

El equipo que al final genere el modelo con el mejor poder predictivo tendrá 10% extra en las tareas. Para evaluar los resultados de la predicción, tienen que entrar a la página de Kaggle (previamente registrarse e iniciar una sesión). En la tarea 3 se explicará cómo se somete la predicción.

La primera tarea consiste en aplicar algunos puntos que hemos visto sobre preprocesamiento, en particular: **importación, exploración inicial, ingeniería de variables, tratamiento de datos faltantes de los datos y análisis de la distribución de la respuesta y relación de la respuesta con las preditoras**.

Estos procedimientos se van a aplicar solo a un subconjunto de variables explicativas. El subconjunto de variables predictoras (es decir, descarten las demás para los análisis y generar los modelos) serán:

- MSZoning
- LotArea
- Street
- Neighborhood
- YearBuilt
- OverallCond
- ExterQual
- GrLivArea
- FullBath
- GarageArea
- BsmtCond
- FireplaceQu
- Electrical
- LotFrontage
- KitchenQual
- PavedDrive

La tarea se entrega el próximo lunes 9 pm (se envía por correo). Por favor pongan explicaciones de todo lo que hacen, no solamente pongan puro código, de lo contrario no serán acreedores a toda la

calificación de esta tarea. La tarea se puede hacer directamente en un archivo de google colab, con las explicaciones respectivas en celdas de texto. El colab se puede pasar a pdf.

Observaciones de la tarea 1

1) Si lo hacen en colab (lo van a pasar a pdf) o en un editor de texto, hagan secciones y subsecciones de lo que van realizado, para que se vea y lea con mejor claridad lo que van haciendo. Tiene que darle un mejor orden visual a todo.

2) Si lo hacen en colab, **usen las celdas de texto para escribir la descripción de lo que hacen**, **no** en código comentado.

3) Sugerencia: Instalen los paquetes de R en una carpeta aparte y mándenla a llamar, así como lo mostramos en el texto.

4) No es momento de eliminar **variables de varianza casi cero** (eso se verá después). Solo las no informativas como llaves.

5) Para analizar la variable de respuesta SalePrice (continua) vs otras variables hagan lo siguiente:

- Vs variables continuas (correlación y generen la gráfica de dispersión)
- Vs. Variables discretas que pueden representar conteo (correlación y generen la gráfica de dispersión)
- Vs variables ordinales o nominales. (Gráfica de cajas)

6) Todos los cambios o transformaciones que hagan en la base de Casas.csv se repiten en los datos de Casas_Kaggle.csv.

7) En el archivo txt (descripción de la base) se indican códigos para los valores faltantes de algunas variables. Para algunas variables los valores faltantes denotan ausencia de una cualidad, por lo que se puede remplazar el dato AUSENTE por una cadena de caracteres que indica la ausencia de esa calidad (puede ser 'None'). Es un **tipo de imputación de forma lógica**.

8) **Imputación por relación evidente con otras variables**
Si hay alguna variable con valor faltante que no se pueda imputar de forma lógica, usa el método de missforest.