Actividad: Árboles y *random forest* para regresión y clasificación

Objetivos

Mediante este laboratorio se pretende que apliques los conocimientos adquiridos en sobre árboles y *random forest* para un problema de regresión y uno de clasificación: predecir el precio de venta y el rango de precio del inmueble.

Pautas de elaboración

El conjunto de datos con el cual vamos a trabajar se encuentra en el siguiente link:

<https://www.kaggle.com/gpandi007/usa-housing-dataset>

housing\_train.csv

Realiza los siguientes pasos:

* Análisis descriptivo de los datos:
  + De las variables numéricas, halla datos estadísticos.
  + De las variables categóricas, lista las diferentes categorías y halla la frecuencia de cada una de ellas.
  + Crea matriz de correlaciones existentes entre las variables numéricas del conjunto de datos y analiza los resultados.
* Tratamiento de missing. Si existen valores faltantes, decide si eliminar los registros o llenarlos con valores como la media, la mediana o la moda, y justifica tu respuesta.
* Aplica árboles y *random forest* al problema de regresión.
* Compara, mediante las medidas que te parezcan adecuadas, la capacidad predictiva de ambos métodos.
* Comenta las ventajas y desventajas de cada modelo. De acuerdo con los resultados, ¿son realmente útiles los modelos creados para el conjunto de datos propuesto?
* Para el ejercicio de clasificación, tanto para árboles como para *random forest*, se crean los siguientes grupos: grupo 1 (SalePrice menor o igual a 100 000), grupo 2 (SalePrice entre 100 001 y 500 000) y grupo 3 SalePrice (mayor o igual a 500 001).
* Compara los resultados de los dos clasificadores y comenta los resultados.
* Otros comentarios que consideres adecuados.

Extensión y formato

Extensión máxima de la actividad: 15 páginas. Formato: Calibri 12, interlineado 1,5 puntos.

Rúbrica

|  |  |  |  |
| --- | --- | --- | --- |
|  | Descripción | Puntuación máxima  (puntos) | Peso  % |
| Criterio 1 | Realiza el análisis descriptivo de los datos de manera adecuada | 2 | 20 % |
| Criterio 2 | El tratamiento de los valores faltantes es correcto | 2 | 20 % |
| Criterio 3 | Analiza muy bien los resultados de aplicar los clasificadores al problema de regresión | 2 | 20 % |
| Criterio 4 | Analiza muy bien los resultados de aplicar los clasificadores al problema de clasificación | 2 | 20 % |
| Criterio 5 | Los comentarios sobre los resultados son pertinentes y adecuados | 2 | 20 % |
|  |  | **10** | **100 %** |