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

**Objetivos**

Mediante este laboratorio se pretende que apliques al conjunto de datos proporcionado los métodos o algoritmos de Árboles y Random Forest para predecir el precio de venta de un inmueble.

**Descripción de la actividad**

Importe los datos del USA Housing Dataset:

<https://www.kaggle.com/gpandi007/usa-housing-dataset>  
housing\_train.csv contiene los datos. La variable respuesta es «SalePrice»

Para el ejercicio de regresión, aplica al conjunto de datos los métodos de Árboles y Random Forest para predecir el precio de venta de un inmueble.

Para el ejercicio de clasificación tanto para Árboles como para Random Forest crea los siguientes grupos: grupo1 SalePrice menor o igual a 100 000, grupo2 SalePrice entre 101 000 y 500 000 y grupo3 SalePrice mayor o igual a 501 000.

Realiza los siguientes pasos:

* Análisis descriptivo de los datos:
  + De las variables numéricas hallar el valor mínimo, el máximo, la mediana y la media.
  + De las variables categóricas, listar las diferentes categorías y hallar la frecuencia de cada una de ellas.
  + Hallar todas las correlaciones existentes entre las variables numéricas del conjunto de datos.
* Determinar el conjunto de modelización y el de validación.
* Tratamiento de *missing.* Si existen valores faltantes, decidir si eliminar los registros, llenarlos con valores como la media, la mediana o la moda y justifique su respuesta.
* Calcular las métricas de evaluación de ajuste adecuadas:
  + Para el ejercicio de regresión hallar valores como el error cuadrático medio o su raíz cuadrada.
  + Para la clasificación realizar una validación cruzada con mínimo 5 folds y calcular la matriz de confusión.
* Comparar mediante las medidas que parezcan adecuadas la capacidad predictiva de ambos métodos. Es decir, comparar el error cuadrático medio de los dos modelos y concluir cuál es mejor. Seguir el mismo procedimiento con la matriz de confusión de los dos modelos para el ejercicio de clasificación.
* Comente las ventajas y desventajas de cada modelo. De acuerdo con los resultados, son realmente útiles los modelos creados para el conjunto de datos propuesto o es mejor investigar otros algoritmos.
* Otros comentarios que parezcan adecuados.
* Se puede usar R o Python.
* Se deben comentar los resultados obtenidos y el código.

**Rúbrica**

|  |  |  |  |
| --- | --- | --- | --- |
| Actividad | Descripción | Puntuación máxima  (puntos) | Peso  % |
| Criterio 1 | Metodología | 6 | 60% |
| Criterio 2 | Resultados | 2 | 20% |
| Criterio 3 | Informe | 2 | 20% |
|  |  | **10** | **100 %** |

**Extensión** 5 páginas en formato Word o PDF con el código adjunto aparte.

**Cuantas Categorias Hay**

**Covariancia dependencia**

**Eliminar variables->Columnas**

**Valores Nulos eliminar o media o moda consideradondo el mejor**

**Aplico arboles de descion para el PISO**

**Hacer pruebas de Regresion**

**Trabajar en etiquetas**

**ERROR**

**https://jakevdp.github.io/PythonDataScienceHandbook/05.03-hyperparameters-and-model-validation.html**

**https://towardsdatascience.com/decision-tree-algorithm-explained-83beb6e78ef4**

1. If the relationship between dependent & independent variables is well approximated by a linear model, linear regression will outperform the tree-based model.
2. If there is a high non-linearity & complex relationship between dependent & independent variables, a tree model will outperform a classical regression method.
3. If you need to build a model that is easy to explain to people, a decision tree model will always do better than a linear model. Decision tree models are even simpler to interpret than linear regression!