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

<https://es.wikipedia.org/wiki/Validaci%C3%B3n_cruzada>

**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!

## <https://www.datacamp.com/community/tutorials/random-forests-classifier-python>

## Random Forests vs Decision Trees

* Random forest es un conjunto de múltiples árboles.
* Random fotes previene el overfitting al crear sub-conjuntos aleatorios.
* El ratio aproximado de éxito es mayor.
* Fácil de entrenar
* Arboles de decisión:
* Mayor tiempo de ejecución.
* Es difícil de interpretar
* Los modelos pueden ser muy complejos
* Deep decision trees may suffer from overfitting, but random forests prevents overfitting by creating trees on random subsets.
* Pueden sufrir de overfitting ya que dividen o segmentan el espacio de las variables predictoras.
* Tienen un bajo coste computacional.
* Es fácil de interpretar y convertir en reglas.
* Puede manejar un gran número de cateréticas.
* Tiene problemas con conjuntos de datos faltantes.
* Tiene problemas con datos complejos.
* Decision trees are computationally faster.
* Random forests is difficult to interpret, while a decision tree is easily interpretable and can be converted to rules.

Ficticias

Las variables ficticias sirven para representar información cualitativa mediante el uso de estas variables.

\* Las variables ficticias sirven para representar información cualitativa mediante el uso de estas variables. Estas sirven tantos para modelos de regresión (ficticias aditivas y multiplicativas) y también sirven para medir niveles por categorías.

# Para seleccionar los tipos de datos numéricos se escoge los tipos que los derriben en una lista, luego se utiliza el método describe(), en donde el percentil 50% equivale a la mediana de los datos.

\* Para determinar el conjunto de modelización se realiza un merge entre el conjunto de variables numéricas y las variables no numéricas (para este caso se toma las variables ficticias). Adicionalmente se excluye la columna Id ya que no es útil para el modelo de estimación. Se separa las variables entre las columnas para realizar la predicción y la columna a predecir el precio de venta de un inmueble(SalePrice).

\* Se segmenta los datos en dos conjuntos uno de entrenamiento con el 80% y de prueba 20% del total de datos respectivamente.

Parametrización del algoritmo de regresión.

\* Para el cálculo de frecuencia se las agrupa para cada variable por cada tipo.

<https://es.wikipedia.org/wiki/Validaci%C3%B3n_cruzada>

Learning the parameters of a prediction function and testing it on the same data is a methodological mistake: a model that would just repeat the labels of the samples that it has just seen would have a perfect score but would fail to predict anything useful on yet-unseen data. This situation is called **overfitting**. To avoid it, it is common practice when performing a (supervised) machine learning experiment to hold out part of the available data as a **test set** X\_test, y\_test.

<https://scikit-learn.org/stable/modules/cross_validation.html>

En general el tratar de que una función aprenda sobre el mismo conjunto de datos sin que exista un conjunto de datos de test se considera como un error metodológico. Es decir que es óptimo para predecir los datos con los que se entrena, pero este generalmente falla al realizar predicciones sobre nuevos conjuntos de datos. Luego podemos utilizar un método de validación cruzada conocido como K-fold cross-validation en donde los datos se dividen en K subconjuntos. Cada subconjunto se utiliza como datos de prueba y el resto (K-1) como datos de entrenamiento.

Se muestra el agrupamiento de datos para cada tipo.

Para determinar el conjunto de modelización se realiza un merge entre el conjunto de variables numéricas y las variables agrupadas por categorías. Adicionalmente se excluye la columna Id ya que no es útil para el modelo de estimación. Se separa las variables entre las columnas para realizar la predicción y la columna a predecir el clasificado por el grupo al que pertenece la venta de un inmueble.

Para el problema y el conjunto de datos para la regresión considero que debe considerarse realizar algún método adicional ya que la efectividad en menor al 92% para ambos métodos.

Para el caso del problema de clasificación el estimador de Random Forest tiene una mejor ratio de éxito sobre los arboles de decisión incluso al excluir el grupo3 en las categorías clasificar. Finalmente sería bueno realizar un contraste contra algún método adicional.

Por otro lado, y en particular considero que esta pregunta depende de ciertos parámetros adicionales que puedan dar una justificación a priori del si los modelos se ajustan a las necesidades del negocio, proyecto, investigación y etcétera, que consideren que estos resultados realmente pueden representar los datos que esperan o que estos modelos pueden representar una ventaja competitiva. Adicionalmente es importante considerar el ámbito de los datos y el rátio de éxito que se espera de los modelos. Por ejemplo, en aplicaciones críticas, citar el buscar estimar la potencia para un láser en una cirugía ocular. Es decir, es importante un criterio adicional basado en la experiencia de los datos y juicio de expertos que puedan sustentar si el modelo se ajusta al contexto en que se presentó como una posible solución.

En conclusión, se sugiere realizar una contrastación contra otros algoritmos y adicionalmente validar los resultados con juicios de expertos en el contexto que el problema este definido.

Presentación de las curvas de aprendizaje para cada modelo

La **curva de aprendizaje es** una gráfica del rendimiento del aprendizaje **modelo** sobre la experiencia o el tiempo. Y estas sirven para diagnosticar un modelo de conjunto, o realizar un ajuste adecuado.

Para las funciones de regresión el método de random forest son menores las métricas error cuadrático medio y R2 y por tanto es el mejor método para predecir un valor.

Para las funciones de clasificación la matriz de confusión, cuyos valores son métricas de las estimaciones, el método de random forest tiene el valor de ratio de éxito mayor y por tanto es el mejor método para clasificar.

decisión