Trabajo práctico 2

La tarea puede realizarse en R o Python, o en otro software. Cada grupo tendrá puntos extra según el rendimiento del Árbol de Decisión que haya construido. Para mejorar el rendimiento de la clasificación se aconsejan métodos de ingeniería de características, como el análisis de correlación o la transformación de características.

Tabla 1. Información del conjunto de datos

Columna	Descripción
año_de_trabajo	El año en que se pagó el salario.
nivel_de_experiencia	El nivel de experiencia en el trabajo durante el año con los siguientes valores posibles: ES Nivel inicial / junior MI Nivel medio / intermedio SE Nivel superior / experto EX Nivel ejecutivo / director
tipo_de_empleo	El tipo de empleo para el papel: PT Tiempo parcial FT Tiempo completo CT Contrato FL Freelance
título_de_trabajo	El papel trabajado durante el año.
salario	El importe total del salario bruto pagado.
moneda_salario	La moneda del salario pagado como un código de moneda ISO 4217.
salaryinusd	El salario en USD (tipo de cambio dividido por el tipo de cambio medio en USD para el año correspondiente a través de fxdata.foorilla.com).
residencia_del_empleado	País principal de residencia del empleado durante el año laboral como código de país ISO 3166.
relación_remota	La cantidad total de trabajo realizado a distancia; los valores posibles son los siguientes: 0 Sin trabajo a distancia (menos del 20%) 50 Parcialmente a distancia 100 Totalmente a distancia (más del 80%)
ubicación_de_la_empresa	El país de la oficina principal del empleador o de la sucursal contratante como código de país ISO 3166.
tamaño_de_la_empresa	El número medio de personas que han trabajado para la empresa durante el año: S menos de 50 empleados (pequeña) M de 50 a 250 empleados (mediana) L más de 250 empleados (grande)

Fuente: https://www.kaggle.com/datasets/ruchi798/data-science-job-salaries

Los deberes

Procesamiento previo

1. Transformar las características categóricas en variables binarias (variables ficticias). Las variables ficticias son aceptadas por casi cualquier clasificador. **(1 punto)**

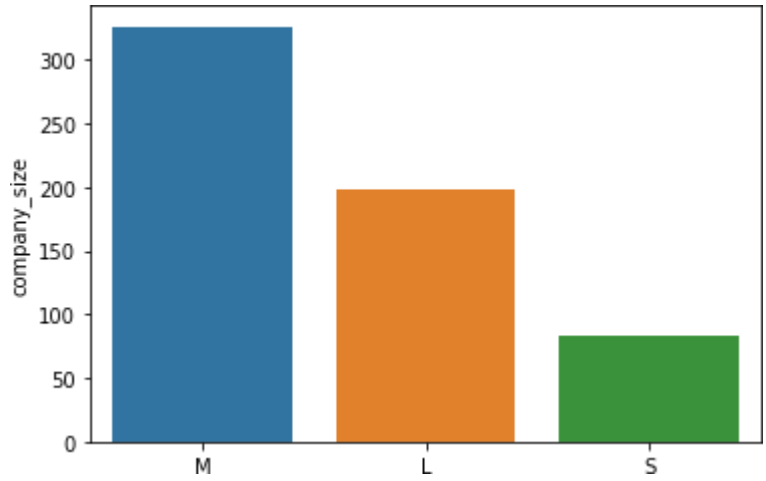
Procesamiento

2. Ajustar un árbol de decisión para clasificar el tamaño de la empresa en la que trabaja cada científico de datos. Esto está representado por la característica "company_size", que es la variable objetivo. (**3 puntos)**
   2.1. Establezca la profundidad máxima del árbol en 3. (1 punto)
   2.2. Trazar el árbol de decisión. (2 puntos)
3. Interpretar los resultados del árbol de decisión **(3 puntos)**
   3.1. Escribe un párrafo interpretando la estructura del árbol de decisión.

Procesamiento posterior

4. Medir el rendimiento del árbol de decisión para hacer predicciones precisas.
   4.1. Por el rendimiento de la validación cruzada con retención **(3 puntos)**
      4.1.1. Estime la precisión, interprete los resultados. (1,5 puntos)
      4.1.2. Estimar el coeficiente de correlación de Matthews, interpretar los resultados (1,5 puntos)

Figura 1. Distribución de las clases para "company_size



- La variable objetivo, o la variable dependiente, que queremos predecir está desequilibrada entre las clases. El mercado laboral de los científicos de datos está compuesto por empresas medianas y grandes (13,67% de los casos).
- Algoritmo de validación cruzada Hold-out:
  - ✓ Seleccione un tamaño de muestra para aprender y probar el Árbol de Decisión. Ejemplo: 80% de la muestra para aprender y 20% para probar.
  - ✓ Dividir aleatoriamente los datos en subconjuntos de entrenamiento y de prueba.
  - ✓ Ajustar un Árbol de Decisión a los datos de entrenamiento y estimar la puntuación en los datos de prueba. Ejemplo: estimar la precisión del árbol de decisión en el subconjunto de prueba.
- La validación cruzada de retención puede repetirse varias veces con diferentes semillas aleatorias para extraer una distribución de puntuaciones y realizar una evaluación robusta de las mismas.
- Se puede probar cualquier otro método de ingeniería de características, o de reducción de la dimensionalidad para obtener mejores resultados.