Pontificia Universidad Católica de Chile
Educación Profesional - Escuela de Ingeniería
Diplomado en Big Data y Ciencias de Datos
Minería de Datos
Relator: Sebastián Raveau

Evaluación 2 – Árboles de Decisión

Fecha de Entrega: domingo 13 de septiembre

El objetivo de esta evaluación es utilizar y analizar árboles de clasificación. La base de datos contiene 891 registros de pasajeros del RMS Titanic. Las variables relevantes son:

- *Survived*, indica si el pasajero sobrevivió al naufragio (variable endógena)
- *Pclass*, indica la clase tarifaria en la que viajaba el pasajero
- *Sex*, indica el sexo del pasajero
- *Age*, indica la edad del pasajero

Usted deberá realizar un análisis comparativo en R de los métodos de clasificación vistos en clases. El análisis debe incluir al menos:

- El impacto del método de división: *information* vs *gini*.
- El impacto del parámetro *cp* asociado al costo de complejidad
- El impacto de los parámetros de parada anticipada *minsplit*, *minbucket* y *maxdepth*.

Cada comparación es independiente entre sí (i.e. no se complique comparando *gini* con cierto *minsplit* vs *information* con otro *minsplit*). Los criterios de análisis y comparación consisten en una evaluación crítica del árbol resultante y la capacidad predictiva del modelo (obteniendo, por ejemplo, matrices de confusión).

Debe generar un breve reporte con su análisis y entregarlo hasta el domingo 13 de septiembre al correo: mineria.datos.PUC@gmail.com. El asunto del correo debe comenzar con [Evaluación 2] seguido con los apellidos de los estudiantes. Ejemplo: "[Evaluación 2] Gutiérrez y Soto".

El objetivo de la tarea es demostrar que entiende lo que está haciendo R al obtener los árboles de clasificación. Por favor presente sólo la información relevante, sin llenar múltiples páginas con gráficos, códigos y estadísticas.

Su entrega será evaluada de acuerdo con los siguientes aspectos:

Claridad [1,5 punto]

Se espera que el reporte sea claro y bien redactado, con las ideas debidamente desarrolladas. Las tablas y figuras deben estar debidamente presentadas y explicadas. El lenguaje utilizado debe ser apropiado para un reporte técnico.

Análisis del método de división [1,5 puntos]

Se presentan y comparan los resultados de distintos métodos de división. Los parámetros utilizados se presentan y fundamentan en forma explícita. Se discuten los resultados desde una perspectiva crítica (e.g. profundidad, tamaño, variables consideradas, divisiones realizadas, etc.). Se presenta y compara la capacidad predictiva de los árboles resultantes.

Análisis del parámetro de complejidad [1,5 puntos]

Se presentan y comparan los resultados de utilizar distintos parámetros de complejidad. Los parámetros utilizados se presentan y fundamentan en forma explícita. Se discuten los resultados desde una perspectiva crítica (e.g. profundidad, tamaño, variables consideradas, divisiones realizadas, etc.). Se presenta y compara la capacidad predictiva de los árboles resultantes.

Análisis de parámetros de parada anticipada [1,5 puntos]

Se presentan y comparan los resultados de considerar distintos parámetros de parada anticipada. Los parámetros utilizados se presentan y fundamentan en forma explícita. Se discuten los resultados desde una perspectiva crítica (e.g. profundidad, tamaño, variables consideradas, divisiones realizadas, etc.). Se presenta y compara la capacidad predictiva de los árboles resultantes.