Samuel Hernández Guerrero DAM2

Clasificador Titanic utilizando Weka

1. ¿Qué atributos son más importantes según el árbol de decisión?

   En J48 (C4.5), se crea el árbol mediante la selección del atributo que presenta una ganancia de información más alta en relación con la clase objetivo (survived).

   Los atributos más significativos en el conjunto de datos del Titanic suelen ser los siguientes:

   sex → La supervivencia está mayormente relacionada con ser mujer.

   class → Los pasajeros de primera clase tuvieron más posibilidades de sobrevivir en comparación con los de tercera o la tripulación.

   age → En el rescate, los niños tuvieron prioridad.

   Por lo tanto, el árbol generalmente comienza con sex como nodo raíz y luego class, a veces también age.

2. ¿Qué ocurre si eliminas un atributo y vuelves a entrenar el modelo?

   Si eliminamos un atributo fundamental (como, por ejemplo, el sexo):
   El árbol se tendrá que basarse en otros atributos, como por ejemplo, age o class .

   La exactitud se reduce, ya que se pierde una variable con alta capacidad predictiva.

   Si suprimimos un atributo de menor importancia (por ejemplo, age):
   El efecto es más pequeño porque "sex" y "class" continúan siendo muy informativos.

   Aunque el efecto depende del valor del atributo que se elimina, en general, la habilidad del modelo para detectar patrones disminuye cuando se quitan atributos.

3. ¿Cuál de los modelos utilizados funciona mejor con este conjunto de datos?

De los cuatro modelos que se han evaluado el que proporciona el balance más adecuado entre simpleza y precisión es J48 (árbol de decisión). Este algoritmo logra una tasa de aciertos bastante alta en el conjunto de datos del Titanic y, por otra parte, facilita una interpretación sencilla de cómo se toman las decisiones. El árbol muestra reglas claras, como que ser mujer o niño aumenta el riesgo de sobrevivir, pero ser tripulante o pertenecer a la tercera clase lo disminuye.

Aunque otros modelos como SMO (SVM) pueden lograr una exactitud ligeramente superior, son más difíciles de comprender y no proporcionan una representación visual tan intuitiva. Naive Bayes es rápido y fácil de usar, pero los conceptos de independencia restringen la exactitud de las predicciones. IBk (KNN) es muy dependiente del número de vecinos y puede tener una estabilidad menor.

Por lo tanto, si tuvieras que elegir uno que sea más efectivo en este ejercicio, lo normal es quedarse con J48, ya que ofrece una buena precisión y se puede explicar y visualizar de manera sencilla.