**Clasificación ->** Es una forma de analizar datos para generar *modelos que describen importantes clases de datos (clasificadores).*

* Datos CATEGORICOS
* Datos NOMINALES y BINARIOS
* No considera el ORDEN JERARQUICO de grupos
* APRENDIZAJE SUPERVISADO, ya que se conocen las etiquetas de clase asociadas a cada tupla de entrenamiento

Es la tarea de aprendizaje que considera una función *f* (*modelo de clasificación*) que asocia cada conjunto de atributos *x* a una de las clases predefinidas y etiquetas en *y*

Un modelo de clasificación puede servir como:

* **Modelo Descriptivo ->** Puede diferenciar objetos de distintas clases
* **Modelo de Predicción ->** Puede predecir la etiqueta de clase de un registro desconocido

**Pasos de la clasificación**

1. **Aprendizaje de Mapeo o Función:** *Se construye el modelo de clasificación*, por medio del uso de un *algoritmo de clasificación* el cual analiza y aprende de un conjunto de entrenamiento compuesto por tuplas y las etiquetas de clase asociadas a cada una de estas.
2. **Clasificación:** *Se utiliza el modelo de clasificación o clasificador para predecir las etiquetas de clase de otros datos.* Se estima la precisión de este modelo, a partir del porcentaje de las tuplas de prueba que el clasificador clasifica CORRECTAMENTE. Si la precisión del clasificador se considera *aceptable*, se puede usar para clasificar tuplas de datos futuras y desconocidas.

A picture containing text, diagram, screenshot

Description automatically generated

**Árbol de decisión ->** Es una estructura de árbol similar a un diagrama de flujo, donde cada nodo interno representa una prueba de atributo, cada rama representa el resultado de una prueba, y cada nodo hoja o nodo terminal tiene una etiqueta de clase.

|  |  |  |
| --- | --- | --- |
| **ID3** | **C4.5** | **CART** |
| Atributos CATEGORICOS | Atributos NUMERICOS | Atributos CATEGORICOS Y NUMERICOS |
| Usa la *ganancia de información (****Information Gain****)* como criterio de división | Usa la *relación de ganancia (****Gain Ratio****)* como criterio de división | Usa el *índice Gini*, para elegir el atributo de división |
| Deja de crecer cuando:   * Todas las instancias de un nodo pertenecen a la misma clase (**nodo puro**), o el mínimo de atributos para dividir no se cumple. * Cuando la mejor *ganancia de información* no es mayor a 0 | Deja de crecer cuando el numero de instancias a dividir esta por debajo de cierto umbral |  |
| No aplica poda | Aplica poda cuando el árbol deja de crecer, y se basa en errores |  |
| No maneja valores numéricos ni faltantes | Maneja valores faltantes | Construye arboles de clasificación y regresión |
| Puede causar un sobreajuste en los datos de entrenamiento | Es **n-ario** con valores discretos, y **binario** con valores continuos | Es puramente **binario** |
| No garantiza una solución optima |  |  |

**Entropía ->** Mide la cantidad de *incertidumbre* en un conjunto de datos.

* Varia de 0 a 1
* En una *muestra totalmente homogénea (todos los elementos pertenecen a la misma clase)*, la incertidumbre es igual a **0.**
* En una *muestra igualmente distribuida (dos clases que tienen exactamente la misma cantidad de elementos)*, la incertidumbre es igual a **1.**

A picture containing font, text, white, line

Description automatically generated

**Ganancia de Información ->** La diferencia entre la cantidad de información que se necesita para hacer una clasificación correcta antes de la división, y la cantidad de información después de la división.

A black text on a white background

Description automatically generated with medium confidence

Para determinar la **bondad de una condición de prueba de atributo** o **la ganancia de pureza**, se debe comparar el grado de impureza del nodo primario con el grado de impureza de los nodos secundarios.



**Gain Ratio ->** Es una extensión de la *ganancia de información.*

A picture containing font, typography, handwriting, calligraphy

Description automatically generated



**Gini Index ->** Mide la impureza de una partición de datos o un conjunto de tuplas de entrenamiento.

A picture containing font, text, white, calligraphy

Description automatically generated

Las medidas para evaluar la *impureza* de un nodo:

A picture containing text, font, receipt, white

Description automatically generatedA picture containing line, diagram, plot

Description automatically generated

**Algoritmo de inducción ->** Es una entidad que obtiene un conjunto de entrenamiento y forma un modeo que generaliza la relación entre los atributos de entrada y el atributo objetivo.

**Error de generalización ->** Tasa de clasificación errónea de una distribución D.

**Exactitud de la clasificación ->** Uno menos el error de generalización.

**Error de entrenamiento ->** Es el porcentaje de ejemplos en el conjunto de entrenamiento clasificados correctamente por el árbol de clasificación.

**Método de retención ->** Un conjunto de datos se divide aleatoriamente en un conjunto de entrenamiento y un conjunto de prueba.

**Submuestreo Aleatorio ->** Los datos se dividen aleatoriamente varias veces en conjuntos de entrenamiento y conjuntos de prueba disjuntos.

**validación Cruzada ->** Los datos se dividen aleatoriamente en n subconjuntos mutuamente excluyentes de aproximadamente el mismo tamaño.

**Comprensibilidad ->** Que tan bien los humanos captan el clasificador inducido

**Navaja de Occam ->** Cuando se busca la explicación de un fenómeno, uno debe hacer la menor cantidad de suposiciones y eliminar aquellas que no hacen ninguna diferencia en las predicciones observables de la hipótesis explicativa.

**Robustez ->** Capacidad del modelo de manejar valores faltantes o con ruido, y hacer predicciones correctas.

**Estabilidad ->** Es el grado en el que un algoritmo genera resultados repetibles, dados diferentes lotes de datos del mismo proceso.

**Sobreajuste ->** Se refiere a la situación en la que el algoritmo de inducción genera un clasificador que se ajusta perfectamente a los datos del entrenamiento, pero que ha perdido la capacidad de generalizar a instancias no presentadas durante el entrenamiento.

**Sesgo ->** Que tanto se alejan en promedio las predicciones de un modelo respecto a los valores reales.

**Varianza ->** Que tanto se dispersan en promedio los datos de su media.

**Métodos de Conjunto:**

**Bagging ->** Genera subconjuntos de prueba seleccionados aleatoriamente y con reemplazado. Dado a que hay un conjunto de clasificadores, la predicción de nuevos de nuevos ejemplos se efectúa por votación mayoritaria.

**Boosting ->** Construye nuevos modelos tratando de corregir errores cometidos previamente.

**Random Forest ->** Utiliza una gran cantidad de árboles de decisión individuales sin podar, los cuales se crean aleatorizando la división en cada nodo del árbol de decisión.

**Soporte ->** Es el porcentaje de transacciones en D, que contienen el antecedente y el consecuente, o A unión B.

“En el tanto % del total de transacciones analizadas, existen A y B juntos”

**Confianza ->** Es el porcentaje de transacciones en D, que tienen tanto a A como a B. Si existe B, existe A.

“El tanto % de las transacciones que tienen A, también tienen B”

**Algoritmo a priori ->** Algoritmo para encontrar conjuntos de elementos frecuente

**Lift ->** Medición que permite cuantificar la utilidad de una regla de asociación.

**Modelo ->** Es una descripción global o explicación de un conjunto de datos, desde una perspectiva de alto nivel.

**Patrones ->** Características locales de los datos.