Modelos Supervisados

Certificación en Ciencia de Datos

Andrés Martínez

28 Junio 2024

Universidad Externado de Colombia
135 Años

Contenido

Árboles de Decisión

- Características
- Regresión.
- Clasificación

Árbol de Decisión

Un **árbol de decisión** es un modelo de aprendizaje automático utilizado para la toma de decisiones y la clasificación de datos. Se asemeja a una estructura de árbol, donde cada nodo representa una decisión o un punto de división, y cada hoja representa una etiqueta o una decisión final.

- Los árboles de decisión se utilizan para resolver problemas de clasificación y regresión.
- El objetivo es dividir el conjunto de datos en ramas que optimicen la toma de decisiones.
- Cada nodo se basa en una característica y un umbral para tomar decisiones.
- Los árboles pueden ser profundos (con muchas divisiones) o poco profundos.

Características Árbol de Decisión

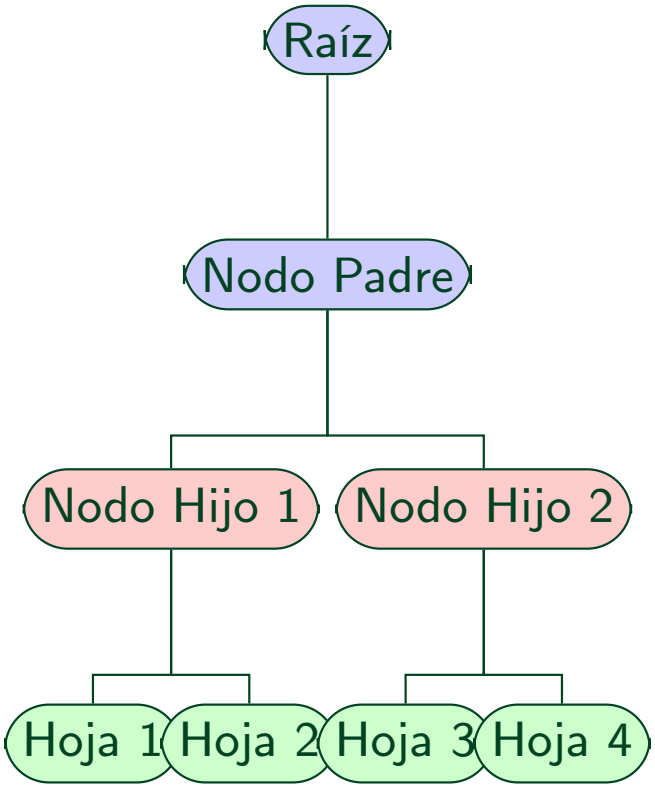
- Los árboles de decisión son interpretables y fáciles de visualizar.
- Pueden manejar datos categóricos y numéricos.
- Sin embargo, pueden ser propensos al sobreajuste en conjuntos de datos ruidosos.
- Estrategias como la poda de árboles se utilizan para evitar el sobreajuste.
- Los árboles de decisión son componentes clave en ensambles como el bosque aleatorio.
- Estructura jerárquica en forma de árbol.
- Cada nodo representa una decisión.
- Cada hoja representa una etiqueta o valor.
- Basado en divisiones de características.
- Puede manejar datos categóricos y numéricos.
- Propenso al sobreajuste en datos ruidosos.

Nivel de Impureza en un Árbol de Decisión

La impureza es una medida fundamental en la construcción de un árbol de decisión. Representa la falta de homogeneidad en un conjunto de datos y es esencial en la toma de decisiones sobre cómo dividir los nodos del árbol.

- Los árboles de decisión buscan minimizar la impureza en los nodos para lograr una clasificación o regresión precisa.
- Las métricas de impureza comunes incluyen el índice Gini, la entropía y el error cuadrático medio (MSE).
- Un nodo con impureza cero significa que todos los puntos de datos en ese nodo pertenecen a la misma clase (en el caso de la clasificación) o tienen el mismo valor de destino (en la regresión).
- La impureza se utiliza para determinar cómo dividir un nodo en dos nodos hijos, seleccionando la división que maximiza la reducción de impureza.
- La elección de la métrica de impureza depende del problema y los datos específicos.

Partes de un Árbol



- **Raíz:** El nodo superior del árbol, desde donde comienza la toma de decisiones.
- **Nodo Padre:** Un nodo que tiene uno o más nodos hijos.
- **Nodo Hijo:** Un nodo que se deriva de un nodo padre.
- **Nivel de Impureza:** En cada nivel del árbol, se evalúa la impureza de los datos. Los nodos se dividen para reducir la impureza.

Árbol de Decisión en Regresión

Un **árbol de decisión en regresión** es un modelo de aprendizaje automático que se utiliza para predecir valores numéricos en función de características de entrada. Matemáticamente, un árbol de decisión en regresión se puede definir como:

$$f(x) = \sum_{i=1}^{N} c_i \cdot \mathbb{I}(x \in R_i)$$

Donde:

- $f(x)$ es la predicción para la entrada x.
- N es el número de nodos hoja en el árbol.
- R_i representa la región del espacio de características asignada al nodo hoja i.
- c_i es el valor promedio de las etiquetas de entrenamiento en la región R_i.
- $\mathbb{I}(x \in R_i)$ es una función indicadora que devuelve 1 si x pertenece a la región R_i, y 0 en caso contrario.

El árbol de decisión divide el espacio de características en regiones y asigna un valor de regresión a cada región.

Medición del Nivel de Impureza en un Árbol de Regresión

El nivel de impureza en un árbol de regresión se mide mediante la varianza o el error cuadrado medio (MSE, por sus siglas en inglés). La fórmula matemática para calcular el MSE es la siguiente:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Donde:

- MSE es el error cuadrado medio, que mide la impureza.
- n es el número de muestras en un nodo hoja.
- y_i es el valor real de la variable objetivo para la muestra i.
- $\hat{y}_i$ es la predicción del modelo para la muestra i.
- La suma se realiza sobre todas las muestras en el nodo hoja.

Un valor bajo de MSE indica que las predicciones del árbol son precisas y que el nodo hoja es puro en términos de regresión.

Árbol de Regresión

La variable dependiente es Costo de Vida

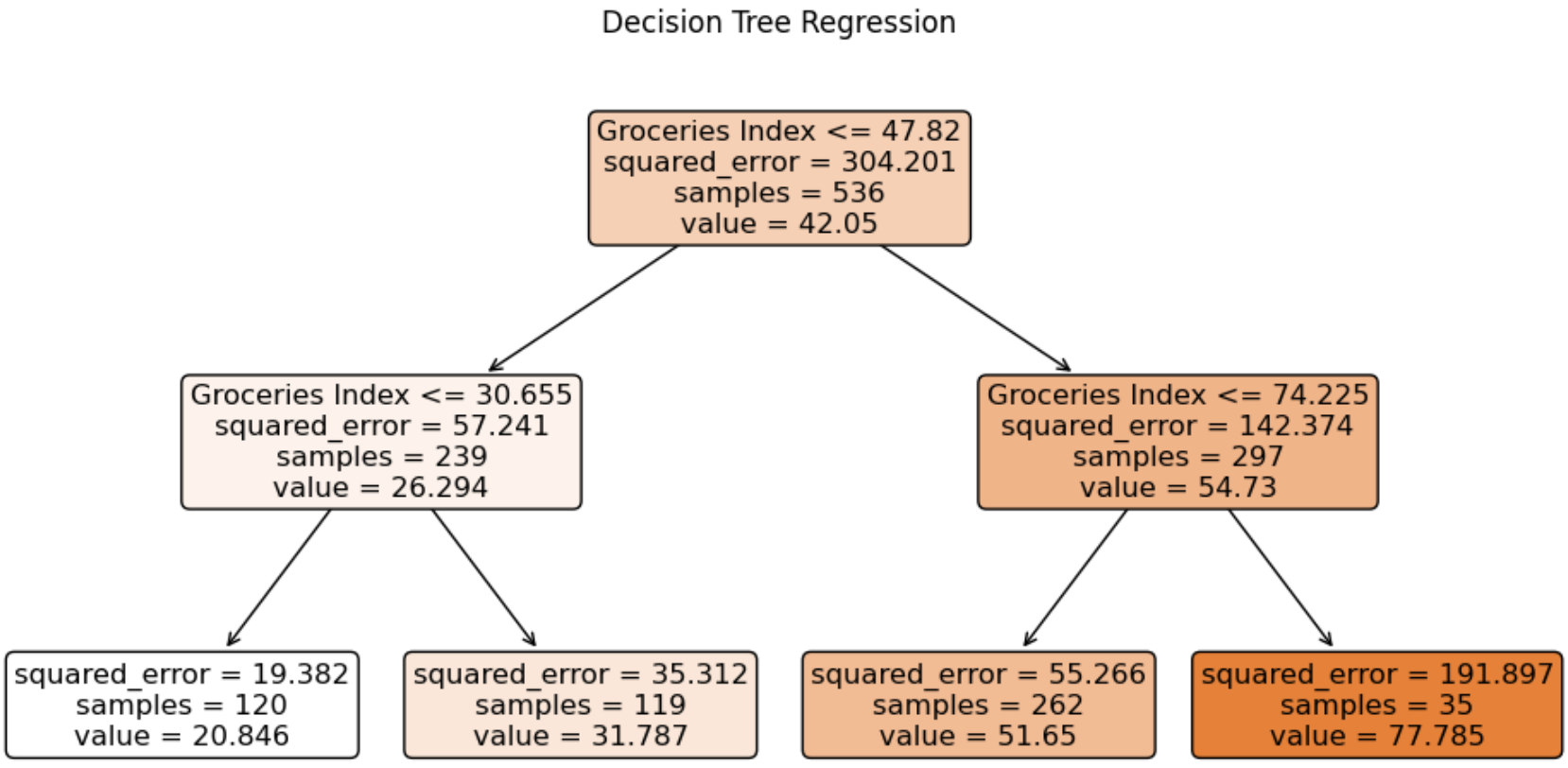
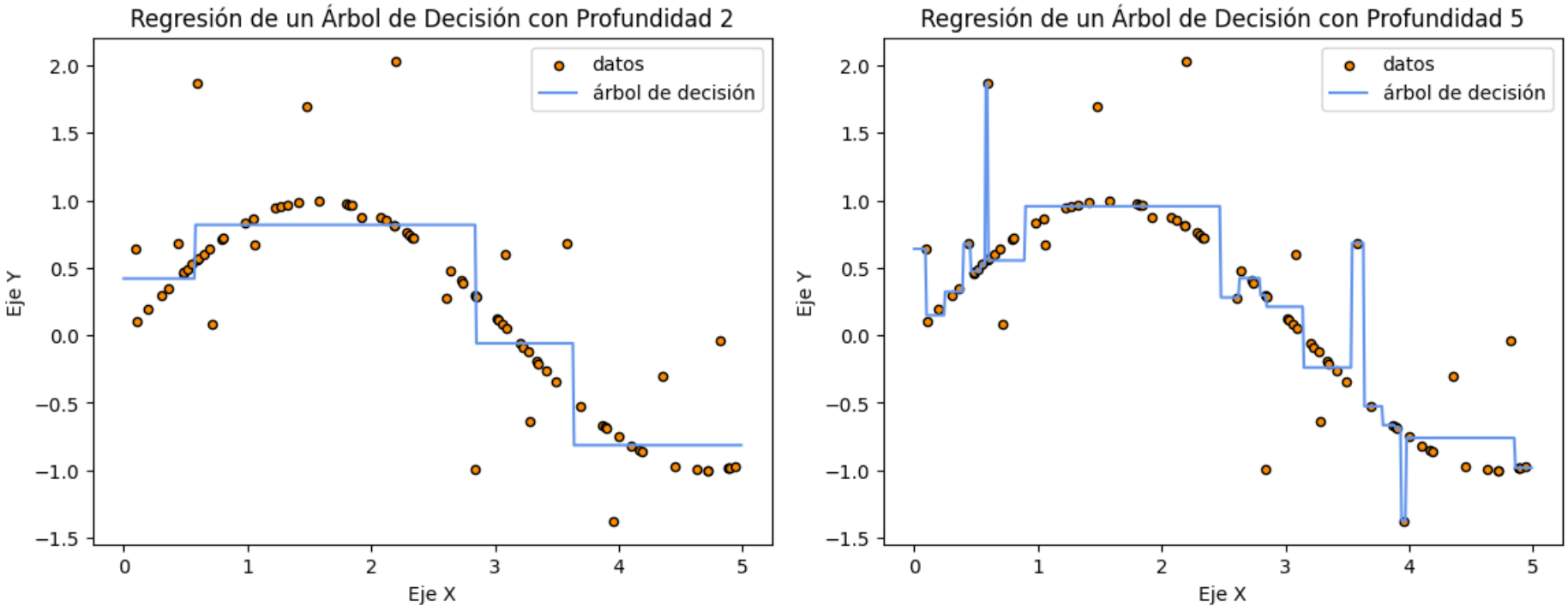


Figure 1: Regresión

Calibración de un árbol de regresión



(a) Nivel de profundidad 2

(b) Nivel de profundidad 5

Figure 2: Comparación de árboles de regresión

Árbol de Clasificación: Explicación Matemática

Un árbol de clasificación es un modelo de aprendizaje automático que se utiliza para predecir una etiqueta de clase (c) a partir de características (X). La construcción del árbol se basa en la minimización de la impureza.

Notación

N : Número de ejemplos en un nodo

H : Impureza de un nodo

c : Clase

$p(c|N)$: Proporción de ejemplos de la clase c en un nodo

Impureza en Áboles de Clasficación

Entropía (H)

La entropía es una medida de impureza en un nodo de un árbol de decisión. Se calcula utilizando la fórmula:

$$H(N) = -\sum_{c} p(c|N) \cdot \log(p(c|N))$$

Donde:

- $H(N)$: Entropía en el nodo N.
- c: Clase o categoría.
- $p(c|N)$: Proporción de ejemplos de la clase c en el nodo N.

La entropía varía de 0 (cuando el nodo es puro) a 1 (cuando el nodo es completamente impuro).

Impureza en Áboles de Clasficación

Impureza Gini (G)

El Gini impurity es otra medida de impureza en un nodo de un árbol de decisión. Se calcula utilizando la fórmula:

$$G(N) = 1 - \sum_{c} (p(c|N))^2$$

Donde:

- $G(N)$: Impureza Gini en el nodo N.
- c: Clase o categoría.
- $p(c|N)$: Proporción de ejemplos de la clase c en el nodo N.

El Gini impurity también varía de 0 (cuando el nodo es puro) a 1 (cuando el nodo es completamente impuro).

Impureza en Áboles de Clasficación

División de Nodos

Para construir el árbol, se selecciona la división que minimiza la impureza en los nodos hijos. Esto se logra evaluando todas las posibles divisiones y eligiendo la que maximiza la reducción de la impureza (usualmente se utiliza la ganancia de información o la reducción de la Gini).

Ganancia de Información (IG):

$$IG(N, A) = H(N) - \sum_{v \in A} \frac{|N_v|}{|N|} \cdot H(N_v)$$

Reducción de Gini (IG):

$$G(N, A) = G(N) - \sum_{v \in A} \frac{|N_v|}{|N|} \cdot G(N_v)$$

A representa una variable que denota una característica o atributo específico que se utiliza para dividir los nodos en un árbol de decisión

Árbol de Clasificación

La variable dependiente es default

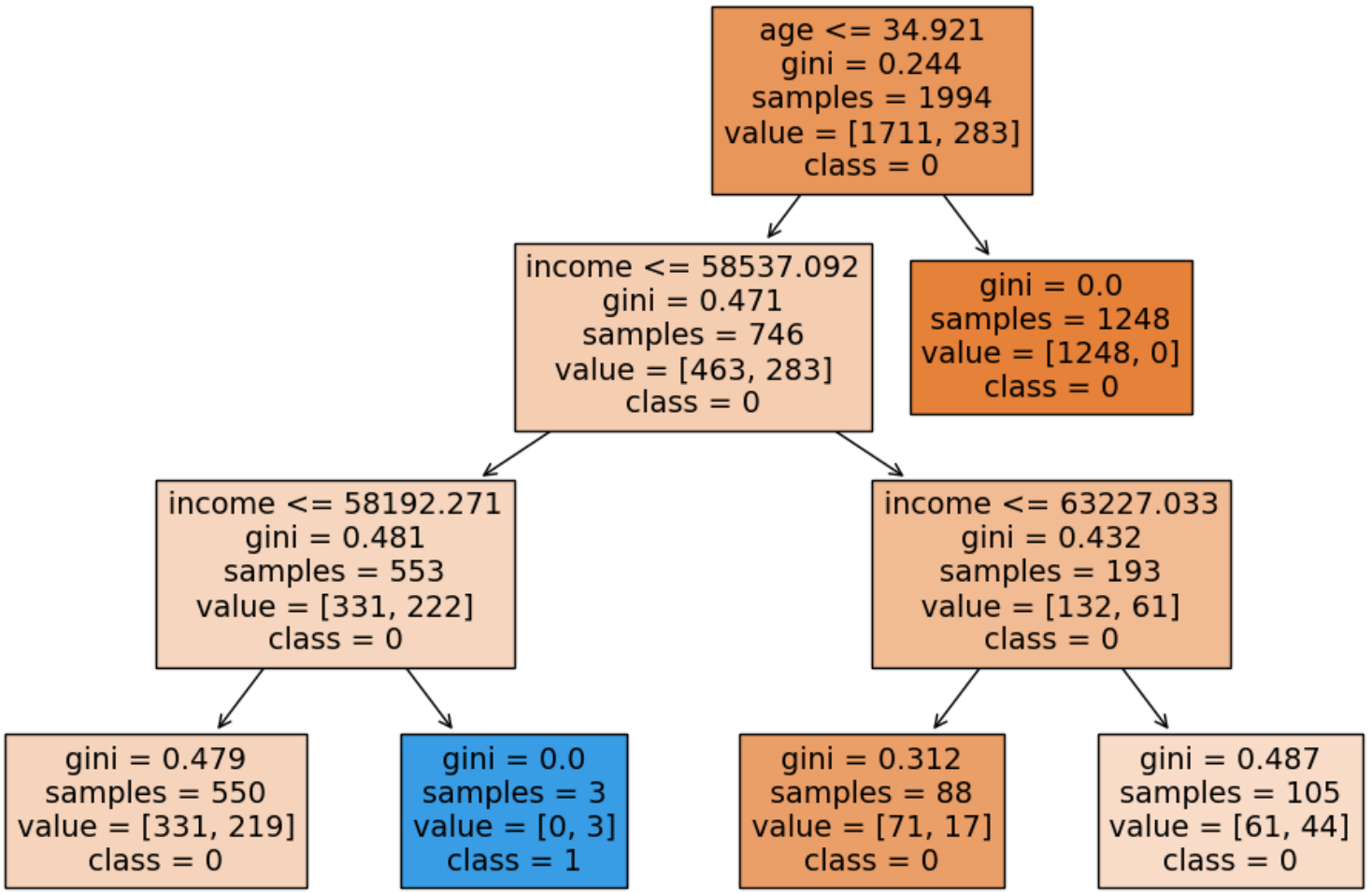


Figure 3: Clasificación

Cálculo del Gini y la Entropía

Entropía y Gini Impurity

La entropía (H) y la impureza Gini (G) son dos métricas comunes para medir la impureza de un nodo en un árbol de clasificación:

Entropía (H):

$$H(N) = -\sum_{c} p(c|N) \cdot \log(p(c|N))$$

Impureza Gini (G):

$$G(N) = 1 - \sum_{c} (p(c|N))^2$$

Cálculo del Gini y la Entropía

Ejemplo de Cálculo de Entropía (H)

Supongamos tenemos un nodo con tres ejemplos de dos clases diferentes:

$$N = 3, \quad p(\text{Clase A}|N) = \frac{1}{3}, \quad p(\text{Clase B}|N) = \frac{2}{3}$$

Calculamos la entropía:

$$H(N) = -\left(\frac{1}{3} \cdot \log\left(\frac{1}{3}\right) + \frac{2}{3} \cdot \log\left(\frac{2}{3}\right)\right)$$

Cálculo del Gini y la Entropía

Ejemplo de Cálculo de Impureza Gini (G)

Para el mismo nodo con tres datos de dos clases diferentes:

$$N = 3, \quad p(\text{Clase A}|N) = \frac{1}{3}, \quad p(\text{Clase B}|N) = \frac{2}{3}$$

Calculamos la impureza Gini:

$$G(N) = 1 - \left(\left(\frac{1}{3} \right)^2 + \left(\frac{2}{3} \right)^2 \right)$$

Algoritmo k-Nearest Neighbors (k-NN)

Introducción

El algoritmo k-Nearest Neighbors (k-NN) es un método de aprendizaje supervisado que se utiliza para clasificación y regresión. Su enfoque se basa en la proximidad de los puntos de datos en un espacio de características.

Clasificación en k-NN

Dado un punto de consulta Y, la clasificación en k-NN se realiza de la siguiente manera:

- Calculamos la distancia entre Y y todos los puntos en el conjunto de entrenamiento.
- Seleccionamos los k puntos más cercanos a Y según la medida de distancia.
- Contamos las clases de estos k puntos.
- Asignamos la clase más común a Y como la predicción.