ESTRUCTURA DE DATOS PARA LA PREDICCIÓN DE RENDIMIENTO EN LAS PRUEBAS SABER PRO

UNIVERSIDAD EAFIT

Presentación del Equipo

Juan Pablo Restrepo

Juan José Sánchez

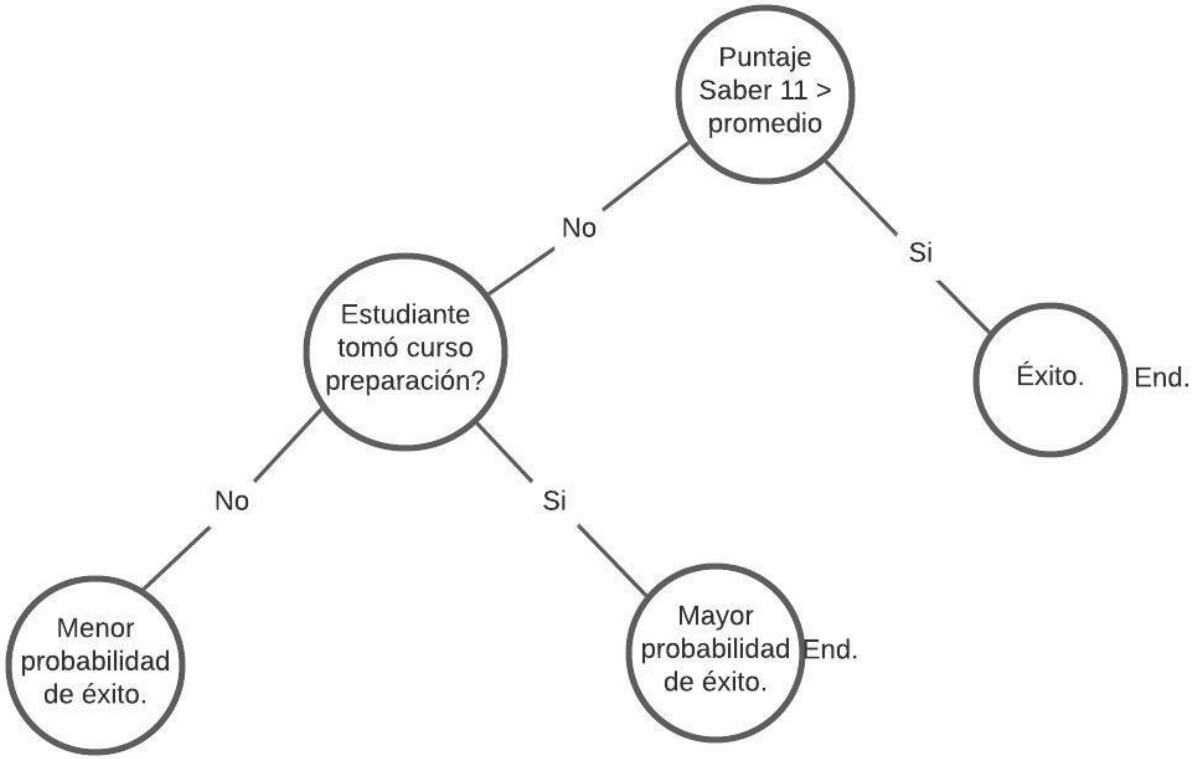
Miguel Correa

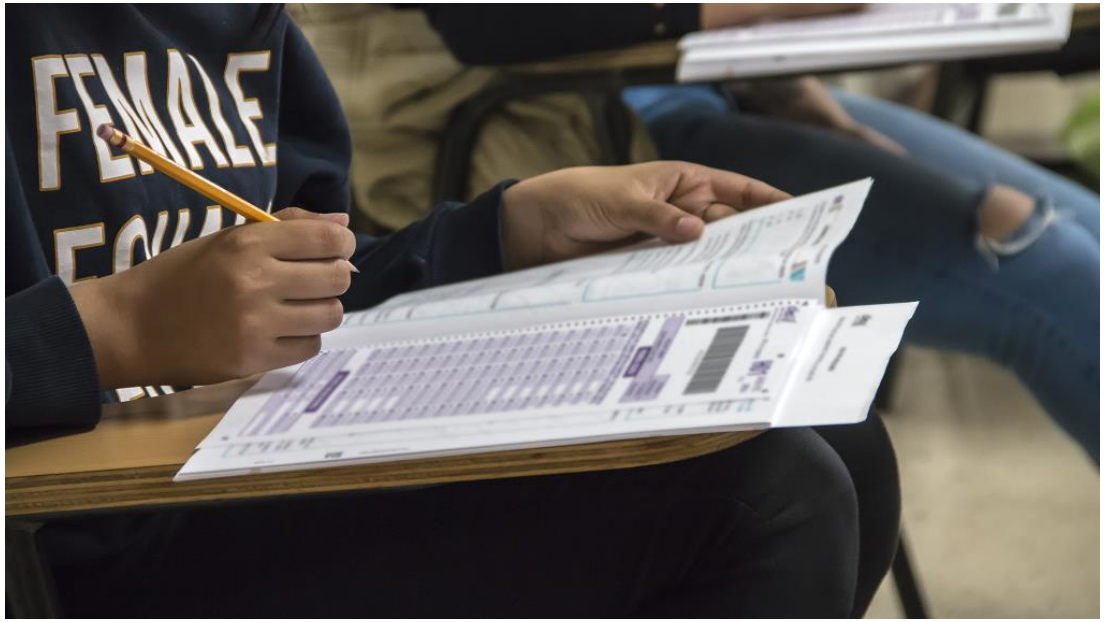
Mauricio Toro

https://github.com/jprestrepe/ST0245002/tree/master/proyecto

UNIVERSIDAD EAFIT

Diseño del Algoritmo

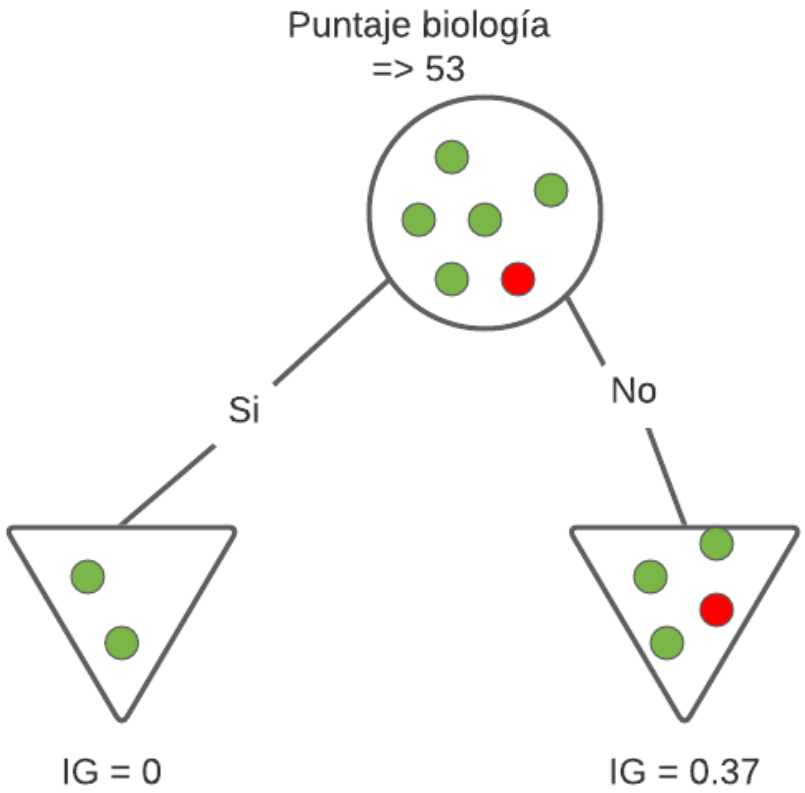




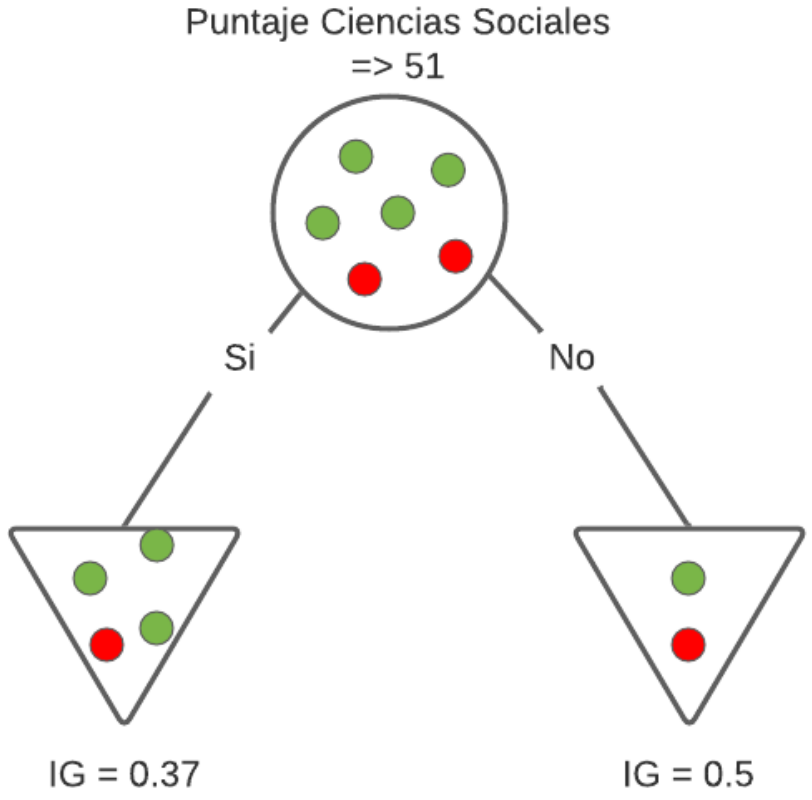
Trabajamos con el algoritmo CART para crear nodos por medio de variables. Utilizando la impureza de Gini podremos generar nodos que determinen más específicamente la probabilidad del éxito para los estudiantes que presenten la prueba Saber Pro.

Imagen tomada de Google -> https://www.icfes.gov.co/documents/20143/1390648/contactenos.png/864ec610-906e-7058-97c7-94c2f9c8e54c?t=1553124728669

División de un nodo



Esta división usa como condición “Puntaje biología => 53” dando como resultado en el nodo izquierdo una Impureza de Gini de 0 y en el nodo derecho una Impureza de Gini de 0.37. La impureza ponderada es 0.24



Esta división usa como condición “Puntaje Ciencias Sociales => 53” dando como resultado en el nodo izquierdo una Impureza de Gini de 0.37 y en el nodo derecho una Impureza de Gini de 0.5. La impureza ponderada es 0.41

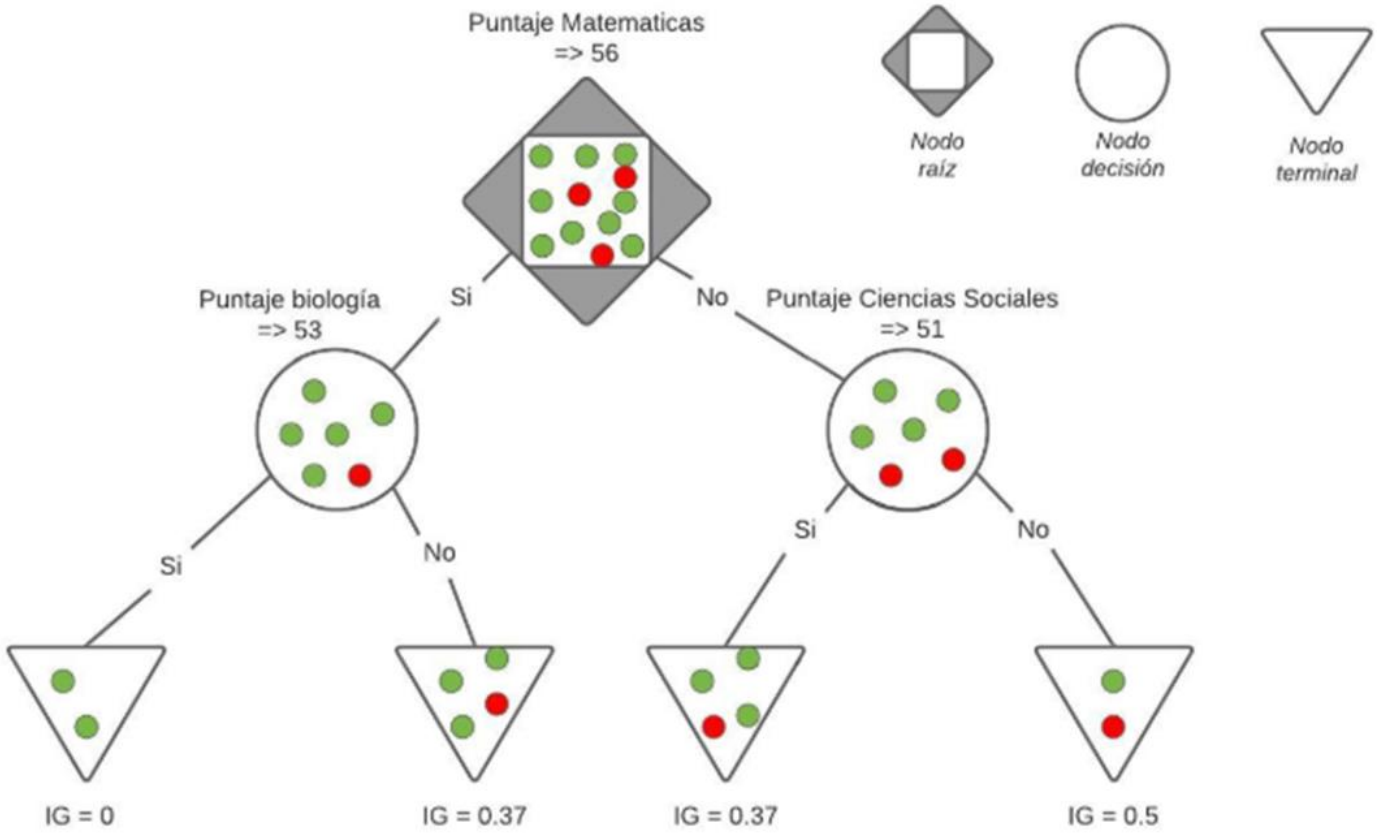
Complejidad del Algoritmo

	Complejidad en tiempo	Complejidad en memoria
Entrenamiento del modelo	$O(N^{\wedge 2}M*2^{\wedge M})$	$O(N*M*2^{\wedge M})$
Validación del modelo	O(N*M)	O(1)

Complejidad en tiempo y memoria del algoritmo del algoritmo CART. Donde N hace referencia a las filas y M a las columnas.

Imagen tomada de Google -> https://unmsm.edu.pe/noticias/templates/noticias2016/junio/d27/ExamenPREII_03_g.JPG

Modelo de árbol de decisión



Este modelo de entrenamiento permite al algoritmo, saber que preguntas va a recibir y por lo tanto poder trabajar con una condición que le permita generar la impureza de Gini según la posición indicada.

Características mas importantes

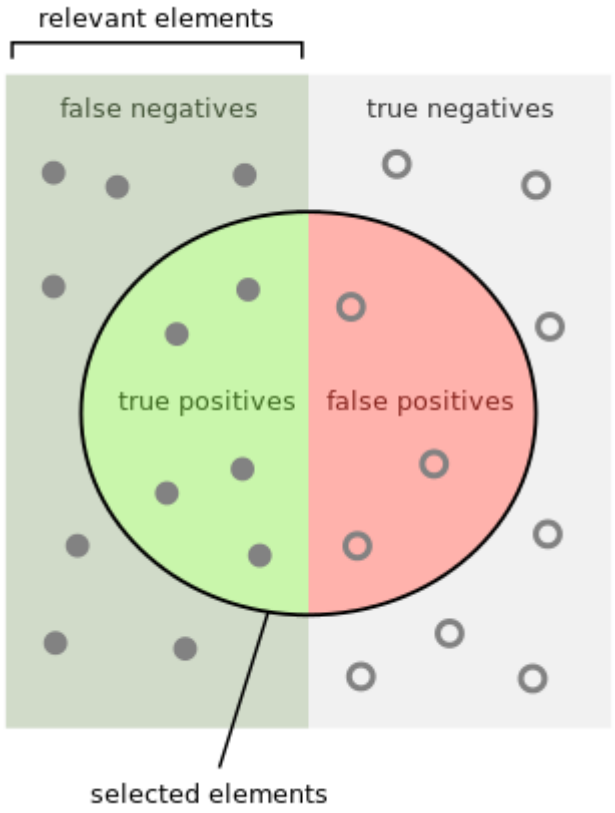
Matemáticas

Biología

Ciencias Sociales

UNIVERSIDAD EAFIT

Métricas de Evaluación



¿Cómo se determinan los datos relevantes?

$$\text{Exactitud} = \frac{\text{true positives} + \text{true negatives}}{\text{relevant elements} + \text{true negatives}}$$

¿Cuántos datos seleccionados son relevantes?

$$\text{Precisión} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

¿Cuántos datos relevantes son seleccionados?

$$\text{Sensibilidad} = \frac{\text{true positives}}{\text{relevant elements}}$$

Métricas de Evaluación

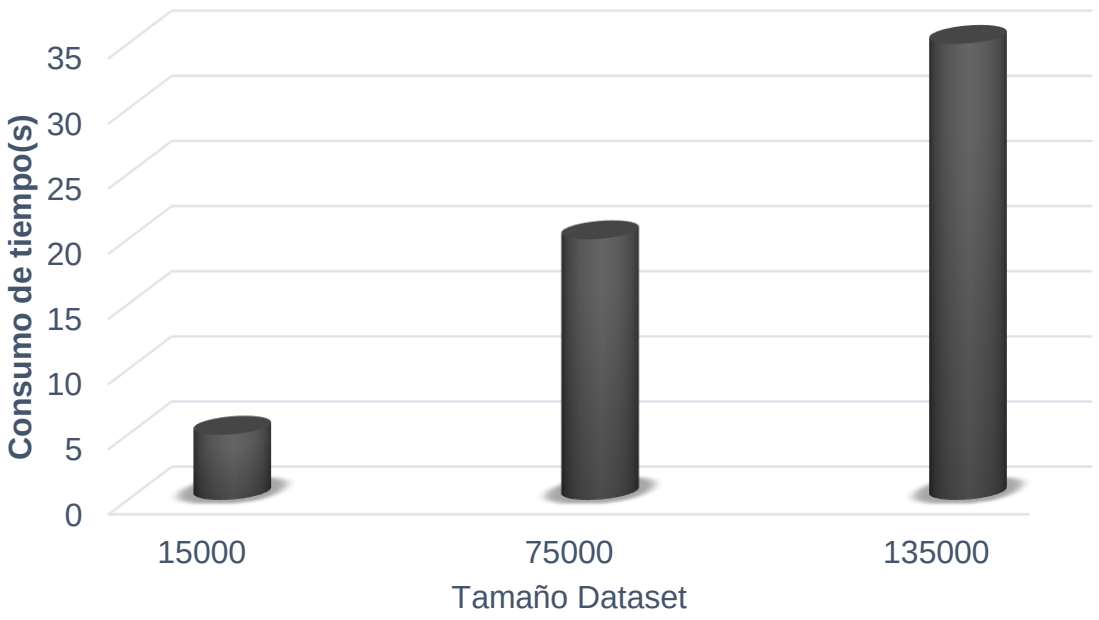
	Conjunto de entrenamiento	Conjunto de validación
Exactitud	0.75	0.78
Precisión	0.78	0.75
Sensibilidad	0.76	0.83

Métricas de evaluación obtenidas con el conjunto de datos de entrenamiento de 135,000 estudiantes y el conjunto de datos de validación de 45,000 estudiantes.

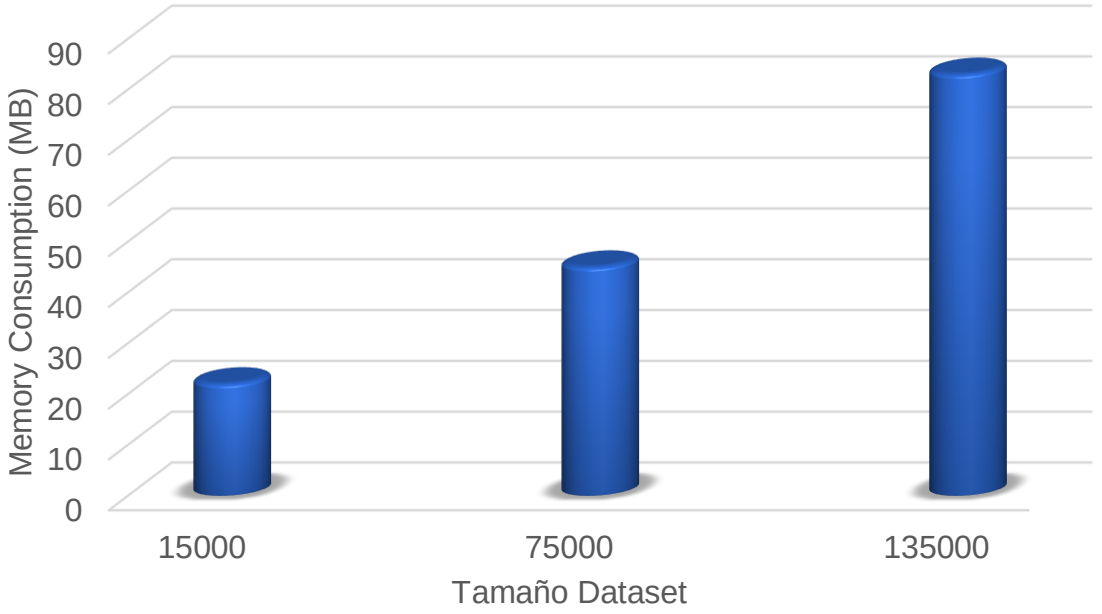
Imagen tomada de Google -> https://www.usbcali.edu.co/sites/default/files/icfes_saper_pro_tablero.png

UNIVERSIDAD EAFIT

Consumo de tiempo y memoria



Consumo de tiempo



Consumo de memoria


¡GRACIAS!