

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



Presentación del Equipo



Juan Pablo
Restrepo



Juan José
Sánchez



Miguel
Correa



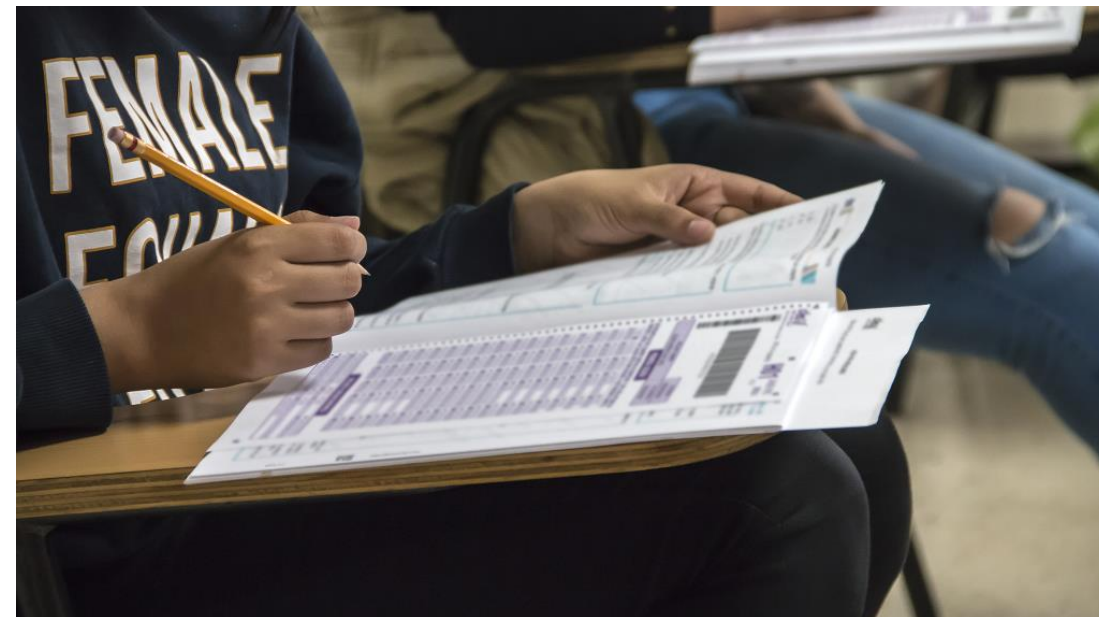
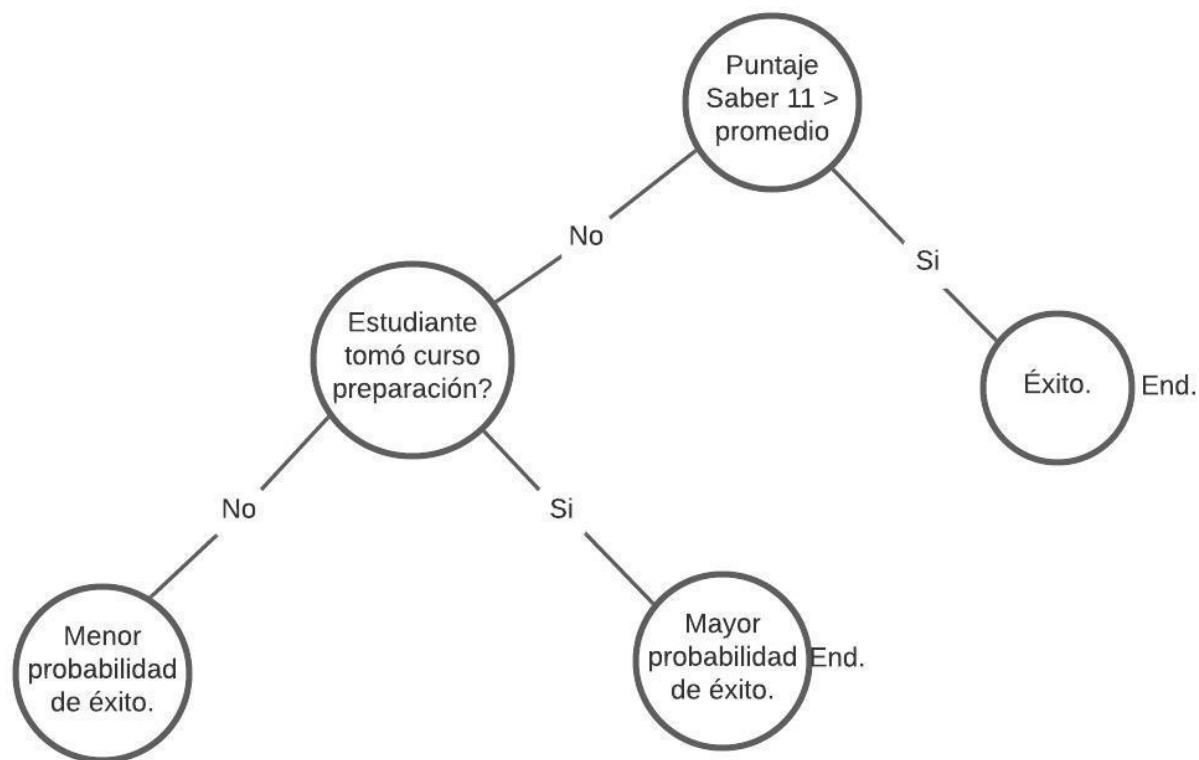
Mauricio
Toro



<https://github.com/jprestrepo/ST0245002/tree/master/proyecto>

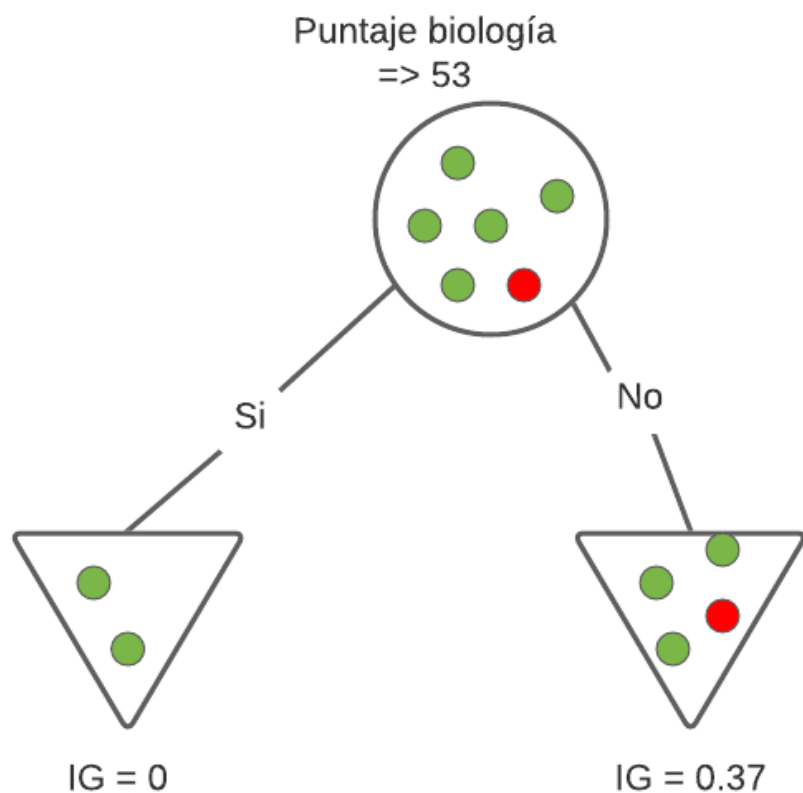


Diseño del Algoritmo

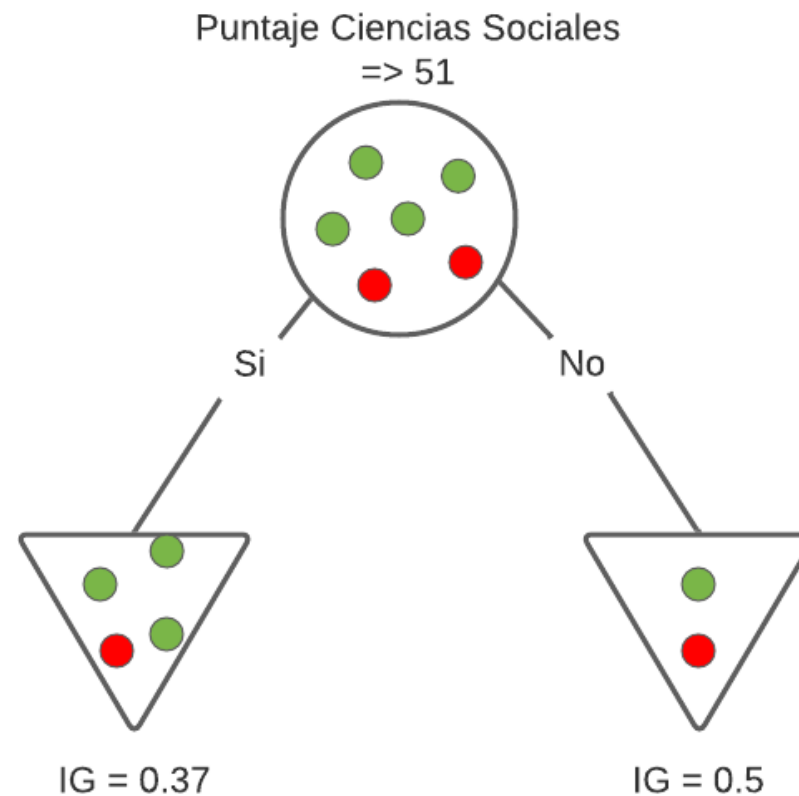


Trabajamos con el algoritmo CART para crear nodos por medio de variables. Utilizando la impureza de Gini podremos generar nodos que determinen más específicamente la probabilidad del éxito para los estudiantes que presenten la prueba Saber Pro.

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



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

Complejidad del Algoritmo

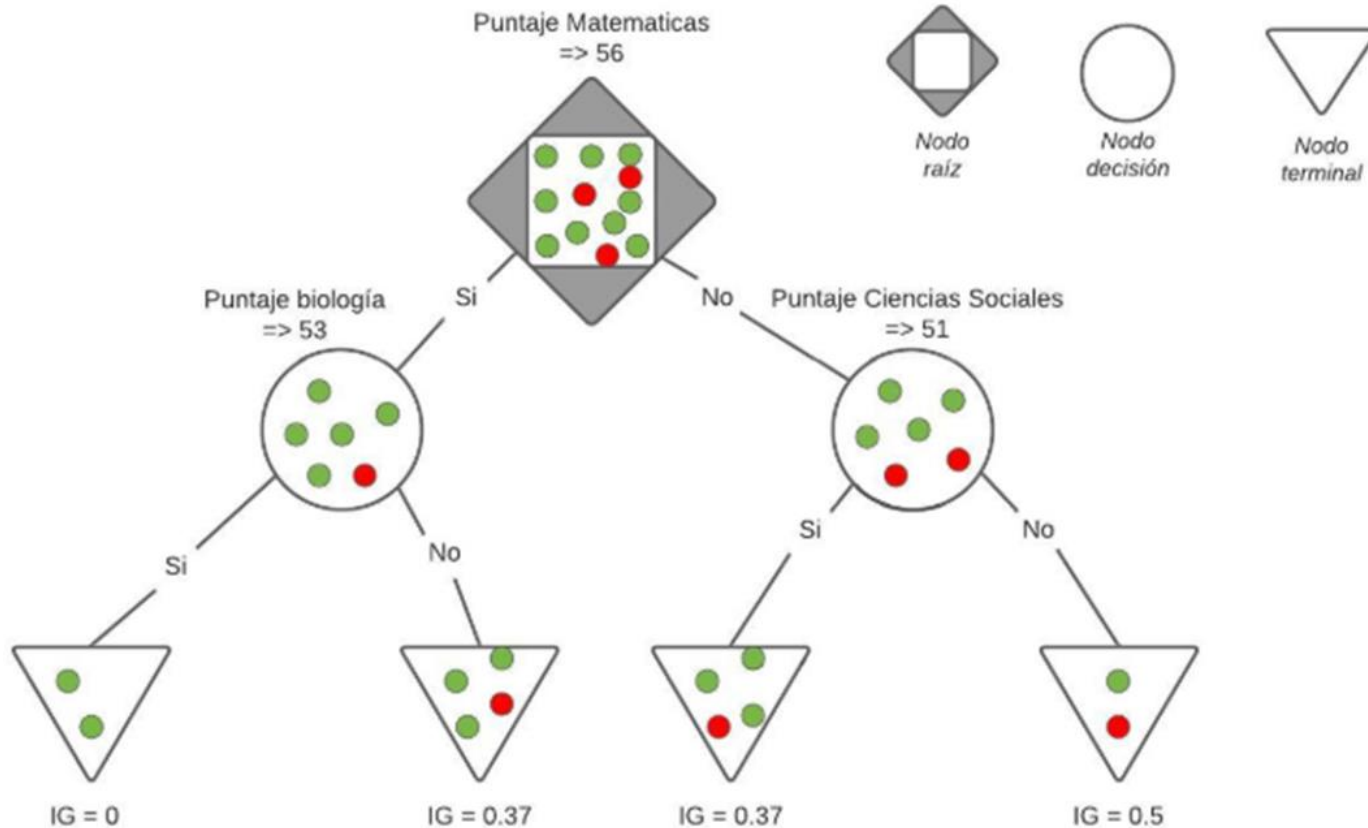


	Complejidad en tiempo	Complejidad en memoria
Entrenamiento del modelo	$O(N^2M \cdot 2^M)$	$O(N \cdot M \cdot 2^M)$
Validación del modelo	$O(N \cdot M)$	$O(1)$

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



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

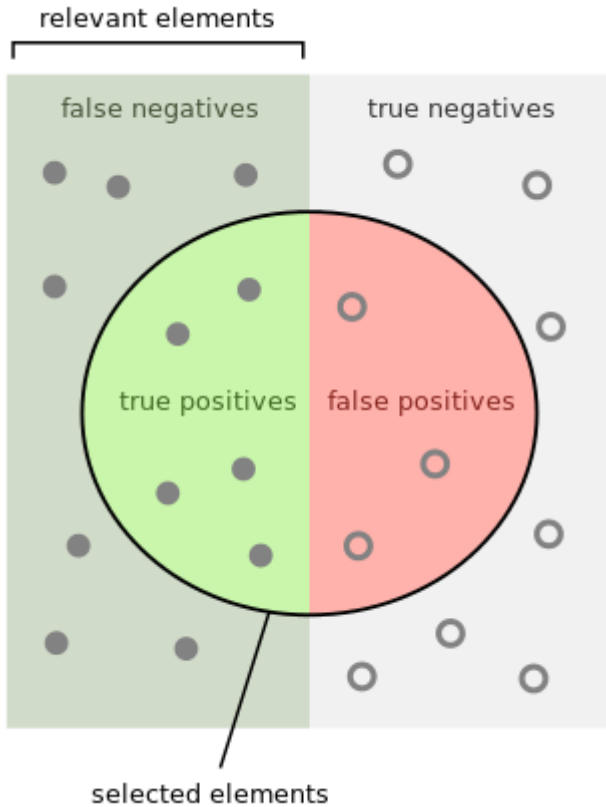


Biología



Ciencias Sociales

Métricas de Evaluación



¿Cómo se determinan los datos relevantes?

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

¿Cuántos datos seleccionados son relevantes?

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

¿Cuántos datos relevantes son seleccionados?

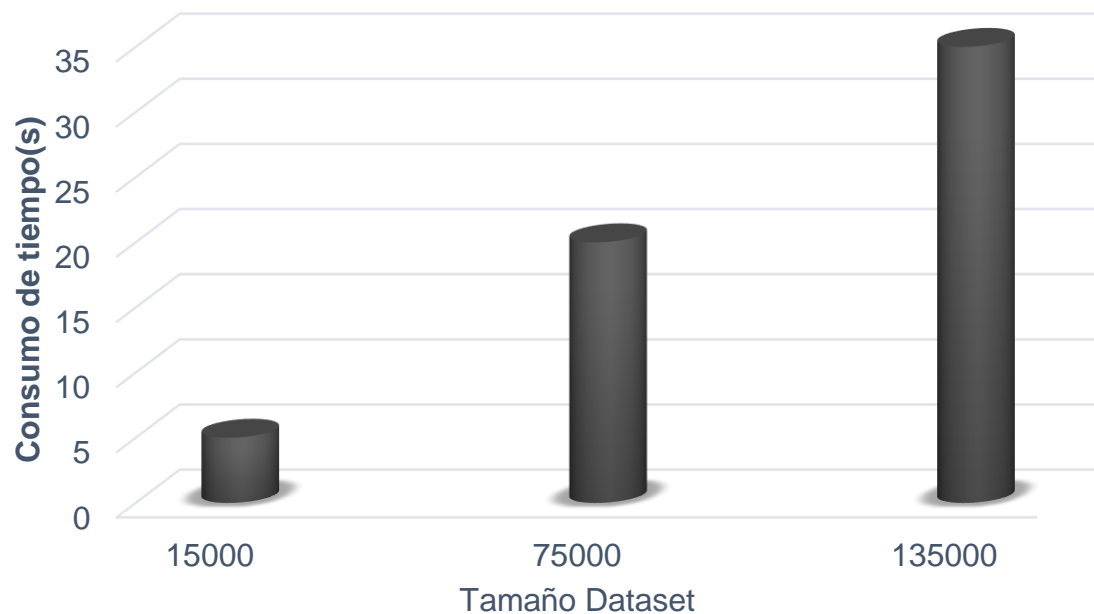
$$\text{Sensibilidad} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

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

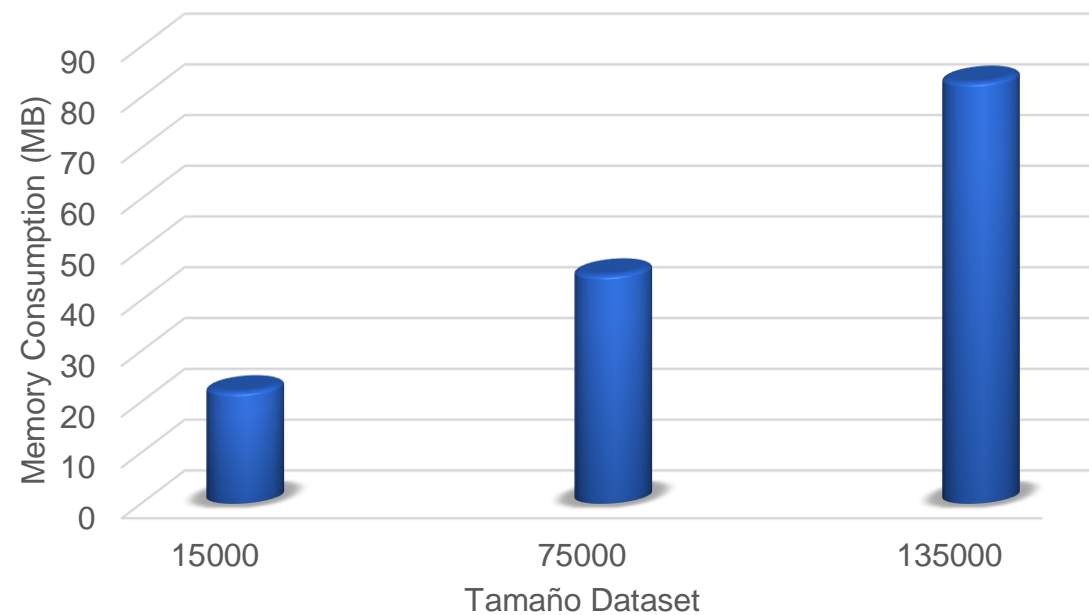


Métricas de evaluación obtenidas con el conjunto de datos de entrenamiento de 135,000 estudiantes y el conjunto de datos de validación de 45,000 estudiantes.

Consumo de tiempo y memoria



Consumo de tiempo



Consumo de memoria



¡GRACIAS!