PREDICCIÓN DE RESULTADOS EN PRUEBAS SABER PRO USANDO DATOS ACADEMICOS Y SOCIALES

UNIVERSIDAD EAFIT

Presentación del Equipo

Miguel Manzur

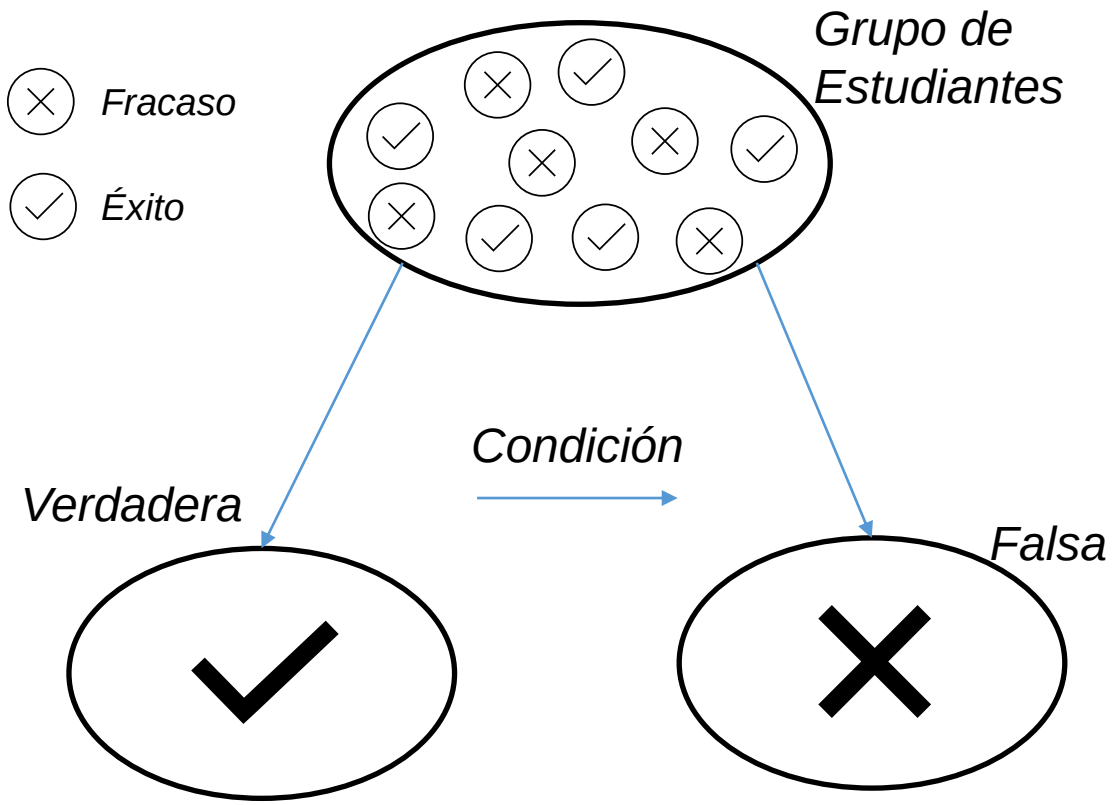
José Aníbal Pinto

Jorge Juan Araujo

Miguel Correa

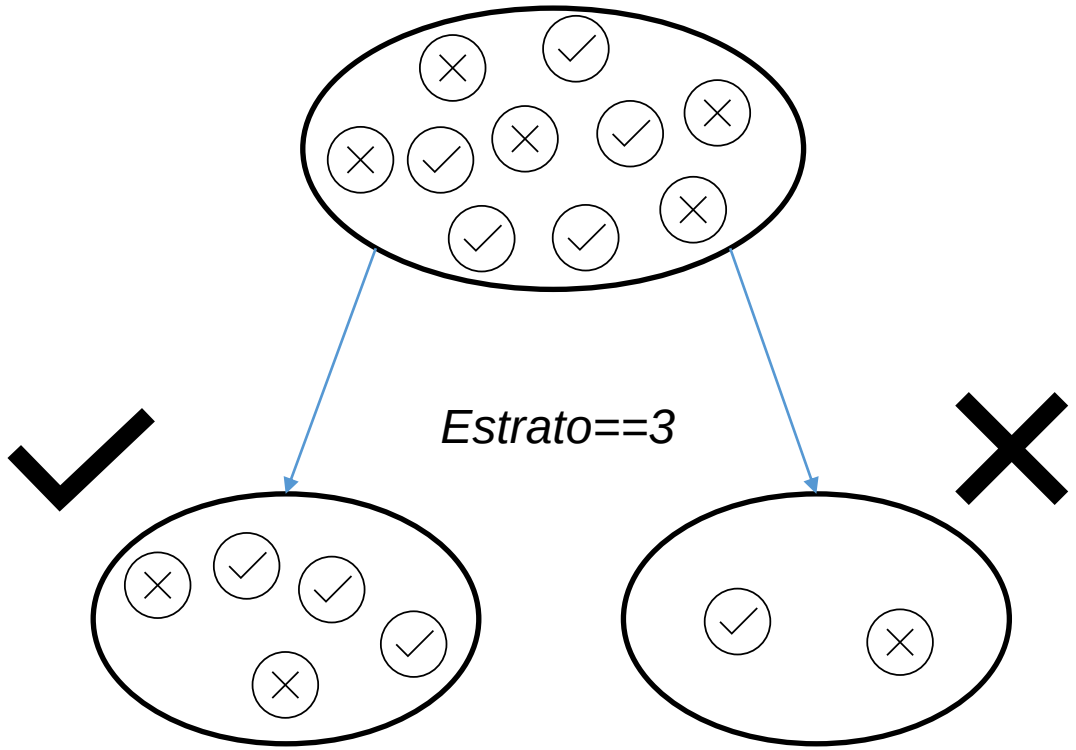
Mauricio Toro

Diseño del Algoritmo

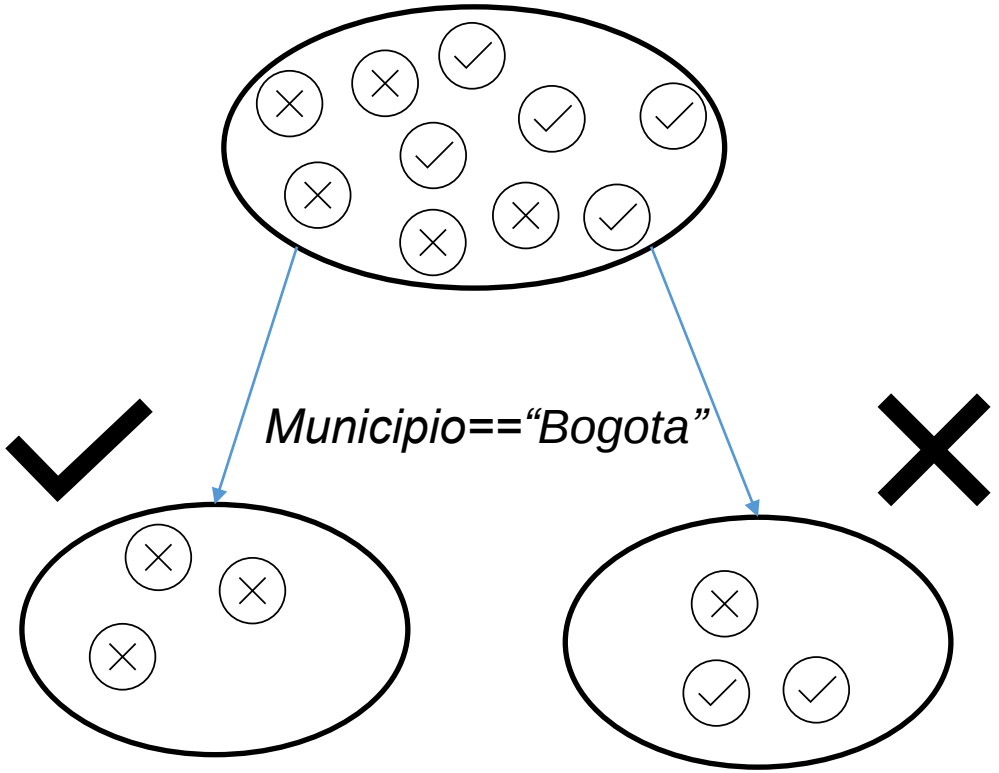


Algoritmo para construir un árbol binario de decisión usando ID3. En este ejemplo, mostramos un modelo para predecir si un estudiante ha tenido buenos resultados (Exito o Fracaso) por lo que buscaremos las condiciones que tengan en común estos estudiantes

División de un nodo



Esta división está basada en la condición “Estrato==3.” Para este caso, la impureza Gini de la izquierda es 0.48, la impureza Gini de la derecha es 0.5 y la impureza ponderada es de 0.49.



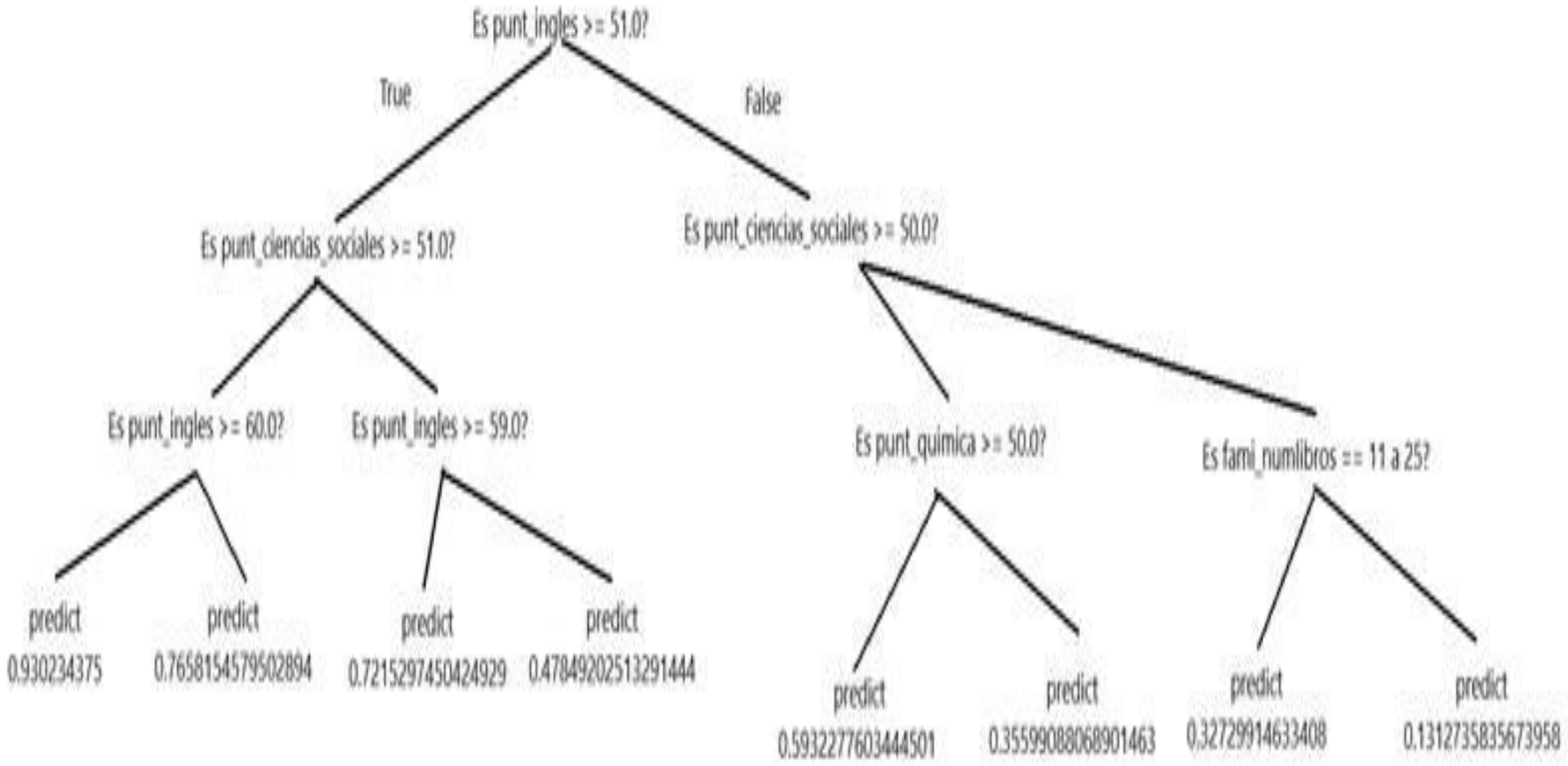
Esta división está basada en la condición “Minucipio==“Bogotá”.” Para este caso, la impureza Gini de la izquierda es 0, la impureza Gini de la derecha es 0.44 y la impureza ponderada es 0.22.

Complejidad del Algoritmo

	Complejidad en tiempo	Complejidad en memoria
Entrenamiento del modelo	O(2^x *n*m)	O(n*m +x)
Validación del modelo	O(Log2 n)	O(n)

Complejidad en tiempo y memoria del algoritmo (En este semestre, una opción puede ser CART. Donde x es el numero de nodos en el árbol, n el numero de columnas y m el número de valores únicos en cada columna.

Modelo de Árbol de Decisión



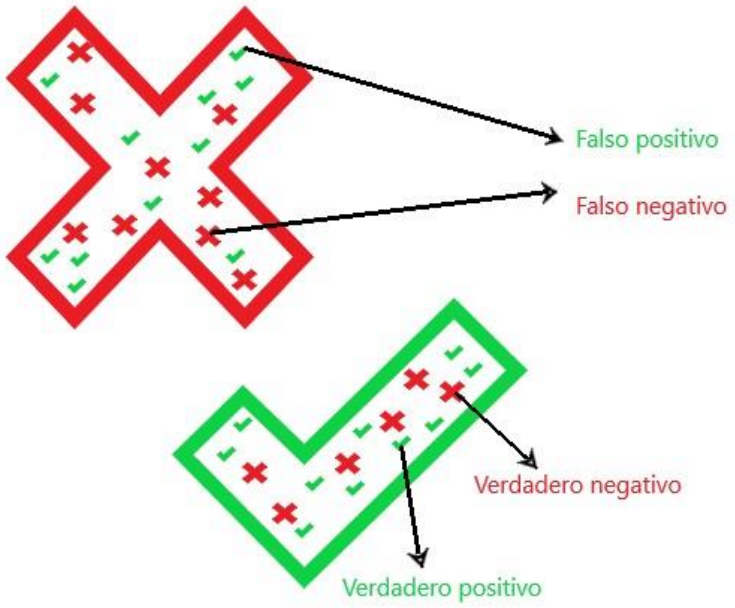
Características Más Relevantes

- Ingles
- Ciencias Sociales
- Quimica
- # de libros

Un árbol de decision para predecir el resultado del Saber Pro usando los resultados del Saber 11. Es un árbol de decisión binario donde los nodos son las preguntas, y cada pregunta divide el marco de datos en dos marcos de datos más pequeños que separan los objetos que cumplen la condición y los que no.

UNIVERSIDAD EAFIT

Métricas de Evaluación



Precisión

$$\frac{Verdadero\ P}{(Verdadero\ P + Falso\ P)}$$

Sensibilidad

$$\frac{Verdadero\ P}{(verdadero\ P + Falso\ N)}$$

Exactitud

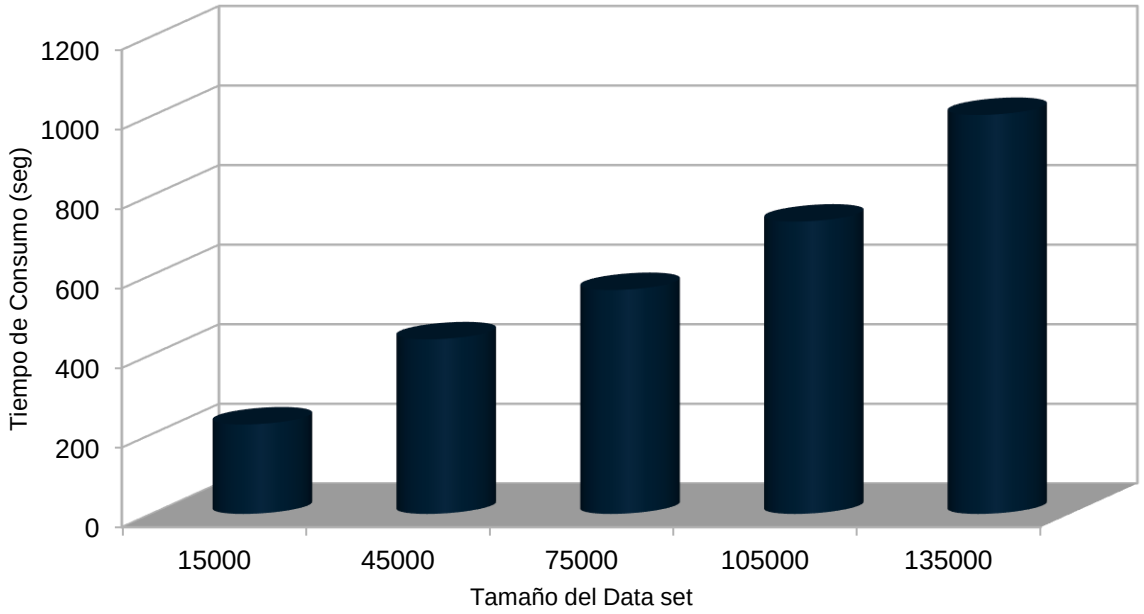
$$\frac{(Verdadero\ P + Verdadero\ N)}{(Verdadero\ P + Verdadero\ N + Falso\ N + Falso\ P)}$$

Métricas de Evaluación

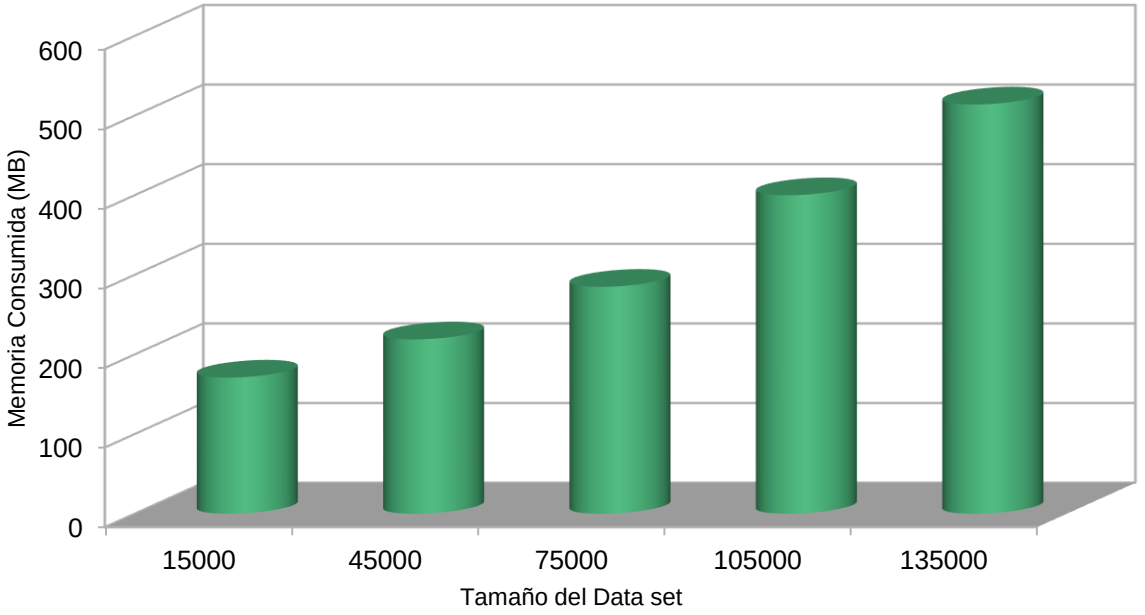
	Conjunto de entrenamiento	Conjunto de validación
Exactitud	80,5%	82%
Precisión	80,3%	80%
Sensibilidad	80,9%	84,6%

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. Este fue el porcentaje que obtenimos según la proporción.

Consumo de tiempo y memoria



Consumo de tiempo



Consumo de memoria


GRACIAS POR SU TIEMPO