PRONOSTICAR EL RESULTADO DEL ICFES UTILIZANDO ARBOLES DE DECISIÓN

Andrés Múnera
Universidad EAFIT
Colombia
ajmunerau@eafit.edu.co

Jorge Daniel Ruiz
Universidad EAFIT
Colombia
jdruizl@eafit.edu.co

Mauricio Toro
Universidad EAFIT
Colombia
mtorobe@eafit.edu.co

RESUMEN
Determinar los factores que determinan los resultados de los estudiantes colombianos en los exámenes institucionales podría ser la clave para la mejoría de la educación en nuestro país. Debido a lo anterior, el objetivo del presente trabajo es presentar un algoritmo que permita ubicar, clasificar y prever los resultados de este grupo de personas en las pruebas, de modo que a partir de los resultados que arroje dicho ejercicio puedan tomarse medidas al respecto.

1. INTRODUCCIÓN

El estudiante promedio en Colombia no sabe leer ni entiende de matemáticas o de ciencias; prueba de ello, los resultados de estos en las pruebas Pisa 2019 [1]. ¿Qué factores determinan estos resultados y cómo trabajar en ellos para mejorar la educación en el país? Conocer las variables que influyen en un buen o un mal resultado del estudiante podría dar lugar a una mejor educación y, por lo tanto, a mejores resultados en pruebas internacionales y aumento de posibles becas y oportunidades para el estudiante.

2. PROBLEMA

¿Cómo anticipar el éxito o el fracaso del estudiante (entendiendo por éxito el haber alcanzado el promedio) en las PRUEBAS SABER PRO, a partir de factores que influyen en su rendimiento (nivel socioeconómico, condiciones familiares, etc.)?

3. TRABAJOS RELACIONADOS

3.1 ID3

Es un algoritmo utilizado en el contexto de la inteligencia artificial, es muy rápido y construye un árbol pequeño.

La estructura de los árboles de decisión está formada por:

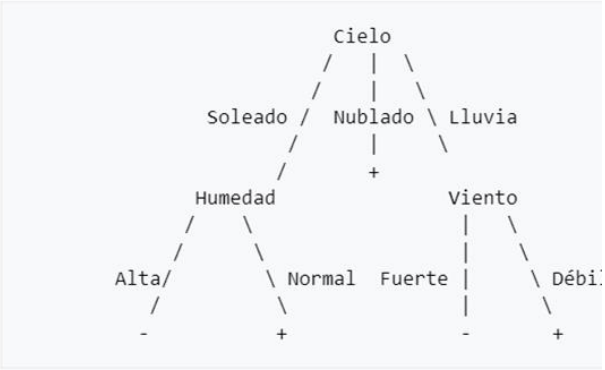
1. **Nodos:** nombres de los atributos.
2. **Ramas:** posibles valores del atributo asociado al nodo.
3. **Hojas:** conjuntos clasificados con el nombre de una clase.

La entropía permite calcular el porcentaje de incertidumbre de una muestra y se calcula así:

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

Ej.	Cielo	Temperatura	Humedad	Viento	Jugar tenis
D1	Sol	Alta	Alta	Débil	-
D2	Sol	Alta	Alta	Fuerte	-
D3	Nubes	Alta	Alta	Débil	+
D4	Lluvia	Suave	Alta	Débil	+
D5	Lluvia	Baja	Normal	Débil	+
D6	Lluvia	Baja	Normal	Fuerte	-
D7	Nubes	Baja	Normal	Fuerte	+
D8	Sol	Suave	Alta	Débil	-
D9	Sol	Baja	Normal	Débil	+
D10	Lluvia	Suave	Normal	Débil	+
D11	Sol	Suave	Normal	Fuerte	+
D12	Nubes	Suave	Alta	Fuerte	+
D13	Nubes	Alta	Normal	Débil	+
D14	Lluvia	Suave	Alta	Fuerte	-

En ese caso el árbol finalmente obtenido sería así:

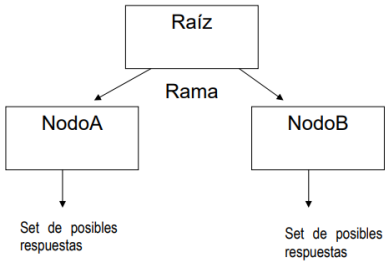


3.2 C4.5

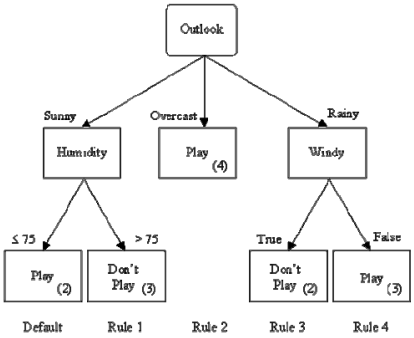
Es uno de los algoritmos más populares de *clasificación basada en reglas*. Es utilizado para resolver *minería de datos*, es decir, puede encontrar patrones repetitivos, tendencias y predecir resultados. El C4.5 se considera un avance del algoritmo ID3 ya que fue desarrollado por el mismo creador años después.

Este genera un árbol mediante particiones de los datos realizadas recursivamente. El algoritmo C4.5 considera todas las posibles pruebas que se pueden dar, divide el conjunto de datos y selecciona la prueba que resulta en la mayor ganancia de información.

• Los nodos de las raíces son los nodos principales de los nodos del árbol. Se consideran todas las muestras y se seleccionan los atributos más importantes.
•La información de muestra es pasada a los nodos subsiguientes, llamados "nodos de la rama", que eventualmente terminan en los "nodos de las hojas", que dan las decisiones.
• Las reglas son generadas mediante la ilustración del camino desde el "nodo de la raíz" hasta el "nodo de la hoja" [3].



Ejemplo aplicado de Árbol de Decisión adaptado para C4.5

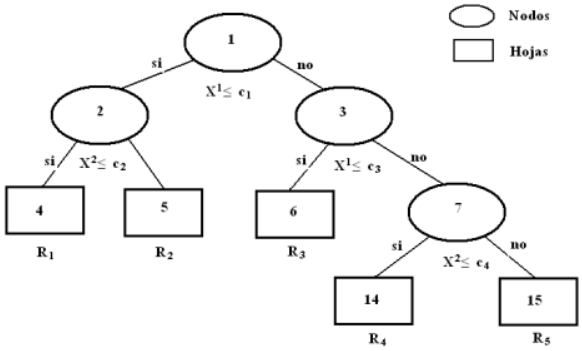


3.3 C5

C5.0 puede generar un árbol de decisión el cual es posible realizar exactamente una predicción para cada registro de datos específico presente en un árbol de decisión.

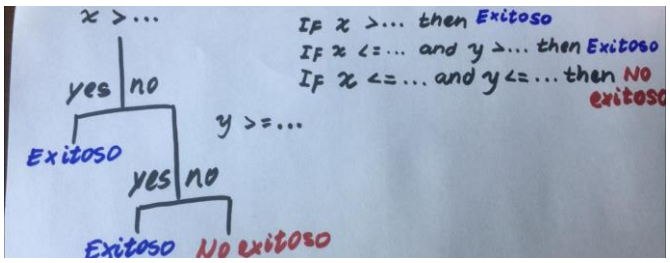
Las distintas submuestras definidas por la primera división se vuelven a dividir, por lo general basándose en otro campo, y el proceso se repite hasta que resulta imposible dividir las submuestras de nuevo. Por último, se vuelven a examinar las divisiones del nivel inferior, y se eliminan o podan las que no contribuyen significativamente con el valor del modelo.

Con esto nos ayudaría significativamente con el proyecto ya que nos ayuda a buscar las posibilidades deseadas.



3.4 CART

Los árboles de clasificación y regresión o también conocidos como la metodología CART, proceden del ámbito de la estadística.
El problema estadístico es establecer una relación entre Y y las X de tal forma que sea posible predecir Y basado en los valores de las X., se quiere estimar la probabilidad condicional de la variable aleatoria Y,
P [Y = y|x1, x2, . . . , xp]
cuando la variable Y es discreta, o un funcional de su probabilidad tal como la condicional
E [Y |x1, x2, . . . , xp].
cuando la variable Y es continua. Esto nos ayudaría a trazar una línea mas directa entres las posibilidades y encontrar patrones de decisiones



REFERENCIAS

1. Colprensa. Colombia, con la peor nota de la Ocde en pruebas Pisa. *El Colombiano*.
2. J.R.Quinlan, Induction of Decision Trees, *Machine Learning*, 1986, pp81-106.
3. Mazid, M., Ali, S. and Tickle K. Improved C4.5 Algorithm for Rule Based Classification. *Central Queensland University.* 296-301
4. Díaz, J. and Correa, J. Comparación entre ´arboles de regresión CART y regresión lineal. *Universidad Santo Tomás.* Vol. 6, No. 2, 175–195