PREDICCIÓN DEL ÉXITO ACADÉMICO EN EDUCACIÓN SUPERIOR UTILIZANDO ÁRBOLES DE DECISIÓN

Miguel Ángel Sarmiento Aguiar
Universidad EAFIT
Colombia
msarmie4@eafit.edu.co

Camilo Villa Tamayo
Universidad EAFIT
Colombia
cvillat@eafit.edu.co

Marlon Pérez Ríos
Universidad EAFIT
Colombia
mperezr@eafit.edu.co

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

RESUMEN
El objetivo de este informe es mostrar una predicción al puntaje de los estudiantes de pregrado en las universidades colombianas en las pruebas saber pro y mejorar lo que se podría sacar en ellas, ya que esto servirá para aconsejar a los jóvenes antes de tomar la decisión de escoger una carrera, y así salgan más preparados para presentarla y ayuden al desarrollo del país aumentando el nivel académico de este a nivel mundial.

Se busca poder predecir el posible éxito de la población estudiantil en algún programa de educación superior, y con esta predicción, aconsejar respecto a una decisión posiblemente más acertada.

1. INTRODUCCIÓN

En la actualidad el nivel de las pruebas saber pro en la población de jóvenes estudiantes de pregrado es baja, con respecto a otras universidades del mundo, debido a que la mayoría al finalizar la educación secundaria no sabe en qué área poseen mejores competencias, por esto toman una decisión con mayor probabilidad de ser errónea respecto a la probabilidad de éxito. Casi en la totalidad de los casos dicha decisión se toma basándose en el gusto por alguna profesión o por el deseo de una gran remuneración monetaria.

Por esto se pretende mostrar la metodología por la cual se va a desarrollar un algoritmo basado en árboles de decisión, con el fin de predecir el posible éxito de los jóvenes en las diferentes carreras universitarias, reflejándose en las pruebas saber pro, todo esto en base a la información de variables sociodemográficas y de las pruebas saber 11 suministrada por el gobierno colombiano en los últimos años.

2. PROBLEMA

El problema a resolver es la creación de un algoritmo basado en árboles de decisión y en los datos suministrados por el ICFES, mencionados anteriormente. La solución a este problema podrá predecir si el puntaje en las pruebas Saber Pro estará o no por encima del promedio, teniendo en cuenta que los datos del ICFES no solo incluyen variables académicas como el puntaje sino también variables sociodemográficas como género, edad, información de los padres, estrato, entre muchas otras.

Así, se pretende disminuir el nivel de deserción de los estudiantes de pregrado y aumentar los puntajes en la prueba saber pro, guiándolos por la opción en la que tienen mayor probabilidad de éxito, así se puede ahorrar una gran cantidad de recursos tanto para ellos como para el gobierno, añadiendo que con dicha predicción se puede aumentar el interés y las ganas por el estudio, incrementando el número de posibles profesionales en el país.

3. TRABAJOS RELACIONADOS

3.1 Algoritmo ID3

Se basa en la búsqueda de hipótesis dado un conjunto de ejemplos de datos, estos deben estar conformados por una serie de listas ordenadas de valores, cada valor es denominado atributo, uno de estos es el atributo a clasificar (objetivo), el cual es de tipo binario.

La elección del mejor atributo se establece mediante la entropía, eligiendo aquel que proporcione una mejor ganancia de información. Para buscar la hipótesis se utiliza un árbol de decisión que mediante nuevas instancias dirá si el ejemplo dado va a ser positivo o negativo.

Sus elementos son:

- Nodos: contienen los atributos.
- Arcos: contienen valores posibles del nodo padre.

- Hojas: nodos que clasifican el ejemplo como positivo o negativo.

Como ejemplo se propone predecir si es buena idea jugar o no tenis dependiendo de atributos como temperatura, humedad, entre otros:

Estado	Humedad	Viento	Juego tenis
¿	Alta	Leve	No
Soleado	Alta	Fuerte	No
Nublado	Alta	Leve	Si
Lluvia	Alta	Leve	Si
Lluvia	Normal	Leve	Si
Lluvia	Normal	Fuerte	No
Nublado	Normal	Fuerte	Si
Soleado	Alta	Leve	No
Soleado	Normal	Leve	Si
Lluvia	Normal	Leve	Si
Soleado	Normal	Fuerte	Si
Nublado	Alta	Fuerte	Si
Nublado	Normal	Leve	Si
Lluvia	Alta	Fuerte	Si

Figura 1 - Ejemplo ID3

El árbol de decisión sería:

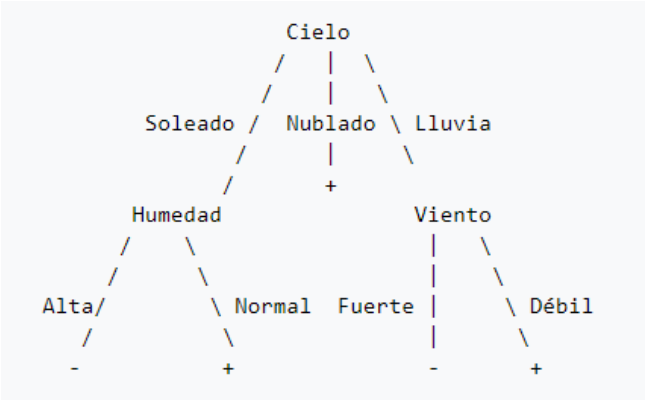


Figura 2 - Árbol de decisión ID3
[1]

3.2 Algoritmo C4.5

Este algoritmo construye árboles de decisión desde un grupo de datos de entrenamiento usando el concepto de entropía de información, donde dichos datos son ejemplos ya clasificados. Cada ejemplo es un vector con los atributos o características de este. Los datos de entrenamiento son aumentados con otro vector cuyos componentes representan la clase a la que pertenece cada muestra.

En cada nodo del árbol elige un atributo de los datos que más eficazmente dividen el conjunto de muestras en subconjuntos enriquecidos en una clase, su criterio es el normalizado para ganancia de información que resulta en la elección de un atributo para dividir los datos. El atributo con la mayor ganancia de información normalizada se elige como parámetro de decisión. [2]

Como ejemplo se toma el mismo caso del algoritmo ID3, donde la distribución de datos para el atributo Estado es:

	Desconocido	Soleado	Nublado	Lluvia
No	1	2	0	1
Si	0	2	4	4
Totales	1	4	4	5

Figura 3 - Ejemplo C4.5

Tomando la división de los datos para el valor Nublado, Lluvia y Soleado del atributo Estado:

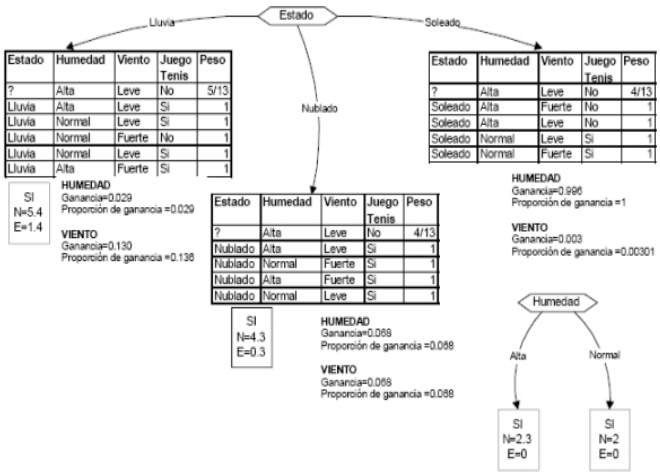


Figura 4 - Árbol de decisión C4.5
[3]

3.3 Algoritmo Random Forest

Este algoritmo consiste en un gran número de árboles de decisión individuales que operan como un conjunto, para así mejorar el rendimiento de la predicción. Cada árbol individual genera un tipo de predicción y el que tenga mejor desempeño es el que va a ser el modelo de predicción de todo el Random Forest.

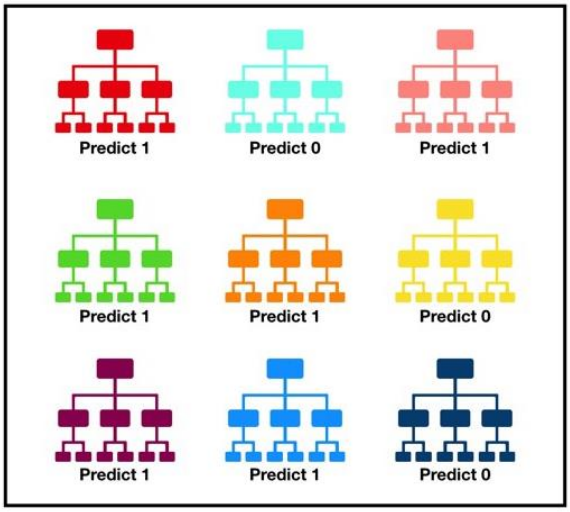


Figura 5 - Algoritmo Random Forest

El aspecto principal del Random Forest es el conocimiento de cada parte individual. Cada árbol individual no es tan potente como cuando se juntan todos los árboles, aún sin estar relacionados o con poca relación, todos se complementan, y si hay algún error en alguno de los individuales, los otros lo pueden corregir, y este es una de las principales partes claves de este algoritmo. Esto es como una empresa que está dividida por diferentes áreas, que muchas no están muy relacionadas entre sí, pero cuando todas se juntan crean una empresa que da grandes resultados.

Algunos prerrequisitos para que funcione bien son:

1- Que haya una señal real para que las predicciones no sean aleatorias sino acertadas.

2- Las predicciones hechas por los árboles individuales deben tener bajas correlaciones entre sí. [4]

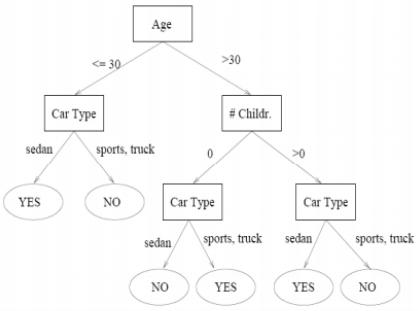
3.4 Algoritmo CART

Es una técnica de aprendizaje de árbol de decisión no paramétrica que produce árboles de clasificación o regresión, dependiendo de si la variable dependiente es categórica o numérica, respectivamente. La palabra binario implica que un nodo en un árbol de decisión solo puede dividirse en dos grupos. CART utiliza el índice de Gini como medida de impureza para seleccionar el atributo. El atributo con la mayor reducción de impurezas se utiliza para dividir los registros del nodo. CART acepta datos con valores numéricos o categóricos y también maneja valores de atributos faltantes. Utiliza la poda de complejidad de costos y también genera árboles de regresión.

El análisis de árboles de clasificación y regresión (CART) generalmente consiste en tres pasos:

1. Construcción del árbol máximo.
2. Poda del árbol.
3. Selección del árbol óptimo mediante un procedimiento de validación cruzada.

Como ejemplo se supone el árbol y los datos en la Figura 6, donde se quiere determinar un conjunto de reglas que indiquen si un conductor vive o no en los suburbios.



Car type	Age	Chd	Sub
sedan	23	0	yes
sports	31	1	no
sedan	36	1	no
truck	25	2	no
sports	30	0	no
sedan	36	0	no
sedan	25	0	yes
truck	36	1	no
sedan	30	2	yes
sedan	31	1	yes
sports	25	0	no
sedan	45	1	yes
sports	23	2	no
truck	45	0	yes

Figura 6 - Ejemplo CART

Se concluye:

- Si Age ≤ 30 y CarType = Sedan entonces Si
- Si Age ≤ 30 y CarType = truck/Sports entonces No
- Si Age > 30, Children = 0 y CarType = Sedan entonces No
- Si Age > 30, Children = 0 y CarType = truck/Sports entonces Si
- Si Age > 30, Children > 0 y CarType = Sedan entonces Si
- Si Age > 30, Children > 0 y CarType = truck/Sports entonces No

[5]

REFERENCIAS

1. Wikipedia. (2019). Algoritmo ID3. [online] Available at: https://es.wikipedia.org/wiki/Algoritmo_ID3

2. Wikipedia. (2020). C4.5. [online] Available at: https://es.wikipedia.org/wiki/C4.5

3. López, B. (2005). Algoritmo C4.5. [ebook] Available at: http://www.itnuevolaredo.edu.mx/takeyas/Apuntes/Inteligencia%20Artificial/Apuntes/tareas_alumnos/C4.5/C4.5(2005-II-B).pdf

4. Yiu, T. (2019). Understanding Random Forest. [online] Towards data science. Available at: https://towardsdatascience.com/understanding-random-forest-58381e0602d2

5. Serna, S. (2009). Comparación de árboles de regresión y clasificación y regresión logística. Maestría. Universidad Nacional de Colombia.