Algoritmo de predicción para la consecución universitaria por medio de árboles de decisión

Juan Felipe Lopez Gutierrez
Universidad Eafit
Colombia
jflopezg@eafit.edu.co

David Restrepo Ramirez
Universidad
Colombia
drestrepor@eafit.edu.co

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

RESUMEN

El contexto social Colombiano es amplio y plural y a su vez este mismo logra ser parte clave del espectro de decisiones posibles que una persona puede tomar a lo largo de su vida. Muchas de ellas se pueden clasificar y ordenar por medio de un estudio de la misma persona, en la mayoría de casos pudiendo incluso llegar a predecir comportamientos o situaciones futuras. Se agrega al estudio el fenómeno estudiantil de alta deserción dentro de la educación superior en el país, esta es una problemática grave dado que se suma y relaciona de manera integral a la poca recepción de jóvenes con la oportunidad de ingresar a la educación universitaria de calidad y causando dentro de la sociedad un déficit de graduados en muchas de las ramas del conocimiento.

Es allí donde se entra entonces al estudio de variables disponibles y ver si se es posible lograr hacer una correlación entre esos datos y la posibilidad de que un estudiante cualquiera pueda tener éxito o no dentro del ámbito universitario.

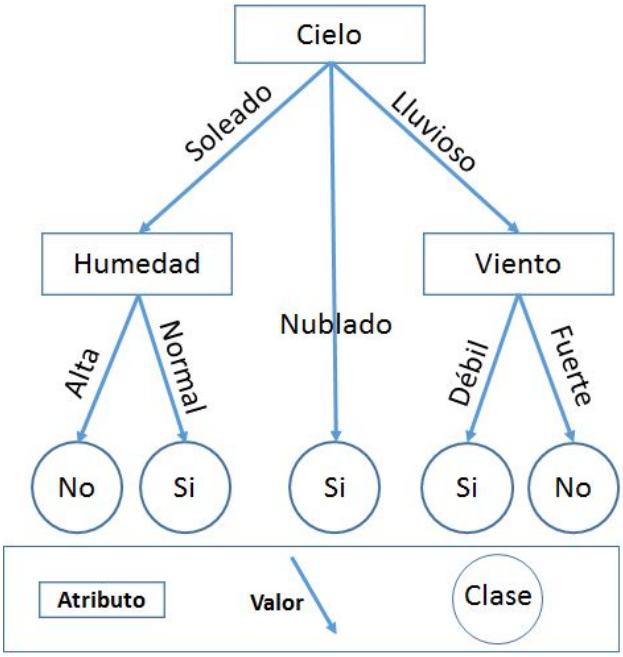
1. INTRODUCCIÓN

El documento busca abordar una solución ante una situación tan preocupante dentro del sistema educativo como lo es la deserción o fracaso durante y después de los estudios universitarios y asi segun los datos generales del estudiante ser capaces de dar una predicción oportuna por medio de un algoritmo y así guiar de manera más óptima el futuro de los jóvenes colombianos dentro del sistema de educación. Además de ello el documento dará un vistazo a problemas relacionados en el contexto de las estructuras de datos.

La herramienta que usaremos para abordar este problema, será la implementación de árboles de decisión, dadas sus ventajas con respecto a otras estructuras de datos; como por ejemplo el "aprendizaje automático" del software; puesto que de este modo el mismo podrá obtener respuestas más rápidas cuando se cumpla alguna de las condiciones. Además, la implementación de este tipo de estructuras no requiere de tanta memoria, ni tanta capacidad de cómputo como sí lo hacen las redes neuronales.

Pero, ¿Qué es un árbol de decisión?
Un árbol de decisión es un modelo de predicción en el que, dado un conjunto de datos, estos se categorizan en base a un criterio "objetivo", de forma que podamos tomar una decisión u otra dependiendo de la situación.

El siguiente es un ejemplo de árbol de decisión:



Imágen tomada de: https://lh3.googleusercontent.com/proxy/NUg_Edc0I6mL8wkLirtXWJGhaWxQCAyho11U5mIqh-o5mQg9e15vvj3LbQ-oHEY-djCATAxOFBrAH9XZvuJVPQsnHozIpZ_i-caesgfDGWDcrPVoi7dWAlI

En el anterior ejemplo, tenemos algunos atributos con distintas posibilidades; dependiendo de la posibilidad con la que nos encontremos, tomaremos un camino u otro hasta tomar una decisión.

Finalmente se dará a explicar el algoritmo, su estructura, formulación y justificación el cual con los datasets otorgados buscar dar de la manera mas optima un solución a la predicción de la deserción por medio del uso de árboles de decisión.

2. PROBLEMA

El problema base a solucionar es, lograr encontrar una predicción sobre el futuro éxito de un estudiante en base al

estudio de variables planteadas en su evaluación de pruebas saber. Gracias a la gran variedad de datos pertenecientes a diferentes ámbitos de los que se rodea el estudiante en cuestión, es posible lograr la resolución de un acercamiento virtual a la futura realidad del mismo por medio de la aplicación de un algoritmo. Sería de gran ayuda la implementación de este tipo de solución para la deserción y así poder guiar de manera más óptima el futuro de la educación superior Colombiana, ayudando a disminuir los preocupantes porcentajes registrados de deserción que logran llegar al 42% en el país (1). Haciendo uso del algoritmo se podría hacer de manera anticipada una correlación entre los datos del estudiante y así predecir su éxito, para así asegurar lo más posible el éxito del estudiante y de manera simultánea de la sociedad Colombiana, teniendo como marco de consecución y logro dentro de sus estudios estar sobre la media de su correspondiente prueba saber pro.

3. TRABAJOS RELACIONADOS

3.1 ID3

Este es uno de los algoritmos pioneros en árboles de decisión. Funciona organizando jerárquicamente los atributos que se tengan, y creando nodos en ellos; siendo los primeros o "mejores" los que tienen menor entropía. Así entonces, lo que hace es dar una respuesta en base a los datos que tenga, primero evaluando el caso en los atributos con menor entropía (para ver si puede obtener una respuesta inmediata) y luego acudiendo a los que siguen en el orden jerárquico; esto, repitiendolo por todo el árbol de decisión.

3.2 C4.5

Es la evolución del ID3, pero con algunas mejoras que optimizan el uso de la memoria. Se diferencia de su antecesor en que, puede crear varios nodos de un mismo atributo. Además, "poda" las ramas del árbol que no le proporcionan información extra, como por ejemplo si todas las alternativas en un atributo son iguales.

Adicionalmente, es de más fácil implementación, dado que tiene pocos casos base.

3.3 CART

Este algoritmo tiene la capacidad de crear árboles de clasificación, para obtener respuestas binarias (positivo/negativo o sí/no), y también, la capacidad de crear árboles de regresión (nos pueden arrojar resultados más relacionados con porcentajes). Se desarrolló por la misma época que el ID3, pero de forma paralela. Su funcionamiento radica en dividir los datos en subconjuntos lo más binarios posible, rigiéndose por la variable de clase. Esto lo hace aplicando una fórmula para hallar el coeficiente de impureza de Gini. Luego, se arroja una respuesta en base al conjunto del que hace parte.

Este algoritmo tiene algunas desventajas contra el C4.5; una de ellas es el uso de la memoria, dado que en el C4.5 se obtiene una gestión más óptima de la memoria. Sin embargo, nos da cierta flexibilidad con respecto al tipo de datos que se pueden arrojar, lo que lo hace un buen candidato para la implementación de nuestra práctica.

3.4 CHAID

Este también es un algoritmo de clasificación. Busca hallar criterios de identificación o diferenciación de conjuntos, en base a una variable categórica (que los separe en dos grupos). De los criterios hallados, elige el mejor en base a lo mucho que pueden diferenciarse en base a la variable categórica los conjuntos que generan estos criterios. Esto, lo hace con una fórmula. Podría decirse que tiene un funcionamiento similar al CART. Es un algoritmo muy usado en estadística

(1)https://www.elespectador.com/noticias/educacion/el-problema-no-es-solo-plata-42-de-los-universitarios-deserta-articulo-827739

(2)https://www.youtube.com/watch?v=J3QwOjSVH8k

(3)https://es.wikipedia.org/wiki/Aprendizaje_basado_en_árboles_de_decisión

(4)https://es.wikipedia.org/wiki/C4.5

(5)https://es.wikipedia.org/wiki/Algoritmo_ID3