Predicción del Éxito Académico Utilizando Arboles de Decisión

Santiago Montoya Tobon
Ingeniería de Sistemas
Universidad EAFIT
Medellín Colombia
smontoyat@eafit.edu.co

Nelson A. Barrios Jiménez
Ingeniería de Sistemas
Universidad EAFIT
Medellín Colombia
nabarriosj@eafit.edu.co

Mauricio Toro Bermúdez (Docente)

RESUMEN

En este documento explicaremos nuestra manera de abordar el proyecto elegido para la materia Estructuras de Datos y Algoritmos l basándonos en problemas ya solucionados, explicando los problemas y las soluciones ya creadas.

1. INTRODUCCION

En la actualidad se utilizan algoritmos para cada cosa, algunos que nos ayudan a predecir, otros que nos ayudan a optimizar tareas diarias. A través del paso del tiempo hemos encontramos como sociedad maneras de agilizar y simplificar procesos con el uso de las tecnologías emergentes.

En los últimos años se ha trabajado especialmente en las Inteligencias Artificiales[1], actualmente tienen distintos usos entre los cuales están: la conducción autónoma, la ayuda de software fotográfico, predicción comercial, búsqueda de malware, mejora en el realismo de los videojuegos, traducción y automatización de fábricas, entre otros.

Con el desarrollo de la IA día a día aparecen nuevos problemas a solucionar nuevas funciones que cumplir y nuevas tareas que optimizar y ahí está la razón de este texto.

En la actualidad Colombia cuenta con la segunda mayor deserción de Latinoamérica, la cobertura de educación persona en el país ronda el 57% de jóvenes entre los 17 y 24 años se estima que el 42% de estos terminan desertando.

2. PROBLEMA

El problema consiste en predecir el éxito[2] universitario de una persona x, utilizaremos variables predictoras como su puntaje en las pruebas icfes saber 11 y variables socioeconómicas

De este modo entrenaremos el programa ingresándole múltiples datos para conseguir predicciones cada vez más precisas.

1 de aquí en adelante será nombrada IA

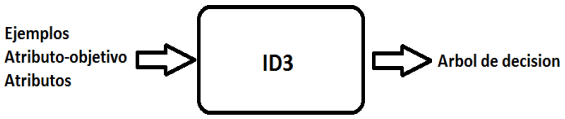
2 la condición de éxito será obtener un puntaje mayor al promedio

3. PROBLEMAS SIMILARES

En esta sección expondremos problemas relacionados que han sido solucionados por distintos algoritmos ya creados, explicaremos su funcionamiento y lo mostraremos gráficamente para facilitar su entendimiento.

3.1 ALGORITMO ID3

El algoritmo ID3 fue creado por J. R. Quinlan en la universidad de Sídney se basa en el algoritmo "Concept Learning System" (CLS). Es un algoritmo creado para crear arboles de decisión a partir de un conjunto fijo de ejemplos, el árbol resultante es usado para clasificar muestras futuras.

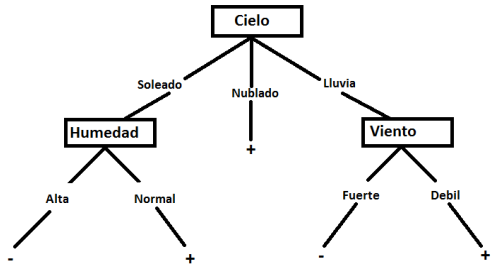


3.1.1 Árbol de decisión

Para demostrar la manera en que este algoritmo crea los árboles utilizaremos un ejemplo[1].

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

Para estos datos el árbol de decisión creado será el siguiente



Como vemos en el árbol el algoritmo toma un dato principal u atributo-objetivo y de ahí empieza a sacar sus ramas.

3.2 ALGORITMO C4.5

El algoritmo C4.5 es una extensión del algoritmo ID3 desarrollado por J. R. Quinlan los datos ingresados para entregar el algoritmo son grupos de ejemplos que ya han sido clasificados. El algoritmo elige un atributo que sea el mas eficaz a la hora de dividir el conjunto dado en subconjuntos centrados en una clase

3.2.1 PSEUDOCODIGO

Utilizaremos pseudocodigo[2] para ayudar a la comprensión de este algoritmo.

```
1    comprobar casos base
2    para cada atributo a
3        encontrar ganancia de la división a
4    a_best atributo con ganancia más alta
5    crear nodo para dividir a_best
6    repetir en las listas obtenidas
```

1 sacado de: https://www.cise.ufl.edu/~ddd/cap6635/Fall-97/Short-papers/2.htm

2 sacado de: http://www2.cs.uregina.ca/~dbd/cs831/notes/ml/dtrees/c4.5/tutorial.html

3.3 ALGORITMO C5

J. R. Quinlan continuo con la creación del C5.0 esta versión implementa las siguientes mejoras:

·la nueva versión mejora significativamente la velocidad en comparación al algoritmo C4.5

·el algoritmo C5.0 es mas eficiente que el algoritmo C4.5

·el algoritmo c5.0 obtiene resultados parecidos al C4.5 con arboles mas compactos

·se agrega soporte para boosting que mejora los arboles y les otorga mas precisión

·al agregar la opción automática "Winnowing" se permite aplicar un algoritmo de clasificación (Winnow) a los atributos para eliminar aquellos que sean de poca ayuda.

El argoritmo se basa en extraer patrones de los datos ingresados para mejorar la predicción en situaciones futuras[1]

1 sacado de https://www.rulequest.com/see5-info.html