Predicción del Éxito Académico Utilizando Arboles de Decisión

Santiago Montoya Tobon
Universidad EAFIT
Medellin Colombia
smontoyat@eafit.edu.co

Nelson A. Barrios
Universidad EAFIT
Medellin Colombia
nabarriosj@eafit.edu.co

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

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.

Palabras clave
Estructura de datos; arboles de decisión; predicción; inteligencia artificial; notación O

Palabras clave de la ACM
Computing methodologies → Artificial intelligence
Applied computing → Operations research → Decision analysis
Computing methodologies → Symbolic and algebraic manipulation → Symbolic and algebraic algorithms
Mathematics of computing → Discrete mathematics → Graph theory → Trees
Computing methodologies → Machine learning → Machine learning argorithms

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 (IA), actualmente estas tienen distintos usos entre los cuales están: la conducción autónoma, la ayuda de software fotográfico, la predicción comercial, búsqueda de malware, mejora en el realismo grafico en entornos como el videojuego, 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 universitario de una persona x, utilizaremos variables predictoras como su puntaje en las pruebas icfes saber 11 y variables socioeconómicas

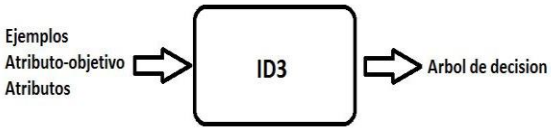
De este modo entrenaremos el programa ingresándole múltiples datos para conseguir predicciones cada vez más precisas.

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.

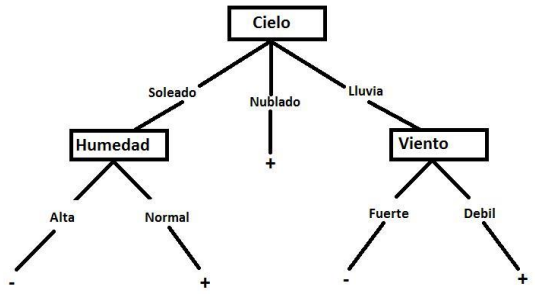


3.1.1 Árbol de decisión

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 más 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
```

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 más eficiente que el algoritmo C4.5

·el algoritmo c5.0 obtiene resultados parecidos al C4.5 con árboles más compactos

·se agrega soporte para boosting que mejora los árboles y les otorga más 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 algoritmo se basa en extraer patrones de los datos ingresados para mejorar la predicción en situaciones futuras

4. ARREGLO DE ARREGLOS

Para solucionar el problema la estructura de datos que decidimos utilizar es un arreglo de arreglo (también llamado matriz) (fig.1), de modo que, el arreglo mas extenso va a ser (por lógica) el conjunto de personas y cada elemento del arreglo de personas corresponde a una característica de ellos

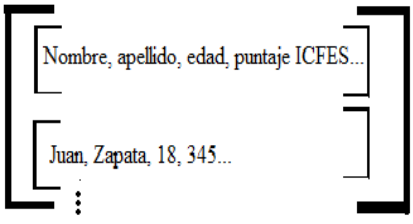


Fig.1: En la matriz cada una de las personas funciona como un arreglo que contiene sus características.

4.1 OPERACIÓNES EN LA ESTRUCTURA

Para solucionar el problema necesitamos que la estructura sea eficiente y para eso debe tener unas operaciones básicas para poder alterar nuestra estructura las cuales son:

4.1.1 Operación append

La primera y la fundamental es aquella que nos permite leer los datos y agregarlos a nuestra matriz. Básicamente nos permite cargar nuestros datos y organizarlos en la matriz (Fig.2).

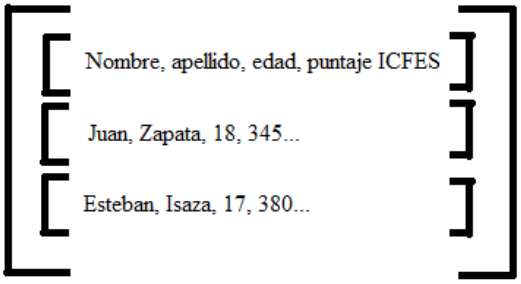
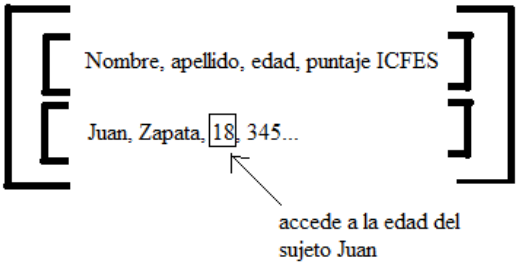


Fig.2: Luego de ejecutar el comando append se agrega al final de la lista el nuevo ítem

4.1.2 Operación Buscar

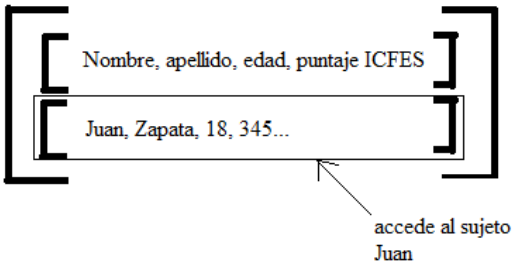
La función recibe como parámetros un arreglo f y un elemento a. La función compara las columnas con el elemento a y retorna el valor de un atributo de una persona.



En este caso se accede a la edad del sujeto f

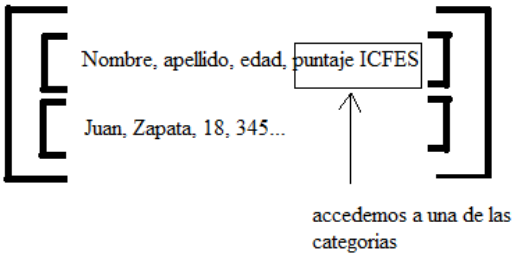
4.1.3 Operación Persona

La función recibe como parámetros un arreglo f y un elemento a, la función nos permitirá ingresar al arreglo de la persona a



4.1.4 Operación Categoría

La función recibe como parámetros un arreglo f y dos elementos b, c, la función nos permite acceder a uno de los elementos en el arreglo 0 de la matriz



4.2 ELECCION DE ESTRUCTURA DE DATOS

Para poder resolver el problema de manera eficiente no solo necesitamos unas operaciones fundamentales si no que, necesitamos que estas sean lo mas eficientes posibles para que los tiempos de ejecución no sean absurdamente largos.

Es por esta razón que decidimos utilizar la matriz, ya que, como lo mencionamos antes la matriz puede ser clasificada como un arreglo de arreglo y esto significa que acceder a un elemento será O(1), lo cual quiere decir que tomara un tiempo constante al tomar esa operación.

4.3 ANALISIS DE COMPLEJIDAD

Esta estructura de datos fue elegida ya que los tiempos de ejecución de las operaciones básicas no es muy elevado, debido a que la complejidad de acceder a un elemento será O(1), la complejidad de añadir un archivo será O(m*n) y la complejidad de buscar será O(n) (donde m se refiere al numero de personas y n a la cantidad de características de la persona).

4.4 TIEMPOS DE EJECUCION

Operaciones	Data set 1 (s)	Data set 2 (s)	Data set 3 (s)	Data set 4 (s)
Append	0.289	0.382	0.874	1.52
Buscar	1.82e-5	7.5e-6	2.32e-5	8.82e-6
Persona	6.02e-6	2.82e-6	5.73e-6	4.86e-6
Categoria	2.54e-5	9.12e-6	9.34e-6	8.52e-6

Tiempo conseguido en el peor de los casos

4.5 USO DE MEMORIA

	Data set 1	Data set 2	Data set 3	Data set 4
Consumo de memoria	71MB	81MB	102MB	138MB

REFERENCIAS

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

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

Casas Mogollon, P. 2018. El problema no es solo plata: 42% de los universitarios deserta | ELESPECTADOR.COM

Eduardo Morales, Hugo Jair Escalante
https://ccc.inaoep.mx/~emorales/Cursos/NvoAprend/Acetatos/sbl.pdf

Esther Vaati
https://ccc.inaoep.mx/~emorales/Cursos/NvoAprend/Acetatos/sbl.pdf