

## ÁRBOLES DE DECISION EN PRUEBAS DE ESTADO

Juan José Wilches Rivas Universidad Eafit Colombia jjwilchesr@eafit.edu.co	Juan José Zuluaga Bedoya Universidad Eafit Colombia jjzuluagab@eafit.edu.co	Miguel Correa Universidad Eafit Colombia macorream@eafit.edu.co	Mauricio Toro Universidad Eafit Colombia mtorobe@eafit.edu.co
---	--	--	--

### RESUMEN

En este documento se abarcará el problema de un posible bajo desempeño y poder, a base de datos y de variables por medio de arboles de decision, solucionarlo, ya que por este método se cuenta con un índice que dará a muestra datos y podrán salir soluciones o conclusiones base a estos datos.

La importancia del problema es lograr un equilibrio basado en un buen desempeño estudiantil y poder predecir lo que pasara en menor tiempo, tomando datos y posibles finales después de relacionarlos.

Si tomamos la variación de desempeño nacional se puede ver claramente que todo tiene un por que. En esto podemos definir las razones que se pueden dar, entre esto podemos decir que influye valores o variables como la edad, nivel de pobreza, o incluso estrato o lugar de vivienda.

### Palabras clave

Árboles de decisión, aprendizaje automático, éxito académico, predicción de los resultados de los exámenes

### 1. INTRODUCCIÓN

Si la predicción de datos y resultados se afianza los jóvenes podrán aspirar por medio de corrección de hábitos a estudiar mas con el fin de buscar mejores e resultados y así poder aspirar a grandes becas o incluso a mejores oportunidades de trabajo, desde aquí comienza un camino basado en el mejoramiento de una vida partiendo de la decision de la persona.

#### 1.1. Problema

En este semestre planeamos, con una predicción de datos ajustar mas a la realidad del futuro que pasará, a base de esto se puede predecir y cambiar mejor los resultados, por ejemplo, si se sabe que una persona no le ira bien y uno de los factores principales es las horas de estudio se le puede aconsejar que incremente este valor para que sus posibilidades de un buen desempeño aumenten.

#### 1.3 Estructura del artículo

En lo que sigue, en la sección 2, presentamos el trabajo relacionado con el problema. Más adelante, en la sección 3, presentamos los conjuntos de datos y métodos utilizados en esta investigación. En la sección 4, presentamos el diseño del algoritmo. Después, en la sección 5, presentamos los resultados. Finalmente, en la sección 6, discutimos los

resultados y proponemos algunas direcciones de trabajo futuras.

### 2. TRABAJOS RELACIONADOS

A continuación, se encontrarán diferentes artículos relacionados con la predicción académica utilizando arboles de decisión.

#### 2.1 Árboles de decisión para predecir desempeño académico Saber 11°

El objetivo de este trabajo fue desarrollar un sistema de árboles de decisión para predecir los patrones relacionados con el desempeño de estudiantes en grado once que presentan las pruebas ICFES, todo esto mediante el algoritmo J48[1] 366 - 376.

Como resultado se obtuvo que los principales factores de atributos que influyeron en el éxito o no de los estudiantes fueron, el estrato socioeconómico medio o alto, la jornada de estudio en la mañana o completa, el índice TIC regular y la edad menor que 18 años. Gracias a estos atributos el algoritmo logró responder con un precisión de un 65% [1] 372.

#### 2.2 Modelo predictivo de deserción estudiantil basado en arboles de decisión

El objetivo fundamental de este artículo fue realizar un modelo basado en el árbol de decisión el cual fuera capaz de determinar la probabilidad de que un estudiante abandone la universidad teniendo en cuenta su rendimiento académico y su entorno personal, basándose principalmente en un modelo CART.[2]

Como resultados se obtuvo que aquellos entre el sexto y el décimo curso son menos propensos a retirarse de sus estudios, también aquellos que tengan un promedio de notas entre 4.0 – 10.0 son menos propensos a desertar. [2]

#### 2.3 Modelo predictivo para la determinación de causas de reprobación mediante Minería de Datos

El objetivo principal de este trabajo fue describir las principales causas de reprobación en las diferentes materias de la universidad politécnica de puebla, dando uso de aspectos actuales y pasados tales como, historial académico, problemas personales y psicológicos. Así gracias al uso de árboles de decisión basados en el algoritmo C4.5, puede predecir la reprobación en diferentes materias, es importante resaltar que los resultados no fueron homogéneos, 5 de las nueve materias dan una certeza entre el 80% – 100%, por otro lado, las 4 restantes oscilan entre 33% y 76%.

## 2.4 Modelo de decision para estudiantes de educación superior en Perú

Debería mencionar el problema que resolvieron, el algoritmo que usaron, la precisión que lograron y la citación.

Se busca con un modelo de árbol de decision identificar variables y datos para el impacto y solución de el bajo desempeño de los estudiantes en la educación superior; el algoritmo fue basado en la recolección de datos de jóvenes de diferentes instituciones y ciudades del país, y entre los datos reunidos se incluyeron variables como nivel de pobreza y se utilizo un modelo de árbol con nodos incluyendo el chaid y quest, que se ve reflejado en la tabla a continuación.

Característica	CHAID	CHAID Exhaustivo	C&RT	QUEST
Tipo de Partición	Múltiple	Múltiple	Binaria	Binaria
Dependiente Continua	Si	Si	Si	No
Predictoras Continuas	Si (*)	Si (*)	Si	Si
Coste de Mala Clasificación (Crecimiento del Arbol)	No	No	Si	Si
Pruebas Estadísticas (Selección del Predictor)	Si	Si	No	Si
Pruebas Estadísticas (Particionar)	Si	Si	No	No
Velocidad	Moderada	Moderada	Lento	Moderada/Lento
Utiliza A priori?	No	No	Si	Si
Valores Faltantes para los Predictores Usados?	Si, como una categoría	Si, como una categoría	No, Sustitutos usados para partición	No, Sustitutos usados para partición

En base a los datos que marca la tabla anterior se hizo el estudio de los datos sacando diferentes tes porcentajes de variables, pero con una acertividad al menos del 80% y así definiendo que incluso algunas variables definían que tanto estudiaría una persona en número de semestres.

## 3. MATERIALES Y MÉTODOS

En esta sección se explica cómo se recopilaron y procesaron los datos y, después, cómo se consideraron diferentes alternativas de solución para elegir un algoritmo de árbol de decisión.

### 3.1 Recopilación y procesamiento de datos

Obtuvimos datos del *Instituto Colombiano de Fomento de la Educación Superior* (ICFES), que están disponibles en línea en <ftp.icfes.gov.co>. Estos datos incluyen resultados anonimizados de Saber 11 y Saber Pro. Se obtuvieron los resultados de Saber 11 de todos los graduados de escuelas secundarias colombianas, de 2008 a 2014, y los resultados de Saber Pro de todos los graduados de pregrados colombianos, de 2012 a 2018. Hubo 864.000 registros para Saber 11 y 430.000 para Saber Pro. Tanto Saber 11 como Saber Pro, incluyeron, no sólo las puntuaciones sino también datos socioeconómicos de los estudiantes, recogidos por el ICFES, antes de la prueba.

En el siguiente paso, ambos conjuntos de datos se fusionaron usando el identificador único asignado a cada estudiante. Por lo tanto, se creó un nuevo conjunto de datos que incluía a los estudiantes que hicieron ambos exámenes estandarizados. El tamaño de este nuevo conjunto de datos es de 212.010 estudiantes. Después, la variable predictora binaria se definió de la siguiente manera: ¿El puntaje del estudiante en el Saber Pro es mayor que el promedio nacional del período en que presentó el examen?

Se descubrió que los conjuntos de datos no estaban equilibrados. Había 95.741 estudiantes por encima de la media y 101.332 por debajo de la media. Realizamos un submuestreo para equilibrar el conjunto de datos en una proporción de 50%-50%. Después del submuestreo, el conjunto final de datos tenía 191.412 estudiantes.

Por último, para analizar la eficiencia y las tasas de aprendizaje de nuestra implementación, creamos al azar subconjuntos del conjunto de datos principal, como se muestra en la Tabla 1. Cada conjunto de datos se dividió en un 70% para entrenamiento y un 30% para validación. Los conjuntos de datos están disponibles en <https://github.com/mauriciotoro/ST0245-EaFit/tree/master/proyecto/datasets>.

	Conjunto de datos 1	Conjunto de datos 2	Conjunto de datos 3	Conjunto de datos 4	Conjunto de datos 5
Entrenamiento	15,000	45,000	75,000	105,000	135,000
Validación	5,000	15,000	25,000	35,000	45,000

**Tabla 1.** Número de estudiantes en cada conjunto de datos utilizados para el entrenamiento y la validación.

### 3.2 Alternativas de algoritmos de árbol de decisión

#### 3.2.1 ID3

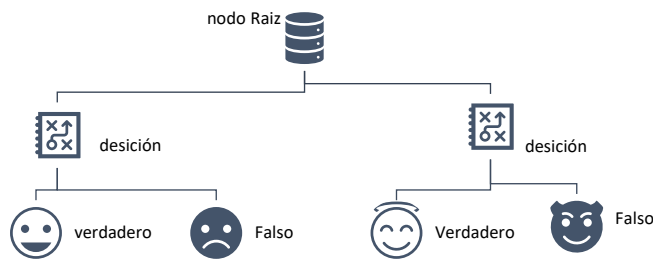
Este algoritmo se base en un proceso de arriba hacia abajo analizando los atributos que dan mejor ganancia. Este algoritmo toma el concepto de entropía

$$Entropia(S) = - \sum_{i=1}^m p_i \log_2 p_i$$

A partir de este el concepto de ganancia

$$Ganancia(S,A) = Entropia(S) - \sum_{v(a)} \frac{|S_v|}{|S|} Entropia(S_v)$$

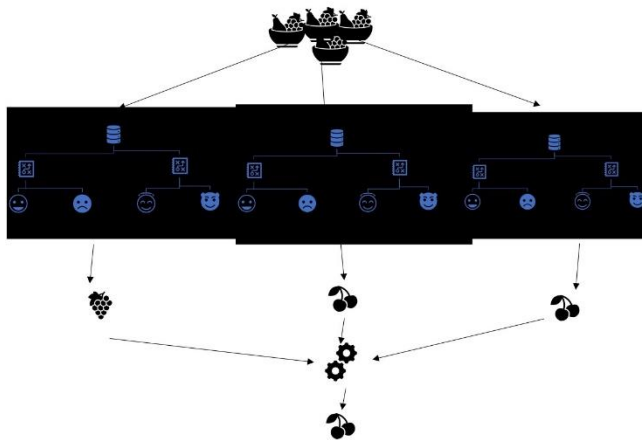
Los que hace el algoritmo es analizar la viabilidad de cada algoritmo a la hora de crear cada nodo aquel nodo atributo que dé más ganancia en el nodo será el elegido.



### 3.2.2

#### Random Forest

Este algoritmo se basa en la creación de diferentes árboles, encargados cada uno de un subconjunto del conjunto inicial,

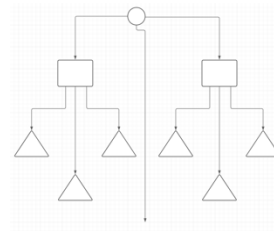


lo que se busca es que cada uno de estos árboles elija un dato a partir del subconjunto aquel dato que haya sido escogido por más árboles será tomado como resultado.

### 3.2.3 CHAID

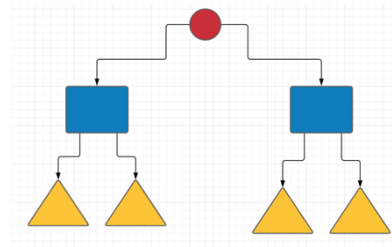
Consiste en generar arboles de decision con estadísticas para buscar unas divisiones de optimización únicas, por otro lado, es diferente a Quest y a C&RT, ya que chaid puede generar arboles de decision no binarios, lo que simboliza que algunas de las divisiones pueden generar mas de dos ramas, y los

campos de entrada y objetivo pueden ser numéricos o continuos y categóricos. Requiere más tiempo para realizar los cálculos. Busca tener más facilidad al hacer estudio y realización de datos, pero esto llevara mas tiempo.



### 3.2.4 QUEST

Proporciona método de clasificación binaria, para él creación de arboles de decision, y se creo principalmente para reducir el tiempo de procesamiento, necesario para realizar el análisis de datos y registros tomados en nodos como C&RT, para favorecer las entradas y que se permitan realizar mas divisiones,y su objetivo es que los rangos sean categóricos, aunque también se pueden rangos numéricos, todas las divisiones son binarias.



### REFERENCIAS

1. Timarán-Pereira, R., Caicedo-Zambrano, J., & Hidalgo-Troya, A., Árboles de decisiones para predecir factores asociados al desempeño académico de estudiantes de bachillerato en las pruebas saber 11°. Rev.investig. desarro.innov., 9 (2), 363-378.
2. Cuji, B., Gavilanes, W., Sánchez, R. Modelo predictivo de deserción estudiantil basado en arboles de decisión. Revista Espacios, 38(55), 17-25.
3. Rodallegas Ramos E., Torres González A., Torres González B., Modelo predictivo para la

determinación de causas de reprobación mediante Minería de Datos, Universidad Tecnológica de Puebla.

4. Quora, what are the differences between ID3, C4.5 and CART? Chinmay Pradhan, December 20, 2016, <https://www.quora.com/What-are-the-differences-between-ID3-C4-5-and-CART>
5. Yangli-ao, G., Ming L., Qingyong L. & Ruisi H. Introduction of machine learning, y The Institution of Engineering and Technology, y, London, United Kingdom, 2019.
6. [http://www.cladea.org/proceeding-2018/pdf/papers/Estrategia/CLADEA\\_2018\\_paper\\_109.pdf](http://www.cladea.org/proceeding-2018/pdf/papers/Estrategia/CLADEA_2018_paper_109.pdf) (references complete of participants Page 13-16)
7. [https://www.ibm.com/support/knowledgecenter/es/SS3RA7\\_sub/modeler\\_mainhelp\\_client\\_ddit\\_a/clementine/nodes\\_treebuilding.html](https://www.ibm.com/support/knowledgecenter/es/SS3RA7_sub/modeler_mainhelp_client_ddit_a/clementine/nodes_treebuilding.html)