# ÁRBOLES DE DECISION EN PRUEBAS DE ESTADO


| Juan José Wilches Rivas | Juan José Zuluaga Bedoya | Miguel Correa | Mauricio Toro |
|---|---|---|---|
| Universidad Eafit | Universidad Eafit | Universidad Eafit | Universidad Eafit |
| Colombia | Colombia | Colombia | Colombia |
| jjwilchesr@eafit.edu.co | jjzuluagab@eafit.edu.co | macorream@eafit.edu.co | 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 árboles de decisión, 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 qué. 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 más 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 decisión de la persona.

### 1.1. Problema

En este semestre planeamos, con una predicción de datos ajustar más 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.

### 1.2 Solución

En este trabajo nos centraremos en una herramienta para la predicción de resultados como lo son lo arboles de decisión, como lo dice Aswath Damodaran los árboles de decisión también son útiles, porque no sólo permiten considerar el riesgo en cada una de las etapas, sino que te ayudan a diseñar la mejor respuesta, dado un resultado determinado (si ocurre x, habría que hacer z). Vincular acciones y opciones a los resultados de eventos inciertos, a través de árboles de decisión [8]. Evitamos los métodos de caja negra como las redes neuronales y las máquinas de soporte vectorial ya estas no permiten un desarrollo en medio de la incertidumbre último, la red neuronal sencillamente carece de información suficiente como para operar con un mínimo de precisión o no está modelada teniendo en cuenta las particularidades de la realidad que intenta capturar [9].

Este proyecto se basa en arboles decisión el cual toma diferentes datos sobre un proceso en específico y mediante una serie de preguntas divide los datos en aquellos que cumplen y aquellos que no.

### 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 em 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 cuanta 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 decisión 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 decisión identificar variables y datos para el impacto y solución del 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 utilizó un modelo de árbol con nodos incluyendo el chaid y quest, que se ve reflejado en la tabla a continuación.

| Característica | CHAID | CHAID Exahustivo | C&RT | QUEST |
|---|---|---|---|---|
| Tipo de Partición | Múltiple | Múltiple | Binaria | Binaria |
| Dependiente Continua | Sí | Sí | Sí | No |
| Predictoras Continuas | Si (*) | Si (*) | Sí | Sí |
| Coste de Mala Clasificación (Crecimiento del Arbol) | No | No | Sí | Sí |
| Pruebas Estadísticas (Selección del Predictor) | Sí | Sí | No | Sí |
| Pruebas Estadísticas (Particionar) | Sí | Sí | No | No |
| Velocidad | Moderada | Moderada | Lento | Moderada/Lento |
| Utiliza A priori? | No | No | Sí | Sí |
| 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 asertividad 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

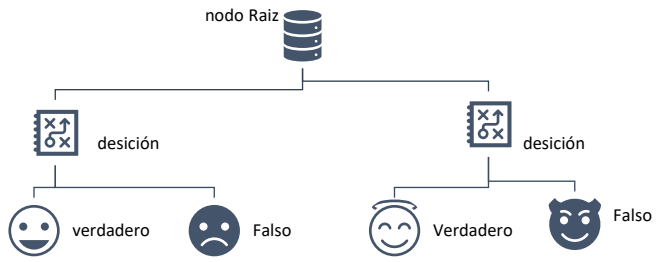
Este algoritmo se base en un proceso de arriba hacia abajo analizando los atributos que dan mejor ganancia. Este algoritmo toma el concepto de entropía

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

A partir de este el concepto de ganancia

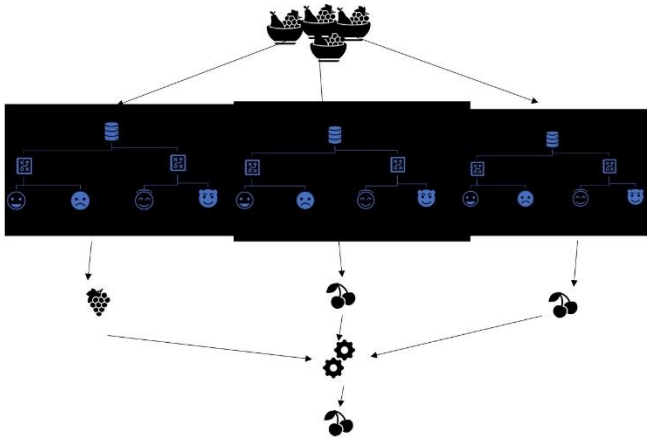
$$Ganancia(S, A) = Entropia(S) - \sum_{V(a)}^{m} \frac{|S_v|}{|S|} Entropia(S_v)$$

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.



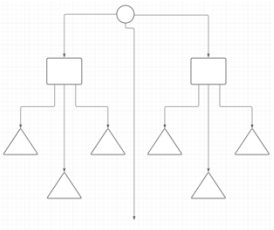

#### 3.2.2 Random Forest

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 arboles será tomado como resultado.



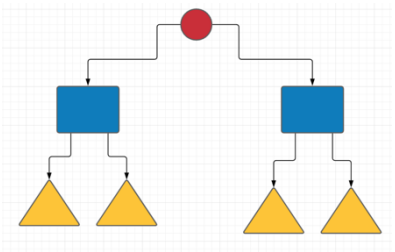
#### 3.2.3 CHAID

Consiste en generar arboles de decisión 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 decisión no binarios, lo que simboliza que algunas de las divisiones pueden generar más 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 más tiempo.




#### 3.2.4 QUEST

Proporciona método de clasificación binaria, para él creación de árboles de decisión, y se creó 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 más divisiones, y su objetivo es que los rangos sean categóricos, aunque también se pueden rangos numéricos, todas las divisiones son binarias.



## 4. DISEÑO DE LOS ALGORITMOS

### 4.1 Estructura de los datos

Para entender como funciona un arbol primero se debe saber de donde viene. Sus bases o fundamentos estan basados en un sistema de reglas de inferencia, las culaes estudian las premisas o situacions y se les da un valor de verdad, asi, de este mismo modo trabaja los arboles de decision, se construyen diagramas de construcciones lógicas y se escogen preguntas optimas paraa llegar a respuestas que puedan dar solucion a un problema.

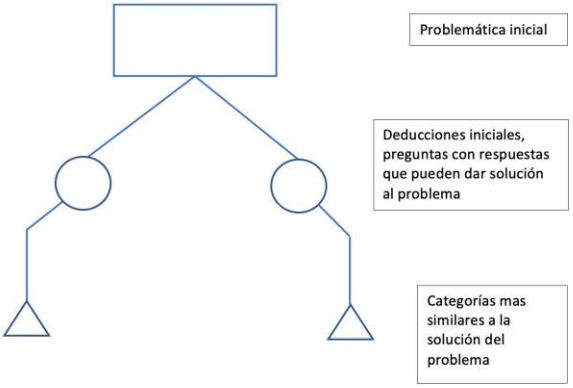
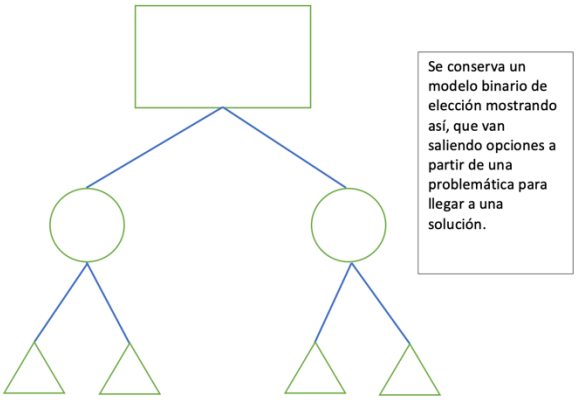



**Figura 1:** Un árbol de decisión binario para predecir Saber Pro basado en los resultados de Saber 11. Los nodos violetas representan a aquellos con una alta probabilidad de éxito, los verdes con una probabilidad media y los rojos con una baja probabilidad de éxito.

### 4.2 Algoritmos

*Árbol de decisión CART:*

*Su nombre se debe a un acronimo: (Classification And Regression Trees), arboles de clasificacion y regresion. Este algoritmo esta basado en metodos netamente binarios admite variables de entrada y de salida nominales, ordinales y continuas, por lo que se pueden resolver tanto problemas de clasificación como de regresión.*




#### 4.2.1 Entrenamiento del modelo

El algoritmo tiene como entrada una probelmatica, de alli comienza a sacar una serie de opciones a las cuales le saca mas nodos de decision, la gracia de trabajar con este algoritmo es que sea usado solamente de forma binaria por lo cual cada nodo tendra 2 ramas abajo y asi continuara hasta llegar sl punto mas sencillo y mas cercano a la solucion de dicha problemática.

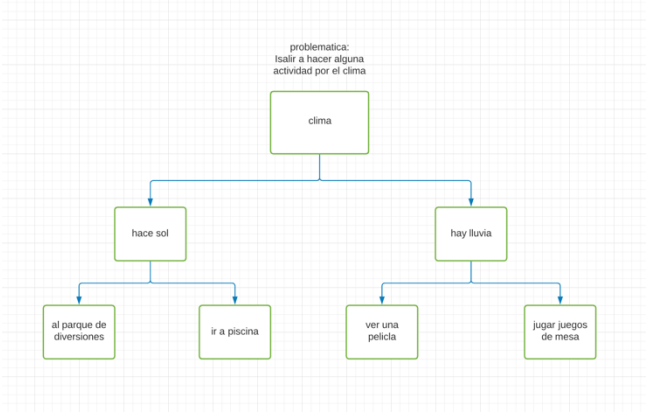



**Figura 2:** Entrenamiento de un árbol de decisión binario usando *(En este semestre, uno podría ser CART, ID3, C4.5... por favor, elija)*. En este ejemplo, mostramos un modelo para predecir si se debe jugar al golf o no, según el clima.

Este algoritmo su base para la decisión es el coeficiente Gini, $I_g = 1 - (p_0^2 + p_1^2)$ donde $p_0 = \frac{n_0}{(n_0+n_1)}$ y $p_1 = \frac{n_1}{(n_1+n_0)}$, sabiendo que $n_i$ es el número de elementos

con la etiqueta i. luego de un exhaustivo proceso de selección encontramos que nuestras preguntas para el árbol son:

| pregunta | Dato | Promedio Gini | Mejor Analisis |
|---|---|---|---|
| punt_ingles | 50 | 0,374715592 | separarMayor/separarMenor |
| punt_ciencias_sociales | 50 | 0,385802042 | separarMayor/separarMenor |
| punt_lenguaje | 52 | 0,387539322 | separarMayor/separarMenor |
| punt_quimica | 51 | 0,392324866 | separarMayor/separarMenor |
| punt_biologia | 50 | 0,395573762 | separarMayor/separarMenor |
| punt_matematicas | 52 | 0,398061174 | separarMayor/separarMenor |
| punt_filosofia | 45 | 0,420626544 | separarMayor/separarMenor |
| desemp_ingles | A- | 0,425973394 | 0 |
| punt_fisica | 50 | 0,426650814 | separarMayor/separarMenor |
| cole_jornada | COMPLETA | 0,476830802 | 0 |
| fami_estratovivienda.1 | Estrato 1 | 0,480354038 | 0 |
| fami_pisoshogar | baldosa, tableta, mÃ¡ | 0,482050868 | 0 |
| fami_tieneinternet.1 | Si | 0,4821446 | 0 |
| fami_numlibros | 0 A 10 LIBROS | 0,482450482 | 0 |
| fami_tienecomputador.1 | No | 0,486019176 | 0 |
| cole_caracter | ACADÃ‰MICO | 0,491228278 | 0 |
| estu_areareside | Cabecera Municipal | 0,495959532 | 0 |
| edad | 17 | 0,49621454 | separarMayor/separarMenor |
| fami_tiene_nevera.1 | Si | 0,497397418 | 0 |
| estu_trabajaactualmente | No | 0,498493658 | 0 |
| estu_tomo_cursopreparacion | Si | 0,499694104 | 0 |
| fami_tiene_celular.1 | 0 | 0,499786414 | 0 |

**Figura 3:** resultado de buscar mejor Gini, cual es la mejor respuesta para cada pregunta el Gini y el método de separación de las preguntas numéricas.

### 4.2.2 Algoritmo de prueba

Primero se tomaron los datos que ante nuestras opiniones eran mas influyentes en el proceso de predicción de resultados para una prueba, luego de tener estos datos los clasificamos en gurpos de tipo de datos, teniedno estos datos leimos que nivel de influencia tenia y usamos la impureza de Gini como apoyo y aun mas filtracion, llegando asi a una serie de datos que nos podrian guiar a una respuesta mas correcta de cómo sería el desempeño de la perosna en dicha prueba.

## 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 (references complete of participants Page 13-16)
7. https://www.ibm.com/support/knowledgecenter/es/SS3RA7_sub/modeler_mainhelp_client_ddita/clementine/nodes_treebuilding.html
8. DAMODARAN, Aswath (2014). “Uno de los mayores errores es asumir que el crecimiento de una compañía es gratis o muy barato”. Entrevista concedida a Javier García para el portal Sintetia.com el 13 de enero de 2014.
9. Consultada el 10 de febrero de 2015. Disponible en: https://www.sintetia.com/aswath-damodaran-stern-finance/ DAMODARAN, Aswath (2014). “Uno de los mayores errores es asumir que el crecimiento de una compañía es gratis o muy barato”. Entrevista concedida a Javier García para el portal Sintetia.com el 13 de enero de 2014. Consultada el 10 de febrero de 2015. Disponible en: https://www.sintetia.com/aswath-damodaran-stern-finance/
10. Hidalgo Pérez M., Las redes neuronales tienen derecho a no poner la mano en el fuego, Retina, 23 de diciembre de 2019, tomado de: https://retina.elpais.com/retina/2019/12/20/innovacion/1576838697_328758.html
11. https://sites.google.com/site/sistemasexpertosunah/home/sistemas-expertos-basados-en-reglas
12. https://bookdown.org/content/2274/metodos-de-clasificacion.html
13.