# LA PREDICCION DEL ÉXITO ACADEMICO


| Juan Pablo Henao<br>Universidad Eafit<br>Colombia<br>jphenaob@eafit.edu.co | Diego Alejandro Vanegas<br>Universidad Eafit<br>Colombia<br>dvanegasg@eafit.edu.co | Miguel Correa<br>Universidad Eafit<br>Colombia<br>macorream@eafit.edu.co | Mauricio Toro<br>Universidad Eafit<br>Colombia<br>mtorobe@eafit.edu.co |
|---|---|---|---|



**RESUMEN**

La problemática ante este proyecto es la búsqueda de predicción de futuros resultados académicos mediante arboles de decisión, es importante ya que aporta mayor control y oportunidad anticipada de los futuros resultados académicos, así asegurando un éxito si así se desea.



**Palabras clave**

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


## 1. INTRODUCCIÓN

Nos motivamos por la excelencia académica que podría ser posible gracias a este proyecto. Donde cada persona tenga la oportunidad de mejorar sus resultados y métodos para cumplir sus objetivos

### 1.1. Problema

El impacto que genera en un continente donde el acceso a educación gratuita suele ser complicada, causa efectos negativos en ciertas ocasiones; impidiendo el avance tanto de sociedad como de entorno. La importancia de resolver esta problemática es bastante, ya que el objetivo es formar una sociedad con más posibilidades y accesibilidades cuando de estudio y formación se trata, apoyando el avance científico, tecnológico, entre otras.

### 1.2 Solución

En este trabajo, nos centramos en los árboles de decisión porque proporcionan una gran explicabilidad. Evitamos los métodos de caja negra como las redes neuronales, las máquinas de soporte vectorial y los bosques aleatorios porque carecen de explicabilidad

### 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

### 2.1 Predicción de Epítopos B

Metodología computacional para la predicción cuantitativa de epítopos de células B. Con la propuesta de desarrollar nuevas vacunas contra el dengue y Chikunguña.

Se realizan 8 predicciones, se agrupan resultados y se determinan computacionalmente los cambios energéticos en cada uno de ellos aminoácidos

$$FL = WL.\ \Delta\Delta\ G.\ S1.\ S2\ /\ B$$

$$FC = WC.\ \Delta\Delta\ G.\ S1.\ S2\ /\ B$$

WL y WC son la frecuencia de aparición de aminoácidos presentes en dicho epítopo. FL y FC corresponde al valor promedio de cada epítopo B.

### 2.2 Predicción del consumo de Cocaína en adolescentes

Evaluar poder predictivo de la impulsividad y búsqueda de sensaciones sobre el consumo de Cocaína.

La muestra se llevó a cabo en 278 adolescentes entre 14 a 18 años. Se tuvieron en cuenta factores de personalidad como impulsividad y búsqueda de sensaciones y si consume actualmente alguna sustancia.

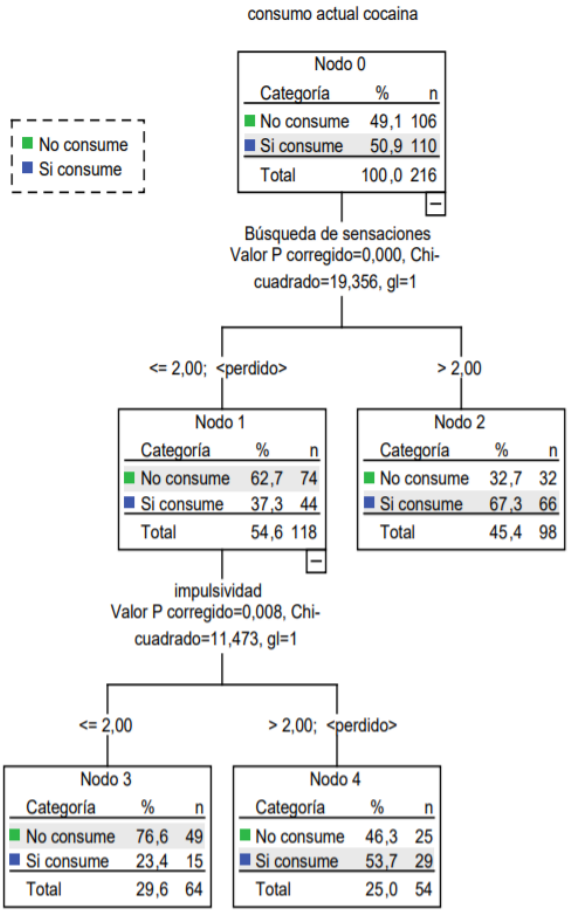



Figura 1. Ejemplo árbol de decisión para predecir el consumo de cocaína mediante el algoritmo CHAID (Kass,1980)

**2.3 Clasificación de dengue hemorrágico**

Hallar reglas de decisión que permitan clasificar un paciente con dengue en diversas formas de la enfermedad.

Síntomas de la enfermedad se representan como fiebre (F), cefalea (C), dolor retroocular (DO), diarreas (D), entre otras.

Se consideran las matrices de perdidas siguientes

(0 1,001-i/100)

(i/100 0 ) donde i= 1,...,100

**2.4 Predicción de dureza en piezas construidas con acero templado y revenido**

Predicción de propiedades mecánicas en puntos de una pieza sometida a un tratamiento de temple y revenido, desarrollando 8un algoritmo que permita obtener valores de dureza en cualquier punto redondo representativo de la sección de la pieza.

Su funcionamiento consiste en ingresar diámetro de pieza, se cargan datos de composición, se introduce tamaño de grano, temperatura y velocidad relativa, temperatura revenido. Se calcula valor sumatorio aleantes y calcula su dureza.

## 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.

## 4. DISEÑO DE LOS ALGORITMOS

### 4.1 Estructura de datos

La estructura de datos utilizada para hacer la prediccion es un Árbol de decisión binario, el cuál consta de un Root Node del que se despliegan varios Decision Node y de estos mismos se pueden desplegar Terminal Nodes o otros Decision Node, estos nos proporcionan gran explicabilidad a los resultados de la predicción

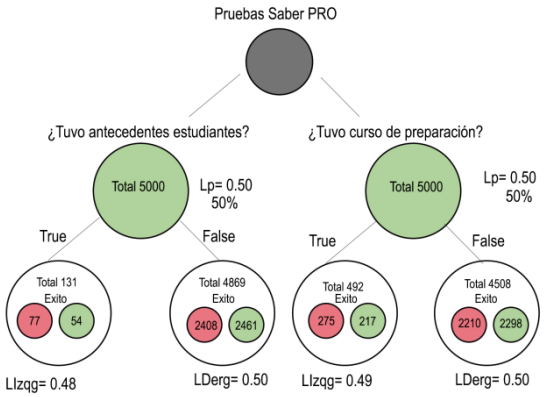



Figura 1: Un árbol de decisión binario para predecir Saber Pro basado en los antecedentes del estudiante y los cursos de preparación que tomó. Los nodos violetas representan a aquellos con una alta probabilidad de éxito, los verdes con una probabilidad media y los rojos con una baja probabilidad de éxito

### 4.2 Algoritmos

Utilizamos el algoritmo CART porque nos permite una clasificación con técnicas tradicionales para el análisis de datos de una forma más eficiente.

| Estudiantes | Antecedentes estudiantiles | | Curso de preparación | |
|---|---|---|---|---|
| | 0 | 1 | 1 | 0 |
| Total | 492 | 4508 | 131 | 4869 |
| Pasaron | 54 | 2461 | 217 | 2298 |
| No pasaron | 77 | 408 | 275 | 2210 |

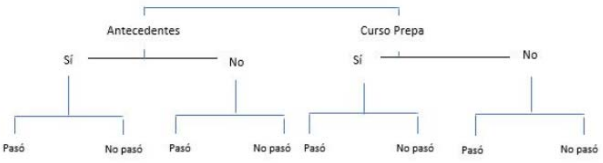
Tabla 1: Número de estudiantes en cada conjunto de datos utilizados para la predicción del éxito y el cálculo de la impureza de Gini


Gráfica 1: Datos graficados de los Antecedentes Estudiantiles

vistos en la Tabla 1.


Gráfica 2: Datos graficados de los estudiantes que tomaron un

Curso de Preparación vistos en la Tabla 1.




#### 4.2.1 Entrenamiento del modelo

Ya que el Algoritmo CART clasifica y construye a partir de condiciones binarias, lo que nos da una forma muy simple de interpretar los resultados

4.2.2 Algoritmo de prueba Lo probamos utilizando las variables “Antecedentes Estudiantiles” y “Curso de preparación”, analizándolos con el respectivo porcentaje de éxito que tuvieron los estudiantes en el examen

Saber Pro y de esta forma, clasificamos los datos y construimos el árbol.

5. RESULTADOS

5.1 Evaluación del modelo

En esta sección, presentamos algunas métricas para evaluar el modelo. La precisión es la relación entre el número de predicciones correctas y el número total de datos de entrada. Precisión. es la proporción de estudiantes exitosos identificados correctamente por el modelo y estudiantes exitosos identificados por el modelo. Por último, Sensibilidad es la proporción de estudiantes exitosos identificados correctamente por el modelo y estudiantes exitosos en el conjunto de datos.

**5.1.1 Evaluación del modelo en entrenamiento**

A continuación, presentamos las métricas de evaluación de los conjuntos de datos de entrenamiento en la Tabla 3.

| | ***Conjunto de datos 1*** | ***Conjunto de datos 2*** |
|---|---|---|
| *Exactitud* | 0.7 | 0.75 |
| *Precisión* | 0.7 | 0.75 |
| *Sensibilidad* | 0.7 | 0.75 |

**Tabla 3.** Evaluación del modelo con los conjuntos de datos de entrenamiento.

**5.1.2 Evaluación de los conjuntos de datos de validación**

A continuación presentamos las métricas de evaluación para los conjuntos de datos de validación en la Tabla 4.

| | ***Conjunto de datos 1*** | ***Conjunto de datos 2*** |
|---|---|---|
| *Exactitud* | 0.5 | 0.55 |
| *Precisión* | 0.5 | 0.55 |
| *Sensibilidad* | 0.5 | 0.55 |

**Tabla 4.** Evaluación del modelo con los conjuntos de datos de validación.

**5.2 Tiempos de ejecución**

Calcular el tiempo de ejecución de cada conjunto de datos en Github. Medir el tiempo de ejecución 100 veces, para cada conjunto de datos, e informar del tiempo medio de ejecución para cada conjunto de datos.

| | ***Conjunto de datos 1*** | ***Conjunto de datos 2*** |
|---|---|---|
| *Tiempo de entrenamiento* | 0.313 s | 0.913 s |
| *Tiempo de validación* | 0.313 s | 0.913 s |

Tabla 5: Tiempo de ejecución del algoritmo *CART* para diferentes conjuntos de datos.

5.3 Consumo de memoria

Presentamos el consumo de memoria del árbol de decisión binario, para diferentes conjuntos de datos, en la Tabla 6.

| | ***Conjunto de datos 1*** | ***Conjunto de datos 2*** |
|---|---|---|
| Consumo de memoria | 30 MB | 90 MB |

**Tabla 6: Consumo** de memoria del árbol de decisión binario para diferentes conjuntos de datos.

**6. DISCUSIÓN DE LOS RESULTADOS**

**6.1 Trabajos futuros**

En el futuro nos gustaría mejorar el algoritmo y hacerle varias implementaciones, tales como, la optimización del tiempo de ejecución y el consumo de memoria, utilizar un mayor número de árboles para mejorar la precisión y exactitud del algoritmo. También, revisar la idea sobre implementar Bosques Aleatorios al algoritmo, de esta forma hacer un gran cambio a este.

**AGRADECIMIENTOS**

Agradecemos la asistencia con metodología a Maria Elena Vanegas González, Estudiante de la Universidad Católica del Norte por los comentarios que mejoraron enormemente el manuscrito

**REFERENCIAS**

1. Isea, R. (2015). Predicción computacional cuantitativa de epítopos de células B. *Vaccimonitor*, *24*(2), 0-0.

2.García, E. G., & Pol, A. L. P. (2009). Predicción del consumo de cocaína en adolescentes mediante árboles de decisión. *Revista de Investigación en Educación*, *6*(1), 7-13.

3.Vega RB, Sánchez VL, Cortiñas AJ, et al. Clasificación de dengue hemorrágico utilizando árboles de decisión en la fase temprana de la enfermedad. Rev Cubana Med Trop. 2012;64(1):35-42.

4.Yanzón, R. C., Bocca, J. C., Rebollo, D., & Sánchez, A. R. (2009). Predicción de dureza en piezas construidas con acero templado y revenido.

5. De Lejarza y Esparducer, I. M. (1998). Árboles de Clasificación y Regresión [Archivo PDF]. Recuperado de https://www.uv.es/mlejarza/actuariales/tam/arbolesdecision.pdf