

## Predicción del éxito estudiantil con árboles de decisión

Samuel David Villegas Bedoya Universidad Eafit Colombia sdvillegab@eafit.edu.co	Julian Andres Ramirez Jimenez Universidad Eafit Colombia jaramirezj@eafit.edu.co	Miguel Correa Universidad Eafit Colombia macorream@eafit.edu.co	Mauricio Toro Universidad Eafit Colombia mtorobe@eafit.edu.co
---	--	--	--

**Para cada versión de este informe: 1. Detalle todo el texto en rojo. 2. Ajustar los espacios entre las palabras y los párrafos. 3. Cambiar el color de todos los textos a negro.**

**Texto rojo = Comentarios**

**Texto negro = Contribución de Miguel y Mauricio**

**Texto en verde = Completar para el 1er entregable**

**Texto en azul = Completar para el 2º entregable**

**Texto en violeta = Completar para el tercer entregable**

### RESUMEN

Este trabajo se centra en el diseño y desarrollo de un algoritmo basado en arboles de decisión que nos permita predecir el éxito académico de los estudiantes en la educación superior, más específicamente en las pruebas Saber Pro. Esto se hará teniendo en cuenta ciertos factores socioeconómicos y los resultados de las pruebas ICFES Saber 11. Es importante dar solución a este problema debido a que podemos identificar las variables que más influyen en que un estudiante pueda lograr un desempeño por encima de la media o no en las pruebas Saber Pro, y a futuro estos datos podrían ayudar a tomar decisiones gubernamentales que impacten sobre la calidad de la educación.

¿Cuál es el algoritmo propuesto? ¿Qué resultados obtuvieron? ¿Cuáles son las conclusiones de este trabajo? El resumen debe tener como máximo **200 palabras**. (En este semestre, usted debe resumir aquí los tiempos de ejecución, el consumo de memoria, la exactitud, la precisión y la sensibilidad)

### Palabras clave

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

## 1. INTRODUCCIÓN

Año tras año muchas de las universidades de todo el mundo buscan la excelencia académica en sus estudiantes, cada una lo hace a su manera, pero en cada país existen formas de ver si una universidad tiene cierto nivel académico en sus estudiantes, esto es por medio de pruebas estandarizadas, en nuestro país, por ejemplo, tenemos a las pruebas saber Pro. Estas pruebas las presentan la mayoría de los estudiantes, e incluso es requisito de graduación en la

mayoría de las universidades. Por lo anterior un alto porcentaje de estudiantes de pregrado lo hacen, dando pie a pensar que puede ser un gran indicador de la calidad académica, de allí el interés de las instituciones de querer sacar un buen puntaje en estas pruebas.

Una de las maneras de buscar el mejoramiento es prediciendo los resultados, es decir, adelantarse a la prueba real, dado que podemos poner en evidencia los factores y variables que influyen en un resultado positivo o negativo, y a partir de esto buscar soluciones que no solamente aporten al resultado de la prueba, sino también al mejoramiento de la educación. Teniendo en cuenta lo anteriormente mencionado, el objetivo de nuestro proyecto es predecir los resultados y buscar aquellos factores que pueden influir en los resultados de las pruebas estandarizadas en las universidades de América latina, para así mejorar la calidad de educación de esta región. Esto lo pretendemos lograr mediante un algoritmo basado en arboles de decisión.

### 1.1. Problema

El problema a solucionar es realizar una predicción del éxito académico de los estudiantes en la educación superior en base a diferentes datos de estos, resolver este problema ayudaría a que se puedan tomar decisiones a tiempo por parte de las Instituciones de educación superior para el mejoramiento y así evitar malos resultados.

### 1.2 Solución

En este trabajo, nos centramos en los árboles de decisión porque proporcionan una gran explicabilidad (¡falta una cita para este argumento!). Evitamos los métodos de caja negra como las redes neuronales, las máquinas de soporte vectorial y los bosques aleatorios porque carecen de explicabilidad (¡Falta una cita para este argumento!).

Explique, brevemente, su solución al problema (En este semestre, la solución es una implementación de un algoritmo de árbol de decisión para predecir el éxito académico. ¿Qué algoritmo elegiste? ¿Por qué?)

### 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**

### **3.1 Árboles de decisión para predecir factores asociados al desempeño académico de estudiantes de bachillerato en las pruebas Saber 11°**

En este artículo se habla de que en base a un modelo de clasificación basado en árboles de decisión se detectaron factores asociados al desempeño académico de los estudiantes de grado undécimo de la educación media, en este caso se aplica el modelo para los estudiantes que presentaron las pruebas Saber 11 en los años 2015 y 2016. Para el modelo de clasificación utilizaron el algoritmo J48, el cual se basa e implementa el C4.5., además la precisión que lograron con el algoritmo fue del 67%. [1]

### **3.2 Using Decision Trees to Understand Student Data**

En este artículo se habla de la aplicación que se le dio a un algoritmo basado en árboles de decisión con el que a partir de datos de estudiantes universitarios se generaron gráficas para la comprensión humana, cuyo análisis permitía predecir la graduación, o mejor, comprender los factores que conducían a la misma y así poder generar conclusiones que aporten a futuras decisiones institucionales. Además se comparó este método de árboles de decisión con redes neuronales, máquinas vectoriales de soporte y regresión del núcleo, para comprobar su efectividad frente a estas. La precisión de este modelo estuvo alrededor del 64%. [2]

### **3.3 Predicción del rendimiento de estudiantes de ingeniería mediante árboles de decisión**

En este artículo se plantearon generar un modelo predictivo utilizando árboles de decisión que buscaba pronosticar los desempeños de estudiantes de ingeniería, permitiendo identificar antes, a los estudiantes que posiblemente tuvieran más riesgo de fracasar, dando la posibilidad de hacer un acompañamiento a estos casos. El algoritmo utilizado fue el J48, el cual es una implementación en el lenguaje de programación Java del algoritmo C4.5, según los datos obtenidos la tasa positiva del modelo para la clase FAIL fue de 0.907, significando que el modelo estaba identificando con éxito aquellos estudiantes que probablemente fracasarían. [3]

### **3.4 Modelo predictivo de deserción estudiantil basado en árboles de decisión**

En este artículo contrario a los otros trabajos relacionados, se buscaba encontrar aquellos factores que influyen en la deserción estudiantil, implementando varios algoritmos en base a árboles de decisión como lo fue el CART y este apoyándose en el algoritmo Hunt. En cuanto a la efectividad del algoritmo realizaron la curva Receiver Operating Characteristic (ROC) obteniendo como resultado un 94% de efectividad. [4]

## **3. MATERIALES Y MÉTODOS**

En esta sección se explica cómo se recopilaban y procesaban 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
<b>Entrenamiento</b>	15,000	45,000	75,000	105,000	135,000
<b>Validación</b>	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

El algoritmo ID3 realizado por Quinlan implementa un sistema de árboles de decisión con dirección de arriba abajo, teniendo en cuenta lo siguiente:

**Ejemplos:** Se denomina ejemplos a aquellos datos que se utilizan para el entrenamiento del algoritmo, buscan aumentar la posibilidad de acierto de este, es decir entre más ejemplos mejor. Cada uno de estos se componen de atributos y valores.

**Atributos:** Se denomina atributo a aquellas características que componen a los datos, estos son los que se toman en cuenta a la hora de crear el árbol. Por ejemplo, si queremos predecir la posibilidad de tener un contagio de COVID un atributo podría ser Tener tos o fiebre.

**Valores:** Denominamos valores a los estados que pueden tomar los atributos, pueden ser cuantitativos y cualitativos. Por ejemplo, si queremos predecir la posibilidad de almorzar en un restaurante, un atributo puede ser si hay muchas personas y sus posibles valores pueden ser ninguna, algunas o muchas.

**Hojas:** Es aquel resultado que se tienen por cada una de las ramas. En algunos casos puede pasar que en varias ramas repitan el mismo resultado. Puede ser un resultado binario, es decir, si o no, verdadero o falso, etc. Aunque también se pueden usar más de dos estados.

**Arcos:** Los cuales contienen valores posibles del nodo inicial.

**Entropía:** Este concepto viene de la teoría de la información, que consiste básicamente en el nivel de desorden de los componentes de un sistema, en nuestro caso el nivel de variaciones que podamos tener a medida que agregamos nuestros atributos.

- Una muestra completamente homogénea tiene entropía 0.
- Una muestra igualmente distribuida tiene entropía 1.

Formulas

$$E(S) = \sum_{i=1}^c -p_i \log_2 p_i$$

$$\text{Entropía}(s) = -P \log_2 (P) - N \log_2 (N)$$

**Figura 1:** Es la fórmula general para la entropía.

**Figura 2:** Es una variante de la fórmula, se utiliza más que todo para calcular la entropía original, donde  $P = \frac{\#EjemplosPos}{TotalEjemplos}$  y  $N = \frac{\#EjemplosNeg}{TotalEjemplos}$ .

**Ganancia (Informacion Gain):** El nivel de información que se gana en cada atributo. De este concepto depende el nivel de cada uno de los atributos, es decir, entre más ganancia tenga un atributo es ubicado en un nivel más alto.

#### Proceso

¿Cómo buscar el nodo principal?

- Se calcula la entropía del total
- Se divide el conjunto de datos en función de los diferentes atributos.
- Se calcula la entropía en cada rama y suma proporcionalmente las ramas para calcular la entropía del total.
- Se resta este resultado de la entropía original
- El resultado es la ganancia de información (Descenso de entropía).
- EL atributo con mayor ganancia es seleccionado como nodo de decisión.

Otros pasos:

- Una rama con entropía 0 se convierte en hoja (Todos los casos clasificados).

- Si no es así, la rama debe seguir subdividiéndose
- El algoritmo se ejecuta recursivamente hasta que se llegue a nodos hoja. [5]

### Ejemplo: PlayGolf

Atributos				Objetivo
Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

Figura 3: Datasets ejemplo golf.

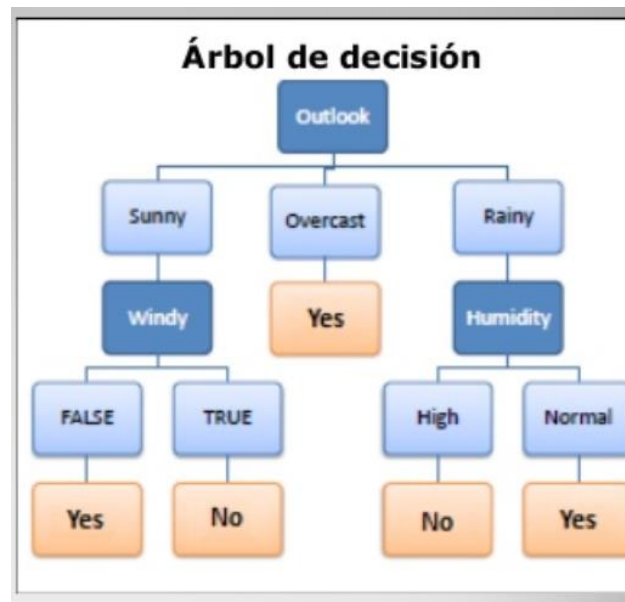


Figura 4: Árbol generado de los datasets mostrados en la figura 1 utilizando ID3.

**Complejidad Espacio:** Lineal/ Polinómico

**Complejidad Tiempo:** Polinómica/Exponencial [6]

### 3.2.2 C4.5

El algoritmo C4.5 podríamos llamarlo como una actualización del algoritmo anterior llamado ID3, por ende, se basa en los mismos conceptos como lo son: La entropía y

la ganancia de información. Pero introduce nuevos conceptos como:

**Split Info (SI):** Permite generar nodos hijos con la menor impureza posible.

$$SI(S) = - \sum_{i=1}^N \frac{S_i}{S} \left( \log\left(\frac{S_i}{S}\right) \right)$$

Figura 5

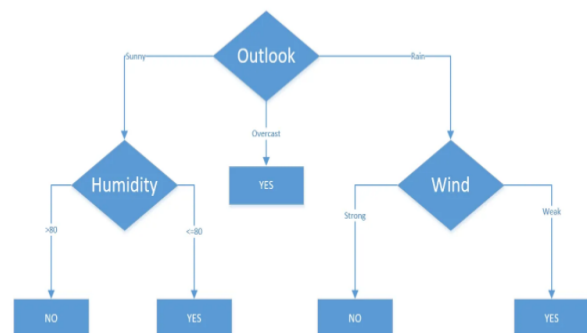
**Ratio (GR):** Es la razón entre el IG que teníamos por ejemplo en el ID3 y el nuevo concepto SI.

$$GR(S, A) = \frac{IG(S, A)}{SI(S)}$$

Figura 6

Eliminamos el concepto de ganancia como ya lo conocíamos dado que nos generaba un problema, Al momento de evaluar los atributos el algoritmo ID3 favorecía a los atributos con gran cantidad de valores. Para arreglar este asunto Quinlan establece la propuesta del Ratio (GR).

También se introduce un nuevo concepto como lo es **La poda**, esta tiene como función eliminar y reemplazar por nodos de hoja a aquellas ramas que tengan una tasa de error más alta establecido por una regla de decisión.



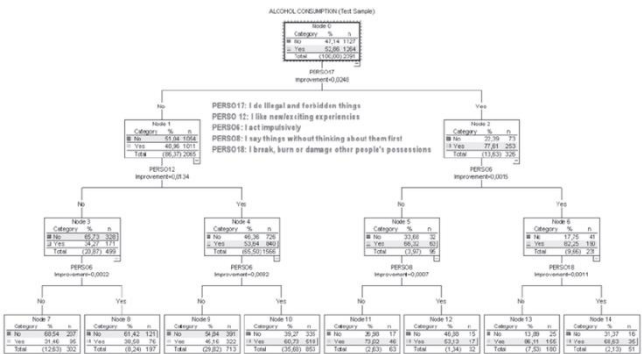
Árbol de decisión generado por C4.5

**Figura 7:** Es un ejemplo del árbol generado por C4.5 en comparación con el árbol hecho en el algoritmo ID3.

En cuanto a la complejidad no se encontraron registros de cambio en memoria o tiempo, pero si mejoraron luego con C5.0/See5 que es una versión mejorada del algoritmo en cuestión.

3.2.3 CART

CART es el acrónimo de Classification and Regresion Trees(Arboles de clasificación y regresión). El algoritmo CART es un método que utiliza datos históricos para construir arboles de clasificación o de regresión los cuales son usados para clasificar o predecir nuevos datos.[7]



**Figura 8:** Árbol de clasificación generado por el algoritmo CART (Breiman et al., 1984)

3.2.4 QUEST

QUEST es el acrónimo de Quick, Unbiased, Efficient Statistical Tree (Árbol Estadístico Eficiente, Insesgado y Rápido). El procedimiento que sigue este algoritmo básicamente es el siguiente: “primero se elige la mejor variable predictora cuyo objetivo es que el número de categorías que poseen las variables no afecte a la elección de la mejor variable, para realizar después la mejor segmentación de la variable que ha seleccionado”[8]

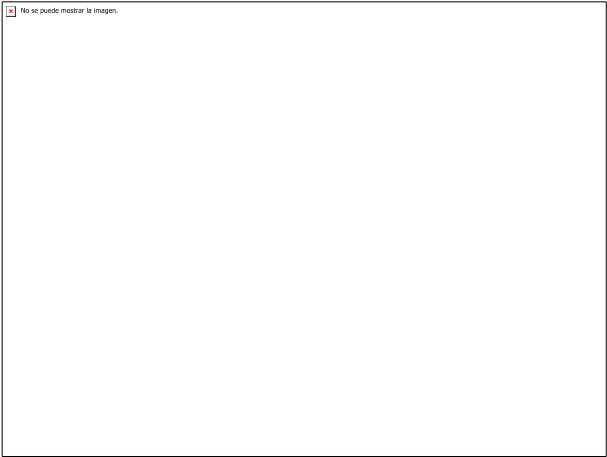
(Para este algoritmo en específico no encontramos una figura)

4. DISEÑO DE LOS ALGORITMOS

En lo que sigue, explicamos la estructura de los datos y los algoritmos utilizados en este trabajo. La implementación del algoritmo y la estructura de datos se encuentra disponible en Github<sup>1</sup>.

4.1 Estructura de los datos

Explique la estructura de datos utilizada para hacer la predicción y haga una figura que la explique. No utilice imágenes de Internet. (En este semestre, la estructura de datos es un árbol de decisión binario)



**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.

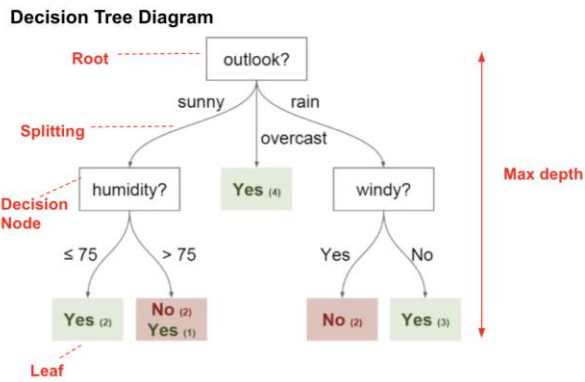
4.2 Algoritmos

Explica el diseño del algoritmo para resolver el problema y haz una figura. No uses figuras de Internet, haz las tuyas propias. (En este semestre, un algoritmo debe ser un algoritmo para entrenar un algoritmo de árbol de decisión como ID3, C4.5, CART y el segundo algoritmo debe ser un algoritmo para clasificar los nuevos datos utilizando dicho árbol).

4.2.1 Entrenamiento del modelo

Explique, brevemente, cómo entrenó a la modelo: Esto equivale a explicar cómo su algoritmo construye automáticamente un árbol de decisión binario.

<sup>1</sup><http://www.github.com/ ?????????? /proyecto/>



**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.

#### 4.2.2 Algoritmo de prueba

Explique, brevemente, cómo probó el modelo: Esto equivale a explicar cómo su algoritmo clasifica los nuevos datos después de que se construya el árbol.

#### 4.3 Análisis de la complejidad de los algoritmos

Explique en sus propias palabras el análisis para el peor caso usando la notación  $O$ . ¿Cómo calculó tales complejidades.

Algoritmo	La complejidad del tiempo
Entrenar el árbol de decisión	$O(N^2 * M^2)$
Validar el árbol de decisión	$O(N^3 * M * 2N)$

**Tabla 2:** Complejidad temporal de los algoritmos de entrenamiento y prueba. (*Por favor, explique qué significan  $N$  y  $M$  en este problema.*)

Algoritmo	Complejidad de memoria
Entrenar el árbol de decisión	$O(N * M * 2N)$
Validar el árbol de decisión	$O(1)$

**Tabla 3:** Complejidad de memoria de los algoritmos de entrenamiento y prueba. (*Por favor, explique qué significan  $N$  y  $M$  en este problema.*)

#### 4.4 Criterios de diseño del algoritmo

Explica por qué el algoritmo fue diseñado de esa manera. Use un criterio objetivo. Los criterios objetivos se basan en

la eficiencia, que se mide en términos de tiempo y consumo de memoria. Ejemplos de criterios no objetivos son: "Estaba enfermo", "fue la primera estructura de datos que encontré en Internet", "lo hice el último día antes del plazo", etc. Recuerde: Este es el 40% de la calificación del proyecto.

## 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	...Conjunto de datos $n$
Exactitud	0.7	0.75	0.9
Precisión	0.7	0.75	0.9
Sensibilidad	0.7	0.75	0.9

**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	...Conjunto de datos $n$
Exactitud	0.5	0.55	0.7
Precisión	0.5	0.55	0.7
Sensibilidad	0.5	0.55	0.8

**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.

	<i>Conjunto de datos 1</i>	<i>Conjunto de datos 2</i>	<i>...Conjunto de datos n</i>
<i>Tiempo de entrenamiento</i>	10.2 s	20.4 s	5.1 s
<i>Tiempo de validación</i>	1.1 s	1.3 s	3.3 s

**Tabla 5:** Tiempo de ejecución del algoritmo (*Por favor, escriba el nombre del algoritmo, C4.5, ID3*) 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.

	<i>Conjunto de datos 1</i>	<i>Conjunto de datos 2</i>	<i>...Conjunto de datos n</i>
Consumo de memoria	10 MB	20 MB	5 MB

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

Para medir el consumo de memoria, debería usar un generador de perfiles (*profiler*). Uno muy bueno para Java es VisualVM, desarrollado por Oracle, <http://docs.oracle.com/javase/7/docs/technotes/guides/visualvm/profiler.html>. Para Python, use C-profiler.

## 6. DISCUSIÓN DE LOS RESULTADOS

Explique los resultados obtenidos. ¿Son la precisión, exactitud y sensibilidad apropiadas para este problema? ¿El modelo está sobreajustado? ¿Es el consumo de memoria y el consumo de tiempo sib apropiados? (*En este semestre, de acuerdo con los resultados, ¿se puede aplicar esto para dar becas o para ayudar a los estudiantes con baja probabilidad de éxito? ¿Para qué es mejor?*)

### 6.1 Trabajos futuros

Respuesta, ¿qué le gustaría mejorar en el futuro? ¿Cómo le gustaría mejorar su algoritmo y su implementación? ¿Qué hay de usar un bosque aleatorio?

### AGRADECIMIENTOS

Identifique el tipo de agradecimiento que quiere escribir: Para una persona o para una institución. Considere las siguientes pautas: 1. El nombre del profesor no se menciona porque es un autor. 2. No debe mencionar sitios web de autores de artículos que no haya contactado. 3. Debe

mencionar estudiantes y profesores de otros cursos que le hayan ayudado.

Como ejemplo: Esta investigación fue apoyada parcialmente por [Nombre de la Fundación, Donante].

Agradecemos la asistencia con [técnica particular, metodología] a [nombre apellido, cargo, nombre de la institución] por los comentarios que mejoraron enormemente el manuscrito.

### REFERENCIAS

Las referencias se hacen con el formato de referencias de la ACM. Lea las directrices de ACM en <http://bit.ly/2pZnE5g>

1. Timarán, R y Caicedo, J y Hidalgo, A . Árboles de decisión para predecir factores asociados al desempeño académico de estudiantes de bachillerato en las pruebas Saber 11°. *Revista De Investigación, Desarrollo E Innovación* 9, 2 (2019). DOI:<http://dx.doi.org/10.19053/20278306.v9.n2.2019.9184>
2. Murray, E. Using Decision Trees to Understand Student Data. 2015.
3. R. R. Kaba y R. S. Bichkar. Performance Prediction of Engineering Students using Decision Trees, *International Journal of Computer Applications* (0975 – 8887), 36(11). 8-12.
4. Cuji, B., Gavilanes, W. y Sanchez, R. Modelo predictivo de deserción estudiantil basado en arboles de decisión. *Revista Espacios*, 38(55). 17- 26.
5. Sancho, F. (Enero 2013). Arboles decisión id3 Recuperado 16 Agosto, 2020, de <https://es.slideshare.net/FernandoCaparrini/arboles-decision-id3>
6. Iglesia, J.L. Descubriendo la Inteligencia Artificial. (2017, 11, 3). N° 177: IA Probabilidad - Árboles de Decisión ID3 02 [Archivo de video]. Recuperado de <https://www.youtube.com/watch?v=PxPKZA87EXU>
7. Díaz Sepúlveda, J. and Correa, J., . Comparación entre árboles de regresión CART y regresión lineal. *Comunicaciones en Estadística*, 2013, 6(2), p.175.
8. Parra, F. Estadística y Machine Learning con R. 2019.