# **TÍTULO (UNA BREVE DESCRIPCIÓN DEL PROYECTO, ENTRE 8 Y 12 PALABRAS)**

|  |  |  |  |
| --- | --- | --- | --- |
| Nombre del primer autor  Universidad Eafit  País  Correo electrónico en Eafit | Nombre del segundo autor  Universidad Eafit  País  Correo electrónico en Eafit | 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 = C**ontribución 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**

Para escribir un resumen, debe responder a las siguientes preguntas en un párrafo: ¿Cuál es el problema? ¿Por qué es importante el problema? ¿Cuáles son los problemas relacionados? ¿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**

Explica la motivación, en el mundo real, que lleva al problema. Incluya la historia de este problema. *(En este semestre, la motivación es la razón por la que necesitamos predecir los resultados de los exámenes o el éxito académico en los pregrados de América Latina)*

# **1.1. Problema**

En pocas palabras, explique el problema, el impacto que tiene en la sociedad y por qué es importante resolver el problema. *(En este semestre, el problema es predecir el éxito académico)*

**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**

## Explique cuatro (4) artículos relacionados con el problema descrito en la sección 1.1. Puede encontrar los problemas relacionados en las revistas científicas. Considere el Google Scholar para su búsqueda. *(En este semestre, el trabajo relacionado es la investigación de árboles de decisión para predecir los resultados de los exámenes de los estudiantes o el éxito académico)*

## **3.1 Un título para el primer problema relacionado**

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

## **3.2 Un título para el segundo problema relacionado**

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

## **3.3 Un título para el tercer problema relacionado**

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

## **3.4 Un título para el cuarto problema relacionado**

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

## **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 gradudados 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**

## En lo que sigue, presentamos diferentes algoritmos usardos para construir automáticamente un árbol de decisión binario. *(En este semestre, ejemplos de tales algoritmos son ID3, C4.5 y CART).*

**3.2.1 Nombre del primer algoritmo**

Por favor, explique el algoritmo, su complejidad e incluya una figura vectorizada.

**3.2.2 Nombre del segundo algoritmo**

Por favor, explique el algoritmo, su complejidad e incluya una figura vectorizada.

**3.2.3 Nombre del tercer algoritmo**

Por favor, explique el algoritmo, su complejidad e incluya una figura vectorizada.

**3.2.4 Nombre del cuarto algoritmo**

Por favor, explique el algoritmo, su complejidad e incluya una figura vectorizada.

## **4. DISEÑO DE LOS ALGORITMOS**

## En lo que sigue, explicamos la estructura de los datos y los algoritmos utilizados en este trabajo.

## **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)*

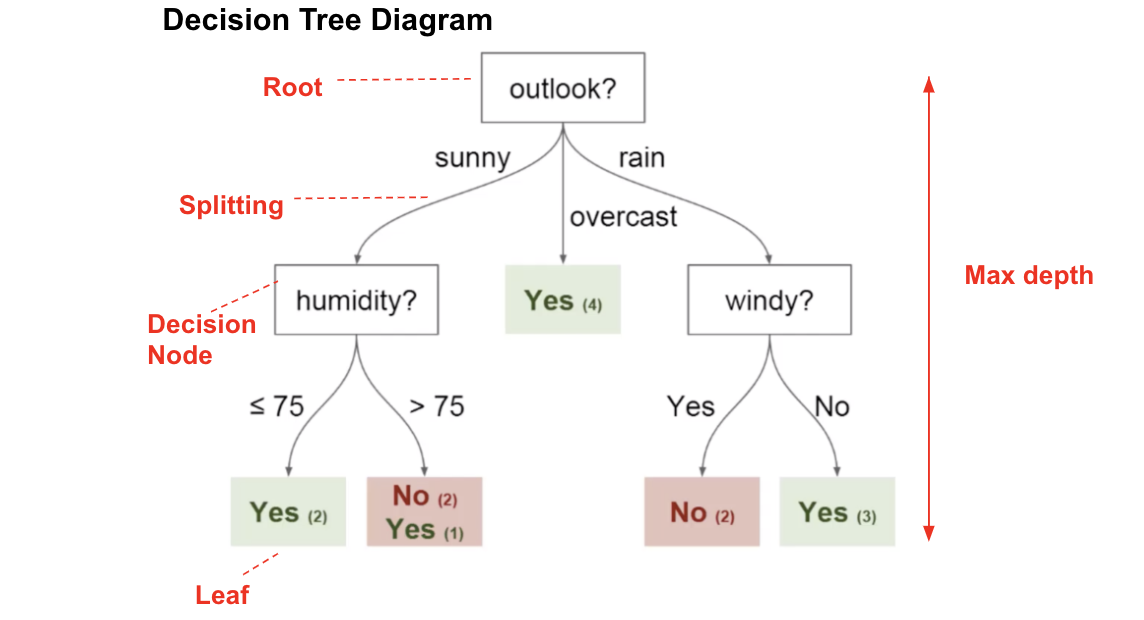
**Figura 1:** Un árbol de decisión binario para predecir Saber Pro basado en los resultados de Saber 11. Los nodos violetas representan a aquellos con una alta probabilidad de éxito, los verdes con una probabilidad media y los rojos con una baja probabilidad de éxito.

**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.



**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(N2\*M2) |
| Validar el árbol de decisión | O(N3\*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.

## 

|  |  |  |  |
| --- | --- | --- | --- |
|  | ***Conjunto de datos 1*** | ***Conjunto de datos 2*** | ***...Conjunto de datos n*** |
| *Tiempo de entrenamiento* | 10.2 s | 20.4 s | 5.1 s |
| *Tiempo de validación* | 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.

|  |  |  |  |
| --- | --- | --- | --- |
|  | ***Conjunto de datos 1*** | ***Conjunto de datos 2*** | ***...Conjunto de datos n*** |
| 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.](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**

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

A modo de ejemplo, consideremos estas dos referencias:

1.Adobe Acrobat Reader 7, Asegúrate de que el texto de las secciones de referencia es está alíneado a la derecha y no justificado.http://www.adobe.com/products/acrobat/.

2. Fischer, G. y Nakakoji, K. Amplificando la creatividad de los diseñadores con entornos de diseño orientados al dominio. en Dartnall, T. ed. Artificial Intelligence and Creativity: An Interdisciplinary Approach, Kluwer Academic Publishers, Dordrecht, 1994, 343-364.