# **Los árboles que predicen el futuro**

|  |  |  |  |
| --- | --- | --- | --- |
| Brahyan Esteban Rios Soto  Universidad Eafit  Colombia   Berioss@eafit.edu.co | Jonathan Smith Julio Diaz  Universidad Eafit  Colombia   jsjuliod@eafit.edu.co | Miguel Correa  Universidad Eafit  Colombia  [macorream@eafit.edu.co](mailto:macorream@eafit.edu.co) | Mauricio Toro  Universidad Eafit  Colombia  [mtorobe@eafit.edu.co](mailto: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**

Los datos son algo de suprema importancia ya que estos nos dan información para tomar decisiones de la manera más segura y precisa posible. Así tener una predicción del éxito académico es de suma importancia; teniendo ese propósito, usaremos arboles de decisión y tomando como éxito el puntaje de la prueba Saber Pro predecir el resultado de los exámenes y con esto tomar medidas pertinentes para el mejoramiento de dicho resultado. ¿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 |

# **INTRODUCCIÓN**

La tecnología es un factor esencial hoy en día, en Colombia la transformación digital en la educación se conoce como educación 4.0 que busca entender a través de medios tecnológicos aspectos de suma importancia en la educación Colombiana, grandes avances se han hecho como lo es la deserción y diferentes factores de motivación en los estudiantes, no obstante al momento de la predicciones el éxito académico se han quedado cortos y con esta problemática en mente buscamos poder llegar a predecir el éxito académico utilizando arboles de decisión ya que es una herramienta excepcional en problemas lineales y no lineales lo nos da más herramientas para lograr nuestro objetivo

# **1.1. Problema**

Predecir el éxito académico, con el fin de tomar medidas que puedan influir de forma significativa y positiva a los resultados académicos de la presente y futuras generaciones

**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)*

## **2.1 Predicción de resultados académicos de estudiantes de informática mediante el uso de redes neuronales**

Gracias a lo que hicieron pudieron predecir los resultados académicos de los estudiantes que cursaban Estructuras de datos y algoritmos I y II del Instituto Superior Politécnico José Antonio Echeverría usando como base calificaciones anteriores entre otros registros previos, como es descrito en el tirulo se utilizaron redes neuronales diseñadas en MATLAB, la precisión fue de un 75% a un 78%. (Álvarez et al. 2016)

## **2.2 Extracción de Conocimiento para la Predicción y Análisis de los Resultados de la Prueba de Calidad de la Educación Superior en Colombia**

## Poder predecir el puntaje de unos estudiantes de Barranquilla en la Saber Pro usando minería de datos, se usa una minería de extracción de datos conocida como KDD, no mencionan una precisión exacta, pero se puede concluir que ronda el 90% (Garcia el al. 2019).

**2.3 ANÁLISIS DEL DESEMPEÑO DE ESTUDIANTES DE INGENIERÍA EN LAS PRUEBAS SABER-PRO**

## Aplicando la metodología de minería de datos CRISP-DM,

## se realiza un estudio de los resultados obtenidos en las pruebas Saber-Pro de estudiantes de ingeniería en

## Antioquia (Colombia).

## A partir de 108 variables académicas, económicas y socio demográficas se realizan 3 modelos analíticos:

## 1) agrupación de los tipos de estudiantes

## 2) selección de los factores que más influyen

## en el desempeño de las pruebas

## 3) predicción del desempeño en las pruebas a partir de las variables seleccionadas.

## **2.4 Evaluación del resultado académico de los**

## **estudiantes a partir del análisis del uso de**

## **los Sistemas de Control de Versiones**

El objetivo de este trabajo es evaluar si el resultado académico de los estudiantes se puede predecir monitorizando su actividad en uno de estos sistemas. Para tal efecto, hemos construido un modelo que predice el resultado de los estudiantes en una práctica de la asignatura Ampliación de Sistemas Operativos, perteneciente al segundo curso del grado en Ingeniería Informática de la Universidad de León. Para obtener la predicción, el modelo analiza la interacción del estudiante con un repositorio Git. Para diseñar el modelo, se evalúan varios modelos de clasificación y predicción utilizando la herramienta MoEv. Esta herramienta permite entrenar y validar diferentes modelos de clasificación y obtener el

más adecuado para un problema concreto.

## **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 usados para construir automáticamente un árbol de decisión binario. *(En este semestre, ejemplos de tales algoritmos son ID3, C4.5 y CART).*

**3.2.1 Algoritmo ID3**

Este algoritmo funciona de forma *recursiva*, los árboles que crea son binarios lo que quiere decir que hay dos opciones por cada atributo y el funcionamiento se basa en la entropía.

Ganancia de atributos = Se refiere a la cantidad de información que se puede extraer de esta, a su vez es inversamente proporcional a la Entropía; al maximizar la ganancia se minimiza la Entropía

El algoritmo funciona insertando un conjunto de ejemplos, un conjunto de *atributos* y un *nodo* el cual puede ser vacío de ese modo se creara uno nuevo, lo que el algoritmo hace es buscar cuál de los *atributos* da mayor información después de eso se creara un nodo con ese atributo para posteriormente crear los *nodos* *hijos* de ese *atributo*, esto utilizando los ejemplos que contengan ese *atributo*; para después volver a llamar la función *recursiva* pero esta vez sin el anterior *atributo* y sus *ejemplos*, así de esta manera se repetirá con todos y cada uno de los *atributos* empezando por los más importantes que son los que mayor *ganancia* proporcionan. El caso base de la función atiende a dos situaciones: 1) Que ya no queden atributos lo que significaría que le árbol está terminado, 2) Llegar a un nodo hoja ósea a una respuesta (Si o No).

La complejidad de este algoritmo se calcula de la siguiente manera:

-Donde n es el número de datos y m el número de

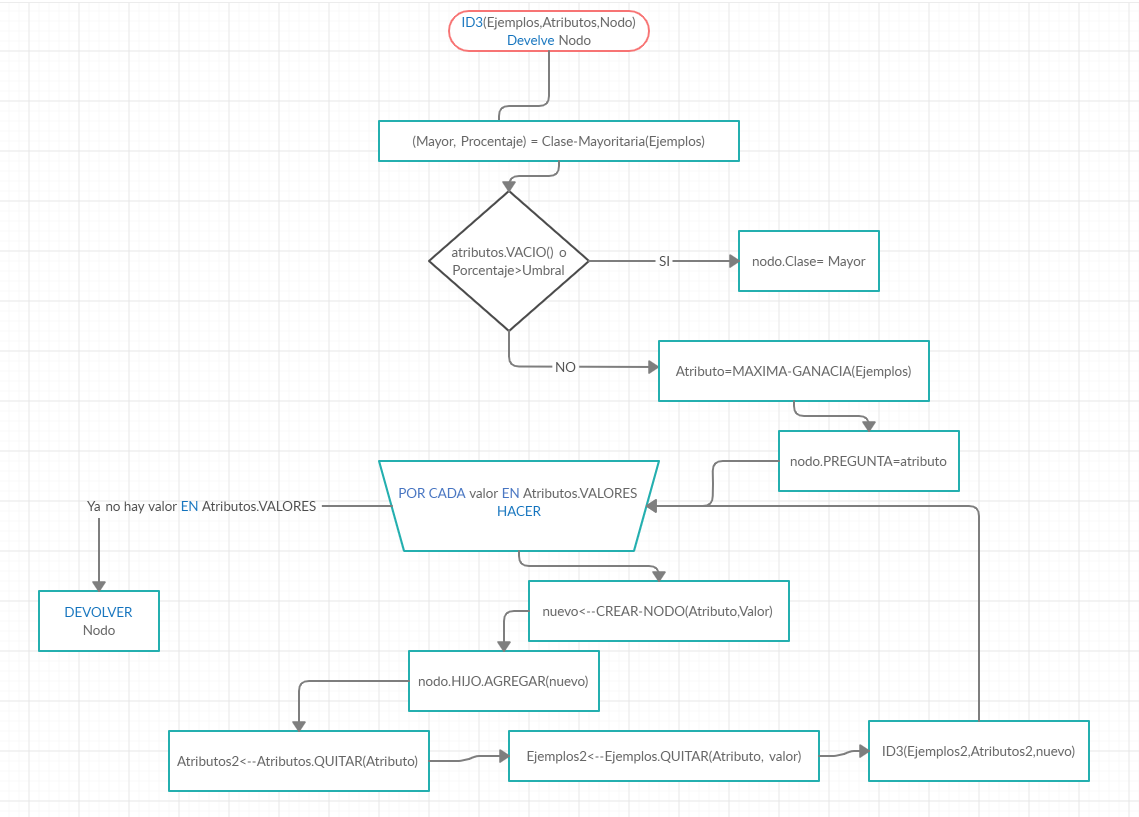
atributos

-El primer término se refiere a construir un árbol de

decisión sin considerar podado. El segundo término se

refiere cuando realizamos podado. Esto es

independiente de que los atributos sean continuos o no.



**3.2.2 Algoritmo C4.5**

Es una mejora del Algoritmo ID3

**3.2.3 Algoritmo CART**

El algoritmo CART es el acrónimo de Classification And Regression Trees (Árboles de Clasificación y de Regresión).

Este modelo admite variables generadas por el entorno, así como también variables de salida que sean, nominales, ordinales y continuas, por lo que se pueden resolver tanto problemas de clasificación como de regresión.

Este mismo hace uso principalmente del Índice GINI para calcular la homogeneidad, es decir, la pureza de un nodo en específico:

Índice GINIdel nodo = Peso relativo\*Índice GINI Sub-nodo

Índice GINI de los Sub-nodos = p2+q2

Donde tenemos que p son aquellos datos que, si cumplen una dicha condición, mientras que q, son aquellos que no.

Ejemplo:

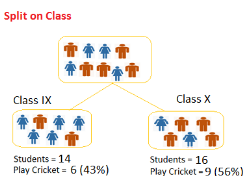
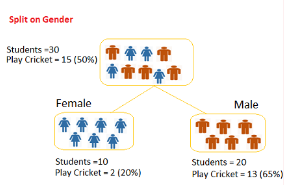
- 30 estudiantes

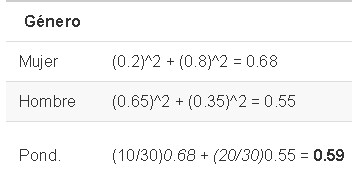
- 2 Variables: Genero (Hombre/Mujer) y Clase (IX/X)

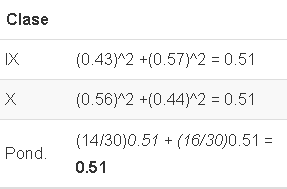
- 15 estudiantes juegan cricket

- Crear un modelo para predecir quien jugará cricket

- Segregar estudiantes basados en todos los valores de las 2 variables e identificar aquella variable que crea los conjuntos más homogéneos de estudiantes y que a su vez son heterogéneos entre ellos.







Dando como mejor variable para segregar los datos, el género, ya que esta es de mayor magnitud que la de Clase.

**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)*

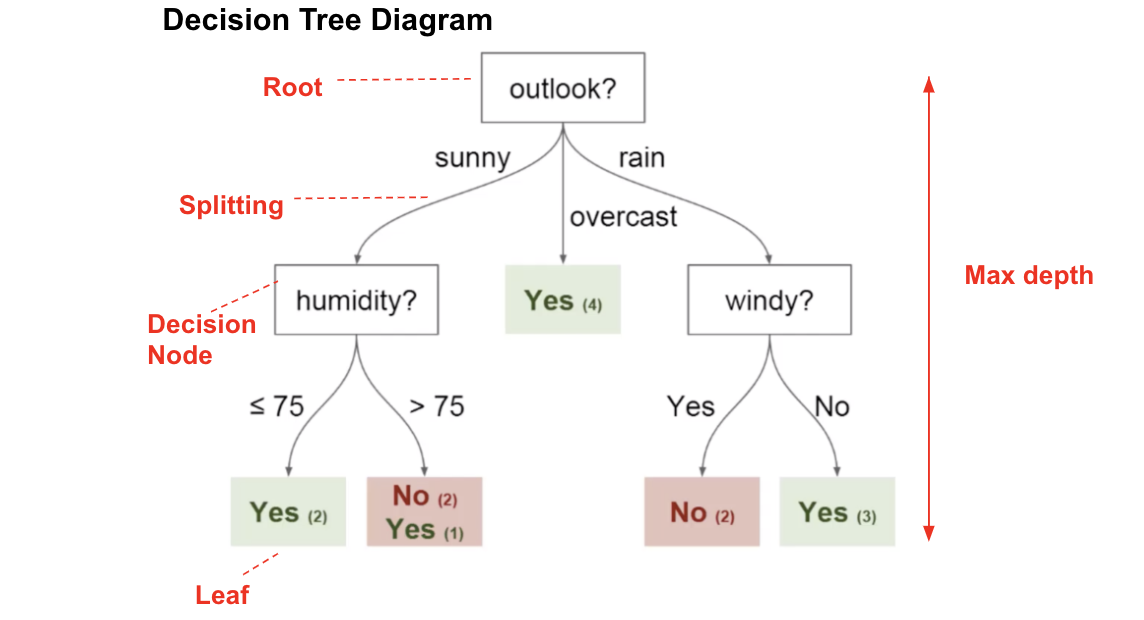
**Figura 1:** Un árbol de decisión binario para predecir Saber Pro basado en los resultados de Saber 11. Los nodos violetas representan a aquellos con una alta probabilidad de éxito, los verdes con una probabilidad media y los rojos con una baja probabilidad de éxito.

**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**

Las 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:

Álvarez Blanco, Jorge, Lau Fernández, Rogelio, Pérez Lovelle, Sonia, & Leyva Pérez, Exiquio C. Predicción de resultados académicos de estudiantes de informática mediante el uso de redes neuronales. *Ingeniare. Revista chilena de ingeniería*, Facultad de Ingeniería Informática. Instituto Superior Politécnico José Antonio Echeverría, La Habana- Cuba, 2016, *24*(4), 715-727.

Jose G.G, Paola S.S, Manuel O.,Sergio O. Extracción de Conocimiento para la Predicción y Análisis de los Resultados de la Prueba de Calidad de la Educación Superior en Colombia. Formación universitaria, Facultad de Ingenierías, Universidad Simón Bolívar, Barranquilla - Colombia. 2019, 12(4), 55-62.

A. Oviedo, J. Jiménez. “Minería de datos educativos: análisis del desempeño de estudiantes de ingeniería en las pruebas SABER-PRO”, Revista Politécnica, vol. 15, no.29 pp.128-140

Gutiérrez Fernández, A., Guerrero Higueras, A. M., Conde González, M. A., y Fernández Llamas. “Evaluación del resultado académico de los estudiantes a partir del análisis del uso de los Sistemas de Control de Versiones”. RIED(Revista Iberoamericana de Educación a Distancia), vol. 23 no.2, pp. 127-145.