ARBOLES DE DECISIÓN PARA PREDICCIÓN DEL ÉXITO EN PRUEBAS SABER PRO

Tomas Atehortua Ceferino Universidad Eafit Colombia tatehortuc@eafit.edu.co	Sebastian Velez Galeano Universidad Eafit Colombia svelezg4@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

El objetivo de este informe es analizar y predecir la probabilidad que tiene un estudiante de obtener un puntaje total, superior al promedio de su cohorte, en las pruebas Saber Pro (pruebas estandarizadas que realiza el gobierno colombiano al final de la carrera) por medio de árboles de decisión y datos de la prueba Saber11.

La solución de este problema es de gran importancia para el futuro del país ya que brindaría los estándares de los próximos profesionales, dando así información para mejorar la formación de estos.

Uno de los problemas relacionados con el planteado es verificar la calidad de educación superior y de futuros en del país.

¿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

En un futuro cercano, el papel de la tecnología será un factor clave en el proceso de transformación digital de la educación en Colombia. En el pasado, se han estudiado qué factores influyen en la deserción académica, cuáles son sus causas y motivaciones, y se han utilizado algoritmos para predecir la deserción. No obstante, es poco lo que se ha logrado para predecir el éxito académico en educación superior.

Para efectos de este proyecto, vamos a definir el éxito académico como la probabilidad que tiene un estudiante de obtener un puntaje total, superior al promedio de su cohorte, en las pruebas Saber Pro (pruebas estandarizadas que realiza el gobierno colombiano al final de la carrera).

1.1. PROBLEMA

El problema al cual nos enfrentamos se basa en predecir el éxito académico diseñando un algoritmo, basado en arboles de decisión y en los datos del saber 11, para predecir si un estudiante tendrá un puntaje total en las pruebas Saber Pro, por encima del promedio o no.

Resolver este problema sería un gran aporte para el futuro del país ya que otorgaría una media de la calidad de sus futuros profesionales, brindando así posibles cambios o mejoras en el sistema de educación superior colombiano.

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

2.1 Predecir la tasa de deserción de los estudiantes

Predijeron la tasa de deserción después del primer semestre de su carrera e identificaron factores críticos de éxito asociados a ese programa de estudios. A través de árboles de decisión (CART) y (C4.5) lograron predecir la deserción estudiantil con una precisión de entre el 75% y 81% [1].

2.2 Predecir la tasa de éxito de los estudiantes de una institución de educación superior en sus cursos

Predijeron las tasas de aprobación de asignaturas cursadas por estudiantes universitarios. Los autores sometieron a consideración los datos de 106 individuos inscritos en asignaturas del área informática y comprobaron que los algoritmos de árboles de decisión son igualmente eficientes para predecir futuros comportamientos en muestras pequeñas. Con el uso de distintos paquetes de software para ejecutar los algoritmos de análisis (C4.5) (J48), encontraron un 90% de precisión en la predicción de la aprobación de las asignaturas [2].

2.3 Predicción de factores asociados al desempeño académico de estudiantes de bachillerato en las pruebas Saber 11°

Se generaron árboles de decisión (J48) que permitieron identificar patrones asociados al buen o mal desempeño académico de los estudiantes en las pruebas Saber 11°. Los patrones descubiertos ayudarán en los procesos de toma de decisiones del Ministerio de Educación Nacional, junto con las instituciones que velan por la calidad de la educación en Colombia.

Se llegó a varias "reglas" o conclusiones que indican el porcentaje de estudiantes que clasifican por encima o debajo de la media [3].

2.4 Predicción temprana del éxito de los estudiantes

Se estudió los factores que predicen el éxito de estudiantes en una institución de educación superior de Nueva Zelandia. Utilizando una muestra de 450 estudiantes que cursaron una clase de Sistemas de Información.

Desafortunadamente, la precisión de la clasificación de los árboles de clasificación no fue muy alta. En el caso del árbol CHAID, la precisión general de la clasificación fue del 59,4% y en el caso del Árbol CART ligeramente superior al 60,5%. Esto sugeriría que la información de antecedentes (género, edad, etnia, discapacidad, escuela secundaria, situación laboral e inscripción temprana) recopilados durante el proceso de inscripción, no contiene suficiente información para una separación precisa de estudiantes exitosos y no exitosos [4].

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 Algoritmo ID3

El ID3 es un algoritmo simple pero potente, cuya misión es la elaboración de un árbol de decisión bajo las siguientes premisas:

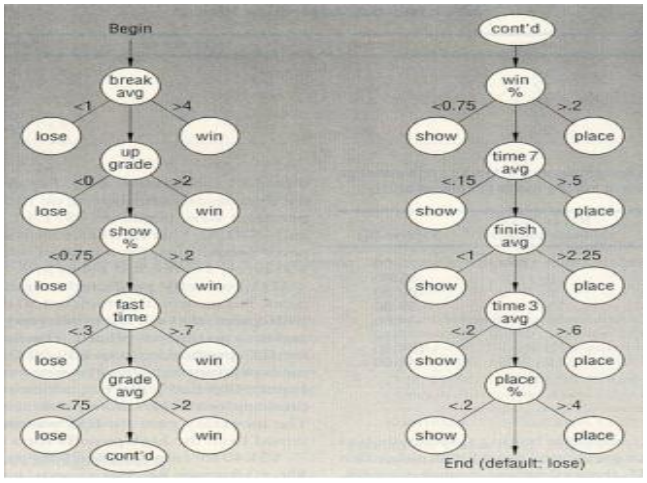
1. Cada nodo corresponde a un atributo y cada rama al valor possible de ese atributo. Una hoja del árbol especifica el valor esperado de la decisión de acuerdo con los ejemplos dados. La explicación de una determinada decisión viene dada por la trayectoria desde la raíz a la hoja representativa de esa decisión.
2. A cada nodo es asociado aquel atributo más informativo que aún no haya sido considerado en la trayectoria desde la raíz.
3. Para medir cuánto de informativo es un atributo se emplea el concepto de entropía. Cuanto menor sea el valor de la entropía, menor será la incertidumbre y más útil será el atributo para la clasificación.

El ID3 es capaz de tratar con atributos cuyos valores sean discretos o continuos. En el primer caso, el árbol de decisión generado tendrá tantas ramas como valores posibles tome el atributo. Si los valores del atributo son continuos, el ID3 no clasifica correctamente los ejemplos dados. Por ello, se propuso el C4.5, como extensión del ID3.

La estructura del arbol estpa compuesta por:

- Nodos: Los cuales contendrán atributos.
- Arcos: Los cuales contiene valores posibles del nodo padre
- Hojas: Nodos que clasifican el ejemplo como positive o negativo

Ejemplo de Árbol ID3:



3.2.2 Algoritmo C4.5

Es un algoritmo de inducción que genera una estructura de reglas o árbol a partir de subconjuntos (ventanas) de casos extraídos del conjunto total de datos de "entrenamiento". En este sentido, su forma de procesar los datos es parecido al de Id3. El algoritmo genera una estructura de reglas y evalúa su "bondad" usando criterios que miden la precisión en la clasificación de los casos. Emplea dos criterios principales para dirigir el proceso dados por:

1. Calcula el valor de la información proporcionada por una regla candidata (o rama del árbol), con una rutina que se llama "info".

2. Calcula la mejora global que proporciona una regla/rama usando una rutina que se llama gain (beneficio).

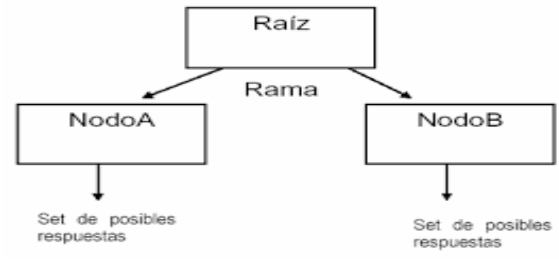
Con estos dos criterios se puede calcular una especie de calor de coste/beneficio en cada ciclo del proceso, que le sirve para decidir si crear, por ejemplo, dos nuevas reglas, o si es mejor agrupar los casos de una sola.

El algoritmo realiza el proceso de los datos en sucesivos ciclos. En cada ciclo se incrementa el tamaño de la "ventana" de proceso en un porcentaje determinado respecto al conjunto total. El objetivo es tener reglas a partir de la ventana que clasifiquen correctamente a un número cada vez mayor de casos en el conjunto total. Cada ciclo de proceso emplea como punto de partida los resultadosconseguidos por el ciclo anterior.

La estructura del árbol esta compuesta por dos tipos de nodos:

- Una hoja (nodo terminal), que indica una clase.
- Un nodo de decisión, que especifica una comprobación a realizar sobre el valor de una variable. Tiene una rama y un subárbol para cada resultado posible de la comprobación.

Ejemplo de Árbol C4.5:



3.2.3 Algoritmo CART

Con este algoritmo, se generan árboles de decisión binarios, lo que quiere decir que cada nodo se divide en exactamente dos ramas

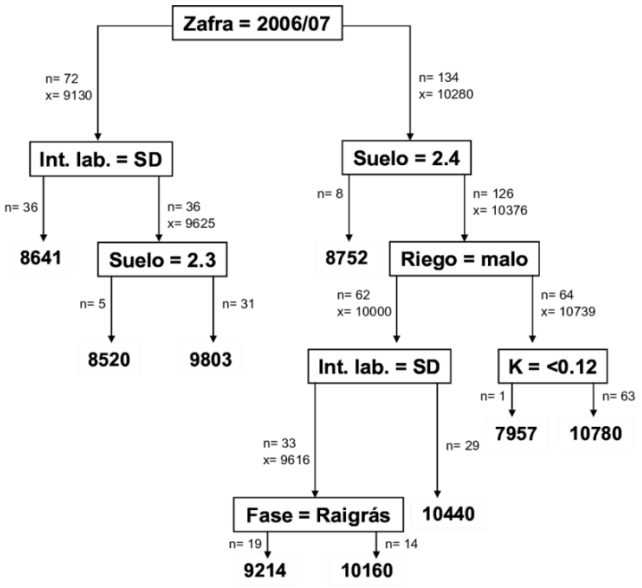
La metodología CART utiliza datos históricos para construir arboles de clasificación o de regresión los cuales son usados para clasificar o predecir nuevos datos. Estos árboles CART pueden manipular fácilmente variables numéricas y/o categóricas. Entre otras ventajas esta su robustez a outliers, la invarianza en la estructura de sus ´arboles de clasificación o de regresión a transformaciones monótonas de las variables independientes, y sobre todo, su interpretabilidad.

Esta metodología consiste en tres pasos:

- Construcción del árbol saturado
- Escogencia del tamaño correcto del árbol
- Clasificación de nuevos datos usando el árbol construido

La construcción del árbol saturado se hace con particionamiento recursivo. La diferencia en la construcción de los árboles de clasificación y los ´arboles de regresión es el criterio de división de los nodos, es decir, la medida de impureza es diferente para los ´arboles de clasificación y de regresión.

Ejemplo de Árbol CART:



3.2.4 Algoritmo CHAID

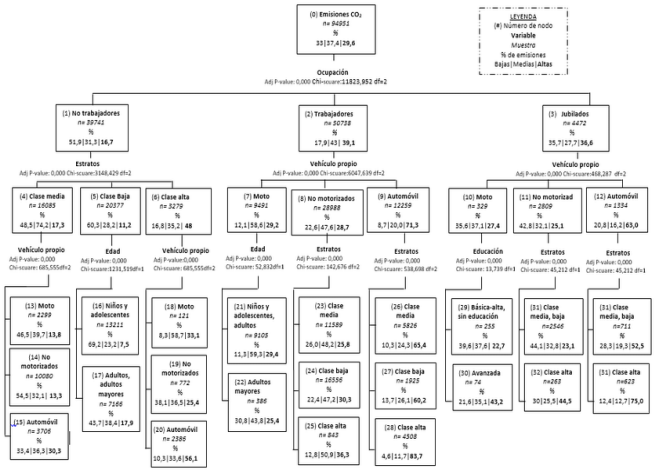
CHAID consiste en un algoritmo para la construcción de árboles de decision basado en el testeo de significancia ajustada. Su nombre proviene de “Chisquared Automatic Interaction Detection”. Fue publicado en 1980 por Gordon V. Kass y proviene del clásico AID, con algunas particularidades añadidas.

Se distingue por el uso de la chi-cuadrado, χ^2, para medir el grado de correlación entre las variables independientes y la clase.

Sus principals caracteristicas son:

- Función de división.
- Tip de variables.
- Valores missing.
- Poda de árbol

Ejemplo Árbol CHAID:



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)*

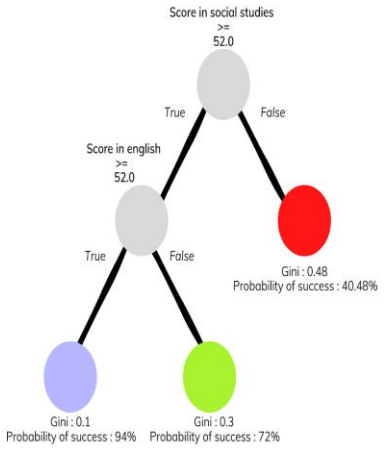


Figura 1: Un árbol de decisión binario para predecir Saber Pro basado en los resultados de Saber 11. Los nodos violetas representan a aquellos con una alta probabilidad de éxito, los verdes con una probabilidad media y los rojos con una baja probabilidad de éxito.

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.

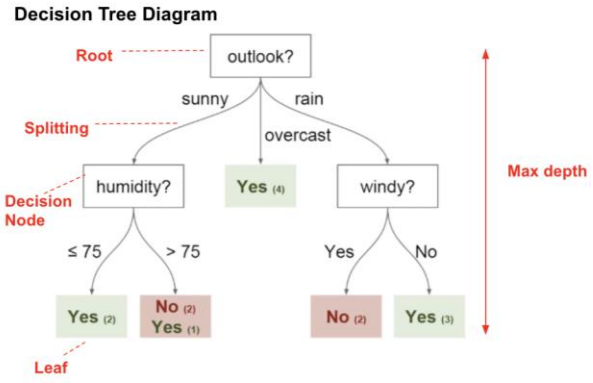


Figura 2: Entrenamiento de un árbol de decisión binario usando *(En este semestre, uno podría ser CART, ID3, C4.5... por favor, elija)*. En este ejemplo, mostramos un modelo para predecir si se debe jugar al golf o no, según el clima.

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. 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. Dekker, G., Pechenizkiy, M. y Vleeshouwers, J. Predicting students drop out: A case study. in International Conference on Educational Data Mining (EDM), (Cordoba, Spain, 2009), International Working Group on Educational Data Mining, 41-50. https://eric.ed.gov/?id=ED539082

2. Natek, S. y Zwilling, M. Student data mining solution-knowledge management system related to higher education institutions, Expert Systems with Applications, 41(14), 6400-6407. http://iranarze.ir/wp-content/uploads/2017/08/7548-English-IranArze.pdf

3. Timarán, R., 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), 363-378. https://doi.org/10.19053/20278306.v9.n2.2019.9184.

4. Kovačić, Z. J., Early prediction of student success: Mining student enrollment data. In Proceedings of Informing Science & IT Education Conference, (Wellington, New Zealand ,2010), The Open Polytechnic Of New Zealand Publications, 647-665. https://pdfs.semanticscholar.org/e48e/ba98bde33586c20442d46ab9a59c411196e5.pdf