DESEMPEÑO DE LOS ESTUDIANTES USANDO ARBOLES DE DESICIO

Kevin Mauricio Loaiza Arango	Miguel Correa	Mauricio Toro
Universidad Eafit	Universidad Eafit	Universidad Eafit
Colombia	Colombia	Colombia
kmloaizaa@eafit.edu.co	macorream@eafit.edu.co	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

Para escribir un resumen, debe responder a las siguientes preguntas en un párrafo: ¿Cuál es el problema? Consiste en usar los datos sacados de los estudiantes y de sus resultados de las pruebas 11, además de otros datos, para así hacer un posible resultado de sus pruebas saber pro. ¿Por qué es importante el problema? Es importante por la información que puede brindar a las instituciones acerca de los avances que tienen en la educación de sus estudiantes ¿Cuáles son los problemas relacionados? Los problemas relacionados son el uso de arboles de decisión para la predicción de resultados, predecir el éxito académico y generar una forma de aprendizaje automático. ¿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. *La motivación que nos lleva a realizar esto es, dar un estudio mediante un algoritmo para calcular el éxito académico en las pruebas saber pro, esto principalmente para darle a la institución y su comunidad la oportunidad de mejorar antes de que se realicen dichas pruebas, además para dar información de antemano de la calidad académica de esta institución*

1.1. Problema

En pocas palabras, explique el problema, el impacto que tiene en la sociedad y por qué es importante resolver el problema. *El problema es básicamente utilizar información de los estudiantes, de sus antecedentes para calcular cuál será su desempeño en las pruebas saber pro, la importancia de este estudio es dar un precedente del avance o camino que está llevando cada estudiante para el desempeño de las pruebas, siendo este, una ayuda para saber si debe prepararse más para la prueba, el impacto de este proyecto en la sociedad creemos que se puede orientar a la preparación de los estudiantes ante las pruebas, como al desarrollo de nuevas técnicas de aprendizaje.*

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!*).

Nuestra solución será la implementación del algoritmo ID3 para tener una mejor predicción del éxito académico en Saber Pro utilizando datos de Saber 11 presentados por graduados de secundaria. Este algoritmo es realmente útil al intentar dividir conjuntos de datos, mejorando la eficiencia en el tiempo de ejecución.

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 Predecir rendimiento académico de los estudiantes

El problema que se estudió y resolvió, fue el rendimiento de los estudiantes de secundaria, se utilizaron dos algoritmos el C4.5 y el ID3, la precisión que se logro con ambos algoritmos fue del 75.145%. ADHATRAO, K., GAYKAR, A., DHAWAN, A., JHA, R. Y HONRAO, V.

1 **PREDICTING STUDENTS' PERFORMANCE USING ID3 AND C4.5 CLASSIFICATION ALGORITHMS**

En el texto: (Adhatrao, Gaykar, Dhawan, Jha & Honrao, 2013)

Bibliografía: Adhatrao, K., Gaykar, A., Dhawan, A., Jha, R., & Honrao, V. (2013). *PREDICTING STUDENTS' PERFORMANCE USING ID3 AND C4.5 CLASSIFICATION ALGORITHMS* [Ebook] (3rd ed., pp. 1-14). Navi Mumbai: International Journal of Data Mining & Knowledge Management Process (IJDKP). Retrieved from https://arxiv.org/ftp/arxiv/papers/1310/1310.2071.pdf

3.2 Minería de datos estudiantiles con árboles de decisión.

El uso de la minería de datos para obtener información del desempeño de los estudiantes y asi cuestionar el éxito de la intitucion, se usaron 3 algoritmos ID3, C4.5 y Naive Bayes, la precisión de estos en 2 pruebas fueron 38.4615% – 28.3186%, 35.8974% – 38.0531%, 33.3333% - 38.03531%, respectivamente. AL-RADAIDEH, Q. A., AL-SHAWAKFA, E. M. Y AL-NAJJAR, M. I.

2 **Mining Student Data Using Decision Trees**

En el texto: (Al-Radaideh, Al-Shawakfa & Al-Najjar, 2006)

Bibliografía: Al-Radaideh, Q., Al-Shawakfa, E., & Al-Najjar, M. (2020). *Mining Student Data Using Decision Trees* [Ebook] (1st ed., pp. 1-5). Jordan: The 2006 International Arab Conference on Information Technology (ACIT'2006). Retrieved from https://www.acit2k.org/ACIT2006/Proceeding/131.pdf

3.3 Rendimiento académico mediante el uso de arboles de clasificación y regresión.

En este estudio la prioridad fue realizar una estadística de rendimiento estudiantil para buscar los causales de deserción y así erradicarla. El algoritmo utilizado fue el C4.5 y tuvo una precisión del 78%.

KRISHNA, M., RANI, B. S. B. P., CHAKRAVARTHI, G. K., MADHAVRAO, B. Y CHOWDARY, S. M. B.

3 **Predicting Student Performance using Classification and Regression Trees Algorithm**

En el texto: (Krishna, Rani, Chakravarthi, Madhavrao & Chowdary, 2020)

Bibliografía: Krishna, M., Rani, B., Chakravarthi, G., Madhavrao, B., & Chowdary, S. (2020). *Predicting Student Performance using Classification and Regression Trees Algorithm* [Ebook] (9th ed., pp. 1-8). International Journal of Innovative Technology and Exploring Engineering (IJITEE). Retrieved from http://www.ijitee.org/wp-content/uploads/papers/v9i3/C8964019320.pdf

3.4 Éxito académico mediante el uso de arboles de decisión.

Este estudio se encargo de hacer un futuro posible del futuro de los estudiantes mediante el desempeño de la secundaria, sus exámenes y su desempeño en primer año y así calcular la posibilidad de éxito académico en el futuro, algunos de los algoritmos usados para esto fueron REPTree decisión tree, ID3, Naive Bayes, C5.0 y C4.5, tuvieron una precisión de 79%, 40.63%, 76.65%, 93%, 67.78% respectivamente.

MESARIĆ, J. Y ŠEBALJ, D.

4 **Decision trees for predicting the academic success of students**

En el texto: (Mesarić & Šebalj, 2016)

Bibliografía: Mesarić, J., & Šebalj, D. (2016). *Decision trees for predicting the academic success of students* [Ebook] (1st ed., pp. 1-22). Osijek: Croatian Operational Research Review. Retrieved from http://bib.irb.hr/datoteka/853222.clanak.pdf

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).*

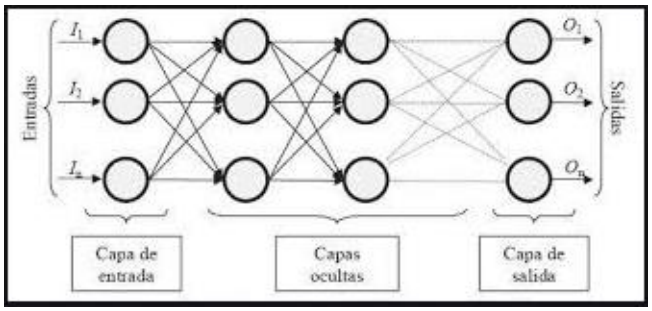
3.2.1 ID3

Este algoritmo consiste en elegir el mejor atributo para compararlo en la raíz, el proceso de selección del mejor atributo cada atributo de instancia se pone en una prueba de estadística para determinar que calificación tiene.

Su complejidad se define por:

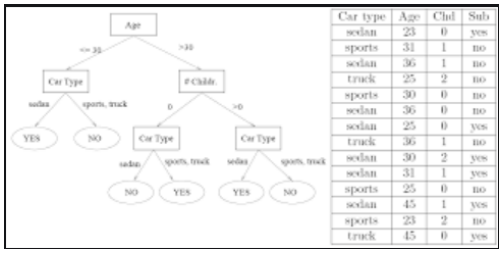
$$\text{Entropia}(S) = \sum_{i=1}^{n} -p_i \log_2 p_i$$

Y una figura vectorizada seria



3.2.2 CARD

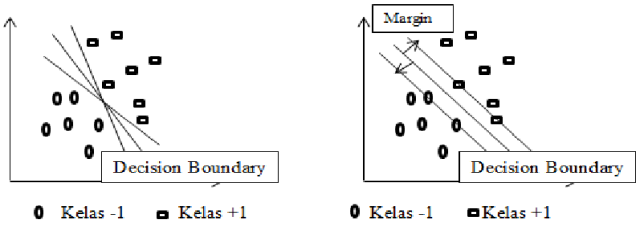
Su funcionamiento como tal es generar múltiples arboles binarios, de estos existe uno con el tamaño correcto que se identifica al evaluar el rendimiento predictivo de cada árbol. Su complejidad está dado por cada variable que crea un grupo para una división.



Car type	Age	Chd	Sub
sedan	23	0	yes
sports	31	1	no
sedan	36	1	no
truck	25	2	no
sports	30	0	no
sedan	36	0	no
sedan	25	0	yes
truck	36	1	no
sedan	30	2	yes
sedan	31	1	yes
sports	25	0	no
sedan	45	1	yes
sports	23	2	no
truck	45	0	yes

3.2.3 C4.5

Es una extensión del ID3, este se basa en usar el criterio de ratio para evitar que las variables de mayor número de posibles valores salgan beneficiados en la selección.



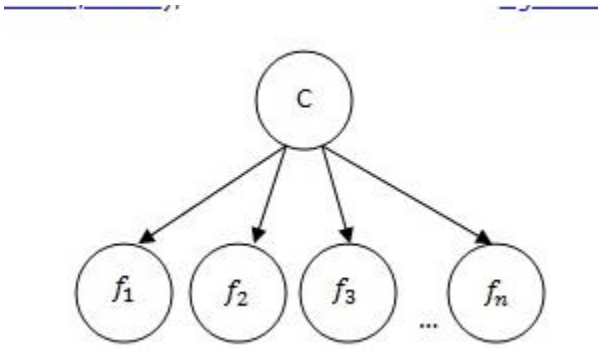
3.2.4 Naive Bayes

Por favor, explique el algoritmo, su complejidad e incluya una figura vectorizada.Se basa en la técnica estadística llamada teorema de Bayes. Funciona creando una tabla de probabilidad, esta calcula los eventos posibles.

La formula es:

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

Su vector seria:



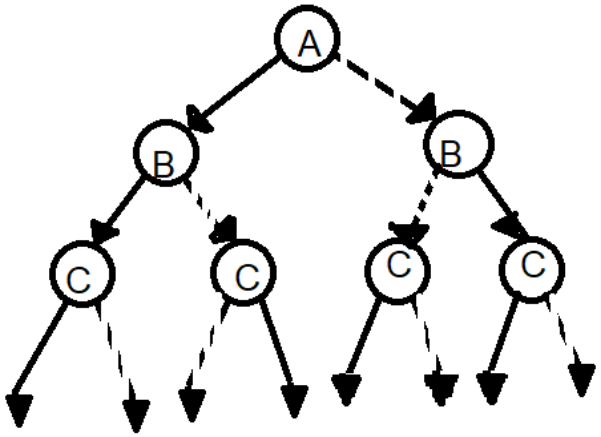
4. DISEÑO DE LOS ALGORITMOS

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[1].

4.1 Estructura de los datos

Usaremos un árbol de decisión binario implementando el algoritmo ID3, con la intención de optimizar el rendimiento buscando dividir el conjunto de datos más grande en subconjuntos con los que sea más fácil trabajar y con mayor capacidad de prueba.

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

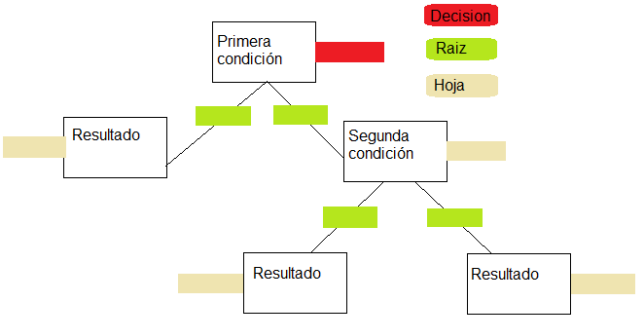


.

4.2 Algoritmos

El algoritmo de clasificación funciona seleccionando determinadas variables (y condiciones que deben cumplir estas variables). Al iterar por cada una de estas variables y obtener diferentes valores, se calculará el índice de Gini (cuanto más bajo, mejor).

El algoritmo de desarrollo se encarga de procesar estos datos clasificados y devolver un resultado (el éxito previsto).



4.2.1 Entrenamiento del modelo

La creación del árbol (y generación de sus hojas) se realizará bajo la premisa de que las “decisiones” que tengan / devuelvan un índice de Gini menor son las más correctas para la continuación del árbol; descartando así los nodos que tienen un índice más alto. Además, el algoritmo evaluará cada uno de los nodos y luego (en caso de obtener

[1]http://www.github.com/ ????????? /proyecto/

un resultado satisfactorio) los dividirá en "sub nodos" y optimizará los resultados..

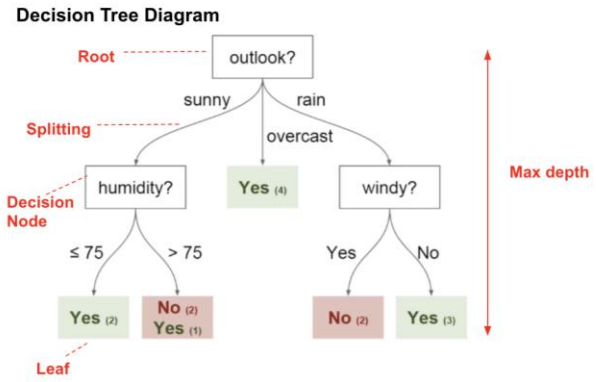


Figura 2: Entrenamiento de un árbol de decisión binario usando *(En este semestre, uno podría ser CART, ID3, C4.5... por favor, elija)*. En este ejemplo, mostramos un modelo para predecir si se debe jugar al golf o no, según el clima.

4.2.2 Algoritmo de prueba

Luego de procesar los datos el algoritmo toma la información para la creación del 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

Adhatrao, K., Gaykar, A., Dhawan, A., Jha, R., & Honrao, V. (2013). Navi Mumbai, PREDICTING STUDENTS' PERFORMANCE USING ID3 AND C4.5 CLASSIFICATION ALGORITHMS (3rd ed., pp. 1-14). Presentado por International Journal of Data Mining & Knowledge Management Process (IJDKP).

Al-Radaideh, Al-Shawakfa & Al-Najjar, 2006. Jordan, Mining Student Data Using Decision Trees (1st ed., pp. 1-5). Presentado por The 2006 International Arab Conference on Information Technology (ACIT'2006).

Krishna, Rani, Chakravarthi, Madhavrao & Chowdary, 2020. Predicting Student Performance using Classification and Regression Trees Algorithm (9th ed., pp. 1-8). Presentado por International Journal of Innovative Technology and Exploring Engineering (IJITEE).

Mesarić & Šebalj, 2016. Osijek, Decision trees for predicting the academic success of students (1st ed., pp. 1-22). Presentado por Croatian Operational Research Review.