PREDICCIÓN DEL ÉXITO ACADÉMICO EN EDUCACIÓN SUPERIOR UTILIZANDO ÁRBOLES DE DECISIÓN

Isabella Pérez Universidad Eafit Colombia iperezs2@eafit.edu.co	Mariana Quintero Universidad Eafit Colombia mquintero3@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 predecir el éxito académico en educación superior teniendo en cuenta las variables sociodemográficas y académicas de cada estudiante obtenidas con anterioridad, para esto se busca desarrollar un programa y estructura de datos que nos permita leer, guardar y organizar los datos de los archivos a usar, con el fin de que se logre aplicar herramientas útiles como los árboles de decisión y algoritmos matemáticos, y con ellos predecir la probabilidad que tiene un estudiante de obtener un resultado por encima del promedio.

La solución a este problema es de gran importancia debido a que a pesar de que la tecnología avanza exponencialmente, es bastante poco lo que se ha logrado predecir respecto al éxito académico en educación superior, y sería de increíble utilidad poder tener un dato de predicción más seguro con el fin de evitar tomar exámenes con poca probabilidad de éxito, por lo que el desarrollo eficiente de este proyecto puede ofrecer la capacidad de saber con mayor seguridad cuál sería esta probabilidad de éxito que tiene cada persona en una prueba de educación superior. Existen varios problemas similares al que se plantea en este informe y algunos de estos serán analizados con el propósito de hallar una solución más efectiva. ¿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

Está claro que el papel de la tecnología es un factor clave en la educación Colombiana que poco a poco se convierte más en una parte vital de este, lo que nos lleva inevitablemente a querer dar pasos nuevos o avances en la situación. Anteriormente se han visto estudios acerca de los factores que influyen en el desempeño académico de estudiantes y otras variables como la deserción y sus causas o motivaciones, y en base a eso se ha buscado obtener resultados de predicción respecto a ese tema. Sin embargo, respecto a la predicción de éxito académico en educación superior es muy poco lo que se ha logrado puesto que para medir el éxito se pueden considerar diversas maneras tales como la empleabilidad del egresado, el salario de los egresados, la felicidad del trabajo de los egresados, entre otros. Teniendo en cuenta lo anterior, la finalidad de este proyecto es, entonces, lograr predicciones respecto al éxito académico en educación superior, tomando este como la probabilidad que tiene un estudiante de obtener un puntaje mayor al promedio en las pruebas saber Pro.

1.1. Problema

Mediante el uso de algoritmos basados en árboles de decisión, y en datos académicos y sociodemográficos de un estudiante diseñar y poner en práctica un programa que permita predecir el éxito académico del estudiante; con este pronóstico el estudiante conoce a definitiva al nivel que se encuentra respecto al promedio, y así tomar decisiones acerca de su vida académica (universitaria), finalmente la

sociedad puede prepararse de manera temprana para estas pruebas gracias al resultado que genera el algoritmo.

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?)*

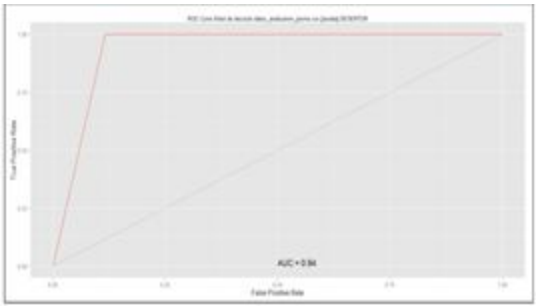
1.3 Estructura del artículo

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 Deserción estudiantil basado en árboles de decisión

Este algoritmo busca pronosticar la probabilidad de que un estudiante abandone su programa académico, a través de técnicas de clasificación, basadas en árboles de decisión; por medio de Classification and Regression Tree (CART) se construyó un árbol con cuatro niveles de profundidad que evalúa a los posibles desertores. Llevando a concluir que las variables nivel y notas tienen mayor influencia en la deserción. La siguiente imagen (figura 1) muestra la curva ROC, en donde se concluye que el modelo cuenta con un 94% de efectividad en predicción.[1]



(figura 1).

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

Este algoritmo busca detectar factores asociados al desempeño académico de los estudiantes colombianos de grado undécimo de educación media en las pruebas Saber 11°, este modelo permite predecir cuales son los factores socioeconómicos, académicos e institucionales asociados al puntaje en las pruebas Saber 11°. Se utilizó un modelo de clasificación basado en árboles de decisión, utilizando el algoritmo J48 de la herramienta WEKA, el cual implementa al algoritmo C.45, el algoritmo J48 se basa en el uso del criterio de ganancia de información, de esta manera se consigue evitar que las variables con mayor número de posibles valores salgan beneficiadas en la selección. El parámetro más importante que se tuvo en cuenta para la poda fue el factor de confianza C (confidence level), que influye en el tamaño y capacidad de predicción del árbol construido. Cuanto más baja se haga esa probabilidad, se exigirá que la diferencia en los errores de predicción antes y después de podar sea más significativa para no podar. Finalmente se obtuvo la precisión del 67%. [2]

2.3 Predicción del rendimiento académico aplicando técnicas de minería de datos

Este algoritmo busca predecir el rendimiento académico de los estudiantes en instituciones educativas superiores, con la finalidad de predecir la clasificación final (Desaprobado o Aprobado) de los futuros estudiantes matriculados de los diferentes cursos. Se utilizan los algoritmos de árboles de decisión y redes bayesianas con el programa Weka y con validación cruzada 10 folds; se comparan varias TMD (C4.5, ID3, CART y J48, Naive de Bayes, Redes Neuronales, k-medias y k-vecino más cercano) para predecir el rendimiento académico. Finalmente se usaron datos de la universidad de Tailandia, esto arrojó que el árbol de decisión tuvo una precisión de 85,2%. [3]

2.4 Predicción de la deserción estudiantil en la Educación Superior Privada.

Este algoritmo busca predecir la deserción o el abandono en la Educación Superior Privada, se utilizó la metodología CRIPS-DM con spss clementine 12.0 de minería de datos árboles de decisión. Se usó 1761 datos de estudiantes de la Universidad Privada César Vallejo de la Escuela profesional de Ingeniería de Sistemas, para el análisis de estos datos se manejó 27 atributos para cada uno de ellos que están relacionadas con la deserción del alumno. Finalmente se hizo entrenamiento, validación y prueba con

100 datos nuevos en donde se obtuvo una precisión de 89%. [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 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

3.2.1 Algoritmo ID3

Este tipo de árbol de decisión se basa en una búsqueda codiciosa, en la cual se elige la opción más beneficiosa, este algoritmo divide continuamente las propiedades en varios grupos. El enfoque codicioso se utiliza cuando cada ganancia puede recogerse en cada paso, por lo que ninguna opción bloquea otra, finalmente el algoritmo escoge como la mejor característica como la que genera más ganancia actual de información.

3.2.2 Classification And Regression Trees (Árboles de Clasificación y de Regresión) - CART

El modelo admite variables de entrada y salida nominales, ordenadas y continuas, así que este puede resolver problemas de clasificación y regresión, selecciona el corte que conduce al mayor decrecimiento de la impureza. Este algoritmo usa el índice de gini para calcular la medida de

impureza (figura 2). El corte en cada nodo viene dado por reglas de tipo binario.

$$G(A_i) = \sum_{j=i}^{M_i} p(A_{ij}) G(C/A_{ij})$$

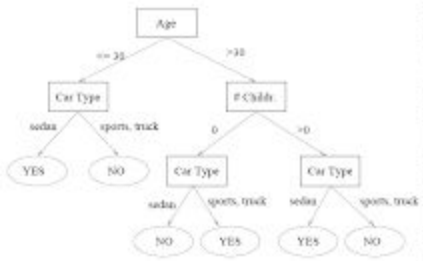
Siendo, $G(C/A_{ij})$ igual a:

$$G(A_{ij}) = -\sum_{j=i}^{M_i} p(C_k/A_{ij})(1 - p(C_k/A_{ij}))$$

- A_{ij} es el atributo empleado para ramificar el árbol,
- J es el número de clases,
- M_i es el de valores distintos que tiene el atributo A_i
- $p(A_{ij})$ constituye la probabilidad de que A_i tome su j-ésimo valor y
- $p(C_k/A_{ij})$ representa la probabilidad de que un ejemplo sea de la clase C_k cuando su atributo A_i toma su j-ésimo valor.

El índice de diversidad de Gini toma el valor cero cuando un grupo es completamente homogéneo y el mayor valor lo alcanza cuando todas las $p(A_{ij})$ son contantes, entonces el valor del índice es $\frac{(J-1)}{J}$.

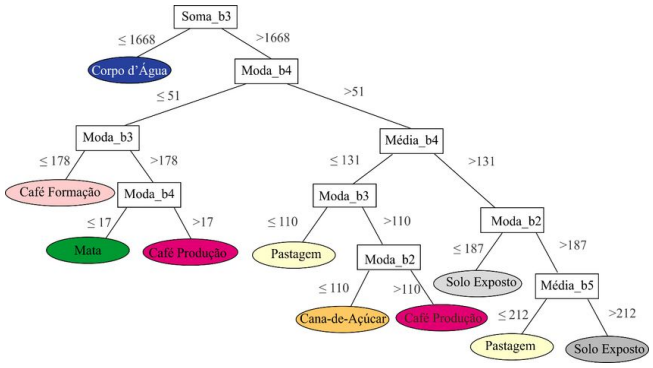
(figura 2).



Car type	Age	Chil	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

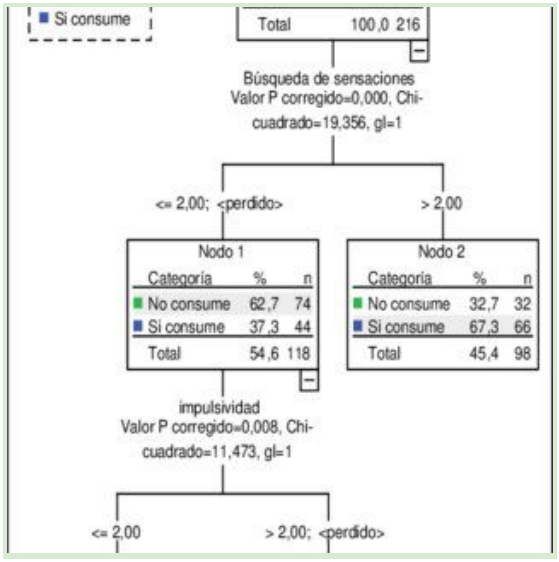
3.2.3 Algoritmo C4.5

Este algoritmo es una extensión del algoritmo ID3, que se utiliza para la clasificación estadística y genera un árbol de decisión a partir de un conjunto de datos de entrenamiento (como ID3); cada ejemplo es un vector, que representa los atributos o características del ejemplo. Este algoritmo selecciona atributos que dividen más eficazmente la muestra en subconjuntos enriquecidos y selecciona el atributo con la mayor ganancia de información como parámetro de decisión.



3.2.4 1.1 Chi-square Automatic Interaction Detector (CHAID)

El algoritmo CHAID se basa en AID (Automatic Interaction Detection). El algoritmo tiene dos limitaciones muy importantes; por un lado, se debe a la gran cantidad de elementos muestrales necesarios para realizar el análisis, y por otro lado, la falta de explicación o determinación de variables dependientes y explicativas. La agrupación de categorías se da probando todas las posibles combinaciones binarias de las variables, la prueba estadística F se utiliza para seleccionar la mayor diferencia posible. En este algoritmo, el proceso de segmentación de la muestra es en grupos dicotómicos.



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

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

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

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.

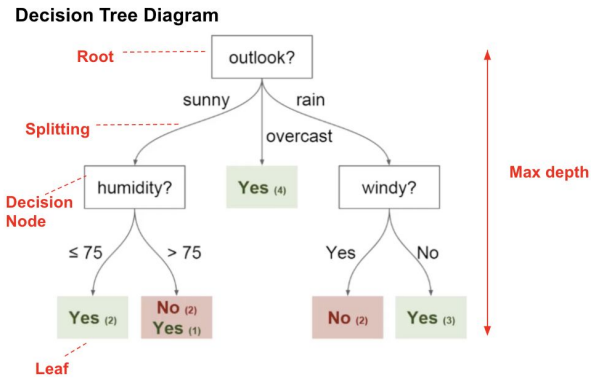


Figura 2: Entrenamiento de un árbol de decisión binario usando *(En este semestre, uno podría ser CART, ID3, C4.5... por favor, elija)*. En este ejemplo, mostramos un modelo para predecir si se debe jugar al golf o no, según el clima.

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

1. Cuji, B.; Gavilanes, W. y Sanchez, R. 2017. Modelo predictivo de deserción estudiantil basado en árboles de decisión. en Revista Espacios. Vol.38(N°55), 17-26. DOI: http://ww.revistaespacios.com/a17v38n55/a17v38n55p17.pdf

2. Caicedo Zambrano, J; Timarán Pereira, R e Hidalgo Troya, A. 2019. Árboles de decisión para predecir factores asociados al desempeño académico de estudiantes de bachillerato en las pruebas Saber 11°. en Revista de Investigación Desarrollo e innovación. Vol.9(N°2), 363-378. DOI: https://dialnet.unirioja.es/servlet/articulo?codigo=6976577

3. Menacho Chiok, C. H. 2017. Predicción del rendimiento académico aplicando técnicas de minería de datos. en Revistas La Molina. Vol.78(N°1), 26-33. DOI: http://revistas.lamolina.edu.pe/index.php/acu/article/view/811/pdf_43

4. Daza, A. 2016. Un modelo basado en árboles de decisión para predecir la deserción estudiantil en la educación superior privada. Scientia. Vol.8(N°1), 59-73. DOI: https://dialnet.unirioja.es/servlet/articulo?codigo=6181472