PREDICCIÓN DEL ÉXITO ACADÉMICO EN EDUCACIÓN SUPERIOR UTILIZANDO ÁRBOLES DE DECISIÓN

Isabella Pérez	Mariana Quintero	Miguel Correa	Mauricio Toro
Universidad Eafit	Universidad Eafit	Universidad Eafit	Universidad Eafit
Colombia	Colombia	Colombia	Colombia
iperezs2@eafit.edu.co	mquintero3@eafit.edu.co	macorream@eafit.edu.co	mtorobe@eafit.edu.co

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.

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.

1.3 Estructura del artículo

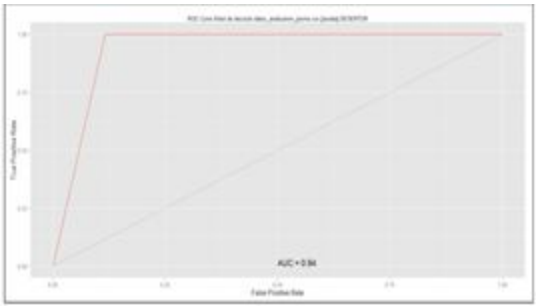
En lo que sigue, en la sección 2, presentamos el trabajo relacionado con el problema. Más adelante, en la sección 3, presentamos los conjuntos de datos y métodos utilizados en esta investigación. En la sección 4, presentamos el diseño del algoritmo. Después, en la sección 5, presentamos los resultados. 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]

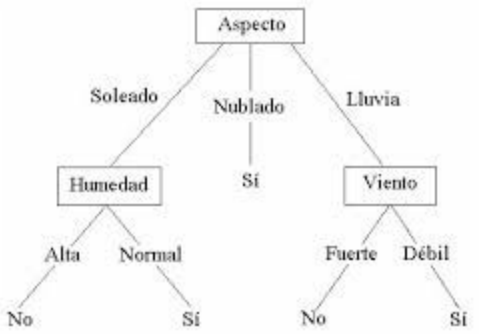
2.4 Predicción de la deserción estudiantil en la Educación Superior Privada.

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]

3.2 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.

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

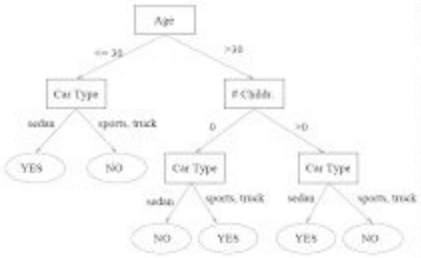
Siendo, $G(C/A_{ij})$ igual a:

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

- 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 $\frac{(J-1)}{J}$.

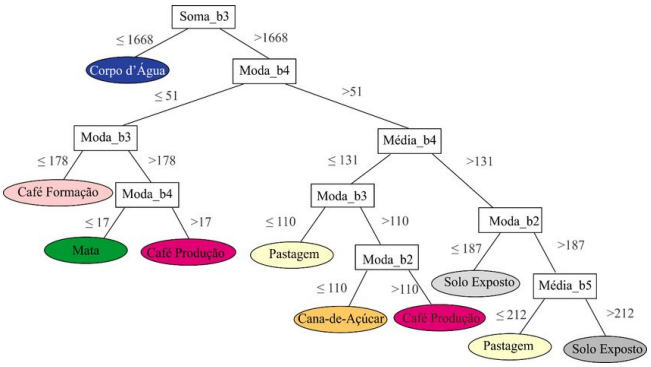
(figura 2).



Car type	Age	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

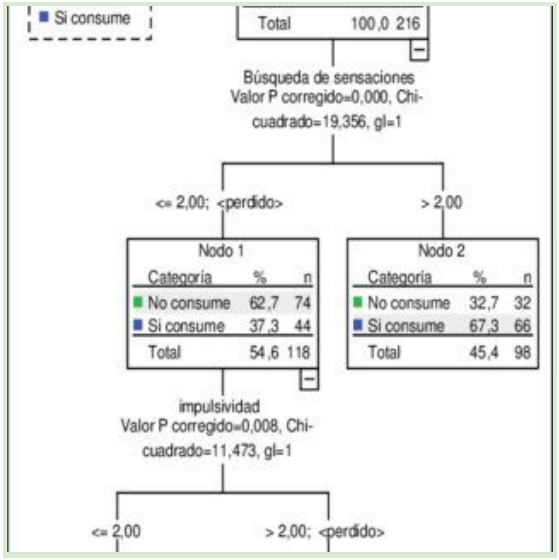
3.2.3 Algoritmo C4.5

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.



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