Modelo predictivo de deserción de estudiantes universitarios utilizando técnicas de minería de datos

Anexos

Kristein Johan Ordoñez López
Jhonathan Astudillo Astudillo

Director: MSc. Jimena Adriana Timaná Peña
Co-Director: PhD. Carlos Alberto Cobos Lozada

Universidad del Cauca

Facultad de Ingeniería Electrónica y Telecomunicaciones
Departamento de Sistemas
Grupo de I+D en Tecnologías de la Información
Línea de Investigación: Sistemas Inteligentes
Popayán, julio de 2021

Anexo A

DETALLE DE LA FASE DEL CICLO DE MUERTE DE LA METODOLOGÍA XP

A continuación, se detalla el menú definido para la creación del modelo Bagging, el cual utilizó la librería **scikit-learn** de Python.

La figura 1 muestra el menú principal para la creación del modelo híbrido.

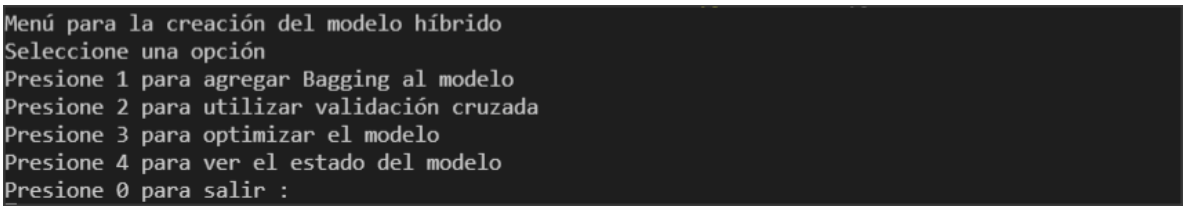

Figura 1. Menú para la creación del modelo híbrido.

El menú principal proporciona todas las diferentes opciones para la creación del modelo híbrido, su evaluación y la visualización de sus resultados.

En la figura 2 se puede observar el menú que se presenta al seleccionar la opción 1 del menú principal.

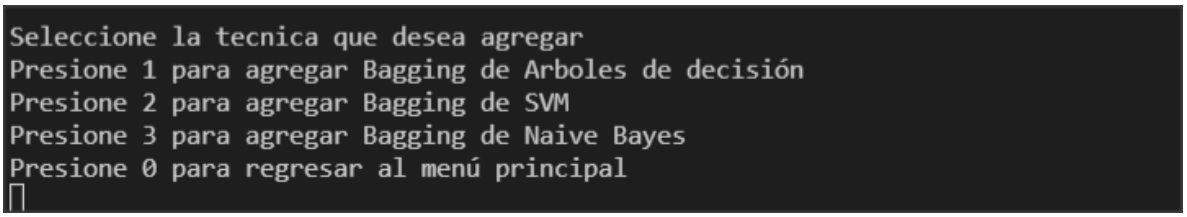

Figura 2. Menú para agregar un componente Bagging al modelo híbrido.

En este menú se puede elegir entre 3 diferentes técnicas: Naive Bayes, Arboles de decisión y Support Vector Machine. Para que posteriormente se cree el componente con ellas y también permite regresar al menú principal.

En la figura 3 se puede observar el menú que se presenta al seleccionar la opción 2 del menú principal.

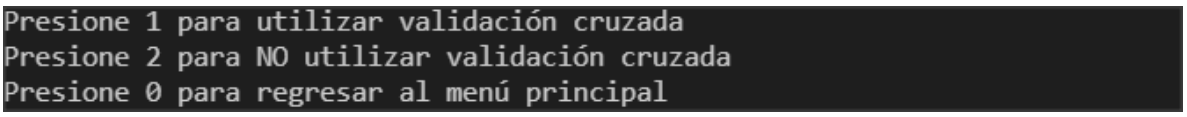

Figura 3. Menú para construir el modelo híbrido con validación cruzada.

En esta opción del menú se tiene la posibilidad de construir el modelo híbrido utilizando o no validación cruzada.

En la figura 4 se pueden observar las opciones que se despliegan al escoger la opción número 3 del menú principal.

```
Seleccione la tecnica con la que desea optimizar el modelo
Presione 1 para optimizar con Hill Climbing
Presione 2 para optimizar con Simulated Annealing
Presione 3 para optimizar con Step Grid
Presione 0 para regresar al menú principal
```

Figura 4. Menú de técnicas para la optimización del modelo híbrido.

En la opción número 3 del menú principal, es posible optimizar el modelo, para ello se puede elegir entre tres opciones diferentes de optimización. La primera utilizando Hill Climbing, la segunda con Simulated Anealing y la última con Step Grid. A continuación, en la figura 5 se presenta un ejemplo del resultado de elegir como optimizador el algoritmo Simulated Anealing.

```
Construyendo Bagging   DecisionTreeClassifier()
Decision Tree :  0.8575
Construyendo Bagging   GaussianNB()
Naive Bayes :  0.8545
Pesos iniciales:  [0.5, 0.5]
Precision inicial: 0.8497
Pesos finales: [0.6570779609314329, 0.3429220390685671]
Precision final: 0.8497
####################################
[[1204  198]
 [ 220 1312]]
Precision verdaderos positivos: 0.8455056179775281
Area bajo la curva 0.8575850240052442
Recuerdo 0.856396866840731
```

Figura 5. Menú de técnicas para la optimización del modelo híbrido.

En la figura 5 se observa el resultado de elegir como técnica de optimización a Simulated Annealing, en la imagen se puede observar (los pesos iniciales con los que fue construido el modelo, la precisión de cada uno de los modelos internos, la precisión inicial y final) que corresponde a la segunda opción del sub menú presentado en la figura 4.

En la figura 6 se puede observar el resultado de seleccionar la opción número 4 del menú principal.

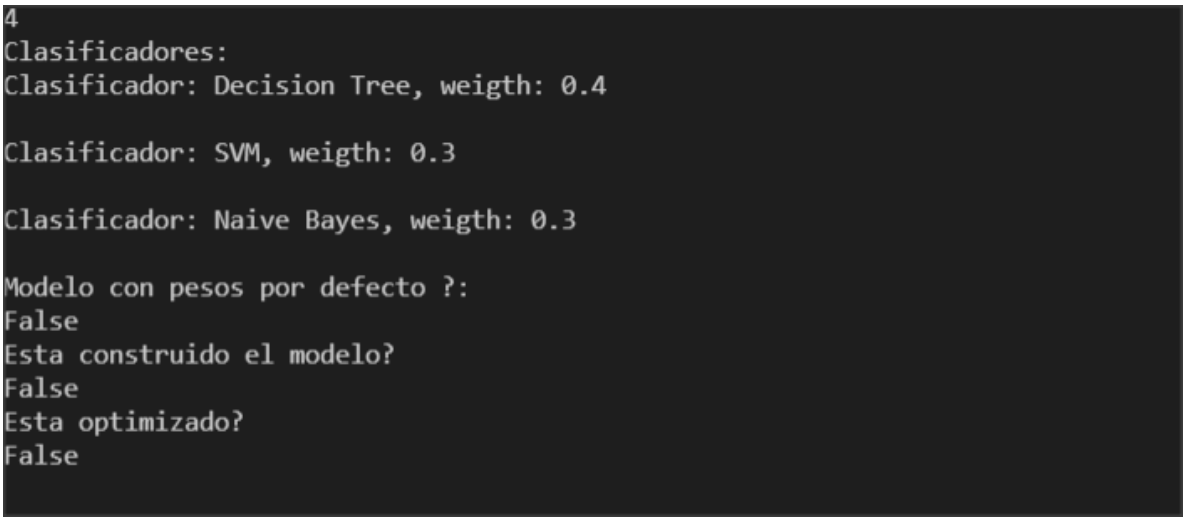

Figura 6. Menú de técnicas para la optimización del modelo híbrido.

La opción numero 4 permite conocer la configuración actual del modelo híbrido, se pueden saber las técnicas utilizadas, si tiene pesos iniciales, si ya fue optimizado o construido, entre otras.

Anexo B

ARTICULO SOBRE REVISION SITEMATICA

Revisión sistemática sobre Modelos predictivos de deserción de estudiantes universitarios utilizando técnicas de minería de datos

Systematic review on predictive models of college student dropout using data mining techniques

Kristein Johan Ordoñez Lopez[1]
Jhonatan Astudillo Astudillo[2]
Jimena Adriana Timaná Peña[3]
Carlos Alberto Cobos Lozada[4]

Resumen

Objetivo: Identificar las técnicas de minería de datos más utilizadas y que obtienen los mejores resultados de precisión para la predicción de deserción estudiantil universitaria, así como los atributos más utilizados para realizar esta predicción. **Metodología**: Se realizó una exhaustiva búsqueda sistemática de artículos científicos, artículos de conferencias, revistas indexadas y libros de ingeniería, en bases de datos de referencias bibliográficas, lo que permitió, una revisión exploratoria y analítica para la recopilación de información significativa, sobre las técnicas de minería de datos usadas para la predicción de este tipo de deserción estudiantil. **Resultados:** Esta revisión permitió establecer que las técnicas de minería de datos más empleadas en la creación de modelos para la predicción de deserción estudiantil universitaria son a saber: Arboles de Decisión, Naive Bayes y Support Vector Machine. **Conclusiones:** Después de haber realizado una exhaustiva revisión sistemática se observó que los conjuntos de datos con los que se construyen los modelos son pequeños y están desbalanceados, los datos desbalanceados

1 Grado académico, filiación institucional, Código ORCID, Correo institucional de correspondencia

2 Grado académico, filiación institucional, Código ORCID, Correo institucional de correspondencia

3 Magister en Computación, Universidad del Cauca, jtimana@unicauca.edu.co
4 Doctor en Ingeniería de Sistemas y Computación, Universidad del Cauca, ccobos@unicauca.edu.co

provocan que los modelos tengan mayor dificultad cuando tratan de clasificar un abandono positivo, pues los algoritmos de clasificación son propensos a ignorar la clase minoritaria (clase con menor número de registros), en este caso los estudiantes que si desertan. Para estos modelos la menor precisión reportada fue de 34 % y la mayor de 94%.
Palabras claves: Arboles de decisión, minería de datos, deserción estudiantil, Naive Bayes, SVM, KNN, revisión sistemática, precisión.

Abstract

Objective: Identify the most used data mining techniques and which obtain the best precision results for the prediction of university student dropout, as well as the most used attributes to make this prediction. **Methodology or method:** A systematic search was carried out for scientific articles, conference articles, indexed journals and engineering books, in databases of bibliographic references, which allowed an exploratory and analytical review for the collection of significant information on mining techniques. data used to predict this type of student dropout. **Results:** This review made it possible to establish that with the help of data mining and based on the sociodemographic and academic data of a student, it is possible to predict dropout with a high rate of precision.

Conclusions: The application of the search procedure to identify the data mining techniques most used in the prediction of university student dropout and the effectiveness of each of them, it was also evidenced that the quality of the prediction largely depends on the volume and composition of the data with which the model is created.
Keywords: Decision trees, data mining, student dropout, Naive Bayes, SVM, KNN, systematic review, precision.

Introducción

La deserción estudiantil no solo debe ser entendida como el abandono definitivo del aula de clase, si no como el abandono de la formación académica, independientemente de las condiciones y modalidades de presencialidad, es decisión personal del sujeto y no obedece a un retiro académico forzoso (por el no éxito del estudiante en el rendimiento académico, como es el caso de expulsión por

bajo rendimiento académico) o el retiro por asuntos diciplinares. Entonces se puede decir que la deserción es opción del estudiante, influenciado positiva o negativamente por circunstancias internas o externas [1]. Debido a la alta tasa de deserción estudiantil universitaria, se han realizado diferentes investigaciones en todo el mundo para desarrollar modelos basados en técnicas de minería de datos que puedan predecir cuándo un estudiante está en riego de desertar. Algunas de las técnicas de minería de datos más usadas son: los árboles de decisión, las redes bayesianas, el algoritmo de los k vecinos más cercanos (K-nn), entre otras, los modelos realizados con estas técnicas tienen en cuenta datos sociodemográficos y académicos de los estudiantes. En general, estos modelos alcanzan una precisión para la predicción de deserción estudiantil que esta entre el 34% y el 94%.

Metodología

En esta sección se presenta el proceso que se llevó a cabo para la realización de la revisión sistemática de la literatura científica.

Proceso de búsqueda de literatura científica

La revisión sistemática de la literatura científica está basada en los lineamientos propuestos por Kitchenham [2] y se realizó usando la base de datos de Scopus. En este proceso se define un protocolo que incluye la definición de: 1. Cadena de búsqueda, 2. Criterios de selección, 3. Criterios de selección de los documentos a incluir en el estudio y 4. Palabras clave.

Cadena de búsqueda

A continuación, se muestra la ecuación de consulta utilizada en la base de datos de Scopus:

("predictive model" OR "predicting model" OR "predict model" OR "prediction model" OR "predictive approach" OR "predicting approach" OR "predict approach" OR "prediction approach" OR "Predicting Academic Performance" OR "Prediction Academic Performance" OR "Predict Academic Performance" OR "Predictive Academic Performance") AND ("student dropout" OR "student retention" OR "student desertion" OR "students dropout" OR "students retention" OR "students desertion" OR "university student dropout" OR "university

student retention" OR "university student desertion" OR "university students dropout" OR "university students retention" OR "university students desertion") AND ("data mining" OR "data mining techniques" OR "educational data mining")

Criterios de selección

La Tabla 1, muestra los criterios de selección (inclusión, exclusión y calidad) identificados y que permitirán escoger aquellos estudios más acordes con el tema de investigación y excluir aquellos que no son relevantes para responder la pregunta de investigación.

Criterios de inclusión	
CI-1	artículos en idioma español e inglés
CI-2	artículos que traten de deserción estudiantil en el ámbito universitario
CI-3	full papers, conference papers, book chapter
Criterios de exclusión	
CE-1	artículos inferiores al año 2010
CE-2	artículos que solo presenten el abstract, revisiones, opiniones, técnicas en su contenido
CE-3	artículos que traten de mooc's o cursos en línea
CE-4	investigaciones similares o duplicadas del mismo autor
Criterios de Calidad	
CQ1	¿El estudio describe adecuadamente el proceso de creación del modelo de minería de datos? Respuestas posibles: (si = 1 / no = 0)
CQ2	¿El estudio utiliza un método de validación (caso de estudio, experimentos) donde se evalúe el modelo de minería de datos? Respuestas posibles: (si=1 / no=0)
CQ3	¿El estudio presenta los resultados detallados obtenidos en una fase de validación del modelo? Respuestas posibles (muy completo=1 medianamente=0.5 y no=0)

CQ4	¿El estudio ha sido publicado en una fuente reconocida? La clasificación del artículo se basa en el cuartil en el que fue publicado y/o en el publindex de la revista en el que fue publicado según Colciencias. Para las conferencias la clasificación se basó en el ranking de Investigación y educación en computación (CORE). (Cuartil 1 o revista/conferencia tipo A =2; cuartil 2 o revista/conferencia tipo B=1.5; cuartil 3 o revista/conferencia tipo C = 1; cuartil 4 =0.5; Sin indexación= 0)

Tabla 1. Criterios de selección

De los criterios de calidad de la Tabla 1, el puntaje total para evaluar una propuesta viene dado por la Fórmula 1.

$$P = CQ1 + CQ2 + CQ3 + CQ4$$

Fórmula 1. Suma de criterios de calidad

Donde *P* es el puntaje total y *CQn* es el puntaje obtenido en cada criterio.

Resultados

La selección de los artículos resultantes de la revisión sistemática se llevó a cabo siguiendo 4 fases a saber: artículos encontrados al probar la cadena de búsqueda, filtrado de los artículos potenciales, aplicación de los criterios de inclusión y exclusión y selección de artículos finales de acuerdo a los criterios de calidad establecidos. En la primera fase, se probó la cadena de búsqueda sobre el indexador de artículos Scopus, dando como resultado una lista de 514 artículos. En la segunda fase, se seleccionó los artículos potenciales, resultado de filtrar aquellas propuestas que contenían al menos dos palabras claves en su título y tres en el resumen (*abstract*), con un resultado de 33 artículos. En la tercera fase, se aplicó los criterios de inclusión y exclusión definidos en la Tabla 1 sobre los artículos potenciales. Finalmente, en la cuarta fase, se aplicó los criterios de calidad a los artículos resultantes de la fase anterior. Dichas propuestas fueron puntuadas a

través de la Fórmula 1. Los artículos finales seleccionados son aquellos que superen el umbral definido en el valor 3.0.

Aplicación cadena de búsqueda

Al aplicar la cadena de búsqueda definida en la sección “Cadena de búsqueda” sobre la base de datos bibliográfica Scopus se obtuvo un total de 514 artículos.

Selección de artículos potenciales

Para la selección de los artículos potenciales se estableció que estos debían tener al menos dos palabras claves en el título y mínimo tres en el resumen. El número total de propuestas que cumplieron estas condiciones se presentan en la Tabla 2.

Palabras clave	Cantidad de artículos encontrados	Total de artículos potenciales
"predictive model" , "predicting model" , "predict model", "prediction model" , "predictive approach", "predicting approach", "predict approach" , "prediction approach", “Predicting Academic Performance”, “Prediction Academic Performance”, “Predict Academic Performance”, “Predictive Academic Performance”, "student dropout", "student retention", "student desertion", "students dropout", "students retention", "students desertion", "university student dropout", "university student retention", "university student desertion", "university students dropout", "university students retention", "university students desertion", "data mining", "data mining techniques", "educational data mining".	514	33

Tabla 2. Resultados artículos potenciales

En la Tabla 3, se presentan los artículos potenciales seleccionados.

Id. artículo	artículo
1	Trends in information models on retention-university dropout [3]
2	University student retention: Best time and data to identify undergraduate students at risk of dropout [4]
3	Predicting Student Retention Among a Homogeneous Population Using Data Mining [5]
4	Review of techniques, tools, algorithms and attributes for data mining used in student desertion [6]
5	Identifying Students at Risk of Academic Failure Within the Educational Data Mining Framework [7]
6	University dropout prediction through educational data mining techniques: A systematic review [8]
7	Predicting academic performance of tertiary students using classification algorithm [9]
8	A machine learning approach to Predict the Engineering Students at risk of dropout and factors behind: Bangladesh Perspective [10]
9	Research and application of grade prediction model based on decision tree algorithm [11]
10	Students' Success Predictive Models Based on Selected Input Parameters Set [12]
11	From Lab to Production: Lessons Learnt and Real-Life Challenges of an Early Student-Dropout Prevention System [13]

12	An Early Feedback Prediction System for Learners At-Risk within a First-Year Higher Education Course [14]
13	Predictive Analysis for Student Retention by Using Neuro-Fuzzy Algorithm [15]
14	Predictive analytic models of student success in higher education: A review of methodology [16]
15	An analysis of student representation, representative features and classification algorithms to predict degree dropout [17]
16	A Hybrid Prediction Model Integrating a Modified Genetic Algorithm to K-means Segmentation and C4.5 [18]
17	The Analysis of Student Performance Using Data Mining [19]
18	Utilizing feature selection in identifying predicting factors of student retention [20]
19	Dropout situation of business computer students, University of Phayao [21]
20	Students' performance prediction model using meta-classifier approach [22]
21	Detection of desertion patterns in university students using data mining techniques: A case study [23]
22	Application of data mining for the detection of variables that cause university desertion [24]
23	An artificial neural network based early prediction of failure-prone students in blended learning course [25]

24	An approach to educational data mining model accuracy improvement using histogram discretization and combining classifiers into an ensemble [26]
25	A survey of machine learning approaches and techniques for student dropout prediction [27]
26	Self-organising maps and student retention: Understanding multi-faceted drivers [28]
27	Finding the best algorithms and effective factors in classification of Turkish science student success [29]
28	University dropout: A prediction model for an engineering program in Bogotá, Colombia [30]
29	Analytical approach for predicting dropouts in higher education [31]
30	The efficacy of learning analytics interventions in higher education: A systematic review [32]
31	Dropout early warning systems for high school students using machine learning [33]
32	Predictive data modeling: Educational data classification and comparative analysis of classifiers using python [34]
33	A predictive model for student outcomes using sparse coding – Hybrid features selection [35]

Tabla 3. Artículos potenciales seleccionados

Aplicación de criterios de inclusión y exclusión

La Tabla 4 presenta la aplicación de los criterios de inclusión y exclusión definidos en la sección “Criterios de selección”, sobre los artículos potenciales.

ID. Articulo	CI-1	CI-2	CI-3	CE-1	CE-2	CE-3	CE-4	Cumplimiento

1	✓	✓	✓					SI
2	✓	✓	✓					SI
3	✓	✓						NO
4	✓	✓	✓					SI
5	✓	✓	✓					SI
6	✓	✓	✓					SI
7	✓		✓					NO
8	✓	✓			X			NO
9	✓	✓	✓					SI
10	✓	✓	✓				X	NO
11	✓	✓	✓					SI
12	✓	✓	✓					SI
13	✓	✓						NO
14	✓							NO
15	✓	✓	✓					SI
16	✓		✓					NO
17	✓		✓					NO
18	✓		✓					NO
19	✓	✓	✓					SI
20	✓		✓					NO
21	✓		✓					NO
22	✓		✓					NO
23	✓	✓	✓					SI
24	✓	✓	✓					SI
25	✓	✓	✓					SI

26	✓							NO
27	✓		✓					NO
30	✓	✓	✓					SI
29	✓	✓	✓					SI
30	✓		✓					NO
31	✓							NO
32	✓	✓	✓					SI
33	✓		✓					NO

Tabla 4. Criterios de inclusión y exclusión

Las propuestas resultantes después de haber aplicado los criterios de inclusión y exclusión son consideradas como estudios primarios y se encuentran resumidos en la Tabla 5.

Id. Articulo	Autor	Tipo de artículo	Año
1	Guerra, Laura; Rivero, Dulce; Díaz, Eleazar; Arciniegas, Stalin.	Full paper	2020
2	José Maria Ortiz; Antonio Rua; Paloma Bilbao, Martí Casadesús.	Full paper	2020
4	K. Y Diaz Pedroza; B. Y Chindoy Chasoy; A. Rosado Gómez.	Conference paper	2019
5	Annalina Sarra; Lara Fontanella; Simone Di Zio.	Full paper	2019
6	Franceso Agrusti; Gianmarco Bonavolontá; Mauro Mezzini.	Full paper	2019
9	Yaling Zhang; Bei Wu.	Conference paper	2019

11	Alvaro Ortigosa; Rosa M. Carro; Javier Bravo.	Full paper	2019
12	David Baneres; M. Elena Rodriguez; Montse Serra.	Full paper	2019
15	Ruben Manrrique; Bernardo Pereira; Olga Marino.	Conference paper	2019
19	Pratya Nuankaew	Full paper	2019
23	Otgontsetseg Sukhbaatar; Lodoiravsal Choimaa; Tsuyoshi Usagawa	Full paper	2019
24	Dejan Rancić; Olivera Pronić-Rancić; Danijela Milošević.	Conference paper	2019
25	Neema Mduma; Khamisi Kalegele; Dina Machuve.	Full paper	2019
28	Andres Acero; Juan Camilo Achury; Juan Morales Piñero.	Conference paper	2019
29	Garima Jaiswal; Arun Sharma; Sumit Kumar Yadav.	Full paper	2019
32	Pratiyush Guleria; Manu Sood.	Conference paper	2018

Tabla 5. Propuestas primarias

Aplicación de criterios de calidad

Los criterios de calidad definidos en la Tabla 1, son aplicados sobre los estudios primarios presentados en la Tabla 5. Sobre cada uno de ellos se aplica la Ecuación 1 para obtener la puntuación final P de cada propuesta, dicha puntuación se presenta en la Tabla 6.

Id. Articulo	Revista de publicación	Puntuación				
		CQ1 si =1 no= 0	**CQ2** si =1 no = 0	**CQ3** muy completo =1 medianamente =0.5 no =0	**CQ4** cuartil 1 o tipo A =2 cuartil 2 o tipo B=1.5 cuartil 3 o tipo C = 1 cuartil 4 = 0.5 Sin indexación= 0	**P**
1	Associacao Iberica de Sistemas e Tecnologias de Informacao	0	1	0,0	1	**2,0**
2	Taylor & Francis Online	1	1	0,5	0	**2,5**
4	IOPScience	0	1	0,5	0	**1,5**
5	Springer Link	1	1	1,0	0	**3,0**
6	Je-LKS	1	1	1,0	0	**3,0**
9	ACM Digital Library	1	1	0,5	0	**2,5**
11	IEEE Xplore	1	1	1,0	0	**3,0**
12	IEEE Xplore	1	1	1,0	0	**3,0**
15	ACM Digital Library	1	1	1	0	**3,0**
19	iJET	1	1	1	0	**3,0**

23	iJET	0	1	0,0	1	**2,0**
24	Springer Link	1	1	0,5	0	**2,5**
25	Data Sicence Journal	0	1	0,5	0	**1,5**
28	ResearchGate	1	1	1,0	0	**3,0**
29	IGI Global	1	1	1,0	0	**3,0**
32	IEEE Explore	1	1	0,5	0	**2,5**

Tabla 6. Evaluación criterios de calidad

Los estudios primarios finalmente seleccionados fueron aquellos que igualaron o superaron un valor mínimo de calidad de 3.0. Las propuestas seleccionadas como estudios primarios de calidad se encuentran resaltadas en la Tabla 6.

Solución a las preguntas de investigación

A continuación, se contestan cada una de las preguntas de investigación específicas definidas en la sección “Preguntas de investigación”.

PI1. ¿Cuáles de estos modelos están soportados bajo las técnicas de minería de datos?

Id. Articulo	Articulo
1	Trends in information models on retention-university dropout
2	University student retention: Best time and data to identify undergraduate students at risk of dropout
4	Review of techniques, tools, algorithms and attributes for data mining used in student desertion
5	Identifying Students at Risk of Academic Failure Within the Educational Data Mining Framework
6	University dropout prediction through educational data mining techniques: A systematic review

9	Research and application of grade prediction model based on decision tree algorithm
11	From Lab to Production: Lessons Learnt and Real-Life Challenges of an Early Student-Dropout Prevention System
12	An Early Feedback Prediction System for Learners At-Risk within a First-Year Higher Education Course
15	An analysis of student representation, representative features and classification algorithms to predict degree dropout
24	An approach to educational data mining model accuracy improvement using histogram discretization and combining classifiers into an ensemble(Conference Paper)

Tabla 7. Modelos predictivos soportados bajo técnicas de minería de datos

PI2. ¿Cuáles son las técnicas de minería de datos usadas en estos modelos predictivos?

Técnica de minería de datos	**Id. artículo**
RANDOM FOREST	4,15,24
ARBOL DECISION	2,4,6,9,11,12,24
NAIVE BAYES	4,5,6,12,15,24
R. LOGISTICA	6
SVM	6,12,15
KNN	6,12
RED NEURONAL	6
No especifica la técnica	1
GRADIEND BOOT TREE	15

Tabla 8. Técnicas de minería de datos usadas por modelos predictivos

PI3. ¿Cuáles son los atributos más comunes usados en estos modelos predictivos?

Tipo de atributo	**atributo**	**Id. artículo**
No describe	No describe	9

Sociodemográfico	General	1,2,5,12
	Identificación	11
	Nacionalidad	6
	Ciudad procedencia	4
	Genero	4,6
	Estado civil	4,6,11
	Edad	4,6,11
	Etnia	4
	Discapacidad física	6
	Ocupación padres	4,6,11
	Tamaño familia	4
	Disciplina	6
Académico	General	1,2,5,12
	Puntaje admisión	11
	Valor matricula	6
	Edad inscripción	2
	Programa académico	4,
	Promedio académico	4,6,11,15,24
	Semestre	4,6,11

	Numero materias	6
	Materias repetidas	4
	Promedio escuela	4
	Tipo admisión	15
	nivel estudios padres	6,11
	nivel inglés	2
	nivel matemáticas	9
	distancia residencia	11
Financiero	general	1
	ayuda familiar	4
	tipo vivienda	6
	trabaja	4
	estrato	4,11
	ingresos familiares	11
	deserción	2

Tabla 9. Atributos identificados en los modelos predictivos

PI4. ¿Cuáles son las técnicas de minería de datos con mejor precisión en estos modelos predictivos?

Técnica	**1**	**2**	**4**	**5**	**6**	**9**	**11**	**12**	**15**

Id. Articulo									
Random Forest									86%
Árbol Decisión		76%		94%	67%	81%	89%		
Naive Bayes					49%				57%
R. Logística					34%				
SVM									67%
KNN									
MLP									
Red Neuronal					40%				
Gradiend Boot Tree									84%
No reporta	X		X					X	

Tabla 10. Precisión modelos predictivos según técnica de minería de datos utilizada

Conclusiones

Después de haber realizado una exhaustiva revisión sistemática se observó que los conjuntos de datos con los que se construyen los modelos son pequeños y están desbalanceados, los datos desbalanceados provocan que los modelos tengan mayor dificultad cuando tratan de clasificar un abandono positivo, pues los algoritmos de clasificación son propensos a ignorar la clase minoritaria (clase con menor número de registros), en este caso los estudiantes que si desertan. Para estos modelos la menor precisión reportada fue de 34 % y la mayor de 94%.

Se identificaron las técnicas de minería de datos más utilizadas en la predicción de la deserción estudiantil universitaria y la efectividad de cada una de ellas, además se evidencio que la calidad de la predicción depende ampliamente del volumen y la composición de los datos con los que se crea el modelo.

Teniendo en cuenta que la deserción estudiantil no es un problema únicamente de bajo rendimiento académico, fue posible identificar otras características con las que se construyeron modelos de predicción con una tasa de precisión superior al 85%.

Referencias

[1] G. J. Paramo y C. A. Correa, “Deserción Estudiantil Universitaria. Conceptualizacion,”*Revista Universidad EAFIT,* vol. 35, nº 114, pp. 65-78, 2012.

[2] B. Kitchenham, “Systematic literature reviews in software engineering – A tertiary study,”de *Information and Software Technology*, 2010, pp. 792-805.

[3] L. Guerra, D. Rivero, E. Díaz y S. Arsiniegas, “Trends in information models on retention-university dropout,”*RISTI - Revista Iberica de Sistemas e Tecnologias de Informacao,* vol. 2020, pp. 55-68, 2020.

[4] J. M. Ortiz-Lozano, A. Rua-Vieites, P. Bilbao-Calabuig y M. Casadesús-Fa, “University student retention: Best time and data to identify undergraduate students at risk of dropout,”*Innovations in Education and Teaching International,* vol. 57, pp. 74-85, 2020, DOI: https://doi.org/10.1080/14703297.2018.1502090.

[5] G. Bilquise, S. Abdallah y T. Kobbaey, “Predicting Student Retention Among a Homogeneous Population Using Data Mining,”vol. 1058, pp. 35-46, 2020, DOI: https://doi.org/10.1007/978-3-030-31129-2_4 .

[6] K. Y. Diaz Pedroza, B. Y. Chindoy Chasoy y A. A. Rosado Gómez, “Review of techniques, tools, algorithms and attributes for data mining used in student desertion,”vol. 1409, 2019, DOI: https://doi.org/10.1088/1742-6596/1409/1/012003 .

[7] A. Sarra, L. Fontanella y S. Di Zio, “Identifying Students at Risk of Academic Failure Within the Educational Data Mining Framework,”vol. 146, pp. 41-60, 2019, DOI: https://doi.org/10.1007/s11205-018-1901-8.

[8] F. Agrusti, G. Bonavolontà y M. Mezzini, “University dropout prediction through educational data mining techniques: A systematic review,”*Journal of E-Learning and Knowledge Society,* vol. 15, pp. 161-182, 2019, DOI: https://doi.org/10.20368/1971-8829/1135017.

[9] S. Jayaprakash y V. Jaiganesh, “Predicting academic performance of tertiary students using classification algorithm,”*International Journal of Recent Technology and Engineering,* vol. 8, pp. 6558-6561, 2019, DOI: https://doi.org/10.35940/ijrte.B2716.078219.

[10] S. A. Ahmed y S. I. Khan, “A machine learning approach to Predict the Engineering Students at risk of dropout and factors behind: Bangladesh Perspective,”*2019 10th International Conference on Computing, Communication and Networking Technologies,* pp. 1-6, 2019, DOI: https://doi.org/10.1109/ICCCNT45670.2019.8944511.

[11] Y. Zhang y B. Wu, “Research and application of grade prediction model based on decision tree algorithm,”*ACM TURC 2019: ACM Turing Celebration Conference,* pp. 1-6, 2019, DOI: https://doi.org/10.1145/3321408.3322857.

[12] S. Milinković y V. Vujović, “Students’ Success Predictive Models Based on Selected Input Parameters Set,”*International Symposium INFOTEH-JAHORINA,* 2019, DOI: https://doi.org/10.1109/INFOTEH.2019.8717654.

[13] A. Ortigosa, R. M. Carro, J. Bravo-Agapito, D. Lizcano, J. J. Alcolea y O. Blanco, “From Lab to Production: Lessons Learnt and Real-Life Challenges of an Early Student-Dropout Prevention System,”*IEEE Transactions on Learning Technologies,* vol. 12, pp. 264-277, 2019, DOI: https://doi.org/10.1109/TLT.2019.2911608.

[14] D. Baneres, M. E. Rodríguez-Gonzalez y M. Serra, “An Early Feedback Prediction System for Learners At-Risk Within a First-Year Higher Education Course,”vol. 12, pp. 249-263, 2019, DOI: https://doi.org/10.1109/TLT.2019.2912167.

[15] M. Adil, F. Tahir y S. Maqsood, “Predictive Analysis for Student Retention by Using Neuro-Fuzzy Algorithm,”*Computer Science and Electronic Engineering*

(CEEC), vol. 10, pp. 41-45, 2018, DOI: https://doi.org/10.1109/CEEC.2018.8674216.

[16] Y. Cui, F. Chen, A. Shiri y Y. Fan, “Predictive analytic models of student success in higher education: A review of methodology,”*Information and Learning Sciences,* vol. 120, pp. 208-227, 2019, DOI: https://doi.org/10.1108/ILS-10-2018-0104.

[17] R. Manrique, B. P. Nunes, O. Marino, M. A. Casanova y T. Nurmikko-Fuller, “An Analysis of Student Representation, Representative Features and Classification Algorithms to Predict Degree Dropout,”*ACM’s International Conference Proceedings Series,* pp. 401-4010, 2019, DOI: https://doi.org/10.1145/3303772.3303800.

[18] M. Y. Orong, A. M. Sison y R. P. Medina, “A Hybrid Prediction Model Integrating a Modified Genetic Algorithm to K-means Segmentation and C4.5,”*TENCON 2018 - 2018 IEEE Region 10 Conference,* pp. 1853-1858, 2019, DOI: https://doi.org/10.1109/TENCON.2018.8650064.

[19] L. W. Santoso y Yulia, “The Analysis of Student Performance Using Data Mining,”*Advances in Computer Communication and Computational Sciences,* vol. 924, pp. 559-573, 2019, DOI: https://doi.org/10.1007/978-981-13-6861-5_48.

[20] J. D. Febro, “Utilizing feature selection in identifying predicting factors of student retention,”*International Journal of Advanced Computer Science and Applications,* vol. 10, pp. 269-274, 2019, DOI: https://doi.org/10.14569/ijacsa.2019.0100934.

[21] P. Nuankaew, “Dropout situation of business computer students, University of Phayao,”*International Journal of Emerging Technologies in Learning,* vol. 14, pp. 117-131, 2019, DOI: https://doi.org/10.3991/ijet.v14i19.11177.

[22] H. Hassan, S. Anuar y N. B. Ahmad, “Students’ performance prediction model using meta-classifier approach,”*Communications in Computer and Information Science,* vol. 1000, pp. 221-231, 2019, DOI: https://doi.org/10.1007/978-3-030-20257-6_19.

[23] D. Vila, S. Cisneros, P. Granda, C. Ortega, M. Posso-Yépez y I. García-Santillán, “Detection of desertion patterns in university students using data

mining techniques: A case study,"*Communications in Computer and Information Science,* vol. 895, pp. 420-429, 2019, DOI: https://doi.org/10.1007/978-3-030-05532-5_31.

[24] X. Palacios-Pacheco, W.Villegas-Ch y S. Luján-Mora, "Application of Data Mining for the Detection of Variables that Cause University Desertion,"*Communications in Computer and Information Science,* vol. 895, pp. 510-520, 2019, DOI: https://doi.org/10.1007/978-3-030-05532-5_38.

[25] O. Sukhbaatar, T. Usagawa y L. Choimaa, "An artificial neural network based early prediction of failure-prone students in blended learning course,"*International Journal of Emerging Technologies in Learning,* vol. 14, pp. 77-92, 2019, DOI: https://doi.org/10.3991/ijet.v14i19.10366qwer.

[26] G. Dimić, D. Rančić, O. Pronić-Rančić y D. Milošević, "An approach to educational data mining model accuracy improvement using histogram discretization and combining classifiers into an ensemble,"*Smart Innovation, Systems and Technologies,* vol. 144, pp. 267-280, 2019, DOI: https://doi.org/10.1007/978-981-13-8260-4_25.

[27] N. Mduma, K. Kalegele y D. Machuve, "A survey of machine learning approaches and techniques for student dropout prediction,"*Data Science Journal,* vol. 18, 2019, DOI: https://doi.org/10.5334/dsj-2019-014.

[28] D. C. Gibson, M. Ambrose y M. Gardner, "Self-organising maps and student retention: Understanding multi-faceted drivers,"*ASCILITE 2015 - Australasian Society for Computers in Learning and Tertiary Education, Conference Proceedings,* pp. 112-120, 2019.

[29] E. Filiz y E. Öz, "Finding the best algorithms and effective factors in classification of Turkish science student success,"*Journal of Baltic Science Education,* pp. 239-253, 2019, DOI: https://doi.org/10.33225/jbse/19.18.239.

[30] A. Acero, J. C. Achury y J. C. Morales, "University dropout: A prediction model for an engineering program in Bogotá, Colombia,"2019.

[31] J. Garima, S. Arun y Y. Sumit Kumar, "Analytical Approach for Predicting Dropouts in Higher Education,"*International Journal of Information and Communication Technology Education,* vol. 15, pp. 89-102, 2019, DOI: https://doi.org/10.4018/IJICTE.2019070107.

[32] A. L. Sønderlund, E. Hughes y J. Smith, “The efficacy of learning analytics interventions in higher education: A systematic review,”*British Journal of Educational Technology,* vol. 50, pp. 2594-2618, 2019, DOI: https://doi.org/10.1111/bjet.12720.

[33] Y. J. Chung y S. Lee, “Dropout early warning systems for high school students using machine learning,”*Children and Youth Services Review,* vol. 96, pp. 346-353, 2019, DOI: https://doi.org/10.1016/j.childyouth.2018.11.030.

[34] P. Guleria y M. Sood, “Predictive Data Modeling: Educational Data Classification and Comparative Analysis of Classifiers Using Python,”*2018 Fifth International Conference on Parallel, Distributed and Grid Computing,* pp. 740-746, 2018, DOI: https://doi.org/10.1109/PDGC.2018.8745727.

[35] M. Zaffar, M. Hashmani, K. S. Savita y A. Qayyum, “A predictive model for student outcomes using sparse coding – Hybrid features selection,”*Journal of Theoretical and Applied Information Technology,* vol. 96, pp. 7124-7138, 218.

Anexo C

ARTICULO SOBRE CREACION DE MODELO HIBRIDO

Modelo híbrido predictivo de deserción de estudiantes universitarios utilizando técnicas de minería de datos

Predictive model of college student dropout using data mining techniques

Kristein Johan Ordoñez Lopez[5]
Jhonatan Astudillo Astudillo[6]
Jimena Adriana Timaná Peña[7]
Carlos Alberto Cobos Lozada[8]

Resumen

Objetivo: Proponer un modelo híbrido predictivo de deserción de estudiantes universitarios que combine las mejores técnicas de minería de datos reportadas en el estado del arte y que use grandes volúmenes de datos balanceados buscando mejorar los reportes de calidad de predicción de deserción de estudiantes reportados a la fecha. **Metodología:** Siguiendo la metodología CRISP-DM, se construyó un modelo híbrido para la predicción de la deserción de estudiantes universitarios a partir de un conjunto de datos balanceados. Además, se creó el pseudocódigo del mismo modelo y se evaluó la calidad de sus resultados frente a otros modelos. **Resultados:** Se logró crear un modelo híbrido que mejoró la calidad en la predicción de la deserción de estudiantes universitarios combinando tres técnicas de minería de datos junto con la función de optimización de Recocido Simulado. **Conclusiones:** A partir de un conjunto de datos balanceado, el modelo híbrido propuesto presentó una mejora en la calidad de la predicción de la deserción de estudiantes universitarios en comparación a otros modelos, basado en la aplicación de las métricas: precisión, precisión de verdaderos positivos, sensibilidad y el área bajo la curva.

Palabras claves: Árboles de decisión, Naive Bayes, Support Vector Machine, modelo híbrido, deserción de estudiantes, datos balanceados, minería de datos, Bagging.

5 Ingeniero de Sistemas, Universidad del Cauca, joan@unicauca.edu.co
6 Ingeniero de Sistemas, Universidad del Cauca, jonas@unicauca.edu.co
7 Magister en Computación, Universidad del Cauca, jtimana@unicauca.edu.co
8 Doctor en Ingeniería de Sistemas y Computación, Universidad del Cauca, ccobos@unicauca.edu.co

Abstract

Objective: Proponer un modelo predictivo de deserción de estudiantes universitarios que hibride las mejores técnicas de minería de datos reportadas en el estado del arte y que use grandes volúmenes de datos balanceados buscando mejorar los reportes de calidad de predicción de deserción de estudiantes reportados a la fecha. **Methodology:** Siguiendo la metodología CRISP-DM, se construyó un modelo híbrido para la predicción de la deserción de estudiantes universitarios a partir de un conjunto de datos balanceados. Además, se creó el pseudocódigo del mismo modelo y se evaluó la calidad de sus resultados frente a otros modelos. **Results:** Se logró crear un modelo híbrido que mejoró la calidad en la predicción de la deserción de estudiantes universitarios combinando tres técnicas de minería de datos junto con la función de optimización de Recocido Simulado (Simulated Annnealing). **Conclusions:** A partir de un conjunto de datos balanceado, el modelo híbrido propuesto presentó una mejora en la calidad de la predicción de la deserción de estudiantes universitarios en comparación a otros modelos, basado en la aplicación de las métricas: precisión, precisión de verdaderos positivos, sensibilidad y el área bajo la curva.

Keywords: Decision Tree, Naive Bayes, Support Vector Machine, Hybrid model, student dropout, balanced data, data mining, Bagging.

Introducción

Uno de los principales desafíos que enfrenta el Sistema de Educación Superior Colombiano es disminuir los altos niveles de deserción académica que se presenta en el pregrado. Pese a que en los últimos años se han implementado estrategias que incluyen entre otros: 1) la mejora de la calidad y el volumen de la información que se entrega a los aspirantes sobre los programas ofrecidos, 2) la creación de programas de ayuda financiera para los estudiantes de bajos recursos o que provienen de otras ciudades, y 3) el acompañamiento psicológico de nivelación y orientación al estudiante, todavía el número de alumnos que no logra culminar sus estudios superiores es alto [1]. Según estadísticas del Ministerio de Educación Nacional, de cada cien estudiantes que ingresan a una institución de Educación Superior, cerca de la mitad no logra culminar su ciclo académico y obtener la graduación [2]. Gran parte de éstos, abandona sus estudios, principalmente en los primeros semestres.

Según El Ministerio de Educación Nacional la deserción universitaria en Colombia para año 2018 alcanzó una tasa del 8.79% [3]. Los expertos indican que la deserción en las instituciones universitarias se centra en los primeros años, concretamente en el primer o segundo semestre del programa de educación superior, bien sea porque los estudiantes no encuentran lo que esperaban al ingresar a determinado programa o porque la situación socioeconómica no les permite continuar.

Las consecuencias de la deserción incluyen entre varios aspectos: 1) pérdidas financieras, tanto para el estudiante como para la institución y 2) menores tasas de graduación e indicadores no favorables para temas relacionados a la acreditación de alta calidad de un programa. Si una institución pierde a un estudiante por cualquier motivo, la institución tiene una tasa de abandono más alta. La identificación temprana de los estudiantes que están en riesgo de abandono, es fundamental para el éxito de cualquier estrategia que busque combatir la deserción [2].

Debido a la alta deserción de estudiantes universitarios, se han realizado diferentes investigaciones en todo el mundo que han llevado al desarrollo de modelos predictivos que tienen en cuenta datos sociodemográficos de estudiantes y que utilizan técnicas de minería de datos como los árboles de decisión, las redes neuronales, el algoritmo de los k vecinos más cercanos (K-nn), el algoritmo Random Forest, entre otros. Sin embargo, algunos de estos modelos implementados trabajan con un número limitado de datos, por lo que sus resultados no son totalmente confiables. Gran parte de estos modelos son alimentados por datos recopilados mediante formularios de registros de la propia universidad o a través de encuestas por parte de organizaciones gubernamentales de cada país. En general, estos modelos alcanzan una precisión que oscila entre el 74% y el 95%, pero utilizan generalmente pocos registros, mientras que los modelos que usan grandes volúmenes de datos su precisión no supera el 86%. Estos grupos de registros están desbalanceados, los datos desbalanceados provocan que los modelos tengan mayor dificultad cuando tratan de clasificar un abandono positivo, pues los algoritmos de clasificación tienden a ignorar las clases con menor número de

apariciones (clase minoritaria), en este caso los estudiantes que si desertan de la universidad [4].

Metodología

Modelo híbrido con función de optimización Recocido Simulado (MHSA)

Un componente Bagging es un método que permite ensamblar múltiples modelos, construidos a partir del mismo algoritmo base, que utilizan el mismo conjunto de datos para tomar la decisión de la clasificación de un nuevo registro por medio del voto mayoritario [5].

El modelo MHSA propuesto, es un modelo híbrido que utiliza tres componentes Bagging como base para predecir la deserción de estudiantes universitarios. Cada componente Bagging usa respectivamente las técnicas de minería de datos: Árbol de decisión, Naive Bayes, Support Vector Machine [6]. Además, para mejorar la calidad del modelo, se utilizó la función de optimización Recocido Simulado (Simulated Annealing) [7]. El diseño del modelo MHSA propuesto se presenta en la Figura 1.

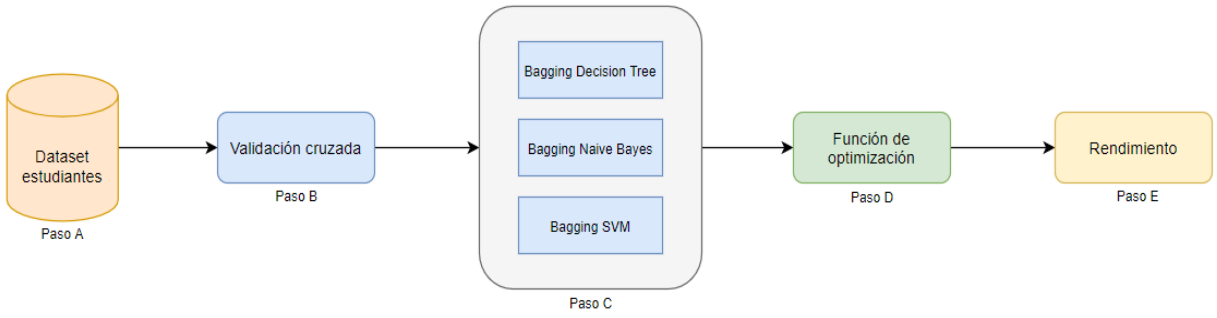

Figura 1. Diseño del modelo MHSA

A continuación, se detalla el proceso de funcionamiento del modelo MHSA. En el paso A se obtiene el conjunto de datos de los estudiantes. En el paso B se realiza el proceso de validación cruzada de 10 Folders para el entrenamiento y prueba del modelo. Seguidamente en el paso C, se crea el modelo Bagging Híbrido que contiene los tres componentes Bagging. En el paso D, Una vez finalizado el proceso

de creación del modelo, los resultados generados se envían a la función de optimización donde se asigna un peso a cada resultado para luego ser evaluado mediante el voto ponderado como lo muestra la Figura 2.

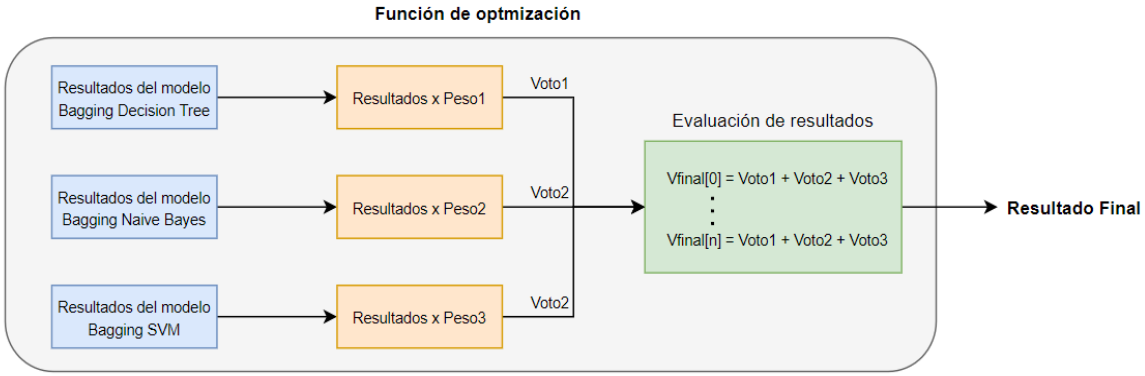

Figura 2. Función de optimización Simulated Annealing

El pseudocódigo del algoritmo del modelo MHSA se presenta en la Figura 3.

```
CreateBaggingModel_MHSA(int IT boolean weight, List[ClasifierModel] clasiffierModels)
1       N ←IT
2       weight_model ← weight
3       array_clasiffier ← clasifierModels
4       dataset ← loadDatasets()
5       train, target ← splitDataset(dataset)
6       array_predicted ← [ ]
7       for item ← 0 to array_clasiffier-1 do
8               model ← buildBagging(dataset, item.classifier)
9               model_predicted ← cross_val_predict(model, train, target, 10)
10              array_predicted.append(model_predicted)
11      end_for
12    current_score ←precisionModelByWeighted(array_predicted, array_weight, target)
13     solution_weight ← array_weight
14     for i ← to N-1 do
15        getNeighbor(array_weight)
16        score_diff ←precisionModelByWeighted(array_predicted, array_weight, target)
- current_score
17        if score_diff > 0
18            solution_weight ← array_weight
19            current_score ← precisionModelByWeighted(array_predicted,
array_weight, target)
20        else
21            if random(0,1) < Math.Exp(-score_diff / current_score)
22                solution_weight ←array_weight
23            end_if
24        end_if
25    end_for
end_function
```

Figura 3. Pseudocódigo del modelo MHSA

A continuación, se presenta la explicación del pseudocódigo del algoritmo del modelo MHSA. En la línea 1, a la variable *N* se asigna el valor del parámetro de entrada *IT* que contiene el número de iteraciones que se realizaran para que el modelo encuentre los mejores valores de pesos y se evalúen con los resultados, esto con el fin de mejorar la calidad de precisión del modelo. Estos elementos se explican detalladamente en una sección posterior. En la línea 2, a la variable *weigth_model* se asigna un valor del parámetro de entrada *weigth* de tipo booleano, en caso de ser verdadero el proceso del modelo se desarrollará con pesos iniciales para cada componente preestablecido, de lo contrario los pesos para los componentes internos se asignarán por partes iguales. En la línea 3, se guarda en *array_clasiffier* los clasificadores de tipo *ClassifierModel* que llegan por parámetros mediante la variable *classifierModels*, a los que se les construirá un ensamble de tipo Bagging. En la línea 4, a la variable *dataset* se asigna el conjunto de datos que retorna la función *loadDatasets*, esta función devuelve el conjunto de datos a usar. En la línea 5, la variable *train* guarda el conjunto de datos de entrenamiento y la variable *target* guarda el conjunto de datos objetivo que retorna la función *splitDataset*, esta función recibe como parámetro la variable *dataset* que se encarga separar la variable clase de los demás atributos y finalmente retorna el arreglo de datos a entrenar y el arreglo de datos objetivo. En la línea 6, se crea la variable *array_predicted* y se inicializa como una lista vacía donde posteriormente se guardarán los modelos entrenados. De la línea 7 a la línea 11, se realiza el proceso iterativo de construcción del ensamble tipo Bagging y entrenamiento de cada modelo guardado en la lista *array_clasiffier*. En la línea 8, la variable *model* guarda el modelo Bagging preparado para su entrenamiento que retorna la función *buildBagging*, esta función recibe como parámetros la variable *dataset* que contiene el conjunto de entrenamiento y el objeto del clasificador de tipo *ClassifierModel*. En la línea 9, a la variable *model_predicted* se asigna el modelo entrenado que retorna la función *cross_val_predict*, a la cual se le envían como parámetros la variable *train* que contiene el conjunto de datos a entrenar, la variable *target* que contiene los datos objetivos y la constante 10 que indica el número de folders. La función

cross_val_predict realiza el proceso de validación cruzada. En la línea 10, el valor de la variable *model_predicted* que contiene el modelo entrenado se agrega la lista de los modelos entrenados *array_predicted*. En la línea 12, a la variable *current_score* se le asigna el valor de la precisión del modelo mediante la función *precisionModelByWeighted,* esta función se encarga de retornar la precisión del modelo. En la línea 13, se crea la variable *solution_weight* la cual se le asigna el arreglo de pesos *array_weight*. De la línea 14 a la línea 25, se da inicio al proceso iterativo que permite calcular la mejor precisión del modelo basándose en el algoritmo de optimización Simulated Annnealing. En la línea 15, se invoca la función *getNeighbor* que recibe como parámetro el arreglo de pesos *array_weight*. La función *getNeighbor* identifica de forma aleatoria los pesos vecinos pertenecientes al arreglo de pesos *array_weight*. En la línea 16, a la variable *score_diff* se le asigna el valor de la nueva precisión que se obtiene mediante la función *precisionModelByWeighted* menos el valor de la variable *current_score*. En la línea 17 pregunta si el valor de *score_diff* es mayor a cero. Si es así, en la línea 18, a la variable *soluction_weight* se le asigna el valor del arreglo de pesos *array_weight*. En la línea 19, a la variable *current_score* se le asigna el nuevo valor de la precisión mediante la función *precisionModelByWeighted* enviando como parámetros el arreglo de modelos *array_predicted* y el nuevo arreglo de pesos *array_weight*. Si no, en la línea 21 pregunta, si el valor decimal aleatorio entre 0 y 1 es menor que el valor de la función exponencial de la variable *score_diff* sobre la variable *current_score*. Si es así, en la línea 22, a la variable *solution_weight* se le asigna el valor del arreglo de pesos *array_weight* que contiene los pesos optimizados de los modelos.

Resultados

El modelo MHSA se entrenó con un conjunto de datos balanceado de 9779 registros que contiene información de los estudiantes matriculados en el año 2018 y 2019 pertenecientes a la Universidad del Cauca, se comparó los resultados del modelo MHSA frente a otros modelos, donde se observó que el modelo MHSA obtiene mejor

desempeño, evidenciado mediante las métricas de precisión, precisión de verdaderos positivos (Precision TP), sensibilidad y el área bajo la curva (AUC) [8, 9].

Modelo	Precisión	Precisión TP	Recall	AUC
Modelo Bagging Híbrido Simulated Annealing	75%	75%	84%	80%
Modelo Bagging Arbol de decisión	67%	70%	64%	75%
Modelo Bagging Naive Bayes	54%	54%	73%	52%
Modelo Bagging Support Vector Machine (SVM).	54%	54%	73%	52%

Figura 4. Resultados de los modelos

Finalmente, el modelo MHSA se validó con un conjunto de datos de prueba de 1472 registros pertenecientes a estudiantes matriculados en la Universidad del Cauca para el año 2020. En la Figura 7, se presentan los resultados de precisión, precisión de verdaderos positivos (Precision TP), sensibilidad y el área bajo la curva (AUC).

Modelo	Precisión	Precisión TP	Recall	AUC
Modelo Bagging Híbrido Simulated Annealing	74%	72%	70%	77%

Figura 7. Resultados del modelo MHSA

Los resultados detallados del modelo MHSA se pueden evidenciar mediante la siguiente matriz de confusión (Confusion Matrix) que es una tabla que permite observar fácilmente que tipos de aciertos y errores tiene el modelo que se está entrenando [10].

		Predicción	
		Positivos	Negativos
Real	Positivos	1349	533
	Negativos	512	1418

Figura 8. Matriz de confusión del modelo MHSA

Según la matriz de confusión, los verdaderos positivos corresponden en el modelo a 1349 estudiantes que fueron clasificados correctamente como desertores. Asimismo, se observó que 512 estudiantes desertores fueron clasificados

incorrectamente por el modelo, esta cantidad se muestra como falsos positivos. Por otro lado, se observó que el valor de 1418 representa la cantidad de verdaderos negativos, que son aquellos estudiantes no desertores clasificados de manera correcta por el modelo. Igualmente, los falsos negativos son 533 registros de estudiantes que el modelo clasifica incorrectamente.

En cuanto a la precisión de los verdaderos positivos, el modelo identificó correctamente el 72% de los estudiantes desertores. En cuanto a la sensibilidad del modelo, este identifica correctamente el 70% de los verdaderos positivos.

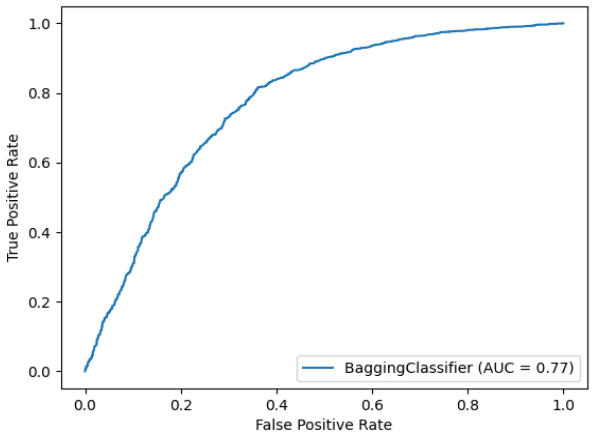

Figura 9. Curva ROC del modelo propuesto

Por otro lado, la métrica del área bajo la curva (AUC) presenta un valor de 77% indicando la calidad de precisión que tiene el modelo para identificar los verdaderos positivos y verdaderos negativos.

Conclusiones

Se pudo evidenciar una mejora en la calidad de la predicción de la deserción estudiantil universitaria utilizando un modelo que combina tres técnicas de minería de datos junto con una función de optimización, el cual fue denominado modelo MHSA. Este a su vez utilizó un conjunto de datos con 9779 registros balanceados. Los resultados del modelo híbrido propuesto fueron los siguientes: precisión de 75%, precisión de verdaderos positivos de 75%, sensibilidad (Recall) de 84% y el área bajo la curva (AUC) de 80%, en comparación al mejor de los otros modelos,

que fue el modelo de Árbol de decisión, el cual obtuvo una precisión de 67%, precisión de verdaderos positivos de 70%, sensibilidad (Recall) de 64% y el área bajo la curva (AUC) de 75%.

Al evaluar los modelos con un conjunto de datos balanceado, se identificó que la métrica de la precisión no es la mas alta, debido a que la cantidad de apariciones de las variables de clase tienden a ser iguales. Sin embargo, al emplear otras métricas como la precisión de verdaderos positivos, la sensibilidad (Recall) y el área bajo la curva (AUC), los resultados de estas métricas propenden a tener un valor alto, esto le permite al modelo no tener una gran dificultad al tratar de clasificar un abandono positivo.

La comparación de los resultados de los modelos permitió evidenciar que existen diferentes atributos que determinan la deserción de un estudiante y que algunas técnicas de minería de datos tienen mejor desempeño con ciertos atributos, como trabajo futuro se plantea crear un modelo que permita escoger las técnicas de minería de datos a usar de acuerdo a la naturaleza de los atributos.

Referencias bibliográficas

[1] MEN, "Estrategias para la Permanencia en Educación Superior: Experiencias Significativas," Bogotá, Colombia: Publicación del Ministerio de Educación Nacional, 2018.

[2] MEN, "Deserción estudiantil en la educación superior colombiana," Bogotá, Colombia: Publicación del Ministerio de Educación Nacional, 2009.

[3] MEN, "Deserción de la Educación Superior", Bogotá, Colombia, 2018. Disponible: https://www.mineducacion.gov.co/sistemasdeinformacion/1735/w3-article-357549.html?_noredirect=1

[4] A. Viloria, et al., "Retraction: Data Mining Applied in School Dropout Prediction", Conference Series, vol. 1432, p. 012107, 2020. DOI: 10.1088/1742-6596/1432/1/012107

[5] L. Breiman, "Bagging predictors," Machine Learning, vol. 24, pp. 123-140, 1996/08/01 1996. DOI: https://doi.org/10.1007/BF00058655

[6] S. Bhojani and N. Bhatt, "Data Mining Techniques and Trends – A Review," 2016. DOI: https://doi.org/10.36106/gjra

[7] C. Millán Páramo, et al., "Propuesta y validación de un algoritmo Simulated annealing modificado para la solución de problemas de optimización," Revista Internacional de Métodos Numéricos para Cálculo y Diseño en Ingeniería, vol. 30, pp. 264-270, 2014/10/01/ 2014. DOI: https://doi.org/10.1016/j.rimni.2013.10.003

[8] L. C. Jaime Cerda, "Uso de curvas ROC en investigación," Revista chilena de infectología, 2012. DOI: http://dx.doi.org/10.4067/S0716-10182012000200003

[9] D. Powers and Ailab, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation," J. Mach. Learn. Technol, vol. 2, pp. 2229-3981, 2011. DOI: https://arxiv.org/abs/2010.16061

[10] S. Visa, et al., "Confusion Matrix-based Feature Selection", vol. 710, 2011.

Anexo D

A continuación, se indica la dirección donde se encuentra el repositorio del código:
https://github.com/johaning91/classifiermodels.git