Obtención de un modelo de minería de datos aplicado a la deserción **universitaria del programa de Ingeniería de Sistemas de la Universidad de Cundinamarca**

Holmes Yesid Ayala-Yaguara[1]
Universidad de Cundinamarca – Extensión Facatativá
hayala@ucundinamarca.edu.co

Gina Maribel Valenzuela-Sabogal[2]
Universidad de Cundinamarca - Extensión Facatativá
gvalenzuela@ucundinamarca.edu.co

Alexander Espinosa-García[3]
Universidad de Cundinamarca - Extensión Facatativá
aespinosa1@ucundinamarca.edu.co

DOI: https://doi.org/10.21158/23823399.v7.n0.2019.2676

Fecha de recepción: 12 de marzo de 2020
Fecha de aprobación: 30 de julio de 2020

Cómo citar este artículo: Ayala-Yaguara, H. Y.; Valenzuela-Sabogal, G. M.; Espinosa-García, A. (2019). Obtención de un modelo de minería de datos aplicado a la deserción universitaria del programa de Ingeniería de Sistemas de la Universidad de Cundinamarca. *Revista Ontare, 7,* 133-150.
DOI: https://doi.org/10.21158/23823399.v7.n0.2019.2676

[1] Estudiante del programa de Ingeniería de Sistemas en la Universidad de Cundinamarca. **ORCID:** https://orcid.org/0000-0003-0528-3161
[2] Investigador Junior (IJ), Min Ciencias. Líder grupo de Investigación GISTFA. Magister en Administración y Planificación Educativa, Universidad Metropolitana de Educación, Ciencia y Tecnología, UMECIT. Especialización en Docencia Universitaria, Universidad Cooperativa de Colombia. Pregrado en Ingeniería de Sistemas, Fundación Universidad INCCA De Colombia. **ORCID:** https://orcid.org/0000-0002-2833-1579
[3] Investigador del grupo de Investigación GISTFA. - Magister en Ingeniería de Sistemas y Computación, Universidad Nacional de Colombia. Pregrado en Ingeniería Electrónica, Universidad Distrital Francisco José de Caldas. **ORCID:** https://orcid.org/0000-0002-6571-7294

RESUMEN

En el presente artículo se describe la obtención de un modelo de minería de datos aplicado al problema de la deserción universitaria en el programa de Ingeniería de Sistemas de la Universidad de Cundinamarca, extensión Facatativá. El modelo se estructuró mediante la metodología de minería de datos KDD (*knowledge discovery in databases*) haciendo uso del lenguaje de programación Python, la librería de procesamiento de datos Pandas y de *machine learning Sklearn*. Para el proceso se tuvieron en cuenta problemas adicionales al proceso de minería, como, por ejemplo, la alta dimensionalidad, por lo cual se aplicaron los métodos de selección de las variables estadístico univariado, *feature importance* y *SelectFromModel* (Sklearn). En el proyecto se seleccionaron cinco técnicas de minería de datos para evaluarlas: vecinos más cercanos (*K nearest neighbors, KNN*), árboles de decisión (*decision tree, DT*), árboles aleatorios (*random forest, RF*), regresión logística (*logistic regression, LR*) y máquinas de vectores soporte (*support vector machines, SVM*). Respecto a la selección del modelo final se evaluaron los resultados de cada modelo en las métricas de precisión, matriz de confusión y métricas adicionales de la matriz de confusión. Por último, se ajustaron los parámetros del modelo seleccionado y se evaluó la generalización del modelo al graficar su curva de aprendizaje.

Palabras clave: deserción universitaria; minería de datos; vecinos más cercanos; árbol de decisión; árbol aleatorio; regresión logística; máquinas de vectores soporte.

Obtaining a data mining model to be applied **to university desertion from the Systems Engineering program of the University of Cundinamarca**

ABSTRACT

This article describes how a data mining model was obtained and applied to the problem of university dropout in the Systems Engineering program of the University of Cundinamarca, in Facatativá. The model was structured by means of the KDD (knowledge discovery in databases) data mining methodology using Python programming language, Pandas data processing library, and the Sklearn machine learning. For the process, we took into account problems that are additional to the ones specific to the mining process, such as high dimensionality, reason why the methods of selection of the univariate statistical variables, feature importance, and SelectFromModel (Sklearn) were applied. In the project, five data mining techniques were selected for evaluation: nearest neighbors (KNN), decision tree (DT), random forest (RF), logistic regression (LR), and support vector machines (SVM). Regarding the selection of the final model, the results of each model were tested on the precision metrics, confusion matrix, and additional metrics of the confusion matrix. Finally, the parameters of the selected model were adjusted and the generalization of the model was evaluated by plotting its learning curve.

Keywords: university dropout; data mining; nearest neighbors; decision tree; random forest; logistic regression; support vector machines.

Obtenção de um modelo de mineração de dados **aplicado a evasões universitárias do programa de Engenharia de Sistemas da Universidade de Cundinamarca**

RESUMO

Este artigo descreve a obtenção de um modelo de mineração de dados aplicado ao problema da evasão universitária no curso de Engenharia de Sistemas da Universidade de Cundinamarca, campus Facatativá. O modelo foi estruturado utilizando a metodologia de mineração de dados KDD (*Knowledge Discovery in Databases*) utilizando também a linguagem de programação Python, a biblioteca de processamento de dados Pandas e a *Machine Learning Sklearn.* Para o processo, foram levados em consideração os problemas adicionais ao processo de mineração, como, por exemplo, a alta dimensão, para a qual foram aplicados os métodos de seleção das variáveis estatísticas univariadas, feature importance e *SelectFromModel* (Sklearn). No projeto, foram selecionadas cinco técnicas de mineração de dados para avaliá-las: vizinhos mais próximos (*K Nearest Neighbors, KNN*), árvores de decisão (DT), árvores aleatórias (*Random Forest, RF*), regressão logística, (*Logistic Regression LR*) e máquinas de vetores de suporte (*Support Vector Machines, SVM*). No que diz respeito à seleção do modelo final, foram avaliados os resultados de cada modelo nas métricas de precisão, matriz de confusão e métricas adicionais da matriz de confusão. Por fim, os parâmetros do modelo selecionado foram ajustados e a generalização do modelo foi avaliada no momento de diagramar sua curva de aprendizado.

Palavras-chave: evasão universitária; mineração de dados; vizinhos mais próximos; árvore de decisão; árvore aleatória; regressão logística; máquinas de vetores de suporte.

Modèle d'analyse de données appliqué au **décrochage universitaire du cursus d'ingénierie informatique de l'université de Cundinamarca**

RÉSUMÉ

Cet article décrit l'obtention d'un modèle d'exploitation des données appliqué à la problématique du décrochage universitaire du cursus d'ingénierie informatique du campus de Facatativá de l'université de Cundinamarca. Le modèle a été structuré à l'aide de la méthodologie d'exploitation de données KDD (*knowledge discovery in databases*) utilisant le langage de programmation Python, la bibliothèque de traitement de données Pandas et l'apprentissage automatique Sklearn. Lors de ce processus, des indicateurs supplémentaires d'extraction ont été retenus comme la dimensionnalité élevée, les méthodes de sélection des variables statistiques univariées ou l'importance des caractéristiques du SelectFromModel (Sklearn). Dans ce projet, cinq techniques d'exploitation des données ont été sélectionnées et soumises à évaluation: les parentés les plus proches (KNN), les matrices de décision (DT), les arbres aléatoires (RF), la régression logistique (LR) et les supports de machines vectorielles (SVM). Pour la sélection du modèle final, chaque résultat des différents modèles a été évalué grâce aux métriques de précision, à la matrice de confusion et aux métriques supplémentaires de la matrice de confusion. Enfin, les paramètres du modèle sélectionné ont été ajustés et la généralisation du modèle évaluée via le traçage graphique de la courbe d'apprentissage.

Mots clés: décrochage universitaire; extraction de données; arbre de décision; arbre aléatoire; régression logistique; support de machines vectorielles.

1. Introducción

Los índices de deserción en la educación superior son altos, ya que la cantidad de alumnos que culminan sus estudios es baja, estimándose, según estadísticas del Ministerio de Educación Nacional, en adelante MEN, que cerca de la mitad de los estudiantes no logran culminar estudios (MEN, 2009). Los elementos que caracterizan la deserción son muy diversos: los estudiantes, las instituciones de educación superior (IES) y las entidades responsables de las políticas de educación nacional (MEN, 2009). Por tanto, la complejidad de la deserción realza la importancia del análisis y el estudio de los factores que afectan este fenómeno a fin de identificar los factores que más lo influyen y estar en capacidad de predecir el posible abandono de un estudiante.

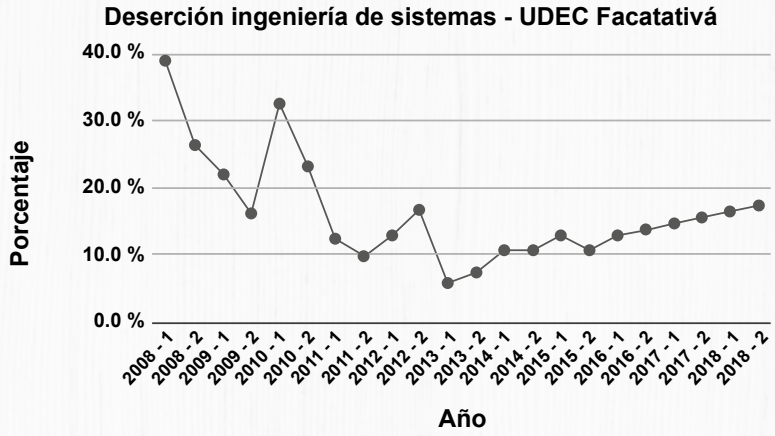
En este documento se expone la obtención de un modelo de minería por medio de la selección de una metodología que estructure el proceso, el desarrollo del modelo según la metodología escogida y, de igual forma, lleva a cabo una comparación de técnicas de minería para seleccionar aquella que presente los mejores resultados según las métricas de rendimiento planteadas. El proceso se desarrolla bajo el lenguaje de programación Python y las librerías Pandas y Sklearn.

2. Contexto del problema

La población objeto de estudio del proyecto comprende los estudiantes del programa de Ingeniería de Sistemas de la Universidad de Cundinamarca (UCundinamarca), extensión Facatativá. De acuerdo con las estadísticas del Sistema para la Prevención de la Deserción en la Educación Superior, Spadies, la Universidad de Cundinamarca, bajo el área de conocimiento del programa de Ingeniería de Sistemas —ingeniería, arquitectura, urbanismo y afines—, presenta en el periodo 2014-2019 un incremento del índice de deserción, pasando del 20 % al 43 %, de manera que se ubica entre las áreas de conocimiento con los índices de deserción más altos.

Del *Boletín Estadístico* —décima edición— entregado por la Dirección de Planeación Institucional de la UCundinamarca, se obtuvieron los índices de deserción para el programa de Ingeniería de Sistemas de la extensión Facatativá (véase la Figura 1), en el cual se evidencia que desde el 2013 hasta el 2018 hubo un incremento del 11,5 % en el índice de deserción.

Figura 1. Deserción del programa de Ingeniería de Sistemas —2008, I P. A.-2018 II P. A.—



Fuente. Universidad de Cundinamarca, 2019.

3. Minería de datos

La minería de datos es el proceso de examinar exhaustiva y minuciosamente grandes cantidades de datos a fin de identificar, extraer y descubrir nuevo conocimiento, de manera automática (Galvis y Martínez, 2004). Las metodologías de minería de datos permiten estructurar el proceso de minería de datos en fases que se encuentran interrelacionadas entre sí, convirtiendo el proceso de minería en un proceso iterativo e interactivo (Fischer, 2012).

Entre las metodologías más utilizadas se encuentra KDD, escogida para este proyecto como el modelo para establecer las etapas principales del proceso de extracción de información; se considera que el término *KDD* y minería de datos se utilizan indistintamente para hacer referencia al proceso completo

de descubrimiento de conocimiento (Moine, Aedo y Gordillo, 2011). Se abordaron, entonces, cinco fases: selección, preprocesamiento, transformación, minería-evaluación e interpretación (Hernández, Ramírez y Ferri, 2004).

4. Obtención del modelo

4.1 Selección

La Universidad de Cundinamarca no dispone de una base de datos institucional que consolide la información necesaria, por tanto, los datos se obtuvieron de repositorios y por medio de encuestas realizadas a estudiantes del programa de Ingeniería de Sistemas; se recopilaron en total de 649 registros, 100 de ellos a partir de las encuestas.

Ahora bien, el Ministerio de Educación Nacional (MEN, 2009) establece un conjunto de variables que permite describir el problema de la deserción, clasificadas en cuatro aspectos: individuales, socioeconómicos, académicos e institucionales. Teniendo en cuenta la disponibilidad de los datos, se seleccionaron 22 variables (véase la Tabla 1).

4.2 Preprocesamiento

En esta fase se hace la conversión de los archivos base que contienen los datos —Excel-XLSX—, a una estructura de datos que pueda ser interpretada por las tecnologías usadas —Dataframe—. También se tratan los datos erróneos que son reemplazados por el promedio de su columna respectiva.

4.3 Transformación

La transformación de variables consiste en determinar los tipos nominal u ordinal de cada variable, y en caso de ser nominal se separan en subatributos. Para las variables de estudio (véase la Tabla 1), el proceso de transformación resultó en 32 variables (véase la Tabla 2), incluido el atributo clase.

Tabla 1. Variables iniciales por categoría, b) variables transformadas

Variables			
Individual	**Socioeconómico**	**Académico**	**Institucional**
Género	Desea posgrado	Razón de elegir la universidad	Estrato
Edad	Preparacion previa		Costeo estudios
Zona residencia	Tipo de colegio		Trabaja actualmente
Estado civil	Tiempo de viaje		Estado estudiante (Desertor-No desetor)
Padres conviven	Tiempo de estudio		
Padre trabaja			
Madre trabaja			
Estudios padre			
Estudios madre			
Vive con familia			
Tamaño familia			
Discapacidad			

Fuente. Elaboración propia.

Tabla 2. Variables transformadas

Variables			
Género masculino	Género femenino	Discapacidad sí	Discapacidad no
Padre trabaja	Madre trabaja	Estudios padre	Estudios madre
Tipo colegio privado	Tipo colegio público	Pareja sí	Pareja no
Razón cercanía	Razón disponibilidad	Razón reputación	Razón otro
Trabaja actual sí	Trabaja actual no	Preparación previa sí	Preparación previa no

Costeo recursos propios	Costeo recursos familiares	Costeo estudios otros	Estrato
Edad	Zona residencia	Padres conviven	Desea posgrado
Vive con familia	Tamaño familia	Tiempo de viaje	Estudio independiente
Estado estudiante (variable clase)			

Fuente. Elaboración propia

4.4 Minería

4.4.1 Atributos relevantes.

En este proyecto se consideró tener en cuenta una mayor cantidad de atributos con el propósito de describir mejor el problema, pero muchos autores coinciden en que esto puede afectar el rendimiento de los modelos de *machine learning* (Raschka, 2015). Se utilizaron, entonces, tres métodos de selección de atributos: el estadístico univariado que calcula si existe una relación estadísticamente significativa entre cada atributo y su variable clase por medio de percentiles (Müller y Guido, 2016); el segundo método es el *feature importance* basado en bosques aleatorios, mide la importancia de la característica a medida que disminuyen los datos poco correlacionados (Raschka, 2015); el tercero es SelectFromModel (Sklearn), el cual utiliza una máquina de aprendizaje supervisado para juzgar la importancia de cada característica por modelo y solo mantiene las más importantes. Los resultados se evidencian en la tabla 1, en la cual se obtuvo un conjunto reducido de diez atributos —atributos resaltados— comunes a los tres métodos de selección.}

Tabla 3. Conjunto de atributos reducido

Atributos	Feature Imprtance	Estadístico univariado	Select From Model
Edad	8.648 %	Sí	Sí
Estudio independiente	6.802 %	Sí	Sí
Estudios madre	6.789 %	Sí	Sí
Estudios padre	6.674 %	Sí	Sí
Tiempo viaje	5.247 %	No	Sí

Desea posgrado	5.106 %	Sí	Sí
Estrato	4.454 %	Sí	Sí
Tipo colegio público	3.845 %	Sí	Sí
Tipo colegio privado	3.773 %	Sí	Sí
Vive con familia	3.092 %	No	Sí
Zona residencia	2.885 %		Sí
Tamaño familia	2.820 %	No	Sí
Madre trabaja	2.797 %	Sí	Sí
Razón disponibilidad de la carrera	2.792 %	Sí	Sí
Genero femenino	2.638 %		Sí
Genero masculino	2.624 %		Sí

Fuente. Elaboración propia.

4.4.2 Evaluación de modelos.

Antes de la evaluación se dividieron los datos en un conjunto de entrenamiento y uno prueba, usando división estratificada a fin de obtener una proporción de clases igual al set de datos completo. Cabe resaltar que cada prueba se llevó a cabo utilizando el set de atributos completo (véase la Tabla 1) y el set de atributos reducido (Tabla 2).

4.4.2.1 Técnicas.

Las técnicas seleccionadas para identificar el prototipo de modelo fueron: vecinos más cercanos (*K nearest neighbors*, KNN), árboles de decisión (*decision tree,* DT), árboles aleatorios (*random forest*, RF), regresión logística (*logistic regression,* LR) y máquinas de vectores soporte (*support vector machines*, SVM).

4.4.2.2 Precisión.

Se aplicó la regresión logística, la cual consiste en dividir los aciertos del modelo sobre el total de muestras. Se obtuvieron los mejores resultados como se aprecia en la tabla 4.

Tabla 4. Resultados, métrica de precisión

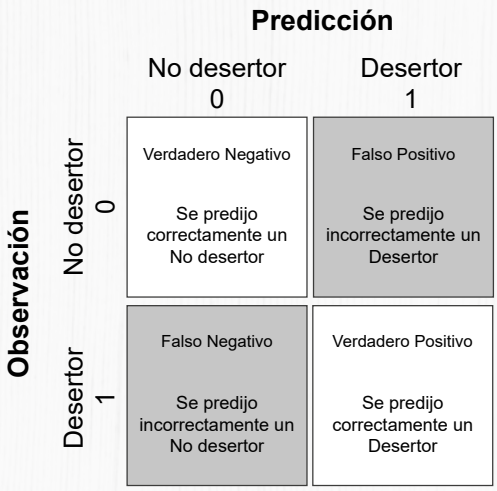
# Atributos	Regresión Logística	SVM	KNN	Random Forest	Arbol de decisión
Todos	76.0 %	74.0 %	75.0 %	73.0 %	66.0 %
32	73.0 %	74.0 %	74.0 %	74.0 %	65.0 %
10	74.0 %	72.0 %	70.0 %	67.0 %	63.0 %
Promedio	74.3 %	73.3 %	73.0 %	71.3 %	64.7 %

Fuente. Elaboración propia.

4.4.2.3 Matriz de confusión.

Permite saber cómo distribuye los errores un modelo clasificador, según la clasificación correcta o incorrecta de cada muestra (véase la Figura 2).

Figura 2. Descripción matriz de confusión



Fuente. Elaboración propia.

Los modelos de regresión logística y SVM obtuvieron la mayor tasa de predicciones acertadas (véase la Tabla 5). También se evidencia que los mejores resultados se obtienen al utilizar el set de atributos completo.

Tabla 5. Resultados, matriz de confusión

	Regresión Lgística		SVM		KNN		Random Forest		Árbol de Decisión	
# Atributos	10	32	10	32	10	32	10	32	10	32
Verdadero Positivo	38	50	32	39	37	55	46	49	52	62
Verdadero Negativo	293	284	296	296	279	280	256	278	238	214
Total	341	366	338	367	326	367	312	359	300	308
Aciertos totales	353.5		352.5		346.5		335.5		304	
Falso Positivo	23	32	20	20	37	36	60	38	78	102
Falso Negativo	100	88	106	99	101	83	92	89	86	76
Total	123	120	126	119	138	119	152	127	164	178
Fallos Totales	121.5		122.5		128.5		139.5		171	

Fuente. Elaboración propia.

4.4.2.4 Métricas adicionales.

Las matrices de confusión permiten evaluar métricas adicionales: precisión, a fin de obtener la probabilidad de clasificar correctamente un registro; alcance, con el propósito de clasificar correctamente un registro en su categoría correspondiente; y puntaje F1, para el promedio ponderado de precisión y sensibilidad. De la tabla 6 se aprecia que el modelo de regresión logística obtuvo los mejores resultados junto con el modelo KNN; el set de atributos completo obtuvo de nuevo los mejores resultados.

Tabla 6. Resultados, métricas adicionales

		Presición		Alcance		F1	
# Atributos		10	32	10	32	10	32
Regresión Logística	No desertor	75 %	76 %	92 %	89 %	83 %	82 %
	Desertor	62 %	60 %	29 %	37 %	40 %	46 %
	MP	71 %	71 %	73 %	73 %	70 %	71 %
SVM	No desertor	73 %	74%	94 %	93 %	82 %	82 %
	Desertor	61 %	61%	22 %	24 %	33 %	34 %
	MP	69 %	70 %	72 %	72 %	67 %	67 %

KNN	No desertor	73 %	77 %	87 %	88 %	79 %	82 %
	Desertor	46 %	59 %	25 %	38 %	32 %	46 %
	MP	65 %	72 %	68 %	73 %	65 %	71 %
Random Forest	No desertor	75 %	74 %	81 %	87 %	78 %	80 %
	Desertor	46 %	51 %	38 %	30 %	42 %	38 %
	MP	66 %	67 %	68 %	70 %	67 %	67 %
Árbol de decisión	No desertor	74 %	75 %	75 %	72 %	74 %	74 %
	Desertor	40 %	42 %	38 %	46 %	39 %	44 %
	MP	64 %	65 %	64 %	64 %	63 %	65 %

Fuente. Elaboración propia.

4.5 Evaluación e interpretación

Con los resultados de las métricas evaluadas, la regresión logística presenta el mayor índice promedio de precisión, indicando que entre las técnicas seleccionadas es la que mejor permite describir el problema de la deserción en el contexto planteado.

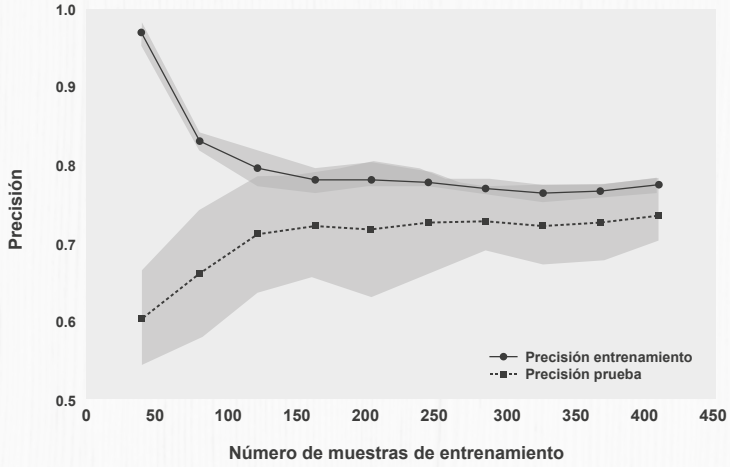
4.5.1 Ajustar modelo.

El modelo final se ajusta a los datos de entrenamiento haciendo uso de la herramienta GridSearchCV —Sklearn— que permite seleccionar los hiperparámetros de un modelo usando validación cruzada (Pedregosa *et al*., 2011).

4.5.2 Curvas de aprendizaje.

Con los parámetros óptimos encontrados con GridSearchCV se procede a comparar la precisión del modelo, teniendo en cuenta el set de entrenamiento y el set de prueba. Se obtiene una curva de aprendizaje por cada set que permite describir si el modelo sufre de alto sesgo —subajuste, curvas muy unidas— o alta varianza —sobreajuste, curvas muy separadas—. El modelo de regresión logística presenta baja varianza y bajo sesgo debido a que las curvas no están muy unidas ni muy separadas (véase la Figura 3).

Figura 3. Curva de aprendizaje, modelo de regresión logística



Fuente. Elaboración propia.

5. Resultados

El modelo obtenido de minería de datos se sustenta bajo la técnica de regresión logística debido a que presentó los mejores índices de precisión en cada una de las métricas de rendimiento planteadas. Además, se encontró que entre más atributos se tengan en cuenta, los índices de precisión de los modelos mejoran.

6. Conclusiones

La disponibilidad de datos es importante para caracterizar el problema según el contexto en el que se aplique, sin embargo, la Universidad de Cundinamarca solo dispone de datos porcentuales que describen la deserción a nivel general, por lo que se necesita de una consolidación más descriptiva respecto a los índices de deserción.

La alta dimensionalidad no fue un problema para el caso de estudio, evidenciando que el uso de un mayor espacio de atributos, según los datos disponibles, permite describir mejor el problema de la deserción.

De las técnicas de minería propuestas, la regresión logística fue la que obtuvo mejores resultados. No obstante, hace falta la obtención de un modelo que permita una mayor interpretabilidad de los resultados, como es el caso de los árboles de decisión que establecen condiciones que describen la relación entre variables.

Referencias

Fischer, E. S. (2012). *Modelo para la automatización del proceso de determinación de riesgo de deserción en estudiantes universitarios.* Santiago de Chile: Universidad de Chile.

Galvis, M.; Martínez, F. (2004). *Confrontación de dos técnicas de minería de datos aplicadas a un dominio específico.* Bogotá: Pontificia Universidad Javeriana.

Hernández, J.; Ramírez, J.; Ferri, C. (2004). *Introducción a la minería de datos.* Madrid: Pearson.

MEN (Ministerio de Educación Nacional). (2009). *Deserción estudiantil en la educación superior colombiana: metodología de seguimiento, diagnóstico y elementos para su prevención.* Bogotá: Ministerio de Educación Nacional.

Moine, J. M.; Gordillo, S.; Haedo, A. (2011). E*studio comparativo de metodologías para minería de datos.* Texto presentado en el XIII Workshop de Investigadores en Ciencias de la Computación. Red de Universidades con Carreras en Informática (RedUNCI), San Juan, Argentina, 5-6 de mayo. Recuperado de http://hdl.handle.net/10915/20034

Müller, A.; Guido, S. (2016). *Introduction to machine learning: a guide for data scientists*. Sebastopol CA: O' Reilly.

Pedregosa, F. et al. (2011). Scikit-learn: machine learning in python. *Journal of Machine Learning Research*, 12, 2826-2830. Recuperado de https://hal.inria.fr/hal-00650905

Raschka, S. (2015). *Python machine learning.* Birmingham: Packt Publishing.

Universidad de Cundinamarca. (2019). *Boletín Estadístico X Edición.* Universidad de Cundinamarca, Dirección Planeación Institucional. Fusagasugá: Fusunga Casa Editorial.