INFORME TÉCNICO DE REVISIÓN BIBLIOGRÁFICA

1. ANTECEDENTES

Dentro de la investigación científica, la metodología es una parte fundamental, ya que provee un conjunto de procedimientos racionales empleados para el logro de los objetivos. Utilizar una metodología durante el proceso de Minería de Datos (MD) permitirá una correcta creación de conocimiento. De tal manera, la buena elección de una metodología facultaría la sistematización de los procedimientos y técnicas que se requieren para resolver un problema.

Existen diferentes metodologías para llevar a cabo un trabajo de extracción de conocimiento, por ende, en la presente revisión bibliográfica se busca identificar la metodología más adecuada para utilizar en la creación de modelos en el desarrollo del proyecto titulado "Machine Learning para el análisis del rendimiento académico de estudiantes de Ingeniería".

2. DESARROLLO

Para identificar una metodología que permita extraer conocimiento de los datos académicos, se utilizó la técnica de investigación bibliográfica de tipo expositiva. El proceso que se desarrolló para cumplir con dicha técnica se presenta en la figura 1, la cual utiliza como base a la metodología descrita por Gómez-Luna et al [1].

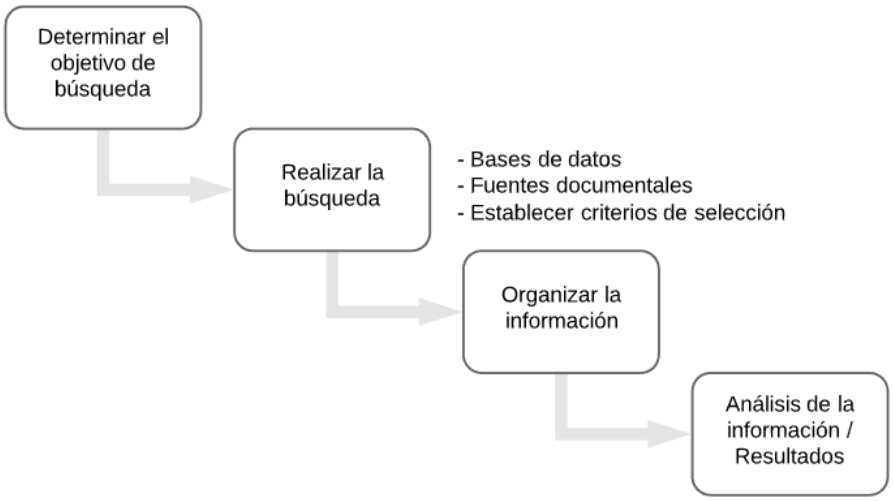

Figure 1 Proceso de la revisión bibliográfica

Para determinar el objetivo se planteó la búsqueda de información en artículos y documentos que aborden lo siguiente "metodologías de minería de datos" y que se relacionen con "predicción del rendimiento académico en la educación superior"; para la búsqueda de bibliografía científico-académica, se utilizó de una forma jerárquica las siguientes bases de datos (donde, 1 es la de mayor jerarquía y 5 la de menor): 1. Scopus, 2. Web of Science, 3. ACM, 4. IEEE Explore y 5. ScienceDirect. Adicionalmente, se utilizó Google Académico para complementar con una mínima cantidad de literatura gris.

Para realizar las cadenas de búsqueda se analizó los tesauros que mayores resultados arrojan, así como también, en la manera de manipular los documentos repetidos, se seleccionó los documentos alojados en las bases de datos de mayor jerarquía. Se utilizó la búsqueda avanzada para encontrar coincidencias en el título, el resumen y las palabras claves.

Las cadenas de búsqueda que se utilizaron en cada una de las bases de datos científicas y en el buscador académico, se muestran en la tabla 1.

Tabla 1 Cadenas de búsqueda.

BASE DE DATOS	CADENA DE BÚSQUEDA	RESULTADOS
Scopus	TITLE-ABS-KEY ("data mining" methodology AND "academic performance" AND ("higher education" OR "university")) AND DOCTYPE (ar OR cp) AND PUBYEAR > 2014	18
Web of Science	TEMA: (data mining methodology and academic performance and higher education) Refinado por: AÑOS DE PUBLICACIÓN: (2020 OR 2018 OR 2016 OR 2019 OR 2017 OR 2015) AND DOMINIOS DE INVESTIGACIÓN: (SCIENCE TECHNOLOGY) AND DOMINIOS DE INVESTIGACIÓN: (SCIENCE TECHNOLOGY) AND TIPOS DE DOCUMENTOS: (ARTICLE) Período de tiempo: Todos los años. Bases de datos: WOS, KJD, RSCI, SCIELO. Idioma de búsqueda=Auto	13
ACM	"query": {AllField:(Data mining methodology) AND Abstract:(higher education) AND Abstract:(predicting academic performance)} "filter": {ACM Pub type: Journals, Published in: ACM Transactions on Intelligent Systems and Technology, Publication Date: (01/01/2015 TO 05/31/2020)}	25
IEEE Explore	((("All Metadata":data mining methodology) AND "All Metadata": academic performance) AND "All Metadata":higher education)	7
ScienceDirect	Find articles with these terms: "data mining" and "academic performance" and "higher education" Refine by: Article type: Research articles (29) Publication title: Procedia Computer Science (18) Computers & Education (11)	29
Google Scholar	(Metodologías de "minería de datos") Y ("predicción del rendimiento académico") Y ("Educación Superior" O "Universidad")	96

Los criterios de inclusión que se establecieron para la búsqueda y la selección de información fueron:

- Fecha de publicación: publicaciones realizadas entre, enero del 2015 - mayo del 2020.
- Tipo: en las bases de datos se seleccionó artículos de revistas y documentos de conferencias. Para la literatura gris, se identificó las tesis de doctorado o maestría.
- Idioma: se recopiló principalmente documentos en inglés, en español y portugués.
- Disciplina: documentos vinculados con la educación superior.

Así mismo, se consideraron los siguientes criterios de exclusión:

- Autores: documentos en las cuales se desconocía su autoría.
- Contenido: Estudios de big data o aprendizaje profundo.

3. RESULTADOS

Luego de recopilar la información de interés, se procedió a organizar los documentos digitales, en donde se examinaron y se seleccionaron los trabajos afines a la investigación. A continuación, se organizó la información ordenada en forma cronológica, conforme se puede ver en la tabla 2; considerando las siguientes propiedades: año de publicación del estudio, el título, metodología de MD utilizada, técnica de MD implementada y los algoritmos que se utilizaron.

Tabla 2 Artículos encontrados

AÑO	TÍTULO	METODOLOGÍA DE MD	TÉCNICA DE MD	ALGORITMOS
2020	Using data mining techniques to predict student performance to support decision making in university admission systems	AJUSTADA	CLASIFICACIÓN	Red Neuronal Artificial
2020	Prediction of academic performance using artificial intelligence techniques	AJUSTADA	CLASIFICACIÓN	Árboles de decisión
2019	Predicción del rendimiento académico de los estudiantes de la UNSAAC a partir de sus datos de ingreso utilizando algoritmos de aprendizaje automático	CRISP-DM	CLASIFICACIÓN	Árboles de decisión
2019	Application of machine learning in predicting performance for computer engineering students: A case study	AJUSTADA	CLASIFICACIÓN	Árboles de decisión
2019	Árboles de decisión para predecir factores asociados al desempeño académico de estudiantes de bachillerato en las pruebas Saber 11°	CRISP-DM	CLASIFICACIÓN	Árboles de decisión
2019	Construcción e implementación de un modelo para predecir el rendimiento académico de estudiantes universitarios mediante el algoritmo Naïve Bayes	KDD	CLASIFICACIÓN	Naïve Bayes
2019	Academic performance based on gender using filter ranker algorithms - An experimental analysis in Sultanate of Oman	AJUSTADA	CLASIFICACIÓN	Árboles de decisión
2019	An ensemble-based model for prediction of academic performance of students in undergrad professional course	AJUSTADA	CLASIFICACIÓN	Árboles de decisión, gradient boost y Naïve Bayes
2019	Using Decision Tree and Artificial Neural Network to Predict Students Academic Performance	CRISP-DM	CLASIFICACIÓN	Árboles de decisión Y Red Neuronal Artificial
2019	Knowledge capture for the prediction and analysis of results of the quality test of higher education in Colombia	KDD	CLASIFICACIÓN	Red Neural Artificial
2019	Modelo de mineria de datos basado en factores asociados para la predicción de deserción estudiantil universitaria	CRISP-DM	CLASIFICACIÓN	Red Neural Artificial
2018	Early segmentation of students according to their academic performance: A predictive modelling approach	AJUSTADA	CLASIFICACIÓN	Árbol de decisión y SVM
2018	Bound Model of Clustering and Classification (BMCC) for Proficient Performance Prediction of Didactical Outcomes of Students	AJUSTADA	CLASIFICACIÓN Y AGRUPAMIENTO	Árbol de decisión y k-mean
2018	Comparative Analysis of Prediction Techniques to Determine Student Dropout: Logistic Regression vs Decision Trees	SAP	CLASIFICACIÓN	Regresión logística y Árboles de decisión

2018	Prediction of course completion by students of a university in Brazil	AJUSTADA	CLASIFICACIÓN	SVM
2018	Applying Predictive Analytics in Elective Course Recommender System while preserving Student Course Preferences	AJUSTADA	CLASIFICACIÓN	SVM
2018	Towards reliable prediction of academic performance of architecture students using data mining techniques	AJUSTADA	CLASIFICACIÓN	Regresión logística y SVM
2018	Predicting student academic performance using multi-model heterogeneous ensemble approach	AJUSTADA	CLASIFICACIÓN	Árboles de decisión Y Red Neuronal Artificial
2018	Predicción del rendimiento académico mediante minería de datos en estudiantes del primer ciclo de la Escuela Profesional de Ingeniería de Computación y Sistemas, Universidad de San Martín de Porres	CRISP-DM	CLASIFICACIÓN	Árboles de decisión
2018	Diseño de un modelo para automatizar la predicción del rendimiento académico en estudiantes del IPN	AJUSTADA	CLASIFICACIÓN	Naïve Bayes
2018	Minería de datos aplicada a la clasificación del rendimiento académico	CRISP-DM	CLASIFICACIÓN	Árboles de decisión, Naïve Bayes
2018	Detección de patrones de bajo rendimiento académico mediante técnicas de minería de datos de los estudiantes de la Universidad Nacional Amazónica De Madre De Dios 2018	CRISP-DM	CLASIFICACIÓN	Random Forest
2017	Analyzing undergraduate students' performance using educational data mining	AJUSTADA	CLASIFICACIÓN	Árboles de decisión
2017	Decision tree learning used for the classification of student archetypes in online courses	AJUSTADA	CLASIFICACIÓN	Árboles de decisión
2017	A review of applications of data mining techniques for prediction of students' performance in higher education	AJUSTADA	CLASIFICACIÓN	Árboles de decisión
2016	Predicting Students' Performance in University Courses: A Case Study and Tool in KSU Mathematics Department	AJUSTADA	CLASIFICACIÓN	CBA
2016	Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas de la formación profesional	CRISP-DM	CLASIFICACIÓN	Árboles de decisión
2015	Predicción del fracaso y el abandono escolar mediante técnicas de minería de datos	AJUSTADA	CLASIFICACIÓN	Árboles de decisión
2015	Student Dropout Predictive Model Using Data Mining Techniques	KDD	CLASIFICACIÓN	Árboles de decisión
2015	Predicción del Rendimiento Académico en carreras de Computación utilizando Árboles de Decisión	KDD	CLASIFICACIÓN	Árboles de decisión

Para exponer los resultados finales, es conveniente explicar cada una de las metodologías encontradas durante la revisión bibliográfica y que se encuentran organizadas en la tabla 2.

1. CRISP-DM

   Es una metodología estandarizada para cumplir con el ciclo de vida de un proyecto de análisis de datos. CRISP-DM se describe en términos de un modelo de proceso jerárquico, que consiste en conjuntos de tareas descritas en cuatro niveles de abstracción (de lo general a lo específico): fase, tarea genérica, tarea especializada e instancia de proceso [2].

2. KDD

   KDD es el proceso no trivial de identificación de patrones válidos, novedosos, potencialmente útiles y finalmente comprensibles en los datos. KDD contiene a la minería de datos como una de sus fases [3].

3. SAP

   Es una metodología de trabajo para tratar y crear modelos de predicción, propia de la empresa desarrolladora del software SAP Predictive Analytics [4].

4. AJUSTADA

   Con esta clasificación se hace referencia a los documentos que utilizaron una combinación de ciertas fases propias de KDD y CRISP-DM.

La Ilustración 1, simboliza el uso de las metodologías encontradas en los estudios analizados en la revisión bibliográfica, en la cual se obtuvo que: de 30 documentos recopilados y organizados, ocho (8) utilizaron la metodología CRISP-DM, cuatro (4) KDD, diecisiete (17) utilizaron una metodología ajustada y en un solo artículo se utilizó SAP.

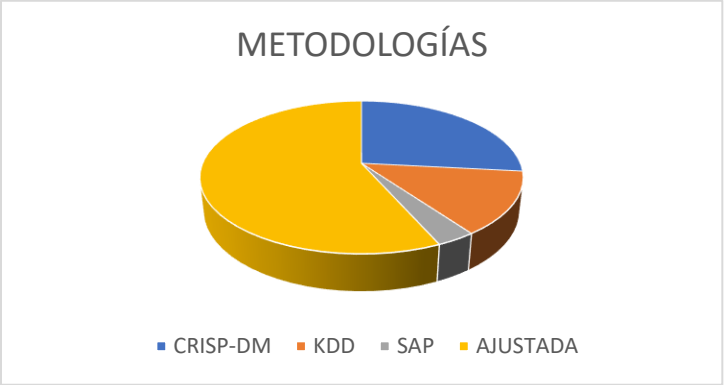

Ilustración 1 Metodologías encontradas

La Ilustración 2, representa el uso de las técnicas de minería de datos encontradas en los estudios analizados en la revisión bibliográfica, en la cual se obtuvo que: de 30 documentos recopilados y organizados, veintinueve (29) utilizaron la clasificación y en una sola investigación se aplicó simultáneamente las técnicas de regresión y clasificación.

Ilustración 2 Técnicas de minería de datos

En la Ilustración 3, representa los algoritmos que fueron utilizados en cada uno de los estudios analizados, en donde se conoció que: los árboles de decisión se implementaron 19 veces, seguido de las redes neuronales con 5 veces, Naive Bayes y SVM, que se implementaron 4 veces. La regresión logística se utilizó 2 veces, mientras que, Gradient Boost, Random Forest, CBA y K-means se utilizaron en una sola vez.

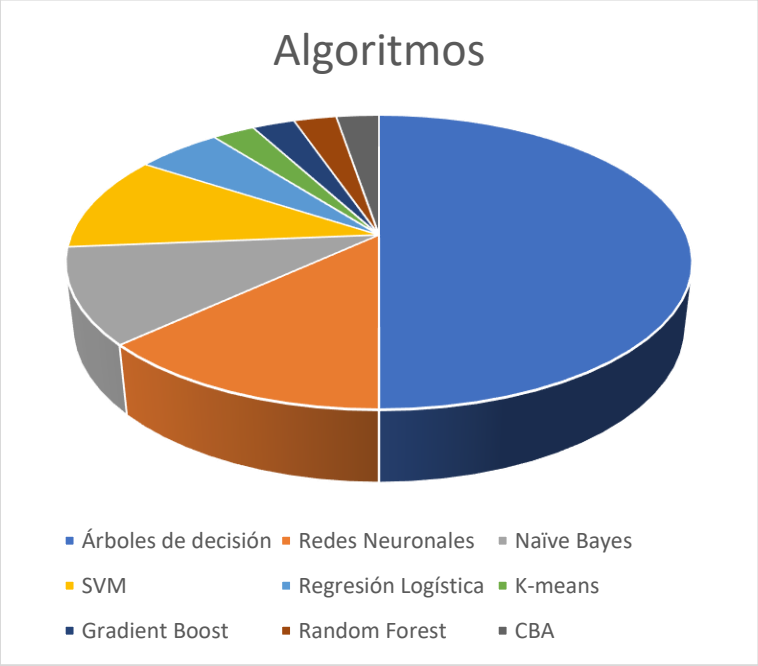

Ilustración 3 Algoritmos utilizados

4. CONCLUSIONES

- En las diversas investigaciones encontradas, distintos actores utilizan ciertas fases o partes específicas de KDD y CRISP-DM. Como la finalidad de la presente revisión bibliográfica es identificar una metodología para realizar una minería de datos educativos; entonces ignoramos a la metodología etiquetada como AJUSTADA.

- El análisis de la información y la presentación de los resultados mediante cuadros estadísticos permitió determinar que la metodología CRIPS-DM es la más utilizada en proyectos de minería de datos relacionados con la educación; por lo tanto, se fundamenta la selección y uso de dicha metodología.

- La metodología CRIPS-DM es la más completa, ya que comienza con la fase importante, que es la conciencia empresarial y abarca aspectos complementarios a los técnicos durante todo el proceso de minería de datos. Esta metodología que se centra en las necesidades de los gerentes para resolver problemas de gestión y además, brinda un marco referencial de los resultados que se deben obtener en cada etapa; por lo tanto, se puede afirmar que es la metodología más apropiada para utilizarla en el trabajo de titulación denominado "Machine Learning para el análisis del rendimiento académico de estudiantes de Ingeniería".

5. BIBLIOGRAFÍA

[1] E. Gómez-Luna, D. Fernando-Navas, G. Aponte-Mayor, and L. A. Betancourt-Buitrago, "Metodología para la revisión bibliográfica y la gestión de información de temas científicos, a través de su estructuración y sistematización," *DYNA*, vol. 81, no. 184, pp. 158–163, 2014, doi: 10.15446/dyna.v81n184.37066.

[2] P. Chapman *et al.*, "CRISP-DM 1.0," DaimlerChrysler, 2000.

[3] R. Pitre, U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, M. V. Chavan, and P. R. N. Phursule, "From Data Mining to Knowledge Discovery in Databases," *AI Mag.*, vol. 17, no. 3, Mar. 1996, doi: 10.1609/aimag.v17i3.1230.

[4] A. Pérez, E. E. Grandón, M. Caniupán, and G. Vargas, "Comparative Analysis of Prediction Techniques to Determine Student Dropout: Logistic Regression vs Decision Trees," in *Proceedings - International Conference of the Chilean Computer Science Society, SCCC*, 2018, vol. 2018-Novem, doi: 10.1109/SCCC.2018.8705262.