REVISTA INCAING
ISSN 2448 9131

Revisión literaria de indicadores que influyen en la educación mediante técnicas de minería de datos.

Luis Felipe Vergara Serrato, David Santiago Mahecha Hernández, y cesar Yesid Barahona Rodríguez

Resumen - Esta revisión literaria integrativa o estado del arte ha sido organizada de manera clara para abordar los temas relacionados con la educación de los estudiantes y las técnicas de minería de datos. La educación es un tema vital en la vida de los seres humanos, en Colombia, esta se divide en educación inicial, la educación preescolar, la educación básica (primaria cinco grados y secundaria cuatro grados), la educación media (dos grados y culmina con el título de bachiller), y la educación superior. Lo que se busca es analizar diferentes indicadores de desempeño en los estudiantes colombianos por medio de técnicas de minería de datos. Lo primero que haremos es encontrar cuales son estos indicadores por medio de diferentes autores, después de esto se hará una búsqueda de los patrones más eficaces y eficientes del análisis de datos.

Documento recibido el 9 de octubre de 2001. (Anote la fecha en que usted presentó su documento para su revisión.) Este trabajo fue apoyado en parte por los U.S. Departamento of Commerce under Grant S123456 (reconocimiento al patrocinador y apoyo financiero va aquí). los títulos del Documento deben ser escritos en letras mayúsculas y minúsculas, no todas las mayúsculas. Evite escribir fórmulas extensas con subíndices en el título; Utilice Fórmulas cortas que identifiquen los elementos (por ejemplo, "Nd-Fe-B"). No escriba "(invitados)" en el título. Escriba los Nombres completos de los autores en el campo autor, pero no es necesario. Ponga un espacio entre los autores.

F. A. Author is with the National Institute of Standards and Technology, Boulder, CO 80305 USA (corresponding author to provide phone: 303-555-5555; fax: 303-555-5555; e-mail: author@ boulder.nist.gov).

S. B. Author, Jr., was with Rice University, Houston, TX 77005 USA. He is now with the Department of Physics, Colorado State University, Fort Collins, CO 80523 USA (e-mail: author@lamar. colostate.edu).

T. C. Author is with the Electrical Engineering Department, University of Colorado, Boulder, CO 80309 USA, on leave from the National Research Institute for Metals, Tsukuba, Japan (e-mail: author@nrim.go.jp).

I. INTRODUCCION

Los indicadores de desempeño son instrumentos que proporcionan información cuantitativa sobre el desenvolvimiento y logros de una institución, programa, actividad o proyecto a favor de la población u objeto de su intervención, en el marco de sus objetivos estratégicos y su Misión. Los indicadores de desempeño establecen una relación entre dos o más variables, que, al ser comparados con periodos anteriores, productos similares o metas establecidas, permiten realizar inferencias sobre los avances y logros de las instituciones y/o programas.[1], para la búsqueda de estos indicadores se toman los conceptos claves como: academic performance, indicators, factors, school, análisys y family.

La minería de datos es el proceso de detectar la información procesable de los conjuntos grandes de datos. Utiliza el análisis matemático para deducir los patrones y tendencias que existen en los datos. Normalmente, estos patrones no se pueden detectar mediante la exploración tradicional de los datos porque las relaciones son demasiado complejas o porque hay demasiado datos. [2], Para el presente estudio se han examinado particularmente cuatro temáticas principales: data mining, decisión trees, y education.

Se usan como instrumentos los motores de búsqueda de las bases de datos de Scopus, ScienceDirect e IEEE, con el fin de determinar las publicaciones más recientes, así como el análisis de los temas que hacen parte del área. Por medio de los conceptos o temáticas principales se realizaron ecuaciones de búsqueda específicas.

La revisión bibliográfica fue un recurso para aplicar técnicas de mapeo VOS con el fin de realizar mediciones bibliométricas y descubrir tendencias en las investigaciones ligadas a la minería de datos e indicadores. Este análisis bibliométrico basado en un

mapa de coocurrencia de palabras claves, permite encontrar conexiones de temáticas propias de la minería de datos, como el uso de diferentes técnicas o algoritmos para la obtención de patrones y además encontrar los factores e incidencias del desempeño de los estudiantes.

II. MÉTODO

En la metodología se consideran las fuentes de los documentos estudiados, así como la manera de llegar a ellos mediante las ecuaciones de búsqueda, han sido elegidos Scopus, ScienceDirect e IEEE para la construcción del estado del arte debido a que son enfocados en la rama de la ingeniería además permite acceder a publicaciones procedentes de más de 5.000 editoriales internacionales revisadas por especialistas. Las publicaciones, revistas y bases de datos que fueron tenidas en cuenta en este trabajo son las expuestas a continuación:

- Elsevier: Es un editorial multimedia internacional de análisis de información general que asiste a instituciones en el avance de la ciencia, cuidados avanzados en temas médicos, así como mejorar la realización de los mismos para el beneficio de la sociedad. Los productos que se ofrecen incluyen revistas, colecciones de revistas electrónicas y el índice de citas bibliográficas de Scopus, entre otros. Es una importante editorial de libros académicos muy reconocida en la comunidad científica internacional por la calidad y su amplia lista de publicaciones, además sus divulgaciones incluyen también autores reconocidos, así como instituciones científicas [3].

- Scopus: Esta es la base de datos más grande de resúmenes y citas de literatura revisada la cual cuenta con revistas, libros y actas de congresos. Scopus proporciona una descripción general completa de los resultados de la investigación global en los campos de la ciencia, la tecnología, la medicina, las ciencias sociales, las artes y las humanidades, y presenta herramientas inteligentes para rastrear, analizar y visualizar la investigación [4].

- Institute of Electrical and Electronics Engineers (IEEE): Es una entidad global que se dedica a promover la creatividad, el desarrollo y la integración dando avance a la tecnología en beneficio de las personas y a la normalización. Es la mayor organización a nivel mundial sin ánimo de lucro de profesionales de las nuevas tecnologías. Esta base de datos ofrece conferencias, publicaciones y herramientas, como la Encuesta y guía de comunicación de IEEE, la Red IEEE, la Revista de comunicación de IEEE y muchas más [3].

El proceso llevado a cabo para la construcción de este documento tuvo en cuenta ecuaciones de búsqueda. En el momento de generar la ecuación se tuvieron en cuenta los siguientes pasos para cada uno de los temas claves:

- Temas principales relacionados con el operador AND.
- Años cubiertos: desde 2016 hasta el 2022.
- Campos de estudio: ciencias de la computación, ingeniería.

Los documentos tenidos en cuenta fueron, artículos, revistas y libros.

Las palabras clave correspondientes a los temas principales mencionados al inicio, son:

A. Minería de datos

A continuación, se realizara la definición de los conceptos usados en la búsqueda de las técnicas de minería de datos en las investigaciones realizadas en las bases de datos Scopus, ScienceDirect e IEEE XPLORE.

- Data mining: La minería de datos incluye un conjunto de técnicas que estudian automáticamente grandes bases de datos, con el objetivo de encontrar patrones, tendencias o reglas de repetición que expliquen el comportamiento de los datos recopilados. Estos patrones se pueden encontrar usando estadísticas o algoritmos de búsqueda cercana a la inteligencia artificial y redes neuronales. Por lo tanto, es de gran importancia ya que estos datos son el medio por el cual se pueden sacar conclusiones y, por lo tanto, se pueden transformar estos datos en información destacada, para que las entidades puedan realizar mejoras y resolver el problema. formas de ayudarlos a lograr sus objetivos. [5].
- Decision trees: Un árbol de decisión es un algoritmo que clasifica la información de una manera que produce un modelo en forma de árbol. Incluye un modelo de información esquemática que representa las diferentes alternativas, así como los posibles resultados de cada una. Este algoritmo se utiliza para clasificar, predecir y segmentar datos para obtener información analizable. [6].
- Education: Es la formación práctica y metódica que se le da a una persona en el camino del desarrollo y la madurez. Es un proceso por el cual la persona recibe las herramientas y conocimientos necesarios para ponerlos en práctica en la vida cotidiana. El aprendizaje de una persona comienza en la niñez, al ingresar a instituciones llamadas escuelas o colegios donde se ha recibido educación y capacitación previa se le inculca al niño identidades culturales, valores y valores para que en el futuro el niño se convierta en una buena persona. [7].

B. Indicadores de desempeño

A continuación, se realizará la definición de los conceptos usados en la búsqueda de los indicadores de

desempeño en las investigaciones realizadas en las bases de datos Scopus, ScienceDirect e IEEE XPLORE.

• Performance indicators: Permite identificar y valorar el estado en que se encuentra el estudiante con referencia a un conocimiento, valor, sentimiento, actitud, habilidad o destreza con lo que se convierte en un verdadero criterio de evaluación [8].

• Academic performance: Es la capacidad del alumno, que expresa lo que él ha aprendido a lo largo del proceso formativo [9].

• Indicators: Un indicador es una característica específica, observable y medible que puede ser usada para mostrar los cambios y progresos que está haciendo un programa hacia el logro de un resultado específico [10].

• Factors: Un factor es lo que contribuye a que se obtengan determinados resultados al caer sobre él la responsabilidad de la variación o de los cambios [11].

• School: Institución destinada a la enseñanza primaria o secundaria

• Análisys: identificar los componentes de un todo, separarlos y examinarlos para lograr acceder a sus principios más elementales [12].

• Family: grupo de personas que poseen un grado de parentesco y conviven como tal [13].

A continuación, las ecuaciones de búsqueda usadas en Scopus, ScienceDirect e IEEE:

C. Ecuación de busqueda scopus

Ahora veremos las diferentes ecuaciones de búsqueda implementadas en la investigación acerca de los indicadores de desempeño y las técnicas de minería de datos para el análisis de datos.

- TITLE-ABS-KEY ("data mining" AND "decision trees" AND "academic performance" AND "education") AND (LIMIT-TO (PUBYEAR , 2022) OR LIMIT-TO (PUBYEAR , 2021) OR LIMIT-TO (PUBYEAR , 2020) OR LIMIT-TO (PUBYEAR , 2019) OR LIMIT-TO (PUBYEAR , 2018) OR LIMIT-TO (PUBYEAR , 2017) OR LIMIT-TO (PUBYEAR , 2016))
- (TITLE-ABS-KEY ("academic performance") AND TITLE-ABS-KEY (indicator) OR TITLE-ABS-KEY (factor) AND TITLE-ABS-KEY (school)) AND (LIMIT-TO (PUBYEAR , 2022) OR LIMIT-TO (PUBYEAR , 2021) OR LIMIT-TO (PUBYEAR , 2020) OR LIMIT-TO (PUBYEAR , 2019) OR LIMIT-TO (PUBYEAR , 2018))

D. Ecuación de busqueda ScienceDirect

- Year(2016-2022) - Title, abstract, keywords("data mining" AND "decision trees" AND "academic performance" AND "education")
- Year(2018-2022)- Title, abstract, keywords (indicators, factors) - Title: Education; AND indicators AND "academic performance" AND factors AND school AND family AND analysis.

E. Ecuación de busqueda IEEE xplore

- ("All Metadata":data mining) AND ("All Metadata":decision trees) AND ("All Metadata":academic performance) AND ("All Metadata":education) Filters Applied: 2016 – 2022

- (All Metadata":"academic performance") AND ("All Metadata":indicator) OR ("All Metadata":factor) AND ("All Metadata":school) AND ("All Metadata":analysis) AND ("Publication Title: performance) Filters Applied: 2018-2022

En las anteriores ecuaciones de búsqueda se aplicaron los temas principales como palabras claves, de este modo, se tendrán en cuenta documentos de las temáticas definidas, además, en la búsqueda realizada los trabajos considerados como revistas, artículos y libros están limitados a los que fueron publicados a partir del 2016.

III. DESARROLLO

A. Búsqueda de indicadores de desempeño

El rendimiento académico es una medida de las capacidades de un estudiante, que se puede expresar con lo aprendido en un periodo de tiempo. Rodríguez, Ordoñez e Hidalgo [14] afirman que, el acceso a herramientas tecnológicas de aprendizaje como computadoras y conexión a internet, el acceso a herramientas tecnológicas de aprendizaje como

computadoras y conexión a internet, el mayor nivel educativo de los padres de familia, la condición de ser varón y el estudiar en una institución educativa oficial urbana aumenta la probabilidad de obtener un mejor rendimiento académico.

El nivel socioeconómico es el conjunto de factores, como lo son los ingresos, el patrimonio y las condiciones generales del entorno en las que vive un ser humano. el componente social es un factor crucial para mejorar la calidad de la educación y que variables como las actividades extracurriculares apoyen la formación integral de los estudiantes para impactar positivamente en el rendimiento [15]. Devi, K., Ratnoo, S., & Bajaj, A. [16] afirman, que las variables socioeconómicas de los estudiantes, como la casta, la residencia y la ocupación del padre, impactan su rendimiento académico en el sexto grado. Por otro lado, Muelle [17] afirma, La condición social del alumno y la composición social de su escuela destacan como los factores que afectan mayormente el bajo rendimiento, asociados a factores contextuales como la repetición, la lengua materna, la matrícula oportuna, la dimensión de la escuela, el ausentismo y el género.

En Colombia la educación escolar se puede dividir en colegios públicos y privados, es decir que los públicos los sostiene económicamente el estado, mientras que en la privada se sostiene gracias a los mismos alumnos, por medio de sus padres o algún acudiente legal. Por otra parte, también se puede dividir en escuelas rurales y urbanas.

En Sudáfrica el sistema educativo se caracteriza por escuelas acomodadas y otras que no lo están. Las escuelas acomodadas se desempeñan a un ritmo mucho mejor, esto conlleva a que se vuelva un factor en el rendimiento académico de los estudiantes. [18] Por otra parte, Sarmiento Espinel, J. A., Silva Arias, A. C., & van Gameren, E, [19] afirman que, la desigualdad de oportunidades educativas ha aumentado con el tiempo. La resiliencia, es la capacidad que tiene una persona para superar circunstancias traumáticas, esta se relaciona positivamente con el rendimiento académico y otros factores como la relación profesor alumno, la participación de los padres y los métodos de estudio [20].

El rendimiento académico aumenta. Según JHU (Johns Hopkins University), aspectos como la comprensión y la fluidez lectora mejoran cuando hay participación de los padres, aún más si los papás dedican tiempo para leer con sus hijos, ya que los alumnos saben que sus papás están al pendiente, tratan de mejorar por ellos, se sienten más motivados a aprender y mejorar sus calificaciones. [21] Además, La relación que tienen las variables sociofamiliares y no cognitivas sobre los estudiantes en cuanto a su rendimiento académico es un elemento muy importante para el éxito en la Educación Secundaria [22]. En rusia un estudio realizado por Kotomina, O. v., & Sazhina, A. [23] afirma que, la revisión radica en la consideración de tres formas en que la familia impacta en desempeño del estudiante: el estatus socioeconómico de la familia, el capital social de la familia, y la participación de los padres en el proceso educativo. Las dos primeras formas han sido ampliamente estudiadas en la investigación, mientras que la participación de los padres a menudo se considera como un factor significativo en el rendimiento escolar.

Algo que afecto al mundo en general fue la pandemia, por temas de contagios, a la gran mayoría de los colegios el anuncio de cierre de clases los tomó por sorpresa y sin previa preparación para desarrollar su programa de educación a distancia. Fueron muy pocos los colegios los que ya tenían un programa de aprendizaje remoto listo para ser implementado [24]. Estudios realizados en china por Wang, Y., Xia, M., Guo, W., Xu, F., & Zhao, Y [25] afirman que, los padres dedicaron más del doble del tiempo normal a apoyar el aprendizaje y el desarrollo de sus hijos durante el período de COVID-19. Se encontró que los factores de apoyo y motivación de los padres son la contribución más efectiva en el desarrollo de las emociones positivas de los niños y el logro del aprendizaje. Por otra parte, otro estudio en china Los resultados indicaron que el aprendizaje en línea no aumentó el rendimiento académico en el bachillerato rural y se observó una caída significativa del rendimiento en matemáticas e inglés [26].

A continuación, se muestra una tabla la cual resume cuales son los indicadores y cuales autores lo afirman:

Indicadores	Autores	Título del articulo
Acceso a tecnologías	Rodríguez, Ordoñez e Hidalgo.	Determinantes del rendimiento académico de la educación media en el Departamento de Nariño, Colombia
Conexión a internet	Rodríguez, Ordoñez e Hidalgo.	Determinantes del rendimiento académico de la educación media en el Departamento de Nariño, Colombia

Nivel educativo de los padres	Rodríguez, Ordoñez e Hidalgo.	Determinantes del rendimiento académico de la educación media en el Departamento de Nariño, Colombia
Genero	Rodríguez, Ordoñez e Hidalgo. Muelle.	Determinantes del rendimiento académico de la educación media en el Departamento de Nariño, Colombia. Factores socioeconómicos y contextuales asociados al bajo rendimiento académico de alumnos peruanos en PISA 2015.
Institución (Urbana o rural)	Rodríguez, Ordoñez e Hidalgo.	Determinantes del rendimiento académico de la educación media en el Departamento de Nariño, Colombia
Ocupación del padre	Devi K, Ratnoo S, & Bajaj A.	Impact of Socio-Economic Factors on Students' Academic Performance: A Case Study of Jawahar Navodaya Vidyalaya
La residencia	Devi K, Ratnoo S, & Bajaj A.	Impact of Socio-Economic Factors on Students' Academic Performance: A Case Study of Jawahar Navodaya Vidyalaya
Condición social	Velásquez, m y Crissien. Muelle	Multidimensional indicator to measure quality in education. Factores socioeconómicos y contextuales asociados al bajo rendimiento académico de alumnos peruanos en PISA 2015.
Composición social de la escuela	Muelle. Adebayo, K. A., Ntokozo, N., & Grace, N. Z.	Factores socioeconómicos y contextuales asociados al bajo rendimiento académico de alumnos peruanos en PISA 2015. Availability of Educational Resources and Student Academic Performances in South Africa.
Ausentismo	Muelle	Factores socioeconómicos y contextuales asociados al bajo rendimiento académico de alumnos peruanos en PISA 2015.
Participación de los padres	Sarmiento Espinel, J. A., Silva Arias, A. C., & van Gameren, E. Kotomina, O. v., & Sazhina	Evolution of the inequality of educational opportunities from secondary education to university. Influencia de los factores familiares sobre el desempeño de escolares y estudiantes:

		revisión de estudios extranjeros.
Desigualdad de oportunidades educativas	Sarmiento Espinel, J. A., Silva Arias, A. C., & van Gameren, E	Evolution of the inequality of educational opportunities from secondary education to university.
Estatus socioeconómico de la familia	Kotomina, O. v., & Sazhina	Influencia de los factores familiares sobre el desempeño de escolares y estudiantes: revisión de estudios extranjeros.
Capital social de la familia	Kotomina, O. v., & Sazhina	Influencia de los factores familiares sobre el desempeño de escolares y estudiantes: revisión de estudios extranjeros.
Pandemia	Wang, Y., Xia, M., Guo, W., Xu, F., & Zhao Zeng, L., & Luo, H.	Academic performance under COVID-19: The role of online learning readiness and emotional competence. Online Academic Performance during the COVID-19: Evidence from a Rural High School in Western China
Condiciones sociofamiliares	Rodríguez-Rodríguez, D., & Guzmán, R	Academic performance of secondary education students in socio-familial risk contexts

Figura 1. Elaboración propia

B. Técnicas de minería de datos para la búsqueda de patrones.

Ahora bien, frente a la medición del desempeño de las diferentes instituciones educativas, se encuentran varios estudios en el cual se relacionan los diferentes indicadores y la manera de analizar dichos datos, según la investigación de Nuankaew y Sararat [27] realizada a 1859 estudiantes de la escuela Manchasuksa en el distrito de La Mancha Khiri, provincia de Khon Kaen, Tailandia, durante el año académico 2015-2020, en donde las herramientas de investigación están separadas en 2 secciones. La primera sección es un paso básico de análisis estadístico, este se compone de análisis de frecuencia, análisis de porcentaje, análisis de media y análisis de desviación estándar. Otra sección es la fase de análisis de minería de datos, que consiste en la técnica de discretización, la técnica de clasificación XGBoost (árbol de decisión, árboles potenciados por gradiente y random forest), análisis de rendimiento de matriz de confusión y análisis de rendimiento de validación cruzada.

Además, uno de los temas de enfoque en la investigación de Vasiliki Matzavela y Efthimios Alepis [28] trata de la Minería de Datos Educativos (EDM) la cual es una aplicación de Técnicas de Minería de Conocimiento a partir de datos educativos, y su objeto es analizar datos, con el fin de resolver problemas de investigación en el campo de la Educación. Sus datos provienen de diferentes fuentes, como bases de datos de sistemas educativos, sistemas de Internet.

Por otro lado, se evidencia un tema de gran importancia, los árboles de decisión pues son una de las técnicas usadas como modelo predictivo del Rendimiento Académico de los Estudiantes en entornos Inteligentes de M-Learning, en el área del aprendizaje automático y la ciencia de datos, es uno de los métodos más populares dentro de las técnicas de clasificación por ser fácil de entender e interpretar por medio de gráficas, así como también el manejo de datos numéricos y categóricos. Los sistemas M-Learning se consolidan recientemente como uno de los métodos de mayor interés para una educación más efectiva y un aprendizaje adaptativo proporcional a las habilidades de aprendizaje de cada estudiante.

Un clasificador de árbol de decisión es uno de los métodos de aprendizaje supervisado más utilizados para la exploración de datos, aproximando una función por regiones constantes a trozos, y no necesita información

previa de la distribución de los datos esto según Mitra S y Acharya T. [29]. En el estudio de Witten et al. [30] los modelos de árboles de decisión son comúnmente utilizados en la minería de datos para examinar los datos e inducir el árbol y sus reglas que serán utilizadas para hacer las predicciones. Según Rud [31] el verdadero propósito de los árboles de decisión es clasificar los datos en grupos distintos o ramas que generen la separación más fuerte en los valores de la variable dependiente, siendo superiores en identificar segmentos con un comportamiento deseado como la respuesta o la activación, proporcionando así una solución fácilmente interpretable.

Teniendo en cuenta lo anterior, la minería de datos cuenta con diferentes sistemas que contribuyen al descubrimiento de los factores principales para un mejor rendimiento académico, en las últimas décadas se han generado avances significativos de las nuevas tecnologías en el ámbito educativo entre las cuales la Minería de Datos Educativo- EDM, juega un papel indispensable para la búsqueda del mejoramiento pedagógico, permitiendo a los investigadores por medio de bases de datos agrupar variables que ayuden a identificar los factores que influyen en el rendimiento académico de los estudiantes, algunos de estos como la información demográfica de los estudiantes, la disposición o voluntad de aprendizaje y la interacción familiar, usando EDM como regresión lineal, regresión árbol, random forest y red neuronal.(Yucheng-jin y Xiaomeng Yang) [32].

Con los conceptos tratados por Han, J. and Kamber, M. [33] en los últimos años, el campo de la minería de datos se vuelve muy importante para diferentes industrias, co-corporaciones y empresas, etc. debido a su capacidad para utilizar una gran cantidad de datos que antes no tenían uso y respecto de los cuales se pueden realizar análisis, predicción de tendencias y patrones.

Para determinar los indicadores de desempeño se tendrán presentes las técnicas de minería de datos implementado la herramienta WEKA como lo hicieron los investigadores Sadiq Hussain y Neama Abdulaziz [34] para la selección de los atributos o factores y de este modo permitiendo clasificar la información donde los autores determinaron mediante los resultados que el algoritmo de clasificación random forest destaca en precisión. Para confirmar que tecnología es más óptima en el estudio del desempeño académico de los estudiantes se pondrán a prueba varios métodos de selección de características para así extraer los indicadores fundamentales tomando como ejemplo el artículo de Talha Mahboob, Mubbashar Mushtaq y Kamran Shaukat [35] donde en este estudio proponen un método novedoso para la medición del desempeño de las instituciones educativas haciendo uso de varios modelos de aprendizaje automático como los árboles de decisión, bosque de rotación, bosque aleatorio, entre otros, ya que esto permite adaptar los diferentes factores a tratar según las necesidades del estudio.

Según en el artículo de Umair Shafique y Haseeb Qaiser [36] se puede centrar la investigación haciendo uso de tres modelos de procesos de minería de datos que son muy populares y que principalmente son empleados por expertos e investigadores en minería de datos los cuales son Knowledge Discovery Databases (KDD), CRISP-DM y SEMMA. De acuerdo con ello, en la investigación de Samsudin [37] propone hacer uso de una máquina de vectores de soporte el cual es un algoritmo de aprendizaje que sirve para determinar patrones de desempeño académico durante la pandemia de COVID-19.

En el siguiente artículo se explica el proceso de aplicación de la metodología CRISP-DM para detectar factores relacionados con el rendimiento académico de estudiantes colombianos quienes presentaron las pruebas saber 11. Principalmente se construye un repositorio o base de datos con la información socioeconómica y académica disponible por el ICFES, para luego realizar una limpieza de este, se realizó un modelo de clasificación basado en arboles de decisión para predecir los patrones asociados con el buen o bajo rendimiento académico Timarán-Pereira, R., Hidalgo-Troya, A., & Caicedo-Zambrano, J. [38].

En la investigación que realizaron Pandey y Pal [39] de minería de datos utilizando clasificación de Naïve Bayes para analizar, clasificar y predecir estudiantes de alto y bajo rendimiento donde La clasificación de Naïve Bayes se usa como técnica probabilística simple, que asume que todos los atributos dados en un conjunto de datos son independientes entre sí, de ahí el nombre "Naïve". Realizaron esta investigación con una muestra de datos de estudiantes matriculados en un Diploma de Postgrado en Aplicaciones Informáticas Aplicaciones informáticas (PGDCA) en la Universidad Dr. R. M. L. Awadh, Faizabad, India.

De acuerdo con el artículo de Ajibade y Bahiah Binti Ahmad [40] para construir un modelo predictivo, se utilizan varias técnicas de la minería de datos, que son la clasificación, la regresión y la agrupación. Para la investigación de los dos autores mencionados eligieron parámetros como las notas internas, las notas de las sesiones y la puntuación de admisión además hicieron uso del algoritmo de aprendizaje SVM (máquina de vectores de soporte) desarrollado por Subaira [41] que sirve para manejar los desafíos del reconocimiento de patrones y la predicción como también para el análisis y mapeo de funciones, asimismo se usó el algoritmo random forest igual que en el trabajo investigativo de Sabzevari M. [42]. Ahora bien, este método trata de una

colección de algoritmos de árboles de decisión que no están correlacionados. Random Forest genera una gran cantidad de árboles de decisión a partir de subconjuntos del conjunto de datos a estudiar donde cada subconjunto proporciona un árbol de decisión. Ahora, cada modelo de árbol de decisión clasifica una instancia en una clase y así la clase más votada se tomada como instancia, es decir, de reiterar o ser insistente en una orden dada, todo esto es conforme a lo dicho por Amrieh E. A. [43]. En la investigación de Moisa V. [44] utilizaron algunas técnicas de conjuntos como Bagging, Adaboosting y random forest para predecir el rendimiento académico de los estudiantes con mayor precisión.

La finalidad del trabajo de Contreras Leonardo y Fuentes Héctor [45] es predecir el rendimiento académico de estudiantes mediante técnicas de aprendizaje automático donde se analizan 324 variables con métodos de selección de características, con el objetivo de determinar las variables más destacadas. El modelo de predicción del rendimiento académico universitario es estudiado por medio de algoritmos supervisados como (KNN, SVC, Naive Bayes y árbol de decisión), los cuales son optimizados mediante lenguaje Python. Además, son implementados algoritmos de ensamble que permiten mejorar la exactitud de los clasificadores previos, también se implementan métodos Bagging (CART, Random Forest), métodos Boosting (AdaBoost, GBM, XGBoost).

A continuación, veremos una tabla resumiendo las técnicas de minerías de datos con sus autores:

Minería de datos	Arbol de decisión	27.Mitra S, Acharya T. Data Mining 2013.
		28.Witten, I. H., Frank, E., & Hall, M. A. (2011).
		36.Timarán-Pereira, R., Hidalgo-Troya, A., & Caicedo-Zambrano, J. (2020).
		29. Rud, O. P. (2012). Data Mining Cookbook
		33. Alam, T. M., Mushtaq, M., & Shaukat, K. (2021).
		41.Amrieh E. A. 2017. Database Theory and Application 9 119-136.
		25.Nuankaew, P., & Sararat, W. (2022). Student Performance Prediction Model for Predicting Academic
		43. Contreras Bravo, L. E., Fuentes López, H. J., & Rivas Trujilllo, E. (2022). Análisis del rendimiento académico
	Random forest	42.Moisa V. 2018 Journal of Mobile Embedded and Distributed Systems 5 70-77
		41.Amrieh E. A. 2017. Database Theory and Application 9 119-136.
		39.Subaira A. 2016 IEEE 8th Int. Conf. on Intelligent System and Control (ISCO) 978 274-280
		[33]. Alam, T. M., Mushtaq, M., & Shaukat, K. (2021).
		[32]. Hussain, S. (2018, 1 febrero). Educational Data Mining and Analysis of Students
		[40.] Sabzevari M. 2018 Cornell Uni. arXiv preprint arXiv:1802.07877
		[25]. Nuankaew, P., & Sararat, W. (2022). Student Performance Prediction Model for Predicting Academic
	Naive Bayes	[43]. Contreras Bravo, L. E., Fuentes López, H. J., & Rivas Trujilllo, E. (2022). Análisis del rendimiento académico
		[37]. Pandey, U.K. and Pal, S., 2011. Data Mining: A prediction of performer or underperformer using classification.
	vectores de soporte	[35].Samsudin, N. A. M. (2021). Modeling Student's Academic Performance.
		[39].Subaira A. 2016 IEEE 8th Int. Conf. on Intelligent System and Control (ISCO) 978 274-280
		[38].Ajibade, S. S. M., & Bahiah Binti Ahmad, N.

		(2019). Educational Data Mining: Enhancement of Student Performance.
	Metodos Bagging	[42].Moisa V. 2018 Journal of Mobile Embedded and Distributed Systems 5 70-77
Minería de datos educativos	Random forest	[26].Matzavela, V., & Alepis, E. (2021, 5 octubre).
		[30]. Jin, Y., & Yang, X. (2021). Educational Data Mining
Metodologias	WEKA	[32].Hussain, S. (2018, 1 febrero). Educational Data Mining and Analysis of Students
	KDD, CRISP-DM y SEMMA	[34].Shafique, U., & Qaiser, H. (2014). A comparative study of data mining process models.
		31. Han, J. and Kamber, M. "Data Mining: Concepts and Techniques.
		[36]. Timarán-Pereira, R., Hidalgo-Troya, A., & Caicedo-Zambrano, J. (2020).

Figura 2. Elaboración propia.

IV. RESULTADOS

En la aplicación de la metodología y la selección de la bibliografía permite realizar un análisis de coocurrencia de las palabras claves usadas para la representación de cada uno de los artículos resultantes. El estudio arrojó un mapa bibliométrico de coocurrencia construido con el software VOSviewer, donde se utiliza la técnica de mapeo VOS, la cual representa la visualización de similitudes basadas en la distancia.

Para la elaboración de los siguientes mapas se usó como recurso la información de los artículos de investigación, que según la metodología de búsqueda y filtrado en los índices de Scopus, ScienceDirect e IEEE, usando la información de 130 artículos donde se encuentran las temáticas a tratar, la técnica de graficación de VOS crea una matriz de similitud basada en una matriz de coocurrencia con las palabras claves.

El resultado de los mapas de concurrencia se representa de forma gráfica, donde cada palabra clave es clasificada como un ítem representado por su etiqueta y por un círculo. El tamaño de la etiqueta y del círculo de un ítem son determinados por su valor; cuanto mayor sea el valor de un ítem, mayor será la etiqueta y el círculo del artículo. La figura 1 representa el mapa de concurrencia de palabras claves.

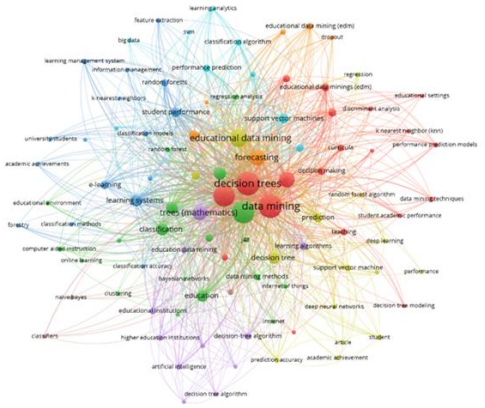

Figura 1. Mapa de concurrencia de palabras clave Scopus

El mapa de la Figura 1 representa una clasificación de las diversas temáticas de investigaciones actuales en torno a la minería de datos, al rendimiento académico, la educación y la interrelación entre dichas temáticas. El análisis de concurrencia deja ver la existencia de temas comunes agrupados en clústeres identificados por diferentes colores donde se puede observar una importante clasificación de investigaciones alrededor de temas como la minería de datos educativa, los algoritmos de clasificación como arboles de decisión, random forest, entre otros. La figura 2 muestra la concurrencia de las palabras claves de la búsqueda realizada en IEEE XPLORE

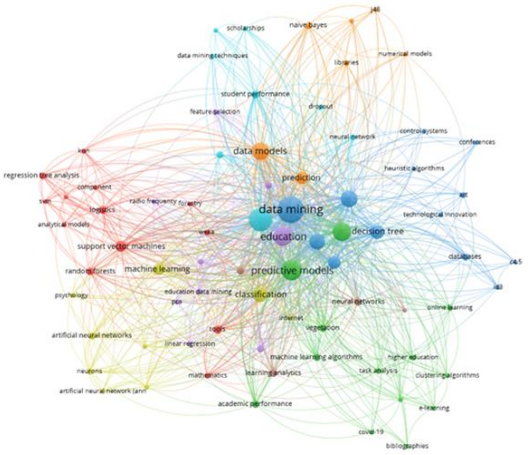

Figura 2. Mapa concurrencia de las palabras clave IEEE XPLORE

Con el mapa de la figura 2 se puede empezar a evidenciar que existe una mayor fuerza en los temas de data mining, modelos predictivos, arboles de decisión entre otros, además, se encuentra la relación de estos con

la educación. Ahora la figura 3 muestra la concurrencia de las palabras claves de la búsqueda realizada en ScienceDirect.

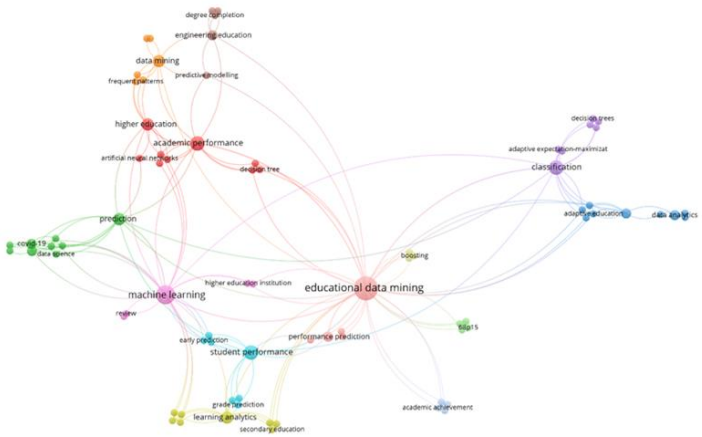

Figura 3. Mapa concurrencia de las palabras clave ScienceDirect.

El mapa de la figura 3 muestra fuerza en los temas de minería de datos educativos y en arboles de decisión, al haber realizado estos mapas con información de diferentes bases de datos de artículos y revistas científicas se encuentra una relación entre las metodologías usadas para el análisis de datos en la educación.

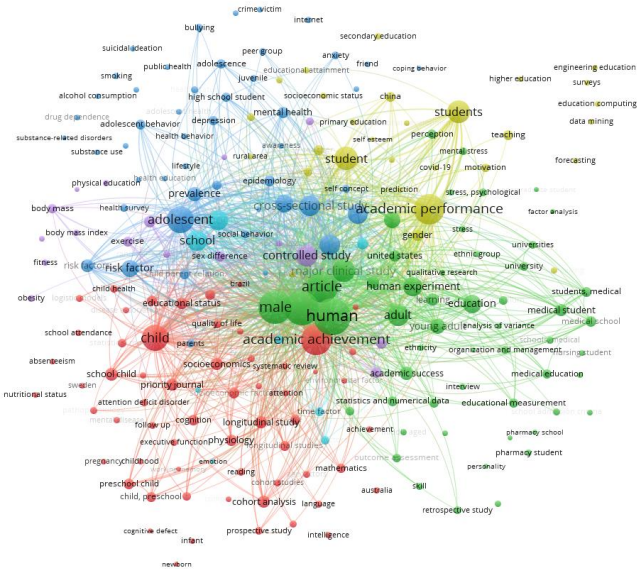

Figura 4. Mapa concurrencia de las palabras claves de Scopus

En el mapa de concurrencia de la figura 4 se elaboro por la base de datos scopus, donde se hizo una busqueda por medio de las palabras claves de los indicadores de desempeño, donde encontramos un total de mas de 1000 resultados donde se exportaron todos los articulos que nos servian para realizar este mapa por medio del software Vosviewer.

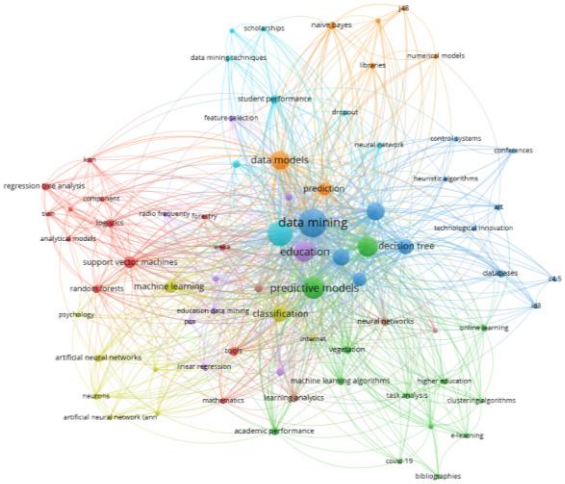

Figura 5. Mapa de concurrencia de las palabras claves de IEEE XPLORE

En el mapa de concurrencia de la figura 5, se realizo por medio de la base de datos IEEE XPLORE por medio de las palabras claves de los indicadores de desempeño, se encontraron mas de 50 articulos donde se exportaron para realizar su respectivo mapa de concurrencia en el software VosViewer.

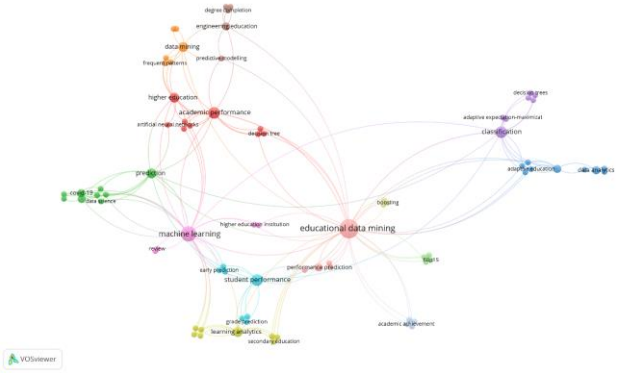

Figura 6. Mapa de concurrencia de las palabras claves de ScienceDirect.

En el mapa de concurrencia de la figura 6, se realizo por medio de las palabras claves sobre los indicadores de desempeño, se realizo una busqueda en la cual obtuvimos mas de 20 articulos, donde se realizo su respectiva exportacion para elaborar el mapa de concurrencia en el software VosViewer.

V. CONCLUSIONES

- Dentro del análisis de datos en la búsqueda de indicadores de desempeño en la educación básica y media de Colombia, se pudo evidenciar

que son varios los factores determinantes del alto o bajo desempeño educativo, dentro de estos se enmarcan el Familiar, Social y Económico, siendo el Familiar un indicador ampliamente determinante pues es el común denominador de la población estudiantil ya que se encuentra en edad joven, es decir, aún dependen de la ayuda económica de su núcleo familiar, partiendo de esta suposición y en concordancia con la información presente en desarrollo de este articulo se puede asumir que la Familia es el indicador principal del cual se desglosa el factor social y económico, de modo que una familia que se encuentra situada en una región poblacional en la cual cuentan con todos los recursos sociales y económicos para desarrollarse íntegramente reúne las condiciones para que el estudiante tenga un mejor rendimiento académico.

- Gracias a toda la investigación realizada en este artículo, podemos decir con certeza que las técnicas más usadas por los autores y que más ayudan al análisis de datos son los árboles de decisión y random forest.

REFERENCIAS

[1] Lima, "MINISTERIO DE ECONOMÍA Y FINANZAS DIRECCIÓN GENERAL DEL PRESUPUESTO PÚBLICO Instructivo para la Formulación de Indicadores de Desempeño," 2010.

[2] "Conceptos de minería de datos | Microsoft Docs." https://docs.microsoft.com/es-es/analysis-services/data-mining/data-mining-concepts?view=asallproducts-allversions (accessed May 26, 2022).

[3] E. Serna, Desarrollo e Innovación en Ingeniería. IAI, 2019. Accedido el 10 de agosto de 2022. [En línea]. Disponible: https://doi.org/10.5281/zenodo.3387679

[4]. Investigación [en línea]. (sin fecha). Publiciencia. [Consultado el 8 de julio de 2022]. Disponible en: https://publiciencia.com/investigación

[5]. Bello. "¿Qué es el minado de Datos o Data Mininig? Técnicas y pasos a seguir". Thinking for Innovation. https://www.iebschool.com/blog/data-mining-mineria-datos-big-data/ (accedido el 10 de agosto de 2022).

[6] "Arboles de decisiones en la minería de datos - Conecta Software." https://conectasoftware.com/analytics/arboles-de-decisiones-en-la-mineria-de-datos/ (accessed May 09, 2022).

[7] A. Sánchez. "¿Qué es la Educación?» Su Definición y Significado 2021". Concepto de - Definición de. https://conceptodefinicion.de/educacion/ (accedido el 10 de agosto de 2022).

[8] "Algunos conceptos importantes sobre educación en Colombia - ESE." https://eservicioseducativos.com/editorial/conceptos-importantes-sobre-la-educacion-en-colombia/ (accessed May 26, 2022).

[9] "Definición de rendimiento académico - Qué es, Significado y Concepto." https://definicion.de/rendimiento-academico/ (accessed May 26, 2022).

[10]"Indicadores."https://www.endvawnow.org/es/articles/336-indicadores.html (accessed May 26, 2022).

[11] "Definición de Factores » Concepto en Definición ABC." https://www.definicionabc.com/general/factores.php (accessed May 26, 2022).

[12] "Definición de análisis - Qué es, Significado y Concepto." https://definicion.de/analisis/ (accessed May 26, 2022).

[13] "Significado de Familia (Qué es, Concepto y Definición) - Significados." https://www.significados.com/familia/ (accessed May 26, 2022).

[14] D. D. R. Rosero, R. E. O. Ortega, and M. E. H. Villota, "Academic performance determinants of high school students in the Department of Nariño, Colombia," Lecturas de Economia, no. 94, pp. 87–126, Jan. 2021, doi: 10.17533/UDEA.LE.N94A341834.

[15] J. V. Rodríguez, D. N. Rodado, T. Crissien Borrero, and A. Parody, "Multidimensional indicator to measure quality in education," International Journal of Educational Development, vol. 89, Mar. 2022, doi: 10.1016/j.ijedudev.2021.102541.

[16] K. Devi, S. Ratnoo, and A. Bajaj, "Impact of Socio-Economic Factors on Students' Academic Performance: A Case Study of Jawahar Navodaya Vidyalaya," Lecture Notes in Networks and Systems, vol. 419 LNNS, pp. 774–785, 2022, doi: 10.1007/978-3-030-96299-9_73.

[17] L. Muelle, "Socioeconomic and contextual factors associated with low academic performance of peruvian students in PISA 2015," Apuntes, vol. 47, no. 86, pp. 111–146, 2020, doi: 10.21678/APUNTES.86.943.

[18] K. A. Adebayo, N. Ntokozo, and N. Z. Grace, "Availability of Educational Resources and Student Academic Performances in South Africa," Universal Journal of Educational Research, vol. 8, no. 8, pp. 3768–3781, Aug. 2020, doi: 10.13189/ujer.2020.080858.

[19] J. A. Sarmiento Espinel, A. C. Silva Arias, and E. van Gameren, "Evolution of the inequality of educational opportunities from secondary education to university," International Journal of Educational

Development, vol. 66, pp. 193–202, Apr. 2019, doi: 10.1016/j.ijedudev.2018.09.006.
[20] G. Bester and N. Kuyper, "The Influence of Additional Educational Support on Poverty-Stricken Adolescents' Resilience and Academic Performance," https://doi-org.ucundinamarca.basesdedatosezproxy.com/10.1080/18146627.2019.1689149, vol. 17, no. 3, pp. 158–174, May 2020, doi: 10.1080/18146627.2019.1689149.
[21] "La participación de los padres en la enseñanza — Observatorio | Instituto para el Futuro de la Educación." https://observatorio.tec.mx/edu-news/la-importancia-de-la-participacion-de-los-padres-en-la-educacion (accessed Aug. 09, 2022).
[22] D. Rodríguez-Rodríguez and R. Guzmán, "Academic performance of secondary education students in socio-familial risk contexts," Suma Psicologica, vol. 28, no. 2, pp. 104–111, 2021, doi: 10.14349/sumapsi.2021.v28.n2.5.
[23] O. v. Kotomina and A. I. Sazhina, "The Influence of Family Factors on the Academic Performance of Schoolchildren and University Students: Review of Foreign Studies," Education and Self Development, vol. 16, no. 4, pp. 74–92, 2021, doi: 10.26907/esd.16.4.07.
[24] "Así ha afectado el Covid-19 la educación en Colombia - Forbes Colombia." https://forbes.co/2020/04/30/actualidad/asi-ha-afectado-el-covid-19-la-educacion-en-colombia/ (accessed Aug. 09, 2022).
[25] Y. Wang, M. Xia, W. Guo, F. Xu, and Y. Zhao, "Academic performance under COVID-19: The role of online learning readiness and emotional competence," Current Psychology, 2022, doi: 10.1007/s12144-022-02699-7.
[26] L. Zeng and H. Luo, "Online Academic Performance during the COVID-19: Evidence from a Rural High School in Western China," Proceedings - 2021 10th International Conference of Educational Innovation through Technology, EITT 2021, pp. 112–116, 2021, doi: 10.1109/EITT53287.2021.00030.
[27]. Nuankaew, P., & Sararat, W. (2022). Student Performance Prediction Model for Predicting Academic Achievement of High School Students. European Journal of Educational Research, 11(2), 949–963. https://doi.org/10.12973/eu-jer.11.2.949
[28]. Matzavela, V., & Alepis, E. (2021, 5 octubre). E-Biblioteca Ucundinamarca. ScienceDirect. Recuperado 17 de abril de 2022,de https://login.ucundinamarca.basesdedatosezproxy.com/login?qurl=https://www.sciencedirect.com%2fscience%2farticle%2fpii%2fS2666920X21000291%3fvia%253Dihub
[29]. Mitra S, Acharya T. Data Mining. Multimedia, Soft Computing, and Bioinformatics. John Wiley & Sons, Inc., Hoboken, New Jersey; 2013.
[30]. Witten, I. H., Frank, E., & Hall, M. A. (2011). Data Mining: Practical Machine Learning Tools and Techniques (3rd ed.). Morgan Kaufmann Publishers.
[31]. Rud, O. P. (2012). Data Mining Cookbook: Modeling Data for Marketing, Risk, and Customer Relationship Management (1.a ed.). John Wiley & Sons Inc.
[32.] Jin, Y., & Yang, X. (2021). Educational Data Mining: Discovering Principal Factors for Better Academic Performance. 2021 the 3rd International Conference on Big Data Engineering and Technology (BDET). https://doi.org/10.1145/3474944.3474945
[33]. Han, J. and Kamber, M. "Data Mining: Concepts and Techniques. Second Edition", Morgan Kaufmann Publishers, San Francisco, 2010.
[34]. Hussain, S. (2018, 1 febrero). Educational Data Mining and Analysis of Students' Academic Performance Using WEKA | Hussain | Indonesian Journal of Electrical Engineering and Computer Science. Indonesian Journal. Recuperado 17 de abril de 2022, de http://ijeecs.iaescore.com/index.php/IJEECS/article/view/9746
[35]. Alam, T. M., Mushtaq, M., & Shaukat, K. (2021). A Novel Method for Performance Measurement of Public Educational Institutions Using Machine Learning Models. Applied Sciences, 11(19), 9296. https://doi.org/10.3390/app11199296
[36]. Shafique, U., & Qaiser, H. (2014). A comparative study of data mining process models (KDD, CRISP-DM and SEMMA). International Journal of Innovation and Scientific Research, 12(1), 217-222.
[37]. Samsudin, N. A. M. (2021). Modeling Student's Academic Performance During Covid- 19 Based on Classification in Support Vector Machine. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(5), 1798-1804.
[38]. Timarán-Pereira, R., Hidalgo-Troya, A., & Caicedo-Zambrano, J. (2020). Academic performance patterns of middle school students in the knowledge natural science test 11 with decision trees. RISTI - Revista De Sistemas e Tecnologias De Informacao, 2020(E32), 190-201. Retrieved from www.scopus.com
[39]. Pandey, U.K. and Pal, S., 2011. Data Mining: A prediction of performer or underperformer using classification. (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 2 (2), 2011, 686- 690.
[40]. Ajibade, S. S. M., & Bahiah Binti Ahmad, N. (2019). Educational Data Mining: Enhancement of Student Performance model using Ensemble Methods.

IOP Conference Series: Materials Science and Engineering, 551(1), 012061. https://doi.org/10.1088/1757-899x/551/1/012061
[41]. Subaira A. 2016 IEEE 8th Int. Conf. on Intelligent System and Control (ISCO) 978 274-280
[42]. Sabzevari M. 2018 Cornell Uni. arXiv preprint arXiv:1802.07877
[43]. Amrieh E. A. 2017 Int'l. Journal of Database Theory and Application 9 119-136.
[44]. Moisa V. 2018 Journal of Mobile Embedded and Distributed Systems 5 70-77
[45]. Contreras Bravo, L. E., Fuentes López, H. J., & Rivas Trujilllo, E. (2022). Análisis del rendimiento académico mediante técnicas de aprendizaje automático con métodos de ensamble. Revista Boletín Redipe, 10(13), 171–190. https://doi.org/10.36260/rbr.v10i13.1737