Apuntes Universitarios, 2023: *13* (2), abril-junio
ISSN: 2304-0335 DOI: https://doi.org/10.17162/au.v13i2.1438

Predicción de repitencias en estudiantes a nivel escolar usando Machine Learning: una revisión sistemática

Predicting student grade repetition at the school level using Machine Learning: a systematic review

Javier Gamboa-Cruzado[1a], Cinthya Y. Alvarez-Cuellar[2], Shirley Martinez-Medina[3], Josue Edison Turpo Chaparro[4], Aníbal Sifuentes Damián[5], María Rodríguez Kong[6]

Universidad Nacional Mayor de San Marcos, Lima, Perú[1]
Universidad Autónoma del Perú, Lima, Perú[2,3]
Universidad Peruana Unión, Lima, Perú[4]
Universidad Nacional José Faustino Sánchez Carrión, Lima, Perú[5]
Universidad Señor de Sipán S.A.C., Chiclayo, Perú[6]

ORCID ID: https://orcid.org/0000-0002-0461-4152[1]
ORCID ID: https://orcid.org/0000-0002-9567-2840[2]
ORCID ID: https://orcid.org/0000-0001-7338-9860[3]
ORCID ID: https://orcid.org/0000-0002-1066-6389[4]
ORCID ID: https://orcid.org/ 0000-0002-8211-9771[5]
ORCID ID: https://orcid.org/0000-0002-9645-2508[6]

Recibido: 09 de noviembre de 2022 **Aceptado:** 05 de febrero de 2023

Resumen

El objetivo principal de la investigación es determinar el estado del arte de la investigación acerca de la Predicción de repitencias en estudiantes a nivel escolar usando Machine Learning. Los resultados obtenidos se han centrado en estudios relacionados a los algoritmos y herramientas de Machine Learning más eficientes para la predicción de repitencia estudiantil. Esto se llevó a cabo mediante una revisión sistemática de la literatura (RSL) en base a Machine Learning, para la predicción de estudiantes con repitencia escolar entre los años 2017-2021. La estrategia de búsqueda identificó 47,490 artículos de bibliotecas digitales como ACM Digital Library, ERIC, Google Scholar, IEEE Xplore, Microsoft Academic, Science Direct y Taylor & Francis Online; de las cuales 90 fueron identificados y seleccionados como adecuados para la revisión. En cuanto a las conclusiones, estas presentan respuestas sobre las categorías de variables más aplicadas en la predicción de repitencia escolar en estudiantes, las métricas utilizadas para evaluar los resultados de la predicción de repitencia escolar, autores con mayor productividad en la predicción de repitencia escolar, y los artículos más citados cuyas discusiones y conclusiones se caracterizan por

[a]Correspondencia al autor: jgamboa65@hotmail.com

su objetividad y polaridad en las investigaciones sobre la predicción de estudiantes con repitencia escolar usando Machine Learning.

Palabras clave: Métodos de predicción, repitencia escolar, Machine Learning, deserción escolar, educación.

Abstract
The main objective of the research is to determine the state of the art of research on the Prediction of repetition in students at school level using machine Learning. It was carried out through a Systematic Literature Review (SLR) based on Machine Learning for the prediction of students with school repetition between the years 2017-2021. The search strategy identified 47490 papers from digital libraries such as ACM Digital Library, ERIC, Google Scholar, IEEE Xplore, Microsoft Academic, Science Direct and Taylor & Francis Online of which 90 were identified and selected as suitable for the review. The results obtained have focused on studies related to the most efficient Machine Learning algorithms and tools for the prediction of student repetition. As for the conclusions, these present answers about the categories of variables most applied in the prediction of school repetition in students, the metrics used to evaluate the results of the prediction of school repetition, the authors with the highest productivity in the prediction of school repetition, and the most cited articles whose discussions and conclusions are characterized by their objectivity and polarity in the research on the prediction of students with school repetition using Machine Learning

Keywords: Prediction methods, school repetition, Machine learning, school dropout, education.

Introducción

La identificación temprana de la probabilidad de repetir un curso o grado escolar usando Machine Learning reduce las tasas de repitencia y/o deserción estudiantil, además de identificar y proporcionar los factores que inducen a los escolares a ser parte de estas estadísticas (Helal et al., 2019). Al obtener la identificación de los estudiantes en potencial riesgo a partir de estas causas, se puede plantear soluciones a los diversos tipos de problemáticas en las que se encuentran (Tamada et al., 2019). Según Imran et al. (2019), en promedio el 25% de estudiantes no logran terminar sus estudios escolares a tiempo, por lo que sufren las consecuencias desde no poder obtener un trabajo rentable, hasta llegar a ser una carga para la sociedad. Las causas más comunes de repitencia escolar son por problemas intrafamiliares o económicos, pero comúnmente no se tiene conocimiento de ellas debido a que no existe un interés entre escuela-alumno y comunicación entre docente-alumno. Por lo tanto, la aplicación de Machine Learning hace evidencia de estos factores para la predicción de Repitencia Escolar. Este artículo se focaliza en identificar factores que indiquen la probabilidad de un estudiante de repetir un grado o curso estudiantil.

Predicción de repitencias en estudiantes a nivel escolar y Machine Learning

En esta investigación se realizó la búsqueda de artículos de revisión sistemática de la literatura con el fin de comparar los resultados obtenidos. los estudios identificados muestran preguntas descriptivas muy sencillas a diferencia de las analíticas presentadas en este artículo, también se enfocaron principalmente en las técnicas que usaron para realizar el estudio. Del mismo modo, hay que considerar que estos artículos oscilan entre el 2017 al 2021.

En primer lugar, Hellas et al. (2018) publicó una revisión sistemática de literatura (RSL) enfocada a predecir el desempeño de los estudiantes utilizando métodos de Machine Learning como Árboles de decisión, Regresión, Clasificación, entre otros; llegando a comparar un total de 357 artículos, concluyendo que una comparación de precisión llevaría al lector a obtener resultados errónea sobre el rendimiento. Luego, Abu Saa et al. (2019), realizaron una RSL para predecir el desempeño de los estudiantes en la educación superior mediante el uso de técnicas de minería de datos predictivos, su objetivo principal es identificar los factores que implican el desempeño del estudiante. El tercer artículo revisado fue de Agrusti et al. (2019) quienes realizaron una revisión bibliográfica sistemática sobre técnicas de minería de datos para predecir la deserción de estudiantes, de los cuales identificaron 73 artículos en función a sus criterios de exclusión, para realizar el análisis de la investigación utilizaron 6 técnicas: Árbol de decisión, K-NN, SVM, Clasificación bayesiana, Redes neuronales y Regresión logística.

En cuarto lugar, se analizó la RSL de Alban y Mauricio (2019) en la cual estudian como predecir la deserción de estudiantes, encontrando factores que relacionaban a la deserción de estudiantes con: Entorno del campus, Tipos de escuelas secundarias, Participación institucional, Infraestructura universitaria; siendo para ellos la conclusión que la deserción de estudiantes afecta a todos tanto como a las universidades porque reducen las matrículas, reduce el ingreso y también a la sociedad y a la familia y eso ayuda para poder diseñar estrategias y así abordar con el problema. Como punto clave se puede decir que esta investigación es una de las primeras revisiones sistemáticas de literatura que realiza la deserción universitaria a través de la minería de datos, con estudios de 2006-2018. Los autores Moreno-Marcos et al. (2019), analizan el estado del arte sobre la predicción de estudiantes MOOC mediante una revisión sistemática de la literatura. Sus principales objetivos son: identificar las características de MOOC para la predicción, describir los resultados de la predicción, clasificar las características de la predicción, determinar las técnicas

utilizadas para predecir las variables e identificar las métricas para evaluar los modelos predictivos. Estos autores evaluaron las métricas a usar y estas fueron: AUC, Accuracy, F- score, Precision, Recall, RMSE, Kappa y Others. Sus técnicas aplicadas son: Regression, SVM, Decision Tree, Random Forest, Naive Bayes, Gradient Boosting, Neural Network y Others.

La metodología aplicada permitió la clara evolución en el número de publicaciones en los últimos años. En el futuro, con la expansión de la educación en línea y la analítica del aprendizaje se esperan más trabajos. Aquí resalta Orellana et al. (2020) que centraron su estudio de investigación sobre el fenómeno del abandono en educación superior en la modalidad virtual, dando como resultado a factores importantes que son motivos para un abandono estudiantil como: el rendimiento académico, la edad, el género, las circunstancias personales y las técnicas para el aprendizaje virtual y autodirigido.

En esta investigación se utilizó el programa gratuito Mendeley el cual es una herramienta que ayudó a la recolección de los artículos y obtener una mayor organización; la aplicación de los criterios de exclusión y de calidad ayudaron a obtener los artículos adecuados y necesarios para continuar con la investigación. Esta RSL se realizó con el objetivo de tener un mejor entendimiento de los principales estudios sobre la predicción de estudiantes con repitencia escolar mediante Machine Learning.

Método de revisión

Protocolo de revisión

Esta revisión se ha realizado utilizando las directrices de Kitchenham & Charters (2007) para así obtener una Revisión Sistemática de la Literatura. Como método de revisión se elaboraron las preguntas y objetivos de la investigación, las fuentes y estrategias de búsqueda, estudios identificados, criterios de exclusión, selección de estudios, evaluación de la calidad, extracción de datos y la síntesis de hallazgos.

Problemas y objetivos de la investigación

Al llevar a cabo la revisión sistemática de la literatura, se formularon exigentes preguntas de investigación que permitieron elaborar las estrategias de búsqueda, la extracción y análisis de datos. Así mismo, también se logró identificar los objetivos que se muestran en la Tabla 1.

Tabla 1
Preguntas y objetivos de la investigación

Preguntas de Investigación	Objetivos
RQ1: ¿Cuáles son los artículos con más citas en la predicción de Repitencia Escolar en estudiantes con Machine Learning?	Determinar cuáles son los artículos con más citas en la predicción de Repitencia Escolar en estudiantes con Machine Learning.
RQ2: ¿Cuáles son las categorías de variables más aplicadas en la predicción de Repitencia Escolar en estudiantes usando Machine Learning?	Identificar las categorías de variables más aplicadas en la predicción de Repitencia Escolar en estudiantes usando aprendizaje automático.
RQ3: ¿Qué métricas se han utilizado para evaluar los resultados de la predicción de Repitencia Escolar en estudiantes aplicando Machine Learning?	Determinar las métricas que se han utilizado, en las investigaciones, para evaluar los resultados de la predicción de Repitencia Escolar en estudiantes aplicando aprendizaje automático.
RQ4: ¿Cuáles son las Palabras Clave (keywords) más utilizadas y con mayor coocurrencia en las investigaciones sobre la predicción de Repitencia Escolar en estudiantes usando Machine Learning?	Conocer las Palabras Clave (keywords) más utilizadas y con mayor coocurrencia en las investigaciones sobre la predicción de Repitencia Escolar en estudiantes usando Machine Learning.
RQ5: ¿ Cuáles son las técnicas de clasificación utilizadas para la predicción de Repitencia Escolar en estudiantes usando Machine Learning?	Identificar las técnicas de clasificación utilizadas para la predicción de Repitencia Escolar en estudiantes usando Machine Learning.
RQ6: ¿Cuáles son los artículos más citados cuyas conclusiones se caracterizan por su objetividad y polaridad en las investigaciones sobre Machine Learning y su impacto en la predicción de estudiantes con Repitencia Escolar?	Identificar los artículos más citados cuyas conclusiones se caracterizan por su objetividad y polaridad en las investigaciones sobre Machine Learning y su impacto en la predicción de estudiantes con Repitencia Escolar.

Fuentes y estrategias de búsqueda

Las fuentes de búsqueda empleadas para obtener los artículos necesarios son: Taylor & Francis Online, ACM Digital Library, Microsoft Academic, ERIC, Google Scholar, IEEE Xplore y Science Direct. Las estrategias de búsqueda se desarrollaron respetando rigurosamente la sintaxis de la fuente en base a las palabras claves y sus sinónimos (Ver Tabla 2). Al término de la búsqueda se obtuvieron 47490 artículos.

Tabla 2

Fuentes y ecuaciones de búsqueda

Fuente	Ecuación de búsqueda
Taylor & Francis Online	[All: "machine learning"] AND [[All: "prediction of school repetition"] OR [All: "school repetition prediction"] OR [All: "at risk student early detection"] OR [All: "grade retention"] OR [All: "dropout school"] OR [All: "student dropouts"] OR [All: "student's dropout"] OR [All: "prediction of dropout in students"] OR [All: "predicting dropout"] OR [All: "student attrition"] OR [All: "dropout"] OR [All: "dropout risk"] OR [All: "early school leaving"]] AND [[All: methodology] OR [All: method] OR [All: model]]
ACM Digital Library	[All: "machine learning"] AND [[All: "prediction of school repetition"] OR [All: "school repetition prediction"] OR [All: "at risk student early detection"] OR [All: "grade retention"] OR [All: "dropout school"] OR [All: "student dropouts"] OR [All: "student's dropout"] OR [All: "prediction of dropout in students"] OR [All: "predicting dropout"] OR [All: "student attrition"] OR [All: "dropout"] OR [All: "dropout risk"] OR [All: "early school leaving"]] AND [[All: methodology] OR [All: method] OR [All: model]]
Microsoft Academic	"Machine Learning" AND ("Prediction of school repetition" OR "School repetition prediction" OR "At risk student early detection" OR "Grade retention" OR "Dropout school" OR "Student dropouts" OR "Student's dropout" OR "Prediction of dropout in students" OR "Predicting dropout" OR "Student attrition" OR "Dropout" OR "Dropout risk" OR "Early school leaving") AND (method OR methodology OR model)
ERIC	"Machine Learning" AND ("Prediction of school repetition" OR "School repetition prediction" OR "At risk student early detection" OR "Grade retention" OR "Dropout school" OR "Student dropouts" OR "Student's dropout" OR "Prediction of dropout in students" OR "Predicting dropout" OR "Student attrition" OR "Dropout" OR "Dropout risk" OR "Early school leaving") AND (method OR methodology OR model)
Google Scholar	"Machine Learning" AND ("Prediction of school repetition" OR "School repetition prediction" OR "At risk student early detection" OR "Grade retention" OR "Dropout school" OR "Student dropouts" OR "Student's dropout" OR "Prediction of dropout in students" OR "Predicting dropout" OR "Student attrition" OR "Dropout" OR "Dropout risk" OR "Early school leaving") AND (method OR methodology OR model)
IEEE Xplore	"All Metadata":"Prediction of school repetition" OR "All Metadata":"School repetition prediction" OR "All Metadata":"At risk student early detection" OR "All Metadata":"Grade retention" OR "All Metadata":"Dropout school" OR "All Metadata":"Student dropouts" OR "All Metadata":"Student's dropout" OR "All Metadata":"Prediction of dropout in students" OR "All Metadata":"Predicting dropout" OR "All Metadata":"Student attrition" OR "All Metadata":"Dropout" OR "All Metadata":"Dropout risk" OR "All Metadata":"Early school leaving" AND "All Metadata":"machine learning" AND ("All Metadata":method OR methodology OR model)
Science Direct	

Criterios de exclusión

En esta etapa, se aplicaron de manera rigurosa los criterios de exclusión con el objeto de evaluar con precisión la calidad de la literatura. Los artículos se examinaron siguiendo los siguientes criterios:

CE1: Los resultados tienen una antigüedad mayor a 5 años.

CE2: Los resultados no están escritos en idioma inglés.

CE3: Los resultados no son artículos.

CE4: Los temas o especialidades no soy muy adecuados.

CE5: Los artículos no son únicos.

CE6: Los títulos y keywords del artículo no son muy adecuados.

CE7: El abstract de los artículos no es muy relevante.

Selección de artículos

En un principio, se encontró 47490 artículos que se encontraron mediante la búsqueda de documentos por medio de palabras claves y sinónimos para el estudio. Posteriormente, se aplicaron cada uno de los criterios de exclusión a todos los artículos identificados. El resultado obtenido es 90 artículos, como se muestra en la Figura 1 (Gráfico PRISMA).

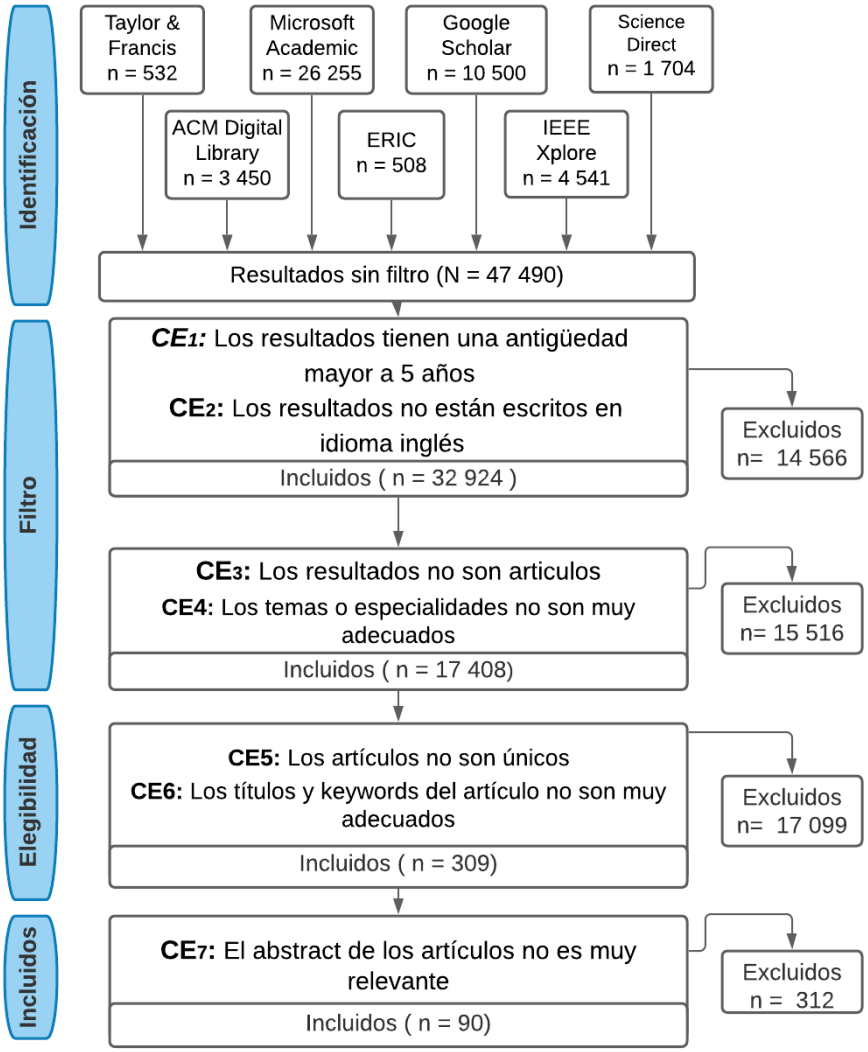

Figura 1. Aplicación de criterios de exclusión

Evaluación de la calidad

El uso de reglas para la evaluación de la calidad fue el siguiente paso aplicado para identificar la colección final de artículos comprendidos en esta revisión sistemática. Con el propósito de evaluar la calidad de los artículos se han aplicado criterios de calidad (QAs) acorde a las preguntas de investigación establecidas. Se identificaron 7 QAs, quedando de la siguiente manera:

QA1. ¿Los objetivos de la investigación se identifican claramente en el documento?

QA2. ¿El experimento realizado es adecuado y aceptable?

QA3. ¿El área específica del tema utilizada está claramente definida?

QA4. ¿El documento está bien organizado?

QA5. ¿Son apropiados los métodos utilizados para analizar los resultados?

QA6. ¿Se identifican e informan claramente los resultados de los experimentos realizados?

QA7. En general, ¿El documento en mención es adecuado?

Para cada documento, se leyó el texto completo y se aplicó los criterios de evaluación para evaluar su calidad. Todos los estudios primarios cumplieron con cada uno de los QAs. Cabe destacar que todos los artículos seleccionados a partir de los criterios de calidad han sido idóneos para el desarrollo de este artículo de investigación. Por consiguiente, se considera que la colección de resultados es de muy buena calidad para el fin de esta investigación.

Estrategias de extracción de datos

En este acápite, se emplea un listado de los artículos finales, del cual se extrae la información indispensable para responder el conjunto de preguntas de investigación. La información extraída de cada artículo incluye lo siguiente: ID del artículo, título del artículo, URL, fuente, año, país, número de páginas, idioma, tipo de artículo, nombre de la publicación, autores, filiación, número de citas, resumen, palabras clave, y tamaño de la muestra. Es preciso recalcar que no todos los artículos ayudan a responder a todas las preguntas de investigación.

Síntesis de datos

Por último, los datos extraídos para responder las preguntas de investigación, se analizaron para obtener resultados en forma de tablas y gráficos y luego realizar una comparación de la información obtenida con investigaciones similares. Estos esquemas demuestran diversos patrones y direcciones de investigación a lo largo del último lustro.

Resultados

Descripción general de los estudios

El conteo final de los trabajos de investigación estudiados obtuvo 90 artículos para su análisis de datos. La Figura 2 muestra la distribución de los artículos publicados por año en el último lustro (2017-2021). Se percibe que con el tiempo las investigaciones sobre la repitencia escolar apoyado en la aplicación de Machine Learning han ido incrementándose gradualmente.

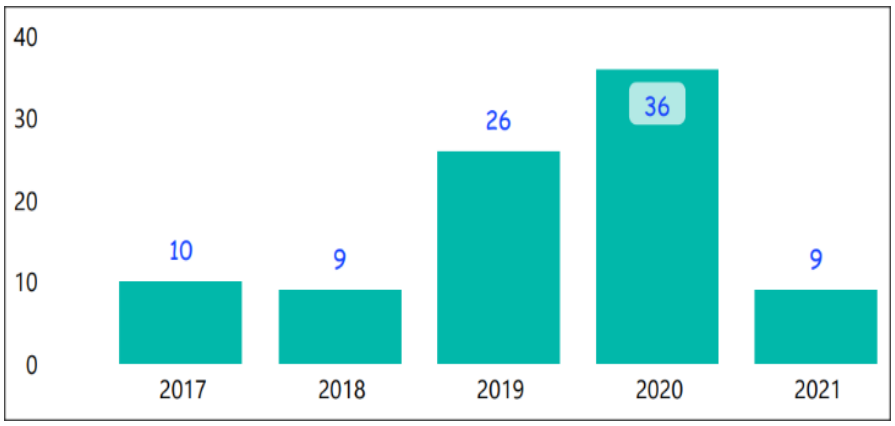

Figura 2. Distribución de los artículos publicados por año

La Figura 3 expone la distribución geográfica de los artículos publicados, entre los cuales destacan Estados Unidos y China como las naciones que más emplean Machine Learning en sus investigaciones para la predicción de repitencia escolar. Ambos países cubren el 33.3 % del total de artículos seleccionados, es decir, el 22.2 % y 11.1 % respectivamente.

Figura 3. Distribución geográfica de los artículos publicados

Los países restantes son productivos, pero con un porcentaje mucho menor. Conforme con los autores Orellana et al. (2020), concuerdan que la mayoría de los artículos publicados pertenecen a Estados Unidos de América, seguidas por las revistas de los Países Bajos.

Respuestas a las preguntas de investigación

Esta sección expone la síntesis de los hallazgos de la revisión sistemática en base a los artículos de investigación seleccionados. Las preguntas cuyo objetivo es obtener información con métodos analíticos de manera automática (RQ4 y RQ6) ayudan a los investigadores a tener una mejor idea de los temas discutidos. Clasificamos las respuestas a dichas preguntas en cuatro aspectos: (1) modelado de mapas georreferenciales, (2) nube de palabras, (3) diagramas de coocurrencia y (4) diagramas estadísticos. Se aplica las técnicas de Procesamiento de Lenguaje Natural (NLP) para las secciones del texto de los artículos. Además, se generan reportes para los siguientes niveles:

1. Títulos
2. Autores
3. Resúmenes
4. Palabras clave
5. Texto completo
6. Conclusiones

Se han convertido todos los artículos de PDF a texto. Para obtener los resultados, los datos se preprocesan para limpiarlos. Se convierte todo el texto a minúsculas, se realiza stemming y lematización, y se eliminan las palabras vacías (stopwords). Además, se excluyeron las secciones de "Agradecimientos" y "Referencias" para lograr resultados a "nivel de texto completo". Todo el texto en los diferentes niveles (título, palabras clave, resumen y texto completo) se tokeniza en palabras individuales, y luego se eliminan las palabras innecesarias. A continuación, se calcula las frecuencias de las palabras y genera la nube de palabras, los nGramas, los diagramas de coocurrencias, la objetividad y polaridad, para cada nivel.

RQ1. ¿Cuáles son los artículos con más citas en la predicción de Repitencia Escolar en estudiantes con Machine Learning?

En la Tabla 3 se muestran los ID, las referencias de los artículos y sus citas en las investigaciones sobre la predicción de Repitencia Escolar en estudiantes usando Machine Learning. En la tabla se puede notar a Costa et al. (2017) como el artículo más citado con 200 citas. Le siguen

Jokhan et al. (2019), Lacave et al. (2018) y Chung & Lee (2019) con 42, 34 y 32 citas respectivamente.

Tabla 3

Artículos y sus citas sobre predicción de repitencia escolar en estudiantes en base a Machine Learning

ID	Referencia	N° Citas
P01	Abu Zohair (2019)	21
P02	Adelman et al. (2018)	5
P03	Adnan et al. (2021)	0
P04	Al-Shabandar et al. (2019)	3
P05	Baker et al. (2020)	6
P06	Berens et al. (2021)	0
P07	Bertolini et al. (2021)	1
P08	Borrella et al. (2019)	2
P09	Botelho et al. (2019)	0
P10	Cagliero et al. (2021)	0
P11	Çam & Özdag (2020)	0
P12	Çetinkaya & Baykan (2020)	2
P13	Chen & Zhang (2017)	6
P14	Chien et al. (2020)	0
P15	Chung & Lee (2019)	32
P16	Cornell-Farrow & Garrard (2020)	0
P17	Corry et al. (2017)	2
P18	Costa et al. (2017)	200
P19	Coussement et al. (2020)	12
P20	de la Fuente-Mella et al. (2020)	0
P21	De Melo et al. (2017)	1
P22	Del Bonifro et al. (2020)	2
P23	Do Nascimento et al. (2019)	0
P24	Ezz & Elshenawy (2020)	1
P25	F. Gontzis et al. (2022)	6
P26	Figueroa-Canas & Sancho-Vinuesa (2020)	1
P27	Freitas et al. (2020)	0
P28	Gitinabard et al. (2018)	0
P29	Gkontzis et al. (2018)	3
P30	Gómez-Pulido et al. (2020)	0
P31	Goopio & Cheung (2021)	9
P32	Hai-tao et al. (2021)	0
P33	Helal et al. (2019)	6
P34	Herodotou et al. (2019)	16
P35	Hlosta et al. (2017)	23
P36	Hmedna et al. (2020)	4
P37	Hoffait & Schyns (2017)	24
P38	Huang et al. (2020)	7
P39	Huberts et al. (2020)	0

P40	Huo et al. (2020)	0
P41	Iatrellis et al. (2021)	2
P42	Imran et al. (2019)	3
P43	Irfan et al. (2019)	0
P44	Jin (2020)	11
P45	Jokhan et al. (2019)	42
P46	Karimi-Haghighi et al. (2021)	0
P47	Kartal et al. (2020)	0
P48	Kemper et al. (2020)	16
P49	Kiss et al. (2019)	0
P50	Ko & Leu (2021)	1
P51	Lacave et al. (2018)	34
P52	Lee & Chung (2019)	9
P53	Lee et al. (2020)	0
P54	Lemay & Doleck (2020)	7
P55	Liao et al. (2019)	17
P56	Lincke et al. (2021)	0
P57	Livieris et al. (2019)	5
P58	Lu et al. (2020)	0
P59	Ma et al. (2018)	1
P60	Martínez-Abad (2019)	0
P61	Moreno-Marcos et al. (2018)	30
P62	Mourdi et al. (2019)	0
P63	Mubarak et al. (2020)	2
P64	Musso et al. (2020)	3
P65	Naicker et al. (2020)	0
P66	Ninrutsirikun et al. (2020)	3
P67	Niu et al. (2018)	1
P68	Oeda & Hashimoto (2017)	5
P69	Pillutla et al. (2020)	3
P70	Qazdar et al. (2019)	5
P71	Rastrollo-Guerrero et al. (2020)	12
P72	Sabri et al. (2021)	0
P73	Shakil Ahamed et al. (2017)	8
P74	Shelton et al. (2018)	3
P75	Tamada et al. (2019)	0
P76	Thomas & Ali (2020)	4
P77	Von Hippel & Hofflinger (2021)	13
P78	Waheed et al. (2020)	20
P79	Wang et al. (2019)	1
P80	Wang et al. (2017)	26
P81	Wang et al. (2020)	0
P82	Whitehill et al. (2017)	21
P83	Wu (2019)	2
P84	Yair et al. (2020)	6
P85	Yang et al. (2020a)	0
P86	Yang et al. (2020b)	1
P87	Yildiz & Borecki (2020)	0
P88	Yousafzai et al. (2020)	0
P89	Zabriskie et al. (2019)	7
P90	Zeineddine et al. (2021)	0

El artículo más citado en la predicción de Repitencia Escolar en estudiantes aplicando Machine Learning se encuentran en la fuente de datos Microsoft Academic. Realizando una búsqueda de revisiones sistemáticas de literatura se concluye que no existe estudios para realizar una comparación.

RQ2. ¿Cuáles son las categorías de variables más aplicadas en la predicción de Repitencia Escolar en estudiantes usando Machine Learning?

En base a las revisiones de los artículos, hay 7 categorías de variables para la predicción de repitencia estudiantil utilizando Machine Learning como se expone en la Tabla 4. Los resultados obtenidos consideraran a Students Previous Grades & Class Performance (17%) y Students eLearning activity (16%) como las categorías más empleadas para la detección de riesgo de repitencia.

Tabla 3
Categorías de variables para la predicción de repitencia escolar

Categorías de variables para predicción de Repitencia Escolar	Referencia	Cant. (%)
Students Previous Grades & Class Performance	P09, P45, P37, P38, P50, P40, P78, P15, P19, P41, P61, P77, P69, P80, P86	15 (17)
Students eLearning activity	P63, P09, P29, P56, P37, P73, P44, P08, P35, P32, P82, P68, P13, P62	14 (16)
Students Demographics	P12, P29, P45, P38, P50, P90, P19, P35, P46, P75, P28, P41, P61	13 (15)
Student Social Information	P70, P74, P88, P34, P89, 35P78, P08, P86, P19, P10, P03	11 (12)
Instructor Attributes	P73, P74, P49, P50, P18, P79, P90, P26	8 (9)
Courses Attributes	P11, P47, P14, P78, P79, P15, P39	7 (8)
Others	P34, P51, P11, P22, P08, P71	6 (7)

El uso de estas variables es indispensable para la evaluación de un estudiante y saber si es potencial candidato para repetir un curso o no. Según Moreno-Marcos et al. (2019) la variable más importante para desarrollar el modelo de predicción es “dropout” siendo un factor razonable ya que de él se divide el conjunto de estudiantes que aprobaron y los que no. En relación, Abu Saa et al.

(2019) concuerdan que las variables más aplicadas son que se encuentran dentro de la clasificación de desempeño y calificaciones anteriores de los estudiantes.

RQ3. ¿Qué métricas se han utilizado para evaluar los resultados de la predicción de Repitencia Escolar en estudiantes aplicando Machine Learning?

Según los resultados finales obtenidos, la Revisión Sistemática de la Literatura demuestra que las métricas utilizadas para medir la efectividad del uso de Machine Learning en la predicción de Repitencia Escolar son 17, tal como se muestra en la Tabla 5. Los resultados de la investigación indican que Accuracy (41%) y el Area Under the Receiver Operating Characteristic (ROC - AUC) (38%) son las métricas que más se aplicaron.

Tabla 5
Métricas de Machine Learning

Métricas de ML	Referencia	Cant. (%)
Accuracy	P63, P29, P25, P45, P37, P42, P73, P38, P74, P36, P44, P51, P54, P58, P47, P27, P22, P60, P40, P78, P79, P90, P06, P76, P15, P85, P48, P02, P03, P64, P75, P72, P87, P65, P41, P32, P04	37 (41)
Area Under the Receiver Operating Characteristic (ROC) (AUC)	P63, P09, P56, P44, P89, P54, P58, P60, P14, P40, P06, P15, P19, P48, P02, P64, P46, P87, P28, P83, P41, P61, P04, P07, P05, P16, P52, P55, P82, P80, P81, P62, P67, P86	34 (38)
Precision	P63, P25, P56, P37, P42, P73, P38, P36, P89, P58, P18, P27, P40, P78, P08, P26, P85, P39, P10, P03, P64, P35, P87, P83, P41, P33, P55, P69, P80, P59, P62, P67, P86	33 (37)
Recall	P63, P25, P56, P42, P73, P38, P36, P88, P58, P18, P27, P40, P78, P06, P26, P39, P10, P03, P64, P35, P87, P83, P41, P05, P33, P52, P69, P80, P62, P86	30 (33)
F1- Score	P63, P56, P38, P36, P44, P58, P47, P27, P40, P26, P10, P03, P64, P87, P28, P83, P61, P04, P80, P59, P62, P86	22 (24)
Confusion Matrix	P42, P88, P89, P21, P08, P06, P15, P85, P39	9 (10)
Sensitivity	P37, P88, P47, P22, P40, P15, P02, P64, P04	9 (10)
Root Mean Squared Error (RMSE)	P09, P70, P88, P30, P39, P61	6 (7)
Specificity	P47, P22, P40, P15, P64, P04	6 (7)
Cohen's Kappa	P37, P54, P60	3 (3)

Mean Absolute Error (MAE)	P29, P70	2 (2)
Overrall	P45, P37	2 (2)
F- measure	P42, P73	2 (2)
Error	P12, P47	2 (2)
Positive Predictive Value (PPV)	P47, P40	2 (2)
Negative Predictive Value (NPV)	P47, P40	2 (2)
Others	P70, P74, P20, P40, P19, P46, P65	7 (8)

La métrica “Accuracy” brindó un mayor impacto como resultado de los criterios de evaluación que se tomaron en cuenta para predecir a estudiantes en riesgo de Repitencia Escolar con Machine Learning, del mismo modo, las demás métricas también lograron buenos valores, aunque con un menor impacto. De acuerdo con Alban & Mauricio (2019) “AUC” es la métrica empleada para obtener una buena predicción de estudiantes con repitencia escolar. Hellas et al. (2018) mencionan que algunas de las métricas ya mencionadas son utilizadas para predecir la repitencia de estudiantes.

RQ4. ¿Cuáles son las Palabras Clave (keywords) más utilizadas y con mayor coocurrencia en las investigaciones sobre la predicción de Repitencia Escolar en estudiantes usando Machine Learning?

En base a la revisión de los 90 artículos, en la parte izquierda de la Figura 4 muestra que las Palabras Clave más utilizadas son “machine learning”, “educational data mining” y “learning analytics”. Mediante la red bibliométrica que se muestra en la parte derecha de la Figura se muestran las coocurrencias de las palabras claves más utilizadas evidenciando a “machine learning” y “educational data mining”. También se muestra conexiones de “machine learning” con “dropout”.

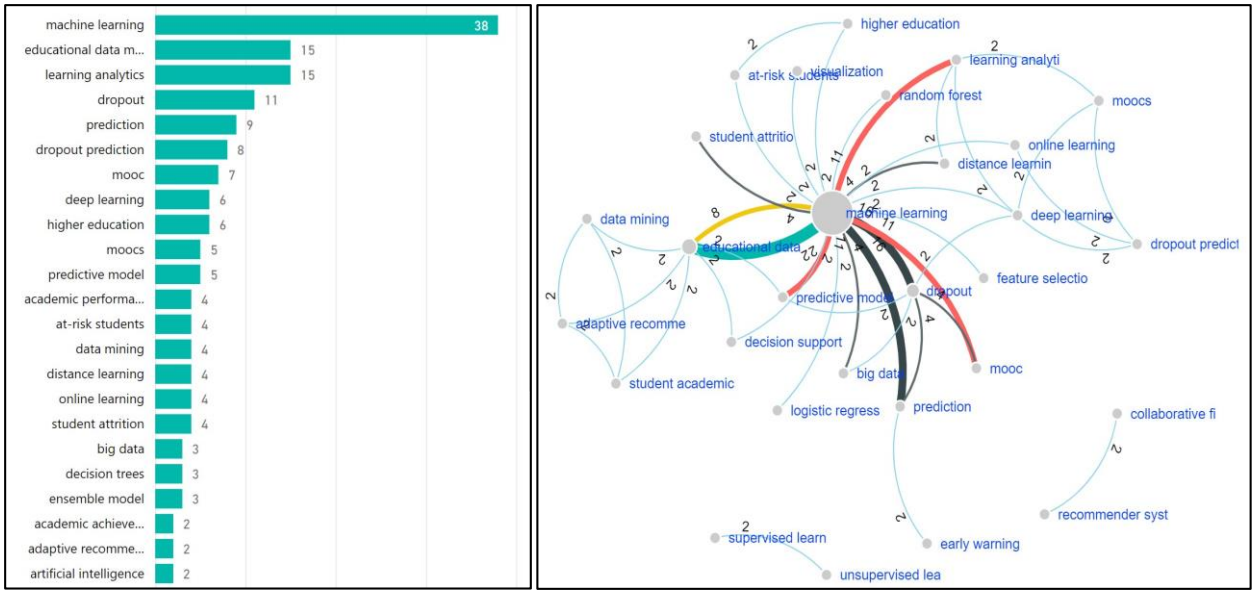

Figura 4. Palabras Clave más usadas y su Red Bibliométrica.

En esta RQ4 se puede mostrar mediante la Red Bibliométrica que “machine learning” con “educational data mining” presentan mayor coocurrencia, así como “machine learning” con “dropout” de menor impacto, pero considerable para la investigación. Después de realizar una rigurosa búsqueda se puede afirmar que no existen otras revisiones sistemáticas para comparar este tipo de representación gráfica.

RQ5. ¿Cuáles son las técnicas de clasificación utilizadas para la predicción de Repitencia Escolar en estudiantes usando Machine Learning?

La Tabla 6 muestra las técnicas de clasificación de Machine Learning. A partir de la revisión bibliográfica estructurada se identificaron 9 técnicas de clasificación aplicadas en Machine Learning con sus respectivos algoritmos para la predicción de repitencia escolar. Se concluye que los algoritmos de Árbol de Decisión (67%) y Algoritmos de Regresión (48%) son las técnicas de clasificación más utilizadas.

Tabla 6

Técnicas de clasificación de Machine Learning

Técnicas de clasificación ML	Referencia	Cant. (%)
Decision Tree Algorithms	[4] [5] [7] [9] [10] [12] [13] [14] [16] [20] [21] [22] [23] [24] [26] [27] [28] [29] [30] [31] [32] [34] [37] [38] [39] [41] [42] [45] [46] [47] [48] [50] [51] [53] [54] [55] [56] [58] [61] [63] [64] [65] [66] [67] [69] [71] [72] [73] [74] [75] [77] [79] [82] [83] [84] [85] [86] [88] [89] [90]	60 (67)
Regression algorithms	[1] [4] [7] [8] [9] [10] [12] [13] [15] [20] [21] [22] [27] [31] [32] [33] [37] [38] [40] [41] [45] [48] [49] [51] [54] [56] [58] [60] [61] [63] [64] [65] [66] [69] [71] [72] [73] [74] [75] [79] [83] [85] [89]	43 (48)
Neural Network Algorithms	[2] [3] [4] [5] [9] [10] [11] [12] [14] [20] [24] [25] [27] [31] [35] [37] [38] [39] [41] [44] [45] [50] [52] [54] [57] [60] [61] [62] [63] [71] [75] [83] [84] [85] [87] [88] [89] [90]	38 (42)
Support Vector Machine Algorithms	[4] [5] [10] [11] [12] [13] [20] [22] [23] [24] [27] [28] [30] [37] [38] [39] [45] [48] [50] [52] [54] [56] [58] [60] [61] [63] [64] [65] [66] [69] [72] [78] [82] [83] [84] [88] [89] [90]	38 (42)
Bayesian classification algorithms	[1] [5] [7] [10] [12] [19] [20] [22] [23] [24] [26] [30] [37] [39] [45] [50] [52] [58] [61] [63] [65] [66] [83] [85] [89]	25 (28)
Instance-Based Algorithms	[4] [5] [11] [14] [16] [20] [22] [23] [27] [37] [45] [49] [50] [52] [54] [56] [61] [73] [88] [89]	20 (22)
Deep Learning Algorithms	[7] [10] [13] [27] [36] [66] [68] [83] [90]	9 (10)
Clustering Algorithms	[10] [20] [25] [37] [61] [67] [80]	7 (8)
Others	[6] [7] [16] [28] [38] [52] [56]	7 (8)

Desde la perspectiva de Moreno-Marcos et al. (2019), en su investigación posicionan a los algoritmos de regresión en primer lugar como técnicas de predicción para detectar si un estudiante está en riesgo de repetir un curso o grado. Por otro lado, Abu Saa, Al-Emran y Shaalan (2019) difieren de opinión y manifiestan que la técnica de árbol de decisión son los algoritmos más comúnmente usados, mientras que sus sucesores se distancian gradualmente entre 1-3 % en base a su total.

RQ6: ¿Cuáles son los artículos más citados cuyas conclusiones se caracterizan por su objetividad y polaridad en las investigaciones sobre Machine Learning y su relación en la predicción de estudiantes con Repitencia Escolar?

La respuesta a esta interrogante se presenta la Tabla 7, en donde se evidencia los índices de objetividad y polaridad según las conclusiones de los artículos revisados en orden descendente respecto al número de citas.

Tabla 7

Artículos más citados caracterizados por su objetividad y polaridad

Título Artículo	NoCitas	Objetividad	Polaridad	Año	Fuente
[25] Evaluating the effectivenes	200	0.65	-0.04	2017	Microsoft Academic
[9] Early warning system as a pr	42	0.55	-0.01	2018	Taylor & Francis Online
[19] Learning Analytics to ident	34	0.57	0.13	2018	Taylor & Francis Online
[41] Dropout early warning syste	32	0.52	0.01	2019	Science Direct
[69] Analysing the predictive po	30	0.56	0.09	2018	Taylor & Francis Online
[83] Deep Model for Dropout Pred	26	0.54	-0.06	2017	ACM Digital Library
[11] Early detection of universi	24	0.51	-0.01	2017	Science Direct
[55] Ouroboros: early identifica	23	0.63	0.17	2017	ACM Digital Library
[51] Prediction of Student's per	21	0.49	0.07	2019	Microsoft Academic
[76] MOOC Dropout Prediction: Le	21	0.66	0.49	2017	ACM Digital Library
[31] Predicting academic perform	20	0.67	0.11	2020	Science Direct
[78] A Robust Machine Learning T	17	0.64	0.03	2019	ERIC

En base a estos resultados, se identifica que el artículo más citado posee una alta objetividad y una polaridad neutral; a diferencia de los artículos siguientes cuyas características son neutrales. De manera similar, en la Figura 5 se aprecia la objetividad y polaridad para cada uno de los de artículos seleccionados y revisados.

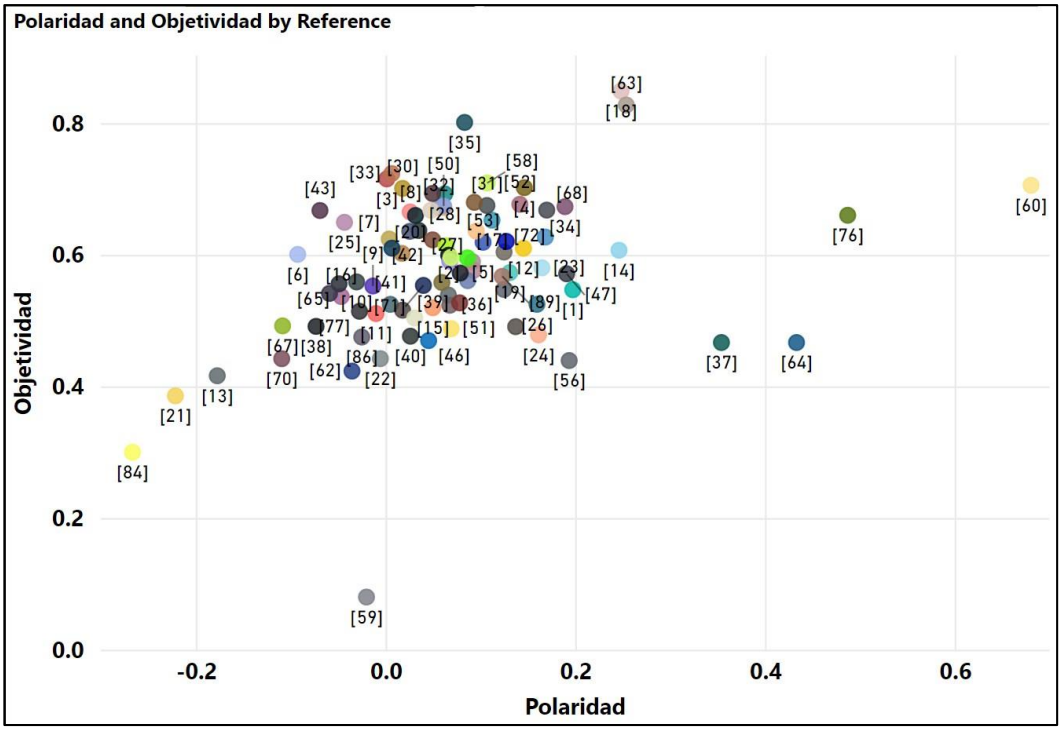

Figura 5. Artículos según su polaridad y objetividad.

Los artículos revisados presentan ecuanimidad en sus características de objetividad y polaridad, debido a que las investigaciones más citadas se basan y apoyan en hechos; además de emplear una escritura moderada hacia el lector. Luego de haber revisado rigurosamente las investigaciones seleccionadas, se puntualiza que no existen documentos con los cuales se puedan realizar comparaciones en relación a este tipo de hallazgos.

Conclusión

En esta investigación se realizó un análisis estadístico sobre la predicción de estudiantes con Machine Learning revisando 90 artículos publicados entre los años 2017 y 2021, a partir de lo cual se formularon cinco preguntas. La mayor cantidad de estudios identificados fueron artículos publicados en Microsoft Academic. También se aplicó criterios de exclusión, en base a la guía de Kitchenham & Charters (2007) y con la herramienta Mendeley se pudo extraer los datos correctos. En cuanto a los hallazgos encontrados para la RQ1 muestra los artículos más citados en la predicción de estudiantes con Repitencia Escolar usando Machine Learning. Para los hallazgos de la RQ2 se puede concluir que la categoría de variable para Machine Learning con mayor impacto es "Students Previous Grades & Class Performance".

Por su parte, la RQ3 muestra que la mayoría de los artículos publicados utilizaron "Accuracy" y "Area Under the Receiver Operating Characteristic (ROC)(AUC)" para los estudios de métricas para las predicciones. En la RQ4 se genera una Red Bibliométrica de las palabras claves más utilizadas y con mayor coocurrencia. La respuesta a la RQ5 detalla las técnicas de clasificación utilizadas en las investigaciones revisadas para la predicción de Repitencia Escolar en estudiantes usando Machine Learning. Y por último, en la RQ6 se detalla los artículos más citados cuyas conclusiones se caracterizan por su alta Objetividad y baja Polaridad. Por tanto, para futuras investigaciones se debería considerar artículos más recientes sobre la predicción de estudiantes con Repitencia Escolar usando Machine Learning.

Referencias

Abu Saa, A., Al-Emran, M., & Shaalan, K. (2019). Factors Affecting Students' Performance in Higher Education: A Systematic Review of Predictive Data Mining Techniques. In *Technology, Knowledge and Learning*, *24* (4). Springer Netherlands. https://doi.org/10.1007/s10758-019-09408-7

Abu Zohair, L. M. (2019). Prediction of Student's performance by modelling small dataset size. *International Journal of Educational Technology in Higher Education*, *16* (1). https://doi.org/10.1186/s41239-019-0160-3

Adelman, M., Haimovich, F., Ham, A., & Vazquez, E. (2018). Predicting school dropout with administrative data: new evidence from Guatemala and Honduras. *Education Economics*, *26* (4), 356–372. https://doi.org/10.1080/09645292.2018.1433127

Adnan, M., Habib, A., Ashraf, J., Mussadiq, S., Raza, A. A., Abid, M., Bashir, M., & Khan, S. U. (2021). Predicting at-Risk Students at Different Percentages of Course Length for Early Intervention Using Machine Learning Models. *IEEE Access*, *9*, 7519–7539. https://doi.org/10.1109/ACCESS.2021.3049446

Agrusti, F., Bonavolontà, G., & Mezzini, M. (2019). University dropout prediction through educational data mining techniques: A systematic review. *Journal of E-Learning and Knowledge Society*, *15* (3), 161–182. https://doi.org/10.20368/1971-8829/1135017

Alban, M., & Mauricio, D. (2019). Predicting University Dropout trough Data Mining: A systematic Literature. *Indian Journal of Science and Technology*, *12* (4), 1–12. https://doi.org/10.17485/ijst/2019/v12i4/139729

Al-Shabandar, R., Hussain, A. J., Liatsis, P., & Keight, R. (2019). Detecting at-risk students with early interventions using machine learning techniques. *IEEE Access*, *7*, 149464–149478. https://doi.org/10.1109/ACCESS.2019.2943351

Baker, R. S., Berning, A. W., Gowda, S. M., Zhang, S., & Hawn, A. (2020). Predicting K-12 Dropout. *Journal of Education for Students Placed at Risk*, *25* (1), 28–54. https://doi.org/10.1080/10824669.2019.1670065

Berens, J., Schneider, K., Görtz, S., Oster, S., & Burghoff, J. (2021). Early Detection of Students at Risk – Predicting Student Dropouts Using Administrative Student Data and Machine Learning Methods. *SSRN Electronic Journal*, *11* (3), 1–41. https://doi.org/10.2139/ssrn.3275433

Bertolini, R., Finch, S. J., & Nehm, R. H. (2021). Testing the Impact of Novel Assessment Sources and Machine Learning Methods on Predictive Outcome Modeling in Undergraduate Biology. *Journal of Science Education and Technology*, *30* (2), 193–209. https://doi.org/10.1007/s10956-020-09888-8

Borrella, I., Caballero-Caballero, S., & Ponce-Cueto, E. (2019). Predict and intervene: Addressing

the dropout problem in a MOOC-based program. *Proceedings of the 6th 2019 ACM Conference on Learning at Scale, L@S 2019*. https://doi.org/10.1145/3330430.3333634

Botelho, A. F., Varatharaj, A., Patikorn, T., Doherty, Di., Adjei, S. A., & Beck, J. E. (2019). Developing Early Detectors of Student Attrition and Wheel Spinning Using Deep Learning. *IEEE Transactions on Learning Technologies*, *12* (2), 158–170. https://doi.org/10.1109/TLT.2019.2912162

Cagliero, L., Canale, L., Farinetti, L., Baralis, E., & Venuto, E. (2021). Predicting student academic performance by means of associative classification. *Applied Sciences (Switzerland)*, *11* (4), 1–22. https://doi.org/10.3390/app11041420

Çam, E., & Özdağ, M. E. (2020). Discovery of Course Success Using Unsupervised Machine Learning Algorithms. *Malaysian Online Journal of Educational Technology*, *9* (1), 26–47. https://doi.org/10.17220/mojet.2021.9.1.242

Çetinkaya, A., & Baykan, Ö. K. (2020). Prediction of middle school students' programming talent using artificial neural networks. *Engineering Science and Technology, an International Journal*, *23* (6), 1301–1307. https://doi.org/10.1016/j.jestch.2020.07.005

Chen, Y., & Zhang, M. (2017). MOOC student dropout: Pattern and prevention. *ACM International Conference Proceeding Series*, *Part F1277*. https://doi.org/10.1145/3063955.3063959

Chien, H., Kwok, O.-M., Yeh, Y.-C., Sweany, N. W., Baek, E., & McIntosh, W. A. (2020). Identifying At-Risk Online Learners by Psychological Variables Using Machine Learning Techniques. *Online Learning*, *24* (4), 131–146. https://doi.org/10.24059/olj.v24i4.2320

Chung, J. Y., & Lee, S. (2019). Dropout early warning systems for high school students using machine learning. *Children and Youth Services Review*, *96*, 346–353. https://doi.org/10.1016/j.childyouth.2018.11.030

Cornell-Farrow, S., & Garrard, R. (2020). Machine learning classifiers do not improve the prediction of academic risk: Evidence from Australia. *Communications in Statistics Case Studies Data Analysis and Applications*, *6* (2), 228–246. https://doi.org/10.1080/23737484.2020.1752849

Corry, M., Dardick, W., & Stella, J. (2017). An examination of dropout rates for Hispanic or Latino students enrolled in online K-12 schools. *Education and Information Technologies*, *22* (5), 2001–2012. https://doi.org/10.1007/s10639-016-9530-9

Costa, E. B., Fonseca, B., Santana, M. A., de Araújo, F. F., & Rego, J. (2017). Evaluating the

effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. *Computers in Human Behavior*, *73*, 247–256. https://doi.org/10.1016/j.chb.2017.01.047

Coussement, K., Phan, M., De Caigny, A., Benoit, D. F., & Raes, A. (2020). Predicting student dropout in subscription-based online learning environments: The beneficial impact of the logit leaf model. *Decision Support Systems*, *135* (12), 113325. https://doi.org/10.1016/j.dss.2020.113325

de la Fuente-Mella, H., Gutiérrez, C. G., Crawford, K., Foschino, G., Crawford, B., Soto, R., de la Barra, C. L., Caneo, F. C., Monfroy, E., Becerra-Rozas, M., & Elórtegui-Gómez, C. (2020). Analysis and prediction of engineering student behavior and their relation to academic performance using data analytics techniques. *Applied Sciences (Switzerland), 10* (20), 1–11. https://doi.org/10.3390/app10207114

De Melo, G., Vasconcelos-Filho, E. P., Oliveira, S. M., Calixto, W. P., Ferreira, C. C., & Furriel, G. P. (2017). Evaluation techniques of machine learning in task of reprovation prediction of technical high school students. *2017 CHILEAN Conference on Electrical, Electronics Engineering, Information and Communication Technologies, CHILECON 2017 - Proceedings*, *2017-Janua* (Ml), 1–7. https://doi.org/10.1109/CHILECON.2017.8229739

Del Bonifro, F., Gabbrielli, M., Lisanti, G., & Zingaro, S. P. (2020). Student dropout prediction. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12163 LNAI*, 129–140. https://doi.org/10.1007/978-3-030-52237-7_11

Do Nascimento, R. L. S., Fagundes, R. A. A., & MacIel, A. M. A. (2019). Prediction of school efficiency rates through ensemble regression application. *Proceedings - IEEE 19th International Conference on Advanced Learning Technologies, ICALT 2019*, *2161-377X* (4), 194–198. https://doi.org/10.1109/ICALT.2019.00050

Ezz, M., & Elshenawy, A. (2020). Adaptive recommendation system using machine learning algorithms for predicting student's best academic program. *Education and Information Technologies*, *25* (4), 2733–2746. https://doi.org/10.1007/s10639-019-10049-7

F. Gontzis, A., Kotsiantis, S., T. Panagiotakopoulos, C., & Verykios, V. S. (2022). A predictive analytics framework as a countermeasure for attrition of students. *Interactive Learning Environments*, *30* (3), 568–582. https://doi.org/10.1080/10494820.2019.1674884

Figueroa-Canas, J., & Sancho-Vinuesa, T. (2020). Early prediction of dropout and final exam performance in an online statistics course. *Revista Iberoamericana de Tecnologias Del Aprendizaje*, *15* (2), 86–94. https://doi.org/10.1109/RITA.2020.2987727

Freitas, F. A., Vasconcelos, F. F. X., Peixoto, S. A., Hassan, M. M., Ali Akber Dewan, M., de Albuquerque, V. H. C., & Rebouças Filho, P. P. (2020). IoT system for school dropout prediction using machine learning techniques based on socioeconomic data. *Electronics (Switzerland)*, *9* (10), 1–14. https://doi.org/10.3390/electronics9101613

Gitinabard, N., Khoshnevisan, F., Lynch, C. F., & Wang, E. Y. (2018). Your actions or your associates? Predicting certification and dropout in MOOCs with behavioral and social features. *Proceedings of the 11th International Conference on Educational Data Mining, EDM 2018*.

Gkontzis, A. F., Kotsiantis, S., Tsoni, R., & Verykios, V. S. (2018). An effective LA approach to predict student achievement. *ACM International Conference Proceeding Series*, 76–81. https://doi.org/10.1145/3291533.3291551

Gómez-Pulido, J. A., Durán-Domínguez, A., & Pajuelo-Holguera, F. (2020). Optimizing latent factors and collaborative filtering for students' performance prediction. *Applied Sciences (Switzerland)*, *10* (16). https://doi.org/10.3390/app10165601

Goopio, J., & Cheung, C. (2021). The MOOC dropout phenomenon and retention strategies. *Journal of Teaching in Travel and Tourism*, *21* (2), 177–197. https://doi.org/10.1080/15313220.2020.1809050

Hai-tao, P., Ming-qu, F., Hong-bin, Z., Bi-zhen, Y., Jin-jiao, L., Chun-fang, L., Yan-ze, Z., & Rui, S. (2021). Predicting academic performance of students in Chinese-foreign cooperation in running schools with graph convolutional network. *Neural Computing and Applications*, *33* (2), 637–645. https://doi.org/10.1007/s00521-020-05045-9

Helal, S., Li, J., Liu, L., Ebrahimie, E., Dawson, S., & Murray, D. J. (2019). Identifying key factors of student academic performance by subgroup discovery. *International Journal of Data Science and Analytics*, *7* (3), 227–245. https://doi.org/10.1007/s41060-018-0141-y

Hellas, A., Ihantola, P., Petersen, A., Ajanovski, V. V., Gutica, M., Hynninen, T., Knutas, A., Leinonen, J., Messom, C., & Liao, S. N. (2018). Predicting academic performance: a systematic literature review. *Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education*, 175–199.

https://doi.org/10.1145/3293881.3295783

Herodotou, C., Hlosta, M., Boroowa, A., Rienties, B., Zdrahal, Z., & Mangafa, C. (2019). Empowering online teachers through predictive learning analytics. *British Journal of Educational Technology*, *50* (6), 3064–3079. https://doi.org/10.1111/bjet.12853

Hlosta, M., Zdrahal, Z., & Zendulka, J. (2017). Ouroboros: Early identification of at-risk students without models based on legacy data. *ACM International Conference Proceeding Series*, 6–15. https://doi.org/10.1145/3027385.3027449

Hmedna, B., El Mezouary, A., & Baz, O. (2020). A predictive model for the identification of learning styles in MOOC environments. *Cluster Computing*, *23* (2), 1303–1328. https://doi.org/10.1007/s10586-019-02992-4

Hoffait, A. S., & Schyns, M. (2017). Early detection of university students with potential difficulties. *Decision Support Systems*, *101*, 1–11. https://doi.org/10.1016/j.dss.2017.05.003

Huang, A. Y. Q., Lu, O. H. T., Huang, J. C. H., Yin, C. J., & Yang, S. J. H. (2020). Predicting students' academic performance by using educational big data and learning analytics: evaluation of classification methods and learning logs. *Interactive Learning Environments*, *28* (2), 206–230. https://doi.org/10.1080/10494820.2019.1636086

Huberts, L. C. E., Schoonhoven, M., & Does, R. J. M. M. (2020). Multilevel process monitoring: A case study to predict student success or failure. *Journal of Quality Technology*, *54* (2), 1–17. https://doi.org/10.1080/00224065.2020.1828008

Huo, H., Cui, J., Hein, S., Padgett, Z., Ossolinski, M., Raim, R., & Zhang, J. (2020). Predicting Dropout for Nontraditional Undergraduate Students: A Machine Learning Approach. *Journal of College Student Retention: Research, Theory and Practice*, *24* (4). https://doi.org/10.1177/1521025120963821

Iatrellis, O., Savvas, I., Fitsilis, P., & Gerogiannis, V. C. (2021). A two-phase machine learning approach for predicting student outcomes. *Education and Information Technologies*, *26* (1), 69–88. https://doi.org/10.1007/s10639-020-10260-x

Imran, A. S., Dalipi, F., & Kastrati, Z. (2019). *Predicting Student Dropout in a MOOC*. 190–195. https://doi.org/10.1145/3330482.3330514

Irfan, M., Alam, C. N., & Tresna, D. (2019). Implementation of Fuzzy Mamdani Logic Method for Student Drop Out Status Analytics. *Journal of Physics: Conference Series*, *1363* (1).

https://doi.org/10.1088/1742-6596/1363/1/012056

Jin, C. (2020). MOOC student dropout prediction model based on learning behavior features and parameter optimization. *Interactive Learning Environments*, 1–19. https://doi.org/10.1080/10494820.2020.1802300

Jokhan, A., Sharma, B., & Singh, S. (2019). Early warning system as a predictor for student performance in higher education blended courses. *Studies in Higher Education, 44* (11), 1900–1911. https://doi.org/10.1080/03075079.2018.1466872

Karimi-Haghighi, M., Castillo, C., Hernandez-Leo, D., & Oliver, V. M. (2021). Predicting Early Dropout: Calibration and Algorithmic Fairness Considerations. *Companion Proceedings 11th International Conference on Learning Analytics & Knowledge, Ml*, 1–10. https://arxiv.org/abs/2103.09068v1

Kartal, E., Özyaprak, M., Özen, Z., Şimşek, İ., Köse Biber, S., Biber, M., & Can, T. (2020). Bir Öğrenciyi Üstün Zekâlı ve Yetenekli Olarak Aday Göstermek İçin Doğru Soruları Sormak: Bir Makine Öğrenmesi Yaklaşımı. *Bilişim Teknolojileri Dergisi*, *13* (4), 385–400. https://doi.org/10.17671/gazibtd.591158

Kemper, L., Vorhoff, G., & Wigger, B. U. (2020). Predicting student dropout: A machine learning approach. *European Journal of Higher Education, 10* (1), 28–47. https://doi.org/10.1080/21568235.2020.1718520

Kiss, B., Nagy, M., Molontay, R., & Csabay, B. (2019). Predicting dropout using high school and first-semester academic achievement measures. *ICETA 2019 - 17th IEEE International Conference on Emerging ELearning Technologies and Applications, Proceedings*, 383–389. https://doi.org/10.1109/ICETA48886.2019.9040158

Kitchenham, B. A., & Charters, S. (2007). *Guidelines for performing Systematic Literature Reviews in Software Engineering. EBSE Technical Report EBSE-2007-01. School of Computer Science and Mathematics, Keele University. January*, 1–57

Ko, C. Y., & Leu, F. Y. (2021). Examining Successful Attributes for Undergraduate Students by Applying Machine Learning Techniques. *IEEE Transactions on Education*, *64* (1), 50–57. https://doi.org/10.1109/TE.2020.3004596

Lacave, C., Molina, A. I., & Cruz-Lemus, J. A. (2018). Learning Analytics to identify dropout factors of Computer Science studies through Bayesian networks. *Behaviour and Information Technology*, *37* (10–11), 993–1007.

https://doi.org/10.1080/0144929X.2018.1485053

Lee, S., & Chung, J. Y. (2019). The machine learning-based dropout early warning system for improving the performance of dropout prediction. *Applied Sciences (Switzerland), 9* (15). https://doi.org/10.3390/app9153093

Lee, Y., Shin, D., Loh, H. Bin, Lee, J., Chae, P., Cho, J., Park, S., Lee, J., Baek, J., Kim, B., & Choi, Y. (2020). Deep attentive study session dropout prediction in mobile learning environment. *CSEDU 2020 - Proceedings of the 12th International Conference on Computer Supported Education, 1*, 26–35. https://doi.org/10.5220/0009347700260035

Lemay, D. J., & Doleck, T. (2020). Predicting completion of massive open online course (MOOC) assignments from video viewing behavior. *Interactive Learning Environments*, *30* (10), 1–12. https://doi.org/10.1080/10494820.2020.1746673

Liao, S. N., Zingaro, D., Thai, K., Alvarado, C., Griswold, W. G., & Porter, L. (2019). A robust machine learning technique to predict low-performing students. *ACM Transactions on Computing Education*, *19* (3), 1–19. https://doi.org/10.1145/3277569

Lincke, A., Jansen, M., Milrad, M., & Berge, E. (2021). The performance of some machine learning approaches and a rich context model in student answer prediction. *Research and Practice in Technology Enhanced Learning*, *16* (1). https://doi.org/10.1186/s41039-021-00159-7

Livieris, I. E., Drakopoulou, K., Tampakas, V. T., Mikropoulos, T. A., & Pintelas, P. (2019). Predicting Secondary School Students' Performance Utilizing a Semi-supervised Learning Approach. *Journal of Educational Computing Research*, *57* (2), 448–470. https://doi.org/10.1177/0735633117752614

Lu, D. N., Le, H. Q., & Vu, T. H. (2020). The factors affecting acceptance of e-learning: A machine learning algorithm approach. *Education Sciences*, *10* (10), 1–13. https://doi.org/10.3390/educsci10100270

Ma, X., Yang, Y., & Zhou, Z. (2018). Using machine learning algorithm to predict student pass rates in online education. *ACM International Conference Proceeding Series*, 156–161. https://doi.org/10.1145/3220162.3220188

Martínez-Abad, F. (2019). Identification of Factors Associated With School Effectiveness With Data Mining Techniques: Testing a New Approach. *Frontiers in Psychology*, *10* (11), 1–13. https://doi.org/10.3389/fpsyg.2019.02583

Moreno-Marcos, P. M., Muñoz-Merino, P. J., Alario-Hoyos, C., Estévez-Ayres, I., & Delgado

Kloos, C. (2018). Analysing the predictive power for anticipating assignment grades in a massive open online course. *Behaviour and Information Technology*, *37* (10–11), 1021–1036. https://doi.org/10.1080/0144929X.2018.1458904

Moreno-Marcos, P. M., Alario-Hoyos, C., Munoz-Merino, P. J., & Kloos, C. D. (2019). Prediction in MOOCs: A Review and Future Research Directions. *IEEE Transactions on Learning Technologies*, *12* (3), 384–401. https://doi.org/10.1109/TLT.2018.2856808

Mourdi, Y., Sadgal, M., El Kabtane, H., & Berrada Fathi, W. (2019). A machine learning-based methodology to predict learners' dropout, success or failure in MOOCs. *International Journal of Web Information Systems*, *15* (5), 489–509. https://doi.org/10.1108/IJWIS-11-2018-0080

Mubarak, A. A., Cao, H., & Zhang, W. (2020). Prediction of students' early dropout based on their interaction logs in online learning environment. *Interactive Learning Environments*, 3*0* (8), 1–20. https://doi.org/10.1080/10494820.2020.1727529

Musso, M. F., Hernández, C. F. R., & Cascallar, E. C. (2020). Predicting key educational outcomes in academic trajectories: a machine-learning approach. *Higher Education*, *80* (5), 875–894. https://doi.org/10.1007/s10734-020-00520-7

Naicker, N., Adeliyi, T., & Wing, J. (2020). Linear Support Vector Machines for Prediction of Student Performance in School-Based Education. *Mathematical Problems in Engineering*, *2020*. https://doi.org/10.1155/2020/4761468

Ninrutsirikun, U., Imai, H., Watanapa, B., & Arpnikanondt, C. (2020). Principal Component Clustered Factors for Determining Study Performance in Computer Programming Class. *Wireless Personal Communications*, *115* (4), 2897–2916. https://doi.org/10.1007/s11277-020-07194-5

Niu, Z., Li, W., Yan, X., & Wu, N. (2018). Exploring causes for the dropout on massive open online courses. *ACM International Conference Proceeding Series*, 47–52. https://doi.org/10.1145/3210713.3210727

Oeda, S., & Hashimoto, G. (2017). Log-Data Clustering Analysis for Dropout Prediction in Beginner Programming Classes. *Procedia Computer Science*, *112*, 614–621. https://doi.org/10.1016/j.procs.2017.08.088

Orellana, D., Segovia, N., & Cánovas, B. R. (2020). El abandono estudiantil en programas de educación superior virtual: revisión de literatura. *Revista de la Educación Superior*, *49*

(194), 45–62. https://doi.org/10.36857/resu.2020.194.1124

Pillutla, V. S., Tawfik, A. A., & Giabbanelli, P. J. (2020). Detecting the Depth and Progression of Learning in Massive Open Online Courses by Mining Discussion Data. *Technology, Knowledge and Learning*, *25* (4), 881–898. https://doi.org/10.1007/s10758-020-09434-w

Qazdar, A., Er-Raha, B., Cherkaoui, C., & Mammass, D. (2019). A machine learning algorithm framework for predicting students performance: A case study of baccalaureate students in Morocco. *Education and Information Technologies*, *24* (6), 3577–3589. https://doi.org/10.1007/s10639-019-09946-8

Rastrollo-Guerrero, J. L., Gómez-Pulido, J. A., & Durán-Domínguez, A. (2020). Analyzing and predicting students' performance by means of machine learning: A review. *Applied Sciences (Switzerland)*, *10* (3). https://doi.org/10.3390/app10031042

Sabri, M., El Bouhdidi, J., & Chkouri, M. Y. (2021). A proposal for a deep learning model to enhance student guidance and reduce dropout. *Lecture Notes in Networks and Systems*, *144*, 158–165. https://doi.org/10.1007/978-3-030-53970-2_15

Shakil Ahamed, A. T. M., Mahmood, N. T., & Rahman, R. M. (2017). An intelligent system to predict academic performance based on different factors during adolescence. *Journal of Information and Telecommunication*, *1* (2), 155–175. https://doi.org/10.1080/24751839.2017.1323488

Shelton, B. E., Yang, J., Hung, J. L., & Du, X. (2018). Two-stage predictive modeling for identifying at-risk students. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11003 LNCS*, 578–583. https://doi.org/10.1007/978-3-319-99737-7_61

Tamada, M. M., Netto, J. F. D. M., & De Lima, D. P. R. (2019). Predicting and Reducing Dropout in Virtual Learning using Machine Learning Techniques: A Systematic Review. *Proceedings - Frontiers in Education Conference, FIE*, *2019-Octob*(October). https://doi.org/10.1109/FIE43999.2019.9028545

Thomas, J. J., & Ali, A. M. (2020). Dispositional Learning Analytics Structure Integrated with Recurrent Neural Networks in Predicting Students Performance. *Advances in Intelligent Systems and Computing*, *1072*, 446–456. https://doi.org/10.1007/978-3-030-33585-4_44

Von Hippel, P. T., & Hofflinger, A. (2021). The data revolution comes to higher education: identifying students at risk of dropout in Chile. *Journal of Higher Education Policy and*

Management, *43* (1), 2–23. https://doi.org/10.1080/1360080X.2020.1739800

Waheed, H., Hassan, S. U., Aljohani, N. R., Hardman, J., Alelyani, S., & Nawaz, R. (2020). Predicting academic performance of students from VLE big data using deep learning models. *Computers in Human Behavior*, *104*. https://doi.org/10.1016/j.chb.2019.106189

Wang, H., Li, G., Wang, G., & Lin, L. (2019). CamDrop: A new explanation of dropout and a guided regularization method for deep neural networks. *International Conference on Information and Knowledge Management, Proceedings*, 1141–1149. https://doi.org/10.1145/3357384.3357999

Wang, W., Yu, H., & Miao, C. (2017). Deep model for dropout prediction in MOOCs. *ACM International Conference Proceeding Series, Part F1306*, 26–32. https://doi.org/10.1145/3126973.3126990

Wang, X., Schneider, H., & Walsh, K. R. (2020). A Predictive Analytics Approach to Building a Decision Support System for Improving Graduation Rates at a Four-Year College. *Journal of Organizational and End User Computing*, *32* (4), 43–62. https://doi.org/10.4018/joeuc.2020100103

Whitehill, J., Mohan, K., Seaton, D., Rosen, Y., & Tingley, D. (2017). *MOOC Dropout Prediction*. 161–164. https://doi.org/10.1145/3051457.3053974

Wu, N. (2019). *CLMS - Net : Dropout Prediction in MOOCs with Deep Learning*. https://doi.org/10.1145/3321408.3322848

Yair, G., Rotem, N., & Shustak, E. (2020). The riddle of the existential dropout: lessons from an institutional study of student attrition. *European Journal of Higher Education*, *10* (4), 436–453. https://doi.org/10.1080/21568235.2020.1718518

Yang, J., Devore, S., Hewagallage, D., Miller, P., Ryan, Q. X., & Stewart, J. (2020). Using machine learning to identify the most at-risk students in physics classes. *Physical Review Physics Education Research*, *16* (2), 20130. https://doi.org/10.1103/PhysRevPhysEducRes.16.020130

Yang, Z., Yang, J., Rice, K., Hung, J. L., & Du, X. (2020). Using Convolutional Neural Network to Recognize Learning Images for Early Warning of At-Risk Students. *IEEE Transactions on Learning Technologies*, *13* (3), 617–630. https://doi.org/10.1109/TLT.2020.2988253

Yildiz, M., & Börekci, C. (2020). Predicting Academic Achievement with Machine Learning Algorithms. *Journal of Educational Technology and Online Learning*, *3* (3).

https://doi.org/10.31681/jetol.773206

Yousafzai, B. K., Hayat, M., & Afzal, S. (2020). Application of machine learning and data mining in predicting the performance of intermediate and secondary education level student. *Education and Information Technologies*, *25* (6), 4677–4697. https://doi.org/10.1007/s10639-020-10189-1

Zabriskie, C., Yang, J., Devore, S., & Stewart, J. (2019). Using machine learning to predict physics course outcomes. *Physical Review Physics Education Research*, *15* (2), 20120. https://doi.org/10.1103/PhysRevPhysEducRes.15.020120

Zeineddine, H., Braendle, U., & Farah, A. (2021). Enhancing prediction of student success: Automated machine learning approach. *Computers and Electrical Engineering*, *89* (11), 106903. https://doi.org/10.1016/j.compeleceng.2020.106903