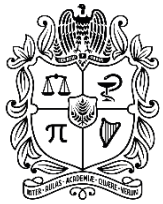


UNIVERSIDAD
NACIONAL
DE COLOMBIA

Metodología para Predecir el Desempeño Estudiantil en Cursos Universitarios Virtuales a Distancia

Víctor Daniel Gil Vera, MSc

Universidad Nacional de Colombia
Facultad de Minas, Área Curricular de Sistemas e Informática
Medellín, Colombia
2021

Metodología para Predecir el Desempeño Estudiantil en Cursos Universitarios Virtuales a Distancia

Víctor Daniel Gil Vera, MSc

Tesis presentada como requisito parcial para optar al título de:

Doctor en Ingeniería - Sistemas

Director:

Juan David Velásquez Henao, PhD

Codirector:

Carlos Jaime Franco Cardona, PhD

Línea de Investigación:

Analítica

Grupo de Investigación:

Big Data & Data Analytics

Universidad Nacional de Colombia

Facultad de Minas, Área Curricular de Sistemas e Informática

Medellín, Colombia

2021

Los problemas no son eternos, siempre tienen solución, lo único que no se resuelve es la muerte. Así que, no permitas que nadie te insulte, te humille o te baje la autoestima. Los gritos son el arma de los cobardes, de los que no tienen la razón. Siempre encontraremos gente que te quiere culpar de sus fracasos y que olvida que cada quien tiene lo que se merece.... Por eso disfruta la vida porque es muy corta, por eso ama, se feliz y siempre sonríe, solo vive intensamente para ti y por ti.
Recuerda:
Antes de discutir, respira
Antes de hablar, escucha
Antes de criticar, examínate
Antes de escribir, piensa
Antes de herir, siente
Antes de rendirte, intenta
Antes de morir... ¡vive!!!

William Shakespeare

Declaración de obra original

Yo declaro lo siguiente:

He leído el Acuerdo 035 de 2003 del Consejo Académico de la Universidad Nacional. «Reglamento sobre propiedad intelectual» y la Normatividad Nacional relacionada al respeto de los derechos de autor. Esta disertación representa mi trabajo original, excepto donde he reconocido las ideas, las palabras, o materiales de otros autores.

Cuando se han presentado ideas o palabras de otros autores en esta disertación, he realizado su respectivo reconocimiento aplicando correctamente los esquemas de citas y referencias bibliográficas en el estilo requerido.

He obtenido el permiso del autor o editor para incluir cualquier material con derechos de autor (por ejemplo, tablas, figuras, instrumentos de encuesta o grandes porciones de texto).

Por último, he sometido esta disertación a la herramienta de integridad académica, definida por la universidad.

_____

Víctor Daniel Gil Vera

Fecha: 26 de septiembre de 2021

Agradecimientos

- A Dios y a mi familia, a los profesores Juan David Velásquez Henao y Carlos Jaime Franco Cardona, por el apoyo, paciencia y tiempo que dedicaron en la dirección de esta investigación.
- A la Universidad Nacional de Colombia, especialmente al Departamento de Ciencias de la Computación y la Decisión de la Facultad de Minas por darme la oportunidad de realizar el doctorado.

Resumen

El incremento masivo de cursos universitarios virtuales a distancia en universidades e Instituciones de Educación Superior a nivel mundial, ha llevado al incremento en la generación de información relacionada con el rendimiento académico estudiantil; esta información puede ser aprovechada para predecir el desempeño académico y prevenir la mortalidad académica y la deserción. A partir de los resultados de la revisión sistemática de literatura se identificó que no existe una metodología que permita a los docentes de cursos universitarios virtuales a distancia predecir el rendimiento académico estudiantil; algunas investigaciones presentan ejercicios de clasificación sobre el desempeño de los estudiantes; pero no establecen un procedimiento formal que pueda ser empleado por docentes de cualquier área de conocimiento que dicten este tipo de cursos. El principal aporte de esta investigación doctoral es la creación de una metodología para predecir el desempeño académico (Aprueba/Reprueba) en cursos universitarios virtuales a distancia. En resumen, la metodología está conformada por los siguientes pasos; determinación de las variables a analizar, construcción de la base de datos, construcción de los modelos de predicción, evaluación de los modelos y visualización de la predicción. La metodología va más allá del Machine Learning dado que esta considera aspectos relevantes del contexto educativo que deben ser considerados para que las predicciones tengan sentido. Se concluye que la metodología formulada tiene una alta precisión e involucra diferentes aspectos relacionados con la vida académica y personal de los estudiantes, ya que el rendimiento académico estudiantil en este tipo de cursos depende de diversos factores.

Palabras clave: Educación, Estudiante, Machine Learning, Predicción, Rendimiento.

Methodology for Predicting Student Performance in Virtual University Distance Learning Courses

Abstract

The massive increase of virtual university distance learning courses in universities and Higher Education Institutions worldwide has led to an increase in the generation of information related to student academic performance; this information can be used to predict academic performance and prevent academic mortality and dropout. From the results of the systematic literature review, it was identified that there is no methodology that allows teachers of virtual distance university courses to predict student academic performance; some researches present classification exercises on student performance; but they do not establish a formal procedure that can be used by teachers of any area of knowledge who teach this type of courses. The main contribution of this doctoral research is the creation of a methodology to predict academic performance (Pass/Fail) in virtual distance university courses. In summary, the methodology consists of the following steps; determination of the variables to be analyzed, construction of the database, construction of the prediction models, evaluation of the models and visualization of the prediction. The methodology goes beyond Machine Learning since it considers relevant aspects of the educational context that must be considered for the predictions to make sense. This research concludes that the formulated methodology has a high accuracy and involves different aspects related to the academic and personal life of the students, since student academic performance in this type of courses depends on several factors.

Keywords: Education, Student, Machine Learning, Performance, Prediction.

Contenido

Pág.

Lista de figuras

Pág.

Lista de tablas

Lista de abreviaturas

IES	Instituciones de Educación Superior
EVA	Entornos virtuales de aprendizaje
EDM	Educational Data Mining
RNA	Redes Neuronales Artificiales
API	Application Programming Interface
LTI	Learning Tools Interoperability
M	Millones
RSL	Revisión Sistemática de Literatura
CGPA	Promedio Acumulativo de Calificaciones
KDD	Knowledge Discovery in Databases
CUVD	Cursos Universitarios Virtuales a Distancia
CUP	Cursos Universitarios Presenciales
ML	Machine Learning (Aprendizaje de Máquinas)
PEV	Plataformas Educativas Virtuales
KNN	K- Nearest Neighbors

1 Definición del proyecto

1.1 Introducción

En el sistema educativo actual la predicción del rendimiento académico estudiantil es de gran utilidad, ya que permite a los docentes identificar a los estudiantes que pueden tener un bajo rendimiento, hacer un seguimiento cuidadoso del proceso de aprendizaje y personalizar las estrategias de enseñanza; el rendimiento académico estudiantil es una parte importante del prestigio de las universidades (Hawlitschek et al., 2019); la capacidad de predecir el desempeño de los estudiantes es importante para mejorar el desempeño de los docentes y se constituye como un conocimiento valioso que se puede utilizar para formular planes estratégicos para desarrollar una educación de calidad (Imran et al., 2019; Dutt et al., 2017).

Universidades e Instituciones de Educación Superior (IES) a nivel mundial han adoptado el uso de plataformas educativas virtuales (PEV); diversas investigaciones han demostrado las grandes ventajas de su implementación y su incidencia en la mejora del rendimiento académico de los estudiantes, ya que permiten almacenar información valiosa sobre el proceso de interacción de los mismos con los recursos, actividades, foros, entre otros (Zhang et al., 2020). La predicción del desempeño de los estudiantes se convirtió en un deseo urgente en la mayoría de universidades e IES para ayudar a los estudiantes en riesgo y asegurar su retención, lo que contribuye al mejoramiento del ranking y de la reputación de las mismas (Abu Zohair, 2019). Las universidades e IES a menudo sienten curiosidad por saber cuántos estudiantes aprobarán o reprobarán los cursos (Imran et al., 2019).

El análisis del aprendizaje hace referencia a la recolección, análisis y presentación de datos sobre los estudiantes y sus contextos con el fin de comprender y optimizar el aprendizaje y los entornos en los que se producen; este surge como una disciplina emergente que busca mejorar la enseñanza y el aprendizaje por medio de una evaluación crítica de datos y la generación de patrones relacionados con los hábitos y respuestas de los estudiantes que permitan proporcionar retroalimentación oportuna (Peña, 2017). Este se centra deliberadamente en abordar las exigentes necesidades de procesamiento y análisis de grandes conjuntos de datos de aprendizaje para proporcionar a los profesores, administrativos e interesados de e-Learning un conocimiento significativo para mejorar las

actividades educativas, el seguimiento de los cursos y el proceso de aprendizaje (Zomaya, 2017).

Actualmente, la predicción del desempeño de los estudiantes está siendo abordada por la minería de datos educativos (EDM). Esta desarrolla métodos para descubrir datos que se derivan del entorno educativo; los cuales son utilizados para comprender al alumno y su entorno de aprendizaje. La aplicación de la EDM ayuda a predecir el rendimiento académico de los estudiantes con el desarrollo de modelos que son construidos con antecedentes académicos de los estudiantes y con el historial de su rendimiento académico (Ihantola et al., 2015). Las técnicas de Machine Learning (ML) de aprendizaje supervisado (Árboles de decisión, Naïve-Bayes, SVM, K-Nearest Neighbor K-NN, Bosques Aleatorios y Regresión logística) son útiles para predecir el rendimiento a varios niveles (Wong, 2017); estas técnicas se utilizan para predecir que estudiantes tienen mayor riesgo de reprobar, lo que permite a los docentes brindarles una orientación personalizada con anticipación. Las técnicas de Árboles de Decisión y Naïve-Bayes han sido empleadas ampliamente en la EDM (Ihantola et al., 2015). Diversas investigaciones presentan criterios para seleccionar la técnica de clasificación más apropiada en términos de precisión; sin embargo, en muchasde ellas ignoran los problemas que surgen durante las fases de minería de datos como la normalización, la alta dimensionalidad, el desequilibrio de clases, la identificación de parámetros óptimos y el error de clasificación (Shahiri et al., 2015).

El aprendizaje automático (ML) se refiere a la capacidad de un sistema para adquirir e integrar conocimientos a través de observaciones a gran escala, y para mejorar y extenderse aprendiendo nuevos conocimientos en lugar de ser programado de manera explícita (Woolf, 2009); el ML busca convertir datos empíricos en modelos utilizables; este surge como resultado de la fusión de la estadística y la inteligencia artificial (IA) (Edgar & Manz, 2017). El ML es un proceso definido, ya que la máquina aprende de experiencias pasadas, este permite analizar grandes volúmenes de datos para encontrar reglas y patrones ocultos, los cuales pueden ser empleados para caracterizar nuevos datos e información (Gedrimiene et al., 2019); gracias al ML los datos digitales pueden ser analizados para encontrar patrones que son complejos de identificar para los humanos (M. Barb, R. Vilanova, J. Lopez Vicario, 2017).

1.2 Revisión Sistemática de Literatura

Esta sección tiene como objetivo detallar la manera como se identificó el vacío en el conocimiento; lo que permitió definir la pregunta, la hipótesis y los objetivos de la investigación. Para lograr lo mencionado anteriormente, se realizó una revisión sistemática de literatura (RSL), la cual tiene como objetivo responder preguntas de investigación claramente definidas; esta emplea métodos sistemáticos explícitos para identificar, seleccionar y evaluar críticamente la investigación relevante de investigaciones científicas (Ten Ham-Baloyi y Jordan, 2016). La RSL es una forma metódica para identificar, evaluar e interpretar estudios empíricos disponibles realizados sobre un tema, pregunta de investigación, o fenómeno de interés (Kitchenham, 2004; Sorrell, 2007; Tranfield et al., 2003). Las fases o etapas para realizar una RSL son:

- Identificación de investigaciones: el objetivo de una RSL es encontrar la mayor cantidad posible de estudios primarios relacionados con una pregunta de investigación, tomando como base criterios y parámetros de búsqueda antes de hacer la selección final. La definición de estos criterios es la principal diferencia que la distingue de la revisión tradicional de literatura (Kitchenham, 2004).
- Selección de estudios: una vez obtenidos los estudios primarios potencialmente relevantes, se debe evaluar su calidad real, para filtrar los estudios que realmente puedan contribuir a responder las preguntas de investigación (Kitchenham, 2004).
- Extracción de datos: el objetivo de esta etapa es diseñar formularios de extracción de datos para registrar con precisión la información que los investigadores obtienen de los estudios primarios. Los formularios de extracción de datos deben ser establecidos al momento de definir el protocolo de búsqueda (Kitchenham, 2004).
- Síntesis: consiste en recopilar y resumir los resultados de los estudios primarios seleccionados. La síntesis debe ser descriptiva (no cuantitativa). Sin embargo, a veces es posible complementar una síntesis descriptiva con un resumen cuantitativo (Kitchenham, 2004).
- Responder las preguntas de investigación.

Específicamente, el objetivo de esta RSL fue responder las siguientes preguntas de investigación:

- P1. ¿Qué investigaciones se han desarrollado para predecir el rendimiento académico estudiantil de cursos universitarios?

1.2.1 Metodología

A continuación, se definen los pasos utilizados para ejecutar la RSL:

- Bases de datos utilizadas: Scopus e ISI Web of Knowledge WoS.
- Período de consulta: toda la información disponible hasta mayo de 2021.
- Cadena de búsqueda:

  TITLE-ABS-KEY (Academic Performance AND (Higher Education OR University OR Variables OR Predict* OR Forecast* OR Models OR Methodologies OR Models OR Techniques))

- Criterios de inclusión: estén relacionados directamente con la predicción del desempeño, que estén aplicados a cursos universitarios, que desarrollen modelos predictivos del desempeño de los estudiantes, que no reproduzcan investigaciones realizadas sin ningún tipo de aporte.
- Criterios de exclusión: que estuvieran enfocados a la educación básica y secundaria, que no desarrollaban modelos predictivos del rendimiento de los estudiantes, no presentan una metodología clara, reproducen investigaciones previas y no aportan nuevo conocimiento.

1.2.2 Resultados

Al aplicar la cadena de búsqueda en las bases de datos seleccionadas se recuperaron automáticamente 95 documentos. Al aplicar los criterios de inclusión y exclusión se descartaron 63 documentos, para obtener un total de 32 que fueron utilizados en esta RSL.

1.2.2.1 Métricas generales de los trabajos seleccionados

La Tabla 1 presenta las principales métricas de los trabajos seleccionados, la mayoría de publicaciones fueron realizadas entre varios investigadores, solo una minoríafueron realizadas por un solo autor, el promedio de citas por documento es superior a9, lo que indica que estas investigaciones han sido citadas en esta área de conocimiento.

Tabla 1. Métricas de los Trabajos Seleccionados I

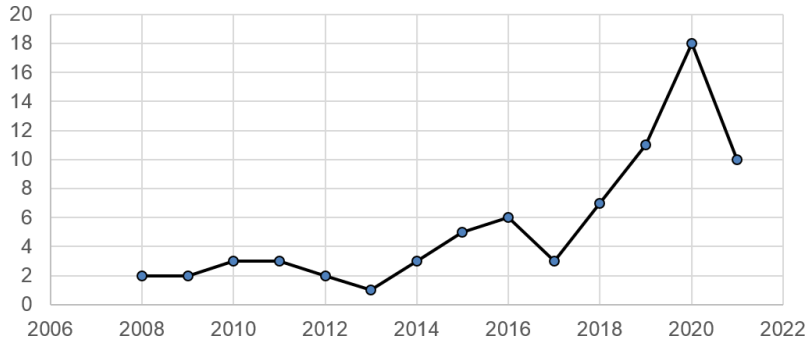
Información Principal Sobre los Datos	
Período de publicación	2008:2021
Cantidad	50
Promedio de citas por documento	9,05
Promedio de citas por año por documento	2,45
Referencias	3.594
Palabras clave	348
Autores	
Publicaciones con un solo autor	6
Publicaciones con varios autores	44
Colaboración de Autores	
Promedio de autores por documento	2,16
Promedio de coautores por documento	2,60
Índice de colaboración	2,24

Fuente: elaboración del autor con Biblioshiny para Bibliometrix

1.2.2.2 Cantidad de publicaciones por año

La Figura 1 presenta el total de publicaciones por año; se evidencia un crecimiento creciente en la cantidad de publicaciones por año; en el año 2020 se registró la mayor cantidad (18 publicaciones), la caída en el año 2021 obedece a que solo se tiene información parcial para este año.

Figura 1. Publicaciones por Año para la RSL del Capítulo 1.



Fuente: elaboración del autor con Biblioshiny para Bibliometrix

1.2.2.3 Países con mayor producción

La Tabla 2 presenta los tres países con la mayor cantidad de publicaciones y la principal contribución de cada uno en el área de conocimiento.

Tabla 2. Top 3 de los países con la mayor cantidad de publicaciones y su principal contribución en la predicción del rendimiento académico estudiantil de cursos universitarios

País	Artículos	Principal contribución
EEUU	11	Las investigaciones realizadas en este país han estado centradas en el desarrollo de modelos predictivos que permitan identificar a estudiantes en riesgo con el ánimo de ayudarles a incrementar sus probabilidades de éxito a través de la implementación de estrategias de diferente naturaleza. Diversas investigaciones señalan la importancia de considerar aspectos como; el género, la raza, la etnia, la clase social generacional, la demografía de los estudiantes, la ubicación geográfica de la universidad y la situación socioeconómica de los estudiantes, para predecir el rendimiento académico y el tiempo que tardarán en graduarse; también, señalan que es esencial establecer planes de intervención temprana y crear redes compactas de estudiantes que incentiven el aprendizaje colectivo (B.-H. Kim et al., 2018; N. Yadav & Srivastava, 2020; J. Kim et al., 2020; Gil-Herrera et al., 2011).
Australia	9	Los trabajos publicados se enfocan en la identificación de los predictores más importantes del rendimiento académico de los estudiantes; analizan tanto aspectos comportamentales como aspectos externos e internos de los estudiantes (conocimientos previos, rendimiento académico, creencias epistemológicas, intereses, edad y género). Diversas investigaciones han estado enfocadas a entrenar y evaluar la calidad de modelos predictivos; además, existe una gran cantidad de investigaciones relacionadas con el análisis del aprendizaje en la educación superior, lo que se constituye como una ventaja a la hora de establecer estrategias encaminadas a evitar el bajo rendimiento, la mortalidad y la deserción académica (Umer et al., 2021; Deo et al., 2020; Virvou et al., 2020; Ahammed & Smith, 2019).

Reino Unido	7	Al igual que en EEUU y Australia, la mayoría de las investigaciones realizadas en este país han estado centradas en la identificación de factores predictores del desempeño académico y en el desarrollo de modelos predictivos. En algunas investigaciones se han enfocado en analizar la relación entre los estilos de aprendizaje de los estudiantes y los resultados finales obtenidos por los mismos. Uno de los aspectos diferenciadores que consideran en la predicción del rendimiento académico de los estudiantes de cursos virtuales en este país es el índice de necesidades insatisfechas (NBI) (calidad de los servicios públicos, condición económica de los hogares, condiciones físicas de lavivienda y hacinamiento), este indicador permite tener un mayor entendimiento de las condiciones físicas en las que vive el estudiante, locual es importante al momento de predecir su rendimiento académico (Xu et al., 2017; Ahammed & Smith, 2019; Cuevas et al., 2018; Sagoo et al., 2016).

Fuente: elaboración del autor con Biblioshiny para Bibliometrix

1.2.2.4 Artículos más citados

La Tabla 3 presenta el Top 10 de las publicaciones con la mayor cantidad de citas en orden descendente y las principales contribuciones de cada trabajo.

Tabla 3. Principales publicaciones sobre la predicción del rendimiento académico

Autores y Año	Título	Citas	Contribución / Conclusión
Lu et al., (2018)	Applying learning analytics for the early prediction of Students' academic performance in blended learning	86	El rendimiento académico estudiantil al finalizar un curso se puede predecir solo con un porcentaje de avance del mismo; los conjuntos de datos que combinan factores críticos personales y factores académicos de los cursos en línea son los que tienen el rendimiento predictivo más alto.
Daud et al., (2017)	Predicting student performance using advanced learning analytics	76	Los modelos predictivos construidos con información relacionada con factores relacionados con gastos familiares e información personal de los estudiantes superan significativamente a los modelos que solo se centran en factores relacionados con el rendimiento académico.
D. A. Gómez-Aguilar et al., (2015)	Tap into visual analysis of customization of grouping of activities in eLearning	71	Existe un patrón recurrente en la frecuencia de los comportamientos y el rendimiento en diferentes cursos virtuales; se recalca la importancia de crear y desarrollar las instrucciones de las actividades evaluativas de los cursos de manera tal que mejore la experiencia y los resultados generales del aprendizaje.
Kovanović et al., (2015)	Penetrating the black box of time-on-task estimation	49	Es importante estimar el tiempo que dedican los estudiantes al desarrollo de tareas y actividades en las plataformas educativas virtuales, ya que éste factor puede influir en su desempeño académico.

Yu & Jo, (2014)	Educational technology approach toward learning analytics: Relationship between student online behavior and learning performance in higher education	47	El tiempo total de estudio en las PEV, la interacción con los compañeros, la regularidad del intervalo de aprendizaje y el número de descargas de recursos son factores significativos que permiten predecir el rendimiento académico estudiantil; este trabajo señala que si los estudiantes que tienen un bajo desempeño académico se esfuerzan por mejorar estos factores pueden mejorar significativamente.
Waheed et al., (2020)	Predicting academic performance of students from VLE big data using deep learning models	44	Las RNA superan las técnicas de Regresión Logística y Máquina de Vectores de Soporte en términos de su precisión; los autores afirman que la inclusión de datos heredados y datos relacionados con la evaluación, influyen significativamente en la precisión de los modelos. También señalan que los estudiantes interesados en acceder al contenido de clases pasadas tienen un mejor rendimiento académico que los que no demuestran interés.
Yang et al., (2018)	Predicting students' academic performance using multiple linear regression and principal component analysis	30	Los modelos de Regresión Lineal Múltiple (MLR) tienen inconvenientes al momento de predecir el rendimiento académico estudiantil; los autores señalan que el Coeficiente de Determinación R^2, el Error Cuadrático Medio (MSE) y la técnica de diagrama cuantil-cuantil (diagrama Q-Q) no son suficientes para evaluar el rendimiento predictivo ni la precisión de este tipo de modelos.
Nguyen et al., (2018)	Linking students' timing of engagement to learning design and academic performance	29	Existe un desajuste entre la forma en que los docentes diseñan las actividades de aprendizaje y la forma en que los estudiantes estudian en la realidad; en este trabajo se indica que los estudiantes pasan menos tiempo estudiando los materiales asignados en el LMS en comparación con el número de horas recomendadas por los docentes. Los autores también afirman que los estudiantes de alto rendimiento pasan mástiempo estudiando por adelantado, mientras que los de bajo rendimiento dedican una mayor cantidad de tiempo a actividades de recuperación. Esta investigación recalca la importancia del contexto pedagógico para transformar el análisis del aprendizaje en conocimientos prácticos.
Joksimović et al., (2016)	Translating network position into performance: Importance of centrality in different network configurations	28	En cursos virtuales, el uso de redes sociales facilita el proceso de aprendizaje de los estudiantes; el rendimiento académico tiene relación con el grado, la cercanía y la interrelación que tienen los estudiantes con otras personas a través de las mismas.
(Khan & Ghosh, 2021)	Student performance analysis and prediction in classroom learning: A review of educational data mining studies	16	A pesar de que en diversas investigaciones hayan logrado predicciones significativas durante la duración de los cursos virtuales, la predicción del rendimiento antes del inicio de los mismos es una temática que no se ha abordado y requiere una atención especial.

Fuente: elaboración del autor con Biblioshiny para Bibliometrix

El promedio de citas de las publicaciones Top 10 es 47,60, la cantidad máxima de

citas es 86, la mínima es 16. Hay publicaciones más recientes que otras con una mayor cantidad de citas, esto obedece al nivel de innovación y aporte al área de conocimiento, al rigor metodológico y a la validez de los resultados alcanzados.

1.2.2.5 Principales autores

La Tabla 4 presenta los tres autores con la mayor cantidad de citas en orden descendente, el índice H y la cantidad de publicaciones.

Tabla 4. Principales Autores del Área

Autor	Cantidad de Publicaciones relacionadas	Índice H	Cantidad de citas	Principales contribuciones
Osmanbegovic, E.	7	5	391	Compara diferentes métodos y técnicas de minería de datos para la predicción del éxito de los estudiantes empleando información sociodemográfica y académica (Osmanbegovic & Suljic, 2012; Osmanbegović et al., 2014; Osmanbegović, Suljić, et al., 2014).
Ramesh, V.	5	9	385	Identifica los factores que influyen en el rendimiento de los estudiantes y establece orientaciones para que los docentes intervengan efectivamente a los que están en riesgo de abandonar sus estudios (Vamanan et al., 2013; Lakshmi & Ramesh, 2017; Jayanthi & Ramesh, 2014).
Gray, G.	4	7	270	Emplea técnicas de clasificación para la predicción del rendimiento de los estudiantes, establece criterios para el pre-procesamiento de datos y ajuste de algoritmos para resolver los problemas de calidad de los datos (Quinn & Gray, 2019; O'Shaughnessy & Gray, 2011; Gray et al., 2016).

Fuente: elaboración del autor con Biblioshiny para Bibliometrix

1.2.2.6 Principales fuentes

La Tabla 5 presenta el Top 10 de las fuentes con la mayor cantidad de publicaciones. La mayoría de las revistas presentadas en la Tabla 5 publican resultados de investigaciones científicas relacionadas con la enseñanza de la ingeniería, la

educación en general y las ciencias de la computación. Algunas presentan desarrollos de aplicaciones y software para su uso en el área de la educación en la ingeniería.

Tabla 5. Principales Fuentes del Área

Fuentes	Cantidad de Publicaciones	Índice H	Total, de Citas
Lecture Notes in Computers Science	4	400	16
International Journal of Engineering Education	4	50	90
Advances in Intelligent Systems and Computing	4	41	26
International Journal of Engineering Education	4	50	90
Journal of Educational Psychology	3	209	11
Educational Technology and Society	3	88	26
ICCE 2018 - 26th International Conference on Computers in Education, Workshop Proceedings	3	3	4
Interactive Learning Environments	3	44	29
International Journal of Advanced Computer Science and Applications	2	18	35
Procedia Computer Science	2	76	31
Total	32		

Fuente: elaboración del autor con Biblioshiny para Bibliometrix

1.2.3 Discusión

En esta sección se responde la pregunta de investigación planteada anteriormente:

- P1. ¿Qué investigaciones se han desarrollado para predecir el rendimiento académico estudiantil de cursos universitarios?

La Tabla 6 presenta las principales investigaciones encontradas en la revisión de la literatura; se detalla a técnica, las variables, la precisión y el principal aporte al conocimiento de cada una de ellas.

Tabla 6. Caracterización de las Investigaciones sobre la Predicción del Rendimiento Académico

Articulo	Técnica y Precisión obtenida	Variables	Principales contribuciones o Conclusiones del trabajo
(Gray et al., 2014)	Árboles de Decisión - 65% Redes Neuronales Artificiales - 69% K-Vecinos Próximos - 69%	Factores psicométricos	La edad es un factor que influye en la precisión de los modelos predictivos del rendimiento académico, señalan que los modelos que son entrenados con datos de estudiantes jóvenes tienen mayor precisión que los modelos que son entrenados con datos de estudiantes de edad avanzada; además, señalan que los primeros

	Bosques Aleatorios – 68%		modelos pueden ser empleados en una nueva cohorte de estudiantes a diferencia de los segundos modelos. Afirman que el rendimiento académico de los estudiantes que tienen una edad avanzada es más difícil de modelar.
(Bunkar et al., 2012)	Árboles de Decisión - 85%	Información Demográfica y Socioeconómica	Los modelos de Árboles de Decisión identifican con mayor éxito a los estudiantes que tienen mayor probabilidad de suspender un curso y señalan que su precisión puede mejorar si tiene una considerables cantidad de información de los estudiantes.
(Jishan et al., 2015)	Árboles de Decisión - 91%	Promedio Acumulativo de Calificaciones (CGPA)	Cuando el número de registros de un conjunto de datos es pequeño, el sobre-muestreo es inminente; además, señalan que la técnica de Naïve Bayes es computacionalmente más rápida que una RNA (*backpropagation)* y que la precisión de cualquier modelo de predicción mejora significativamente cuando el "Optimal Equal Width Binning" y la técnica de sobremuestreo "Synthetic Minority Over-Sampling" (SMOTE) se utilizan conjuntamente para pre-procesar conjuntos de datos de tamaño pequeñocon variables continuas. También afirman que el nivel de error de clasificación puede minimizarse si se tienen en cuenta atributos como las calificaciones de los estudiantes en los cursos que son prerrequisito.
(Osmanbegovic & Suljic, 2012)	Árboles de Decisión - 73% Naïve Bayes - 76%	Promedio Acumulativo de Calificaciones (CGPA), Demografía de los estudiantes, Formación académica, Becas, Interacción con redes sociales	La identificación y evaluación de variables asociadas al proceso de aprendizaje (no solo variables demográficas) y con un tamaño de muestra considerable, es posible construir modelos que sirvan de base para el desarrollo de sistemas de apoyo para la toma de decisiones en la educación superior.
(Mayilvaganan & Kalpanadevi, 2014)	Árboles de Decisión - 66% K-Vecinos Próximos - 83% Bosques Aleatorios - 83%	Información Demográfica, Socioeconómica, Promedio Acumulativo de Calificaciones (CGPA), Actividades extracurriculares	La técnica de K-Vecinos Próximos tiene la mayor precisión en comparación con otras técnicas y se constituye como una herramienta útil para identificar a los estudiantes lentos, identificar a tiempo fallas que puedan estar cometiendo y tomar acciones que contribuyan a mejorar su rendimiento académico.
(Ramesh et al., 2013)	Árboles de Decisión - 65% Redes Neuronales Artificiales - 72% Naïve Bayes - 50%	Datos demográficos del estudiante, Antecedentes en la escuela secundaria	El tipo de universidad (pública / privada) no influye en el rendimiento académico de los estudiantes, y que la ocupación de los padres si influye. También señalan que factores psicológicos, el nivel de motivación que infunden los profesores y el tipo de materiales de aprendizaje

			electrónico, influyen en el rendimiento académico de los estudiantes.
(Natek & Zwilling, 2014)	Árboles de Decisión - 90%	Evaluación externa, CGPA, Demografía de los estudiantes, Actividades extraescolares	La minería de datos no se limita en general a grandes volúmenes de datos, también pueden emplearse en pequeños conjuntos de datos y obtener resultados útiles. También señalan que su aplicación en la educación superior facilita el desarrollo de sistemas de gestión del conocimiento en universidades e IES.
(Mishra et al., 2014)	Árboles de Decisión - 88%	Factores psicométricos, actividades extracurriculares, habilidades sociales	El cociente emocional (determina el nivel de éxito que se tiene en la vida y la sensación de satisfacción), es una de las variables más importantes para predecir el éxito o el fracaso académico de estudiantes universitarios.
(Pal & Pal, 2013)	Árboles de Decisión - 67%	Información demográfica, económica, académica y familiar	La técnica ID3 es la mejor en términos de precisión y error para clasificar el rendimiento académico estudiantil y señalan que esta técnica tiene el potencial de superar significativamente a otras técnicas de clasificación convencionales empleadas en la predicción del rendimiento académico estudiantil.
(Imran et al., 2019)	Árboles de Decisión - 83%	Información demográfica, social, familiar y académica	La importancia del pre-procesamiento de datos y el ajuste de las técnicas de clasificación para resolver problemas de calidad de los datos y señalan que los Árboles de Decisión (J48) superan en precisión a otras técnicas empleadas en la predicción del rendimiento académico estudiantil.
(Quadri y Kalyankar, 2010)	Árboles de Decisión - 80%	Información demográfica, social, familiar y académica	Las bases de datos de los estudiantes, si se gestionan, analizan y explotan adecuadamente, son activos únicos y valiosos para las universidades. También señalan que se pueden desarrollar yaplicar estrategias que permitan a las universidades procesar esta valiosa información en beneficios tangibles.
(Li et al., 2013)	Árboles de Decisión - 86%	Información demográfica, social, familiar y académica	Los programas de ingeniería a nivel mundial tienen una tasa de deserción relativamente alta, por lo general, alrededor del 35% de los estudiantes de primer año en varios programas de ingeniería no llegan al segundo año y el resto de estudiantes, a menudo abandonan o fracasan en su segundo o tercer año de estudios. Señalan que si los predictores de desempeño de los estudiantes se identifican de manera temprana, se pueden usar de manera efectiva en la formulación de planes de acción correctiva para disminuir la mortalidad y las tasas de deserción.
(Lau et al., 2019)	Árboles de Decisión - 91%	Información demográfica, Antecedentes académicos, información familiar	Las RNA son una técnica útil para la clasificación del rendimiento académico de los estudiantes y afirman que éstas pueden presentar un bajo rendimiento cuando las clases están desbalanceadas. Afirman que estas deficiencias pueden mitigarse

			mediante la introducción de nuevas muestras con un tipo de población muestral más equilibrada.
(Arsad et al., 2013)	Redes Neuronales Artificiales - 97%	Promedio Acumulativo de Calificaciones (CGPA), Historial Académico Previo, Evaluaciones externas	El rendimiento de los estudiantes en cursos previos, específicamente en cursos de primeros semestres influyen de manera directa en el rendimiento académico de cursos de semestres más avanzados.
(Jishan et al., 2015)	Redes Neuronales Artificiales - 75% Naïve Bayes - 75%	Promedio Acumulativo de Calificaciones (CGPA)	Para mejorar la precisión de los modelos de predicción del rendimiento de los estudiantes es conveniente pre-procesar los datos utilizando un método de discretización llamado "Optimal Equal Width Binning" y la técnica de sobremuestreo "Synthetic Minority Over-Sampling" (SMOTE).
(Osmanbegovic & Suljic, 2012)	Redes Neuronales Artificiales - 71%	Promedio Acumulativo de Calificaciones (CGPA), Datos demográficos del estudiante, antecedentes de la escuela secundaria, becas, interacción de la red social	La técnica de clasificación Naïve Bayes supera en precisión a las técnicas de Árboles de Decisión y RNA, y señalan que un modelo clasificación debe ser a la vez preciso y comprensible para los docentes. También recalcan la necesidad de establecer criterios que permitan a los docentes conseguir que los modelos de predicción sean fáciles de usar y enfatizan en la necesidad de crear herramientas que faciliten la recolección de datos de los estudiantes.
(Oladokun et al., 2008)	Redes Neuronales Artificiales - 74%	Evaluación externa, Estudiante Demográfico, Antecedentes en la escuela secundaria	Las RNA tienen el potencial para superar en precisión otras técnicas de clasificación y señalan que una de las principales limitaciones de esta técnica para predecir el rendimiento académico estudiantil obedece a que no todos los factores que son relevantes son considerados.
(Kumar et al., 2012)	Redes Neuronales Artificiales - 98%	Evaluaciones internas y externas	Los estudiantes con habilidades de aprendizaje autorregulado obtienen un alto rendimiento académico y mejores puestos de trabajo que los que no y afirman que las técnicas de clasificación son útiles no sólo para predecir el rendimiento académico estudiantil, sino también para analizar el comportamiento de los estudiantes.
(Moucary et al., 2011)	Redes Neuronales Artificiales - 79%	Información demográfica, socioeconómica y familiar	Las RNA tienen un alto nivel de precisión y eficiencia en la identificación de estudiantes lentos, moderados y rápidos, y señalan que éstas también se constituyen como una herramienta potente para orientar a los docentes en el análisis de la trayectoria académica de los estudiantes.
(S. K. Yadav & Pal, 2012)	Redes Neuronales Artificiales - 76%	Información demográfica, socioeconómica y familiar	Los modelos predictivos del rendimiento académico que emplean las técnicas de RNA pueden emplearse en la predicción del rendimiento de estudiantes nuevos que ingresan a la universidad, lo que facilita la identificación temprana de los que necesitan una atención especial.

(Mayilvaganan & Kalpanadevi, 2014)	Naïve Bayes - 73%	Evaluación interna, Promedio Acumulativo de Calificaciones (CGPA), Actividades extracurriculares	La técnica de K-NN multi-etiquetada tiene el mejor rendimiento en cuanto a tiempo de clasificación en comparación con otras técnicas empleadas para predecir el rendimiento académico estudiantil y señalan la importancia de considerar los resultados de exámenes y actividades que desarrollen los estudiantes durante el curso.
(Vamanan et al., 2013)	Naïve Bayes - 72%	Datos demográficos del estudiante, antecedentes de la escuela secundaria, becas, interacción de la red social	Los hábitos de estudio de los estudiantes, las relaciones sociales y factores económicos, son los tres factores que más influyen en su rendimiento académico.
(Minaei-Bidgoli et al., 2003)	K-Vecinos Próximos - 82%	Evaluación interna	La combinación de múltiples técnicas de clasificación conduce a una mejora significativa en los resultados de las predicciones del rendimiento académico de los estudiantes, ya que se mejora la precisión; esto es útil para identificar a los estudiantes en riesgo temprano, especialmente en cursos con una gran cantidad de estudiantes y también para permitir que los docentes brinden un asesoramiento adecuado de manera oportuna.
(Sembiring et al., 2011a)	Máquinas de Vectores de Soporte - 83%	Factores psicométricos	La Máquina de Vectores de Soporte tiene una buena capacidad de generalización y afirman que existen una fuerte correlación entre la salud mental del estudiante y su rendimiento académico.
(Mahmoud Abu Zohair, 2019)	Máquinas de Vectores de Soporte - 89%	Información Demográfica, Condición Socioeconómica, Estado Educativo de los padres, Actividades Extracurriculares, y Comportamientos de los Estudiantes	El principal problema en tareas de clasificación es el tamaño pequeño de los conjuntos de datos de los estudiantes; sin embargo, afirman que con un pre-procesamiento cuidadoso y una selección de paradigmas de modelado adecuadas se pueden entrenar clasificadores con un alto nivel de precisión.
(Ünal, 2020)	Bosques Aleatorios - 91%	Información Demográfica, Condición Socioeconómica, Estado Educativo de los padres, Actividades Extracurriculares, y Comportamientos de los Estudiantes	Es importante considerar factores internos y externos de los estudiantes y no solo información académica (estado demográfico, condición socioeconómica, estado educativo de los padres, actividades extracurriculares y comportamientos de los estudiantes); también señalan que la técnica de Bosques Aleatorios y el método WrapperSubsetEval para la selección de atributos son los más indicados para esta tarea.
(Parack et al., 2012)	Bosques Aleatorios - 78%	Información Demográfica, Socioeconómica y Familiar	La técnica de Bosques Aleatorios es adecuada no solo para predecir el rendimiento académico de los estudiantil, sino también para ayudar a predecir la carrera adecuada a los estudiantes de media vocacional, lo que puede ayudar a

			las instituciones a ofrecer un adecuado asesoramiento y orientar mejor a los estudiantes en su orientación profesional.

Fuente: elaboración del autor

1.3 Definición del problema de investigación

1.3.1 Justificación

De la RSL se evidencia que la predicción del rendimiento de los estudiantes es muy útil para ayudar a los docentes y estudiantes a mejorar el proceso de enseñanza y aprendizaje;la mayoría de los investigadores han utilizado el promedio acumulativo de calificaciones (CGPA) y la evaluación interna como conjuntos de datos de entrenamiento; sin embargo, en diversas investigaciones, el CGPA es el principal atributo para predecir el rendimiento académico de los estudiantes (Angeline, 2013; Quadri y Kalyankar, 2010; Osmanbegovic y Suljic, 2012); la idea principal de por qué la mayoría de los investigadores utilizan el CGPA es porque tiene un valor tangible para la futura movilidad educativa y profesional; además, puede considerarse como un indicador del potencial académico de los estudiantes (Bin Mat et al., 2013). Según Ibrahim y Rusli, (2007), el CGPA es la variable más significativa en comparación con otras variables. En la investigación desarrollada por Christian y Ayub, (2014), el CGPA fue el atributo que más influyó en la determinación de la permanencia de los estudiantes en sus estudios, es decir, determinó si podían completar o no sus estudios.

En (Parack et al., 2012) y (Naren, 2014), las variables que se emplearon para predecir el rendimiento estudiantil fueron: evaluación demográfica de los estudiantes y evaluaciones externas. La demografía estudiantil incluía: género, edad, antecedentes familiares y discapacidades (Bin Mat et al., 2013; Osmanbegovic y Suljic, 2012); mientras que la evaluación externa fue determinada por la puntuación obtenida en el examen final de un curso en particular (Angeline, 2013; Parack et al., 2012). La razón por la que la mayoría de los investigadores utilizan información demográfica como el género, obedece a que los hombres tienen estilos de aprendizaje diferentes a las mujeres (Bin Mat et al., 2013); en la investigación realizada por (Meit et al., 2004), se encontró que las mujeres tenían mejores estilos de aprendizaje y comportamientos más positivos en comparación con los hombres. Según Simsek y Balaban, (2010), las mujeres son más disciplinadas y obedientes en sus estudios, auto-dirigidas, preservadas y enfocadas, adoptan estrategias de aprendizaje más

efectivas, tienen mayor automotivación, organización y disciplina; por lo tanto, el género es una de las variables más importantes que influyen en el desempeño de los estudiantes.

En otras investigaciones, las variables empleadas para predecir el desempeño de los estudiantes fueron: actividades extracurriculares (Mayilvaganan y Kalpanadevi, 2014; Natek y Zwilling, 2014; Mishra et al., 2014), antecedentes en la escuela secundaria (Osmanbegovic y Suljic, 2012; Oladokun et al., 2008; Abu Tair y El-Halees, 2012) e interacciones sociales (Romero et al., 2013 ; Bogarín et al., 2014). En las investigaciones desarrolladas por Sembiring et al., (2011) y (Gray et al., 2014), se utilizó el factor psicométrico para predecir el rendimiento de los estudiantes; un factor psicométrico es identificado como el interés del estudiante, el comportamiento a la hora de estudiar, el tiempo de estudio personal y el apoyo de la familia (Brito et al., 2019). Esta información ayuda a evaluar los logros de los estudiantes en base a su interés personal y a su comportamiento; sin embargo, estos atributos raramente se aplican para predecir el rendimiento de los estudiantes porque se centran más en datos cualitativos y también es difícil obtener datos válidos de los encuestados (Mayilvaganan y Kalpanadevi, 2014).

Se debe aclarar que los cursos universitarios presenciales que se desarrollan al interior de un campus universitario permiten que los docentes tengan la posibilidad de interactuar más con los estudiantes, conocer aspectos personales, familiares, económicos, psicosociales y demás, lo que hace posible construir modelos predictivos que contemplen información adicional a la académica y obtener predicciones más precisas. En cursos universitarios virtuales a distancia (CUVD) híbridos (encuentros sincrónicos y actividades asincrónicas) en los cuales la interacción entre los docentes y los estudiantes se realiza a través de mediaciones tecnológicas, es complejo que los docentes conozcan información relacionada con problemas que estén teniendo los estudiantes y que puedan afectar su desempeño académico (problemas económicos, sentimentales, laborales, familiares, etc.), lo que conlleva a que este tipo de información no sea considerada en la construcción de los modelos predictivos.

En la mayoría de universidades e IES existe una ausencia de sistemas de información o plataformas orientadas a la web que permitan a los estudiantes de CUVD reportar problemáticas personales; debido a la modalidad de este tipo de cursos solo un bajo porcentaje de estudiantes se dirige de manera voluntaria a las unidades de bienestar universitario o permanencia académica para solicitar apoyo en caso de que se les presente

algún tipo de dificultad personal. Debido a lo anterior, es importante que las universidades e IES creen sistemas de información que permitan a los docentes de este tipo de cursos consultar información adicional a la académica de los estudiantes, ya que al conocer información adicional a la que se puede obtener directamente de una plataforma educativa virtual (PEV) se puede mejorar la precisión de las predicciones del rendimiento académico de los estudiantes, identificar con mayor facilidad a los que necesitan apoyo y formular estrategias en pro de prevenir la mortalidad y la deserción estudiantil.

Por otra parte, son diversas las plataformas educativas virtuales (PEV) empleadas por universidades e IES a nivel mundial en el desarrollo de CUVD. Rivas et al., (2021), identificaron las cuatro PEV más utilizadas en la actualidad (Moodle, Blackboard, Edmodo y Schoology); estos autores analizaron seis aspectos de las PEV: si eran de código abierto, si tenían disponibilidad de APIs para desarrolladores e integración con sistemas de comercio electrónico, cantidad de usuarios a nivel mundial, compatibilidad con herramientas de aprendizaje con interoperabilidad y si contaban con plugins o extensiones para la predicción del rendimiento académico estudiantil, estos resultados se presentan en la Tabla 7.

Tabla 7.Comparativo Plataformas Educativas Virtuales

	Moodle	Blackboard	Edmodo	Schoology
Código abierto	Sí	Sí	Sí	Sí
API para desarrolladores	Sí	Sí	Sí	Sí
Apoyo a LTI	Sí	Sí	No	Sí
Comercio electrónico	Sí	No	Sí	No
Usuarios	124 M	25 M	58 M	20 M
Predicción del rendimiento estudiantil	No	No	No	No

Fuente: Adaptado de (Capterra, 2020) y (Moodle, 2020).

Rivas et al., (2021) encontraron que ninguna PEV empleaba herramientas o plugin que permitiera hacer la predicción del rendimiento estudiantil para que los docentes de CUVD pudieran identificar qué estudiantes aprobarán o reprobarán; además, identificaron que el único recurso que podía ser utilizado para ello eran los registros (Logs). Si bien es cierto que los docentes pueden analizar los registros de los estudiantes, construir modelos predictivos y sacar conclusiones a partir de los registros (Logs), en la revisión del estado del arte no se evidencia la existencia de una metodología universal que indique como analizar, agrupar y procesar esta información, de modo que permita contemplar información

adicional a la relacionada con la interacción de los estudiantes con las PEV y que permita a los docentes identificar a los estudiantes que tienen mayor riesgo de reprobar garantizando la calidad de la predicción de los modelos.

Debido a lo anterior, para la construcción de esta metodología es de vital importancia identificar los factores clave que contribuyen a un bajo rendimiento académico estudiantil y no enfocarse únicamente en información académica de los estudiantes (Logs), también considerar aspectos personales, socioeconómicos, hábitos de estudio, ambientales, sociales y tecnológicos de los estudiantes; además, es importante establecer la manera como se debe determinar la técnica adecuada para construir los modelos predictivos, ya que no es correcto afirmar que una técnica sea superior a otras en términos de precisión, dado que la cantidad y tipo de información que se tenga de los estudiantes desempeña un papel de gran importancia al momento de seleccionar el modelo más adecuado entre un conjunto de técnicas. A continuación, se presenta la pregunta y la hipótesis de investigación de este trabajo.

1.3.2 Problema

El problema que se identificó a partir de los resultados de la RSL es que no existe una metodología universal que permita a los docentes de CUVD de diferentes áreas de conocimiento predecir el rendimiento académico de los estudiantes de manera temprana para que identifiquen a los que están en riesgo de reprobar e implementen estrategias y medidas correctivas a tiempo para evitar la mortalidad y la deserción los mismos. A pesar de que diversas investigaciones han empleado técnicas de ML para predecir el rendimiento académico estudiantil en cursos universitarios y han alcanzado altos niveles de precisión, entre las que se destacan los Árboles de Decisión y las RNA, estos son ejercicios aislados que no establecen un procedimiento formal y estándar; tampoco establecen criterios orientados a mejorar la precisión de las predicciones, y solo una baja cantidad de trabajos considera que el desempeño académico de los estudiantes puede depender de diversos factores (internos y externos), más allá de los factores académicos (calificaciones finales, promedio acumulativo, notas obtenidas en actividades, pruebas y trabajos); éste también depende de factores relacionados con la vida personal.

La información académica y la relacionada con factores externos e internos de los estudiantes pueden ser estudiadas y analizadas en conjunto para identificar con mayor

precisión a los estudiantes de CUVD que tienen mayor riesgo de reprobar. Además, estas investigaciones se enfocan principalmente en cursos universitarios presenciales, en los cuales existe una interacción presencial entre los docentes y los estudiantes, y no consideran los CUVD que se realizan a través de mediaciones tecnológicas (clases sincrónicas y actividades asincrónicas), en los cuales el proceso de enseñanza aprendizaje no es igual al de los cursos presenciales (Talebian et al., 2014; Brasche & Harrington, 2012; Orduña et al., 2012).

En los CUVD, la interacción de los estudiantes con docentes y compañeros de curso es más baja comparada con la modalidad presencial (Radović-Marković, 2010), razón por la cual el rendimiento académico de los estudiantes puede ser menor e inferior al de cursos presenciales (Dung, 2020; Pakhomova et al., 2021), lo que contribuye al incremento de la mortalidad y la deserción académica; además, hay una mayor cantidad de fraude debido a la baja implementación de controles por parte de la universidades que impidan la suplantación de identidad en pruebas evaluativas y la tercerización de trabajos o actividades (Erokhina & Anikina, 2020).

1.3.3 Pregunta de investigación

A partir de lo mencionado anteriormente, la pregunta de investigación en esta tesis doctoral es:

> ¿Es posible desarrollar una metodología para predecir el desempeño académico (Aprueba / Reprueba) estudiantil en cursos universitarios virtuales a distancia (CUVD)?

1.3.4 Hipótesis

La hipótesis de esta tesis doctoral es:

> Es posible construir una metodología para predecir el desempeño académico (Aprueba / Reprueba) estudiantil en cursos universitarios virtuales a distancia (CUVD).

1.4 Objetivos

Esta sección presenta el objetivo general y específicos que se esperan alcanzar con el desarrollo de esta investigación.

1.4.1 Objetivo General

Desarrollar una metodología para predecir el desempeño académico (Aprueba / Reprueba) estudiantil en cursos universitarios virtuales a distancia (CUVD).

1.4.2 Objetivos específicos

Los objetivos específicos de esta tesis de doctorado son los siguientes:

1. Determinar los factores clave que causan un bajo rendimiento académico estudiantil en CUVD.
2. Determinar la técnica adecuada para predecir el rendimiento académico estudiantil en CUVD
3. Construir un modelo para predecir el rendimiento académico estudiantil en CUVD.

1.5 Aportes y contribuciones

Esta tesis doctoral pretende contribuir en esta área de conocimiento en los siguientes aspectos:

- Creación de una metodología estandarizada que permita a los docentes de CUVD predecir el rendimiento académico de sus estudiantes en cualquier momento de avance del curso.
- Identificación de los factores que contribuyen a que los estudiantes de este tipo de cursos tengan un bajo rendimiento académico.
- Creación de una herramienta virtual que permita al personal administrativo y a los docentes emplear la metodología propuesta directamente desde el sistema de gestión de aprendizaje (LMS) Moodle.

- Establecimiento de orientaciones para seleccionar la mejor técnica para predecir el rendimiento académico de los estudiantes de CUVD.
- Formulación de consejos y orientaciones para que los docentes de CUVD eviten el bajo rendimiento académico y la deserción en CUVD.

1.6 Mapa del documento

El resto de este documento está organizado de la siguiente manera: el Capítulo 2 presenta las definiciones y conceptos básicos. El Capítulo 3 presenta los resultados de una RSL enfocada a determinar los factores que causan un bajo rendimiento académico en cursos universitarios virtuales a distancia (CUVD). El Capítulo 4 presenta los resultados de una RSL enfocada a determinar las técnicas que se han empleado en la predicción del rendimiento académico en CUVD. El Capítulo 5 presenta la metodología propuesta, la cual está conformada por 5 pasos (determinación de las variables a analizar, construcción de la base de datos, construcción de los modelos de predicción, evaluación de los modelos y visualización de la predicción). El Capítulo 6 presenta el caso de aplicación de la metodología con bases de datos de estudiantes de CUVD de universidades de Inglaterra, Uruguay y Colombia, y por último en el Capítulo 7 se presentan las conclusiones obtenidas.

2 Factores que causan un bajo rendimiento académico estudiantil en los CUVD

En el Capítulo 1 se presentó el proyecto de tesis doctoral y se plantearon sus objetivos. En este capítulo se dará desarrollo al primer objetivo específico que corresponde a:

> Determinar los factores clave que causan un bajo rendimiento académico estudiantil en CUVD.

Para dar cumplimiento a este objetivo se realizó una RSL enfocada a responder la siguiente pregunta de investigación:

> ¿Qué factores contribuyen al bajo rendimiento académico estudiantil en CUVD?

2.1 Introducción

El bajo rendimiento académico es un tema común en la mayoría de universidades e IES a nivel mundial; son diversos los factores que pueden contribuir a que los estudiantes de los CUVD tengan un bajo rendimiento; como se mencionó anteriormente, en cursos universitarios presenciales (CUP) es más factible que los docentes conozcan información de los estudiantes relacionada con problemáticas o situaciones personales que estén perjudicando su rendimiento académico, ya que interactúan de manera presencial; además, en los CUP es más fácil que los estudiantes acudan a las dependencias de bienestar universitario o permanencia académica de manera voluntaria en caso de tener algún tipo de dificultad, ya que los cursos son desarrollados al interior de un campus universitario. En los CUVD es más complejo y difícil conocer este tipo de información, ya que los docentes únicamente cuentan con la información relacionada con la interacción de los estudiantes con las plataformas educativas virtuales (PEV) y no disponen de plataformas o sistemas de información estudiantil que les permitan consultar problemáticas o situaciones personales que estén afectando el rendimiento académico de los estudiantes. Este capítulo presenta una RSL enfocada a determinar los principales factores que pueden contribuir al bajo

rendimiento estudiantil en los CUVD; en las siguientes secciones se detalla la metodología empleada, los resultados y la discusión.

2.2 Metodología

A continuación, se definen los pasos utilizados para ejecutar la RSL:

- Bases de datos utilizadas: Scopus, IEEE Xplore e ISI Web of Knowledge WoS.
- Período de consulta: toda la información disponible hasta mayo de 2021.
- Cadena de búsqueda:

  TITLE-ABS-KEY (Low AND Perform* AND (Academic OR University OR Virtual Courses OR Virtual Education))

- Criterios de inclusión: que fueran investigaciones que especificaran los factores que contribuyen al bajo rendimiento académico estudiantil en cursos universitarios virtuales a distancia, que explicaran el porqué de esos factores y que respaldaran sus hallazgos a partir de bases de datos reales.
- Criterios de exclusión: que se enfocaran en cursos de básica secundaria y media vocacional, que no explicaran el porqué de esos factores y que no respaldaran sus hallazgos a partir de bases de datos reales.

2.3 Resultados

Al aplicar la cadena de búsqueda en las bases de datos seleccionadas se recuperaron automáticamente 35 documentos; al aplicar los criterios de inclusión y exclusión se descartaron 18 documentos, para así obtener un total de 17 documentos para ser utilizados en esta RSL.

2.3.1 Métricas generales de los trabajos seleccionados

La Tabla 8 presenta las principales métricas de los resultados obtenidos, la mayoría de publicaciones fueron realizadas entre varios investigadores, solo una minoríafueron realizadas por un solo autor, el promedio de citas por documento es superior a11, lo que indica que estas investigaciones han sido citadas en esta área de conocimiento.

Tabla 8. Investigaciones sobre Predicción del Rendimiento Académico

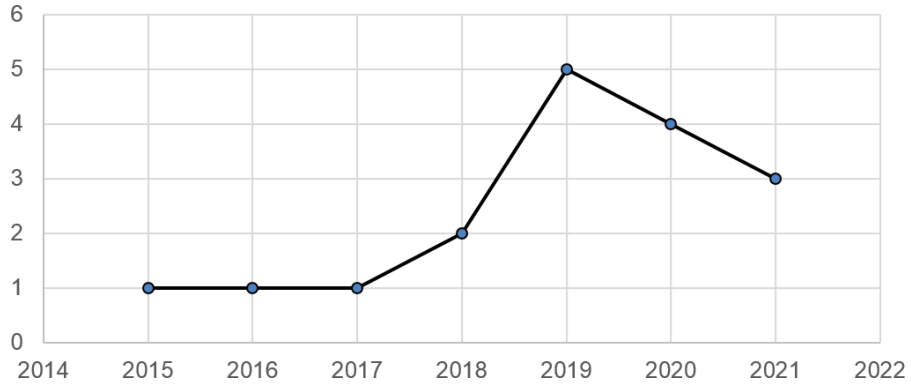
Información Principal Sobre los Datos	
Período de publicación	2004:2021
Cantidad	17
Promedio de citas por documento	11,54
Promedio de citas por año por documento	1,24
Referencias	1.168
Palabras clave	448
Autores	
Publicaciones con un solo autor	4
Publicaciones con varios autores	13
Colaboración de Autores	
Promedio de autores por documento	2,94
Promedio de coautores por documento	2,94
Índice de colaboración	3,19

Fuente: elaboración del autor con Biblioshiny para Bibliometrix

2.3.2 Cantidad de publicaciones por año

De la Figura 2 se puede evidenciar un crecimiento no continuo en la cantidad de investigaciones, en el año 2019 se registró la mayor cantidad (5 publicaciones), la caída en el año 2021 obedece a que solo se tiene información parcial para este.

Figura 2. Publicaciones por Año para la RSL del Capítulo 3.



Fuente: elaboración del autor con Biblioshiny para Bibliometrix

2.3.3 Países con mayor producción

La Tabla 9 presenta los tres países con la mayor cantidad de publicaciones y su principal contribución en el área de conocimiento.

Tabla 9. Top 3 de los países con la mayor cantidad de publicaciones y su principal contribución

País	Artículos	Principal contribución
EEUU	24	Las investigaciones desarrolladas se han enfocado en identificar los factores individuales de los estudiantes de cursos universitarios presenciales y virtuales que se relacionan con el bajo rendimiento académico, con la deserción y con el retraso en la graduación; las investigaciones señalan que factores individuales (el género, la edad, los ingresos familiares, el nivel educativo de los padres, la calidad de la docencia y el nivel de satisfacción), antecedentes académicos y factores ambientales, juegan un papel en la explicación de la graduación tardía y en el bajo rendimiento académico. También hacen hincapié en que el apoyo constante del personal académico a los estudiantes tiene un efecto positivo en el rendimiento académico de los mismos (Baggs et al., 2015; Chang & Brickman, 2018; Miranda et al., 2013; Miskioglu, 2016).
Reino Unido	14	Los trabajos se enfocan en la identificación de los factores internos y externos que contribuyen a que los estudiantes de los CUVD tengan un bajo rendimiento, recalcan la importancia de considerar los registros académicos de los estudiantes. Señalan que el principal problema de los modelos predictivos radica en el hecho de que los estudiantes difieren enormemente en términos de antecedentes y cursos seleccionados; además, los CUVD no son igualmente informativos para hacer predicciones precisas y el progreso evolutivo de los estudiantes no se incorpora en la predicción (Al-Sudani & Palaniappan, 2019; Brook & Roberts, 2021; Bussu et al., 2020; Malau-Aduli, 2011)
Australia	12	Al igual que en los otros países los investigadores se enfocan en determinar los factores que causan un bajo rendimiento y una baja motivación en los estudiantes durante el desarrollo de los cursos virtuales; particularmente hacen hincapié en que el uso de chats y foros no son suficientes para motivar a los estudiantes en el desarrollo de actividades relacionadas con elproceso de aprendizaje. Proponen estrategias de motivación basadas en la gamificación y el constructivismo y señalan que el uso de las TIC en el ámbito educativo no es solo la forma de contactar con el estudiante, es el medio para desarrollar enfoques pedagógico-didácticos que finalmente enriquezcan el proceso educativo y en consecuencia motiven al estudiante a desempeñarse bien (David et al., 2016; Hudson & Treagust, 2013; Malau-Aduli, 2011; Papageorgiou & Halabi, 2014)

Fuente: elaboración del autor con Biblioshiny para Bibliometrix

2.3.4 Artículos más citados

La Tabla 10 presenta el Top 10 de las publicaciones con la mayor cantidad de citas y

su principal contribución al área de conocimiento.

Tabla 10. Principales investigaciones y su contribución

Autores y Año	Título	Citas	Contribución / Conclusión
(Q. Nguyen et al., 2018)	Linking students' timing of engagement to learning design and academic performance	140	Los estudiantes no estudian siguiendo la ruta de aprendizaje que recomiendan los docentes; existen discrepancias entre la manera como los docentes diseñan las actividades de aprendizaje y la manera como los estudiantes lo hacen en la realidad. A pesar de que todos los estudiantes de cursos virtuales tienen acceso a los mismos recursos en las PEV, el tiempo promedio de estudio es inferior al número de horas estipulado por los docentes; los estudiantes con un rendimiento excelente dedican más tiempo que los que tienen un rendimiento bueno, aceptable e insuficiente; razón por la cual los docentes deben examinar y ajustar con frecuencia sus prácticas de enseñanza, eliminar materiales de estudio que se utilicen poco y no diseñar las actividades evaluativas como eventos aislados.
(Stephens, Brannon, et al., 2015)	Feeling at home in college: Fortifying school-relevant selves to reduce social class disparities in higher education	57	Los docentes y académicos deben crear estrategias para fomentar la motivación y el rendimiento académico de los estudiantes de cursos virtuales que trabajan, para evitar que deserten de sus estudios y tengan oportunidades de salir adelante.
(Patterson & Patterson, 2017)	Computers and productivity: Evidence from laptop use in the college classroom	27	El rendimiento académico de los estudiantes que utilizan ordenadores portátiles en cursos virtuales es mayor que el rendimiento de los estudiantes que usan ordenadores de escritorio, principalmente porque los ordenadores portátiles permiten la multi-ubicuidad de los estudiantes al momento de estudiar.
(Correa-Burrows et al., 2016)	Nutritional quality of diet and academic performance in Chilean students	23	El consumo excesivo de alimentos ricos en energía, bajos en fibra y ricos en grasas se relacionan con un bajo rendimiento académico.
(Perin, 2006)	Academic progress of community college nursing aspirants: An institutional research profile	22	Muchos estudiantes de nivel socioeconómico bajo que ingresan a programas de formación profesional en la modalidad virtual tienen bajos niveles de preparación académica; para abordar las serias dificultades que ésta problemática genera a las universidades y para poder garantizar la retención y un buen rendimiento académico de este tipo de estudiantes, se recomienda un acompañamiento continuo por parte de los docentes durante los primeros semestres académicos.
(Napoli & Raymond, 2004)	How reliable are our assessment data?: A comparison of the reliability of data produced in graded and un-graded conditions	20	La evaluación inconsecuente llevada a cabo por separado del proceso real de enseñanza-aprendizaje en cursos universitarios, es un ejercicio quijotesco y poco realista que no provoca una respuesta de máximo esfuerzo

			por parte de los estudiantes; como consecuencia, los resultados obtenidos por los estudiantes en estas pruebas no son un buen indicador de los verdaderos conocimientos y capacidades que tienen.
(T. L. H. Nguyen, 2013)	Middle-level academic management: A case study on the roles of the heads of department at a Vietnamese university	15	La función de los jefes de áreas académicas en las universidades debe estar enfocada no solo a la estructuración y formulación de contenidos, sino también a la supervisión de la ejecución de los cursos, razón por la cual las directivas de las universidades deben conferirles mayor cantidad de tiempo y autonomía para esta tarea.
(Black et al., 2015)	Can you leave high school behind?	12	La calidad de la formación en la escuela secundaria tiene una estrecha relación con el rendimiento académico en la educación superior; la influencia de la escuela secundaria se prolonga mucho más allá de la graduación. A pesar de que las universidades ofrecen programas para superar las desigualdades entre los estudiantesadmitidos como: programas de acompañamiento, tutorías especializadas, grupos de estudio, entre otros; éstas carecen de herramientas para diagnosticar su rendimiento académico.
(Rodríguez Ayán & Ruiz Díaz, 2011)	Indicadores de rendimiento de estudiantes universitarios: calificaciones versus créditos acumulados	12	Se debe motivar a los estudiantes a leer las lecciones de un curso antes del comienzo del mismo; cuando estos comprenden la causa, el efecto y la solución de los problemas comunes, pueden trabajar de forma más productiva, lo que conduce a una mejor experiencia de aprendizaje.
(Smith et al., 2011)	Enhancing undergraduate education in aerospace engineering and planetary sciences at MIT through the development of a CubeSat mission	11	Los estudiantes de cursos universitarios deben aprender a autoevaluarse, a desarrollar la capacidad de reconocer cuándo están haciendo bien o mal las cosas, a identificar cuándo su rendimiento está por debajo de lo esperado; además, muchos estudiantes no son capaces de conectar las temáticas de clases anteriores con temas que desarrollan en sus cursos, razón por la cual, los docentes deben diseñar tareas que impliquen relacionar explícitamente lo que han aprendido en cursos pasados.
(Phan et al., 2018)	A Study of the Effects of Daily Physical Activity on Memory and Attention Capacities in College Students	8	La capacidad de atención y memorización de los estudiantes universitarios puede verse afectada por la falta de actividad física, esta problemática es generada principalmente por el sedentarismo ocasionado por el uso excesivo de dispositivos tecnológicos.

Fuente: elaboración del autor con Biblioshiny para Bibliometrix

El promedio de citas de las publicaciones Top 10 es 31,54, la cantidad máxima de citas es 140, la mínima es 8. Hay publicaciones más recientes que otras con una mayor cantidad de citas, esto obedece al nivel de innovación y aporte al área de

conocimiento, al rigor metodológico y a la validez de los resultados alcanzados.

2.3.5 Principales autores

La Tabla 11 presenta los tres autores con la mayor cantidad de citas en orden descendente, el índice H, la cantidad de publicaciones y la principal contribución al área de conocimiento.

Tabla 11. Principales autores y su contribución

Autor	Cantidad de Publicaciones relacionadas	Índice H	Cantidad de citas	Contribución
Nguyen, Q.	6	14	624	Los datos de las PEV y la información demográfica de los estudiantes son identificadores efectivos de su desempeño académico; sin embargo, las predicciones de los modelos que solo emplean este tipo de información no son adecuadas ni suficientes para formular intervenciones pedagógicas efectivas, debido a que no tienen la capacidad de explicar por qué los estudiantes tienen patrones de comportamiento inadecuados, como la procrastinación y elaburrimiento (Deo et al., 2020; Q. Nguyen et al., 2018; D. Tempelaar et al., 2018; D. T. Tempelaar et al., 2017)
Stephens, N.	4	26	605	Los estudiantes de bajos ingresos económicos que trabajan o que no tienen padres con títulos universitarios son los más propensos a fracasar y abandonar sus estudios; las universidades deben formular planes de intervención para apoyar a este tipo de estudiantes (Stephens et al., 2014; Stephens, Brannon, et al., 2015; Stephens, Townsend, et al., 2015)
Helal, S	2	41	577	Los estudiantes, que tienen antecedentes socioeconómicos de pobreza extrema o que han sido admitidos con base en requisitos especiales de entrada o que tienen un bajo nivel de acceso a los recursos y a los foros del curso tienen más probabilidades de fracasar (Helal et al., 2018, 2019)

Fuente: elaboración del autor con Biblioshiny para Bibliometrix

2.3.6 Principales fuentes

La Tabla 12 presenta el Top 10 de las fuentes con la mayor cantidad de publicaciones, la mayoría de ellas publican resultados de investigaciones científicas relacionadas con

la enseñanza de la ingeniería, la educación en general y las ciencias de la computación. Algunas presentan desarrollos de aplicaciones y software para su uso en el área de la educación en la ingeniería.

Tabla 12. Principales fuentes

Fuentes	Cantidad de Publicaciones relacionadas	Índice H	Total, de Citas
IEEE Access	3	127	43
Economics of Education Review	2	85	147
Education and Information Technologies	2	41	21
Assessment and Evaluation in Higher Education	2	81	56
Journal of Information Technology Education: Research	2	19	30
Informatics in Education	2	19	14
Journal of Educational and Behavioral Statistics	1	59	14
International Journal of Educational Research	1	63	9
Educational Sciences: Theory and Practice	1	20	11
International Review of Research in Open and Distance Learning	1	68	8
Total	17		

Fuente: elaboración del autor con Biblioshiny para Bibliometrix

2.4 Discusión

A continuación, se responde la pregunta de investigación planteada al inicio de este capítulo que es:

> ¿Qué factores contribuyen al bajo rendimiento académico estudiantil en cursos universitarios virtuales a distancia?

La Tabla 13 presenta los principales factores que contribuyen al bajo rendimiento académico estudiantil en los CUVD. En dicha tabla se puede observar que son múltiples los factores que influyen en el desempeño académico de los estudiantes de los CUVD, y estos están relacionados con aspectos académicos y personales; diversas investigaciones coinciden en que el bajo rendimiento académico en este tipo de cursos obedece principalmente a factores socioeconómicos; la mayoría de estudiantes de CUVD trabajan, son cabeza de familia y tienen un mayor promedio de edad, estos factores condicionan directa o indirectamente su rendimiento académico. Sumado a lo anterior, el bajo efecto o influencia que ejercen las universidades bajo la modalidad virtual en comparación con la modalidad presencial es un factor de gran relevancia; bajo un entorno educativo virtual es

más complejo que los estudiantes de los CUVD comprendan y adquieran nuevo conocimiento sin ayuda del docente, ya que muchos se cohíben de preguntar sus inquietudes, se dispersan en el desarrollo de los encuentros sincrónicos y no vuelven a acceder o consultar las grabaciones de los mismos (Al-Nofaie, 2020), por lo cual, es esencial que los docentes realicen intervenciones de apoyo oportunas y efectivas.

Tabla 13. Factores que causan un bajo rendimiento académico en los CUVD

Tipo de Factor	Fuente	Descripción
Físicos	Bravo-Agapito et al., (2021)	• Edad avanzada. • Problemas de salud. • Desnutrición. • Malos hábitos alimenticios. • Cansancio y desgaste mental.
	Beyens et al., (2015)	• Adicción al alcohol y consumo de estupefacientes. • Adicción a la pornografía.
Ambientales y Tecnológicos	Maltby y Mackie (2009)	• Poca dedicación al estudio, lo que se ve reflejado en una baja cantidad de clics, en un bajo nivel de acceso a los recursos del curso y en la poca participación en los foros.
	Bravo-Agapito et al., (2021)	• Problemas de accesibilidad a internet. • Problemas técnicos y/o tecnológicos de las PEV y/o de los computadores que emplean los estudiantes. • Equipos de computación obsoletos o carencia de los mismos. • Pocas habilidades en el uso de las TIC. • Ruido y distracciones en el lugar de estudio.
Personales y Familiares	Salazar & de León (2008)	• Dificultades en el núcleo familiar o con las personas que convive. • Dificultades laborales. • Problemas de salud. • Problemas económicos de la familia del estudiante. • Obligación de cumplir horarios de trabajo. • Ser cabeza de familia.
	Lanzat et al., (2018)	• Problemas y conflictos con la pareja, ruptura amorosa, infidelidad, divorcio. • Muerte de algún familiar o persona significativa. • Problemas económicos severos en la familia.
Psicológicos	Ribeiro et al., (2018)	• Altos niveles de estrés. • Desórdenes del sueño. • Depresión. • Presencia del síndrome de Burnout. • Falta de motivación. • Problemas emocionales. • Dificultades de aprendizaje. • Estrés postraumático y malestar emocional. • Ansiedad.
	Alarcón et al., (2016)	• Baja motivación para culminar los estudios universitarios por temas difíciles de los cursos. • Indiferencia, desinterés y baja motivación por falta de gusto o temáticas con un gran nivel de dificultad. • Falta de apoyo por parte de la universidad y de la familia. • Depresión por falta de recursos económicos. • Incapacidad de autoevaluación del estudiante.
Académicos	Jena (2016)	• Malos hábitos y estilos de aprendizaje al momento de estudiar. • Mala preparación para la presentación de pruebas.

e Institucionales		• Falta de integración del estudiante con los compañeros del curso. • Falta de habilidad del estudiante para pedir ayuda a los profesores y compañeros. • Baja tolerancia a la frustración.
	Palvia et al., (2018)	• Diseño curricular inadecuado. • Falta de implementación de estrategias pedagógicas y didácticas en los cursos. • Falta de compromiso por falta de expectativas del estudiante. • Irresponsabilidad en el desarrollo de las actividades. • Baja automotivación.
	Padua Rodríguez (2019)	• Falta de autogestión y mal manejo del tiempo. • Poca o nula autonomía para el desarrollo de las actividades. • Poca orientación al logro. • Baja o nula espiritualidad. • Alto locus de control interno. • Incapacidad para comprender el material de los cursos. • Mala calidad de las relaciones con los profesores, compañeros y familiares.
	Helal et al., (2019)	• Cantidad excesiva de créditos matriculados. • Entorno socioeconómico de pobreza extrema. • Estudiar a tiempo parcial. • Mal manejo del tiempo y falta de disciplina. • Incapacidad para concentrarse al momento de desarrollar las actividades de aprendizaje. • Procastinación y ansiedad en el desarrollo de las actividades.
	Muljana y Luo (2019)	• Falta de adaptación por ser estudiante de primeros semestres. • Aplicación inadecuada de un estilo de aprendizaje (sensorial, intuitivo, visual, verbal, activo, reflexivo, secuencial, global, multimodal). • Bajo sentido de pertenencia del estudiante. • Bajo desempeño académico en el bachillerato. • Mayor cantidad de trabajo independiente. • Bajo efecto o influencia que ejerce la universidad bajo esta modalidad.

Fuente: elaboración del autor

2.5 Conclusiones y recomendaciones del capítulo

Es recomendable que las universidades e IES que dictan los CUVD creen plataformas virtuales que permitan a los docentes conocer información adicional a la relacionada con la interacción de los estudiantes con las PEV; factores personales de la vida del estudiante pueden influir significativamente en su desempeño académico; esta información puede ser recolectada por unidades de bienestar universitario o de permanencia académica; mientras más información se conozca relacionada con problemáticas, hábitos de estudio y demás, se pueden identificar con mayor precisión a los estudiantes que están en riesgo de reprobar o desertar de los CUVD.

3 Técnicas para predecir el rendimiento estudiantil en CUVD

En el Capítulo 3 se presentaron los resultados de una RSL enfocada a determinar los factores clave que causan un bajo rendimiento académico estudiantil en CUVD y se dio cumplimiento al primer objetivo específico de esta tesis. En este capítulo se dará desarrollo al segundo objetivo específico que corresponde a:

> Determinar la técnica adecuada para predecir el rendimiento académico estudiantil en CUVD

Para dar cumplimiento a este objetivo se realizó una RSL enfocada a responder la siguiente pregunta de investigación:

> ¿Qué técnicas se han empleado para predecir el rendimiento académico estudiantil en CUVD?

3.1 Introducción

Existen diversas técnicas del aprendizaje supervisado y no supervisado que se pueden aplicar para resolver problemas de clasificación y regresión, las cuales han sido empleadas en la predicción del rendimiento académico en CUVD. La mayoría de estas técnicas han sido utilizadas en el área de la educación para la construcción de modelos predictivos a partir de información de los estudiantes; algunas alcanzan mejores niveles de precisión que otras; sin embargo, en muchas de estas investigaciones no realizan agrupaciones por área de conocimiento, tampoco consideran el nivel o grado de dificultad que tienen ni el nivel o semestre en que se dictan; además, tampoco consideran que la estructura evaluativa de los cursos es diferente. Todo lo mencionado anteriormente puede dar lugar a prediccionespoco fiables.

3.2 Metodología

A continuación, se definen los pasos utilizados para ejecutar la RSL:

- Bases de datos utilizadas: Scopus e ISI Web of Knowledge WoS.
- Período de consulta: toda la información disponible hasta mayo de 2021.
- Cadena de búsqueda:

  TITLE-ABS-KEY ((Forecast* OR Predict) AND (Machine Learning OR Big Data) AND (Higher Education OR Virtual Education))

- Criterios de inclusión: que las investigaciones emplearan técnicas de ML en bases de datos de estudiantes de CUVD, que evaluaran la precisión obtenida.
- Criterios de exclusión: que emplearan técnicas diferentes a las de ML, que emplearan bases de datos de estudiantes de secundaria y media vocacional, que replicaran investigaciones pasadas con bases de datos diferentes.

3.3 Conceptos Básicos

A continuación, se definen brevemente las técnicas que se han empleado para predecir el rendimiento académico estudiantil en CUVD.

3.3.1 Regresión logística

En este modelo de clasificación se asume que y puede tomar los valores discretos $\{0,1\}$. La relación entre las variables independientes y la variable dependiente está dada por:

$$y = \sigma \sum_{i=1}^{n} (b_0 + \sum_{i=1}^{n} b_i x_i))$$

$$\sigma(u) = \frac{1}{1 + e^{-u}}$$

La regresión logística se utiliza principalmente para tareas de clasificación: *h(x)* se interpreta como la probabilidad de que la etiqueta de *x* sea 1. La hipótesis de clase asociada con la regresión logística es la composición de una función sigmoidal φ_{sig}: R → [0, 1] sobre la clase de funciones lineales L_d, el nombre “sigmoidal” significa “En forma de S” (Shalev- Shwartz & Ben-David, 2014).

K-Vecinos Próximos (K-NN): técnica de aprendizaje automático supervisado que almacena todos los casos disponibles de la muestra de entrenamiento y clasifica los casos nuevos de la muestra de prueba con base en una medida de similitud (distancia euclidiana) (Shalev-Shwartz & Ben-David, 2014):

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Esta técnica encuentra la distancia euclidiana entre los puntos de datos del conjunto de entrenamiento, luego, selecciona las K entradas más cercanas al nuevo punto de datos (Shalev-Shwartz & Ben-David, 2014). La etiqueta con la mayor frecuencia en las K entradas será la etiqueta de clase del nuevo punto de datos (Shalev-Shwartz & Ben-David, 2014).

Máquinas de Vectores de Soporte (SVM): técnica de aprendizaje automático supervisado que segrega de forma óptima dos clases buscando el mayor margen entre los puntos más cercanos del conjunto de entrenamiento de cualquier clase (hiperplano de frontera), denominados vectores de soporte. Las características de un espacio dimensional finito se pueden mapear en un espacio de mayor dimensión, lo que hace posible la separación lineal a pesar del espacio dimensional (Shalev-Shwartz & Ben-David, 2014). Esta es una técnica potente en comparación con otras técnicas de aprendizaje automático (Brownlee, 2016), ya que proporciona la mejor frontera (línea que separa las clases de puntos) de decisión que separa el espacio en clases.

Naive Bayes: técnica de aprendizaje automático supervisado que se basa en el Teorema de Bayes, método para hallar la probabilidad cuando se conocen otras probabilidades determinadas (Shalev-Shwartz & Ben-David, 2014):

$$P(Y|X) = \frac{P(Y)P(X|Y)}{P(X)}$$

Donde:

- *P(Y |X)* es la probabilidad de que se produzca *Y* cuando se produce *X*.
- *P(X |Y)* es la probabilidad de que ocurra *X* cuando ocurre *Y*.
- *P(Y)* es la probabilidad de que ocurra *Y* sin depender de que ocurra *X*.

- *P(X)* es la probabilidad de que ocurra *X* sin depender de que ocurra *Y*.

La variable *X* representa el conjunto de características y está dada como *X = (X1, X2, X3, ... Xn)*. Así, la ecuación adopta la forma:

$$P(Y|X1, \dots\dots, Xn) = \frac{P(X1|Y).P(X2|Y) \dots P(Xn|Y)}{P(X1).P(X2)\dots P(Xn)}$$

Los valores de la parte derecha de la ecuación se obtienen analizando el conjunto de datos y sustituyéndolo de la solución de la parte izquierda (H. Zhang et al., 2020).

Árboles de decisión: son clasificadores, *h: X → Y*, que predicen la etiqueta asociada a una instancia de variables, al pasar del nodo raíz a una hoja; estos se construyen como fragmentos en forma de rama. El árbol de decisión contiene nodos raíz y nodos hoja que significan las etiquetas de clase (Naganandhini & Shanmugavadivu, 2019). El atributo de datos con mayor prioridad en la toma de decisiones se selecciona como nodo raíz. El proceso de división de un árbol de decisión se decide en función de los valores de los datos de los respectivos nodos; al igual que otras técnicas, los árboles de decisión aprenden durante la fase de entrenamiento y su eficacia se evalúa durante la fase de prueba. La profundidad y la distribución de la información de entrenamiento y prueba del árbol de decisión, influyen dinámicamente en el rendimiento y la eficacia del clasificador (Naganandhini & Shanmugavadivu, 2019). La división se realiza de forma que se maximicen las características similares basadas en la entropía informativa mínima en cada división; las divisiones se denominan hojas y la división final se llama hoja terminal (IIOT, 2016).

Bosques Aleatorios: técnica de aprendizaje automático supervisado basada en árboles de decisión, los cuales se ensamblan por medio de bolsas, los árboles se entrenan de forma independiente (Golden et al., 2019); esta técnica utiliza un conjunto de árboles de decisión para predecir una salida en función de las características. La predicción es el resultado de decisiones secuenciales y binarias que se dividen de forma ortogonal en el espacio multivariado de variables (Golden et al., 2019). En esencia, es un meta-aprendizaje de varios árboles construidos de forma independiente.

RNA – MLP: técnica que se utiliza para resolver problemas de inteligencia artificial (IA), a menudo superan las técnicas de clasificación de ML, porque tienen las ventajas de la no linealidad, las interacciones variables y la personalización (Yegnanarayana, 2009). Como se mencionó en el capítulo I, las RNA son técnicas simplificadas que emulan la forma en que el cerebro humano procesa la información, es decir, un grannúmero de unidades de procesamiento interconectadas las cuales desempeñan el rol de las neuronas biológicas, las cuales trabajan simultáneamente para procesar información. Las RNA están conformadas por capas, cada una de las cuales contieneneuronas. La función de activación (*softmax, tanh, relu*) es la que se encarga de devolver una salida a partir de un valor de entrada, normalmente el conjunto de valores de salida en un rango determinado como *(0,1) o (-1,1).*

3.4 Resultados

Al aplicar la cadena de búsqueda en las bases de datos seleccionadas se recuperaron automáticamente 35 documentos. Al aplicar los criterios de inclusión y exclusión se descartaron 15 documentos, para obtener un total de 20 que fueron utilizados en esta RSL

3.4.1 Métricas generales de los trabajos seleccionados

La Tabla 14 presenta las principales métricas de los resultados obtenidos, la mayoría de publicaciones fueron realizadas entre varios investigadores, solo una minoríafueron realizadas por un solo autor, el promedio de citas por documento es superior a8, lo que indica que estas investigaciones han sido citadas en esta área de conocimiento.

Tabla 14. Investigaciones sobre predicción del rendimiento académico

Información Principal Sobre los Datos	
Período de publicación	2014:2021
Cantidad	20
Promedio de citas por documento	8,12
Promedio de citas por año por documento	2,23
Referencias	1.168

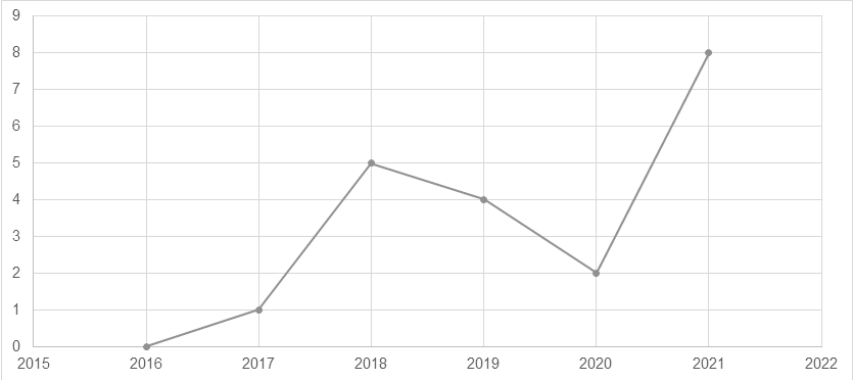
Palabras clave	1.423
Autores	
Publicaciones con un solo autor	3
Publicaciones con varios autores	17
Colaboración de Autores	
Promedio de autores por documento	0,28
Promedio de coautores por documento	3,46
Índice de colaboración	3,67

Fuente: elaboración del autor con Biblioshiny para Bibliometrix

3.4.2 Cantidad de publicaciones por año

De la Figura 3 se puede evidenciar un crecimiento no continuo en la cantidad de investigaciones, en el año 2019 se registró la mayor cantidad (4 publicaciones), la caída en el año 2021 obedece a que solo se tiene información parcial para este.

Figura 3. Publicaciones por año para la RSL.



Fuente: elaboración del autor con Biblioshiny para Bibliometrix

3.4.3 Países con mayor producción

La Tabla 15 presenta los tres países con la mayor cantidad de publicaciones y la principal contribución de cada uno en el área de conocimiento.

Tabla 15. Top 3 de los países con la mayor cantidad de publicaciones y su principal contribución

País	Artículos	Contribución
EEUU	5	Desarrollaron investigaciones enfocadas a comparar el desempeño de diferentes técnicas de ML de aprendizaje supervisado como: Árboles de Decisión, Naïve Bayes, Regresión Logística, Máquinas de Vectores de Soporte, K- Vecinos Próximos y RNA y señalan que la predicción del éxito o fracaso estudiantil permite a los docentes evitar que los estudiantes abandonen los estudios antes de los exámenes finales, identificar a aquellos que necesitan ayuda adicional y mejorar la clasificación y el prestigio de las instituciones; los antecedentes demográficos y académicos, y las características de comportamiento de los estudiantes son los factores esenciales que se deben analizar al momento de la predicción (Huang et al., 2021; Jensen et al., 2021; Marbouti et al., 2021; Wu, 2020).
Australia	3	Las indicaciones tempranas sobre el progreso de los estudiantes ayudan a los docentes a optimizar sus estrategias de aprendizaje y a enfocarse en diversas prácticas educativas para que la experiencia de aprendizaje sea exitosa; la aplicación del ML puede ayudar a los docentes a predecir las debilidades esperadas en los procesos de aprendizaje y, como resultado, pueden involucrar proactivamente a dichos estudiantes en una mejor experiencia de aprendizaje. Al igual que en EEUU, las investigaciones emplean técnicas de ML con información de los estudiantes y realizan comparativos para identificar la técnica óptima en términos de precisión (Hellas et al., 2018; Jia et al., 2019; Surenthiran et al., 2021; Suresh et al., 2021).
India	3	Al igual que en EEUU y Australia, las investigaciones realizadas en este país emplean técnicas de ML para predecir el rendimiento de estudiantes de CUVD, la mayoría de ellas solo consideran información demográfica de los estudiantes e información que se puede extraer de las PEV; y no incluyen aspectos relacionados con problemáticas personales de los estudiantes (Rana et al., 2021; Surenthiran et al., 2021; Suresh et al., 2021). Algunas investigaciones coinciden en la técnica óptima en términos de precisión para predecir el rendimiento en este tipo de cursos, específicamente perceptrones multicapa y árboles de decisión.

Fuente: elaboración del autor con Biblioshiny para Bibliometrix

3.4.4 Artículos más citados

La Tabla 16 presenta el Top 10 de las publicaciones con la mayor cantidad de citas y sus principales contribuciones el área de conocimiento.

Tabla 16. Principales investigaciones y su contribución

Autores y Año	Título	Citas	Contribución / Conclusión
(Hussain et al., 2018)	Student engagement predictions in an e-learning system and their impact on student course assessment scores	47	La predicción de los estudiantes que tendrán un bajo rendimiento académico es importante en los sistemas de e-learning porque permite a los docentes entender el comportamiento que tendrán los mismos en las diferentes actividades del curso. El nivel de compromiso de los estudiantes es un problema complejo que depende de factores como: la experiencia docente, el diseño del curso, el estilo de enseñanza, los conceptos y teorías. Estos factores deben investigarse más a fondo y analizar su relación con el nivel de compromiso de los estudiantes.
(Waheed et al., 2020)	Predicting academic performance of students from VLE big data using deep learning models	44	Las características demográficas y la actividad de los estudiantes después del inicio de un curso en las PEV tienen un impacto significativo en su rendimiento académico; la participación de los estudiantes en las PEV antes de que comiencen los módulos no tiene ninguna relación con su rendimiento.
(Buenaño-Fernández et al., 2019)	Application of machine learning in predicting performance for computer engineering students: A case study	38	Se deben crear planes de acción para evitar el abandono en las aulas y personalizar al máximo el seguimiento de los estudiantes de cursos virtuales, así como poner a su disposición información valiosa que les permita evaluar su rendimiento académico para que realicen acciones de mejora en las asignaturas que tienen mayor riesgo de fracaso; además, se deben agrupar a los estudiantes según diferentes criterios como lo son: afinidades por área de conocimiento, rendimiento por semestre, etc.
(Cazarez & Martin, 2018)	Neural Networks for Predicting Student Performance in Online Education	35	El desempeño en las actividades académicas durante el primer cuarto del semestre (25%), es un indicador del futuro desempeño del estudiante, por lo cual esnecesario identificar los factores que se involucran durante ese breve periodo de tiempo que puedan impactar en la obtención de una calificación aprobatoria al final del curso.
(Toskova et al., 2018)	Neural Networks in the Intelligent Educational Space	29	Es posible generar modelos de predicción para facilitar la aplicación de acciones preventivas enfocadas en minimizar la

			reprobación de los cursos virtuales por parte de los estudiantes; estos modelos pueden ser generados recolectando los datos de las calificaciones obtenidas por los mismos en un momento temprano del curso.
(Saltos & Maldonado, 2019)	Predictive models for the detection of problems in autonomous learning in higher education students virtual modality	25	Para las instituciones de educación superior, sobre todo en una modalidad virtual, sigue siendo un desafío prioritario lograr desarrollar competencias fuertes de autonomía y autorregulación en sus estudiantes; ya que muchos de ellos no se autorregulan demanera adecuada en su autoaprendizaje, lo que les impide completar con éxito los estudios en esta modalidad.
(B.-H. Kim et al., 2018)	Student performance prediction with deep learning	21	Diferentes técnicas y herramientas del Análisis del Aprendizaje (LA) ayudan a entender y optimizar el proceso de aprendizaje y contribuir al logro y éxito de los estudiantes de cursos virtuales; por lo cual se deben formular estrategias basadas en LA para indagar y medir variables asociadas a la autorregulación de los estudiantes y predecir en forma anticipada problemas académicos de los mismos.
(F. Yang & Li, 2018)	Study on student performance estimation, student progress analysis, and student potential prediction based on data mining	18	Los indicadores de progreso de los estudiantes proporcionan diferentes formas de agrupación, estas agrupaciones pueden utilizarse para analizar su progreso y hacer que los resultados del análisis y la estimación sean más precisos.
(Sekeroglu et al., 2019)	Student Performance Prediction and Classification Using Machine Learning Algorithms	16	El análisis de los datos educativos, especialmente el efecto delentorno social y la familia en elrendimiento de los estudiantes, es muy importante para mejorar la calidad de la educación de las generaciones futuras; por esta razón, es importante analizar diferente tipo de información para predecir y clasificar el comportamiento de los estudiantes de manera precisa y proporcionar una intervención adecuada a tiempo.
(Al-Shehri et al., 2017)	Student performance prediction using support vector machine and k-nearest neighbor.	12	Los modelos predictivos construidos con información de la interacción de los estudiantes con los LMS generan mejores predicciones a mitad de curso (50%) que al final (100%), estos modelos pueden ser útiles para la

			detección temprana de estudiantes en riesgo y para proporcionar bases para adoptar acciones tempranas para prevenir el fracaso de los estudiantes. Es posible que las universidades creen oficinas de análisis especializadas dedicadas al seguimiento y a la predicción del rendimiento de los estudiantes, estas unidades pueden proporcionar información a mediados de semestre a los docentes, coordinadores de programas y profesionales de apoyo del aprendizaje para que trabajen con los estudiantes que tienen dificultades.

Fuente: elaboración del autor con Biblioshiny para Bibliometrix

El promedio de citas de las publicaciones Top 10 es 28,50, la cantidad máxima de citas es 47, la mínima es 12. Hay publicaciones más recientes que otras con una mayor cantidad de citas, esto obedece al nivel de innovación y aporte al área de conocimiento, al rigor metodológico y a la validez de los resultados alcanzados.

3.4.5 Principales autores

La Tabla 17 presenta los tres autores con la mayor cantidad de citas en orden descendente, el índice H, la cantidad de publicaciones y la principal contribución al área de conocimiento.

Tabla 17. Principales autores y su contribución

Autor	Cantidad de Publicaciones relacionadas	Índice H	Cantidad de citas	Contribución
Zacharis, N. Z	5	12	825	El análisis de los datos de los estudiantes extraídos de las PEV permite predecir su rendimiento, identificar patrones de comportamiento y desarrollar intervenciones para los que tengan dificultades, las RNA del tipo Perceptrón Multicapa, es una de las mejores técnicas para esta tarea (Zacharis, 2010, 2015, 2016).
Macfadyen, L.P	3	19	289	Existe una estrecha relación entre el tiempo empleado para el desarrollo de actividades en

				línea, la estructura y evaluación de los cursos, y las percepciones de los estudiantes; toda esta información debe ser consideraba para que las predicciones sean fiables (Macfadyen et al., 2017; Macfadyen & Dawson, 2010; Roll et al., 2015).
Waheed, H	2	6	213	La abundancia de datos educativos accesibles que pueden ser obtenidos de las PEV brindan oportunidades para analizar el comportamiento del proceso de aprendizaje de los estudiantes, identificar sus problemas, optimizar el entorno educativo y facilitar la toma de decisiones. Las PEV complementan el paradigma de la analítica del aprendizaje al proporcionar de manera efectiva conjuntos de datos para analizar y reportar el proceso de aprendizaje de los estudiantes (Waheed et al., 2020, 2021; Wasif et al., 2019)

Fuente: elaboración del autor con Biblioshiny para Bibliometrix

3.4.6 Principales fuentes

La Tabla 18 presenta el Top 10 de las fuentes con la mayor cantidad de publicaciones. la mayoría de ellas publican resultados de investigaciones científicas relacionadas con la enseñanza de la ingeniería, la educación en general y las ciencias de la computación. Algunas presentan desarrollos de aplicaciones y software para su uso en el área de la educación en la ingeniería.

Tabla 18 . Principales fuentes y su contribución

Fuentes	Cantidad de Publicaciones relacionadas	Índice H	Total, de Citas
Educational Technology and Society	3	88	156
IEEE Access	3	127	34
IEEE Transactions on Learning Technologies	3	47	146
Education and Information Technologies	3	41	29
International Journal of Advanced Computer Science and Applications	2	18	7
International Journal of Engineering Education	2	50	10
Journal of Education for Business	1	46	15
Teaching of Psychology	1	48	50
Applied Intelligence	1	66	1
Australian Journal of Basic and Applied Sciences	1	33	10
Total	20		

Fuente: elaboración del autor con Biblioshiny para Bibliometrix

3.5 Discusión

A continuación, se responde la pregunta de investigación planteada al inicio de este capítulo que es:

> ¿Qué técnicas se han empleado para predecir el rendimiento académico estudiantil en CUVD?

En esta sección se da respuesta y se discute la pregunta de investigación. La Tabla 19 presenta algunas investigaciones que han empleado las técnicas mencionadas anteriormente para predecir el desempeño académico estudiantil en los CUVD.

Tabla 19.Caracterización de las Investigaciones sobre las técnicas aplicadas en la predicción del desempeño académico en los CUVD

Articulo	Técnica y Precisión obtenida	Variables	Contribución
(Morris et al., 2005)	Regresión Lineal Múltiple – 49%	• Nº de mensajes de discusión vistos. • Nº de páginas de contenido vistas. • Nº de mensajes originales. • Nº de mensajes de seguimiento. • Segundos de visualización de las páginas de debate. • Segundos para ver el contenido. • Segundos de creación de mensajes originales. • Segundos de creación de mensajes de seguimiento.	El tiempo dedicado a la realización de tareas y la frecuencia de participación en las PEV son importantes para tener éxito en el aprendizaje en línea. Los estudiantes que completaron el curso participaron en actividades de aprendizaje en línea con mayor frecuencia y mayor cantidad de tiempo que los alumnos que se retiraron.
(Macfadyen & Dawson, 2010)	Regresión Lineal Múltiple – 57%	• Número total de sesiones en línea. • Tiempo total en línea. • Mensajes de correo leídos. • Mensajes de correo enviados. • Mensajes de discusión leídos. • Total de mensajes de discusión enviados. • Nuevos mensajes de discusión publicados. • Mensajes de respuesta publicados.	La cantidad de tiempo total que pasan los estudiantes en línea se correlacionan débilmente con la calificación final de los estudiantes, el tiempo total en línea no es una variable predictiva significativa de la calificación del estudiante en los modelos de regresión. Información pedagógicamente significativa de los estudiantes no se puede extraer directamente de los registros del LMS.

(Yu & Jo, 2014)	Regresión Lineal Múltiple – 66%	• Frecuencia total de inicio de sesión frecuencia en LMS. • Tiempo total de estudio en el LMS. • Regularidad de intervalo de aprendizaje en LMS. • Número de descargas. • Interacciones con compañeros. • Interacciones con el instructor.	Los docentes deben prestar más atención a mejorar el rendimiento académico de los estudiantes en lugar de predecirlo. El tiempo total de estudio en el LMS, la regularidad de accesos al LMS, el número de descargas y la interacción con compañeros de clase se correlacionan significativamente con la nota final.
(Zacharis, 2015)	Regresión Lineal Múltiple – 52%	• Lectura y publicación de mensajes. • Edición de la wiki. • Lectura de mensajes de correo. • Asignaciones enviadas. • Participación en concursos. • Actualización del blog. • Archivos vistos. • Enlaces web vistos. • Tiempo total en línea. • Total de visitas al LMS. • Vista de la Wiki. • Sesiones en línea. • Mensaje añadidos al foro. • Evaluaciones iniciadas. • Vista del blog. • Intentos de continuación de Test. • Vista de foros. • Vista de recursos. • Intento de cierre del cuestionario. • Acceso a la herramienta de calificaciones. • Intento de cuestionario. • Vista del foro de discusión. • Vista del cuestionario. • Revisión del cuestionario. • Charla en el chat. • Vista de la tarea.	Desarrollar un CUVD de manera exitosa no siempre es una tarea fácil. Además de la necesidad de abordar los requisitos tecnológicos, seleccionar el mejor contenido y diseñar actividades atractivas, es necesario monitorear y rastrear las actividades en línea de los estudiantes de manera constante. La extracción y análisis de los datos de uso del LMS almacenadosen archivos de registro, proporciona a los docentes los medios para supervisar su progreso y planificar intervenciones a tiempo.
(Bravo-Agapito et al., 2021)	Regresión Lineal Múltiple – 76%	• Número total de accesos a la plataforma Moodle. • Frecuencia de acceso de los estudiantes a todos los foros. • Número total de mensajes añadidos por el alumno en los foros. • Frecuencia de acceso de los estudiantes a los materiales didácticos. • Frecuencia de acceso de los estudiantes a todos los glosarios.	Las implicaciones de la educación virtual están principalmente relacionadas con las estrategias de intervención; estas deben ser diseñadas para mejorar el rendimiento académico de los estudiantes en una etapa temprana de sus estudios. La uniformidad entre cursos evita posibles errores de sesgo de medición.

		• Número total de tareas del curso. • Frecuencia de consulta de las tareas por parte del alumno. • Número total de trabajos entregados. • Frecuencia de acceso de los estudiantes a todos los cuestionarios. • Frecuencia con la que el alumno intenta resolver los cuestionarios. • Total de preguntas contestadas en todos los cuestionarios. • Frecuencia con la que el estudiante observa los cuestionarios. • Frecuencia con la que el alumno envía un cuestionario. • Frecuencia con la que el alumno revisa los cuestionarios. • Número de días hasta el primer acceso al aula virtual. • Número total de entradas al curso. • Edad del alumno. • Sexo del estudiante . • Titulación en la que está matriculado el alumno . • Calificación final obtenida en el curso. • GPA	
(Zaporozhko et al., 2019)	RNA - MLP Feedforward 3-L – 98.81%	• Puntuación del examen estatal unificado o de las pruebas de acceso. • Puntuación de los resultados de los diagnósticos de entrada (prueba de diagnóstico, prueba de entrada • trabajo, etc.). • Puntuación de los resultados de la evaluación formativa . • Puntuación de los resultados de la evaluación sumativa (pruebas finales, trabajos escritos).	La necesidad de predicción es un componente importante del sistema de soporte de decisiones para la gestión inteligente del proceso educativo.

		• Puntuación media de los resultados de la certificación intermedia del curso. • Puntuación de los resultados de la aprobación y protección de todos los tipos de prácticas. • Puntuación de los resultados de los exámenes estatales. • Puntuación del resultado de la defensa del trabajo final de calificación (según la declaración de la comisión de examen).	
(Rivas et al., 2021)	RNA - MLP (2 HL- SoftMax) – 78% Potenciación del Gradiente (Extreme gradient boosting) – 76% Bosques Aleatorios - 75% Árboles de Decisión – 70%	• Cantidad de clics. • Cantidad de recursos descargados. • Recursos visualizados. • Cantidad de discusión en foros. • Cantidad de tareas presentadas	Es importante medir el tiempo que los estudiantes dedican a interactuar con las PEV para poder evaluar sus beneficios con certeza. El factor que más influye en el rendimiento de los estudiantes de los CUVD es el número de clics en los recursos de la PEV, es decir, la interacción con los recursos y tareas con las PEV.
(Vasudevan et al., 2018)	OneR – 61% PART – 64% J48 – 62%	• Identificación del estudiante. • Edad. • Sexo. • Resultado final. • Puntuación. • Tipo de evaluación. • Peso de la evaluación. • Suma de clics. • Créditos estudiados	Debido a que cada vez más personas buscan una educación superior orientada a la virtualidad, los sistemas de gestión del aprendizaje (LMS) deben mejorar la estructura para gestionar mejor los datos de los estudiantes.
(Quinn & Gray, 2020)	Bosques Aleatorios – 60.5% Potenciación del Gradiente (Extreme gradient boosting) – 57.1% Análisis Discriminante Lineal (LDA) – 59.5% K-Vecinos Próximos – 59%	• Módulos Vistos. • URL vistas. • Páginas vistas. • Envíos de tareas. • Intentos de Quizzes. • Quizzes enviados. • Lecciones vistas. • Recursos descargados. • Participación en foros. • Cantidad de accesos en franjas horarios (00:00 – 06:00 AM, 6:00 AM – 12:00 M, 12:00 M – 6:00 PM, 6:00 PM – 12:00 PM)	Aunque los datos de Moodle pueden ser útiles como componente de los modelos de alerta temprana, es poco probable que sean suficientes por sí solos para predecir con precisión a la mayoría de los estudiantes que suspenden un CUVD. Para tener un uso más práctico es deseable saber al principio de los CUVD si un estudiante puede estar en riesgo de suspenderlo o reprobarlo.

Fuente: elaboración del autor

Según los resultados de la RSL, se puede afirmar que las RNA - MLP, es la mejor técnica para predecir el desempeño estudiantil en los CUVD debido al alto nivel de precisión; esta afirmación coincide con la investigación desarrollada por Namoun y Alshanqiti (2021), en donde clasifican las mejores y las peores técnicas de predicción del desempeño estudiantil. Estos autores señalan que las peores técnicas son en su orden: regresión lineal, bagging, análisis de funciones discriminantes y regresión logística. Por su parte, señalan que las mejores técnicas son en su orden: redes neuronales artificiales RNA - MLP, Bosques Aleatorios y Naïve Bayes.

A pesar de que el tamaño de la muestra era adecuado en la mayoría de investigaciones y realizan un adecuado pre-procesamiento de los datos, no consideraron el nivel o grado de dificultad que tenían los CUVD, ni el nivel o semestre en que se dictaban; además, la mayoría de variables analizadas únicamente se centraron en información relacionada con la interacción de los estudiantes con las PEV y no contemplaron que los estudiantes de CUVD también pueden atravesar por situaciones o dificultades personales que afectan su desempeño académico.

3.6 Conclusiones y recomendaciones del capítulo

Los modelos predictivos de la mayoría investigaciones analizadas en la RSL alcanzan altos niveles de precisión; sin embargo, no consideran aspectos relevantes que pueden afectar la fiabilidad de las predicciones. El objetivo de los modelos de clasificación es que aprendan patrones que generalicen bien los datos que no fueron analizados en lugar de memorizar datos que aprendieron durante el entrenamiento; no es correcto afirmar con certeza que una técnica de clasificación sea mejor que otras, la cantidad de información y el objetivo que tenga el investigador desempeñan un papel importante. Es necesario evaluar todas las métricas de precisión para decidir cuál es la mejor y no solo centrarse en la precisión. Es necesario observar los modelos que se separen más del caso aleatorio, y no solo fiarse de precisiones altas, ya que es posible que las clases estén desbalanceadas y se presenten problemas de sub-entrenamiento o sobre-entrenamiento. Para poder determinar la mejor técnica para predecir el rendimiento académico en los CUVD se deben identificar los

parámetros óptimos que maximizan la precisión, en este sentido, es importante considerar tanto los valores de precisión y error sobre los datos de validación, como la relación o peso entre Falsos Negativos (FN) y Verdaderos Negativos (VN), ya que en ocasiones es conveniente sacrificar precisión en la predicción y dar peso a alguno de los casos de Falsos Positivos (FP) o Falsos Negativos (FN) con el objetivo de producir un modelo con mayor capacidad de generalización.

4 Metodología para predecir el desempeño estudiantil en los CUVD

4.1 Introducción

Como se mencionó anteriormente, la metodología que se propone en esta investigación doctoral tiene como objetivo facilitar a los docentes de los CUVD la identificación de los estudiantes que tienen mayor riesgo de reprobar, para que de esta manera implementen estrategias enfocadas a minimizar la mortalidad y la deserción académica a tiempo en este tipo de cursos; en la metodología intervienen tres actores:

- Primer actor: un experto en ML, el cual se encarga de entrenar diferentes modelos de clasificación para determinar el más adecuado a partir de una base de datos colectiva de información personal y de las interacciones con las PEV de los estudiantes de los CUVD (accesos al curso, accesos a los recursos de la PEV, URLs vistas, cuestionarios vistos, tareas presentadas, cuestionarios y quices presentados, participación en foros, visualización de discusiones, etc.), pertenecientes a una misma área de conocimiento (Agronomía, Veterinaria y afines, Bellas Artes, Ciencias de la Educación, Ciencias de la Salud, Ciencias Sociales y Humanas, Economía, Administración, Contaduría y afines, Ingeniería, Arquitectura, Urbanismo y afines, y Matemáticas y Ciencia Naturales), que se dicten en los programas académicos de las universidades, y si es posible de una misma asignatura para tener mayor precisión; además, se encarga de configurar la PEV para que los docentes visualicen la predicción en cualquier momento del curso.

- Segundo actor: los docentes de los CUVD de los programas académicos de las universidades, los cuales deben visualizar la predicción realizada para cada uno de los estudiantes en cualquier momento del curso. Se debe aclarar que, no es obligatorio que los docentes tengan conocimientos en programación, como se mencionó anteriormente, su rol únicamente se limita a la formulación, implementación y seguimiento de estrategias pedagógicas y didácticas

encaminadas a evitar la mortalidad y la deserción académica de los estudiantes que fueron identificados en riesgo de reprobar.

- Tercer actor: los estudiantes en riesgo de reprobar, los cuales deberían acogerse a la implementación de las estrategias pedagógicas y didácticas formuladas por el docente de los CUVD; se debe señalar que, en caso de que sea necesario, estos estudiantes pueden ser remitidos a las unidades de permanencia académica o bienestar institucional de las universidades para que reciban ayuda o acompañamiento especializado; el objetivo es que los estudiantes implementen las estrategias recibidas para evitar que fracasen o deserten de los CUVD.

Específicamente, esta metodología está compuesta por los siguientes pasos:

1. Determinación de las variables a analizar
2. Construcción de la base de datos
3. Construcción de los modelos de predicción
4. Evaluación de los modelos
5. Visualización de la predicción

La Figura 4 resume el rol que desempeña cada uno de los actores en la metodología propuesta:

Figura 4. Actores y roles

Experto en ML	Docente del CUVD	Estudiante en Riesgo
Se encarga de determinar las variables de la PEV, construir la base de datos y los modelos, evaluarlos, seleccionar el óptimo y configurar la PEV a los docentes del CUVD.	Se encargan de visualizar los resultados de la predicción, formular estrategias para los estudiantes en riesgo, socializarlas y hacer el seguimiento de las mismas.	Se encarga de realizar las actividades propuestas por el docente del CUVD.

Fuente: elaboración del autor

En la actualidad, gracias al uso de PEV, entornos virtuales de aprendizaje (EVA), sistemas de gestión de aprendizaje (LMS), portales web especializados, entre otros, es posible conocer información relacionada con el acceso, la interacción y el desempeño de los

estudiantes de los CUVD. La Tabla 21 presenta un resumen de la información disponible en PEV que se puede emplear para predecir el rendimiento académico estudiantil en los CUVD:

A continuación, se resumen las principales variables que se pueden extraer de las PEV (Quinn & Gray, 2019, Vasudevan & Almuhanna, 2018, Bravo-Agapito et al., 2021).

- Nº total de ingresos al CUVD
- Primera fecha de ingreso al CUVD
- Tiempo total de permanencia en la PEV (minutos)
- Períodos de inactividad (minutos)
- Tiempo transcurrido hasta el desarrollo de la primera actividad (minutos)
- Tiempo promedio por sesión (minutos)
- Nº de recursos y enlaces vistos
- Nº vistas a recursos de páginas del curso
- Nº vistas a foros de discusión
- Total, de participaciones en foros de discusión
- Nº de quizzes iniciados, intentados, superados y vistos
- Resultado final (aprueba/reprueba)
- Tipo de prueba
- Peso de la prueba (valor porcentual)
- N° total de clics en el curso
- Cursos matriculados en la PEV
- Nº de ingresos
- Nº de cursos vistos
- Nº de recursos vistos
- Nº de tareas vistas, enviadas
- Nº de lecciones vistas, iniciadas, terminadas y respondidas
- Nº de vistas a archivos
- Nº de accesos y participaciones en foros
- Nº de accesos durante la semana (lunes-viernes)
- Nº de accesos durante el fin de semana (sábados-Domingos)
- Nº de accesos entre las 00:00 am y las 6 am.
- Nº de accesos entre las 06:00 am y las 12 m.
- Nº de accesos entre las 12:00 m y las 6:00 pm.
- Nº de accesos entre las 06:00 pm y las 12:00 pm.
- Nº de ingresos con una IP del campus universitario
- Nº de ingresos con una IP por fuera del campus universitario
- Nº de accesos a unidades didácticas
- Nº de accesos a glosarios
- Nº de accesos, intentos y preguntas respondidas a cuestionarios
- Nº de cuestionarios enviados
- Nº de revisiones a cuestionarios
- Género
- Fecha de nacimiento

4.2 Metodología Propuesta

A continuación, se detalla cada uno de los seis pasos que la conforman:

4.2.1 Paso 1. Determinación de las variables a analizar

Como se mencionó anteriormente, en los CUVD es complejo conocer información de los estudiantes relacionada con problemas o situaciones personales que estén afectando su desempeño académico (problemas de salud, económicos, sentimentales, familiares, etc.); además, existe una ausencia de plataformas o sistemas de información en las universidades que permitan que los estudiantes las registren y los docentes las consulten. Solo un bajo porcentaje de estudiantes se dirigen de manera presencial a las dependencias de bienestar universitario o permanencia académica a solicitar apoyo, principalmente por falta de tiempo, confianza o timidez. Como se mencionó en el Capítulo 3, son diversos los factores que pueden contribuir a un bajo rendimiento académico en los CUVD; en esta metodología se propone hacer uso de la información derivada de la interacción de los estudiantes con las PEV (accesos al curso, accesos a los recursos de la PEV, URLs vistas, cuestionarios vistos, tareas presentadas, cuestionarios presentados, quices presentados, participación en foros, visualización de discusiones, etc.) y de la información personal y del entorno del estudiante, lo cual fue identificado como una debilidad de los modelos predictivos actuales.

4.2.1.1 Descripción del paso

Identificar y descargar la información que se puede obtener de las PEV (Ver sección 3.4.3). Para garantizar una alta precisión en los modelos predictivos, debe existir una coherencia en el área de formación de los CUVD, en el semestre que adelantan los estudiantes y en la estructura (cantidad de actividades evaluativas, recursos y nivel de dificultad). No es correcto emplear modelos construidos con registros de estudiantes de los CUVD de mayor grado de complejidad o semestres avanzados para predecir el rendimiento de estudiantes de los CUVD de menor grado de dificultad o de semestres iniciales, y mucho menos si los CUVD pertenecen a diferentes áreas de formación y tienen una estructura evaluativa diferente.

4.2.1.2 Recomendaciones para hacer el paso

Construir modelos para cada CUVD de manera independiente para garantizar la coherencia Se debe aclarar que, no se debe mezclar información de los CUVD de posgrado con los CUVD de pregrado porque tienen una duración y forma de evaluación diferente. Las bases de datos de los estudiantes que se empleen para la construcción de los modelos deben ser de una misma área de conocimiento (ingeniería, ciencias exactas y naturales, ciencias sociales y humanas, derecho, etc.), y si es posible de una misma asignatura para tener mayor precisión. Así, por ejemplo, no es correcto emplear un modelo construido con información de los CUVD de estudiantes de ingeniería con estudiantes de los CUVD de psicología, ya que no hay equivalencia entre las competencias que deben desarrollar los estudiantes en cada uno de los cursos; además, la estructura evaluativa es diferente, al igual que el nivel o grado de dificultad, y los tipos de recursos pedagógicos y didácticos que emplean en las PEV.

Por otra parte, la estructura evaluativa de los CUVD empleados para construir los modelos predictivos debe ser similar en cuanto a cantidad y tipo de recursos, tareas, evaluaciones, quices y foros; además, el porcentaje sobre la nota final y el nivel de nivel de dificultad de las actividades evaluativas debe ser similar, de lo contrario la precisión de los modelos puede verse afectada. Si los modelos predictivos se construyen con toda la información de la interacción de los estudiantes de los CUVD al finalizar un semestre académico, no sería correcto emplearlo en el semestre siguiente si no se conserva la misma estructura, por lo tanto, si hay variabilidad en los aspectos mencionados anteriormente, la precisión de los modelos puede verse afectada.

En caso de que no se tenga acceso a la información completa de la interacción de los estudiantes con la PEV (Logs), se requiere solicitarla al administrador de la PEV de la universidad. El experto o administrativo de la facultad o programa académico debe encontrar e identificar las fuentes para recoger los datos relevantes y en caso de que sea posible recoger información adicional de los estudiantes del CUVD (discapacidad, enfermedad, si trabaja o no, ingresos familiares, si tiene problemas económicos o sentimentales, farmacodependencia, si vive con los padres o no, entre otros) en las dependencias de bienestar universitario o permanencia académica. Como trabajo futuro se propone crear una plataforma que permita registrar información relacionada con problemas personales de los estudiantes de CUVD (físicos, sentimentales, económicos etc.).

4.2.2 Paso 2. Construcción de la base de datos

La consistencia interna de las variables que conforman las bases de datos empleadas para construir modelos predictivos en cualquier área de conocimiento garantiza la calidad de las predicciones; valores por fuera de los rangos normales de variación (outliers), información faltante, baja cantidad de información, entre otros, pueden afectar la precisión de los modelos.

4.2.2.1 Descripción del paso

Crear una base de datos integral con la información de los estudiantes del CUVD.

4.2.2.2 Recomendaciones para hacer el paso

Asegurarse de que toda la información de las variables que conforman la base de datos esté completa; hacer la depuración de la información, categorizar las variables cualitativas para poder proceder con el entrenamiento de los modelos y completar los campos vacíos que no fueron obtenidos con valores nulos; el objetivo de realizar la limpieza de datos es eliminar el ruido y las inconsistencias en el ajuste de los datos.

4.2.3 Paso 3. Construcción de los modelos de predicción

En la Sección 4.1 se presentaron algunas técnicas que han sido empleadas para predecir el rendimiento académico en CUP, cada una de ellas se caracteriza por tener sus propias particularidades; en la revisión del estado del arte se muestra que algunas de ellas alcanzan niveles más altos de precisión que otras, esto no quiere decir que unas sean mejores que otras, la cantidad de datos disponible y el objetivo que tenga el investigador (clasificación o regresión / aprendizaje supervisado o no supervisado) también desempeñan un papel de gran importancia.

4.2.3.1 Descripción del paso

El experto o administrativo del departamento, facultad o programa académico entrenará los modelos de clasificación (Árboles de decisión, Bosques Aleatorios, Naïve Bayes, RNA-

MLP, Máquinas de Vectores de Soporte, Regresión Logística y K-NN) con la base de datos de los estudiantes del CUVD.

4.2.3.2 Recomendaciones para hacer el paso

Los datos pre-procesados que fueron recolectados automáticamente se deben clasificar en dos subconjuntos; entrenamiento y prueba. Se recomienda hacerlo con el 70% y el 30% respectivamente, se deja bajo consideración del experto modificar o conservar estos porcentajes en pro de obtener una mayor precisión.

4.2.4 Paso 4. Evaluación de los modelos

Existen diferentes métricas (Accuracy, F1-Score, ROC-AUC, Recall y Precision) que permiten determinar cuál es el modelo óptimo para resolver un problema de clasificación o regresión; cada una de las técnicas presentadas en la sección 4.1 difieren en cuanto a cantidad y tipo de parámetros; la identificación de los valores de los parámetros que hacen óptimo un modelo permiten que estos sean entrenados de manera tal que alcancen la mayor precisión posible, lo que facilita la comparación y la selección del modelo óptimo.

4.2.4.1 Descripción del paso

Después de que el experto o administrativo del departamento, facultad o programa académico entrene todos los modelos con los parámetros óptimos de cada uno, realizará un comparativo de las métricas (Accuracy, F1-Score, ROC-AUC, Recall y Precision) obtenidas por los modelos y seleccionará el mejor.

4.2.4.2 Recomendaciones para hacer el paso

La métrica Accuracy no funciona bien cuando las clases están desbalanceadas; si la mayoría de estudiantes son clasificados en la etiqueta de “Aprueba”, es muy fácil acertar diciendo que cualquier estudiante del CUVD va a aprobar, cuando esto sucede es mejor valerse del resultado de las métricas; Precision, Recall y F1-Score, ya que dan una mejor idea de la calidad del modelo.

Es necesario evaluar todas las métricas para decidir cuál es la mejor y no solo centrarse en Accuracy, también es importante observar los modelos que se separan más del caso aleatorio, y no solo fiarse de Accuracy altas, ya que es posible que se presenten problemas de sub-ajuste o sobre-ajuste (overfitting y underfitting); el modelo fallará en clasificar a un estudiante porque no tiene estrictamente los mismos valores de la muestra de entrenamiento o por falta de suficientes muestras. La determinación del mejor modelo debe pasar por una serie de técnicas y métricas que deben tenerse en cuenta para mejorar el desempeño de los mismo y obtener el más adecuado.

En este sentido, es importante considerar tanto los valores de precisión y error sobre los datos de validación, como la relación o peso entre Falsos Negativos (FN) y Verdaderos Negativos (VN), ya que en ocasiones es conveniente sacrificar precisión en la predicción y dar peso a alguno de los casos de Falsos Positivos (FP) o Falsos Negativos (FN) con el objetivo de producir un modelo con mayor capacidad de generalización. También se recomienda utilizar técnicas de aprendizaje supervisado para clasificación y no para regresión; el objetivo de los modelos no es predecir la nota exacta del estudiante en las tareas de la PEV, sino detectar a los que están en riesgo de reprobar.

4.2.5 Paso 5 Visualización de la predicción

El nivel de precisión de los modelos garantiza que los estudiantes que realmente están en riesgo de reprobar sean identificados e intervenidos a tiempo; la formulación de estrategias pedagógicas y didácticas personalizadas por parte de los docentes de los CUVD para los estudiantes que tienen un bajo rendimiento, contribuye a que estos tengan la posibilidad de desarrollar actividades empleando el estilo de aprendizaje que más les guste (visual, auditivo, verbal y kinestésico), lo que contribuye al mejoramiento de su desempeño académico.

4.2.5.1 Descripción del paso

Los docentes visualizan las predicciones de los estudiantes del CUVD en las PEV con un avance inferior al 75% e identifican a los que fueron clasificados en riesgo de reprobar; luego, analizan a detalle el desempeño de estos estudiantes en el desarrollo de las actividades y definen estrategias para evitar que reprueben el curso o lo cancelen. En caso

de que sea necesario, estas estrategias pueden ser formuladas en compañía de las dependencias de bienestar universitario o permanencia académica. Finalmente, recibirán la retroalimentación por parte del docente del CUVD e implementarán sus estrategias. Cabe aclarar que los estudiantes también pueden recibir apoyo de las dependencias de bienestar universitario o permanencia académica y los docentes del CUVD deben hacer seguimiento de la implementación de las estrategias formuladas.

4.2.5.1 Recomendaciones para hacer el paso

Es recomendable que los docentes visualicen de manera periódica las predicciones del rendimiento académico de los estudiantes de los CUVD, ya que, si se identifica a tiempo a los estudiantes en riesgo de reprobar, es más factible que estos implementen de manera efectiva estrategias pedagógicas y didácticas en pro de mejorar su proceso de aprendizaje.

4.3 Conclusiones del capítulo

La metodología presentada es lineal, no deben saltarse o evadirse pasos, debe serrealizada en orden y considerar las recomendaciones establecidas en cada paso. Elestudiante es el responsable de implementar las estrategias socializadas por el docente delCUVD, el cual debe hacer un seguimiento periódico para determinar si se están ejecutandoo no; aspectos relacionados con problemas personales de los estudiantes deben ser re- direccionados a las dependencias de bienestar universitario, las cuales se encargarán de hacer un seguimiento al estudiante.

Debido a la cantidad limitada de información que se puede extraer de las PEV, las universidades deberían crear un sitio (base de datos web) donde los estudiantes registren problemas personales, familiares, económicos, entre otros de manera periódica; este sitio lo puede administrar las unidades de bienestar universitario o permanencia estudiantil. La información relacionada con la interacción del estudiante con la plataforma educativa puede ser obtenida con el administrador de la plataforma.

5 Caso de Aplicación

5.1 Introducción

En el Capítulo 3 y Capítulo 4 se dio desarrollo a los dos primeros objetivos específicos, en este capítulo se dará desarrollo al tercer objetivo específico que corresponde a:

> Construir un modelo para predecir el rendimiento académico estudiantil en CUVD.

Predecir el desempeño estudiantil en CUVD puede llegar a ser una tarea compleja si no se cuenta con suficiente información para entrenar y validar modelos predictivos; en este capítulo se presenta el caso de aplicación de la metodología propuesta usando tres bases de datos de CUVD de tres países; Inglaterra, Uruguay y Colombia; el objetivo de este capítulo es aplicar la metodología presentada en el Capítulo 5 a las bases mencionadas anteriormente. Se debe aclarar que la disponibilidad de bases de datos públicas en la web de estudiantes de CUVD es muy limitada debido a la confidencialidad que deben manejar las universidades con la información de los estudiantes; específicamente en Colombia, en el área de la educación, la Ley 1266 de 2008 Habeas Data, establece que el tratamiento de información confidencial de manera inadecuada puede dar lugar a sanciones. A continuación, se detalla la metodología empleada.

5.2 Metodología

En primer lugar, se seleccionaron las variables a analizar de las bases de datos obtenidas, posteriormente, se realizó la depuración de la información, se categorizaron las variables cualitativas para poder entrenar los modelos, se completaron campos vacíos con valores nulos para eliminar el ruido y las inconsistencias en el ajuste de los datos; para la preparación de los conjuntos de datos se empleó MS Excel y para aplicar las diferentes técnicas de clasificación se usó Google Colaboratory. La definición conceptual de cada una de las técnicas se presenta en el Capítulo 2, por lo cual no se abordará en este capítulo.

La base de datos de Inglaterra (n=27.550) se dividió en una muestra de entrenamiento (70%, n=19.285) y en una muestra de prueba (30%, n=8.265). La base de datos de Uruguay

(n=25.260) se dividió en una muestra de entrenamiento (70%, n=17.682) y en una muestra de prueba (30%, n=7.578). La base de datos de Colombia (n=1.340) se dividió en una muestra de entrenamiento (70%, n=938) y en una muestra de prueba (30%, n=402).

Los datos de las tres bases de datos fueron normalizados antes del entrenamiento con la librería MinMaxScaler de Python; los hiperparámetros que optimizaban los resultados del entrenamiento fueron identificados haciendo uso de GridSearchCV, clase disponible en Scikit-Learn que permite evaluar y seleccionar de forma sistemática los parámetros de un modelo de clasificación o regresión; los códigos empleados para el desarrollo de los modelos en Google Colab están disponibles en los siguientes enlaces:

Uruguay	https://colab.research.google.com/drive/13nfW9E4fcMZNCgwb8aBpwOmIQiA-eU3Q?usp=sharing
Colombia	https://colab.research.google.com/drive/1kidAhV_kNSBjgAU8qHAaWBhfRB5u3Fvj?usp=sharing
Inglaterra	https://colab.research.google.com/drive/1nwp1BzdeuGd9QRb1UGgzUXs6TzGqZ-pB?usp=sharing

En cuanto a la evaluación de la calidad de los diferentes modelos de clasificación de ML para predecir el rendimiento estudiantil en CUVD, se emplearon las siguientes métricas:

- Accuracy: mide cuántas observaciones, tanto positivas como negativas se clasificaron correctamente (Chicco & Jurman, 2020).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}} \qquad (9)$$

- ROC-AUC: permite evaluar qué tan bien el técnica puede separar los casos positivos y negativos e identificar el mejor umbral para separarlos (Chicco & Jurman,2020).

$$\text{ROC} - \text{AUC} = \int \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN}} \, \text{d}(\text{FPR}) \qquad (10)$$

- F1-Score: combina precisión y recuperación en una métrica calculando la media armónica entre esos dos, es la media ponderada de la precisión y la recuperación (Chicco & Jurman, 2020).

$$\text{F1} - \text{Score} = 2\,\frac{\text{Precision.\,Recall}}{\text{Precision} + \text{Recall}} \qquad (11)$$

- Recall: número de verdaderos positivos dividido por el número de valores positivos en los datos de prueba. Un valor bajo indica un alto número de falsos negativos (Chicco & Jurman, 2020).

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad (12)$$

- Precision: número de verdaderos positivos dividido por todas las predicciones positivas. La precisión también se denomina predicción positiva. El valor es una medida de la exactitud de un clasificador. Una precisión baja indica un alto número de falsos positivos (Chicco & Jurman, 2020).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad (13)$$

5.3 Descripción de las bases de datos

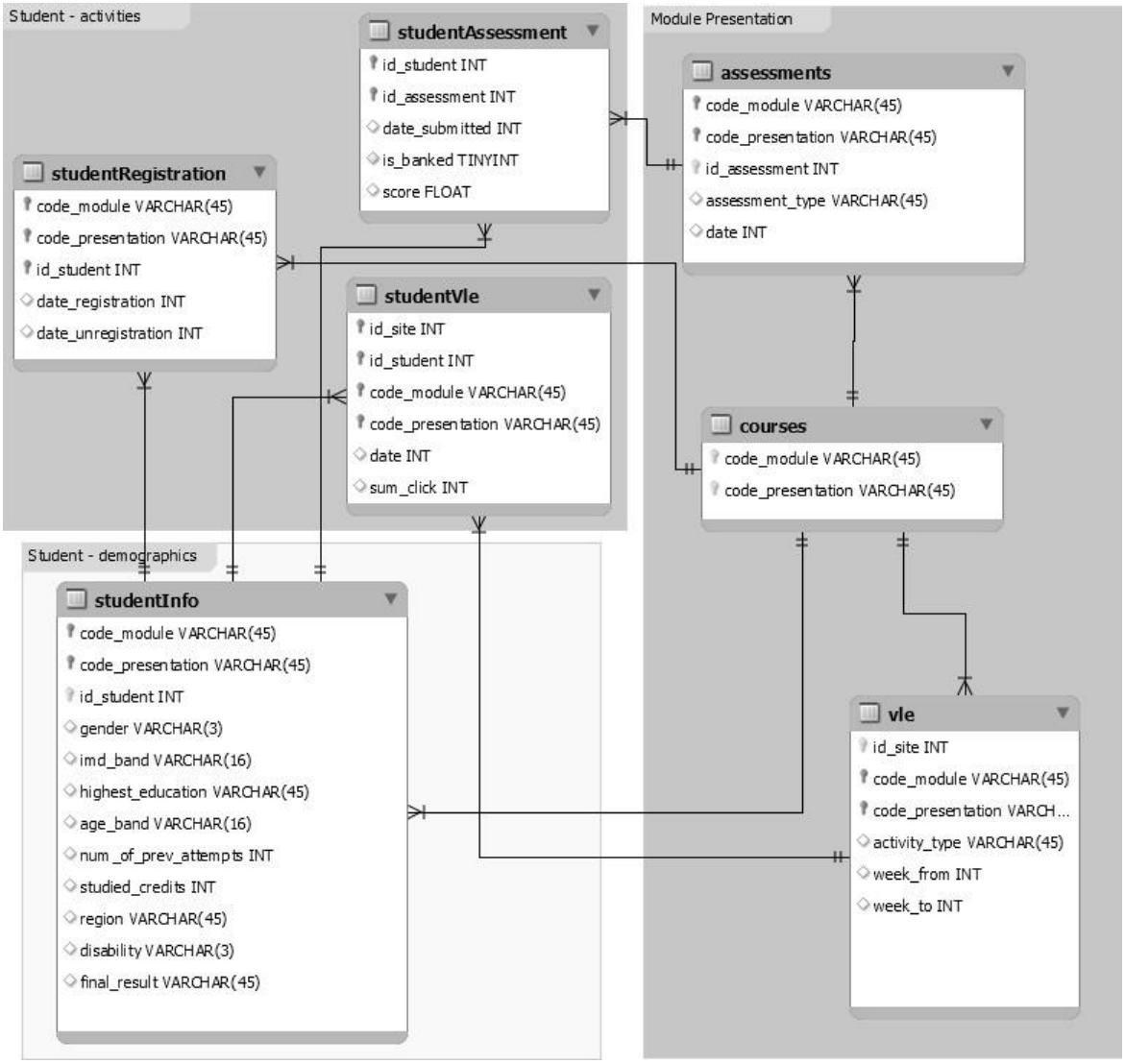
En esta sección se presenta la descripción de las bases de datos analizadas de CUVD de tres países; Inglaterra, Uruguay y Colombia. Se describen las variables que contienen y su tamaño muestral.

5.3.1 Base de datos – Caso Inglaterra

Esta base de datos es pública y anónima y está disponible en el siguiente enlace: https://analyse.kmi.open.ac.uk/open_dataset. Los datos provienen de un grupo de n= 27.550 estudiantes que han participado en una serie de cuatro CUVD (encuentros sincrónicos y tareas asincrónicas); la base de datos incluye información personal de los

estudiantes, tales como su código de identificación, el sexo, la región, el nivel educativo, el rango de edad, índice de privación múltiple (IMD), el número de veces que han participado previamente en el curso, los créditos matriculados, el resultado de exámenes finales y el número de veces que el estudiante ha interactuado con cualquiera de los contenidos del curso en línea. La mayoría de variables son cualitativas, por lo que se categorizaron previamente antes de su uso. La Figura 5 presenta el esquema de esta base de datos.

Figura 5. Base de Datos - Caso Inglaterra



Fuente: OULAD https://analyse.kmi.open.ac.uk/open_dataset.

A continuación, se detallan cada una de las variables de esta base de datos:

courses

- code_module: nombre de código del módulo (identificador).
- code_presentation: nombre en código de la presentación, B (febrero) y J (octubre).
- Longitud: duración de la presentación del módulo en días.

assessments

- code_module: código de identificación del módulo al que pertenece la evaluación.
- code_presentation: código de identificación de la presentación a la que pertenece la evaluación.
- id_assessment: número de identificación de la evaluación.
- Assessment_type: tipo de evaluación, evaluación marcada por el tutor (TMA), evaluación marcada por computadora (CMA) y examen final (examen).
- fecha: información sobre la fecha de envío final de la evaluación calculada como el número de días desde el inicio de la presentación del módulo. La fecha de inicio de la presentación tiene el número 0.
- % Peso de la evaluación: los exámenes se tratan por separado y tienen un peso del 100%, la suma de todas las demás evaluaciones es 100%.

Vle

- id_site: número de identificación del material.
- code_module: código de identificación para el módulo.
- code_presentation: código de identificación de presentación.
- activity_type: rol asociado con el material del módulo.
- week_from: semana a partir de la cual se planea usar el material.
- week_to _week: semana hasta la cual se planea utilizar el material.

StudentInfo

- code_module: código de identificación para un módulo en el que está registrado el estudiante.
- code_presentation: código de identificación de la presentación durante la cual el estudiante está registrado en el módulo.

- id_student: número de identificación único para el estudiante.
- gender: género del estudiante.
- region: identifica la región geográfica donde vivió el estudiante mientras realizaba la presentación del módulo.
- educación_más alta: nivel de educación más alto del estudiante al ingresar a la presentación del módulo.
- imd_band: índice de necesidades insatisfechas del lugar donde el estudiante desarrolló el módulo (equivalente al estrato social en Colombia).
- age_band: edad del estudiante.
- num_of_prev_attempts: número de veces que el estudiante ha intentado curso.
- study_credits: número total de créditos que el estudiante ha superado.
- discapacidad: indica si el estudiante ha declarado una discapacidad o no.
- final_result: resultado final del alumno en la presentación del módulo.

studentRegistration

- code_module: código de identificación para un módulo.
- code_presentation: código de identificación de la presentación.
- id_student: número de identificación único para el estudiante.
- date_registration: fecha de registro del estudiante en la presentación del módulo, (número de días medidos en relación con el inicio de la presentación del módulo).
- date_unregistration: fecha en la que el estudiante se da de baja de la presentación del módulo (número de días medidos con respecto al inicio de la presentación del módulo).

studentAssessment

- id_assessment: número de identificación de la evaluación.
- id_student: número de identificación único para el estudiante.
- date_submitted: fecha de envío del estudiante, medida como el número de días desde el inicio de la presentación del módulo.
- is_banked: bandera de estado que indica que el resultado de la evaluación se ha transferido de una presentación anterior.

- puntaje: puntaje del estudiante en esta evaluación. El rango es de 0 a 100. La puntuación inferior a 40 se interpreta como Falla. Las marcas están en el rango de 0 a 100.

studentVle

- code_module: código de identificación para un módulo.
- code_presentation: código de identificación de la presentación del módulo.
- id_student: número de identificación único para el estudiante.
- id_site: número de identificación para el material VLE.
- fecha: fecha de la interacción del estudiante con el material medido como el número de días desde el inicio de la presentación del módulo.
- sum_click: cantidad total de clics durante el desarrollo del módulo.

Se debe aclarar que, para el análisis únicamente se consideraron las variables que realmente podían afectar el desempeño de los estudiantes de CUVD; las variables relacionadas con la finalización y la calificación de pruebas en línea no fueron consideradas. La Tabla 20 presenta las variables que finalmente fueron empleadas.

Tabla 20. Variables - Caso Inglaterra

	Nombre	Descripción
Variables	Resultado final	Aprueba=0, Reprueba=1
	Genero	Masculino = 0, femenino =1
	Región	Variable categórica [1-13]
	IPM	Variable categórica [1-10]
	Edad	Número entero
	Créditos aprobados a la fecha	Número entero
	Discapacidad	Si = 0, No =1
	Total de clics	Número entero

Fuente: elaboración del autor

5.3.2 Base de datos - Caso Uruguay

Esta base de datos empleada proviene de un grupo de n = 25.261 estudiantes de CUVD (encuentros sincrónicos y tareas asincrónicas) de la Universidad de Oviedo de programas de pregrado de las áreas de humanidades, ciencias, ingeniería, salud, ciencias sociales y derecho; los datos fueron recogidos entre el año 2017 y 2018; las variables analizadas fueron:

Variables relacionadas con la asignación de tareas

- Cumplimiento de lo obligatorio: porcentaje de presentación de las tareasobligatorias.
- Promedio de notas obligatorias: promedio de las notas de los trabajos obligatorios, considerando los trabajos no presentados como nota cero.
- Cumplimiento de las tareas obligatorias: porcentaje de tareas obligatorias entregadas hasta el momento de la predicción, en relación con el número total de tareas obligatorias del curso.
- Cumplimiento opcional: porcentaje de tareas opcionales presentadas.
- Nota promedio de la realización de tareas optativas: nota promedio de los trabajos optativos, considerando los trabajos no presentados como nota cero.
- Cumplimiento de tareas opcionales: porcentaje de tareas opcionales entregadas por el alumno hasta el momento de la predicción.

Variables relacionadas con la interacción de los estudiantes con el LMS

- CourseViewPct: porcentaje de accesos al curso, en relación al total de accesos de todos los alumnos de ese curso (División de la cantidad de accesos del estudiante entre el número total de accesos de todos los estudiantes del curso, multiplicado *100).
- CourseViewTime {1}: primera vez que el estudiante accedió al curso, en relación con su duración (División del tiempo que duró conectado el estudiante en los 5 primeros accesos entre el tiempo de duración del curso, multiplicado *100).
- CourseViewTime {2}: segunda vez que el estudiante accedió al curso, en relación con su duración (División del tiempo que duró conectado el estudiante en los 5 primero accesos entre el tiempo de duración del curso, multiplicado *100).
- CourseViewTime {3}: tercera vez que el estudiante accedió al curso, en relación con su duración (División del tiempo que duró conectado el estudiante en los 5 primero accesos entre el tiempo de duración del curso, multiplicado *100).
- CourseViewTime {4}: cuarta vez que el estudiante accedió al curso, en relación con su duración (División del tiempo que duró conectado el estudiante en los 5 primero accesos entre el tiempo de duración del curso, multiplicado *100).

- CourseViewTime {5}: quinta vez que el estudiante accedió al curso, en relación con su duración (División del tiempo que duró conectado el estudiante en los 5 primero accesos entre el tiempo de duración del curso, multiplicado *100).
- CourseViewTimePct: media geométrica de los cinco valores de CourseViewTime {1, 2, 3, 4, 5}. (Raíz n-ésima del producto de todos los porcentajes anteriores).
- ResourceViewPct: porcentaje de accesos a los recursos del profesor, en relación con el total de accesos a los recursos del profesor para todos los estudiantes de ese curso. (División de los accesos del estudiante a los recursos del profesor entre el total de accesos de todos los estudiantes a los recursos del profesor, multiplicado *100).
- ResourceViewTime {1}: primera vez que el estudiante ve un recurso del profesor, en relación con la duración del curso (porcentaje).
- ResourceViewTime {2}: segunda vez que el estudiante ve un recurso del profesor, en relación con la duración del curso (porcentaje).
- ResourceViewTime {3}: tercera vez que el estudiante ve un recurso del profesor, en relación con la duración del curso (porcentaje).
- ResourceViewTime {4}: cuarta vez que el estudiante ve un recurso del profesor, en relación con la duración del curso (porcentaje).
- ResourceViewTime {5}: quinta vez que el estudiante ve un recurso del profesor, en relación con la duración del curso (porcentaje).
- ResourceViewTimePct: media geométrica de los cinco valores de ResourceViewTime {1, 2, 3, 4, 5}. (Raíz n-ésima del producto de todos los porcentajes anteriores).
- ResourceViewUniquePct: porcentaje de recursos del profesor vistos por el estudiante, en relación con el número total de recursos del profesor en el curso.
- UrlViewPct: porcentaje de URLs vistas, en relación con el total de URLs vistas de todos los alumnos de ese curso.
- UrlViewTime {1}: primera vez que el estudiante vio una URL, en relación con la duración del curso (porcentaje).
- UrlViewTime {2}: segunda vez que el estudiante vio una URL, en relación con la duración del curso (porcentaje).
- UrlViewTime {3}: tercera vez que el estudiante vio una URL, en relación con la duración del curso (porcentaje).

- UrlViewTime {4}: cuarta vez que el estudiante vio una URL, en relación con la duración del curso (porcentaje).
- UrlViewTime {5}: quinta vez que el estudiante vio una URL, en relación con la duración del curso (porcentaje).
- UrlViewTimePct: media geométrica de los cinco valores de UrlViewTime{1, 2, 3, 4, 5}. (Raíz n-ésima del producto de todos los porcentajes anteriores)
- UrlViewUniquePct: porcentaje de URLs vistas por el estudiante, en relación al número total de URLs del curso.
- AssignViewPct: porcentaje de visualizaciones de la tarea, en relación con el total de visualizaciones de la tarea para todos los estudiantes de ese curso.
- AssignViewTime {1}: primera vez que el estudiante vio una tarea, en relación con la duración del curso (porcentaje).
- AssignViewTime {2}: segunda vez que el estudiante vio una tarea, en relación con la duración del curso (porcentaje).
- AssignViewTime {3}: tercera vez que el estudiante vio una tarea, en relación con la duración del curso (porcentaje).
- AssignViewTimePct: media geométrica de los tres valores de AssignViewTime{1, 2, 3}. (Raíz n-ésima del producto de todos los porcentajes anteriores)
- AssignViewUniquePct: porcentaje de tareas vistas por el estudiante, en relación con el número total de tareas del curso.
- QuizViewPct: porcentaje de visualizaciones de cuestionarios, en relación con el total de visualizaciones de cuestionarios de todos los estudiantes de ese curso.
- QuizViewTime {1}: primera vez que el estudiante vio una prueba, en relación con la duración del curso (porcentaje).
- QuizViewTime {2}: segunda vez que el estudiante vio una prueba, en relación con la duración del curso (porcentaje).
- QuizViewTime {3}: tercera vez que el estudiante vio una prueba, en relación con la duración del curso (porcentaje).
- QuizViewTimePct: media geométrica de los tres valores de QuizViewTime{1, 2, 3}. (Raíz n-ésima del producto de todos los porcentajes anteriores)
- QuizViewUniquePct: porcentaje de cuestionarios vistos por el estudiante, en relación con el número total de cuestionarios del curso.

- AssignSubmitPct: porcentaje de envíos de tareas, en relación con el total de envíos de tareas de todos los estudiantes de ese curso.
- AssignSubmitTime {1}: primera vez que el estudiante presentó una tarea, en relación con la duración del curso (porcentaje).
- AssignSubmitTime {2}: segunda vez que el estudiante presentó una tarea, en relación con la duración del curso (porcentaje).
- AssignSubmitTime {3}: tercera vez que el estudiante presentó una tarea, en relación con la duración del curso (porcentaje).
- AssignSubmitTimePct: media geométrica de los tres valores de AssignSubmitTime{1, 2, 3}. (Raíz n-ésima del producto de todos los porcentajes anteriores)
- AssignSubmitUniquePct: porcentaje de tareas presentadas por el estudiante, en relación con el número total de tareas del curso.
- QuizAttemptPct: un intento de cuestionario es cuando el estudiante comienza un cuestionario, por lo que esta variable mide el porcentaje de intentos de cuestionario, en relación con el total de intentos de cuestionario para todos los estudiantes de ese curso.
- QuizAttemptTime {1}: primera vez que el estudiante comenzó un cuestionario, en relación con la duración del curso.
- QuizAttemptTime {2}: segunda vez que el estudiante comenzó un cuestionario, en relación con la duración del curso.
- QuizAttemptTime {3}: tercera vez que el estudiante comenzó un cuestionario, en relación con la duración del curso.
- QuizAttemptTimePct: media geométrica de los tres valores de QuizAttemptTime{1, 2, 3}. (Raíz n-ésima del producto de todos los porcentajes anteriores)
- QuizAttemptUniquePct: porcentaje de cuestionarios iniciados por el estudiante, en relación con el número total de cuestionarios del curso.
- QuizCloseAttemptPct: porcentaje de envíos de cuestionarios, en relación con el total de envíos de cuestionarios de todos los estudiantes de ese curso.
- QuizCloseAttemptTime {1}: primera vez que el estudiante presentó un cuestionario, en relación con la duración del curso.
- QuizCloseAttemptTime {2}: segunda vez que el estudiante presentó un cuestionario, en relación con la duración del curso.

- QuizCloseAttemptTime {3}: tercera vez que el estudiante presentó un cuestionario, en relación con la duración del curso.
- QuizCloseAttemptTimePct: media geométrica de los tres valores QuizCloseAttemptTime {1, 2, 3}. (Raíz n-ésima del producto de todos los porcentajes anteriores)
- QuizCloseAttemptUniquePct: porcentaje de cuestionarios presentados por el estudiante, en relación con el número total de cuestionarios del curso.
- ForumViewForumPct: porcentaje de visualizaciones del foro, en relación con el número total de visualizaciones del foro para todos los estudiantes del curso.
- ForumViewDiscussionPct: porcentaje de visualizaciones de discusiones, en relación con el número total de visualizaciones de discusiones para todos los estudiantes del curso.

En este caso todas las variables fueron consideradas, ya que todas podían afectar el desempeño académico de los estudiantes de CUVD.

5.3.3 Base de datos – Caso Colombia

Esta base de datos fue de n= 1.340 estudiantes de CUVD pertenecientes a diferentes programas de formación de la Universidad Católica Luis Amigó, Medellín, Colombia. La base de datos contiene información relevante que se extrajo del campus virtual universitario https://virtual.ucatolicaluisamigo.edu.co/campus/, el cual fue construido con Moodle, uno de los LMS más empleados por las universidades a nivel mundial. Cabe resaltar que, era una base de datos anónima, los nombres, apellidos y documentos de identificación (cédula de ciudadanía, tarjeta de identidad, etc.) de los estudiantes no fueron considerados; los campos que conformaron la base de datos fueron:

- ID del estudiante: consecutivo alfanumérico
- Edad: número entero
- Género: masculino / femenino
- Estrato: 1,2,3,4,5,6
- Resultado final: aprueba / reprueba
- Total, de clics: cantidad total acumulada durante todo el curso.

- Créditos aprobados: cantidad de créditos que el estudiante ha aprobado al momento de iniciar el CUVD.

Al igual que en el caso de Inglaterra, las variables que se consideraron fueron las que realmente podían afectar el desempeño de los estudiantes; los tipos de eventos relacionados con la finalización y la calificación de los cuestionarios en línea no fueron considerados. La Tabla 21 presenta las variables que se emplearon.

Tabla 21. Variables - Caso Colombia

	Nombre	Descripción
Variables	Resultado final	Aprueba=0, Reprueba=1
	Genero	Masculino = 0, femenino =1
	Estrato	Variable categórica [1-6]
	Edad	Número entero
	Créditos aprobados a la fecha	Número entero
	Total, tiempo dedicado en minutos	Número entero
	Cantidad de cursos matriculados	Número entero
	Total, de clics	Número entero

Fuente: elaboración del autor

5.4 Resultados

Es necesario aclarar que, los datos en las tres bases de datos estaban balanceados, es decir, había una similitud en la cantidad de estudiantes que Aprobaban y Reprobaban, razón por la cual se priorizó la métrica de Accuracy sobre las métricas de Precision, Recally F1-Score. Es necesario aclarar que solo se analizaron los resultados de las métricas de las técnicas con el mayor puntaje (Score) identificado por la función GridSearchCV de Python, los códigos y los resultados del entrenamiento de los modelos se pueden visualizaren los siguientes enlaces de Google Colaboratory.

Uruguay https://colab.research.google.com/drive/13nfW9E4fcMZNCgwb8aBpwOmIQiA-eU3Q?usp=sharing

Colombia https://colab.research.google.com/drive/1kidAhV_kNSBjgAU8qHAaWBhfRB5u3Fvj?usp=sharing

Inglaterra https://colab.research.google.com/drive/1nwp1BzdeuGd9QRb1UGgzUXs6TzGqZ-pB?usp=sharing

La Tabla 22 presenta los mejores puntajes y el valor de los parámetros óptimos de cada una de las bases de datos analizadas.

Tabla 22. Resultados de la aplicación de GridSearchCV

País	Modelo	Mejor Puntaje	Mejores Parámetros
Inglaterra	Bosque Aleatorios	0.92	{'n_estimators': 70}
Uruguay	Bosque Aleatorios	0.89	{'n_estimators': 80}
Colombia	Máquinas de Vectores de Soporte	0.89	{'C': 100, 'kernel': 'rbf', 'tol': 0.01}

Fuente: elaboración del autor

La Tabla 23 presenta un comparativo de los resultados de las métricas mencionadas anteriormente de los modelos con el mejor puntaje (best_score) identificados al aplicar la función GridSearchCV de Scikit-Learn de Python en las tres bases de datos.

Tabla 23. Comparativo general de los mejores modelos de predicción

País	Inglaterra	Uruguay	Colombia
Técnica óptima	Bosques Aleatorios	Bosques Aleatorios	Máquinas de Vectores de Soporte
Accuracy	0.59	0.90	0.96
F1-Score	0.65	0.94	0.95
Recall	0.71	0.97	0.94
Precision	0.60	0.91	0.92

Fuente: elaboración del autor

En el caso de Inglaterra y Uruguay, la técnica de Bosques Aleatorios fue la mejor, en el caso de Colombia las Máquinas de Vectores de Soporte (SVM). Los valores más altos en cuanto a métricas de calidad (Accuracy, F1-Score, Recall y Precision) se obtuvieron con labase de datos de Colombia, a pesar de que el tamaño muestral de la base de datos era considerablemente menor que el tamaño muestral de los CUVD de Inglaterra y Uruguay. Como se mencionó anteriormente, las clases estaban balanceadas en cuanto a la cantidad de estudiantes que aprobaban y reprobaban.

Únicamente en el caso de Colombia fue posible realizar el último paso de la metodología propuesta, el cual hace referencia a la visualización de la predicción (Ver Anexo 1). En el caso de Inglaterra y Uruguay no fue posible realizarlo, ya que no se tenía la autorización de ingresar a las PEV en donde se desarrollaron.

5.5 Discusión

Es necesario aclarar que, en los tres casos analizados la variable Resultado_final se categorizó de la siguiente forma: Aprueba=0 y Reprueba=1; en el caso de Inglaterra, el 49% de los estudiantes reprobaron, el 51% aprobaron. Si un estudiante tiene aprobados menos de 53 créditos tiene una probabilidad de reprobar del 0.39, si tiene aprobados más de 53 créditos, tiene una probabilidad de aprobar del 55%. Se empleó la prueba de chi-cuadrado para analizar la relación de dependencia o independencia entre la variable Resultado final y las otras variables de la base de datos. Los resultados se presentan en la Tabla 24:

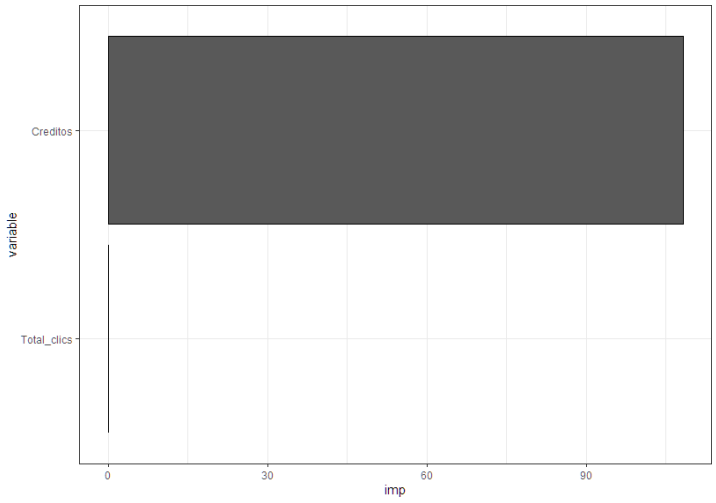
Tabla 24. Análisis de independencia - Caso Inglaterra

	Resultado final		
	X-squared	df	Valor p
Edad	101.98	2	2.2e-16
Índice de privación múltiple (IPM)	502.85	9	2.2e-16
Género	29.645	1	5.189e-08
Crédito aprobados	629.14	54	2.2e-16
Total clics	69.046	63	0.02805
Región	165.57	12	2.2e-16
Discapacidad	76.739	1	2.2e-16

Fuente: elaboración del autor

Con un nivel de significación de 0.05, se rechaza la hipótesis nula en los todos los escenarios, es decir hay dependencia, en el caso de Inglaterra el resultado final está relacionado con la edad, el IPM, el género, los créditos aprobados, el total de clics, la región y la discapacidad. A mayor número de créditos superados al momento del inicio del CUVD, mayor probabilidad de aprobarlo, el número total de clics es la segunda variable en orden de importancia, ver Figura 6. En las técnicas de clasificación, cada predictor tiene una importancia independiente (Max Kuhn, 2021).

Figura 6. Importancia de las variables – Caso Inglaterra



Fuente: elaboración del autor

No quiere decir esto que con solo la información de los créditos sea suficiente para predecir el rendimiento académico de los estudiantes de los CUVD, puede presentarse el caso que a un estudiante que haya aprobado una cantidad de créditos significativa de su programa académico se le presenten dificultades personales que le impidan dedicar el tiempo suficiente al desarrollo del CUVD, lo que sin duda alguna puede afectar su rendimiento académico. Mientras más información se conozca de los estudiantes las predicciones tiene mayor fiabilidad.

En el caso de Uruguay el 20% de los estudiantes reprobaron, el 80% aprobaron. La variable de mayor importancia para determinar si un estudiante aprobará o no es el porcentaje de cumplimiento con lo obligatorio. Si no ha cumplido con el 85% de las asignaciones obligatorias tiene una probabilidad de reprobar el curso del 0.92, si tiene aprobado más del 85% tiene una probabilidad de aprobar el curso del 0.80. Considerando que la variable objetivo Resultado final era categórica, se empleó la prueba de chi-cuadrado para analizar su relación de dependencia o independencia con las otras variables de la base de datos. Los resultados se presentan en la Tabla 25:

Tabla 25. Análisis de independencia - Caso Uruguay

	Resultado final		
	X-squared	df	Valor p
Accomplish_mandatory	504.08	2	2.2e-16
Accomplish_mandatory_grade	1212	993	2.2e-16
Accomplish_mandatory_pct_graded	1452.1	117	2.2e-16

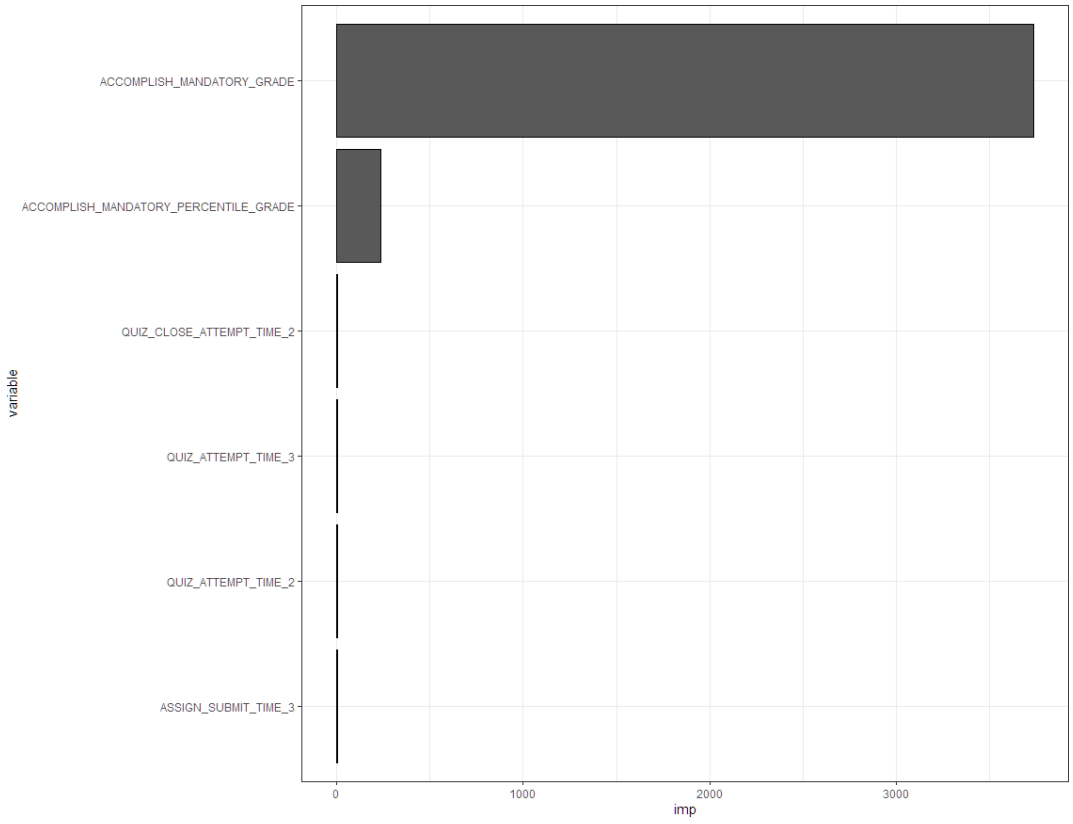
Accomplish_mandatory_percentile_grade	6744	985	2.2e-16
Accomplish_optional	174.21	2	2.2e-16
Accomplish_optional_grade	829.18	619	2.903e-08
Accomplish_optional_pct_graded	545.8	115	2.2e-16
Accomplish_optional_percentile_grade	1649.4	690	2.2e-16
Course_view_pct	1320.1	278	2.2e-16
Course_view_time_1	847.29	874	0.7355
Course_view_time_2	1003.3	915	0.02192
Course_view_time_3	1048.2	934	0.005294
Course_view_time_4	1164.1	961	6.631e-06
Course_view_time_5	1168.7	975	1.728e-05
Course_view_time_pct	1106.1	959	0.000649
Resource_view_pct	1103.1	272	2.2e-16
Resource_view_time_1	1154.3	962	1.764e-05
Resource_view_time_2	1188.8	982	5.68e-06
Resource_view_time_3	1294.8	988	1.404e-10
Resource_view_time_4	1244.6	994	8.821e-08
Resource_view_time_5	1252.3	995	4.424e-08
Resource_view_time_pct	1195.1	993	9.588e-06
Resource_view_unique_pct	2274.9	854	2.2e-16
Url_view_pct	768.2	253	2.2e-16
Url_view_time_1	1240.9	972	9.107e-09
Url_view_time_2	1340.3	974	4.456e-14
Url_view_time_3	1295.7	964	3.881e-12
Url_view_time_4	1337.4	958	4.749e-15
Url_view_time_5	1286.1	911	2.571e-15
Url_view_time_pct	1340.9	989	4.506e-13
Url_view_unique_pct	932.53	230	2.2e-16
Assign_view_pct	990.33	277	2.2e-16
Assign_view_time_1	841.88	960	0.9974
Assign_view_time_2	948.51	981	0.7663
Assign_view_time_3	923.95	985	0.9177
Assign_view_time_pct	963.03	986	0.6937
Assign_view_unique_pct	1060.6	214	2.2e-16
Quiz_view_pct	826.26	195	2.2e-16
Quiz_view_time_1	810.47	918	0.9953
Quiz_view_time_2	949.6	947	0.4701
Quiz_view_time_3	1102.5	967	0.001521
Quiz_view_time_pct	908.79	959	0.8753
Quiz_view_unique_pct	1004.8	161	2.2e-16
Assign_submit_pct	1081.7	188	2.2e-16
Assign_submit_time_1	1125.5	976	0.0005982
Assign_submit_time_2	1216.6	992	1.179e-06
Assign_submit_time_3	1004.8	981	6.241e-06
Assign_submit_time_pct	1183.4	978	2.2e-16
Assign_submit_unique_pct	1461.8	143	2.2e-16
Quiz_attempt_pct	1384.3	144	3.07e-05
Quiz_attempt_time_1	1147.9	962	0.0001063
Quiz_attempt_time_2	1137.3	966	2.2e-16
Quiz_attempt_time_3	1137.3	966	2.2e-16
Quiz_attempt_time_pct	1149.6	981	0.0001453
Quiz_attempt_unique_pct	1113.4	144	2.2e-16
Quiz_close_attempt_pct	1489.7	146	2.2e-16
Quiz_close_attempt_time_1	1102.2	965	0.001345
Quiz_close_attempt_time_2	1162.1	973	2.538e-05
Quiz_close_attempt_time_3	1162.1	973	2.538e-05
Quiz_close_attempt_time_pct	1181.7	979	8.068e-06
Quiz_close_attempt_unique_pct	1184.9	140	2.2e-16
Forum_view_forum_pct	639.66	342	2.2e-16

Forum_view_discussion_pct	717.61	283	2.2e-16

Fuente: elaboración del autor

Con un nivel de significación de 0.05, el resultado final no está relacionado con las variables: Course_View_Time_1, Assign_View_Time_1, Assign_View_Time_2, Assign_View_Time_3, Assign_View_Time_Pct, Quiz_View_Time_1, Quiz_View_Time_2 Y Quiz_View_Time_Pct. En estos casos no se rechaza la hipótesis nula, es decir hay independencia. Con el resto de variables de la base de datos hay dependencia, es decir, se rechaza la hipótesis nula. A continuación, se presenta el análisis de importancia de la aplicación de la técnica de árboles de decisión a la base de datos de Uruguay, ver Figura 7.

Figura 7. Importancia de las variables – Caso Uruguay



Fuente: elaboración del autor

La variable de mayor importancia para determinar si un estudiante aprobará o no es el porcentaje de cumplimiento con lo obligatorio, esto obedece a que los estudiantes que han

cumplido con todas las asignaciones han dedicado mayor cantidad de tiempo de estudio personal.

En el caso de Colombia el 33% de los estudiantes reprobaron, el 77% aprobaron. Sí un estudiante tiene aprobados 51 créditos o más, tiene una probabilidad de reprobar de 0.06, si un estudiante tiene aprobados menos de 51 créditos tiene una probabilidad de aprobar el curso de 0.89. Considerando que la mayoría de las variables eran categóricas y solo había una variable discreta, se empleó la prueba de chi-cuadrado para analizar la relación de dependencia o independencia entre la variable Resultado final y las otras variables de la base de datos. Los resultados se presentan en la Tabla 26:

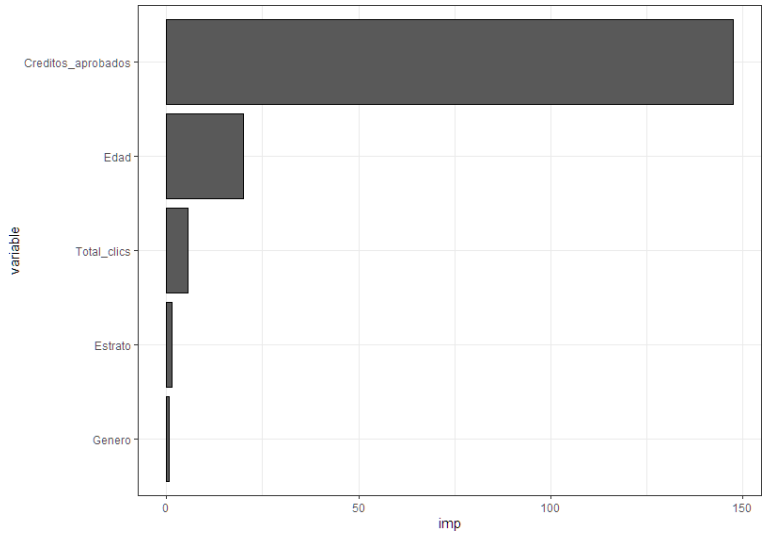
Tabla 26. Análisis de independencia - Caso Colombia

	Resultado final		
	X-squared	df	Valor p
Edad	6.3258	10	0.7872
Estrato	3.6534	5	0.6003
Género	0.64908	1	0.4204
Créditos aprobados	750.36	130	< 2.2e-16
Total clics	328.59	104	< 2.2e-16

Fuente: elaboración del autor

Con un nivel de significación de 0.05, no se rechaza la hipótesis nula en los tres primeros escenarios, es decir hay independencia, en el caso de Colombia el resultado final no está relacionado con la edad, el estrato o el género. Se rechaza la hipótesis nula para los dos escenarios finales, es decir, el resultado final está relacionado con el total de créditos aprobados y el total de clics. A continuación, se presenta el análisis de importancia de la aplicación de la técnica de árboles de decisión. Ver Figura 8:

Figura 8. Importancia de las variables – Caso Colombia



Fuente: elaboración del autor

La cantidad de créditos aprobados es la variable de mayor importancia, esto obedece a que los estudiantes que han superado mayor cantidad de créditos tienen más experiencia y adaptación a la vida universitaria que los estudiantes de primeros semestres que no han superado una cantidad de créditos considerable. Posteriormente, la edad, el total de clics, el estrato y el género.

5.6 Conclusiones del capítulo

No se puede afirmar con certeza que una técnica de predicción sea mejor que otras para predecir el rendimiento académico en CUVD, la cantidad de información que se tiene disponible, las variables analizadas y la consistencia interna de los datos juegan un papel de gran importancia. Verificar que los modelos hayan sido entrenados con los parámetros óptimos es un factor de gran relevancia para poder garantizar la fiabilidad de las predicciones. En las bases de datos de CUVD analizadas en los tres países se pudo evidenciar que los valores de las métricas de calidad eran diferentes; a pesar de que el modelo óptimo identificado por la función GridSearchCv de Python era igual para Inglaterra y Uruguay, los valores de las métricas de Accuracy, F1-Score, Recall y Precision diferían considerablemente; es necesario señalar que la base de datos de Inglaterra era la que tenía

el mayor tamaño muestral; sin embargo, las métricas de calidad fueron considerablemente inferiores a las de los otros dos países analizados.

Respecto al análisis de independencia entre la variable objetivo resultado final y las demás variables independientes en las tres bases de datos de CUVD analizadas en este trabajo, en el caso de Inglaterra el resultado final está relacionado con la edad, el índice de necesidades insatisfechas (IPM) del lugar donde el estudiante desarrolla el CUVD, el género, los créditos que ha aprobado, el total de clics, la región de donde proviene y por la presencia o ausencia de discapacidades.

En el caso de Uruguay la variable de mayor importancia para determinar si un estudiante aprobará o no es el porcentaje de cumplimiento con lo obligatorio, esto obedece a que los estudiantes que han cumplido con todas las asignaciones han dedicado mayor cantidad de tiempo de estudio personal, lo que se ve reflejado en un buen desempeño académico.

Por último, en el caso de Colombia el resultado final no está relacionado con la edad, el estrato o el género, la variable de mayor importancia para determinar si un estudiante aprobará o no es el total de créditos aprobados y el total de clics, al igual que en el caso de Inglaterra.

6 Conclusiones

En esta sección se da respuesta a la pregunta de investigación y se detalla la manera como se dio cumplimiento a los objetivos general y específicos planteados en esta investigación:

6.1 Respuesta a la pregunta de investigación

La pregunta de investigación planteada en esta tesis doctoral fue:

> ¿Es posible desarrollar una metodología para predecir el desempeño académico (Aprueba / Reprueba) estudiantil en cursos universitarios virtuales a distancia (CUVD)?

La respuesta es que sí es posible, esta está conformada por los siguientes pasos; Determinación de las variables a analizar, Construcción de la base de datos, Construcción de los modelos de predicción, Evaluación de los modelos y Visualización de la predicción. Como se mencionó anteriormente, en el caso de aplicación de Inglaterra y Uruguay no fue posible realizar el último paso de la metodología, ya que no fue posible ingresar a las PEV de las universidades en donde se desarrollaron los CUVD. En el Anexo 1 se detalla la manera como se realizó este pasó en el caso colombiano.

La determinación de las variables a analizar tiene como propósito identificar qué tipo de información identificada como relevante en la RSL se debe conocer de los estudiantes. La construcción de la base de datos tiene como objetivo depurar la información obtenida, categorizar variables cualitativas y completar valores faltantes para evitar el ruido y garantizar la calidad de las predicciones. La construcción de los modelos de predicción tiene como propósito entrenar un conjunto de modelos empleando diversas técnicas de ML con los parámetros óptimos. La evaluación de los modelos tiene como objetivo identificar el mejor en términos de precisión, comparando todas las métricas (Accuracy, F1-Score, Recall y Precision) y la visualización de la predicción tiene como propósito que los docentes identifiquen a tiempo los estudiantes en riesgo de Reprobar, para que implementen estrategias encaminadas a evitar que reprueben o deserte del CUVD.

6.2 Cumplimiento de objetivos

A continuación, se detalla la manera como se dio cumplimiento al objetivo general y a los objetivos específicos planteados en esta investigación.

6.2.1 Cumplimiento del Objetivo Específico 1

> Determinar los factores clave que causan un bajo rendimiento académico estudiantil en CUVD

Se dio cumplimiento al objetivo en el Capítulo 3, se desarrolló una RSL enfocada a determinar los factores que contribuyen a que los estudiantes de CUVD tengan un bajo rendimiento académico, en resumen, estos factores se agrupan en: físicos, ambientales y tecnológicos, personales y familiares, psicológicos, académicos e institucionales. Debido a la complejidad en la adquisición de información de los estudiantes relacionada con problemas o situaciones personales (problemas de salud, económicos, sentimentales, familiares, etc.), como trabajo futuro se propone la creación de una plataforma orientada a la web que permita registrar problemáticas a los estudiantes de los CUVD para que los docentes puedan consultarla cuando la necesiten. Es importante señalar que es deseable que las universidades destinen recursos para apoyar a los estudiantes de CUVD que tengan problemas económicos, ya sea brindándoles bonos de apoyo alimentario, de acceso a internet o facilitarles un ordenador o herramienta tecnológica en caso de que no posean.

6.2.2 Cumplimiento del Objetivo Específico 2

> Determinar la técnica adecuada para predecir el rendimiento académico estudiantil en CUVD

Se dio cumplimiento al objetivo en el Capítulo 4 se presentan los resultados de una RSL enfocada a identificar las técnicas que se han empleado para predecir el rendimiento académico estudiantil en CUVD. En resumen, las técnicas identificadas fueron; Regresión Lineal Simple y Múltiple, OneR (Una Regla), PART, J48, Bosques Aleatorios, Análisis Discriminante Lineal (LDA), K-Vecinos Próximos, GBM, Redes Neuronales Artificiales (RNA – MLP Perceptrón Multicapa y Árboles de Decisión. No se puede afirmar con certeza que una técnica sea mejor que otra, la cantidad y calidad de los datos juega un papel importante.

Se debe realizar una selección comparando los puntajes obtenidos con los parámetros óptimos de cada modelo para poder identificar el mejor; además, es necesario considerar otras métricas de precisión como; F1-Score, Recall y Precision, en caso tal de que las muestras estén desbalanceadas (mayor cantidad de estudiantes que Aprueban en comparación con los que Reprueban y viceversa).

6.2.3 Cumplimiento del Objetivo Específico 3

> Construir un modelo para predecir el rendimiento académico estudiantil en CUVD

Se dio cumplimiento al objetivo en el Capítulo 6 se aplica la metodología propuesta en tres bases de datos de CUVD de tres países; Inglaterra, Uruguay y Colombia; además, se construyó un prototipo funcional compatible con Moodle (Ver Anexo 1) el cual permite aplicar la metodología propuesta. Se concluye que, la metodología formulada permite mejorar la fiabilidad de la predicción del rendimiento académico de estudiantes de los CUVD, ya que involucra aspectos relacionados con la vida personal de los mismos, considerando que el rendimiento académico estudiantil en este tipo de cursos depende de diversos factores. Además, la metodología contempla que las bases de datos tengan consistencia interna en cuanto al área de formación de los estudiantes, al grado de dificultad de las asignaturas y al nivel o semestre en el que se dictan, no como lo hacen en la mayoría de investigaciones en donde agrupan la información de los estudiantes para poder tener un tamaño de muestra grande y así poder obtener valores altos en las métricas de precisión (Accuracy, F1-Score, Recall y Precision).

6.2.4 Cumplimiento del Objetivo General

> Desarrollar una metodología para predecir el desempeño académico (Aprueba / Reprueba) estudiantil en cursos universitarios virtuales a distancia (CUVD).

Se dio cumplimiento al objetivo en el Capítulo 5 en donde se detallan cada uno de los cinco pasos que la conforman; Determinación de las variables a analizar, Construcción de la base de datos, Construcción de los modelos de predicción, Evaluación de los modelos y Visualización de la predicción; además, en cada paso se presentan recomendaciones en

pro de mejorar la precisión de los modelos. Uno de los factores relevantes de esta metodología es la posibilidad de incorporar información personal de los estudiantes y de su entorno, a diferencia de la mayoría de investigaciones identificadas en la revisión del estado del arte, en las cuales se enfocan solo en información que puedan extraer de las PEV.

6.3 Trabajo futuro

Futuras investigaciones pueden enfocarse a la predicción del rendimiento académico de estudiantes de los CUVD antes del inicio de los mismos, esta temática no se ha abordado antes y requiere de una atención especial, ya que los docentes al identificar a los posibles estudiantes que tendrán un bajo rendimiento antes de que comiencen a dictar los CUVD pueden diseñar actividades que involucren diferentes estilos de aprendizaje, implementar estrategias pedagógicas y didácticas que incentiven la motivación y el autoaprendizaje y remitir a las unidades de bienestar universitario o permanencia académica de ser necesario.

7 Anexo 1: Instalación y configuración del asistente virtual

Este capítulo presenta el caso de aplicación de la metodología presentada en el Capítulo 5. Como se mencionó anteriormente, se desarrollaron dos asistentes virtuales, el primero para el experto denominado Cyber-CUVD Xpert y el segundo para los docentes Cyber-CUVD Teacher. Ambos asistentes son compatibles con el sistema de gestión de aprendizaje Moodle, y pueden ser integrados a cualquier CUVD como paquetes SCORM (Gonen & Basaran, 2008). Para ello solo es necesario ingresar al campus o plataforma educativa virtual (PEV) de su universidad, activar edición y añadir una actividad o recurso, como se presenta a continuación:

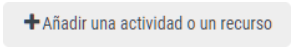
- Paso 1: Después de haber ingresado dar clic en Activar edición:

Figura 9. Importación I

Fuente: elaboración del autor

- **Paso 2:** En la sección del curso que desee, dar clic en Añadir una actividad o u recurso:

Figura 10. Importación II



Fuente: elaboración del autor

- **Paso 3:** Seleccionar la opción Paquete SCORM:

Figura 11. Importación III

Fuente: elaboración del autor

- **Paso 4**: Añadir los archivos SCORM disponibles en los siguientes enlaces:

  Cyber-CUVD Xpert:

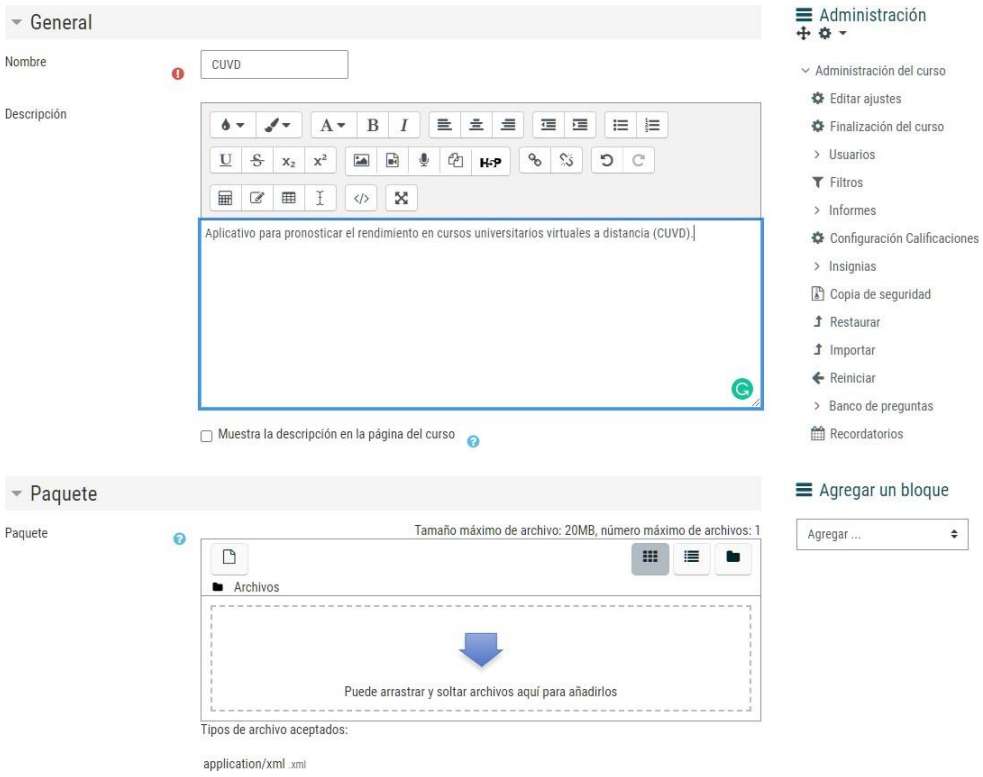
  https://drive.google.com/file/d/1AiI0MbwTdu0JjjbOPsMm9DnWWn42jz-l/view?usp=sharing

  Cyber-CUVD Teacher:

  https://drive.google.com/file/d/1BSf5-Y4Y_Uumo7xtTOjHOku_a22R4QkP/view?usp=sharing

- **Paso 5**: Diligenciar la información solicitada en cada uno de los campos:

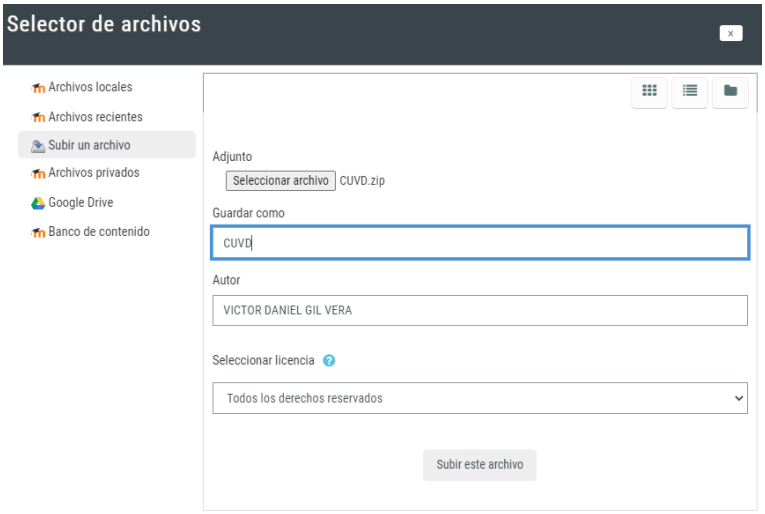
Figura 12. Importación IV



Fuente: elaboración del autor

- **Paso 6**: Dar clic en Subir este archivo:

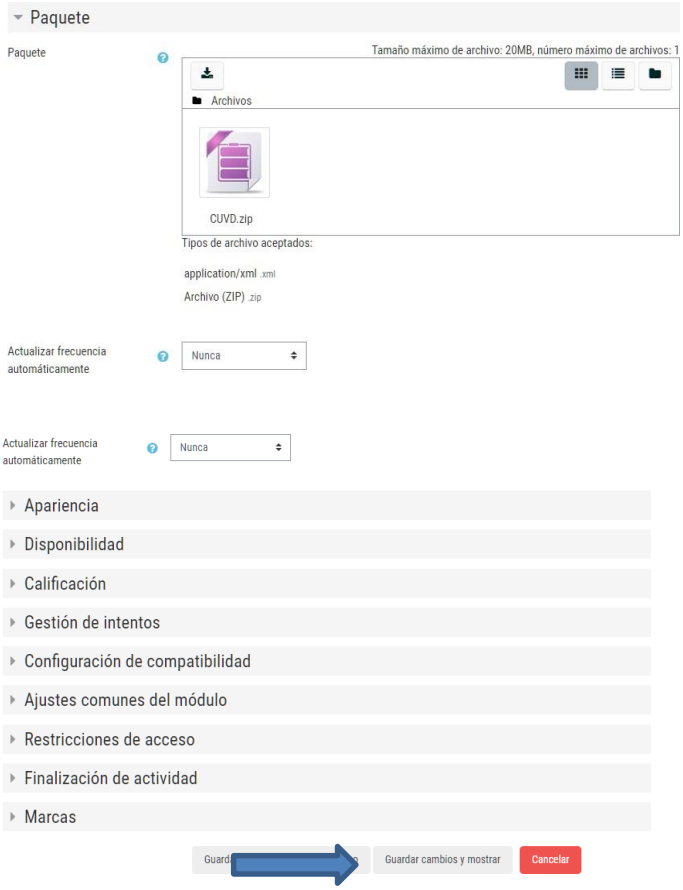
Figura 13. Importación V



Fuente: elaboración del autor

- **Paso 7**: Cuando termine de subir el archivo deje los campos de configuración por defecto y dé clic en Guardar cambios y mostrar:

Figura 14. Importación VI



Fuente: elaboración del autor

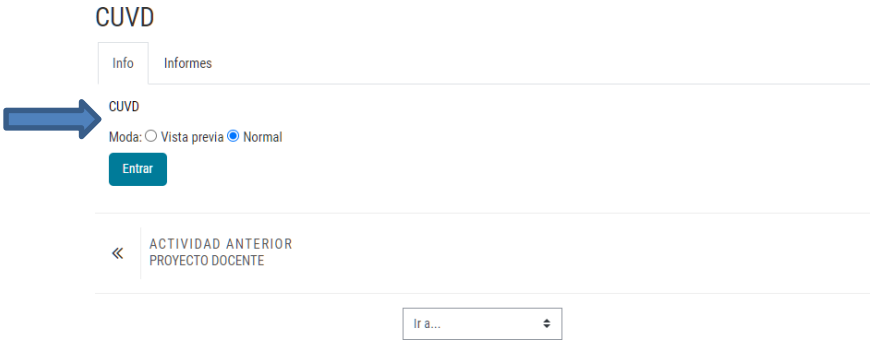
- **Paso 8**: Ingrese a Cyber-CUVD Xpert o a Cyber-CUVD Teacher dependiendo de su rol haciendo doble clic sobre el ícono correspondiente:

Figura 15. Acceso I

Fuente: elaboración del autor

- **Paso 9**: Seleccione Normal y luego presione Entrar y listo:

Figura 16. Acceso II

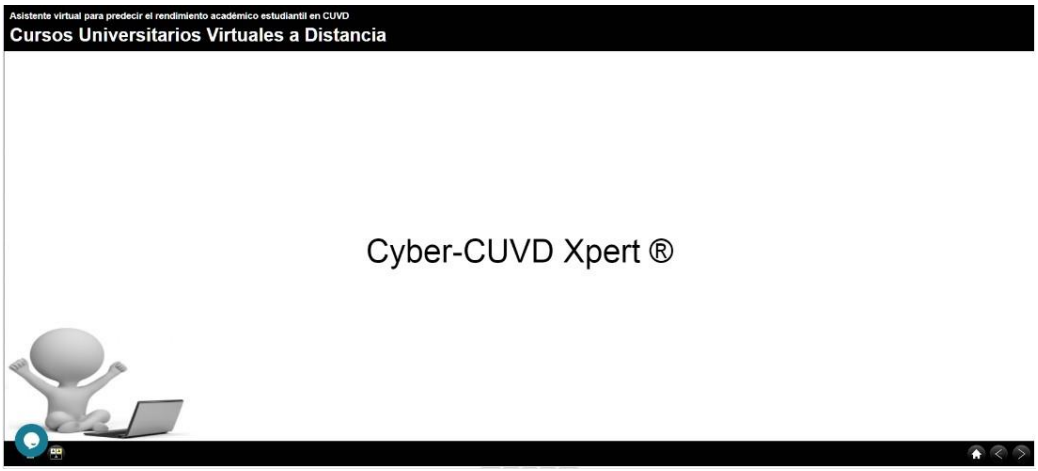


Fuente: elaboración del autor

7.1 Modo de uso del asistente virtual

Para navegar en Cyber-CUVD Xpert presione las flechas que se presentan en la parte inferior derecha. El ícono del home retorna a la pantalla inicial, ver Figura 17:

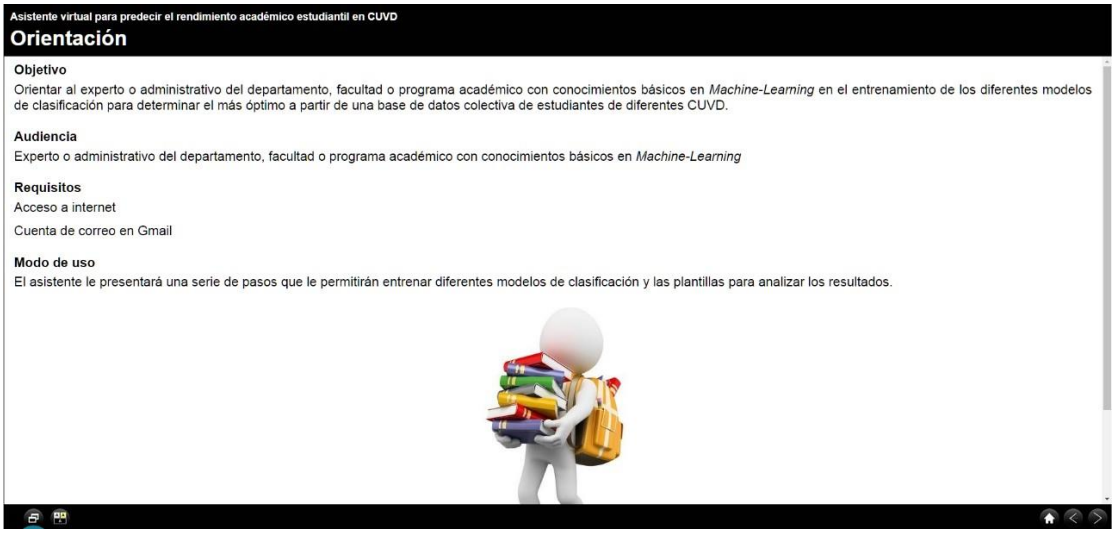
Figura 17. Cyber – CUVD Xpert I.



Fuente: elaboración del autor

En este apartado se presenta la orientación, el objetivo, el tipo de audiencia a la que va dirigido, los requisitos y cómo utilizarlo:

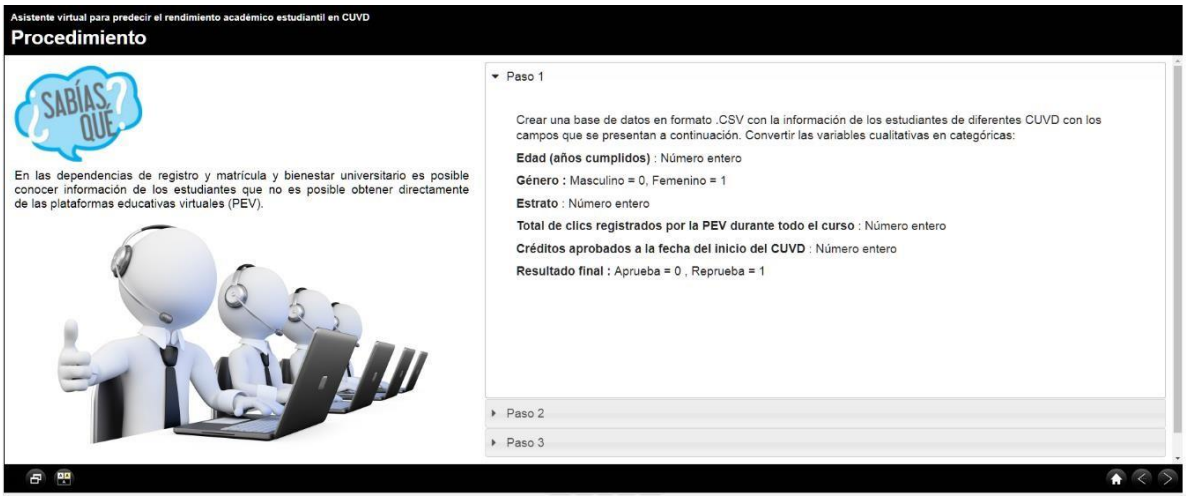
Figura 18. Cyber – CUVD Xpert II



Fuente: elaboración del autor

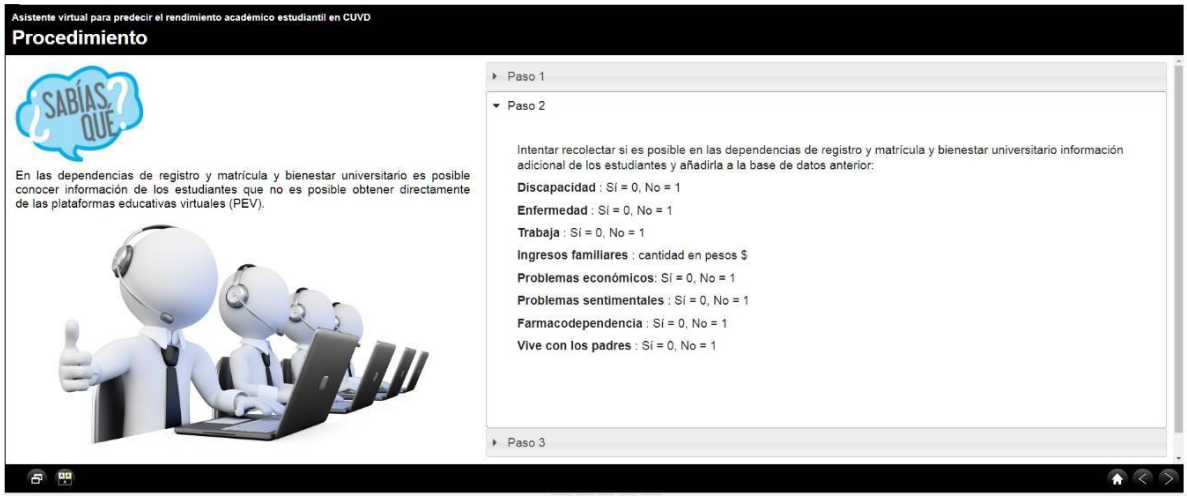
El asistente está conformado por 12 pasos. El experto debe desarrolle cada paso en orden y acceder a los enlaces con el navegador Google Chrome:

Figura 19. Paso 1 - Cyber CUVD Xpert.



Fuente: elaboración del autor

Figura 20. Paso 2 - Cyber CUVD Xpert.

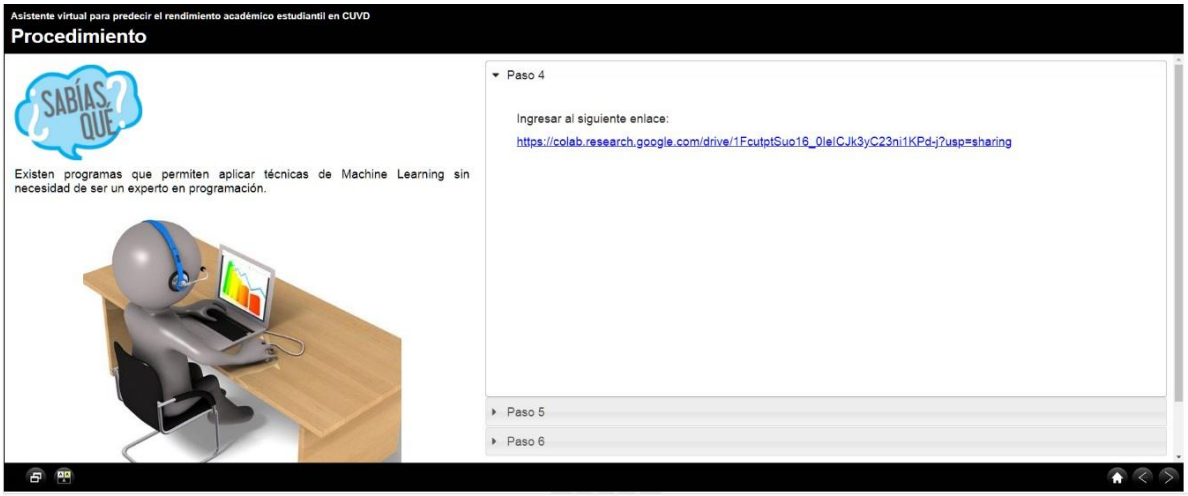


Fuente: elaboración del autor

Figura 21. Paso 3 - Cyber CUVD Xpert.

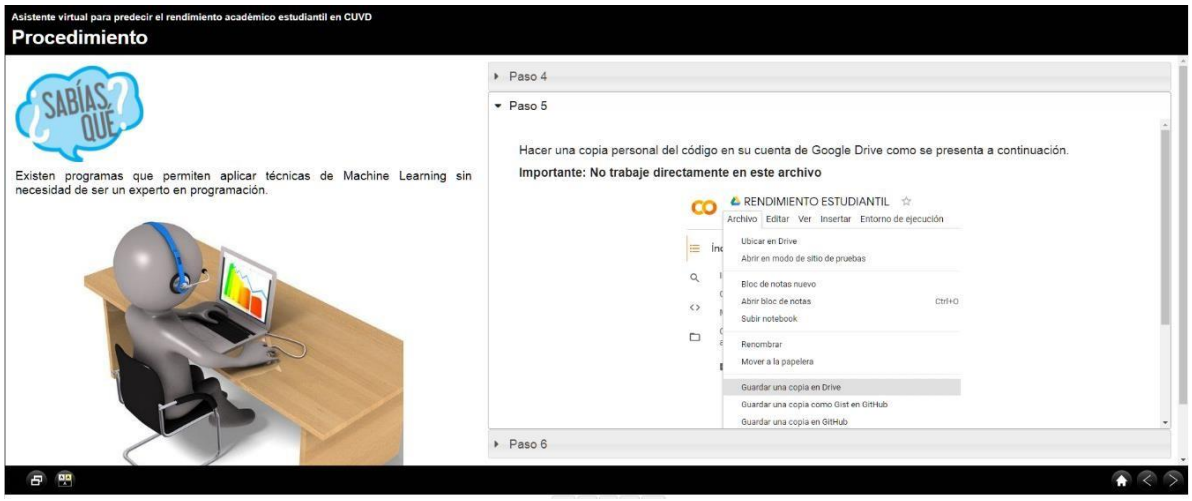
Fuente: elaboración del autor

Figura 22. Paso 4 - Cyber CUVD Xpert.



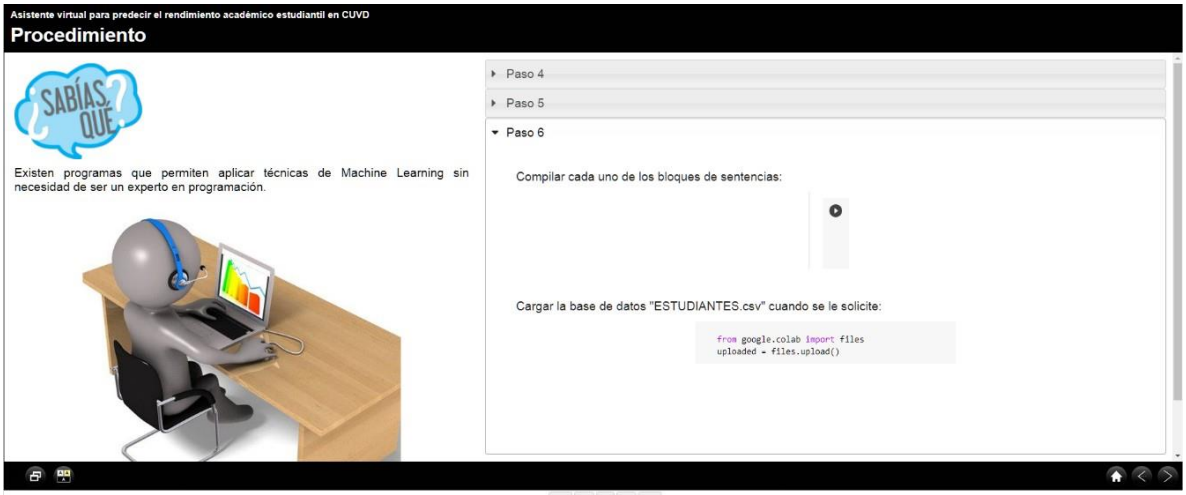
Fuente: elaboración del autor

Figura 23. Paso 5 - Cyber CUVD Xpert.



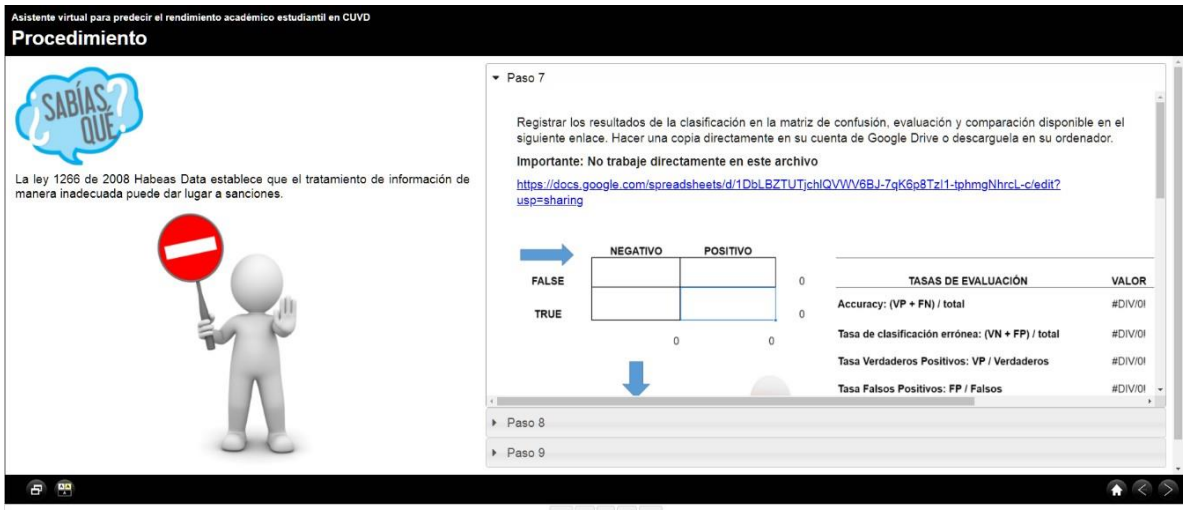
Fuente: elaboración del autor

Figura 24. Paso 6 - Cyber CUVD Xpert.



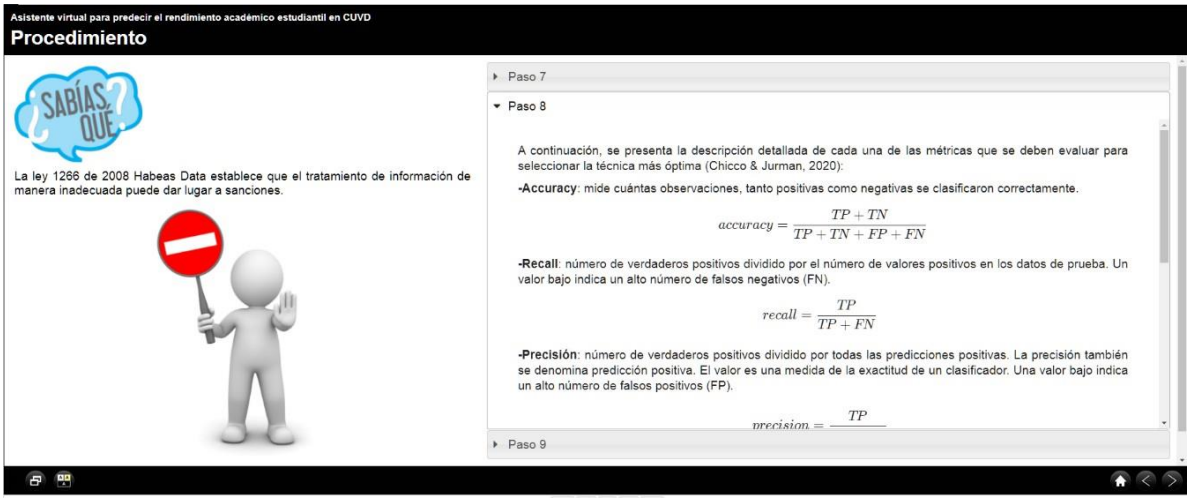
Fuente: elaboración del autor

Figura 25. Paso 7 - Cyber CUVD Xpert.



Fuente: elaboración del autor

Figura 26. Paso 8 - Cyber CUVD Xpert.



Fuente: elaboración del autor

Figura 27. Paso 9 - Cyber CUVD Xpert.

Fuente: elaboración del autor

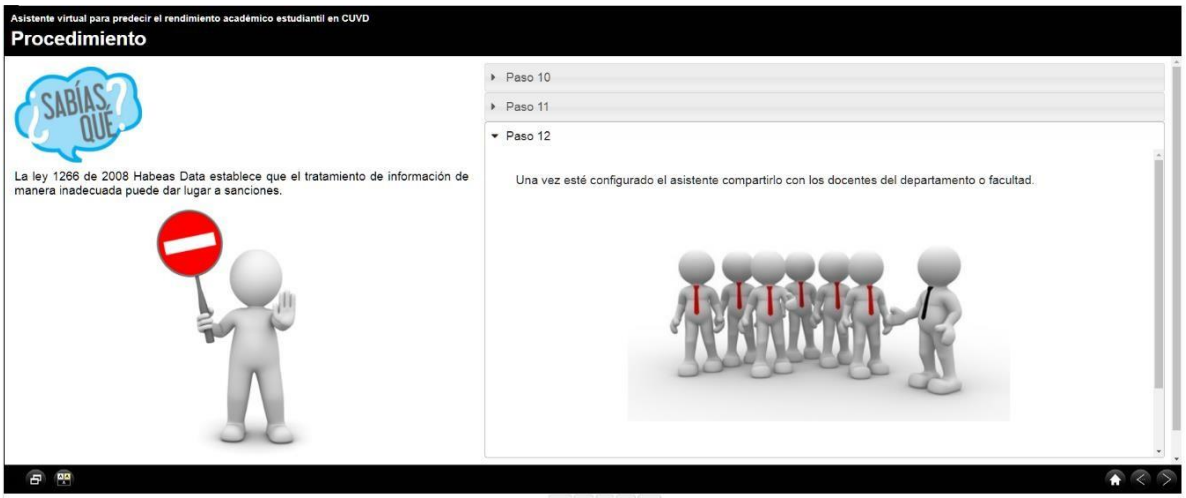
Figura 28. Paso 10 - Cyber CUVD Xpert

Fuente: elaboración del autor

Figura 29. Paso 11 - Cyber CUVD Xpert

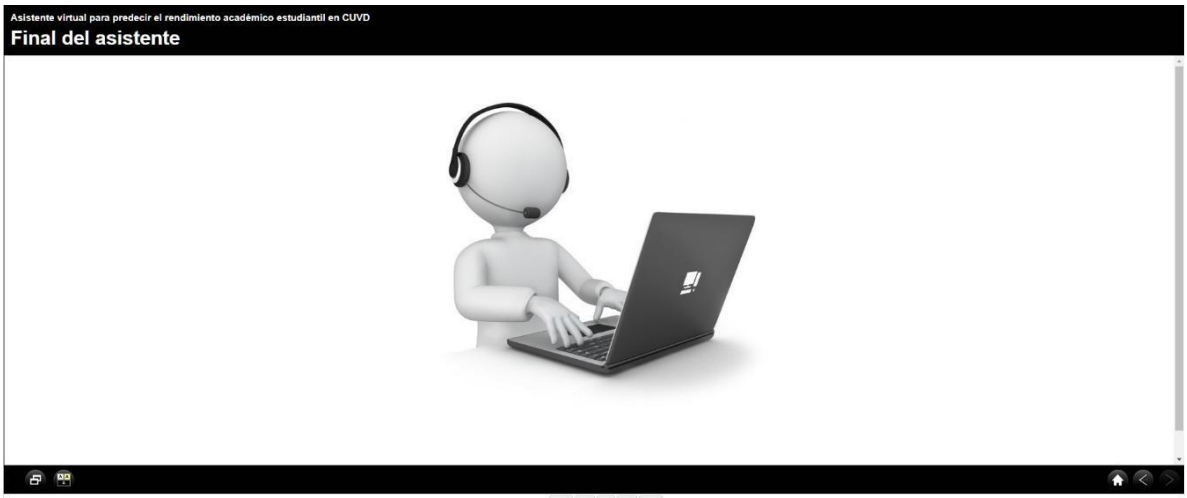
Fuente: elaboración del autor

Figura 30. Paso 12 Cyber – CUVD Xpert



Fuente: elaboración del autor

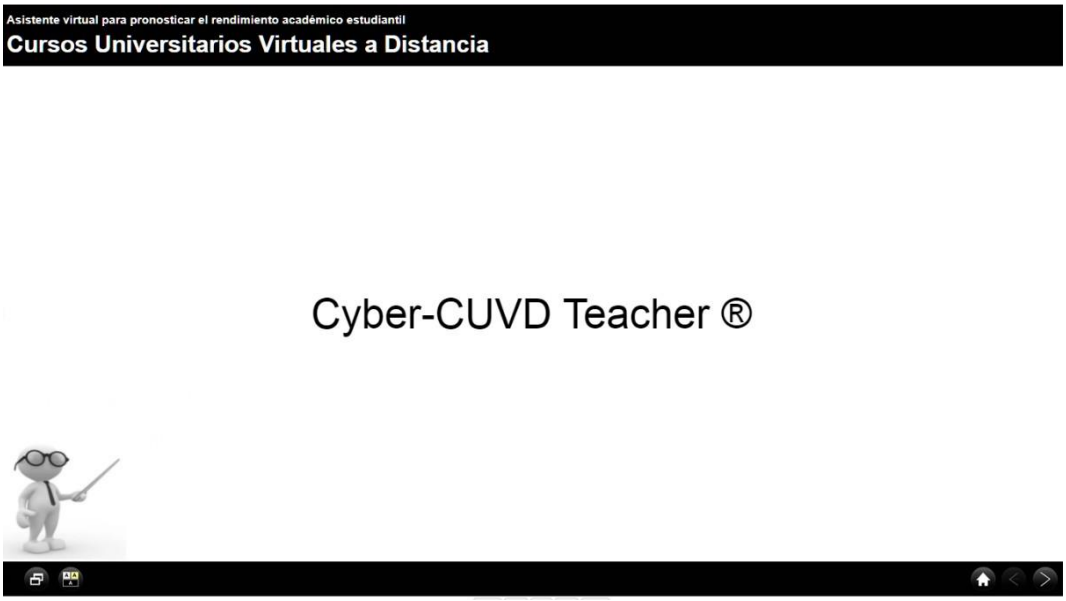
Figura 31. Fin Cyber – CUVD Xpert



Fuente: elaboración del autor

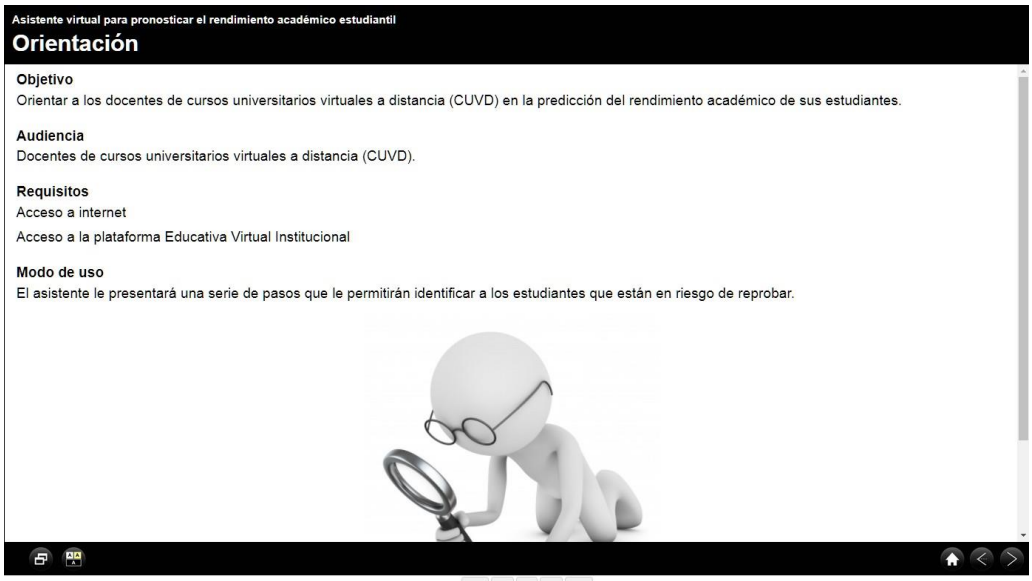
A continuación, se presenta el asistente para el experto Cyber-CUVD Teacher:

Figura 32. Cyber – CUVD Teacher I



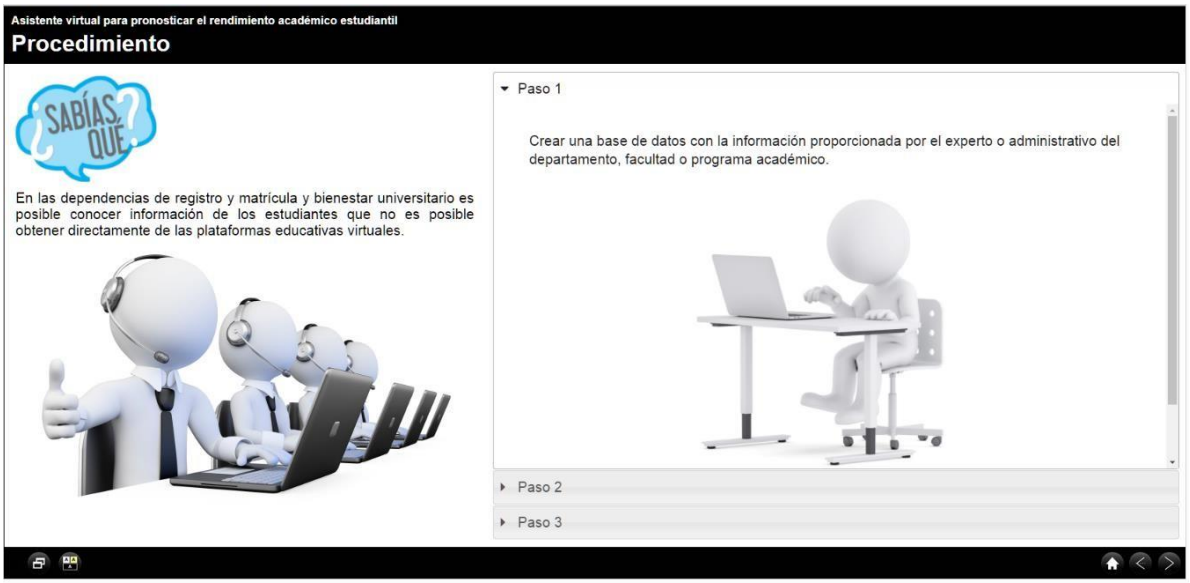
Fuente: elaboración del autor

Figura 33. Cyber - CUVD Teacher II



Fuente: elaboración del autor

Figura 34. Cyber - CUVD Teacher III



Fuente: elaboración del autor

Figura 35. Cyber – CUVD Teacher IV

Fuente: elaboración del autor

Figura 36. Paso 3 Cyber – CUVD Teacher V

Fuente: elaboración del autor

Figura 37. Paso 4 Cyber – CUVD Teacher VI

Fuente: elaboración del autor

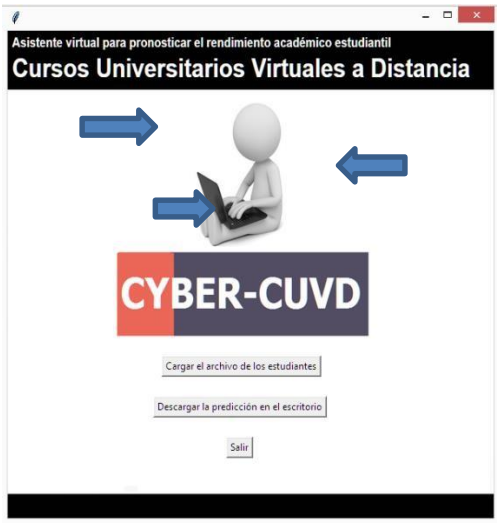
Figura 38. Paso 5 Cyber – CUVD Teacher VII

Fuente: elaboración del autor

El asistente para los docentes tiene tres botones; uno para cargar la base de datos de los estudiantes, otro para descargar la base de datos con el resultado de la predicción y otro para salir. El docente debe descargar la base de datos directamente de la plataforma virtual que emplee la Universidad o IES con los campos que el experto haya configurado por

defecto. Se debe aclarar que, en caso de que el docente obtenga información adicional de los estudiantes y el experto no la incluyó, tampoco se debe considerar al momento de hacer la predicción. Ver Figura 42:

Figura 39. Cyber – CUVD Teacher VIII



Fuente: elaboración del autor

A la base de datos original el asistente le añade una columna con el resultado de la predicción. El docente debe identificar a los que fueron clasificados con la etiqueta Reprueba y tratar de comunicarse con ellos, ya sea a través de un correo electrónico u otro medio de comunicación. Ver Figura 44:

Figura 40. Cyber – CUVD Teacher IX

Edad	Estrato	Genero	Creditos aprobados	Clics	Predicción
20	5	0	50	71	Aprueba
21	3	0	45	94	Aprueba
20	5	1	23	43	Reprueba
24	2	1	49	78	Reprueba
24	3	0	70	156	Aprueba
25	2	1	50	76	Reprueba
19	2	0	42	112	Aprueba
24	2	0	31	43	Reprueba
21	2	0	80	120	Aprueba
23	3	1	45	78	Reprueba
24	5	0	75	200	Aprueba
22	5	1	65	110	Aprueba
19	5	1	40	80	Aprueba
21	4	1	12	66	Reprueba
22	5	0	45	110	Aprueba
19	5	1	49	99	Aprueba
22	4	0	46	70	Reprueba
20	2	1	16	170	Reprueba
24	2	1	45	71	Reprueba
20	4	1	45	75	Aprueba
24	3	1	21	115	Reprueba
20	5	0	23	56	Reprueba
23	2	1	46	124	Aprueba
22	4	1	50	76	Aprueba
24	4	1	45	130	Aprueba
20	4	1	11	53	Reprueba
24	5	0	41	26	Reprueba
24	5	1	49	37	Reprueba
20	2	1	57	140	Aprueba
22	4	1	43	78	Reprueba
23	4	1	43	72	Reprueba
24	4	1	57	157	Aprueba
22	4	0	47	71	Aprueba
23	3	1	42	47	Reprueba
25	4	1	35	157	Reprueba

Fuente: elaboración del autor

Al finalizar, el asistente le presentará recomendaciones generales para contribuir a mejorar el rendimiento académico de los estudiantes de CUVD. Ver Figura 46 y Figura 47:

Figura 41. Recomendaciones Cyber – CUVD Teacher VIII

Fuente: elaboración del autor

Figura 42. Recomendaciones Cyber – CUVD Teacher IX

Fuente: elaboración del autor

Se debe aclarar que, los cursos CUVD sobre los que se puede emplear el asistente deben tener fechas de inicio y finalización fijas, no está diseñado para ser utilizado con CUVD de inscripción continua.

Se recomienda emplear la matriz de confusión que se presenta en el asistente para analizar la precisión del modelo (Falsos Positivos (FP), Falsos Negativos (FN), Verdaderos Negativos (VN) y Verdaderos Positivos (VP)) y calcular las tasas de evaluación estándar, como se presenta en la Tabla 27 y Tabla 28:

Tabla 27. Matriz de confusión estándar

	Predicción	
Real	Negativo	Positivo
Falso	TN = VN	FP = FP
Verdadero	FN = FN	TP = VP

Fuente: elaboración del autor

Tabla 28. Tasas de evaluación estándar

Tasa	Utilidad	Fórmula
Accuracy	¿Con qué frecuencia es correcto el clasificador?	(TP + TN) / total
Clasificación errónea	¿Con qué frecuencia es incorrecta?	(FP + FN) / total
Verdaderos positivos	¿Con qué frecuencia predice que sí?	TP / Actual Sí
Falsos positivos	¿Con qué frecuencia predice que sí?	FP / Actual No
Negativa verdadera	¿Con qué frecuencia predice que no?	TN / Actual No
Precisión	¿Con qué frecuencia es correcto?	TP / Predicho Sí
Prevalencia	¿Con qué frecuencia ocurre realmente la condición de sí en nuestra muestra?	Actual Sí / Total

Fuente: elaboración del autor

7.2 Recomendaciones de uso

Se recomienda que los docentes revisen semanalmente los ingresos de los estudiantes a la PEV. Si bien la metodología funciona hay información que no se tiene a la mano y que es importante, considerando que el rendimiento académico estudiantil es un resultado multifactorial. No puede afirmarse con certeza que el uso de PEV garanticen un mejor desempeño académico. Futuras investigaciones deben analizar si el uso de PEV propician un cambio positivo en el aprendizaje o no, y evaluar su efecto sobre el rendimiento académico de manera precisa y objetiva. Entre las diferentes estrategias que pueden implementar los docentes para mejorar el rendimiento académico estudiantil de estudiantes de CUVD se proponen las siguientes:

- Periodicidad entre ingresos: revisar semanalmente el último acceso de los estudiantes a la plataforma e identificar los que registren intervalos de tiempo considerables y enviarles un mensaje a través de la plataforma.
- Interacción sincrónica: crear espacios de escucha de manera sincrónica con los estudiantes para compartir experiencias, problemas, dificultades y demás.
- Generación de alertas: enviar recordatorios sobre la realización de trabajos o exámenes a través de la plataforma con anticipación.
- Uso de recursos interactivos: crear herramientas que incentiven el aprendizaje autónomo de los estudiantes como Objetos Virtuales de Aprendizaje (OVA).
- Apoyo a estudiantes: identificar estudiantes que tengan problemas para acceder a internet o que no dispongan de equipos de computación para que busquen apoyo en las dependencias de bienestar universitario.
- Retroalimentación: socializar con los estudiantes los resultados de las actividades evaluativas de manera oportuna.

8 Anexo 2: Definiciones y Conceptos Básicos

A continuación, se definen brevemente algunas definiciones y conceptos básicos que aparecen a lo largo del documento:

Adivinación: afirmar un suceso o acontecimiento futuro sin utilizar procedimientos basados en la razón ni en conocimientos científicos (Bastús, 1855).

Algoritmo: secuencia finita de instrucciones definidas e implementables por el ordenador para resolver un problema o un cálculo (Webster, 2006).

Análisis del aprendizaje: medición, recopilación, análisis y comunicación de datos sobre los estudiantes y sus contextos, con el fin de comprender y optimizar el aprendizaje y los entornos en los que se produce (Siemens, 2010).

Análisis predictivo: tipo de análisis que emplea datos históricos para predecir eventos futuros; los datos históricos son utilizados para crear un modelo matemático que capture las tendencias importantes (MathWorks, 2021).

Árboles de decisión: técnica de ML que consiste en agrupar elementos en función de las posibles interacciones que se pueden formar con variables explicativas; en cada categoría resultante la frecuencia observada se mide con respecto a la variable objetivo, luego, se abren las ramas según las frecuencias más relevantes, repitiendo el proceso hasta encontrar categorías en las que las frecuencias resultantes ya no son relevantes (Li et al., 2013).

Blended Learning (B-Learning): enfoque educativo que combina recursos e interacción en línea con los métodos tradicionales de las clases presenciales, requiere la presencia física del docente y del estudiante (Banditvilai, 2016).

Bosques Aleatorios: técnica de aprendizaje automático supervisado basada en árboles de decisión, los cuales se ensamblan por medio de bolsas; los árboles se entrenan de forma independiente (Golden et al., 2019).

Curso sincrónico: modalidad de aprendizaje en la cual el docente y el estudiante interactúan en tiempo real, independiente de que se encuentren en espacios físicos diferentes (Moyano, 2018).

Curso virtual: modalidad de aprendizaje en la cual los docentes y estudiantes participan en un entorno digital a través de nuevas tecnologías y redes de computadoras, haciendo uso intensivo de las facilidades que proporciona la Internet y las tecnologías de la información y la comunicación (TIC) (Moyano, 2018).

Curso asincrónico: modalidad de aprendizaje en la cual los estudiantes aprenden estando desconectados a través de videos, materiales o recursos educativos previamente proporcionados por el docente (Moyano, 2018).

Curso híbrido: modalidad de aprendizaje en la cual los componentes en línea están destinados a reemplazar una parte del tiempo de las clases presenciales; las interacciones en línea pueden ser sincrónicas, es decir, los estudiantes interactúan en tiempo real a través de sesiones de clase realizadas mediante Webinars (AMIR, 2020)

Curso telepresencial: curso que se desarrolla en un horario determinado de manera virtual con un docente o experto, con contenidos y objetivos formativos idénticos a los de la formación presencial (Nanfor, 2021).

Content Management System (CMS): programa informático que permite crear un entorno de trabajo para la creación y administración de contenidos, principalmente en páginas web (Mauthe & Thomas, 2004).

Deep Learning: conjunto de algoritmos de aprendizaje automático que intentan modelar abstracciones de alto nivel en datos usando arquitecturas computacionales que admiten transformaciones no lineales múltiples e iterativas de datos expresados en forma matricial o tensorial (Bengio et al., 2013).

e-Learning: enseñanza y aprendizaje en línea a través de la Internet y recursos tecnológicos (D.-A. Gómez-Aguilar et al., 2014)

Flipped classroom: técnica pedagógica que plantea la necesidad de transferir parte del proceso de enseñanza y aprendizaje fuera del aula con el fin de utilizar el tiempo de

clase para el desarrollo de procesos cognitivos de mayor complejidad que favorezcan el aprendizaje (Tucker, 2012).

Google Colaboratory: plataforma virtual que permite la escritura y ejecución de código elaborado en Python, permite entrenar técnicas de aprendizaje automático, análisis de datos y educación. Es un servicio de notebook alojado de Jupyter que no requiere configuración para usarlo y brinda acceso gratuito a recursos computacionales, incluidas GPU (Google, 2021).

Inteligencia Artificial: inteligencia demostrada por máquinas que involucra conciencia y emocionalidad (Mutascu, 2021).

K-Vecinos Próximos (K-NN): técnica de ML que almacena los casos disponibles y clasifica los casos nuevos con base en una medida de similitud (funciones de distancia) (Gray et al., 2014).

Learning Management System (LMS): software instalado en un servidor web que permite administrar, distribuir y controlar actividades de formación no presencial de una institución u organización (Subirà & Catasús, 2011).

Locus de control: grado en que las personas sienten que tienen el control de lo que ocurre en sus vidas, desde un evento rutinario hasta una situación de peligro Padua Rodríguez (2019).

M-Learning: forma de aprendizaje que facilita la construcción del conocimiento, la resolución de problemas y el desarrollo de destrezas y habilidades diversas de manera autónoma y ubicua, gracias a la mediación de dispositivos móviles (Ehsanpur & Razavi, 2020).

Machine Learning: estudio de algoritmos informáticos que mejoran automáticamente a través de la experiencia y mediante el uso de datos (Mitchell, 1997).

Máquina de Vectores de Soporte (SVM): técnica de aprendizaje supervisado utilizada principalmente para clasificación (Sembiring et al., 2011).

Naïve-Bayes: técnica de clasificación basada en el Teorema de Bayes, esta técnica asume que la presencia de una característica particular en una clase no está relacionada con la presencia de ninguna otra; esta técnica utiliza todos los atributos contenidos en los

datos, luego, analiza cada uno de ellos para mostrar la importancia e independencia de cada variable (Osmanbegovic y Suljic, 2012).

Técnica: función que utiliza datos de entrenamiento como entrada para generar una salida (Jones, 2018).

Objeto virtual de aprendizaje (OVA): colección de elementos de contenido, práctica y valoración virtual que se combinan en función de un solo objetivo de aprendizaje (Polsani, 2003).

Perceptrón Multicapa (MLP): tipo de RNA que se utiliza para clasificar conjuntos que no son linealmente independientes, esta técnica es una extensión de un perceptrón al que se le agregan series de capas para transformar las variables de entrada (González-Briones et al., 2018).

Plataforma educativa virtual (PEV): herramienta que brinda la capacidad de interactuar, a través de la red, con fines pedagógicos. Complementan y presentan alternativas a las prácticas de educación tradicional (Zhang et al., 2020).

Predecir: estimar resultados ajustando una técnica a un conjunto de datos deentrenamiento, generando como resultado un estimador que puede hacer prediccionespara nuevas muestras (Döring, 2018).

Pronosticar: hacer predicciones sobre el futuro, tomando como base datos de series de tiempo; se diferencia de la predicción por la dimensión temporal (Döring, 2018).

Redes neuronales Artificiales (RNA): modelos paramétricos de regresión no lineal que emulan la forma en que el cerebro humano procesa la información, es decir, un gran número de unidades de procesamiento interconectadas las cuales desempeñan el rol de las neuronas biológicas, estas trabajan simultáneamente para procesar información (Corchado et al., 2005).

Red de Hopfield: tipo de RNA de una sola capa las cuales se utilizan principalmente en el reconocimiento de patrones, especialmente en economía (Hartati y El-Hawary, 2000) y en la creación de imágenes (Tatem et al., 2002).

Redes de Aprendizaje Competitivas: tipo de RNA en la cual las neuronas compiten entre sí para representar patrones, este tipo de RNA es una variación de un perceptrón

multicapa; sin embargo, debido a que su aplicación se realiza en matrices bidimensionales, son empleadas principalmente para tareas de visión por ordenador, como la clasificación y segmentación de imágenes, entre otras aplicaciones (Rivas et al., 2018).

Red de Kohonen: tipo de RNA que pertenece a la categoría de redes no supervisadas, se diferencia con otros tipos de RNA en el hecho de que las neuronas que representan patrones similares aparecen juntas en el espacio de salida, el cual puede ser unidimensional, una línea, bidimensional, un plano o n-dimensional (E. Corchado et al., 2011).

R cran: lenguaje y entorno de libre acceso para la computación estadística y gráfica; modelización lineal y no lineal, pruebas estadísticas, análisis de series temporales, clasificación, agrupación, etc. (R, 2021)

Técnica: en aprendizaje de máquinas hace referencia a un conjunto de procedimientos que permiten realizar predicciones futuras basadas en comportamientos o características analizadas en datos históricos etiquetados (Ikonomakis et al., 2005).

9 Referencias

Abu Tair, M. M., & El-Halees, A. M. (2012). Mining educational data to improve students' performance: a case study. *Mining Educational Data to Improve Students' Performance: A Case Study*, *2*(2).

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

Ahammed, F., & Smith, E. (2019). Prediction of students' performances using course analytics data: A case of water engineering course at the university of south australia. *Education Sciences*, *9*(3). https://doi.org/10.3390/educsci9030245

Al-Nofaie, H. (2020). Saudi University Students' Perceptions towards Virtual Education during COVID-19 Pandemic: A Case Study of Language Learning via Blackboard. *Arab World English Journal*, *11*(3), 4-20.

Al-Shehri, H., Al-Qarni, A., Al-Saati, L., Batoaq, A., Badukhen, H., Alrashed, S., Alhiyafi, J., & Olatunji, S. O. (2017). Student performance prediction using support vector machine and k-nearest neighbor. *2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE)*, 1-4.

Al-Sudani, S., & Palaniappan, R. (2019). Predicting students' final degree classification using an extended profile. *Education and Information Technologies*, *24*(4), 2357-2369. https://doi.org/10.1007/s10639-019-09873-8

Alarcón, G. G., Añorve, J. R., Sánchez, M. del R. G., & Salgado, T. A. (2016). Los factores psicosociales como impacto en el bajo rendimiento escolar de los estudiantes de la Universidad Autónoma de Guerrero/Psychosocial factors as impact on poor school performance of the students of the Autonomous University of Guerrero. *RIDE Revista Iberoamericana Para La Investigación y El Desarrollo Educativo*, *7*(13), 107-125.

AMIR. (2020). *Curso Híbrido*. https://academiamir.com/cursos-amir/hibrido/

Angeline, D. M. D. (2013). Association rule generation for student performance analysis using apriori algorithm. *The SIJ Transactions on Computer Science Engineering & Its Applications (CSEA)*, *1*(1), 12-16.

Arsad, P. M., Buniyamin, N., & Manan, J. A. (2013). A neural network students' performance prediction model (NNSPPM). *2013 IEEE International Conference on Smart Instrumentation, Measurement and Applications (ICSIMA)*, 1-5.

Baggs, T., Barnett, D., & McCullough, K. (2015). The value of traditional cognitive variables for predicting performance in graduate speech-language pathology programs. *Journal of Allied Health*, *44*(1), 10-16. https://www.scopus.com/inward/record.uri?eid=2-s2.0-84924898093&partnerID=40&md5=3c7591fefc8ae763d803459a9a87dc00

Banditvilai, C. (2016). Enhancing students language skills through blended learning. *Electronic Journal of E-Learning*, *14*(3), pp223-232.

Bastús, V. J. (1855). *Diccionario histórico enciclopédico, 2*. Roca.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(8), 1798–1828.

Beyens, I., Vandenbosch, L., & Eggermont, S. (2015). Early adolescent boys' exposure to Internet pornography: Relationships to pubertal timing, sensation seeking, and academic performance. *The Journal of Early Adolescence*, *35*(8), 1045–1068.

Bin Mat, U., Buniyamin, N., Arsad, P. M., & Kassim, R. (2013). An overview of using academic analytics to predict and improve students' achievement: A proposed proactive intelligent intervention. *2013 IEEE 5th Conference on Engineering Education (ICEED)*, 126–130.

Black, S. E., Lincove, J., Cullinane, J., & Veron, R. (2015). Can you leave high school behind? *Economics of Education Review*, *46*, 52–63.

Bogarín, A., Romero, C., Cerezo, R., & Sánchez-Santillán, M. (2014). Clustering for improving educational process mining. *Proceedings of the Fourth International Conference on Learning Analytics and Knowledge*, 11–15.

Brasche, I., & Harrington, I. (2012). Promoting teacher quality and continuity: Tackling the disadvantages of remote Indigenous schools in the Northern Territory. *Australian Journal of Education*, *56*(2), 110–125.

Bravo-Agapito, J., Romero, S. J., & Pamplona, S. (2021). Early prediction of undergraduate Student's academic performance in completely online learning: A five-year study. *Computers in Human Behavior*, *115*(January 2020), 106595. https://doi.org/10.1016/j.chb.2020.106595

Brito, M., Medeiros, F., & Bezerra, E. (2019). A report-type plugin to indicate dropout risk in the virtual learning environment moodle. *2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT)*, *2161*, 127–128.

Brook, S., & Roberts, M. (2021). What are the determinants of student performance on an undergraduate accounting degree? *Journal of Further and Higher Education*. https://doi.org/10.1080/0309877X.2021.1882666

Brownlee, J. (2016). *Support Vector Machines for Machine Learning*. Machine Learning Mastery. https://machinelearningmastery.com/support-vector-machines-for-machine-learning/

Buenaño-Fernández, D., Gil, D., & Luján-Mora, S. (2019). Application of machine learning in predicting performance for computer engineering students: A case study. *Sustainability*, *11*(10), 2833.

Bunkar, K., Kumar, S., PandyaBhupendra, K., & Bunkar, R. (2012). Data mining: Prediction for performance improvement of graduate students using classification. *2012 Ninth International Conference on Wireless and Optical Communications Networks (WOCN)*, 1–5.

Bussu, A., Detotto, C., & Serra, L. (2020). Indicators to prevent university drop-out and

delayed graduation: an Italian case. *Journal of Applied Research in Higher Education*, *12*(2), 230–249. https://doi.org/10.1108/JARHE-10-2018-0201

Capterra. (2020). *Best Learning Management System Software for Windows*. https://www.capterra.com/learning-management-system-software/s/windows/

Cazarez, R. L. U., & Martin, C. L. (2018). Neural Networks for Predicting Student Performance in Online Education. *IEEE Latin America Transactions*, *16*(7), 2053–2060. https://doi.org/10.1109/TLA.2018.8447376

Chang, Y., & Brickman, P. (2018). When group work doesn't work: Insights from students. *CBE Life Sciences Education*, *17*(3). https://doi.org/10.1187/cbe.17-09-0199

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1), 1–13.

Christian, T. M., & Ayub, M. (2014). Exploration of classification using NBTree for predicting students' performance. *2014 International Conference on Data and Software Engineering (ICODSE)*, 1–6.

Correa-Burrows, P., Burrows, R., Blanco, E., Reyes, M., & Gahagan, S. (2016). *Nutritional quality of diet and academic performance in Chilean students Qualité nutritionnelle de l'alimentation et résultats scolaires des lycéens chiliens Calidad nutricional de la dieta y rendimiento académico de los estudiantes chilenos*.

Cuevas, R., Ntoumanis, N., Fernandez-Bustos, J. G., & Bartholomew, K. (2018). Does teacher evaluation based on student performance predict motivation, well-being, and ill-being? *Journal of School Psychology*, *68*, 154–162. https://doi.org/10.1016/j.jsp.2018.03.005

Daud, A., Aljohani, N. R., Abbasi, R. A., Lytras, M. D., Abbas, F., & Alowibdi, J. S. (2017). Predicting student performance using advanced learning analytics. *Proceedings of the 26th International Conference on World Wide Web Companion*, 415–421.

David, M. C., Eley, D. S., Schafer, J., & Davies, L. (2016). Risk assessment of student performance in the international foundations of medicine clinical science examination by the use of statistical modeling. *Advances in Medical Education and Practice*, *7*, 653–660. https://doi.org/10.2147/AMEP.S122841

Deo, R. C., Yaseen, Z. M., Al-Ansari, N., Nguyen-Huy, T., Langlands, T. A. M., & Galligan, L. (2020). Modern Artificial Intelligence Model Development for Undergraduate Student Performance Prediction: An Investigation on Engineering Mathematics Courses. *IEEE Access*, *8*, 136697–136724. https://doi.org/10.1109/ACCESS.2020.3010938

Döring, M. (2018). *Prediction vs Forecasting*. Data Science Blog. https://www.datascienceblog.net/post/machine-learning/forecasting_vs_prediction/

Dung, D. T. H. (2020). The advantages and disadvantages of virtual learning. *IOSR Journal of Research & Method in Education*, *10*(3), 45–48.

Dutt, A., Ismail, M. A., & Herawan, T. (2017). A systematic review on educational data mining. *Ieee Access*, *5*, 15991–16005.

Edgar, T. W., & Manz, D. O. (2017). Chapter 6 - Machine Learning. In T. W. Edgar & D. O. Manz (Eds.), *Research Methods for Cyber Security* (pp. 153–173). Syngress. https://doi.org/https://doi.org/10.1016/B978-0-12-805349-2.00006-6

Ehsanpur, S., & Razavi, M. R. (2020). A Comparative analysis of learning, retention, learning and study strategies in the traditional and M-learning systems. *European Review of Applied Psychology*, *70*(6), 100605. https://doi.org/https://doi.org/10.1016/j.erap.2020.100605

Erokhina, E. A., & Anikina, E. M. (2020). Advantages and Disadvantages of Remote Education in the Russian Federation During the Pandemic Period. *Непрерывное Профессиональное Образование: Теория и Практика*, 121–123.

Gedrimiene, E., Silvola, A., Pursiainen, J., Rusanen, J., & Muukkonen, H. (2019). Learning Analytics in Education: Literature Review and Case Examples From Vocational Education. *Scandinavian Journal of Educational Research*, *3831*. https://doi.org/10.1080/00313831.2019.1649718

Gil-Herrera, E., Tsalatsanis, A., Yalcin, A., & Kaw, A. (2011). Predicting academic performance using a Rough Set Theory-based knowledge discovery methodology. *International Journal of Engineering Education*, *27*(5), 992–1002. https://www.scopus.com/inward/record.uri?eid=2-s2.0-80052985455&partnerID=40&md5=94ee1759e02b6c5659f7637b9e67aa7a

Golden, C. E., Rothrock Jr, M. J., & Mishra, A. (2019). Comparison between random forest and gradient boosting machine methods for predicting Listeria spp. prevalence in the environment of pastured poultry farms. *Food Research International*, *122*, 47–55.

Gómez-Aguilar, D.-A., García-Peñalvo, F.-J., & Therón, R. (2014). Analítica visual en e-learning. *Profesional de La Información*, *23*(3), 236–245.

Gómez-Aguilar, D. A., Hernández-García, Á., García-Peñalvo, F. J., & Therón, R. (2015). Tap into visual analysis of customization of grouping of activities in eLearning. *Computers in Human Behavior*, *47*, 60–67.

Gonen, S., & Basaran, B. (2008). The New Method of Problem Solving in Physics Education by Using SCORM-Compliant Content Package. *Online Submission*, *9*(3), 112–120.

Google. (2021). *Colaboratory*. Preguntas Frecuentes. https://research.google.com/colaboratory/faq.html#:~:text=Colaboratory%2C or “Colab” for,learning%2C data analysis and education.

Gray, G., McGuinness, C., & Owende, P. (2014). An application of classification models to predict learner progression in tertiary education. *2014 IEEE International Advance Computing Conference (IACC)*, 549–554.

Gray, G., McGuinness, C., Owende, P., & Hofmann, M. (2016). Learning factor models of students at risk of failing in the early stage of tertiary education. *Journal of Learning Analytics*, *3*(2), 330–372.

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

Helal, S., Li, J., Liu, L., Ebrahimie, E., Dawson, S., Murray, D. J., & Long, Q. (2018). Predicting academic performance by considering student heterogeneity. *Knowledge-Based Systems*, *161*, 134–146.

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. *Annual Conference on Innovation and Technology in Computer Science Education, ITiCSE*, 175–199. https://doi.org/10.1145/3293881.3295783

Huang, C., Zhou, J., Chen, J., Yang, J., Clawson, K., & Peng, Y. (2021). A feature weighted support vector machine and artificial neural network algorithm for academic course performance prediction. *Neural Computing and Applications*. https://doi.org/10.1007/s00521-021-05962-3

Hudson, R. D., & Treagust, D. F. (2013). Which form of assessment provides the best information about student performance in chemistry examinations? *Research in Science and Technological Education*, *31*(1), 49–65. https://doi.org/10.1080/02635143.2013.764516

Hussain, M., Zhu, W., Zhang, W., & Abidi, S. M. R. (2018). Student engagement predictions in an e-learning system and their impact on student course assessment scores. *Computational Intelligence and Neuroscience*, *2018*.

Ibrahim, Z., & Rusli, D. (2007). Predicting students' academic performance: comparing artificial neural network, decision tree and linear regression. *21st Annual SAS Malaysia Forum, 5th September*.

IIOT, A. &. (2016). *A Complete Tutorial on Tree Based Modeling from Scratch (in R & Python)*. Analytics. https://analyticks.wordpress.com/2016/07/05/a-complete-tutorial-on-tree-based-modeling-from-scratch-in-r-python/

Ikonomakis, M., Kotsiantis, S., & Tampakas, V. (2005). Text classification using machine learning techniques. *WSEAS Transactions on Computers*, *4*(8), 966–974.

Imran, M., Latif, S., Mehmood, D., & Shah, M. S. (2019). Student academic performance prediction using supervised learning techniques. *International Journal of Emerging Technologies in Learning*, *14*(14), 92–104. https://doi.org/10.3991/ijet.v14i14.10310

Jayanthi, G., & Ramesh, V. (2014). Design of Academic Performance Prediction System Using Multi-Layer Perceptron. *International Journal of Computer Science and Software Engineering*, *1*(1), 9–15.

Jena, R. K. (2016). Investigating the interrelation between attitudes, learning readiness, and learning styles under virtual learning environment: a study among Indian students. *Behaviour and Information Technology*, *35*(11), 946–957. https://doi.org/10.1080/0144929X.2016.1212930

Jensen, E., Umada, T., Hunkins, N. C., Hutt, S., Huggins-Manley, A. C., & D'mello, S. K. (2021). What you do predicts how you do: Prospectively modeling student quiz performance using activity features in an online learning environment. *ACM*

International Conference Proceeding Series, 121–131. https://doi.org/10.1145/3448139.3448151

Jia, B., Niu, K., Hou, X., Li, N., Peng, X., Gu, P., & Jia, R. (2019). Prediction for Student Academic Performance Using SMNaive Bayes Model. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11888 LNAI*, 712–725. https://doi.org/10.1007/978-3-030-35231-8_52

Jishan, S. T., Rashu, R. I., Haque, N., & Rahman, R. M. (2015). Improving accuracy of students' final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique. *Decision Analytics*, *2*(1), 1–25.

Joksimović, S., Manataki, A., Gašević, D., Dawson, S., Kovanović, V., & De Kereki, I. F. (2016). Translating network position into performance: Importance of centrality in different network configurations. *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, 314–323.

Jones, T. (2018). *Supervised learning models*. IBM Developer. https://developer.ibm.com/technologies/artificial-intelligence/articles/cc-supervised-learning-models/#:~:text=Supervised learning is a method,then generalize for new examples.&text=In supervised learning%2C you create,data and a wanted output.

Khan, A., & Ghosh, S. K. (2021). Student performance analysis and prediction in classroom learning: A review of educational data mining studies. *Education and Information Technologies*, *26*(1), 205–240.

Kim, B.-H., Vizitei, E., & Ganapathi, V. (2018). GritNet: Student performance prediction with deep learning. *ArXiv Preprint ArXiv:1804.07405*.

Kim, J., Park, J. Y., & van Den Noortgate, W. (2020). On the use of Bayesian probabilistic matrix factorization for predicting student performance in online learning environments. *International Conference on Higher Education Advances*, *2020-June*, 751–759. https://doi.org/10.4995/HEAd20.2020.11137

Kitchenham, B. . (2004). *Procedures for undertaking systematic reviews*.

Kovanović, V., Gašević, D., Dawson, S., Joksimović, S., Baker, R. S., & Hatala, M. (2015). Penetrating the black box of time-on-task estimation. *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*, 184–193.

Kumar, A., Vijayalakshmi, M., & Kumar, D. A. (2012). Appraising the significance of self-regulated learning in higher education using neural networks. *International Journal of Engineering Research and Development*, *1*(1), 9–15.

Lakshmi, V., & Ramesh, D. V. (2017). Evaluating students' descriptive answers using natural language processing and artificial neural networks. *International Journal of Creative Research Thoughts (IJCRT)*, *5*(4), 3168–3173.

Lanzat, A. M. A., López, A. J. G., González, M. L. C., & Navío, E. P. (2018). Causas del fracaso escolar: Un análisis desde la perspectiva del profesorado y del alumnado. *Enseñanza & Teaching: Revista Interuniversitaria de Didáctica*, *36*(1), 129–149.

Lau, E. T., Sun, L., & Yang, Q. (2019). Modelling, prediction and classification of student academic performance using artificial neural networks. *SN Applied Sciences*, *1*(9), 982. https://doi.org/10.1007/s42452-019-0884-7

Li, K. F., Rusk, D., & Song, F. (2013). Predicting student academic performance. *2013 Seventh International Conference on Complex, Intelligent, and Software Intensive Systems*, 27–33.

Lu, O. H. T., Huang, A. Y. Q., Huang, J. C. H., Lin, A. J. Q., Ogata, H., & Yang, S. J. H. (2018). Applying learning analytics for the early prediction of Students' academic performance in blended learning. *Journal of Educational Technology & Society*, *21*(2), 220–232.

M. Barb, R. Vilanova, J. Lopez Vicario, R. (2017). Data Mining Tool for Academic Data Exploitation. In *Romania* (Issue June).

Macfadyen, L. P., & Dawson, S. (2010). Mining LMS data to develop an "early warning system" for educators: A proof of concept. *Computers and Education*, *54*(2), 588–599. https://doi.org/10.1016/j.compedu.2009.09.008

Macfadyen, L. P., Groth, D., Rehrey, G., Shepard, L., Greer, J., Ward, D., Bennett, C., Kaupp, J., Molinaro, M., & Steinwachs, M. (2017). Developing institutional learning analytics' communities of transformation'to support student success. *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, 498–499.

Mahmoud Abu Zohair, L. (2019). *Prediction of Student's performance by modelling small dataset size*. *16*, 18. https://doi.org/10.1186/s41239-019-0160-3

Malau-Aduli, B. S. (2011). Exploring the experiences and coping strategies of international medical students. *BMC Medical Education*, *11*(1). https://doi.org/10.1186/1472-6920-11-40

Maltby, A., & Mackie, S. (2009). Virtual learning environments – help or hindrance for the 'disengaged' student? *ALT-J*, *17*(1), 49–62. https://doi.org/10.1080/09687760802657577

Marbouti, F., Ulas, J., & Wang, C.-H. (2021). Academic and Demographic Cluster Analysis of Engineering Student Success. *IEEE Transactions on Education*, *64*(3), 261–266. https://doi.org/10.1109/TE.2020.3036824

MathWorks. (2021). *Análisis predictivo*. Predictive Analytics. https://es.mathworks.com/discovery/predictive-analytics.html

Mauthe, A., & Thomas, P. (2004). Professional content management systems. *John Whiley & Sons, Ltd*.

Max Kuhn. (2021). *15 Variable Importance*. https://topepo.github.io/caret/variable-importance.html

Mayilvaganan, M., & Kalpanadevi, D. (2014). Comparison of classification techniques for predicting the performance of students academic environment. *2014 International Conference on Communication and Network Technologies*, 113–118.

Meit, S. S., Borges, N. J., Cubic, B. A., & Seibel, H. R. (2004). Personality Differences in Incoming Male and Female Medical Students. *Online Submission*.

Minaei-Bidgoli, B., Kashy, D. A., Kortemeyer, G., & Punch, W. F. (2003). Predicting student performance: an application of data mining methods with an educational web-based system. *33rd Annual Frontiers in Education, 2003. FIE 2003.*, *1*, T2A-13.

Miranda, G. J., Casa Nova, S. P. C., & Cornacchione Jr., E. B. (2013). To Sir with Love: The relations between teacher qualification and student performance in accounting [Ao mestre com carinho: Relações entre as qualificações docentes e o desempenho discente em contabilidade]. *Revista Brasileira de Gestao de Negocios*, *15*(48), 462–480. https://doi.org/10.7819/rbgn.v15i48.1351

Mishra, T., Kumar, D., & Gupta, S. (2014). Mining students' data for prediction performance. *2014 Fourth International Conference on Advanced Computing & Communication Technologies*, 255–262.

Miskioglu, E. (2016). Unseen influences on student performance: Instructor assessment styles. *ASEE Annual Conference and Exposition, Conference Proceedings*, *2016-June*. https://www.scopus.com/inward/record.uri?eid=2-s2.0-84983266259&partnerID=40&md5=538ffaaf39189337ccab815bef9db963

Mitchell, T. M. (1997). *Machine learning*.

Moodle. (2020). *Course dedication*. Moodle Course Dedication. https://moodle.org/plugins/block_dedication

Morris, L. V, Finnegan, C., & Wu, S.-S. (2005). Tracking student behavior, persistence, and achievement in online courses. *The Internet and Higher Education*, *8*(3), 221–231.

Moucary, C. El, Khair, M., & Zakhem, W. (2011). Improving student's performance using data clustering and neural networks in foreign-language based higher education. *The Research Bulletin of Jordan ACM*, *2*(3), 27–34.

Moyano, M. (2018). *¿Curso a distancia sincrónico o asincrónico?* Online Educaton Center. https://onlineeducation.center/es/curso-a-distancia-sincronico-o-asincronico

Muljana, P. S., & Luo, T. (2019). Factors contributing to student retention in online learning and recommended strategies for improvement: A systematic literature review. *Journal of Information Technology Education: Research*, *18*.

Mutascu, M. (2021). Artificial intelligence and unemployment: New insights. *Economic Analysis and Policy*, *69*, 653–667. https://doi.org/https://doi.org/10.1016/j.eap.2021.01.012

Naganandhini, S., & Shanmugavadivu, P. (2019). Effective diagnosis of alzheimer's disease using modified decision tree classifier. *Procedia Computer Science*, *165*, 548–555.

Namoun, A., & Alshanqiti, A. (2021). Predicting student performance using data mining and learning analytics techniques: A systematic literature review. *Applied Sciences (Switzerland)*, *11*(1), 1–28. https://doi.org/10.3390/app11010237

Nanfor. (2021). *Telepresencial*. https://nanfor.com/pages/avanzado

Napoli, A. R., & Raymond, L. A. (2004). How reliable are our assessment data?: A comparison of the reliability of data produced in graded and un-graded conditions. *Research in Higher Education*, *45*(8), 921–929.

Naren, J. (2014). *Application of data mining in educational database for predicting behavioural patterns of the students*.

Natek, S., & Zwilling, M. (2014). Student data mining solution–knowledge management system related to higher education institutions. *Expert Systems with Applications*, *41*(14), 6400–6407.

Nguyen, Q., Huptych, M., & Rienties, B. (2018). Linking students' timing of engagement to learning design and academic performance. *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, 141–150.

Nguyen, T. L. H. (2013). Middle-level academic management: A case study on the roles of the heads of department at a Vietnamese university. *Tertiary Education and Management*, *19*(1), 1–15.

O'Shaughnessy, S., & Gray, G. (2011). Development and evaluation of a dataset generator tool for generating synthetic log files containing computer attack signatures. *International Journal of Ambient Computing and Intelligence (IJACI)*, *3*(2), 64–76.

Oladokun, V. O., Adebanjo, A. T., & Charles-Owaba, O. E. (2008). *Predicting students academic performance using artificial neural network: A case study of an engineering course*.

Orduña, P., García-Zubia, J., Rodriguez-Gil, L., Irurzun, J., López-de-Ipiña, D., & Gazzola, F. (2012). Using LabVIEW remote panel in remote laboratories: Advantages and disadvantages. *Proceedings of the 2012 IEEE Global Engineering Education Conference (EDUCON)*, 1–7.

Osmanbegović, E., AGIĆ, H., & Suljić, M. (2014). PREDICTION OF STUDENTS'SUCCESS BY APPLYING DATA MINING ALGORITHAMS. *Journal of Theoretical & Applied Information Technology*, *61*(2).

Osmanbegovic, E., & Suljic, M. (2012). Data mining approach for predicting student performance. *Economic Review: Journal of Economics and Business*, *10*(1), 3–12.

Osmanbegović, E., Suljić, M., & Agić, H. (2014). Determining dominant factor for students performance prediction by using data mining classification algorithms. *Tranzicija*, *16*(34), 147–158.

Padua Rodríguez, L. M. (2019). Factores individuales y familiares asociados al bajo rendimiento académico en estudiantes universitarios. *Revista Mexicana de Investigación Educativa*, *24*(80), 173–195.

Pakhomova, T. O., Komova, O. S., Belia, V. V, Yivzhenko, Y. V, & Demidko, E. V. (2021). Transformation of the pedagogical process in higher education during the quarantine. *Linguistics and Culture Review*, *5*(S2), 215–230.

Pal, A. K., & Pal, S. (2013). Analysis and mining of educational data for predicting the performance of students. *International Journal of Electronics Communication and*

Computer Engineering, *4*(5), 1560–1565.

Palvia, S., Aeron, P., Gupta, P., Mahapatra, D., Parida, R., Rosner, R., & Sindhi, S. (2018). *Online education: Worldwide status, challenges, trends, and implications*. Taylor & Francis.

Papageorgiou, K., & Halabi, A. K. (2014). Factors contributing toward student performance in a distance education accounting degree. *Meditari Accountancy Research*, *22*(2), 211–223. https://doi.org/10.1108/MEDAR-08-2013-0032

Parack, S., Zahid, Z., & Merchant, F. (2012). Application of data mining in educational databases for predicting academic trends and patterns. *2012 IEEE International Conference on Technology Enhanced Education (ICTEE)*, 1–4.

Patterson, R. W., & Patterson, R. M. (2017). Computers and productivity: Evidence from laptop use in the college classroom. *Economics of Education Review*, *57*, 66–79.

Perin, D. (2006). Academic progress of community college nursing aspirants: An institutional research profile. *Community College Journal of Research and Practice*, *30*(8), 657–670.

Phan, D.-V., Chan, C.-L., Pan, R.-H., Yang, N.-P., Hsu, H.-C., Ting, H.-W., & Lai, K. R. (2018). A Study of the Effects of Daily Physical Activity on Memory and Attention Capacities in College Students. *Journal of Healthcare Engineering*, *2018*, 2942930. https://doi.org/10.1155/2018/2942930

Polsani, P. R. (2003). Use and abuse of reusable learning objects. *Journal of Digital Information*, *3*(4).

Quadri, M. M. N., & Kalyankar, N. V. (2010). Drop out feature of student data for academic performance using decision tree techniques. *Global Journal of Computer Science and Technology*.

Quinn, R. J., & Gray, G. (2019). Prediction of student academic performance using Moodle data from a Further Education setting. *Irish Journal of Technology Enhanced Learning*, *5*(1). https://doi.org/10.22554/ijtel.v5i1.57

Quinn, R. J., & Gray, G. (2020). Prediction of student academic performance using Moodle data from a Further Education setting. *Irish Journal of Technology Enhanced Learning*, *5*(1).

R. (2021). *The Comprehensive R Archive Network*. https://cran.r-project.org/

Radović-Marković, M. (2010). Advantages and disadvantages of e-learning in comparison to traditional forms of learning. *Annals of the University of Petroşani, Economics*, *10*(2), 289–298.

Ramesh, V., Parkavi, P., & Ramar, K. (2013). Predicting student performance: a statistical and data mining approach. *International Journal of Computer Applications*, *63*(8).

Rana, P., Raj Gupta, L., Dubey, M. K., & Kumar, G. (2021). Review on evaluation techniques for better student learning outcomes using machine learning. *Proceedings of 2021 2nd International Conference on Intelligent Engineering and Management, ICIEM 2021*,

86–90. https://doi.org/10.1109/ICIEM51511.2021.9445294

Ribeiro, Í. J. S., Pereira, R., Freire, I. V, de Oliveira, B. G., Casotti, C. A., & Boery, E. N. (2018). Stress and Quality of Life Among University Students: A Systematic Literature Review. *Health Professions Education*, *4*(2), 70–77. https://doi.org/https://doi.org/10.1016/j.hpe.2017.03.002

Rivas, A., Gonzalez-Briones, A., Hernandez, G., Prieto, J., & Chamoso, P. (2021). Artificial neural network analysis of the academic performance of students in virtual learning environments. *Neurocomputing*, *423*, 713–720.

Rodríguez Ayán, M. N., & Ruiz Díaz, M. Á. (2011). Indicadores de rendimiento de estudiantes universitarios: calificaciones versus créditos acumulados. *Revista de Educación*.

Roll, I., Macfadyen, L. P., & Sandilands, D. (2015). Evaluating the Relationship Between Course Structure, Learner Activity, and Perceived Value of Online Courses. *Proceedings of the Second (2015) ACM Conference on Learning@ Scale*, 385–388.

Romero, C., López, M.-I., Luna, J.-M., & Ventura, S. (2013). Predicting students' final performance from participation in on-line discussion forums. *Computers & Education*, *68*, 458–472.

Sagoo, M. G., Smith, C. F., & Gosden, E. (2016). Assessment of anatomical knowledge by practical examinations: The effect of question design on student performance. *Anatomical Sciences Education*, *9*(5), 446–452. https://doi.org/10.1002/ase.1597

Salazar, M. E. L., & de León, A. L. E. (2008). *Desempeño académico de estudiantes en educación virtual. Algunos factores negativos*.

Saltos, W. R. F., & Maldonado, C. G. (2019). Predictive models for the detection of problems in autonomous learning in higher education students virtual modality. *2019 14th Iberian Conference on Information Systems and Technologies (CISTI)*, 1–6.

Sekeroglu, B., Dimililer, K., & Tuncal, K. (2019). Student Performance Prediction and Classification Using Machine Learning Algorithms. *Proceedings of the 2019 8th International Conference on Educational and Information Technology*, 7–11. https://doi.org/10.1145/3318396.3318419

Sembiring, S., Zarlis, M., Hartama, D., Ramliana, S., & Wani, E. (2011a). *Prediction of student academic performance by an application of data mining techniques*. *6*, 110–114.

Sembiring, S., Zarlis, M., Hartama, D., Ramliana, S., & Wani, E. (2011b). Prediction of student academic performance by an application of data mining techniques. *International Conference on Management and Artificial Intelligence IPEDR*, *6*(1), 110–114.

Shahiri, A. M., Husain, W., & Rashid, N. A. (2015). A Review on Predicting Student's Performance Using Data Mining Techniques. *Procedia Computer Science*, *72*, 414–422. https://doi.org/10.1016/j.procs.2015.12.157

Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory*

to algorithms. Cambridge university press.

Siemens, G. (2010). *Call for papers of the 1st international conference on learning analytics & knowledge (lak 2011)*.

Simsek, A., & Balaban, J. (2010). Learning Strategies of Successful and Unsuccessful University Students. *Online Submission*, *1*(1), 36–45.

Smith, M. W., Miller, D. W., & Seager, S. (2011). Enhancing undergraduate education in aerospace engineering and planetary sciences at MIT through the development of a CubeSat mission. *UV/Optical/IR Space Telescopes and Instruments: Innovative Technologies and Concepts V*, *8146*, 81460S.

Sorrell, S. (2007). Improving the evidence base for energy policy: the role of systematic reviews. *Energy Policy*, *35*(3), 1858–1871.

Stephens, N. M., Brannon, T. N., Markus, H. R., & Nelson, J. E. (2015). Feeling at home in college: Fortifying school- relevant selves to reduce social class disparities in higher education. *Social Issues and Policy Review*, *9*(1), 1–24.

Stephens, N. M., Hamedani, M. G., & Destin, M. (2014). Closing the social-class achievement gap: A difference-education intervention improves first-generation students' academic performance and all students' college transition. *Psychological Science*, *25*(4), 943–953.

Stephens, N. M., Townsend, S. S. M., Hamedani, M. G., Destin, M., & Manzo, V. (2015). A difference-education intervention equips first-generation college students to thrive in the face of stressful college situations. *Psychological Science*, *26*(10), 1556–1566.

Subirà, M. P.-M., & Catasús, M. G. (2011). *Aprender y enseñar en línea*. Universitat Oberta de Catalunya.

Surenthiran, S., Rajalakshmi, R., & Sujatha, S. S. (2021). Student Performance Prediction Using Atom Search Optimization Based Deep Belief Neural Network. *Optical Memory and Neural Networks (Information Optics)*, *30*(2), 157–171. https://doi.org/10.3103/S1060992X21020119

Suresh, K., Meghana, J., & Pooja, M. E. (2021). Predicting the E-Learners Learning Style by using Support Vector Regression Technique. *Proceedings - International Conference on Artificial Intelligence and Smart Systems, ICAIS 2021*, 350–355. https://doi.org/10.1109/ICAIS50930.2021.9396018

Talebian, S., Mohammadi, H. M., & Rezvanfar, A. (2014). Information and communication technology (ICT) in higher education: advantages, disadvantages, conveniences and limitations of applying e-learning to agricultural students in Iran. *Procedia-Social and Behavioral Sciences*, *152*, 300–305.

Tempelaar, D., Rienties, B., Mittelmeier, J., & Nguyen, Q. (2018). Student profiling in a dispositional learning analytics application using formative assessment. *Computers in Human Behavior*, *78*, 408–420. https://doi.org/https://doi.org/10.1016/j.chb.2017.08.010

Tempelaar, D. T., Rienties, B., & Nguyen, Q. (2017). Towards actionable learning analytics

using dispositions. *IEEE Transactions on Learning Technologies*, *10*(1), 6–16.

Ten Ham-Baloyi, W., & Jordan, P. (2016). Systematic review as a research method in post-graduate nursing education. *Health {SA} Gesondheid*, *21*, 120–128. https://doi.org/http://dx.doi.org/10.1016/j.hsag.2015.08.002

Toskova, A., Toskov, B., Doukovska, L., Daskalov, B., & Radeva, I. (2018). Neural Networks in the Intelligent Educational Space. *ANNA '18; Advances in Neural Networks and Applications 2018*, 1–6.

Tranfield, D., Denyer, D., & Smart, P. (2003). Towards a methodology for developing evidence informed management knowledge by means of systematic review. *British Journal of Management*, *14*(3), 207–222.

Tucker, B. (2012). The flipped classroom. *Education Next*, *12*(1), 82–83.

Umer, R., Susnjak, T., Mathrani, A., & Suriadi, L. (2021). Current stance on predictive analytics in higher education: opportunities, challenges and future directions. *Interactive Learning Environments*. https://doi.org/10.1080/10494820.2021.1933542

Vamanan, R., Parkavi, P., & Ramar, K. (2013). Predicting Student Performance: A Statistical and Data Mining Approach. *INTERNATIONAL JOURNAL OF COMPUTER APPLICATIONS*, *63*, 975–8887. https://doi.org/10.5120/10489-5242

Vasudevan, V., Almozini, B., Almuhanna, S., & Aljubair, A. (n.d.). *University Student's Learning Pattern Analysis and Prediction in LMS Using Data Mining Techniques*.

Vasudevan, V., & Almuhanna, S. (2018). *University Student's Learning Pattern Analysis and Prediction in LMS Using Data Mining Techniques*. 7.

Velásquez, J. . (2014). Una guía corta para escribir Revisiones Sistemáticas de Literatura. *DYNA*, *81*(187), 9–10. http://www.bdigital.unal.edu.co/44811/1/46758-226992-1-PB.pdf

Virvou, M., Alepis, E., Tsihrintzis, G. A., & Jain, L. C. (2020). Machine learning paradigms: Advances in learning analytics. In *Intelligent Systems Reference Library* (Vol. 158, pp. 1–5). https://doi.org/10.1007/978-3-030-13743-4_1

Waheed, H., Anas, M., Hassan, S.-U., Aljohani, N. R., Alelyani, S., Edifor, E. E., & Nawaz, R. (2021). Balancing sequential data to predict students at-risk using adversarial networks. *Computers & Electrical Engineering*, *93*, 107274.

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*, 106189.

Wasif, M., Waheed, H., Aljohani, N. R., & Hassan, S.-U. (2019). Understanding student learning behavior and predicting their performance. In *Cognitive Computing in Technology-Enhanced Learning* (pp. 1–28). IGI Global.

Webster, M. (2006). Merriam-Webster online dictionary. *Retrieved June*, *20*, 2013.

Woolf, B. P. (2009). Chapter 7 - Machine Learning. In B. P. Woolf (Ed.), *Building Intelligent*

Interactive Tutors (pp. 221-297). Morgan Kaufmann. https://doi.org/https://doi.org/10.1016/B978-0-12-373594-2.00007-1

Wu, J. (2020). Machine Learning in Education. *Proceedings - 2020 International Conference on Modern Education and Information Management, ICMEIM 2020*, 56–63. https://doi.org/10.1109/ICMEIM51375.2020.00020

Xu, J., Moon, K. H., & Van Der Schaar, M. (2017). A Machine Learning Approach for Tracking and Predicting Student Performance in Degree Programs. *IEEE Journal on Selected Topics in Signal Processing*, *11*(5), 742–753. https://doi.org/10.1109/JSTSP.2017.2692560

Yadav, N., & Srivastava, K. (2020). Student Performance Prediction from E-mail Assessments Using Tiny Neural Networks. *2020 9th IEEE Integrated STEM Education Conference, ISEC 2020*, *2020-Janua*. https://doi.org/10.1109/ISEC49744.2020.9397817

Yadav, S. K., & Pal, S. (2012). *Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification*. *2*(2), 51–56. http://arxiv.org/abs/1203.3832

Yang, F., & Li, F. W. B. (2018). Study on student performance estimation, student progress analysis, and student potential prediction based on data mining. *Computers & Education*, *123*, 97–108. https://doi.org/https://doi.org/10.1016/j.compedu.2018.04.006

Yang, S. J. H., Lu, O. H. T., Huang, A. Y. Q., Huang, J. C. H., Ogata, H., & Lin, A. J. Q. (2018). Predicting students' academic performance using multiple linear regression and principal component analysis. *Journal of Information Processing*, *26*, 170–176.

Yegnanarayana, B. (2009). *Artificial neural networks*. PHI Learning Pvt. Ltd.

Yu, T., & Jo, I.-H. (2014). Educational technology approach toward learning analytics: Relationship between student online behavior and learning performance in higher education. *Proceedings of the Fourth International Conference on Learning Analytics and Knowledge*, 269–270.

Zacharis, N. Z. (2010). The impact of learning styles on student achievement in a web-based versus an equivalent face-to-face course. *College Student Journal*, *44*(3), 591–598.

Zacharis, N. Z. (2015). A multivariate approach to predicting student outcomes in web-enabled blended learning courses. *The Internet and Higher Education*, *27*, 44–53.

Zacharis, N. Z. (2016). Predicting student academic performance in blended learning using artificial neural networks. *International Journal of Artificial Intelligence and Applications*, *7*(5), 17–29.

Zaporozhko, V. V, Parfenov, D. I., & Shardakov, V. M. (2019). Development Approach of Formation of Individual Educational Trajectories Based on Neural Network Prediction of Student Learning Outcomes. *International Conference of Artificial Intelligence, Medical Engineering, Education*, 305–314.

Zhang, Q., Wang, K., & Zhou, S. (2020). Application and Practice of VR Virtual Education Platform in Improving the Quality and Ability of College Students. *IEEE Access*, *8*, 162830–162837.

Zomaya, A. Y. (2017). Foreword. In J. Miguel, S. Caballé, & F. Xhafa (Eds.), *Intelligent Data Analysis for e-Learning* (pp. xvii-xviii). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-804535-0.09995-0