DESDE 2013
https://repository.uaeh.edu.mx/revistas/index.php/icbi/issue/archive
Pädi Boletín Científico de Ciencias Básicas e Ingenierías del ICBI

ISSN: 2007-6363

Publicación Semestral Pädi Vol. 10 No. Especial 2 (2022) 75-81

Aplicación de técnicas de minería de datos para la caracterización de estudiantes bajo el efecto de la COVID-19

Application of data mining techniques to characterize students under the effect of COVID-19

L. Ramírez-Melo [a], E. R. Delgado-Ávila [a], M. A. Montúfar-Benítez [a,*]

[a] *Área Académica de Ingeniería y Arquitectura, Universidad Autónoma del Estado de Hidalgo, 42184, Pachuca, Hidalgo, México.*

Resumen

En esta investigación se integraron técnicas de minería de datos (MD), tales como el análisis de clúster jerárquico y la regresión logística, para caracterizar alumnos de un programa educativo de la Universidad Autónoma del Estado de Hidalgo (UAEH) que se pueden ver vulnerables a causa de la rápida migración a la educación en línea y otras repercusiones de las medidas adoptadas para evitar la propagación de la COVID-19. Para la aplicación de las técnicas de MD se realizó un sondeo a los alumnos para obtener datos acerca de las condiciones sociodemográficas, económicas, técnicas, de salud mental y académicas que permitan encontrar patrones que inciden en el desempeño académico del alumnado. Con los resultados obtenidos se puede observar la existencia de 2 grupos (clústeres), uno de ellos con mejores condiciones que otro y también un listado de ítems que influyen tanto negativamente como positivamente a que un alumno vea afectado su rendimiento escolar.

Palabras Clave: Covid-19, Minería de datos, Educación en línea, Educación superior.

Abstract

In this research, data mining (DM) techniques were integrated, such as hierarchical cluster analysis and logistic regression, to characterize students of an educational program of the Autonomous University of the State of Hidalgo (UAEH) who may be seen as vulnerable due to the fast migration to online education and other repercussions of the actions taken to prevent the spread of COVID-19. For the application of the DM techniques, a survey was carried out on the students to obtain data about the sociodemographic, economic, technical, mental health and academic conditions that allow finding patterns that affect the academic performance of the students. With the results obtained, the existence of 2 groups (clusters) can be observed, one of them with better conditions than the other and also a list of items that carry both weight negatively and positively a student's academic performance being affected.

Keywords: Covid-19, Data Mining, Online Education, Higher Education.

1. Introducción

El 11 de marzo de 2020 la OMS declaró oficialmente al COVID-19 como una pandemia. Las afectaciones en temas de salud pública, donde se contabilizan más de 190 millones de casos de contagios confirmados y 4 millones de muertes acumuladas en todo el mundo (OMS, 2021), se suman a las de otros ámbitos como el económico y social, esto como efecto colateral de las medidas que cada gobierno ha adoptado para mitigar y controlar la enfermedad, que en su mayoría han sido el confinamiento, el distanciamiento social/físico, la suspensión de actividades escolares presenciales, el cierre parcial del comercio y la industria, así como la restricción a la inmigración (González et al., 2020).

Se ha reportado que 220 millones de estudiantes de educación terciaria en el mundo se vieron obligados a continuar sus estudios en modalidad online debido al COVID-19 (UNESCO, 2021a), en México fueron más de 4 millones de alumnos en ese nivel educativo los afectados por la misma situación (UNESCO, 2021b), lo cual mostró que los sistemas educativos no se encontraban preparados en su totalidad para sobrellevar aspectos técnicos, de ayuda psicológica y/o económica hacia los estudiantes, provocando que muchos de ellos enfrentarán el riesgo de desertar de sus estudios o ser vulnerables a otro tipo de afectaciones causadas por la

*Autor para la correspondencia: montufar@uaeh.edu.mx
Correo electrónico: ra382386@uaeh.edu.mx (Lucero Ramírez-Melo), de298779@uaeh.edu.mx (Eder Renato Delgado-Ávila), montufar@uaeh.edu.mx (Marco Antonio Montúfar-Benítez)

Fecha de recepción: 07/02/2021 Fecha de aceptación: 26/04/2022 Fecha de publicación: 24/06/2022
https://doi.org/10.29057/icbi.v10iEspecial2.8669

pandemia. Para hacer frente a dicho problema es necesario primeramente identificar a los estudiantes vulnerables y posteriormente a través de la implementación de políticas de ayuda poder mitigar las consecuencias negativas de la pandemia (Murphy, 2020). Una de las herramientas para identificar, clasificar y analizar relaciones de manera no trivial, es el proceso de descubrimiento de conocimiento en bases de datos (KDD por sus siglas en inglés), en donde uno de los pasos más importantes para obtener este conocimiento es la minería de datos de donde se pueden obtener modelos predictivos y descriptivos dependiendo el tipo de aprendizaje, por lo tanto el uso de esta metodología puede ser de gran ayuda para el problema de caracterizar estudiantes para posteriormente mitigar su vulnerabilidad.

Se realizaron dos modelos, uno de aprendizaje no supervisado donde se implementó el análisis de clúster jerárquico para agrupar alumnos de acuerdo a la similitud entre sus respuestas y otro de aprendizaje supervisado que fue la regresión logística, con el fin de clasificar a los alumnos de acuerdo a sus características y finalmente medir la funcionalidad del modelo supervisado como predictor de desempeño académico. Estos dos modelos se utilizaron como complementarios ya que cada modelo permite responder distintas preguntas, en el caso del modelo de aprendizaje no supervisado permite responder si existen grupos en los cuáles los individuos compartan características similares y cuáles son esas características, mientras que el modelo de aprendizaje supervisado permite responder la pregunta "¿cuáles son los factores que influyen tanto positivamente como negativamente a clasificar un alumno como reprobado?".

El modelo de análisis de clúster jerárquico, sugirió que, entre los dos clústeres formados, el clúster dos se caracteriza porque existen altos porcentajes en la reprobación de los alumnos, se tiene un ingreso familiar mensual bajo, los alumnos dedicaron menos tiempo al estudio y realizaron actividades extra al estudio (trabajo) durante la pandemia, en comparación con los alumnos del clúster uno. Mientras que, por parte del modelo de regresión logística se identificaron variables que influyen en el rendimiento académico del alumno como percibir sus calificaciones durante la pandemia como más bajas, conectarse a clases por medio de un dispositivo de renta y/o considerar la baja temporal o definitiva durante el periodo de clases en línea.

Con el uso de los resultados generados se prevé que las técnicas de minería de datos podrán clasificar de manera eficaz a los alumnos de tal forma que los tomadores de decisiones podrán tener mayor certeza de que sus acciones están soportadas en información obtenida de manera metódica y científica.

2. Materiales y Métodos

2.1. Estado del arte

Son muchas las variables que existen para que un alumno repruebe alguna materia en condiciones normales y puede que con la actual emergencia sanitaria estos factores aumenten o se potencien debido a la rápida migración de alumnos a la educación en línea (Wang et al., 2020).

Alrededor del globo diversos autores han aplicado distintas técnicas cuantitativas y cualitativas para estudiar los fenómenos provocados durante la pandemia en el ámbito escolar, son de distinta índole y objetivo. Abidah et al. (2020) comenta que los cierres de las instituciones educativas por el brote de COVID-19 provocaron un impacto sin precedentes en la educación y han dejado al descubierto muchas ineficiencias y debilidades del sistema educativo de Indonesia. En su estudio Kapasia et al. (2020) encontró algunas causas de los desafíos de la educación en pregrado y posgrado durante la pandemia como sensación de estrés, depresión y ansiedad en los estudiantes, falta de dispositivos electrónicos para estudiar en línea, entre otras. Dong (2020) encontró que las principales razones de los problemas durante las clases online entre alumnos de Bangladesh con profesores chinos fueron que los alumnos no contaban con la infraestructura tecnológica necesaria y la diferencia de horarios entre ambos países hacía difícil la comunicación.

En otro estudio efectuado por Odriozola-González et al. (2020) utilizando una escala de Depresión Ansiedad Stress (DASS-21) determinan los porcentajes de la población encuestada de alumnos y trabajadores de una institución educativa en España, que se encuentra en las categorías de moderado a severo extremo en la respectiva medición.

En la búsqueda bibliográfica realizada, no se encontraron estudios donde se hayan aplicado métodos semejantes como los aquí desarrollados para caracterizar y clasificar a los estudiantes afectados y vulnerables durante la pandemia. Sin embargo, se han realizado estudios utilizando técnicas de minería de datos para la predicción del desempeño académico en estudios como: Rico et al. (2014), Martínez et al. (2020) y en Heredia et al. (2014), donde utilizan la regresión logística como herramienta de aprendizaje supervisado. También Cirami et al. (2018) aplica la técnica de análisis de clúster jerárquico como método para evaluar alumnos universitarios frente a situaciones de estrés.

2.2. Diseño y aplicación del instrumento de medición

Se aplicó un muestreo por conveniencia donde la población de estudio establecida fueron los estudiantes de segundo a noveno semestre de la Licenciatura en Ingeniería Industrial de la UAEH. La muestra generó la participación de 396 alumnos.

El instrumento de medición utilizado fue una encuesta elaborada con la herramienta Google Forms, integrada por 30 preguntas para recabar información de variables de índole sociodemográfica, económica, condiciones técnicas referentes a la educación a distancia, salud mental, así como académica. El enlace del cuestionario se envió a los estudiantes a través del correo electrónico institucional. El formulario está disponible en https://forms.gle/47QAv8CALVL5vySp8.

La selección de la población de estudio fue definida debido a que 1) se busca un grupo de alumnos lo suficientemente grande para aplicar las técnicas estadísticas para la caracterización de alumnos, 2) los alumnos deben de pertenecer al mismo programa educativo para no provocar un aumento de la dimensionalidad y que por tanto el método de aprendizaje supervisado no converja y 3) los alumnos de primer semestre fueron excluidos del estudio debido a que al ser de nuevo ingreso no han cursado ningún semestre en línea en el programa de estudios seleccionado.

2.3. Modelo de aprendizaje no supervisado

De acuerdo a Winston y Albright (2019) los métodos de aprendizaje no supervisado, también llamado descriptivo, no dirigido o de agrupación, buscan patrones y estructuras complejas en todas las variables sin ninguna guía específica del analista. Algunas aplicaciones comunes del aprendizaje no supervisado son la reducción de dimensiones, el reconocimiento de patrones y extracción de la mayor cantidad posible de información (Jaggia et al., 2021; Bramer, 2013).

Las técnicas de aprendizaje no supervisado más comunes son análisis de clúster y análisis de reglas de asociación. Tras un análisis previo de estas técnicas se determinó aplicar análisis de clúster.

El objetivo del análisis de clúster es encontrar similitudes entre los datos estudiados y agruparlos en conglomerados de observaciones que mantienen características similares (Jaggia et al., 2021). Se busca que los elementos que se encuentran en cada conglomerado tengan características lo más similares posibles, mientras que entre conglomerados sean lo más diferentes posibles. Dentro del análisis de clúster existen dos técnicas de agrupación muy comunes que son el clúster jerárquico y agrupación k-means.

La agrupación en clústeres jerárquica utiliza un proceso iterativo para agrupar datos en una jerarquía de clústeres. Para lograrlo existen dos estrategias que son el agrupamiento aglomerativo (AGNES) y el agrupamiento divisivo (DIANA). El algoritmo implementado fue AGNES, el cual considera de manera inicial que cada observación es un grupo, el algoritmo fusiona de manera iterativa grupos que son similares entre sí a medida que uno asciende en la jerarquía (Jaggia et al., 2021).

2.3.1. Pre-procesamiento de datos

Como primer paso se realizó el pre-procesamiento de datos, la base de datos (instancia) tipo ".csv" se obtuvo directamente de las respuestas de la encuesta aplicada.

Fue necesario realizar una limpieza de datos, así como una eliminación y extracción de características (se formaron algunos subconjuntos de variables para resumir la información). De esta manera se pasó de una instancia con alta dimensionalidad (66 variables) a una con menor dimensionalidad (30 variables). Este proceso denominado reducción de dimensionalidad permite que el algoritmo utilizado tenga mayor exactitud, que se disminuya la carga computacional y se mejora la interpretación de los clasificadores (Braga-Neto, 2020).

2.3.2. Implementación del modelo

Una vez que el conjunto de datos estaba preparado se procedió a realizar el agrupamiento de clúster jerárquico, para ello se utilizó el software estadístico libre R, versión 4.1.0 y el entorno de desarrollo integrado libre RStudio, versión 1.4.1717.

En el repositorio de Github https://github.com/lu-ramirez/DM_Techniques.git se pueden encontrar los scripts desarrollados tanto para el modelo de aprendizaje no supervisado como el de aprendizaje supervisado, las bases de datos necesarias para ambos modelos se pueden consultar en 10.21227/3jc2-ja07.

Se instalaron y cargaron las librerías de *cluster*, para poder utilizar los algoritmos de agrupación; *ggplot2* y *factoextra* para editar los gráficos resultantes; y *openxlsx* para exportar un data frame con resultados a un archivo ".xlsx".

La manera de determinar si una observación es similar o diferente a otra es a través de las medidas de similitud. Estas se basan en la distancia entre observaciones por pares (registros) de las variables. Cuanto menor es la distancia entre observaciones mayor es su similitud y viceversa. Las medidas de similitud se tipificaron de acuerdo a la clase de datos con los que se va a trabajar: numéricos, categóricos o mixtos. Debido a la naturaleza de los datos (mixtos) se utilizó el coeficiente de Gower, el cual calcula la distancia para cada variable, al convertirla en una escala [0,1] y calcula un promedio ponderado de las distancias escaladas (Jaggia et al., 2021).

El coeficiente de similaridad entre dos elementos i y j, con base a la variable h se define como una función S_{ijh} no negativa y simétrica que satisface la ecuación 1:

$$S_{jii} = 1, 0 \leq S_{ijh} \leq 1, \quad S_{ijh} = S_{jih} \tag{1}$$

Una vez obtenida la similaridad entre elementos se transformaron los coeficientes en distancias. Al definirse $d_{ij} = 1 - S_{ij}$ en algunos casos puede no satisfacer la desigualdad triangular, por lo tanto, se define la matriz de similaridades como positiva y la distancia se define de acuerdo a la ecuación 2:

$$d_{ij} = \sqrt{a(1 - S_{ij})} \tag{2}$$

Esta satisface la desigualdad triangular y por lo tanto satisface todas las características de una medida de distancias. El coeficiente de similitud de Gower propuesto en 1971 está definido como se muestra en la ecuación 3.

$$d_{ij}{}^2 = 1 - S_{ij} \tag{3}$$

$$S_{ij} = \frac{\sum_{h-1}^{p1}\left(1-\frac{|x_{ih}-x_{jh}|}{Gh}\right)+a+\alpha}{p_1+p_2-d+p_3} \tag{4}$$

La ecuación 4 muestra el coeficiente de similaridad, donde p_1 es el número de variables cuantitativas, a y d son el número de coincidencias en 1 (presencia de la característica) y el número de coincidencias en 0 (ausencia de la característica) respectivamente para las p_2 variables binarias, α es el número de coincidencias para las p_3 variables cualitativas y Gh es el rango de la h-ésima variable cuantitativa.

El algoritmo AGNES utiliza un método de vinculación que le permite evaluar la similitud entre grupos, esto sucede durante el proceso iterativo. Algunos métodos de vinculación son el método de enlace único, el método de enlace completo, el método del centroide, el método de vinculación promedio y el método de Ward. Este último es el que se empleó en el modelo, ya que utiliza la suma de cuadrados del error (ESS por sus siglas en inglés) para medir la pérdida de información que ocurre cuando las observaciones se agrupan (Jaggia et al., 2021).

Sean A y B dos clases no vacías y disjuntas y sean P_A, P_B y g_A, g_B sus pesos y centros de gravedad, respectivamente. La distancia de Ward entre los dos grupos, en función de la distancia euclidiana canónica d, viene dada por la ecuación 5.

$$W(A,B) = \frac{P_A P_B}{P_A+P_B} d^2(g_A - g_B) \tag{5}$$

De acuerdo a los detalles anteriores, la descripción general del algoritmo de análisis de clúster jerárquico se muestra en la Tabla 1.

Tabla 1: Descripción completa del algoritmo de análisis de clúster jerárquico

Entrada: conjunto de datos D, el número total de objetos de datos n.
Salida: Los objetos de datos se dividen en diferentes grupos.
Paso 1: preprocesamiento de los atributos de D.
Paso 2: calcular la similaridad mediante (4) a los atributos.
Paso 3: emplear valores del paso 1 para obtener las distancias con (3) entre los atributos.
Paso 4: realización de AGNES sobre la base del paso 2 haciendo uso de (5).
Paso 5: Tratar los resultados del agrupamiento de AGNES como un nuevo conjunto de datos.
Paso 6: Obtener el resumen de los clústeres.
Paso 7: Identificar a qué clúster pertenece cada individuo.

Una manera de medir la fuerza de la estructura de agrupación natural en los datos es con el coeficiente de aglomeración, en este caso el coeficiente de aglomeración obtenido fue de 0.94, cuyo valor se considera muy bueno, ya que, desde 0.75 se puede decir que las agrupaciones tienen una estructura buena y natural.

Cuando se completó el proceso de agrupamiento realizado por el algoritmo los datos se representaron en una estructura de dendograma. Es muy útil debido a que permite inspeccionar de manera visual el resultado de la agrupación y determinar el número de clústeres a formar, aunque esta es una tarea muy subjetiva (Jaggia et al., 2021). En la figura 1 se puede observar el dendograma resultante de correr el algoritmo AGNES.

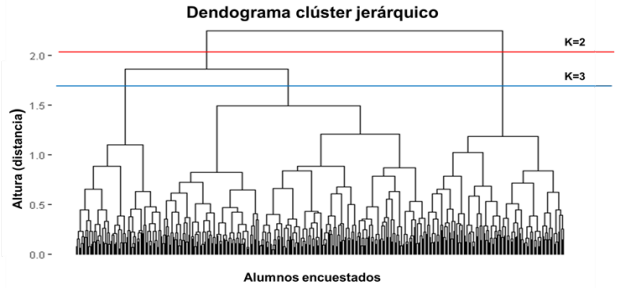

Figura 1: Dendograma de clúster jerárquico obtenido con AGNES.

Como se puede apreciar en la figura 1, las líneas que cortan de manera horizontal las "ramas" del gráfico indican la cantidad de clústeres (k) que se pueden formar en esos puntos (2 y 3 respectivamente). Si se continúa cortando las ramas más abajo, la cantidad de clústeres aumenta. Cabe mencionar que en un dendograma entre mayor sea la altura (distancia) de las ramas más distintivo es un grupo del otro. Para poder determinar el valor de k, se realizó un análisis de los promedios de las variables en cada clúster formado al utilizar distintos valores para k. Estos valores fueron 2 y 3, los cuales fueron suficientes ya que se observó que, al aumentar el valor de k, los grupos que se formaban no mostraban información de interés de acuerdo al propósito del estudio.

Tabla 2: Promedios de algunas variables de los clústeres con k=2

Clúster	*N**	*Calidad de la señal*	*Nivel de ansiedad*	*Reprobó materias*
1	284	3	3	0.29
2	112	2	4	0.76

*N es el número de alumnos que pertenece a cada clúster.

Tabla 3: Promedios de algunas variables de los clústeres con k=3

Clúster	*N**	*Calidad de la señal*	*Nivel de ansiedad*	*Reprobó materias*
1	70	3	3	0.37
2	214	3	3	0.26
3	112	2	4	0.76

*N es el número de alumnos que pertenece a cada clúster.

En las tablas 2 y 3 se puede observar que al utilizar k=2 y k=3 el último clúster que se forma es el mismo (agrupa a los mismos elementos, en este caso son alumnos). En él se incluyen aquellos alumnos que al parecer tienen mayor tendencia a reprobar materias, factor que se puede considerar indicador de vulnerabilidad. Bajo este análisis se decidió conservar como k=2.

Se obtuvo un resumen de las variables para cada clúster formado, así como la identificación del clúster correspondiente de cada individuo, una parte de dichos resultados se pueden observar en la tabla 4.

Tabla 4: Relación de clúster al que pertenece cada encuestado

Encuestado	*Clúster al que pertenece*
1	1
2	1
3	1
4	1
5	1
6	1
7	2
8	1
9	2
⋮	⋮
396	1

2.4. Modelo de aprendizaje supervisado

El aprendizaje supervisado consiste en recibir datos para los que se sabe la solución al problema que se está planteando y el modelo se va entrenando con estos datos que son pares, es decir, que el modelo se va entrenando con datos de entrada para los que ya se conoce la salida o resultado (Mathivet, 2018).

Dentro de los modelos y algoritmos de aprendizaje supervisado se encuentran la regresión logística, el clasificador Naïve Bayes, redes neuronales, entre otros (Winston y Albrigth, 2019).

El método que se decidió usar en el presente es la regresión logística (Cox & Snell, 1989) que es un método estadístico que permite clasificar y modelar algunos fenómenos cuya variable dependiente Y_i es binaria, estima la probabilidad P_i de que un individuo se encuentre en una categoría. Utiliza una función no lineal (sigmoidal) de las variables explicativas X_i que se muestra en la ecuación 6 para clasificar a un individuo con base en la probabilidad obtenida (Winston y Albrigth, 2019).

$$P_i = \frac{1}{1+e^{-\beta X_i}} \quad (6)$$

Para poder encontrar los valores del vector β de pesos o coeficientes es necesario maximizar la función de máxima verosimilitud, que es lo equivalente a minimizar la desviación total (Robles et al., 2020). Como se muestra en la ecuación (7), siendo i el número de observación y N el tamaño de la muestra.

$$l = \prod_{i=1}^{N} P_i^{Y_i}(1-P_i)^{1-Y_i} \quad (7)$$

Del modelo de regresión logística se obtiene la salida binaria Y_i en donde se clasifica como grupo 0 a las personas que no reprueban y grupo 1 a las que sí lo harían, con base en la probabilidad obtenida de la ecuación 6 para cada observación, de la siguiente manera.

$$Y_i \begin{cases} 0 \; si \; P_i < umbral \\ 1 \; si \; P_i \geq \; umbral \end{cases} \quad (8)$$

El umbral que generalmente se usa en la ecuación 8 es de 0.5, pero se puede ajustar a las exigencias del problema, en este caso el valor considerado fue de 0.25.

2.4.1. Pre-procesamiento de datos

Una vez recolectados los datos se realizó un proceso mediante el que se convirtieron en binarios los valores de la columna "Reprobadas" que posteriormente fue utilizada como variable dependiente Y_i, ya que en la base de datos la columna elegida tenía valores numéricos entre cero y diez. En este caso la columna seleccionada pasó de representar el número de materias reprobadas por un alumno en la modalidad a distancia a representar si un alumno había reprobado (representado con un 1) o no (representado por un 0) en dicha modalidad. Este proceso se realizó utilizando las librerías *numpy* y *pandas* de Python.

Posteriormente, se convirtieron todas las variables categóricas a variables ficticias de tal manera que se obtuvieron $c-1$ variables ficticias de cada variable categórica tomando en cuenta que c es el número de categorías que poseía cada variable categórica. Después se eliminaron las variables categóricas originales, es decir, se eliminaron las variables a partir de las cuales se crearon las nuevas variables ficticias, inmediatamente se definió la columna "Reprobadas" como variable dependiente Y_i y se eliminó una de las variables ficticias creadas para cada columna para reducir la dimensionalidad del dataset. Se determinó de manera preliminar el conjunto de variables candidatas a formar parte del modelo final, que al momento eran todas las variables restantes con excepción de la variable dependiente.

2.4.2. Selección de rasgos para el modelo

Después de la creación de variables ficticias, se obtuvo un conjunto de datos de 396 filas y 35 columnas. Para seleccionar las variables que formarán parte del modelo se ocupó el método "Clasificación de características con eliminación de características recursivas" que viene en la librería *sklearn* de Python para hacer una clasificación de las variables independientes X_i y posteriormente seleccionar las de mayor relevancia de acuerdo a la clasificación. Para realizar esto se le dio al método los datos de entrada de cuantas variables forman parte del modelo, y qué tipo de modelo es.

Una vez obtenido el Ranking de las variables que tienen mayor relevancia dentro del modelo se procedió a guardar las variables para posteriormente implementar el modelo.

2.4.3. Implementación del modelo

Para implementar el modelo se ocupó la librería *statsmodels.api* de Python a la cual se introdujeron las variables independientes y la variable dependiente que ya se habían definido anteriormente X_i e Y_i correspondientemente. Posteriormente se resuelve el método de máxima verosimilitud para poder obtener los coeficientes y otros estadísticos que se muestran más adelante en la tabla 8 de la sección 3.2.

2.4.4. Validación del modelo

Para poder validar el modelo, del conjunto final de los datos se separó en un conjunto de entrenamiento tomando el 80% de los datos de manera aleatoria y el otro 20% fue utilizado para probar el modelo, y se utilizaron las siguientes métricas.

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (9)$$

$$precision = \frac{TP}{TP+FP} \quad (10)$$

$$recall = \frac{TP}{TP+FN} \quad (11)$$

$$specificity = \frac{TN}{TP+FN} \quad (12)$$

$$f1 = 2 * \frac{precision*recall}{precision+recall} \quad (13)$$

Las métricas anteriores se obtuvieron de la matriz de confusión que se muestra en la tabla 5, donde los TN son verdaderos negativos, TP son verdaderos positivos, mientras los FP son falsos positivos y por último los FN son falsos negativos.

Tabla 5: Matriz de confusión

	Predicción	
Realidad	0	1
0	TN	FP
1	FN	TP

Para el caso de estudio realizado es más grave pronosticar a un alumno como no reprobado y que en realidad lo resulte siendo, es por esto que la métrica más importante será la sensibilidad (recall), ya que, esta tiene un valor más bajo cuando se tienen muchos falsos negativos, en otras palabras, la métrica recall castiga falsos negativos.

Para realizar la validación del modelo se seleccionó un umbral de 0.25 para sustituir en la ecuación 8 para después realizar la clasificación conforme a la misma ecuación 8 y así obtener la matriz de confusión que se muestra en la tabla 9. Para finalizar la validación del modelo se obtienen las métricas de las ecuaciones (9 - 13). A continuación, en la tabla 6 se presenta la descripción general del algoritmo utilizado para realizar la regresión logística y su validación.

Tabla 6: Descripción completa del algoritmo de regresión logística

Entrada: conjunto de datos de D.
Salida: Coeficientes de la ecuación de clasificación y métricas de desempeño de la misma.
Paso 1: Preprocesamiento de los atributos de D.
Paso 2: Selección de la variable dependiente.
Paso 3: Aplicación del método RFE para ranquear variables independientes de mayor relevancia.
Paso 4: Seleccionar las n primeras variables independientes.
Paso 5: Partir D en dos subconjuntos aleatoriamente, uno de entrenamiento y otro de prueba.
Paso 6: Aplicar el método de máxima verosimilitud con las variables independientes y la variable dependiente del conjunto de entrenamiento.
Paso 7: Repetir el paso 4-5 con n-1 variables independientes hasta que el método de máxima verosimilitud converja.
Paso 8: Aplicar la ecuación de clasificación sobre el conjunto restante y clasificar aciertos y errores en la matriz de confusión.
Paso 9: Calcular métricas de desempeño.

3. Resultados

3.1. Modelo de aprendizaje no supervisado

Al realizar la agrupación de clúster jerárquico al conjunto de datos se obtuvo información relevante de los dos clústeres formados. Con los datos del resumen proporcionado por R se realizó una caracterización de cada clúster, los resultados para los atributos más representativos se pueden observar en la tabla 7.

Algunas variables no representaron valores tan diferentes entre clústeres por lo que se puede considerar que no son tan críticas, algunas de estas fueron: el sexo, el estado civil, el nivel de educación de sus padres, si algún familiar ha sido positivo a covid-19 y/o si tuvo un deceso por la misma causa.

3.2. Modelo de aprendizaje supervisado

Una vez realizado el modelo de regresión logística se obtuvo el resumen que se muestra en la tabla 8, cabe destacar que el ordenamiento de las variables es de mayor a menor dependiendo del valor de la columna llamada exp [Coef], esto indica el orden en el que inciden de manera positiva a pertenecer a la clase "1" (alumnos vulnerables), por lo que si el alumno percibe que tiene calificaciones más bajas es porque está en riesgo de reprobar, lo mismo sucede si el alumno accede a las clases mediante un dispositivo de acceso público o de renta. En cambio, que se conecte por medio de una laptop o computadora de escritorio disminuye sus probabilidades de ser vulnerable. Al validar el modelo se obtuvo la tabla 9 donde se muestra la matriz de confusión en donde se puede observar que el error que menos se comete es el tipo II (falsos negativos).

Tabla 9: Matriz de confusión

	Predicción	
Realidad	No reprobado	Reprobado
No reprobado	21	25
Reprobado	5	29

Por lo que se muestra en la tabla 10, si bien no se tiene una accuracy muy alto es mayor que si se clasificara a todos como reprobados o no reprobados, además, el valor más alto es el de recall lo que indica que se están clasificando muy pocos alumnos como no reprobados cuando sí lo son. Este tipo de error sería el más costoso en este modelo debido a que en caso de que se decidiera apoyar a los clasificados como reprobados a estos no se les tomaría en cuenta.

Tabla 10: Métricas de desempeño

Accuracy	***Precision***	***Recall***	***Specifity***	***F1***
0.63	0.54	0.85	0.40	0.66

4. Conclusiones

Se puede concluir que los clústeres formados al aplicar análisis de clúster jerárquico tienen una marcada diferencia que demuestra que los alumnos que pertenecen al clúster dos tienen valores que indican una desventaja en cuestiones sociodemográficas, económicas, técnicas, de salud y académicas; por ejemplo, en el clúster dos existen altos porcentajes en la reprobación de los alumnos, se tiene un ingreso familiar mensual bajo, los alumnos dedicaron menos tiempo a estudiar y realizaron actividades extra al estudio (como trabajar) durante la pandemia, en comparación con los alumnos del clúster uno.

Sobre el modelo de regresión logística aplicado a la clasificación de alumnos se puede decir que es un modelo que puede clasificar medianamente bien si un alumno pertenece al grupo de reprobados durante la nueva modalidad de estudio. Además, el modelo da un marco de referencia sobre los factores que influyen en que un alumno repruebe en la modalidad de estudios en línea y por tanto es de ayuda para detectar y tratar de mejorar sus condiciones para que su desempeño académico sea mejor. El modelo funcionaría como base para el tomador de decisiones, ya que, dependiendo de cada alumno se puede decidir cuál es la forma de apoyarlo de acuerdo a sus condiciones. Claramente un algoritmo de clasificación como lo es la regresión logística suele no ser suficiente y la mejor opción puede ser el que se revise cada caso por separado para identificar a los alumnos en riesgo de reprobar. También se puede decir que existe cierto perfil de riesgo de los alumnos al que hay que poner atención, el cual es: quienes perciben sus calificaciones como más bajas, se conectan a clases por medio de un dispositivo de acceso público y/o han considerado la baja temporal o definitiva durante el periodo de clases en línea.

Referencias

Abidah, A., Hidaayatullaah, H. N., Simamora, R. M., Fehabutar, D., Mutakinati, L. (2020). The impact of Covid-19 to Indonesian education and its relation to the philosophy of "MerdekaBelajar". SiPoSE: Studies in Philosophy of Science and Education, 1(1), 38–49.

Braga-Neto, U. (2020). Fundamentals of Pattern Recognition and Machine Learning. Springer. https://doi.org/10.1007/978-3-030-27656-0.

Bramer, M. (2013). Principles of Data Mining. (2a ed.) Springer. DOI 10.1007/978-1-4471-4884-5.

Cirami, L., Ursino, D. J., Beltramino Persoglia, A., Mancevich, L. A. y Andreau. (2018). Afrontamiento emocional en situaciones de estrés académico: resultados preliminares de un análisis de clúster jerárquico con estudiantes universitarios. IX Congreso Internacional de Investigación y Práctica Profesional en Psicología XXIV Jornadas de Investigación XIII Encuentro de Investigadores en Psicología del MERCOSUR. Facultad de Psicología - Universidad de Buenos Aires, Buenos Aires.

Cox, D. R., Snell, E. J. (1989). Analysis of Binary Data. (2nd ed.). London: Chapman and Hall Ltd.

Dong, J. (2020). Online learning and teaching experiences during the covid-19 pandemic: a case study of Bangladeshi students receiving China's higher education. Sciedu Press. 9(2). 37- 45. https://doi.org/10.5430/elr.v9n2p37

González, T., De la Rubia, M. A., Hincz, K. P., Comas-Lopez, M., Subirats, L., Fort, S., Sacha, G. M. (2020). Influence of COVID-19 confinement on students' performance in higher education. PLoS ONE, 15(10 October). https://doi.org/10.1371/journal.pone.0239490.

Heredia, R., Rodríguez, H., Vilalta, A. (2014). Predicción del rendimientoen una asignatura empleando la regresión logística ordinal. SciElo. https://scielo.conicyt.cl/scielo.php?pid=S0718-07052014000100009&script=sci_arttext.

Jaggia, S., Kelly, A., Lertwachara, K., Chen, L. (2021). Business analytics. New York, USA: McGraw Hill Education.

Kapasia, N., Paul, P., Roy, A., Saha, J., Zaveri, A., Mallick, R., Barman, B., Das, P., Chouhan, P.(2020). Impact of lockdown on learning status of undergraduate and postgraduate students during COVID-19 pandemic in West Bengal, India. Children and Youth Services Review,Volume 116. https://doi.org/10.1016/j.childyouth.2020.105194.

Martínez, J. R., Ferrás, Y., Bermudez, L. L., Ortiz, Y., Pérez, H. E. (2020, 5 junio). Regresión logística y predicción del bajo rendimiento académico de estudiantes en la carrera Medicina. Revista Electrónica Dr. Zoilo E. Marinello Vidaurreta http://www.revzoilomarinello.sld.cu/index.php/zmv/article/view/2230/pdf_691.

Mathivet, V. (2018). Inteligencia Artificial para desarrolladores. (2a ed.). Ediciones Eni.

Murphy, M. P. A. (2020). COVID-19 and emergency eLearning: Consequences of the securitization of higher education for post-pandemic pedagogy. Contemporary Security Policy, 41(3), 492–505. https://doi.org/10.1016/j.jad.2020.06.034

Tabla 7: Resultados de los atributos de cada clúster

Factor	*Atributo*	*Clúster 1 N = 284*	*Clúster 2 N = 112*
Sociodemográfico	Edad	21	21
	Semestre	6	5
	Personas que viven en el hogar	4	5
	Pertenencia una minoría social[1]	0.07	0.14
	Residencia en zona urbana[1]	0.58	0.42
Económicos	Ingreso mensual en el hogar menor a $6,000[1]	0.67	0.88
	Tuvo afectación económica a causa de la pandemia[1]	0.79	0.92
Técnicos	Posee una laptop para tomar sus clases[1]	0.68	0.46
	Posee al menos un dispositivo electrónico propio[1]	0.85	0.68
	Funcionalidad de su dispositivo electrónico[2]	4	3
	Calidad de la señal de internet para estudiar[2]	3	2
Salud	Nivel de ansiedad durante la pandemia[2]	3	4
	Tratamiento psicológico durante la pandemia[1]	0.08	0.22
Situación académica	Al menos una materia reprobada durante la pandemia[1]	0.29	0.76
	Calificaciones inferiores durante la pandemia[1]	0.09	0.58
	Dedicó menos tiempo al estudio durante la pandemia[1]	0.1	0.59
	Realiza actividades extra al estudio[1]	0.69	0.83
	Consideró darse de baja temporal durante la pandemia[1]	0.52	0.86

[1] Los valores representan porcentajes promedio para cada clúster.
[2] Los valores están medidos en la escala de Likert de 1 a 5, donde 1 es nada y 5 demasiado.

Tabla 8: Resumen de la Regresión Logística

Variable	*Coef.*	*Std.Err.*	*z*	*P>|z|*	*[0.025*	*0.975]*	*EXP (coef)*
Tipo de posesión_renta	0.6712	0.7092	0.9465	0.3439	-0.7187	2.0611	1.95658381
Baja t o d	0.5101	0.2505	2.0367	0.0417	0.0192	1.001	1.66545773
ESTADO CIVIL_Soltera(o)	0.299	0.9136	0.3272	0.7435	-1.4917	2.0896	1.34850962
Reacción uaeh	0.1766	0.1143	1.5456	0.1222	-0.0474	0.4006	1.19315374
Edad	0.1244	0.0508	2.4486	0.0143	0.0248	0.224	1.13246877
Semestre	-0.1612	0.0602	-2.678	0.0074	-0.2792	-0.0432	0.85112183
Sexo_femenino	-0.2924	0.2447	-1.195	0.2322	-0.772	0.1873	0.74646989
Dispositivo_laptop	-0.341	0.2717	-1.255	0.2094	-0.8734	0.1915	0.71105891
Tiempo de estudio	-0.3806	0.1202	-3.167	0.0015	-0.6161	-0.1451	0.68345122
Afectación económica	-0.7056	0.3261	-2.164	0.0305	-1.3447	-0.0664	0.4938122
DISPOSITIVO_Computadora de escritorio	-0.733	0.4054	-1.808	0.0706	-1.5275	0.0616	0.48046543
Calificaciones	-1.1081	0.2192	-5.055	0	-1.5378	-0.6785	0.33018572

Odriozola-González, P., Planchuelo-Gómez, A., Irurtiaa, M. J., Luis-García, R. (2020). Psychological effects of the COVID-19 outbreak and lockdown among students and workers of a Spanish university. Psychiatry Research, Volume 290. https://doi.org/10.1016/j.psychres.2020.113108.

Organización Mundial de la Salud (OMS). (2021). COVID-19 Dashboard. Geneva: World Health Organization, 2020. Consultado el 19 de julio de 2021 en: https://covid19.who.int/.

Rico, J. J., Rodríguez, A. G., & Vilalta, J. A. (2014, 9 enero). Empleo de la regresión logística ordinal para la predicción del rendimiento académico. http://www.invoperacional.uh.cu/index.php/InvOp/article/view/415.

Robles, A., Cortés, P., Muñuzuri, J., Barbadilla, E. (2020, 1 abril). Aplicación de la regresión logística para la predicción de roturas de tuberías en redes de abastecimiento de agua. Dirección y Organización. 70, 78-85. https://www.revistadyo.es/index.php/dyo/article/view/570.

UNESCO (2021a). New UNESCO global survey reveals impact of COVID-19 on higher education. Recuperado de https://en.unesco.org/news/new-unesco-global-survey-reveals-impact-covid-19-higher-education.

UNESCO. (2021b). Education: From disruption to recovery. Consultado el 19 de julio de 2021 en: https://en.unesco.org/covid19/educationresponse/.

Wang, Z, Hai-Lian Y. Yun-Qing, Y., Dan, L., Zhi-Hao, L., Xi-Ru, Z., Yu-Jie, Z., Dong, S., Pei-Liang, C., Wei-Qi, S., Xiao-Meng, W., Xian-Bo, W., Xing-Fen, Y., Chen, M. (2020, 1 octubre). Prevalence of anxiety and depression symptom, and the demands for psychological knowledge and interventions in college students during COVID-19 epidemic: A large cross-sectional study. ScienceDirect, 275. https://www.sciencedirect.com/science/article/abs/pii/S0165032720323922?via%3Dihub

Winston, W. L., Albright, S. C. (2019). Practical management science. (6a ed.). Cengage.