**Artículo Original**

# Patrones de desempeño docente en Educación a Distancia utilizando aprendizaje automático

## Teacher Performance Patterns in Distance Education Using Machine Learning

Kárilyn Brunett Zarza[1,2], Rosa María Valdovinos Rosas[2], Patricia del Cármen Montaño Reyes[2], José Jacobo Gómez Quiróz[1]

[1] Universidad Nacional Rosario Castellanos. México
[2] Universidad Autónoma del Estado de México. México

kbrunettz@uaemex.mx, rvaldovinosr@uaemex.mx, pamr@uaemex.mx, jgq2405irc@gmail.com

**Recibido:** 06/06/2025 | **Aceptado:** 11/02/2026

**Cita sugerida**: K. Brunett Zarza, R. M. Valdovinos Rosas, P. del C. Montaño Reyes, J. Jacobo Gómez Quiróz, "Patrones de desempeño docente en Educación a Distancia utilizando aprendizaje automático," *Revista Iberoamericana de Tecnología en Educación y Educación en Tecnología*, no. 43, pp. 41-54, 2026. doi:10.24215/18509959.43.e4.

**Esta obra se distribuye bajo LicenciaCreative Commons CC-BY-NC 4.0**

## Resumen

El crecimiento acelerado de la Educación a Distancia ha impulsado la adopción de plataformas que generan grandes volúmenes de datos, pero estos no se analizan suficientemente y la información podría perderse. Este estudio perfila al docente virtual de la plataforma SEDUCA de la UAEMéx empleando aprendizaje automático y el ciclo CRISP-DM. Se realizó un análisis cuantitativo con 1 041 registros y 24 variables obtenidas del reporte de desempeño docente de la plataforma. En el preprocesamiento se realizó limpieza, imputación con K-NN y detección de *outliers*. Para el análisis descriptivo se aplicó *K-means* y Agrupamiento Jerárquico; se validó con coeficiente de silueta; la fase predictiva utilizó *Random Forest* optimizado con *GridSearch* y validado mediante matriz de confusión y validación cruzada. Se identificaron tres perfiles: docentes de interacción mínima (15 %), moderada (66 %) y alta (19 %). Las variables más influyentes fueron total de actividades calificadas y porcentaje de evidencias entregadas a tiempo; otras, como correos y número de alumnos, mostraron menor relevancia. El modelo alcanzó una exactitud del 95,2 % tras optimización. Este enfoque ofrece un marco para que las universidades aprovechen datos de sus plataformas en la planificación de la formación y selección de docentes virtuales.

*Palabras clave*: Educación a distancia; Minería de datos educativos; Desempeño docente; Aprendizaje automático; Sistema de gestión de aprendizaje.

## Abstract

The rapid expansion of distance education has driven the adoption of learning platforms that generate vast volumes of data; however, these data often go unanalyzed, risking the loss of valuable insights. This study profiles virtual instructors on the UAEMéx SEDUCA platform using machine learning within the CRISP-DM framework. A quantitative analysis of 1,041 records and 24 variables extracted from the platform's teaching performance report was conducted. Preprocessing involved data cleaning, K-NN imputation, and outlier detection. For descriptive analysis, *K-means* and Hierarchical *Clustering* was applied and validated with silhouette scores; the predictive phase employed a

*Random Forest* classifier optimized with *GridSearch*CV and validated via confusion matrix and cross-validation. Three instructor profiles were identified—minimal interaction (15 %), moderate interaction (66 %), and intensive interaction (19 %). The most influential predictors were total graded activities and the percentage of on-time submitted evidence, whereas features such as email exchanges and student count proved less relevant. The model achieved 95.2 % accuracy after optimization. This approach offers universities a reproducible framework for leveraging platform data to inform virtual instructor training and selection.

***Keywords:*** Distance education; Educational data mining; Teaching performance; Machine learning; Learning management system.

# 1. Introducción

La educación a distancia (EaD) en el nivel superior ha crecido significativamente, impulsada por los avances tecnológicos que han facilitado el acceso a la educación desde cualquier lugar con conexión a Internet. Los informes Horizon de 2020, 2021 y 2022 destacan que la educación en línea de alta calidad, los modelos híbridos de enseñanza y la formación docente para un desempeño óptimo en estos contextos son algunas de las tendencias principales en el ámbito de la educación superior [1], [2], [3]. Particularmente, el informe del 2023 resalta el aumento de la inteligencia artificial (IA) en la educación y la difuminación de las fronteras entre las modalidades de aprendizaje presencial y en línea [4].

De este modo, la EaD se presenta como una solución clave para ampliar la cobertura educativa de forma equitativa, especialmente en regiones marginadas o aisladas, posicionándose como una herramienta eficiente y accesible para garantizar este derecho [5]. Esto se alinea con los objetivos de la Agenda 2030, específicamente en la meta 4.4, que busca asegurar el acceso igualitario del 100% de hombres y mujeres a una formación técnica, profesional y superior de calidad, incluida la enseñanza universitaria [6] aspecto que sería difícil sin la implementación de la EaD.

En México, de acuerdo con la Asociación Nacional de Universidades e Instituciones de Educación Superior (ANUIES) [7]; la matrícula de estudiantes de educación superior y posgrado en modalidades no escolarizada y mixta ha mostrado un crecimiento constante. Entre los ciclos 2016–2017 y 2022–2023, se registró un aumento del 73.72% (Figura 1).

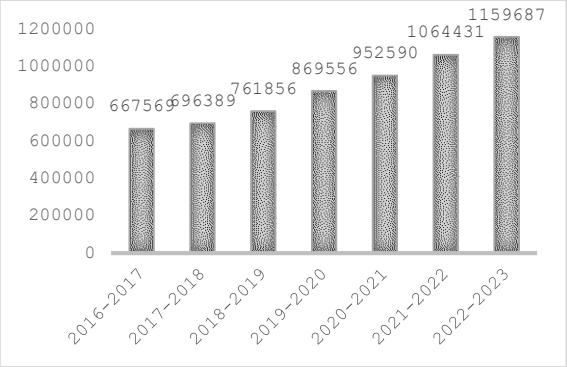



*Figura 1.* Evolución de la matrícula en modalidades no escolarizada y mixta en Educación Superior (2016–2023).

*Nota:* Las matrículas consideran a los estudiantes de modalidad no escolarizada y mixta de licenciatura, especialidad, maestría y doctorado de instituciones integradas a la ANUIES. Fuente: Elaboración propia con datos de los Anuarios Estadísticos de Educación Superior de los ciclos escolares del 2016-2017 al 2022-2023 [7].

La Universidad Autónoma del Estado de México (UAEMéx), cuenta con planteles distribuidos en diversas regiones del Estado de México, ofrece programas de bachillerato, licenciatura y posgrado en modalidades escolarizada (presencial), mixta y no escolarizada (en línea) [8] . En el caso de las licenciaturas en modalidad mixta, estas están obligadas a incluir al menos una materia que utilice un guion instruccional en línea. Para este fin, desde 2004, la UAEMéx., utiliza el Sistema de gestión de educación llamado SEDUCA como plataforma tecnológica para la educación a distancia y mixta [9]. Desde sus inicios hasta el día de hoy se ha observado un crecimiento constante en la matrícula de estudiantes de nivel superior en modalidad a distancia, lo que refleja un interés cada vez mayor por la educación en línea (Tabla1).

Tabla 1. Matrícula de estudios profesionales por ciclo escolar en la UAEMéx

| Licenciatura | Mód. | 16-17 | 17-18 | 18-19 | 19-20 | 20-21 | 21-22 | 22-23 | 23-24 | Δ (%) |
|---|---|---|---|---|---|---|---|---|---|---|
| Administración | E | 87 | 91 | 138 | 164 | 129 | 167 | 197 | 193 | 122 |
| Logística | E | 26 | 33 | 40 | 23 | 30 | 62 | 67 | 45 | 73 |
| Enseñanza del inglés | E | 23 | 21 | 22 | 21 | 21 | 19 | 32 | 27 | 17 |
| Derecho internacional | E | 26 | 43 | 29 | 32 | 34 | 64 | 52 | 39 | 50 |
| Informática administrativa | E | 44 | 74 | 76 | 55 | 51 | 75 | 70 | 66 | 50 |
| Negocios internacionales | E | 13 | 24 | 28 | 24 | 23 | 45 | 43 | 47 | 262 |
| Relaciones econ. internacionales | E | 24 | 28 | 33 | 31 | 38 | 30 | 33 | 25 | 4 |
| Administración | M | – | – | – | 919 | 984 | 1 008 | 1 125 | 1 087 | 18 |
| Contaduría | M | – | – | – | 1 018 | 1 081 | 1 082 | 1 126 | 1 099 | 8 |
| Enfermería* | M | 50 | 46 | 55 | 767 | 796 | 833 | 854 | 899 | 17 |
| Turismo | M | – | – | – | 686 | 609 | 579 | 576 | 550 | −20 |
| Trabajo social | M | – | – | – | 92 | 106 | 86 | 104 | 98 | 7 |
| Informática administrativa | M | – | – | – | 472 | 255 | 346 | 334 | 449 | −5 |
| Ing. en computación | M | – | – | – | 738 | 613 | 750 | 821 | 743 | 1 |
| Gestión info. en Redes Sociales | M | – | 32 | – | 39 | 46 | 38 | 68 | 49 | 26 |
| Medios alt. de sol. de conflictos | M | – | 7 | 17 | 74 | 65 | 74 | 74 | 70 | 900 |
| Gastronomía | M | – | – | – | 206 | 196 | 192 | 184 | 172 | |
| Sumatoria general | | 293 | 399 | 438 | 5 361 | 5 077 | 5 450 | 5 760 | 5 658 | |

*Nota:* E = En Línea; M = Mixta; "–" indica que el programa no se ofertó en ese ciclo. Δ (%) corresponde al cambio porcentual acumulado desde 2016-2017 hasta 2023-2024. Todos los valores son números de estudiantes matriculados en cada ciclo escolar. Para el cálculo del

crecimiento en la Licenciatura en enfermería se consideró como inicio el valor de 767 correspondiente a la matricula del 2019-2020. [10].

Como puede verse en la figura 2, la matrícula de las Licenciaturas de la UAEMéx a lo largo de los últimos años muestra un crecimiento significativo, con un aumento notable entre los periodos 2018-2019 y 2019-2020. A pesar de algunas caídas en ciertos periodos, la matrícula se ha mantenido en niveles superiores a los iniciales, reflejando un interés creciente en la oferta académica (Figura 2).

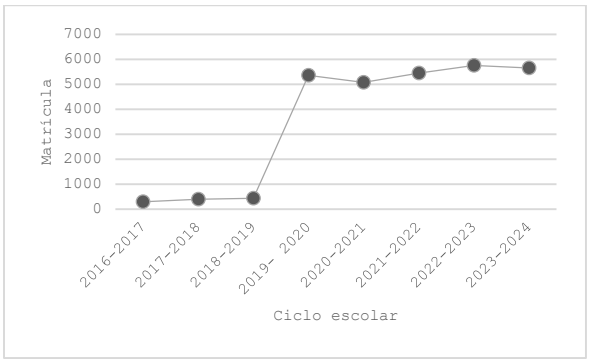



Figura 2. Evolución de la Matrícula de estudiantes de licenciatura de la UAEMéx por ciclo escolar.

*Nota.* Estadísticas de matrícula profesional por ciclo escolar [10].

En un entorno virtual, un LMS (*Learning Management System*) o Sistema de Gestión del Aprendizaje es clave, ya que permite administrar cursos, distribuir materiales, evaluar el desempeño y mantener la comunicación entre docentes y estudiantes. El uso de un LMS como SEDUCA no solo facilita la gestión del aprendizaje, sino que también genera y almacena una gran cantidad de datos sobre la actividad de los estudiantes, sus interacciones, entregas, calificaciones y patrones de participación. Estos datos son susceptibles de análisis, proporcionando información valiosa para mejorar los procesos educativos. Sin embargo, el mero acopio de datos académicos —como calificaciones, niveles de asistencia o resultados de evaluación— no basta para impulsar mejoras pedagógicas; es necesario aplicar un proceso metodológico sistemático que permita interpretarlos con rigor. [11] subraya que la depuración, transformación y modelado de la información son etapas clave para revelar tendencias y patrones útiles cuyo adecuado tratamiento y análisis transforman esa complejidad en ventajas estratégicas para las organizaciones e instituciones. A su vez, [12]. advierte que el análisis debe orientarse a responder preguntas de investigación específicas, utilizando métodos estadísticos que garanticen la validez y reproducibilidad de los resultados en proyectos científicos.

Las preguntas que guiaron este trabajo fueron las siguientes: ¿Cómo se puede caracterizar al docente virtual a partir de su interacción en una plataforma de aprendizaje?, ¿Cuáles son los indicadores clave de la interacción docente-alumno en una plataforma de aprendizaje virtual y cómo pueden utilizarse para caracterizar al docente?, ¿Qué patrones de comportamiento pueden identificarse en la interacción del docente virtual a partir del análisis de los datos de la plataforma de aprendizaje? Para responder, fue imprescindible aplicar técnicas de minería de datos y procesar grandes volúmenes de información pues era necesario revelar patrones de comportamiento.

La minería de datos, o descubrimiento de conocimiento en bases de datos, se define como el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y comprensibles en grandes conjuntos de datos [13]. Por su parte, la minería de datos educativos, reconocida como disciplina desde los estudios pioneros de [14] se entiende como la aplicación de algoritmos de minería de datos y técnicas de aprendizaje automático al análisis de datos generados en entornos de enseñanza-aprendizaje, con el propósito de descubrir patrones de comportamiento estudiantil, anticipar resultados académicos y optimizar los procesos pedagógicos.

En el ámbito de la educación superior, algunos estudios han comenzado a aplicar técnicas de minería de datos educativos y aprendizaje automático para analizar información digital vinculada con la práctica docente en plataformas institucionales. Estas investigaciones han utilizado, por ejemplo, sistemas expertos basados en reglas y árboles de decisión para clasificar el uso docente de sistemas de gestión del aprendizaje [29], modelos de simulación apoyados en algoritmos como K-NN para explorar decisiones pedagógicas en entornos tipo MOOC [30], enfoques de aprendizaje profundo para analizar indicadores emocionales asociados a la actividad docente [31] y técnicas de minería de textos combinadas con clasificación automática para examinar evaluaciones estudiantiles sobre el desempeño del profesorado [32]. No obstante, aunque estos trabajos evidencian el potencial de la minería de datos para estudiar dimensiones específicas de la docencia universitaria, aún son escasos los estudios que realicen una caracterización sistemática del docente a partir de múltiples variables de comportamiento registradas en un LMS institucional, lo que sugiere un campo de investigación todavía en desarrollo.

En ese sentido, el objetivo de este estudio fue caracterizar al docente virtual en la plataforma SEDUCA durante el semestre 2023B a partir de sus interacciones —identificando los indicadores clave de la relación docente-alumno y los patrones de comportamiento emergentes— mediante técnicas de minería de datos educativos, con el fin de ofrecer un marco analítico replicable que apoye a la UAEMéx y a otras instituciones de educación superior en la evaluación y mejora continua del desempeño docente en entornos virtuales.

## 2. Método

Para el logro del objetivo planteado se utilizó el Proceso Estándar de la Industria para Minería de Datos (*Cross-Industry Standard Process for Data Mining*, CRISP-

DM) mostrado en la figura 3.

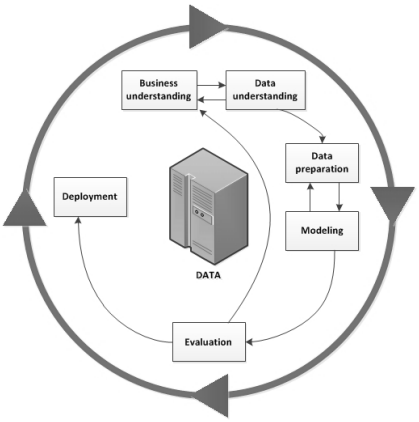



*Figura 3*. Metodología de Ciencia de Datos: CRISP-DM

*Nota:* Las seis fases de la Metodología de Ciencia de Datos [15].

La metodología CRISP-DM es una de las más empleadas actualmente para el desarrollo de proyectos de minería de datos y organiza los proyectos de minería de datos en seis fases secuenciales [16]:

1) **Comprensión del negocio:** Define el problema, objetivos y contexto actual para establecer metas claras y un punto de comparación.

2) **Comprensión de los datos:** Recopila, describe y explora los datos para conocer su estructura, calidad y propiedades estadísticas.

3) **Preparación de datos:** Limpia, transforma e ingenia indicadores a partir de los datos crudos, garantizando su idoneidad para el modelado.

4) **Modelado:** Selecciona técnicas y divide datos de entrenamiento y prueba para generar y ajustar iterativamente el mejor modelo.

5) **Evaluación:** Mide la calidad del modelo con métricas estadísticas y validación experta; decide si avanzar, revisar etapas previas o reiniciar.

6) **Implementación:** Despliega el modelo y documenta resultados, asegurando su uso práctico y monitoreando su desempeño para futuras mejoras.

En las siguientes subsecciones se brinda una explicación detallada de la forma en que fueron abordadas cada una de las etapas en esta investigación.

## 2.1 Entendimiento del negocio

SEDUCA es una plataforma que respalda las diferentes modalidades educativas de la UAEMex, a través de la cual se imparten programas de bachillerato a distancia, siete licenciaturas en modalidad a distancia, diez en modalidad mixta, una maestría y un doctorado [17]. La plataforma se organiza en "comunidades" las cuales representa un grupo distinto de clase, en éstas se alojan materiales de apoyo, páginas de contenido organizadas por índices temáticos, actividades de tipo: cuestionario, portafolio, foro, wikis basados en los programas de estudio de cada unidad de aprendizaje (materia).

La UAEMéx evalúa en escala 0–10 mediante dos parciales: si la estudiante promedia ≥ 8 en ambos exenta el examen final; de lo contrario debe pasar por los exámenes ordinario, extraordinario y de suficiencia antes de recursar la materia. Comprender el proceso de evaluación en la UAEMéx resulta clave, pues el reporte de desempeño docente incorpora variables que reflejan el rendimiento en el primer y el segundo parcial.

## 2.2 Entendimiento de los datos

Los datos utilizados en este estudio provienen del "Reporte de Desempeño Docente", una hoja de cálculo generada por SEDUCA que recopila información sobre el uso de la plataforma por parte del profesorado. Cada comunidad cuenta con 35 características registradas. Durante el semestre 2023B (de agosto 2023 a enero 2024) se generaron 1773 registros que corresponden a grupos de 336 materias diferentes impartidas en 18 diferentes carreras universitarias y distribuidas en 23 espacios académicos de la Universidad Autónoma del Estado de México.

En SEDUCA, cada grupo de clase se registra como una comunidad única. Aunque pueda compartir el mismo nombre con otras comunidades que correspondan a la misma materia, cada comunidad se distingue por el docente que la imparte y los estudiantes que la conforman. En este sentido, cada uno de los 1,773 registros corresponde a una comunidad diferente, asegurando que cada unidad de análisis corresponde a un grupo de clase independiente. Es importante señalar que, de estas comunidades no todas fueron analizadas, más adelante se verá que ese conjunto inicial de 1773 registros se redujo a 1041 en la etapa del preprocesamiento con la intención de garantizar que los datos fueran confiables y adecuados para cualquier análisis o modelado.

## 2.3 Preparación de los datos (preprocesamiento)

Para garantizar la integridad de los datos y la fiabilidad de los modelos, se llevó a cabo un riguroso proceso de limpieza, una de las etapas más extensas del ciclo de ciencia de datos. En primer lugar, se aplicó una reducción de dimensionalidad eliminando aquellas variables que presentaban un alto porcentaje de valores nulos. A continuación, se inspeccionaron los reportes de desempeño docente de SEDUCA y se detectaron ocho columnas cuyos valores eran sistemáticamente cero, sin justificación aparente. Estas variables se descartaron, ya que su inclusión habría incrementado la complejidad del modelo sin aportar información significativa. Como resultado, se eliminaron ocho columnas:

1. Porcentaje-Eficiencia-Terminal.
2. Aprobados.
3. Aprobados ordinarios.
4. Aprobados extraordinarios.
5. Aprobados título.
6. Histórico de socializa por comunidad.
7. Asesoría académica Respondidas/Recibidas.
8. Histórico de chat personal.

Además, por razones de privacidad, se excluyeron tres columnas categóricas: Asesor, Bloque y Grupo, ya que contenían información directamente relacionada con la identidad del profesorado.

También se descartaron las variables 1er_P_Ponderación_semanas y 2oP_Ponderación-semanas, ya que no resultaban necesarias para los objetivos del estudio. Se mantuvieron las columnas: Carrera, Comunidad y Espacio Académico pero no fueron consideradas para el modelado presentado en este estudio, sino para un análisis posterior .

Tras este proceso, el conjunto de datos quedó con 24 características organizadas en dos tipos: categóricas (3) y numéricas (21). Además, se unificó la nomenclatura de las 24 variables para que cada nombre describiera con claridad su contenido.

*Categóricas:*

1. Carrera: Nombre de la licenciatura
2. Comunidad: Nombre de la comunidad (materia, no considerada para el entrenamiento)
3. Espacio_Académico: Nombre del espacio académico

*Numéricas:*

4. alumnos: Cantidad de alumnos en el grupo

Para el primer parcial (identificadas al inicio por "1p") se tienen:

5. 1p_total_actividades_asignadas
6. 1p_total_actividades_evaluativas
7. 1p_porcentaje_evidencias_calificadas
8. 1p_total_actividades_calificadas
9. 1p_total_actividades_entregadas
10. 1p_porcentaje_evidencias_calificadas_tiempo
11. 1p_total_actividades_calificadas_tiempo
12. 1p_porcentaje_aprobados

Para el segundo parcial (identificadas al inicio por "2p") se tienen:

13. 2p_total_actividades_asignadas
14. 2p_total_actividades_evaluativas
15. 2p_porcentaje_evidencias_calificadas
16. 2p_total_actividades_calificadas
17. 2p_total_actividades_entregadas
18. 2p_porcentaje_evidencias_calificadas_tiempo
19. 2p_total_actividades_calificadas_tiempo
20. 2p_porcentaje_aprobados

Variables que miden las características finales del semestre:

21. total_actividades_calificadas
22. trabajos_por_calificar
23. correos_enviados
24. correos_recibidos

Concluida esta primera fase de limpieza —en la que el número de variables se redujo de 37 a 24—, se revisaron los registros del conjunto de datos. De los 1 773 registros iniciales, 620 correspondían a comunidades abiertas que no presentaron actividad alguna; dichos registros se eliminaron, dejando un total de 1 153 registros válidos.

Gracias a esta depuración, el total de valores faltantes del conjunto de datos disminuyó un 16,01 %. Esta depuración responde a la flexibilidad de uso de plataformas digitales permitida en la modalidad mixta de la UAEMéx. El conjunto final contiene exclusivamente registros con evidencia de uso de la plataforma.

Las dimensiones del conjunto de datos resultante son:

- Registros válidos: 1 153
- Variables (características): 24
- Valores faltantes: 2 100 (4,94 % del total de celdas)

### 2.3.1 Imputación de datos

Para manejar los registros con datos faltantes, se realizó su imputación utilizando el algoritmo de los K-vecinos más cercanos (*K-Nearest Neighbors*, KNN). KNN es un método ampliamente utilizado para imputación de datos faltantes, ya que estima los valores en función de las observaciones más similares dentro del conjunto de datos, preservando su estructura interna [18]. El funcionamiento del algoritmo es como sigue:

a) Selección del parámetro *k*: Se determinó que el valor óptimo era *k*=4, ya que un número demasiado bajo o alto podría afectar la precisión de la imputación.

b) Cálculo de distancia: Se utilizó la distancia euclidiana modificada, considerando valores nulos como infinitamente lejanos para evitar sesgos.

c) Identificación de vecinos cercanos: Se seleccionaron los 4 registros más cercanos al punto con datos faltantes.

d) Estimación de valores: Se calculó el valor faltante como el promedio de las características de los vecinos seleccionados.

Tras la imputación, el conjunto de datos quedó con 1,153 registros y 24 columnas sin valores faltantes.

### 2.3.2 Tratamiento de valores atípicos

El siguiente paso del preprocesamiento consistió en la detección y eliminación de valores atípicos *(outliers),* ya que estos pueden distorsionar los resultados del análisis. Se utilizó el método z-score considerando como *outliers* aquellos valores cuyo puntaje absoluto fuera mayor a 3, asumiendo una distribución normal de los datos. El procedimiento consistió en:

a) Cálculo del z-score para cada observación en las columnas numéricas.

$$z = \frac{x-\mu}{\sigma} \quad \text{.....(1)}$$

Donde: x es el valor observado, μ es la media muestral y σ la desviación estándar de la muestra.

Tras este procedimiento se eliminaron 112 registros con *outliers* para evitar sesgos en el análisis, dejando un conjunto de datos final de 1,041 registros.

## 2.4. Modelado

El estudio emplea tanto modelos descriptivos como predictivos. Para la fase descriptiva, dado que el conjunto de datos original carece de categorías predefinidas sobre tipos de docentes en SEDUCA, se recurrió a algoritmos de aprendizaje no supervisado. Este enfoque permite descubrir, de manera autónoma, patrones y estructuras ocultas en datos "en bruto" sin necesidad de etiquetas previas. A diferencia del aprendizaje supervisado, que requiere conjuntos de datos previamente categorizados; en el aprendizaje no supervisado se aplican técnicas de *clustering* (agrupamiento) por ejemplo: *K-means,* jerárquico [19].

Para la etapa de predicción, los perfiles docentes obtenidos mediante técnicas de *clustering* (etapa descriptiva) se utilizan como etiquetas supervisadas para entrenar un modelo de *Random Forest* (RF). Acorde a [20], RF es un algoritmo que tiene la finalidad de construir un conjunto de árboles de decisión entrenados sobre muestras y subconjuntos de características seleccionados aleatoriamente, y luego combinar sus predicciones para mejorar la precisión y estabilidad del modelo. Al aplicar aprendizaje supervisado en la fase predictiva, se garantiza que el sistema aproveche la estructura descubierta en los datos para generar predicciones precisas y rigurosas, ampliando así la utilidad práctica de los resultados en entornos educativos virtuales.

Con el fin de optimizar el rendimiento, se utilizó búsqueda en cuadrícula (*Grid Search*) para ajustar exhaustivamente hiperparámetros clave (número de estimadores, profundidad máxima y tamaño mínimo de muestra por división), [21]. Esta optimización garantizó una configuración de parámetros que maximizó la exactitud del modelo y mejoró su capacidad de generalización en nuevos registros.

Para facilitar la interpretación gráfica de la segmentación obtenida, las 24 variables finales se sometieron a un Análisis de Componentes Principales (PCA) reduciendo la dimensionalidad a los dos primeros componentes, que capturan la mayor parte de la varianza. Sobre esa proyección bidimensional se representaron las asignaciones de *K-means* y de agrupamiento jerárquico.

## 2.5 Evaluación del modelo

Para verificar la robustez del enfoque se emplearon técnicas específicas tanto para *clustering* como para clasificación

### 2.5.1 Validación del clustering descriptivo

Para evaluar la calidad del agrupamiento generado (fase descriptiva) se empleó el coeficiente de silueta el cual mide la cohesión interna de cada clúster y su separación respecto a los demás [22]. Se calculó el valor de silueta para cada observación y se obtuvo un promedio global por método (*K-means* y agrupamiento jerárquico). Un coeficiente cercano a 1 indica agrupamientos bien definidos, mientras que valores próximos a 0 o negativos sugieren solapamiento entre clústeres. Los resultados de silueta de ambos métodos se compararon para determinar cuál ofrecía una segmentación más consistente antes de proceder a la fase predictiva.

### 2.5.2 Validación del modelo predictivo

Para garantizar la robustez y la capacidad de generalización del *Random Forest* se aplicaron las siguientes técnicas:

a) Validación cruzada (5-fold): durante el proceso de optimización de hiperparámetros con *GridSearch* se particionó el conjunto de entrenamiento en cinco subconjuntos, de modo que en cada iteración uno se destinó a prueba y los demás a entrenamiento, minimizando el riesgo de sobreajuste.

b) Matriz de confusión: Esta matriz, de dimensión equivalente al número de clases, permite identificar tanto las instancias correctamente clasificadas (ubicadas en la diagonal principal) como los errores de clasificación (fuera de la diagonal). La Tabla 2 ilustra este concepto para un problema binario.

c)

*Tabla 2.* Matriz de confusión

| | **Predicción Clase (+)** | **Predicción Clase (-)** |
|---|---|---|
| **Clase real (+)** | Verdadero positivo (VP) | Falso negativo (FN) |
| **Clase real (-)** | Falso positivo (FP) | Verdadero negativo (VN) |

*Nota:* La matriz de confusión muestra la comparación entre las clases reales y las predicciones del modelo. En la diagonal principal se representan las instancias correctamente clasificadas para cada clase, mientras que los valores fuera de la diagonal indican las instancias mal clasificadas.

De este modo, las siguientes medidas pueden ser obtenidas:

1. Exactitud (Accuracy): Proporción de predicciones correctas realizadas por el modelo.
$$\frac{VP + VN}{(VP + VN + FP + FN)} \quad (2)$$
2. Precisión (Precision): Proporción de predicciones positivas correctas entre todas las predicciones positivas realizadas por el modelo.
$$\frac{VP}{(VP + FP)} \quad (3)$$
3. Sensibilidad (Recall): Es la proporción de ejemplos positivos reales correctamente identificados por el modelo.
$$\frac{VP}{(VP + FN)} \quad (4)$$
4. F1-score: sirve para evaluar el equilibrio entre precisión y sensibilidad del modelo, especialmente útil cuando las clases están desbalanceadas y se requiere una métrica que penalice tanto los falsos positivos como los falsos negativos.
$$\frac{(2 * Precisión * Recall)}{(Precisión + Recall)} \quad (5)$$

Para completar la fase descriptiva y preparar la validación estadística de los perfiles identificados, se incorporó un análisis de varianza de una vía (ANOVA) sobre las variables seleccionadas el cual mide la significación estadística de las diferencias entre las medias de los grupos determinados en la variable dependiente por los valores de las variables independientes [23]. Este procedimiento permite confirmar que las segmentaciones no son atribuibles al azar, sino que reflejan diferencias reales en el comportamiento docente.

# 3. Resultados

En esta sección se presentan los datos obtenidos tras la aplicación de la metodología, correspondientes a la fase de despliegue (*deployment).*

## 3.1 Agrupamiento descriptivo

Para explorar la estructura latente en los datos se aplicaron dos algoritmos de agrupamiento no supervisado: *K-means*: método de partición que asigna puntos a *k* centroides iterativamente para minimizar la inercia y Agrupamiento jerárquico**:** técnica aglomerativa que construye un dendrograma fusionando progresivamente los grupos según su distancia. Con estos dos enfoques se obtuvo una visión comparativa de la cohesión y separación de los grupos antes de proceder a elegir el número óptimo de clústeres y validar la segmentación.

Para determinar el número óptimo de clústeres se aplicó el método del codo (*Elbow Method*) sobre la inercia (WCSS) para valores de k entre 1 y 10. Como puede verse en la Figura 4, la reducción de WCSS se desacelera notablemente en k=3, lo que sugiere esa partición.

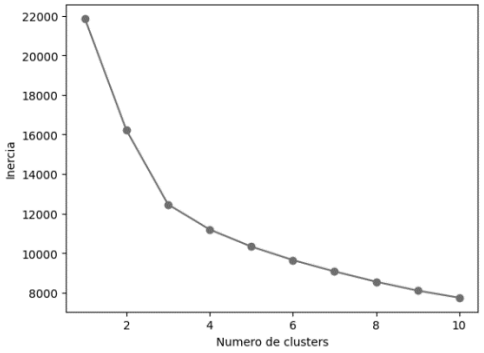



Figura 4. Método del Codo

*Nota:* Se ubica al 3 como mejor número de *clusters.*

Se construyó un dendrograma usando enlace promedio. Al cortar el árbol en la distancia umbral de máxima separación (nivel 50), surgieron también tres clústeres bien definidos (Figura 5).

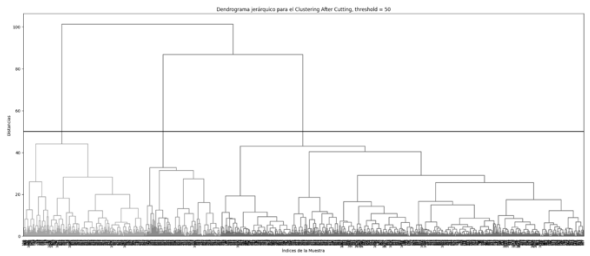


*Figura 5*. Dendrograma con corte a tres clústeres

Nota: la línea indica el nivel de corte.

El coeficiente de silueta promedio reforzó la elección de k=3: 0,3155 para *K-means* y 0,2920 para jerárquico, lo que señala una segmentación ligeramente más nítida con *K-means* (ver figura 6). Estos resultados indican que, aunque ambos métodos identificaron una estructura de clústeres en los datos, *K-means* logró una mejor cohesión interna y separación entre los grupos.

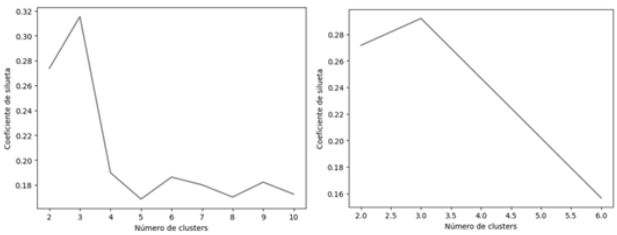



*Figura 6*. Coeficiente de silueta

*Nota:* Coeficiente de silueta con agrupamiento k medias (izquierda) y con agrupamiento jerárquico (derecha).

La validación del agrupamiento mediante el coeficiente de silueta sugiere que la estructura con tres clústeres es adecuada y que el método de *K-means* proporciona una

segmentación más clara del desempeño docente. A continuación, se muestra cómo se distribuyen los 1041 grupos gestionados por los docentes en los tres clústeres identificados (Tabla 3). Las etiquetas se mantienen genéricas —Clúster 1, Clúster 2 y Clúster 3— hasta completar el análisis de sus características y determinar posteriormente el perfil de cada uno.

*Tabla 3*. Distribución de registros por clúster (n=1041)

| Clúster | Registros | % |
|---|---|---|
| 1 | 158 | 15% |
| 2 | 687 | 66% |
| 3 | 196 | 19% |

*Nota:* Fuente: elaboración propia (2024)

Estos resultados confirman la existencia de tres perfiles básicos de gestión de grupos docentes y sirven como fundamento para la etapa predictiva.
Se aplicó un Análisis de Componentes Principales (PCA) a las 24 variables preprocesadas, reduciendo la dimensionalidad a dos ejes (PC1 y PC2) que capturan la mayor parte de la varianza.

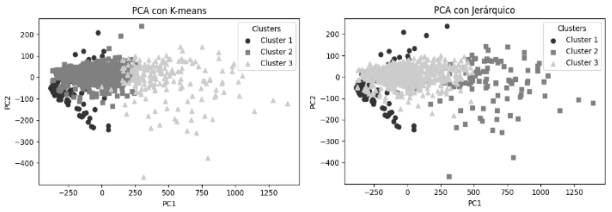



*Figura 7*. PCA de perfiles docentes (*K-means* vs. jerárquico)

*Nota.* Cada símbolo (círculo, cuadrado, triángulo) representa uno de los tres clústeres; los tonos de gris permiten la lectura en escala de grises. Los ejes PC1 y PC2 corresponden a las dos primeras componentes principales del análisis de 24 variables preprocesadas, usadas únicamente para visualización de la distribución de los grupos. Fuente: elaboración propia (2024)

En conjunto, la proyección PCA confirma visualmente la cohesión y separación de los tres perfiles docentes identificados por ambos métodos de agrupamiento (Figura 7). La clara delimitación observada en el plano PC1–PC2 refuerza la elección de k=3 y valida la estructura de clústeres obtenida.

## 3.2 Clasificación predictiva

Para asignar cada grupo nuevo a uno de los perfiles descubiertos se entrenó un Random Forest. Con el fin de garantizar la mejor configuración del modelo se empleó Grid Search con validación cruzada de cinco pliegues, explorando combinaciones de: Número de árboles (n_estimators), Profundidad máxima (max_depth), Mínimo de muestras por división (min_samples_split).

La Tabla 4 resume los valores que maximizaron la exactitud sobre el 20% de datos reservados:

*Tabla 4*. Hiperparámetros óptimos y exactitud (%)

| Método | No. árboles | Profundidad máx | Mínimo de muestras | Exactitud |
|---|---|---|---|---|
| *K-means* | 200 | 10 | 2 | 95.2% |
| Agrupamiento jerárquico | 50 | 10 | 2 | 97.6% |

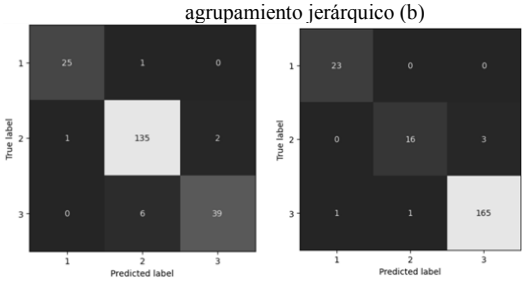
*Nota:* Validación cruzada (cv = 5) sobre el conjunto de prueba. Fuente: elaboración propia (2024)

Con estos parámetros se obtuvo un modelo robusto y listo para su evaluación final. El *Random Forest* entrenado se probó sobre datos no vistos y su desempeño se cuantificó mediante:

a) Matriz de confusión (Figura 8), que muestra aciertos y errores por clúster.

b) Reporte de clasificación (Tablas 5 y 6), con precisión, recall y F1-score para cada grupo.

*Figura 8*. Matriz de confusión obtenido con K-medias (a) y con agrupamiento jerárquico (b)




*Nota.* Mapa de calor de la matriz de confusión generada por *K-means* (a) y agrupamiento jerárquico (b), validado mediante *Random Forest*. Indica la distribución de clasificaciones correctas e incorrectas en los grupos formados. Fuente: elaboración propia (2024)

*Tabla 5*. Reporte de clasificación del algoritmo K-medias

| grupo | precisión | recall | F1 score |
|---|---|---|---|
| 1 | 96% | 96% | 96% |
| 2 | 95% | 98% | 96% |
| 3 | 95% | 87% | 91% |

*Nota.* Resultados de precisión, recall y F1-score para cada grupo generado por el algoritmo K-medias, validados mediante *Random Forest*. Fuente: Elaboración propia (2024)

*Tabla 6*. Reporte de clasificación del algoritmo de Agrupamiento Jerárquico

| Grupo | Precisión | Recall | F1 score |
|---|---|---|---|
| **1** | 96% | 100% | 98% |
| **2** | 94% | 84% | 89% |
| **3** | 98% | 99% | 99% |

*Nota.* Resultados de precisión, recall y F1-score para cada grupo generado por el algoritmo de Agrupamiento Jerárquico, validados mediante *Random Forest*. Fuente: Elaboración propia (2024)

Estos indicadores —todos por encima de 89% y una exactitud global por encima de 95%— demuestran la

capacidad del modelo para asignar nuevos registros a los perfiles definidos.

Tras comparar ambas técnicas de agrupamiento, se decidió continuar el análisis con los clústeres obtenidos mediante *K-means*, dado que presentaron un mayor coeficiente de silueta y ofrecieron una separación más clara entre perfiles. Todas las comparaciones de medias (ANOVA) y el entrenamiento del modelo predictivo se basan en esta segmentación de *K-means*.

## 3.3 Caracterización estadística de los perfiles docentes

Se aplica un análisis de varianza (ANOVA) para contrastar de forma rigurosa las diferencias de medias entre los grupos y confirmar su significación estadística. Finalmente, mediante diagramas de caja se visualiza la dispersión, mediana y rangos intercuartílicos de las principales métricas de desempeño, lo que facilita una comprensión clara y comparativa de las características distintivas de cada perfil de gestión de grupos virtuales.

### 3.3.1 Importancia de variables

El análisis de las variables más relevantes en un proceso de agrupamiento es esencial para entender los factores que contribuyen a la segmentación de los datos. Identificar las variables que más influyen en la formación de estos clústeres permite no solo interpretar mejor los resultados, sino también optimizar y ajustar los modelos de agrupamiento, especialmente cuando se busca mejorar la precisión y la consistencia de las agrupaciones. Las variables más importantes según el agrupamiento realizado con el algoritmo de *K-means* se muestran en la Figura 9. Hay que recordar que el prefijo 1p hace alusión al primer parcial y el 2p al segundo parcial.

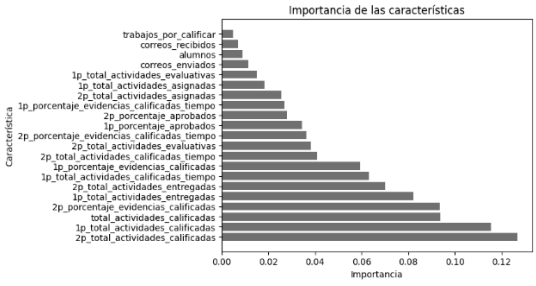



*Figura 9*. Características más relevantes

*Nota:* Las tres variables más relevantes para definir la clasificación de la docencia virtual son: total de variables calificadas en el 2° y 1er parcial, y el total de actividades calificadas.

Al analizar las 21 variables cuantitativas, se encontró que las variables más relevantes son: el total de actividades calificadas en el segundo y primer parcial, y la variable total de actividades calificadas. Lo que indica que el número total de actividades calificadas, tanto en los parciales como en su totalidad, juega un papel fundamental en la caracterización de la docencia virtual en el contexto estudiado. Y las tres variables menos importantes para el agrupamiento resultaron ser: trabajos por calificar, correos recibidos y número de alumnos.

### 3.3.2. Análisis de varianza (ANOVA)

Para evaluar la significancia de las diferencias entre los tres clústeres obtenidos, se aplicó un ANOVA de una vía con $\alpha$=0.05. Inicialmente, este análisis se extendió a las 21 variables numéricas, registrándose valores $p<3{,}34\times10^{-11}$ en todas ellas, lo que permite rechazar la hipótesis nula de igualdad de medias.

A continuación, se centró el ANOVA en los tres indicadores clave que sustentan la segmentación: Total de actividades calificadas en el primer parcial, Total de actividades calificadas en el segundo parcial, Porcentaje de evidencias calificadas a tiempo. Los resultados se muestran en la Tabla 7.

Tabla 7. Resultados del ANOVA para indicadores clave de desempeño docente

| Variable | F | p-valor |
|---|---|---|
| Total actividades calificadas (1er parcial) | 907.12 | < 0.001 |
| Total actividades calificadas (2º parcial) | 974.01 | < 0.001 |
| % evidencias calificadas a tiempo (1er parcial) | 229.43 | < 0.001 |

*Nota*. Valores de F y p valor aplicación de ANOVA

### 3.3.3. Interpretación de clústeres y análisis estadístico

En la visualización de los datos, se emplearon diagramas de caja y gráficos de barras para examinar las diferencias entre los clústeres. El diagrama de caja (Figura 10) de las tres variables más representativas para la clasificación ilustra cómo las medias de las actividades calificadas varían significativamente entre los clústeres, donde el tercer clúster muestra un número considerablemente mayor de actividades calificadas en comparación con los dos primeros. Esta diferencia destaca las variaciones en la carga de trabajo asignada a cada grupo.

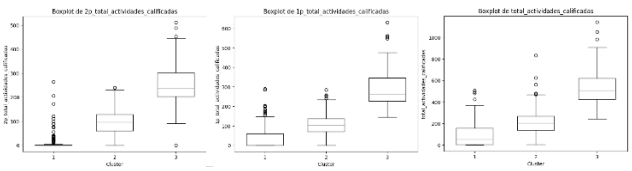



*Figura 10.* Diagrama de caja de las 3 variables más relevantes catalogadas por número de clúster

*Nota:* Se observa una diferencia significativa en los valores de cada grupo para cada una de las variables con valores más altos en las 3 características para el cluster No. 3.

Hasta este punto, la clasificación podría interpretarse únicamente en función de la cantidad de actividades calificadas durante el primer y segundo parcial. El clúster 3, en los tres casos analizados, destacó por un desempeño superior al calificar un mayor número de tareas.

Sin embargo, existen otras variables clave que permiten analizar más profundamente el desempeño de los docentes y clasificar a los clústeres en categorías de áreas de mejora, promedio y superior: porcentaje de evidencias calificadas a tiempo en el primer y segundo parcial y total de actividades entregadas.

La variable *porcentaje de evidencias calificadas a tiempo* mide si las actividades fueron retroalimentadas dentro del plazo estipulado, independientemente de la cantidad de actividades asignadas. En el caso de SEDUCA, el tiempo ideal sugerido para retroalimentar es de 72 horas posteriores a la recepción de la actividad. En el clúster 3, esta variable presenta un rango intercuartilar menor, indicando menos dispersión, y puntajes más altos. Como se observa en la Figura 11, el primer cuartil se encuentra cercano a 0.8 (0.74 en el primer parcial (a) y 0.73 en el segundo parcial (b)), lo que indica que el 75% de los docentes en este clúster califican más del 70% de las evidencias asignadas dentro del plazo establecido en ambos parciales. Esto sugiere que la mayoría de los docentes en este grupo cumplen con el tiempo esperado para la retroalimentación.

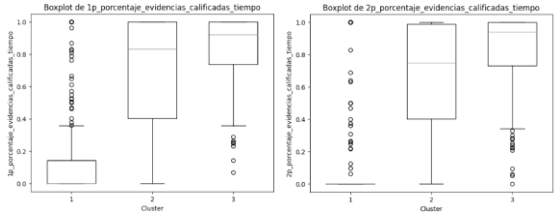



*Figura 11*. Diagrama de caja de la variable: porcentaje de evidencias calificadas a tiempo del primer y segundo parcial

*Nota.* Los datos de la variable porcentaje de evidencias entregadas a tiempo del primer y segundo parcial son muy similares en los grupos 2 y 3.

Otra variable que refuerza la categorización del clúster 3 como el de mejor desempeño es el total de actividades entregadas. Este clúster (el 3) tiene el valor más alto, lo que indica que, aunque asignan más actividades, los estudiantes responden entregándolas en tiempo y forma. Esto refleja una participación más alta por parte del estudiantado en la entrega de tareas (Figura 12).

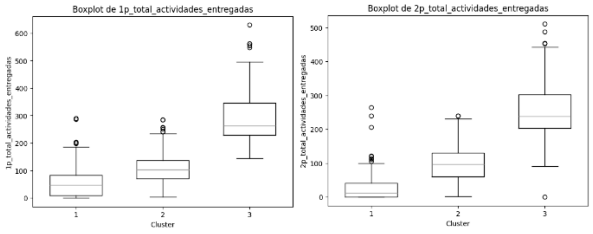



*Figura 12*. Valores por grupo de la variable: total de actividades entregadas primer y segundo parcial.

*Nota.* Se observa que el grupo 3 es el que tiene una mayor cantidad de evidencias entregadas tanto en el primer como en el segundo parcial, esto refleja participación por parte de los estudiantes.

Por último, las Figuras 13 y 14 presentan los resultados variables que, según la Figura 8, fueron identificadas como las menos relevantes para la clasificación: número de estudiantes por grupo y cantidad de correos enviados/recibidos. Contrario a lo que se podría esperar, estos factores no tienen una influencia significativa en la clasificación del desempeño docente.

En la Figura 13, se observa que el clúster 3 tiene en promedio 30 estudiantes por grupo, un valor similar al del clúster 1, mientras que el clúster 2 presenta un promedio de menos de 20 estudiantes por grupo. Esto indica que el tamaño del grupo no es un factor determinante en el desempeño docente.

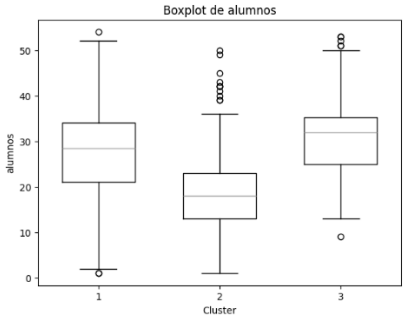



*Figura 13*. Número de estudiantes por grupo

*Nota.* Se observa que el grupo 1 y 3 tienen un número similar de estudiantes. Fuente: Elaboración propia (2024)

Por otra parte, la comunicación a través del correo electrónico tampoco parece ser una variable clave en la clasificación. Como se muestra en la Figura 14, los clústeres 2 y 3 tienen niveles similares de interacción mediante correo electrónico, aunque pertenecen a categorías diferenciadas. Sin embargo, el clúster 1, que representa el grupo con menor desempeño, muestra una notable reducción en la cantidad de correos enviados y recibidos. Esto sugiere que los docentes en este clúster tienen una menor interacción con los estudiantes dentro de la plataforma.

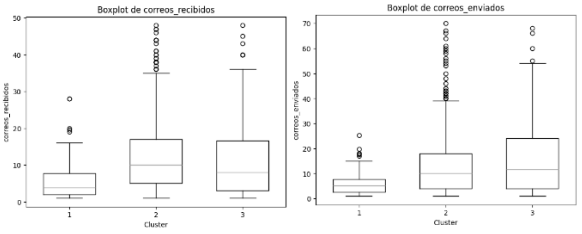



*Figura 14*. Correos enviados y correos recibidos clasificados por clúster

A partir del análisis de las interacciones registradas en la plataforma SEDUCA, se logró clasificar el desempeño docente en tres tipos diferenciados, a partir de la aplicación de técnicas de minería de datos. Este procedimiento permitió identificar patrones comunes en las prácticas de los docentes, dando lugar a una tipología que puede resumirse de la siguiente manera:

*Tabla 8*. Clasificación de docentes

| Clúster | Nombre del perfil | n=1 041 | % |
|---|---|---|---|
| **Con áreas de mejora** | Comunidades con docentes virtuales en desarrollo | 158 | 15% |
| **Desempeño Promedio** | Comunidades con docentes virtuales moderados | 687 | 66% |
| **Desempeño sobresaliente** | Comunidades con docentes líderes de interacción | 196 | 19% |

*Nota:* Elaboración propia (2024)

Los docentes del Perfil 1 – Área de mejora (15 %) se caracterizan por un bajo volumen de interacción: califican pocas actividades en ambos parciales y muestran rangos muy ajustados alrededor de valores bajos. Su porcentaje de evidencias calificadas a tiempo rara vez supera el 40 %, lo que indica que la retroalimentación suele ser tardía. Además, registran la menor cantidad de entregas estudiantiles, reflejando un compromiso limitado tanto de los profesores como de sus estudiantes con las tareas asignadas.

En contraste, los del Perfil 2 – Desempeño promedio (66 %) mantienen un nivel de actividad intermedio pero algo dispar: sus medianas de actividades calificadas se sitúan entre los dos extremos y presentan una dispersión notable, lo que revela heterogeneidad interna. Su porcentaje de evidencias a tiempo roza el 70 %, con docentes que tanto superan como quedan por debajo de ese umbral. Del mismo modo, el número de entregas de los alumnos es moderado, aunque con varios casos que se acercan al perfil más activo o, por el contrario, al más bajo.

Por último, los del Perfil 3 – Desempeño sobresaliente (19 %) destacan por una actividad intensiva y constante: califican un alto número de actividades en ambos parciales y mantienen porcentajes de evidencias calificadas a tiempo superiores al 90 %. La participación estudiantil es también muy elevada, con casi la totalidad de las actividades entregadas y escasos valores atípicos hacia abajo. Este grupo refleja un uso riguroso de la plataforma y un fuerte compromiso pedagógico.

Más allá de la segmentación metodológica en clústeres, lo relevante de este análisis es la posibilidad de caracterizar perfiles docentes distintos a partir de sus prácticas reales en la plataforma. Estos perfiles permiten proponer estrategias diferenciadas de mejora que respondan a las necesidades específicas de cada grupo, promoviendo así una enseñanza en línea más eficaz y equitativa.

# Conclusiones

La consolidación de la minería de datos educativos (EDM) como disciplina interdisciplinaria ha permitido pasar de simples análisis descriptivos a marcos metodológicos robustos que integran estadística, aprendizaje automático y preguntas pedagógicas. En este sentido, [24] definen EDM como un área emergente centrada en el desarrollo y aplicación de métodos para analizar conjuntos de datos generados en contextos educativos, destacando metodologías clave como clasificación, regresión, *clustering* y minería de reglas de asociación para extraer conocimiento accionable. De manera complementaria, [25] subrayan la imperativa necesidad de incorporar EDM en el ámbito formativo para orientar a docentes y autoridades en la mejora de la enseñanza virtual.

Al explorar entornos de educación en línea, [26] ilustran cómo EDM facilita la identificación de patrones de interacción de estudiantes con materiales y tutores, al tiempo que advierten sobre la frecuente omisión de detalles críticos en el procesamiento de los datos. El presente estudio responde a esa deficiencia con un preprocesamiento exhaustivo —selección y depuración de variables, imputación por K-NN y eliminación de outliers— que garantiza la fiabilidad de las etapas de *clustering* y clasificación.

La evaluación docente ha cobrado un papel central en EDM, [27] revisan sistemáticamente técnicas que incluyen lógica difusa, *clustering*, redes neuronales, árboles de decisión y *Random Forest* para predecir la calidad docente a partir de datos de estudiantes, evidenciando la eficacia de los clasificadores supervisados en este campo. Esta perspectiva conecta directamente con la fase predictiva de la presente investigación, en la que un modelo *Random Forest* asigna nuevos casos a los perfiles docentes identificados.

Desde el punto de vista metodológico, una revisión de la aplicación de CRISP-DM confirma que, a pesar de contar con más de dos décadas de vigencia, sigue siendo el estándar de facto en proyectos de minería de datos, aunque muchas investigaciones omiten la fase de despliegue [28]. El trabajo aquí descrito completa de forma coherente todas las etapas de CRISP-DM —desde la comprensión del negocio hasta la evaluación estadística y predictiva—, fortaleciendo así la replicabilidad y la utilidad práctica de sus resultados.

Este estudio confirma la capacidad de EDM para perfilar al docente virtual en SEDUCA y predecir su comportamiento, consolidando a la minería de datos como un verdadero método de investigación educativa que aporta conocimiento para ser considerado por las Instituciones de Educación Superior; por ejemplo, al mapear y predecir perfiles se puede:

- Anticipar necesidades de formación: identificar tempranamente a los docentes que requieren acompañamiento.
- Personalizar intervenciones: diseñar cursos y recursos ajustados a cada estilo de conducción virtual.
- Optimizar procesos de selección y contratación,

orientándolos hacia perfiles de alto desempeño y reforzando las áreas de mejora.

En el contexto institucional de la UAEMéx, esta clasificación puede servir como herramienta de gestión académica, permitiendo identificar a docentes que requieren apoyo o capacitación adicional, así como reconocer y sistematizar buenas prácticas. Su aplicación podría extenderse al diseño de programas de formación diferenciados según el perfil de desempeño, con enfoque en la mejora continua.

No obstante, el estudio presenta algunas limitaciones. La información analizada proviene exclusivamente de SEDUCA, por lo que no se consideran otras estrategias pedagógicas aplicadas fuera de la plataforma. La inclusión de datos cualitativos, como encuestas a estudiantes o análisis de contenido en foros, permitiría enriquecer la interpretación de los patrones detectados. Por último, la utilización de las variables: Carrera y Espacio académico podrían considerarse en trabajos posteriores con la intención de identificar la relación existente entre ellas.

## Agradecimientos

A la Universidad Autónoma del Estado de México (UAEMex) por proporcionar los datos correspondientes a los reportes de desempeño docente del semestre 2023B, los cuales resultaron fundamentales para el desarrollo del presente estudio.

El presente trabajo se desarrolló en el marco de los estudios del doctorado "Ambientes y Sistemas Educativos Multimodales" de la Universidad Nacional Rosario Castellanos para la obtención del grado.

## Referencias

[1] ANUIES, *Anuario Estadístico de Educación Superior*, 2024. [Online]. Available: https://www.anuies.mx/informacion-y-servicios/informacion-estadistica-de-educacion-superior/anuario-estadistico-de-educacion-superior/

[2] M. Brown, M. McCormack, J. Reeves, D. C. Brooks, and S. Grajek, *2020 EDUCAUSE Horizon Report: Teaching and Learning Edition*. Louisville, CO, USA: EDUCAUSE, 2020. [Online]. Available: https://library.educause.edu/-/media/files/library/2020/3/2020_horizon_report_pdf

[3] K. Pelletier, M. Brown, D. C. Brooks, M. McCormack, J. Reeves, and N. Arbino, *2021 EDUCAUSE Horizon Report: Teaching and Learning Edition*. Louisville, CO, USA: EDUCAUSE, 2021. [Online]. Available: https://library.educause.edu/resources/2021/4/2021-educause-horizon-report-teaching-and-learning-edition

[4] K. Pelletier, M. McCormack, J. Reeves, J. Robert, and N. Arbino, *2022 EDUCAUSE Horizon Report: Teaching and Learning Edition*. Louisville, CO, USA: EDUCAUSE, 2022. [Online]. Available: https://library.educause.edu/resources/2022/4/2022-educause-horizon-report-teaching-and-learning-edition

[5] K. Pelletier, J. Robert, N. Muscanell, M. McCormack, J. Reeves, N. Arbino, and S. Grajek, *2023 EDUCAUSE Horizon Report: Teaching and Learning Edition*. Louisville, CO, USA: EDUCAUSE, 2023. [Online]. Available: https://library.educause.edu/resources/2023/5/2023-educause-horizon-report-teaching-and-learning-edition

[6] S. Vargas Ávila and M. L. Cáceres Mesa, "Algunos preliminares sobre los factores socioeconómicos que determinan la aceptación de la educación superior a distancia," *Revista Metropolitana de Ciencias Aplicadas*, vol. 2, no. 1, 2019. [Online]. Available: https://remca.umet.edu.ec/index.php/REMCA/article/view/65

[7] UNESCO, *Educación 2030: Declaración de Incheon y Marco de Acción para la realización del Objetivo de Desarrollo Sostenible 4*. Paris, France: UNESCO, 2016. [Online]. Available: https://agenda2030lac.org/es/ods/4-educacion-de-calidad

[8] UAEMéx., "Modalidades educativas y sistemas de enseñanza," 2024. [Online]. Available: https://campusvirtual.uaemex.mx/modalidadeseducativasySE/

[9] UAEMéx., "Sistema de gestión del aprendizaje. Campus Virtual UAEMéx," 2024. [Online]. Available: https://campusvirtual.uaemex.mx/sistemadegestiondelaprendizaje/

[10] UAEMéx., "Universidatos," 2024. [Online]. Available: https://sdfi.uaemex.mx/estadisticas/universidatos/agendas-estad%C3%ADsticas.html

[11] M. K. Bhatia, "Data Analysis and Its Importance," *International Research Journal of Advanced Engineering and Science*, vol. 2, no. 1, pp. 166-168, 2017. [Online]. Available: https://irjaes.com/wp-content/uploads/2020/10/IRJAES-V2N1P58Y17.pdf

[12] Z. Liu, *Data Analysis for Scientific Research*. [Online]. Available: https://bae.k-state.edu/~zifeiliu/files/fac_zifeiliu/Zifeiliu/BAE815_Liu_06_Data%20Analysis%20for%20Scientific%20Research.pdf

[13] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery in Databases," *AI Magazine*, vol. 17, no. 3, pp. 37-54, 1996, doi:10.1609/aimag.v17i3.1230.

[14] C. Romero and S. Ventura, "Data Mining in Education," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 3, no. 1, pp. 12-27, 2013, doi:10.1002/widm.1075.

[15] IBM, "Conceptos básicos de ayuda de CRISP-DM," Aug. 17, 2021. [Online]. Available: https://www.ibm.com/docs/es/spss-modeler/saas?topic=dm-crisp-help-overview

[16] J. J. Espinosa-Zúñiga, "Aplicación de metodología CRISP-DM para segmentación geográfica de una base de datos pública," *Ingeniería, Investigación y Tecnología*, vol. 21, no. 1, Art. no. e00008, 2020, doi:10.22201/fi.25940732e.2020.21n1.008.

[17] UAEMéx., "Licenciaturas en línea – Campus Virtual UAEMéx," 2024. [Online]. Available: https://campusvirtual.uaemex.mx/licenciaturasenlinea/

[18] N. D. N. Nuñez Sánchez, "Análisis de algoritmos para la imputación de datos y modelos de predicción," Bachelor's thesis, 2023. [Online]. Available: http://hdl.handle.net/11349/39424

[19] S. Naeem, A. Ali, S. Anam, and M. Ahmed, "An Unsupervised Machine Learning Algorithms: Comprehensive Review," *International Journal of Computing and Digital Systems*, 2023, doi:10.12785/IJCDS/130172.

[20] N. Donges, "Random Forest: A Complete Guide for Machine Learning," *Built In*, Nov. 26, 2024. [Online].

[21] F. Pedregosa *et al*., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, 2011.

[22] A. Géron, *Aprende Machine Learning con Scikit-Learn, Keras y TensorFlow*, 3rd ed. Madrid, Spain: Anaya Multimedia, 2023.

[23] C. Pérez López, *Técnicas de análisis multivariante de datos*. México: Pearson Educación, 2008.

[24] I. Papadogiannis, M. Wallace, and G. Karountzou, "Educational Data Mining: A Foundational Overview," *Encyclopedia*, vol. 4, pp. 1644-1664, 2024, doi:10.3390/encyclopedia4040108.

[25] R. García Sánchez and J. E. Gil Mateos, "Minería de datos educacionales: descubrir tesoros ocultos durante el aprendizaje," *Revista Científica Ecociencia*, vol. 10, Special Issue, pp. 18-41, 2023, doi:10.21855/ecociencia.100.830.

[26] A. Rosado Gómez and A. Verjel Ibáñez, "Aplicación de la minería de datos en la educación en línea," *Revista Colombiana de Tecnologías de Avanzada*, vol. 1, no. 29, pp. 92-98, 2017, doi:10.24054/rcta.v1i29.194.

[27] R. Ordoñez-Ávila, N. Salgado Reyes, J. Meza, and S. Ventura, "Técnicas de minería de datos para predecir la evaluación docente en la educación superior: Una revisión sistemática de la literatura," *Heliyon*, vol. 9, no. 3, Art. no. e13939, 2023, doi:10.1016/j.heliyon.2023.e13939.

[28] C. Schröer, F. Kruse, and J. C. Marx Gómez, "A Systematic Literature Review on Applying the CRISP-DM Process Model," *Procedia Computer Science*, vol. 181, pp. 526-534, 2021, doi:10.1016/j.procs.2021.01.199.

[29] L. M. Regueras, M. J. Verdú, and J. P. de Castro, "A Rule-Based Expert System for Teachers' Certification in the Use of Learning Management Systems," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 7, pp. 75-81, 2022, doi:10.9781/ijimai.2022.11.004.

[30] F. Sciarrone and M. Temperini, "A Web-Based System to Support Teaching Analytics in a MOOC's Simulation Environment," in *Proc. International Conference on Information Visualisation (IV)*, 2020, pp. 491-495, doi:10.1109/IV51561.2020.00085.

[31] S. Gao, Y. Bai, X. Sun, and F. Zhao, "Feedback System for Teaching Activities Based on Deep Learning," in *Proc. 17th International Congress on Advanced Applied Informatics (IIAI-AAI-Winter)*, 2024, pp. 91-96, doi:10.1109/IIAI-AAI-Winter65925.2024.00026.

[32] K. Okoye, A. Arrona-Palacios, C. Camacho-Zuñiga, J. A. G. Achem, J. Escamilla, and S. Hosseini, "Towards Teaching Analytics: A Contextual Model for Analysis of Students' Evaluation of Teaching Through Text Mining and Machine Learning Classification," *Education and Information Technologies*, vol. 27, no. 3, pp. 3891-3933, 2022, doi:10.1007/s10639-021-10751-5.

*Información de contacto de los autores*

**Kárilyn Brunett Zarza**
Universidad Autónoma del Estado de México
Toluca, Estado de México. México
kbrunettz@uaemex.mx
https://orcid.org/0000-0002-6718-9771

**Rosa María Valdovinos Rosa**
Universidad Autónoma del Estado de México
Toluca, Estado de México. México
rvaldovinosr@uaemex.mx
https://orcid.org/0000-0001-9954-0653

**Patricia del Carmen Montaño-Reyes**
Universidad Autónoma del Estado de México
Metepec, Estado de México, México
pamr@uaemex.mx
https://orcid.org/0000-0001-8649-394X

**José Jacobo Gómez Quiróz**
Universidad Nacional Rosario Castellanos
Ciudad de México, México
jgq2405irc@gmail.com
https://orcid.org/0009-0006-9052-3965

**Kárilyn Brunett Zarza**
Egresada del Doctorado en Ambientes y Sistemas Educativos Multimodales de la Universidad Nacional Rosario Castellanos y Técnico Académico de Tiempo Completo en la Universidad Autónoma del Estado de México (UAEMéx)

**Rosa María Valdovinos Rosas**
Profesora en la Facultad de Ingeniería de la Universidad Autónoma del Estado de México (UAEMéx) y miembro del Sistema Nacional de Investigadores (SNII) Nivel II. Sus líneas de investigación incluyen algoritmos genéticos, redes neuronales artificiales, reconocimiento de patrones y minería de datos.

**Patricia del Carmen Montaño-Reyes**
Académica afiliada a la Universidad Autónoma del Estado de México (UAEMéx). Su trayectoria incluye importantes contribuciones en el ámbito de la educación, especialmente en métodos de enseñanza y aprendizaje en entornos virtuales.

**José Jacobo Gómez Quiróz**
Doctor en Ciencias Sociales por FLACSO Argentina. Actualmente se desempeña como docente de posgrado en la Universidad Nacional Rosario Castellanos, en la Universidad Autónoma Metropolitana y es asesor pedagógico en el Instituto Politécnico Nacional.