

Learning Profiles in Introductory Programming through Learning Analytics in a Gamified CodeCombat-Based Environment

Denis González Herrera¹  and Keleny Francel Zamora Díaz ² and Kendal Trejos Cubero ³

¹ Costa Rica University, San José, Costa Rica,
dennis.gonzalezherrera@ucr.ac.cr,
<https://orcid.org/0000-0003-2313-2332>

² Costa Rica University, San José, Costa Rica,
keleny.zamora@ucr.ac.cr

³ Costa Rica University, San José, Costa Rica,
kendal.trejos@ucr.ac.cr

Abstract. This study aims to identify and characterize learning profiles in introductory programming students through the analysis of interaction data in a gamified environment based on CodeCombat. A quantitative, exploratory-descriptive approach was adopted, using Learning Analytics techniques and unsupervised clustering. The analysis was conducted on data from 144 first-year university students, considering two key variables: average problem-solving time and completion rate, representing efficiency and persistence, respectively. K-means clustering was applied to normalized variables, and the optimal number of clusters ($k = 4$) was determined using the elbow method and silhouette coefficient. The results revealed four distinct learning profiles, reflecting heterogeneous learning trajectories. Differences between profiles were examined using descriptive statistics and the Kruskal-Wallis test, interpreted with caution due to the variables being part of the clustering model. The findings indicate that students exhibit diverse patterns of progress and temporal engagement, particularly in more complex content such as control structures and variables. Importantly, efficiency, measured as time, does not directly imply better learning outcomes, highlighting the need to consider multiple dimensions of student behavior. This study contributes to the integration of Learning Analytics and clustering techniques to support data-driven understanding of learning processes, offering insights for the design of adaptive instructional strategies in programming education.

Keywords: Programming Education, CodeCombat, Learning Analytics, Clustering.

Perfiles de aprendizaje en programación inicial mediante Learning Analytics en un entorno gamificado basado en CodeCombat

Resumen. El presente estudio tiene como objetivo identificar y caracterizar perfiles de aprendizaje en estudiantes de programación inicial mediante el análisis de datos de interacción en un entorno gamificado basado en CodeCombat. Se adoptó un enfoque cuantitativo de tipo exploratorio-descriptivo, utilizando técnicas de Learning Analytics y clustering no supervisado. El análisis se realizó sobre datos de 144 estudiantes universitarios de primer ingreso, considerando dos variables clave: tiempo promedio de resolución y tasa de finalización, asociadas a las dimensiones de eficiencia y persistencia. Se aplicó el algoritmo K-means sobre variables normalizadas, determinándose una solución óptima de cuatro clusters ($k = 4$) mediante el método del codo y el coeficiente de silueta. Los resultados evidencian la existencia de cuatro perfiles diferenciados, lo que refleja trayectorias de aprendizaje heterogéneas. Las diferencias entre perfiles fueron analizadas mediante estadística descriptiva y la prueba de Kruskal-Wallis, interpretada con carácter descriptivo. Los hallazgos indican que los estudiantes presentan patrones diversos de progreso y comportamiento temporal, especialmente en contenidos de mayor complejidad como estructuras de control y variables. Asimismo, se observa que la eficiencia, medida en términos de tiempo, no implica necesariamente un mejor aprendizaje. El estudio aporta evidencia para la comprensión del aprendizaje basado en datos y ofrece insumos para el diseño de estrategias educativas adaptativas en la enseñanza de la programación.

Palabras clave: Enseñanza de la programación, CodeCombat, Analítica de aprendizaje, Clustering.

1 Introducción

La enseñanza de la programación inicial constituye un desafío en la educación en informática debido a la complejidad de sus conceptos y a la diversidad de habilidades, experiencias previas y ritmos de aprendizaje de los estudiantes. La literatura ha señalado que los enfoques tradicionales, centrados en la transmisión de contenidos y la evaluación de resultados finales, limitan la comprensión de las dificultades que emergen durante el proceso de aprendizaje (Ramírez Prada et al., 2020). Asimismo, factores como la motivación, la comprensión conceptual y las estrategias pedagógicas influyen en el desempeño estudiantil, lo que demanda enfoques didácticos más adaptativos (Pineda & Reyes, 2025).

Esta problemática se intensifica en contextos heterogéneos, donde coexisten estudiantes con distintos niveles de conocimiento previo, lo que se asocia con bajo rendimiento y deserción en programas de tecnologías de la información (Carrillo Montoya & Chávez Baltodano, 2026). En este escenario, comprender la interacción de los estudiantes con los contenidos resulta clave para mejorar los procesos de enseñanza.

En los últimos años, Learning Analytics ha emergido como una alternativa para analizar datos educativos y comprender los procesos de aprendizaje. Estas técnicas permiten identificar patrones de comportamiento y apoyar la toma de decisiones pedagógicas basadas en evidencia (Papamitsiou & Economides, 2014; Sønderlund et al., 2018).

De manera complementaria, los entornos gamificados, como CodeCombat (CodeCombat Inc., 2026), han mostrado efectos positivos en la motivación y el aprendizaje de la programación (Kroustalli & Xinogalos, 2021), además de generar registros detallados de interacción. Sin embargo, el potencial de estos datos para identificar perfiles de aprendizaje sigue siendo limitado, particularmente mediante enfoques de segmentación.

En este contexto, el análisis de variables como el tiempo de resolución y la tasa de finalización permite aproximarse a dimensiones clave del aprendizaje, como la eficiencia y la persistencia. A partir de ello, surge la siguiente pregunta de investigación: ¿qué perfiles de aprendizaje pueden identificarse en estudiantes de programación inicial a partir de datos de interacción en un entorno gamificado, y cómo se diferencian en términos de eficiencia y persistencia?

El objetivo de este estudio es identificar y caracterizar perfiles de aprendizaje mediante técnicas de Learning Analytics y clustering no supervisado, utilizando datos de interacción en CodeCombat. La contribución principal radica en la construcción de perfiles empíricos de aprendizaje y su interpretación desde una perspectiva pedagógica, aportando evidencia para el diseño de estrategias educativas adaptativas.

Con base en estos elementos, se formulan las siguientes hipótesis:

1.1 Hipótesis general

Existen perfiles diferenciados de aprendizaje en estudiantes de programación inicial que pueden identificarse a partir del tiempo de resolución y la tasa de finalización de niveles.

H1a: Existen diferencias en el tiempo promedio de resolución entre los perfiles identificados.

H1b: Existen diferencias en la tasa de finalización de niveles entre los perfiles identificados.

Adicionalmente, se plantea como supuesto interpretativo que los perfiles reflejan distintos niveles de eficiencia y persistencia en la resolución de tareas de programación.

2 Marco teórico

2.1 Learning Analytics en educación

El campo de Learning Analytics se ha consolidado como una línea de investigación orientada al análisis sistemático de datos educativos con el propósito de comprender y optimizar los procesos de aprendizaje. De acuerdo con Almosallam & Ouertani (2013), Learning Analytics se define como la medición, recopilación, análisis y reporte de datos sobre los estudiantes y sus contextos, con el fin de mejorar tanto el aprendizaje como los entornos en los que este se desarrolla.

Diversos estudios han evidenciado el potencial de estas técnicas para transformar grandes volúmenes de datos en información significativa sobre el comportamiento y desempeño estudiantil. En particular, la revisión sistemática de Papamitsiou y Economides (2014) destaca que la analítica de aprendizaje permite identificar patrones de interacción que no son visibles mediante métodos tradicionales de evaluación. Asimismo, Sønderlund et al. (2018) señalan que las intervenciones basadas en Learning Analytics pueden mejorar significativamente el rendimiento académico y la retención en educación superior.

En el contexto de la enseñanza de la programación, Learning Analytics adquiere especial relevancia debido a la naturaleza procesual del aprendizaje, donde la resolución de problemas y la interacción continua con sistemas computacionales generan datos detallados del comportamiento del estudiante. En este sentido, Llanos Mosquera y Bucheli (2021) evidencian que la analítica de aprendizaje facilita la identificación de dificultades específicas y patrones de interacción, contribuyendo al diseño de estrategias pedagógicas más efectivas y adaptativas.

2.2 Enseñanza de programación y dificultades de aprendizaje

La enseñanza de la programación inicial representa un desafío significativo en la educación en informática, debido a la diversidad de factores que influyen en el aprendizaje, tales como las habilidades previas, los estilos cognitivos y la capacidad de abstracción de los estudiantes. Según Pineda & Reyes (2025), la motivación y la comprensión conceptual constituyen elementos clave en el aprendizaje de la programación, lo que exige la implementación de estrategias pedagógicas innovadoras. Por su parte, Ramírez Prada et al. (2020) señalan que los enfoques tradicionales suelen centrarse en la evaluación de resultados finales, lo que limita la comprensión de los procesos cognitivos involucrados en la resolución de problemas. Esta limitación dificulta la identificación de dificultades durante el aprendizaje y restringe la posibilidad de intervenir de manera oportuna.

Asimismo, estudios como el de Carrillo Montoya & Chávez Baltodano (2026) evidencian que la deserción y el bajo rendimiento en programas de tecnologías de la información están asociados a la incapacidad de detectar tempranamente patrones de comportamiento que reflejen dificultades en el aprendizaje. Estos hallazgos refuerzan la necesidad de adoptar enfoques basados en datos que permitan analizar el proceso de aprendizaje de manera más precisa y continua.

En este contexto, comprender las diferencias individuales en la forma en que los estudiantes enfrentan los problemas de programación se convierte en un elemento clave para mejorar la enseñanza, lo que requiere herramientas analíticas capaces de capturar dichas variaciones.

2.3 Gamificación y entornos de aprendizaje como CodeCombat

La gamificación ha emergido como una estrategia pedagógica relevante para incrementar la motivación, el compromiso y la participación activa de los estudiantes. En el ámbito de la enseñanza de la programación, los entornos gamificados permiten integrar elementos lúdicos con actividades de resolución de problemas, favoreciendo un aprendizaje más dinámico e interactivo.

Plataformas como CodeCombat (CodeCombat Inc., 2026) constituyen un ejemplo representativo de este enfoque, al combinar mecánicas de juego con el desarrollo de habilidades de programación. El estudio de Kroustalli y Xinogalos (2021) demuestra que el uso de juegos serios puede mejorar la comprensión de conceptos de programación en estudiantes de nivel secundario, al facilitar la experimentación y la retroalimentación inmediata.

De manera complementaria, Choi & Choi (2024) analiza el impacto de CodeCombat en la carga cognitiva, evidenciando que la estructura gamificada influye en la forma en que los estudiantes procesan la información. Asimismo, Li & Ye (2025) destacan que la interactividad percibida en este tipo de entornos contribuye al desarrollo del pensamiento computacional, mediado por competencias digitales y estilos de pensamiento. Más allá de sus beneficios pedagógicos, los entornos gamificados generan grandes volúmenes de datos de interacción, los cuales reflejan de manera detallada el comportamiento de los estudiantes durante el proceso de aprendizaje. Estos datos constituyen una fuente valiosa para la aplicación de técnicas de Learning Analytics, permitiendo analizar no solo los resultados, sino también las dinámicas de aprendizaje.

2.4 Identificación de perfiles de aprendizaje mediante clustering

El análisis de perfiles de aprendizaje mediante técnicas de clustering permite agrupar estudiantes según patrones de comportamiento similares, facilitando la identificación de perfiles diferenciados. Estas técnicas han demostrado ser útiles para segmentar estudiantes y apoyar el diseño de estrategias pedagógicas adaptativas (Bayazit et al., 2022; Pansri et al., 2024).

En particular, el algoritmo K-means permite agrupar observaciones en función de variables cuantitativas derivadas de datos de interacción. En el contexto de la programación, variables como el tiempo de resolución y la tasa de finalización permiten operacionalizar dimensiones clave del aprendizaje, como la eficiencia y la persistencia. En este sentido, la aplicación de clustering sobre datos de interacción en entornos gamificados constituye una aproximación pertinente para identificar perfiles de aprendizaje y analizar la diversidad de comportamientos en el proceso de aprendizaje.

3 Metodología

El presente estudio se desarrolló bajo un enfoque cuantitativo, con un alcance exploratorio-descriptivo, orientado a la identificación de perfiles de aprendizaje en estudiantes de programación inicial a partir del análisis de datos de interacción en un entorno gamificado. Desde la perspectiva de Learning Analytics, este tipo de aproximaciones permite analizar trazas de comportamiento para comprender procesos de aprendizaje y apoyar la toma de decisiones educativas (Almosallam & Ouertani, 2013; Papamitsiou & Economides, 2014).

3.1 Descripción de los datos

La muestra estuvo compuesta por 144 estudiantes universitarios de primer ingreso, con edades entre 17 y 24 años, inscritos entre los años 2021 y 2026 en un curso introductorio

de programación en la Universidad de Costa Rica, impartido en modalidad presencial y de carácter semestral.

La actividad en la plataforma CodeCombat (CodeCombat Inc., 2026) se realizó en un único momento al inicio del curso cada año, como parte de una evaluación diagnóstica, con el propósito de registrar el comportamiento inicial de los estudiantes sin intervención pedagógica previa. Se incluyeron en el análisis aquellos estudiantes que generaron registros completos de interacción en la plataforma, considerando como unidad de análisis al estudiante individual. Los estudiantes enfrentaron un total de 38 niveles, organizados en cuatro bloques temáticos: sintaxis básica (6 niveles), argumentos y propiedades (9 niveles), repeticiones (9 niveles) y variables (14 niveles). Para cada nivel se registró el tiempo de resolución al momento de su finalización. En aquellos casos en los que el nivel no fue completado, se asignó un valor de cero, el cual indica la no finalización de este. Por su parte, los valores mayores a cero corresponden al tiempo invertido en la resolución del nivel.

3.2 Preprocesamiento de datos

El preprocesamiento incluyó la verificación de consistencia del dataset y la definición de criterios para el tratamiento de los datos. En particular, los valores de tiempo iguales a cero fueron interpretados como niveles no completados y excluidos del cálculo de promedios, con el fin de evitar la subestimación del tiempo de resolución.

No obstante, esta decisión metodológica implica que el tiempo promedio de resolución se calcula únicamente sobre los niveles completados, lo que introduce una dependencia parcial con la persistencia del estudiante. En consecuencia, los valores de tiempo promedio deben interpretarse como una medida condicionada al avance, ya que estudiantes con menor tasa de finalización pueden presentar tiempos promedios relativamente bajos al haber resuelto únicamente niveles menos complejos.

3.3 Construcción de variables analíticas

A partir de los datos originales se definieron las siguientes variables:

- Niveles completados: número de niveles con tiempo mayor a cero.
- Tasa de finalización: proporción de niveles completados respecto al total de niveles: $\text{Tasa de finalización} = \text{niveles completados} / \text{Total de niveles (38)}$.
- Tiempo total: suma de los tiempos en niveles completados.
- Tiempo promedio de resolución: tiempo total dividido entre niveles completados.

Estas variables permitieron operacionalizar dos dimensiones centrales del estudio: persistencia, representada por la tasa de finalización y eficiencia, representada por el tiempo promedio de resolución. Es importante señalar que la eficiencia se interpreta como una medida relativa del ritmo de resolución y no como un indicador absoluto del aprendizaje.

3.4 Selección de variables para el clustering

Para la identificación de perfiles se utilizaron tiempo promedio de resolución y tasa de finalización. Esta decisión respondió a criterios de parsimonia y coherencia conceptual, dado que ambas variables representan directamente las dimensiones teóricas del

estudio. Variables como el tiempo total o los niveles completados no fueron incluidas en el clustering debido a su dependencia estructural con las variables seleccionadas.

3.5 Normalización de las variables

Dado que las variables analíticas presentan diferentes escalas, se aplicó un proceso de normalización mediante estandarización (Z-score), para garantizar la comparabilidad entre variables y evitar que aquellas con mayor magnitud influyan de manera desproporcionada en el proceso de clustering.

Este tipo de normalización es ampliamente utilizado en técnicas de segmentación, especialmente en algoritmos como K-means, donde la distancia entre observaciones es un factor determinante (Pansri et al., 2024).

3.6 Identificación de perfiles mediante clustering

Para la identificación de perfiles de aprendizaje se utilizó el algoritmo de clustering no supervisado K-means, aplicado sobre variables analíticas normalizadas (tiempo promedio de resolución y tasa de finalización). El uso de técnicas de clustering en Learning Analytics ha demostrado ser efectivo para agrupar estudiantes en función de patrones de comportamiento similares, facilitando la identificación de perfiles de aprendizaje (Bayazit et al., 2022). En particular, K-means se emplea ampliamente en contextos educativos para segmentar estudiantes según su desempeño y comportamiento (Pansri et al., 2024).

La determinación del número óptimo de clusters se realizó mediante el método del codo y el coeficiente de silueta, evaluando valores de k entre 2 y 8. Los resultados mostraron un punto de inflexión en $k = 4$ (WCSS = 84.64), a partir del cual la reducción de la inercia se vuelve marginal. De manera consistente, el coeficiente de silueta alcanzó su valor máximo en $k = 4$ (0.469), lo que indica una adecuada cohesión intra-cluster y separación inter-cluster. En conjunto, estos resultados justifican la selección de una solución de cuatro clusters, al ofrecer un equilibrio adecuado entre compacidad y diferenciación de los perfiles.

Con el fin de evaluar la estabilidad de la solución, el algoritmo K-means fue ejecutado múltiples veces ($n_{\text{init}} \geq 10$) con diferentes inicializaciones, observándose consistencia en la asignación de los estudiantes a los clusters y en la estructura general de los perfiles identificados. Este comportamiento sugiere que la segmentación es robusta y no depende de condiciones iniciales específicas.

Los perfiles identificados se interpretaron desde una perspectiva pedagógica, considerando las dimensiones de eficiencia y persistencia. Este enfoque permite caracterizar distintos tipos de estudiantes en función de su comportamiento de aprendizaje y analizar sus implicaciones en el contexto de la enseñanza de la programación. La interpretación de perfiles basada en datos constituye un elemento clave en Learning Analytics, ya que facilita la traducción de resultados cuantitativos en información útil para el diseño de estrategias educativas adaptativas (Bayazit et al., 2022).

3.7 Análisis estadístico

Los perfiles identificados fueron caracterizados mediante estadística descriptiva y comparados inferencialmente utilizando la prueba no paramétrica de Kruskal-Wallis,

adecuada para evaluar diferencias entre múltiples grupos independientes sin asumir normalidad. Esta prueba se aplicó sobre el tiempo promedio de resolución y tasa de finalización con un nivel de significancia de $\alpha = 0.05$.

Los análisis se realizaron en jamovi (The jamovi project, 2024) y R (R Core Team, 2024).

3.8 Análisis por contenido

Con el fin de profundizar en la comprensión del desempeño de los estudiantes, se realizó un análisis desagregado por contenido, considerando la estructura del entorno gamificado. Los niveles se agrupan en cuatro categorías: sintaxis básica, argumentos y propiedades, repeticiones y variables.

Para cada contenido se calcularon métricas específicas, incluyendo: tiempo promedio por contenido y tasa de finalización por contenido

Estas métricas permiten analizar diferencias en el comportamiento de los estudiantes según el tipo de contenido, facilitando la identificación de áreas de mayor dificultad o mayor dominio dentro del proceso de aprendizaje.

Cabe destacar que estas variables se utilizaron únicamente para análisis complementario e interpretación, y no como entrada del modelo de clustering, con el fin de preservar la interpretabilidad del modelo.

3.9 Alcance y limitaciones

El análisis se basa en indicadores globales derivados de los datos de interacción, lo que favorece la interpretabilidad de los resultados. Sin embargo, la ausencia de variables más específicas, como intentos de resolución, errores o secuencias temporales, limita la profundidad del análisis. Esta restricción es común en estudios basados en registros de plataformas educativas y no invalida el enfoque adoptado, sino que delimita su alcance.

3.10 Consideraciones éticas y manejo de datos

Los datos utilizados en este estudio fueron recolectados de manera anónima, garantizando en todo momento la confidencialidad de la información de los estudiantes. Los registros analizados corresponden exclusivamente a interacciones con la plataforma CodeCombat, sin incluir datos personales identificables.

El uso de los datos se realizó con fines estrictamente académicos y de investigación, respetando principios éticos asociados al análisis de información educativa. Asimismo, los datos fueron tratados de forma agregada, evitando la identificación individual de los participantes. Con el fin de garantizar la transparencia y reproducibilidad del estudio, los datos anonimizados y los materiales de análisis se encuentran disponibles en un repositorio abierto en Zenodo (González Herrera, 2026).

4 Resultados

4.1 Identificación de perfiles de aprendizaje

Se aplicó el algoritmo K-means sobre las variables estandarizadas de tiempo promedio de resolución (eficiencia) y tasa de finalización (persistencia), determinándose una solución de cuatro clusters mediante el método del codo y el coeficiente de silueta. La distribución de la muestra evidenció heterogeneidad en los patrones de aprendizaje: Cluster 0 (41.7%), Cluster 3 (31.9%), Cluster 1 (13.9%) y Cluster 2 (12.5%). Tabla 1.

Tabla 1. Estadísticos descriptivos de las variables analíticas por cluster.

Cluster	N	Tiempo promedio (M $\pm$ DE)	Tasa de finalización (M $\pm$ DE)	Tiempo total (M $\pm$ DE)
0	60	153.72 $\pm$ 95.88	0.59 $\pm$ 0.16	3461.75 $\pm$ 899.15
1	20	370.50 $\pm$ 220.34	0.76 $\pm$ 0.14	10602.05 $\pm$ 2106.49
2	18	180.47 $\pm$ 120.11	0.23 $\pm$ 0.12	1735.28 $\pm$ 1112.66
3	46	239.46 $\pm$ 140.92	0.70 $\pm$ 0.13	6301.35 $\pm$ 1059.94

Nota. M = media; DE = desviación estándar. El tiempo promedio y la tasa de finalización fueron utilizadas como variables de entrada para el clustering. El tiempo total se incluye únicamente con fines descriptivos.

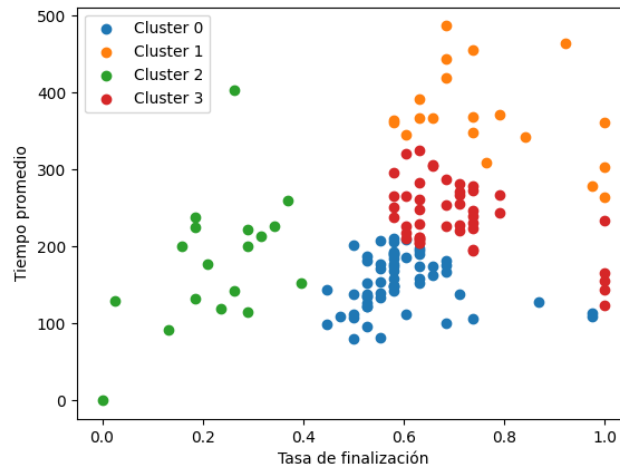


Figura 1. Distribución de clusters en función de eficiencia y persistencia.

La representación de la Figura 1 permite visualizar la separación entre los perfiles, destacando las diferencias en términos de eficiencia y persistencia.

Los resultados descriptivos mostraron diferenciación clara entre perfiles. El Cluster 0 presentó el menor tiempo promedio con una persistencia intermedia, reflejando un patrón de resolución ágil con avance moderado. El Cluster 1 combinó alta persistencia con el mayor tiempo promedio, indicando progreso sostenido con elevado costo

temporal. El Cluster 2 evidenció la menor tasa de finalización entre los grupos. Por su parte, el Cluster 3 mostró alta persistencia con tiempos intermedios, constituyendo un perfil de avance relativamente equilibrado.

El análisis del tiempo total reforzó esta diferenciación, evidenciando que los perfiles no solo difieren en el grado de avance, sino también en el esfuerzo temporal requerido. En conjunto, estos hallazgos confirman la existencia de trayectorias de aprendizaje heterogéneas dentro del entorno gamificado.

4.2 Diferencias estadísticas entre perfiles

Con el objetivo de caracterizar las diferencias entre los perfiles identificados, se aplicó la prueba no paramétrica de Kruskal-Wallis sobre las variables analíticas consideradas en el proceso de clustering (tiempo promedio de resolución y tasa de finalización).

Los resultados evidenciaron diferencias estadísticamente significativas entre clusters tanto en el tiempo promedio ($H = 93.16$, $p < .001$) como en la tasa de finalización ($H = 82.48$, $p < .001$), Tabla 2. No obstante, dado que estas variables constituyen la base del modelo de clustering, estos resultados deben interpretarse en términos descriptivos, como confirmación de la diferenciación interna de los perfiles, y no como una validación independiente de la segmentación.

La inspección de los valores descriptivos sugiere que la diferenciación entre perfiles no es homogénea en ambas dimensiones: algunos clusters se distinguen principalmente por eficiencia (Cluster 0 vs. Cluster 2), mientras que otros lo hacen por persistencia (Cluster 1 vs. Cluster 3). Esto indica que los estudiantes pueden alcanzar niveles similares de avance mediante estrategias temporales distintas, o bien presentar ritmos similares con progresos divergentes.

Tabla 2. Resultados de la prueba de Kruskal-Wallis para variables analíticas según cluster.

Variable	H	gl	p
Tiempo promedio	93.16	3	< .001
Tasa de finalización	82.48	3	< .001

Nota. Se utilizó la prueba no paramétrica de Kruskal-Wallis para comparar diferencias entre los cuatro clusters. Nivel de significancia $\alpha = 0.05$.

4.3 Caracterización de perfiles

La integración de los resultados descriptivos e inferenciales permitió identificar cuatro perfiles diferenciados. El Cluster 1 se caracterizó por una alta persistencia acompañada de baja eficiencia relativa, reflejando un elevado costo temporal para completar las actividades. El Cluster 0 presentó la mayor eficiencia, aunque con una persistencia intermedia. Por su parte, el Cluster 2 evidenció baja persistencia y escaso progreso global, mientras que el Cluster 3 mostró un patrón más equilibrado, combinando alta persistencia con una eficiencia moderada.

Esta tipología refleja combinaciones diferenciadas de avance y ritmo de resolución, evitando clasificaciones simplistas y manteniendo una interpretación basada en evidencia empírica.

4.4 Análisis por contenido

El análisis por bloques temáticos evidenció que las diferencias entre perfiles se mantienen y amplifican según el contenido (Tabla 3). El Cluster 1 presentó sistemáticamente los mayores tiempos en todos los bloques, confirmando su patrón de baja eficiencia relativa. El Cluster 2 mostró una disminución progresiva en la finalización, alcanzando ausencia de niveles completados en el bloque de variables, lo que indica dificultades críticas en contenidos de mayor complejidad.

El Cluster 0 mantuvo tiempos bajos, pero con avance moderado, mientras que el Cluster 3 evidenció un desempeño consistente y equilibrado. De manera transversal, los contenidos de repeticiones y variables emergen como los más demandantes, especialmente para perfiles de menor persistencia.

Tabla 3. Tiempo promedio de resolución y tasa de finalización por contenido según cluster.

Cluster	Sintaxis (tiempo / tasa)	Argumentos (tiempo / tasa)	Repeticiones (tiempo / tasa)	Variables (tiempo / tasa)
0	94.02 / 0.79	123.27 / 0.72	152.38 / 0.66	188.59 / 0.63
1	199.08 / 0.93	269.29 / 0.88	387.80 / 0.84	461.05 / 0.81
2	133.80 / 0.42	169.08 / 0.31	177.89 / 0.18	— / 0.00
3	134.01 / 0.88	181.11 / 0.81	250.60 / 0.78	294.78 / 0.75

Nota. El tiempo corresponde al promedio de resolución (en segundos) calculado únicamente sobre niveles completados. Cuando no se registran niveles completados en un contenido específico, el tiempo promedio se indica como no disponible (—). La tasa corresponde a la proporción de niveles completados dentro de cada contenido.

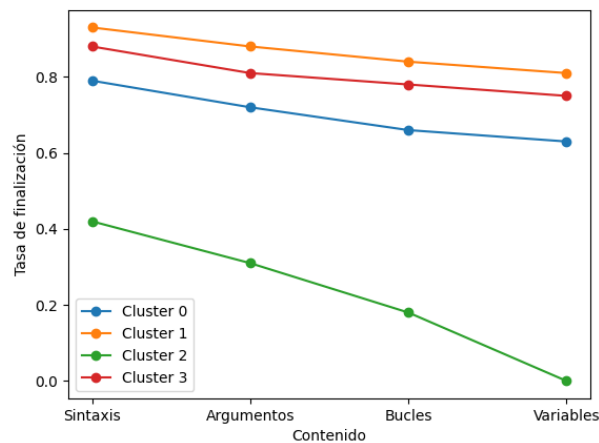


Figura 2. Tasa de finalización por contenido según cluster.

4.5 Síntesis de resultados

Los hallazgos confirman la existencia de cuatro perfiles, definidos por combinaciones específicas de persistencia y eficiencia. La segmentación evidencia que el aprendizaje de la programación no sigue trayectorias homogéneas y que las diferencias entre estudiantes se intensifican en contenidos de mayor complejidad. En conjunto, los resultados demuestran que el análisis combinado de tasa de finalización y tiempo de resolución ofrece indicios útiles para identificar perfiles potencialmente vulnerables y orientar estrategias pedagógicas adaptativas en entornos gamificados.

5 Discusión

Los resultados evidencian perfiles diferenciados de interacción y desempeño inicial en estudiantes de programación, identificados mediante datos de interacción en un entorno gamificado. Estos hallazgos respaldan la hipótesis general y confirman la utilidad de Learning Analytics para captar la heterogeneidad del comportamiento estudiantil (Papamitsiou & Economides, 2014). Las diferencias observadas en el tiempo promedio de resolución y la tasa de finalización permiten caracterizar distintos niveles de eficiencia y persistencia; sin embargo, al constituir la base del modelo de clustering, deben interpretarse como evidencia descriptiva de segmentación y no como una validación independiente de los perfiles.

La integración de eficiencia y persistencia constituye un aporte relevante al superar enfoques unidimensionales del desempeño. Aunque se adopta la denominación perfiles de aprendizaje en consonancia con la literatura, los datos provienen de una actividad diagnóstica inicial; por ello, los perfiles deben interpretarse principalmente como patrones de desempeño e interacción temprana, y no como evidencias concluyentes de aprendizaje. En este sentido, la eficiencia, medida mediante el tiempo de resolución, representa una dimensión relativa del comportamiento de los estudiantes y no un indicador directo de aprendizaje.

Desde una perspectiva pedagógica, los perfiles sugieren necesidades diferenciadas de apoyo. La baja persistencia refleja interrupción temprana del progreso, mientras que la combinación de alta persistencia y menor eficiencia relativa puede asociarse con estrategias menos optimizadas y la necesidad de fortalecer habilidades metacognitivas (Pansri et al., 2024). Asimismo, el perfil equilibrado muestra un avance sostenido con un costo temporal moderado. Aunque estos hallazgos sugieren el potencial de los perfiles para apoyar la detección temprana de estudiantes que requieran acompañamiento, su relación con resultados académicos posteriores no fue analizada. Por ello, futuras investigaciones deberían evaluar su capacidad predictiva respecto a aprobación, reprobación o abandono.

El análisis por contenido mostró que las diferencias entre perfiles se intensifican en temas de mayor complejidad. En particular, las variables emergen como una barrera crítica para los estudiantes con menor progreso (Pineda & Reyes, 2025), mientras que el aumento del tiempo en repeticiones sugiere una mayor carga cognitiva (Choi & Choi, 2024). Metodológicamente, los resultados respaldan el uso de clustering para identificar perfiles diferenciados a partir de datos de interacción. El uso de un conjunto reducido de variables favorece la interpretabilidad, aunque no captura completamente la complejidad del aprendizaje. La ausencia de indicadores como intentos, errores o

secuencias temporales, así como el uso de una única plataforma, limitan el alcance y la generalización de los resultados. A pesar de estas limitaciones, los hallazgos aportan evidencia sobre el potencial de los entornos gamificados como fuentes de datos y sobre el uso combinado de Learning Analytics y clustering para comprender patrones de interacción, explorar posibles trayectorias de aprendizaje y orientar estrategias educativas adaptativas.

6 Conclusiones

El presente estudio tuvo como objetivo identificar y caracterizar perfiles de aprendizaje en estudiantes de programación inicial a partir del análisis de datos de interacción en un entorno gamificado. Mediante la aplicación de técnicas de Learning Analytics y clustering no supervisado, fue posible modelar el comportamiento de los estudiantes en función de las dimensiones de eficiencia y persistencia.

Los resultados evidencian la existencia de cuatro perfiles diferenciados de aprendizaje, lo que sugiere que los estudiantes no siguen trayectorias homogéneas en su proceso formativo. La segmentación permitió identificar combinaciones específicas de progreso y tiempo de resolución, lo que respalda el uso de variables derivadas de la interacción como base para el análisis del comportamiento de aprendizaje.

Desde una perspectiva empírica, las diferencias observadas entre los perfiles en el tiempo promedio de resolución y la tasa de finalización son consistentes con la estructura del modelo de clustering. Asimismo, el análisis por contenido mostró que las dificultades no se distribuyen de manera uniforme, concentrándose en temas de mayor complejidad conceptual, como estructuras de control y variables.

En términos metodológicos, el estudio muestra que la utilización de un conjunto reducido de variables, tiempo promedio de resolución y tasa de finalización, constituye una estrategia parsimoniosa y útil para identificar perfiles de aprendizaje a partir de datos de interacción. Este enfoque favorece la interpretabilidad de los resultados y facilita su aplicación en contextos educativos reales.

Desde una perspectiva pedagógica, los hallazgos aportan evidencia relevante para comprender la heterogeneidad del aprendizaje en programación inicial y pueden servir como base para el diseño de estrategias de apoyo diferenciadas. En particular, la identificación de perfiles con baja persistencia o con altos tiempos de resolución permite orientar futuras intervenciones pedagógicas, aunque el presente estudio no evalúa directamente su efectividad.

Finalmente, como línea de investigación futura, se sugiere incorporar variables más específicas del proceso de aprendizaje, como el número de intentos, los errores cometidos y las secuencias temporales de interacción, así como explorar otros métodos de clustering y validar los perfiles en diferentes contextos educativos. Estas líneas permitirían profundizar en la comprensión de los perfiles de aprendizaje y ampliar el alcance de los resultados.

Referencias

- Almosallam, E. A., & Ouertani, H. C. (2013). Learning analytics: Definitions, applications and related fields. En T. Herawan, M. Deris, & J. Abawajy (Eds.), *Proceedings of the First*

- International Conference on Advanced Data and Information Engineering (DaEng-2013)* (Lecture Notes in Electrical Engineering, Vol. 285). Springer. https://doi.org/10.1007/978-981-4585-18-7_81
- Bayazit, A., Ilgaz, H., Gönüllü, İ., & Erden, S. (2022). Profiling students via clustering in a flipped clinical skills course using learning analytics. *Medical Teacher*, 45(7), 724–731. <https://doi.org/10.1080/0142159X.2022.2152663>
- Carrillo Montoya, X., & Chávez Baltodano, F. (2026). Deserción universitaria en la carrera de tecnologías de la información: Caso de estudio para la identificación de factores y tendencias en la Universidad Técnica Nacional, sede Guanacaste. *Yulök Revista de Innovación Académica* <https://doi.org/10.47633/9zwrs832>
- CodeCombat Inc. (2026). *CodeCombat* [Plataforma educativa]. <https://codecombat.com>
- Choi, W. C., & Choi, I. C. (2024). Investigating the effect of the serious game CodeCombat on cognitive load in Python programming education. En *2024 IEEE World Engineering Education Conference (EDUNINE)* (pp. 1–6). IEEE. <https://doi.org/10.1109/edunine60625.2024.10500551>
- Gonzalez Herrera, D. (2026). Dataset, Learning Profiles in Introductory Programming through Learning Analytics in a Gamified CodeCombat-Based Environment [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.19376172>
- Kroustalli, C., & Xinogalos, S. (2021). Studying the effects of teaching programming to lower secondary school students with a serious game: A case study with Python and CodeCombat. *Education and Information Technologies*, 26, 6069–6095. <https://doi.org/10.1007/s10639-021-10596-y>
- Li, L., & Ye, Z. (2025). Impact of perceived interactivity in CodeCombat on vocational students' computational thinking: Mediating effects of digital competence and thinking styles. *European Journal of Education*. <https://doi.org/10.1111/ejed.70341>
- Llanos Mosquera, J. M., & Bucheli Guerrero, V. A. (2021). Analítica de aprendizaje como estrategia de apoyo al aula invertida en cursos de programación: Una revisión sistemática de literatura. *Revista Investigación e Innovación en Ingenierías*, 9(1), 114–135. <https://doi.org/10.17081/invinno.9.1.4464>
- Pineda, J. R. L., & Reyes, R. N. L. (2025). Estrategias pedagógicas innovadoras para mejorar el aprendizaje y la motivación en la enseñanza de la programación de computadores en estudiantes del ciclo básico de ingeniería. *Encuentro Internacional de Educación en Ingeniería*. <https://acofipapers.org/index.php/eiei/article/download/4377/2846>
- Pansri, B., Sharma, S., Timilsina, S., Choonhapong, W., Kurashige, K., Watanabe, S., & Sato, K. (2024). Understanding student learning behavior: Integrating the self-regulated learning approach and K-means clustering. *Education Sciences*, 14(12), 1291. <https://doi.org/10.3390/educsci14121291>
- Papamitsiou, Z. K., & Economides, A. A. (2014). Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. *Journal of Educational Technology & Society*, 17(4), 49–64.
- R Core Team. (2024). *R: A language and environment for statistical computing (Version 4.4)* [Computer software]. <https://www.r-project.org/>
- Ramírez Prada, P. C., Ortiz Beltrán, A. O., & Lobo Quintero, R. A. (2020). Diagnóstico y reestructuración pedagógica en la enseñanza de programación de computadores en ingeniería de sistemas. *Revista Interamericana de Investigación, Educación y Pedagogía*, 13(2), 239–264. <https://doi.org/10.15332/25005421.6000>
- Sønderlund, A. L., Hughes, E. S., & Smith, J. (2018). The efficacy of learning analytics interventions in higher education: A systematic review. *British Journal of Educational Technology*, 50(5), 2594–2618. <https://doi.org/10.1111/bjet.12720>
- The jamovi project. (2024). *jamovi (Version 2.5)* [Computer software]. <https://www.jamovi.org>