Identificación de patrones entre hábitos estudiantiles y rendimiento académico mediante análisis multivariado

Resumen

El presente trabajo de investigación tiene como objetivo aplicar técnicas de análisis multivariado para identificar los factores que inciden en el desempeño académico de los estudiantes.

A partir de un conjunto de datos sintéticos que comprende 1.000 registros de alumnos y más de quince variables —entre ellas, horas de estudio, patrones de sueño, uso de redes sociales, calidad de la dieta, estado de salud mental y calificaciones finales— se busca explorar cómo los hábitos de vida influyen en el rendimiento académico.

Para ello, se implementaron herramientas estadísticas exploratorias y de reducción de la dimensionalidad, tales como el Análisis Factorial, el Análisis de Componentes Principales (PCA), el Análisis de Conglomerados (o Clúster) y el Análisis de Varianza (ANOVA). Estas técnicas permitieron identificar patrones comunes de comportamiento y asociaciones significativas entre las variables observadas.

Los resultados obtenidos evidencian una relación estadísticamente significativa entre el rendimiento académico —medido a través de las calificaciones en los exámenes finales— y determinados hábitos individuales, destacándose variables como el tiempo dedicado al estudio, el uso de redes sociales, las horas de sueño y el cuidado físico y mental.

Introducción

En la actualidad, el rendimiento académico de los estudiantes se ve influenciado por múltiples factores vinculados a sus hábitos cotidianos. El modo en que los jóvenes gestionan su tiempo, descansan, se alimentan o utilizan las redes sociales puede repercutir directamente en su desempeño educativo. Surge así la pregunta: ¿en qué medida actividades como mirar series, dormir poco o usar TikTok impactan en las calificaciones?

El desarrollo de hábitos saludables y disciplinados promueve el fortalecimiento de las capacidades cognitivas, mejorando la comprensión de los contenidos y, en consecuencia, el desempeño académico. En este marco, la presente investigación tiene como propósito analizar la influencia de los hábitos de vida y estudio sobre el rendimiento académico mediante el uso de un conjunto de datos que simula la conducta de 1.000 estudiantes.

El análisis contempla variables vinculadas con la disciplina académica —como el tiempo de estudio o la asistencia a clases— y otras relacionadas con el bienestar personal —como el sueño, la alimentación y la salud mental—, contrastándolas con las puntuaciones obtenidas en los exámenes finales.

Para cumplir con dicho objetivo, se aplicaron diversas técnicas de análisis multivariado:

- Análisis Factorial, para explicar las correlaciones entre variables observadas mediante factores latentes no directamente medibles.
- Análisis de Componentes Principales (PCA), con el propósito de reducir la dimensionalidad del conjunto de datos conservando la mayor parte de la información original.
- Análisis de Conglomerados (Clúster), para agrupar a los individuos según similitudes en sus características o comportamientos.
- Análisis de Varianza (ANOVA), con el fin de comparar las medias de diferentes grupos y evaluar la existencia de diferencias significativas entre ellos.

Estas herramientas permiten no solo describir la realidad educativa desde una perspectiva estadística, sino también detectar patrones de conducta estudiantil asociados a un mayor o menor rendimiento académico.

Objetivo del análisis

El objetivo de la investigación es determinar la relación existente entre los hábitos de estudio y el rendimiento académico sobre una población de 1.000 estudiantes, a fin de proponer alternativas que favorezcan el incremento del rendimiento académico y personal del estudiante.

Alcance

El alcance del estudio implica analizar los comportamientos y características vinculadas a los hábitos de los estudiantes, y cómo éstos influyen en los resultados académicos.

Para ello, se emplea una base de datos sintética que simula información relevante sobre variables tales como el tiempo de estudio, el sueño, el uso de redes sociales, la calidad de la dieta, la salud mental percibida, entre otras, todas en relación con la calificación obtenida en el examen final.

El enfoque metodológico adoptado es de carácter exploratorio y descriptivo. En este sentido, el estudio no busca establecer relaciones de causalidad entre las variables, sino identificar patrones, tendencias y asociaciones potenciales presentes en los datos.

Para tal fin, se utilizan técnicas de análisis multivariado, que permiten observar la estructura subyacente del conjunto de datos y determinar cómo se agrupan o relacionan las variables entre sí.

Hipótesis

La hipótesis principal del trabajo plantea que el rendimiento académico está influenciado significativamente por los hábitos estudiantiles.

Para su comprobación, se recurre a la aplicación de métodos de análisis multivariado, los cuales permiten reducir la dimensionalidad del conjunto de datos, sintetizando la información original en componentes principales.

Posteriormente, se determina si existe correlación significativa entre las variables y se clasifican los individuos observados en grupos homogéneos en función de las variables identificadas como más relevantes.

Marco teórico

Antes de comenzar con el desarrollo, resulta necesario introducir ciertos conceptos para contextualizar la temática.

El rendimiento académico se refiere al nivel de éxito que un estudiante alcanza en sus estudios. Este se evalúa mediante diversos instrumentos —principalmente exámenes— que reflejan su grado de aprendizaje y la adquisición de conocimientos. Constituye un indicador fundamental del progreso educativo, susceptible de verse afectado por factores personales, sociales y ambientales, tales como la salud, el entorno familiar o las estrategias de estudio.

Por su parte, los hábitos de estudio comprenden las rutinas, técnicas y estrategias que los estudiantes adoptan con el propósito de optimizar su proceso de aprendizaje y facilitar la asimilación de conocimientos. Ambos conceptos se encuentran estrechamente vinculados: buenos hábitos de estudio, como la planificación del tiempo, la creación de un ambiente adecuado o el uso de técnicas de repaso y memorización, pueden mejorar significativamente el desempeño académico. A su vez, un alto rendimiento académico tiende a reforzar la motivación y la constancia del estudiante para mantener o perfeccionar dichos hábitos, configurando un círculo virtuoso de mejora continua.

Análisis metodológico

1. *Análisis factorial*

El Análisis Factorial es una técnica estadística de reducción de la dimensionalidad de los datos originales, basada en un modelo formal de generación de la muestra. Su aplicación permite explicar las covarianzas o correlaciones entre un conjunto de variables observadas mediante un número reducido de factores latentes, los cuales resumen la estructura subyacente de los datos.

El objetivo es obtener factores que sean estadísticamente consistentes y fácilmente interpretables, reduciendo la cantidad de variables sin perder información relevante.

2. *Análisis de componentes principales*

El Análisis de Componentes Principales (PCA, por sus siglas en inglés) es un método multivariante que tiene por finalidad reducir la dimensionalidad de un conjunto de datos con múltiples variables correlacionadas.

Este procedimiento transforma las variables originales en un nuevo conjunto de componentes principales, no correlacionados entre sí, que retienen la mayor proporción de la varianza total de la información original.

De esta manera, el PCA permite sintetizar los datos en un número menor de dimensiones, facilitando su interpretación y análisis posterior.

3. *Análisis de clúster o conglomerados*

El Análisis de Clúster (o Análisis de Conglomerados) es una técnica estadística multivariante de clasificación no supervisada, que busca agrupar individuos o variables en categorías o conglomerados según su grado de similitud.

Su objetivo principal es formar grupos internamente homogéneos (es decir, con elementos similares entre sí) y externamente heterogéneos (diferenciados de otros grupos), en función de las variables observadas. Esta técnica resulta útil para identificar perfiles o patrones comunes dentro del conjunto de estudiantes analizados.

4. *Análisis de la varianza simple (ANOVA)*

El Análisis de la Varianza (ANOVA) es una técnica estadística que permite evaluar la relación entre una variable dependiente (endógena) y dos o más variables independientes (exógenas). En el presente estudio, su propósito es determinar si existen diferencias estadísticamente significativas entre las medias de distintos grupos definidos a partir de los hábitos estudiantiles.

El modelo ANOVA permite así contrastar la hipótesis de igualdad de medias y medir la significación estadística de las diferencias observadas en el rendimiento académico, atribuibles a los distintos grupos de variables independientes.

Obtención de los datos

El análisis se llevó a cabo a partir de un conjunto de datos obtenido del sitio web Kaggle, bajo el título "Student Habits vs Academic Performance" (Hábitos Estudiantiles vs. Rendimiento Académico). El dataset, de carácter sintético pero realista, contiene 1.000 registros correspondientes a estudiantes, en los cuales cada fila representa un individuo y cada columna refleja una característica o hábito cotidiano asociado a su desempeño académico.

Las variables de tipo categórico incluidas en el conjunto de datos son las siguientes:

- Identificación del estudiante: La muestra contempla 1.000 estudiantes.
- Género: Se cuestionaron a 481 estudiantes clasificados como "Femenino", 477 de género "Masculino" y 42 estudiantes percibidos como "Otro".
- Trabajo medio tiempo, respondiendo como "Si / No". La mayoría de los estudiantes no posee un trabajo a medio tiempo.
- Calidad de la dieta, caracterizado como "Regular, Bueno, Malo". En su mayoría definen a su dieta como "Regular".
- Nivel de educación parental, si los padres accedieron a estudios "Secundarios, Universitarios, Posgrados o Ninguno".
- Calidad de internet, poseen acceso a internet "Promedio, Bueno, Malo". En general poseen una buena calidad a internet.
- Participación extracurricular, si forman parte de actividades extracurriculares respondido como "Si / No". Siendo que la mayor parte de los estudiantes no participa en actividades fuera del curso académico.

Mientras que las variables numéricas o continuas son:

- La edad de los estudiantes que participan en la muestra es de 17 a 24 años.
- Horas de estudio por día, siendo el promedio de 3.5 horas de estudio por día.
- Horas que usa redes sociales, en promedio pasan 2.5 horas al día utilizando las redes sociales.

- Horas que pasa viendo Netflix, con un promedio de 1.8 horas al día.
- Porcentaje de asistencia, en promedio poseen un 85% de asistencia a clases.
- Horas de sueño, los estudiantes duermen 6.5 horas cada día.
- Frecuencia de ejercicio, medido en días a la semana que realizan ejercicio se obtiene una media de 3 días.
- Calificación de la salud mental, con una puntuación de 5 promedio.
- Puntuación de examen, en promedio los estudiantes obtienen un 70% de aprobación en los exámenes.

Este conjunto de datos permite contar con una visión integral de los hábitos y condiciones de vida de los estudiantes, posibilitando la aplicación de técnicas de análisis multivariado orientadas a explorar relaciones entre conductas, bienestar y rendimiento académico.

Desarrollo

1. *Importación de la base de datos*

Se importó un archivo en formato XLSX extraído del sitio web Kaggle, titulado "Student Habits vs Academic Performance" (Hábitos Estudiantiles vs. Rendimiento Académico). El dataset contiene una muestra de 1.000 registros de estudiantes, con variables relevantes que describen cómo los hábitos cotidianos influyen en las calificaciones obtenidas en los exámenes finales.

2. Análisis previo de los datos

En una primera instancia, se realizó la inspección estructural del conjunto de datos, identificando el tipo de formato de las variables incluidas.
Se distinguieron las variables categóricas —Género, trabajo a medio tiempo, calidad de la dieta, nivel educativo parental, calidad del servicio de internet y participación extracurricular— y las variables numéricas o continuas —Edad, horas de estudio por día, horas de uso de redes sociales, horas de visualización de Netflix, porcentaje de asistencia, horas de sueño, frecuencia de ejercicio, calificación de salud mental y puntuación de examen.

A continuación, se calcularon los estadísticos descriptivos básicos. Para las variables numéricas: media, desviación estándar, valores mínimos y máximos, y percentiles (25%, 50% y 75%). Para las variables categóricas: valores únicos, moda y frecuencia relativa.

Durante esta exploración se detectaron variables irrelevantes o con datos incompletos. En consecuencia, se eliminaron:

- Nivel de educación parental, por presentar valores nulos; y
- Identificación del estudiante, por no aportar información analíticamente significativa.

3. *Aplicación de los métodos de análisis multivariados*

a. Análisis factorial

Se aplicó el Análisis Factorial Exploratorio (AFE) con el objetivo de identificar relaciones subyacentes entre las variables. Para ello, se plantearon las siguientes hipótesis:

- Hipótesis nula (H_0): no existe correlación significativa entre las variables numéricas o continuas.
- Hipótesis alternativa (H_1): existe correlación significativa entre las variables numéricas o continuas.

Se calculó la matriz de correlación de Pearson, la cual permitió examinar el grado de relación lineal entre las variables analizadas. Posteriormente, se aplicó el test de esfericidad de Bartlett, que arrojó un estadístico Chi-cuadrado = 2314.71 y un valor p = 0.0000, indicando la existencia de correlación significativa entre los atributos numéricos. De este modo, se rechaza la hipótesis nula, justificando la aplicación del análisis factorial.

La matriz de correlación (de dimensión 9x9) se descompuso en valores singulares (autovalores), los cuales evalúan la proporción de varianza explicada por cada factor. La suma total de los autovalores fue igual a 9, coincidente con el número de variables originales. Mediante el método de Máxima Verosimilitud, se estimó la estructura factorial y se determinó que la cantidad óptima de factores debía ser menor o igual a 5, con el fin de explicar adecuadamente las correlaciones observadas.

En consecuencia, se aplicó el método de factores principales, utilizando las nueve variables numéricas y extrayendo cinco componentes (PA1, PA2, PA3, PA4 y PA5). Las cargas factoriales obtenidas se interpretaron de la siguiente manera:

- PA1: Rendimiento académico (altas cargas en horas de estudio y puntuación de exámenes).
- PA2: Asistencia o participación (alta carga de la variable porcentaje de asistencia a clases).
- PA3: Sueño y descanso (alta carga de la variable horas de sueño).
- PA4: Actividad física (alta carga de la frecuencia diaria de ejercicio).
- PA5: Salud mental (alta carga de la variable calificación de salud mental).

Posteriormente, se calcularon las comunalidades y unicidades para evaluar qué proporción de la varianza individual de cada variable es explicada por los factores extraídos. Los resultados indicaron que la mayoría de las variables presentan comunalidades superiores a 0.6, lo que demuestra que el modelo factorial logra captar una proporción considerable de la varianza total. Se destacaron especialmente:

- Puntuación de examen (0.97) y Horas de estudio al día (0.89), con comunalidades muy elevadas, lo cual indica que están fuertemente representadas por los factores identificados.
- En contraste, Horas de Netflix (0.37) y Horas de redes sociales (0.53) mostraron comunalidades más bajas, reflejando que su comportamiento no es completamente explicado por los factores comunes, posiblemente por la influencia de variables externas no contempladas en el modelo.

En conjunto, los resultados respaldan la validez y robustez del modelo factorial, evidenciando que los factores extraídos capturan patrones latentes relevantes entre las variables analizadas.

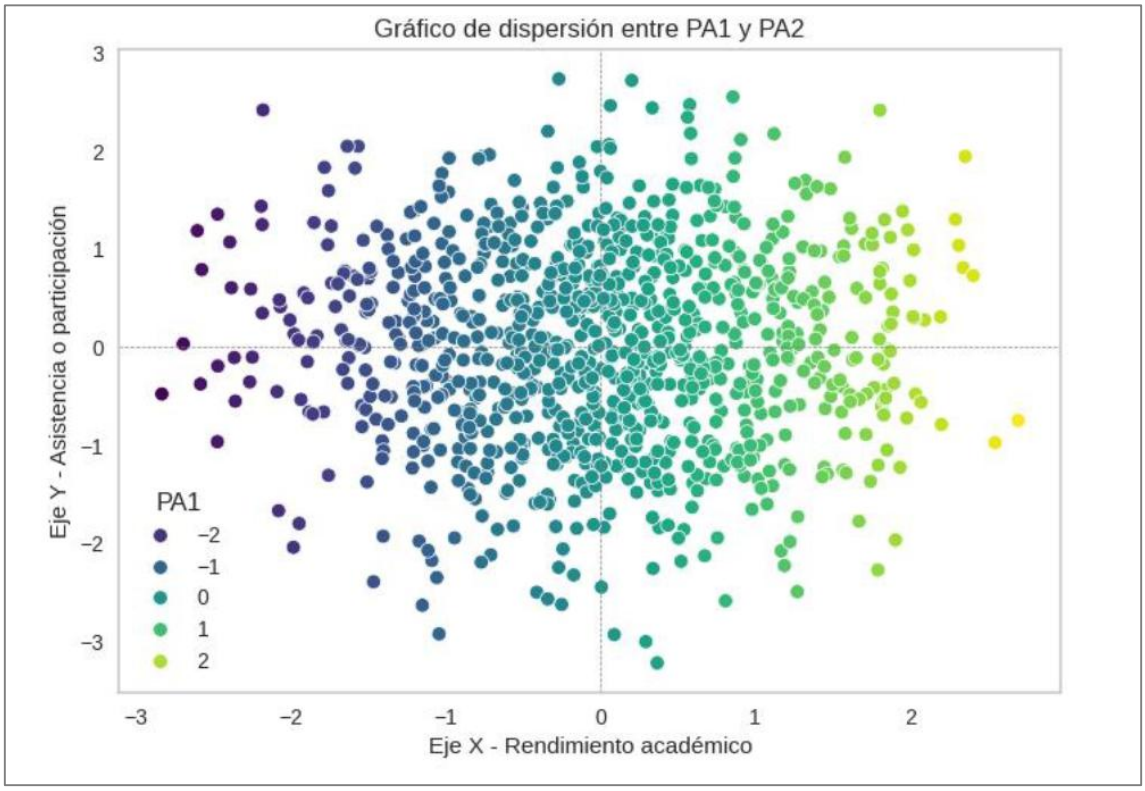

Gráfico 1: Dispersión entre Rendimiento académico y Asistencia o participación.

En el gráfico correspondiente, cada punto representa una observación individual. El factor PA1 refleja el rendimiento académico, con altas cargas en las variables Horas de estudio y Puntuación de exámenes, lo que evidencia una relación directa entre el tiempo dedicado al estudio y los resultados obtenidos. Por su parte, el factor PA2 está asociado con la asistencia o participación en clase.

En consecuencia, se observa que los estudiantes con mayor dedicación al estudio y mayor participación en clases obtienen un rendimiento académico superior, mientras que aquellos con menor asistencia y dedicación tienden a alcanzar calificaciones más bajas.

De forma complementaria, el análisis del segundo gráfico, donde PA4 (Actividad física) se representa en el eje X y PA5 (Salud mental) en el eje Y, muestra la influencia positiva de la actividad física sobre el bienestar psicológico. Los estudiantes que practican ejercicio regularmente presentan menores niveles de estrés y mejor estado de ánimo, lo cual repercute favorablemente en su rendimiento académico.

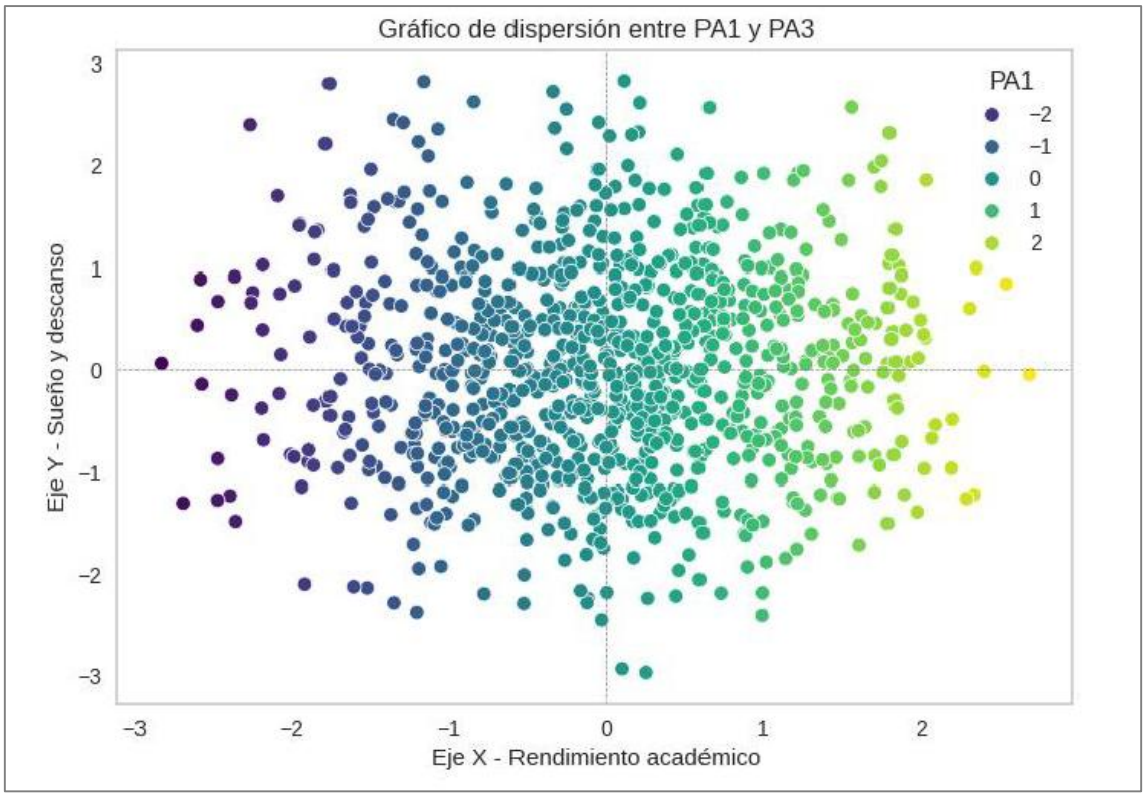

Gráfico 2: Dispersión entre Salud mental y Actividad física.

b. Análisis de componentes principales (PCA)

Una vez determinadas las correlaciones entre las variables, se procedió con la aplicación del método de Análisis de Componentes Principales (PCA, por sus siglas en inglés), con el propósito de reducir la dimensionalidad del conjunto de datos, procurando conservar la mayor proporción posible de varianza y, por ende, la información más relevante para el análisis.

El procedimiento se inicia con el cálculo de las estadísticas básicas y las matrices de varianza-covarianza y de correlaciones, que describen las relaciones entre las variables del dataset.

Dado que las variables originales presentan diferentes escalas y rangos de medición, se optó por utilizar la matriz de correlación en lugar de la de varianzas, con el fin de estandarizar las variables (media cero y varianza uno). Esta estandarización permite una evaluación equilibrada de las contribuciones de cada variable y facilita la interpretación de los componentes extraídos.

El análisis determinó que, de un total de nueve componentes originales, la dimensionalidad puede reducirse a cinco, los cuales capturan el 67,88 % de la variabilidad total del conjunto de datos, sin pérdida sustancial de información significativa.

Una vez aplicado el método, se obtuvieron las siguientes cargas factoriales principales, que reflejan el peso relativo de cada variable en los componentes extraídos:

- PC1: Horas de estudio por día (0.86); puntuación de examen (0.98).
- PC2: Horas en redes sociales (0.64); Porcentaje de presencialidad (0.50); Frecuencia ejercicio (0.52).
- PC3: Edad (0.65); Escala de salud mental (0.56).

- PC4: Horas de sueño (0.63).
- PC5: Horas netflix (0.49); Porcentaje de presencialidad (0.46); Escala de salud mental (0.40).

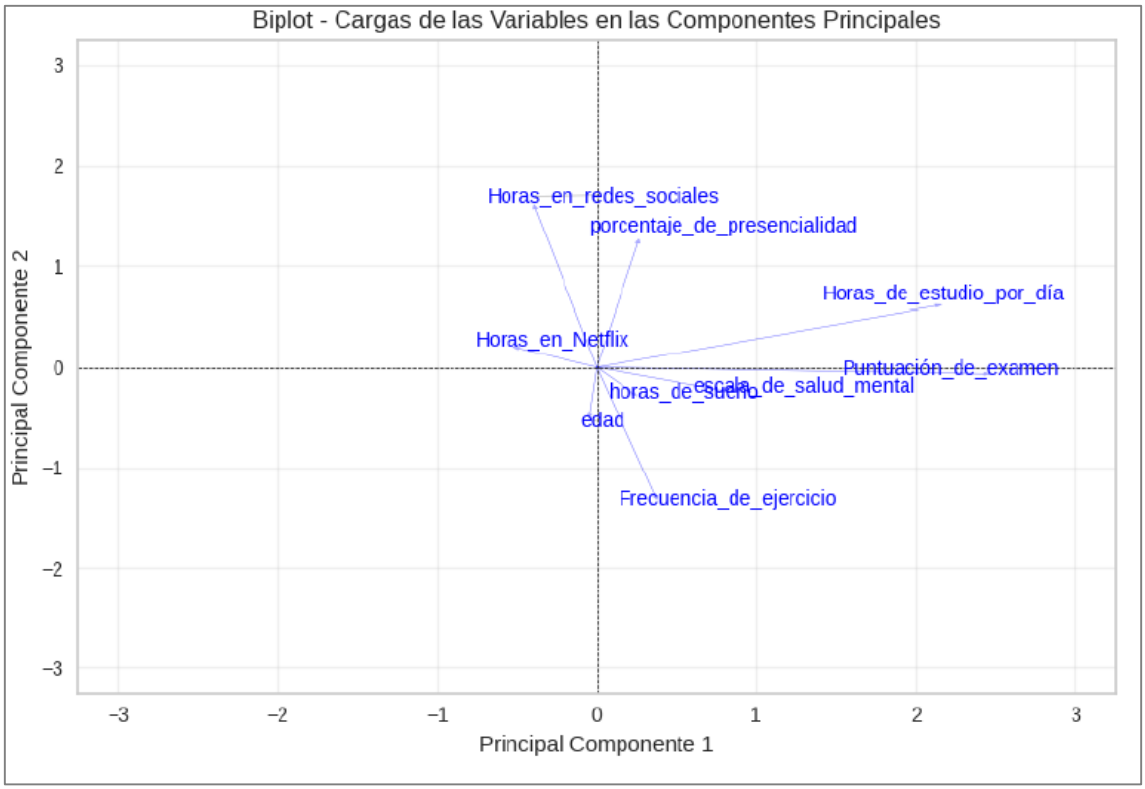

Gráfico 3: Biplot de cargas factoriales de las variables en PC1 Y PC2.

En el gráfico biplot de cargas factoriales —donde el Eje X representa el Componente Principal 1 (PC1) y el Eje Y el Componente Principal 2 (PC2)— se observa que estos dos componentes concentran la mayor parte de la varianza total del conjunto de datos. Cada flecha del gráfico representa una variable original, indicando la dirección y magnitud de su contribución a los componentes.

Las variables Horas de estudio por día y Puntuación de examen presentan flechas largas y orientadas positivamente hacia el eje de PC1, lo que evidencia una asociación positiva y fuerte entre ambas. En consecuencia, a mayor tiempo de estudio, mejor es el desempeño académico reflejado en la puntuación de los exámenes.

En cntraposición, las variables Horas en redes sociales y Horas en Netflix apuntan en dirección opuesta a las de Horas de estudio y Puntuación de examen, lo que sugiere una relación inversa o negativa. Esto implica que un mayor tiempo dedicado a las redes o al entretenimiento digital se asocia con una menor dedicación al estudio y un rendimiento académico inferior.

c. *Análisis de clusters*

A continuación, se aplicó el método multivariado denominado Análisis de Clúster o Conglomerados, con el objetivo de agrupar un conjunto de observaciones en grupos

homogéneos (clústeres), de manera que los elementos dentro de cada grupo sean similares entre sí y, a su vez, diferentes respecto a los pertenecientes a otros grupos.

Para la implementación de este método se definió como variable dependiente la puntuación de los exámenes, y como variables independientes las horas de estudio, las horas dedicadas al uso de redes sociales y las horas de sueño.

Previo al análisis, las variables independientes fueron estandarizadas (media = 0, desviación estándar = 1), con el fin de homogeneizar las escalas de medición y evitar que las diferencias en magnitud influyan en la distancia euclidiana empleada para la clasificación.

En la etapa exploratoria, se aplicó el Método del Codo (Elbow Method) para determinar el número óptimo de clústeres. Este procedimiento consiste en evaluar la Suma de los Errores Cuadráticos Internos (SSE) para distintos valores de k (número de clústeres), y seleccionar aquel punto donde la disminución del error comienza a ser marginal —el denominado punto de inflexión o codo de la curva—.

El gráfico resultante permitió identificar el punto óptimo en k = 3, lo cual indica que la división de los datos en tres clústeres representa adecuadamente la estructura interna del conjunto de observaciones, logrando un equilibrio entre simplicidad y capacidad explicativa.

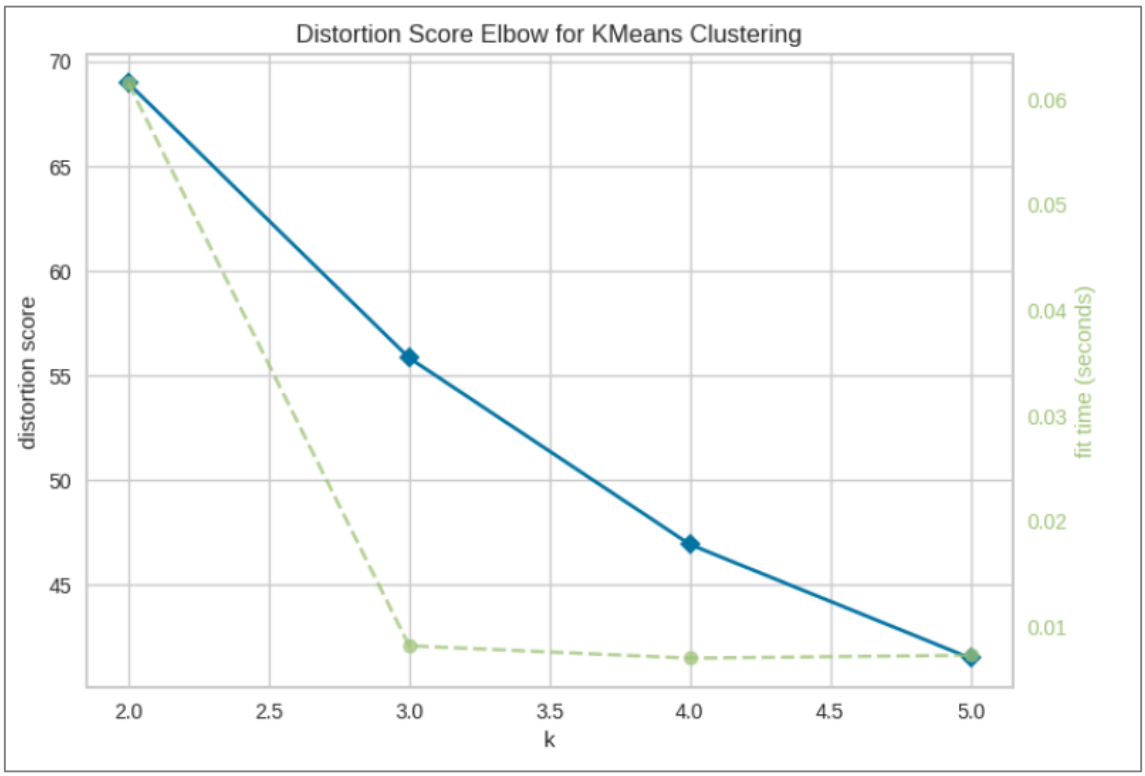

Gráfico 4: Método del codo basado en la distorsión para la agrupación K-means.

Una vez determinado el número óptimo de grupos, se procedió a ejecutar el modelo K-Means, asignando a cada observación el clúster correspondiente según su proximidad al centroide más cercano. Para facilitar la interpretación, se elaboró un gráfico de dispersión en el cual:

- el Eje X representa las Horas de estudio por día (normalizadas),
- el Eje Y muestra las Horas en redes sociales (normalizadas), y
- los colores diferencian los tres grupos obtenidos.

Cada punto del gráfico corresponde a un estudiante, y los centroides indican el centro geométrico de cada grupo, es decir, la zona donde se concentra la mayor densidad de observaciones.

El análisis visual evidencia una clara separación entre los grupos, reflejando perfiles de comportamiento diferenciados entre los estudiantes. Los resultados sugieren la existencia de patrones conductuales bien definidos:

- un grupo con alta dedicación al estudio y bajo uso de redes, asociado a mejores puntuaciones de examen;
- un segundo grupo con equilibrio moderado entre estudio, descanso y ocio digital; y
- un tercer grupo con mayor tiempo en redes y menor dedicación al estudio, vinculado a rendimientos académicos inferiores.

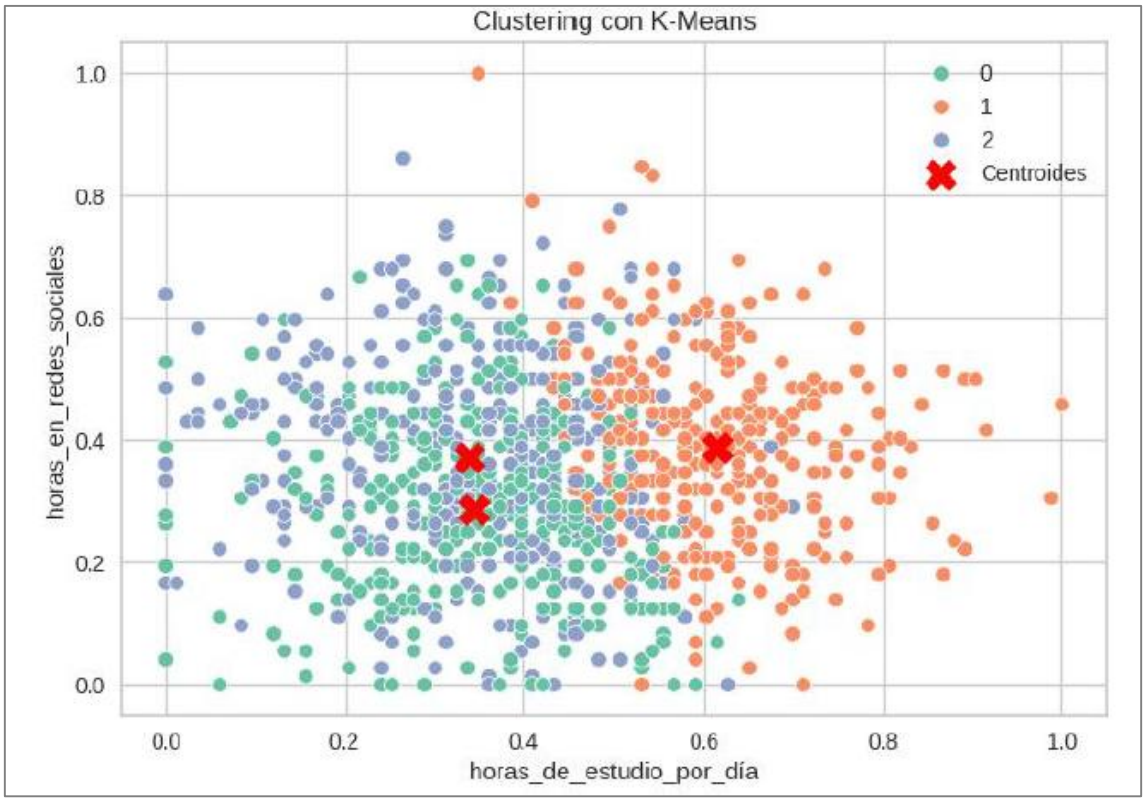

Gráfico 5: Dispersión entre Horas en redes sociales y Horas de estudio por día.

d. *Análisis de Varianza (ANOVA)*

Finalmente, se aplicó la técnica de Análisis de la Varianza (ANOVA) con el propósito de evaluar si existen diferencias estadísticamente significativas entre las medias de los tres clústeres identificados, en relación con los principales hábitos de estudio: horas de estudio por día, horas en redes sociales y horas de sueño. Este procedimiento permite determinar

si las diferencias observadas entre los grupos se deben a variaciones reales en los datos o si podrían explicarse por azar.

Para ello, se formularon las siguientes hipótesis:

- Hipótesis nula: Las medias de los 3 clusters son iguales entre sí.
- Hipótesis alternativa: Al menos una de las medias de los 3 clusters es diferente.

Este método proporciona dos resultados por cada una de las variables, un "Estadístico F" el cual compara la variación entre grupos con la variación dentro de los grupos, y un "Valor P" que es la probabilidad de que los resultados observados hayan ocurrido por azar. Como resultado por cada agrupación de los atributos se obtiene:

- Horas de estudio por día: $F = 521.09$, $p = 0.0000$
- Horas en redes sociales: $F = 39.57$, $p = 0.0000$
- Horas de sueño: $F = 543.98$, $p = 0.0000$

Dado que en todos los casos el valor p es menor que 0.05, se rechaza la hipótesis nula. Esto indica que existen diferencias estadísticamente significativas entre al menos dos de los tres grupos en cada una de las variables analizadas.

En consecuencia, se concluye que el proceso de clustering fue efectivo para segmentar a los estudiantes en grupos con hábitos de estudio significativamente distintos. Los resultados del ANOVA confirman que las variables horas de estudio, uso de redes sociales y horas de sueño presentan una influencia diferenciada sobre el rendimiento académico, validando la estructura de agrupación obtenida en el análisis previo.

Resultados

A partir del análisis multivariado realizado sobre la base de datos simulada de 1.000 estudiantes, se obtuvieron hallazgos relevantes que evidencian la influencia significativa de los hábitos estudiantiles en el rendimiento académico.

El análisis factorial permitió identificar una estructura latente compuesta por cinco factores principales: rendimiento académico, asistencia o participación, sueño y descanso, actividad física y salud mental. Las comunalidades resultaron superiores a 0.6 en la mayoría de las variables, destacándose la puntuación de examen (0.97) y las horas de estudio (0.89), lo cual indica una adecuada representación de la varianza por los factores comunes. Los gráficos de dispersión mostraron relaciones positivas entre una mayor dedicación al estudio y un mejor desempeño académico, así como entre la práctica de actividad física y mejores indicadores de bienestar mental.

El análisis de componentes principales (PCA) permitió reducir la dimensionalidad del conjunto de datos conservando la mayor cantidad posible de información relevante. Se identificaron cinco componentes que explican un 67,88% de la variabilidad total, concentrando los mayores pesos en las variables horas de estudio, puntuación de examen, uso de redes sociales, porcentaje de asistencia, frecuencia de ejercicio, edad, salud mental, horas de sueño y horas dedicadas a plataformas de streaming. El biplot obtenido evidenció que mayores horas de estudio se asocian positivamente con calificaciones más altas, mientras que el uso intensivo de redes sociales o plataformas como Netflix mantiene una relación negativa con el rendimiento académico.

Mediante la aplicación del método K-Means y la utilización del criterio del codo, se determinó que el número óptimo de clústeres era tres. La representación gráfica de los

grupos mostró una clara diferenciación entre los hábitos estudiantiles, confirmando la existencia de patrones de comportamiento distintivos entre los estudiantes.

Por último, la técnica de análisis de la varianza (ANOVA) permitió contrastar las medias de los tres clústeres respecto de las variables clave: horas de estudio, horas en redes sociales y horas de sueño. Los resultados obtenidos (F y $p < 0.05$ en todos los casos) posibilitaron rechazar la hipótesis nula, demostrando que las diferencias entre los grupos son estadísticamente significativas. Este resultado valida la segmentación efectuada y confirma que los hábitos seleccionados son determinantes en el rendimiento académico, sustentando empíricamente la relación entre las rutinas de estudio, el descanso, el uso del tiempo libre y el desempeño educativo.

Conclusiones

El análisis integral realizado permitió confirmar que los hábitos estudiantiles ejercen una influencia significativa sobre el rendimiento académico. Los resultados obtenidos evidencian que las conductas cotidianas de los estudiantes —particularmente el tiempo dedicado al estudio, las horas de descanso, la frecuencia de actividad física y el uso de redes sociales— conforman un conjunto de factores interrelacionados que explican en gran medida las diferencias de desempeño observadas.

El estudio demuestra que la dedicación al estudio constituye el factor más determinante del rendimiento académico, ya que existe una relación directa y positiva entre la cantidad de horas de estudio y las calificaciones obtenidas en los exámenes. En contraposición, el uso excesivo de redes sociales y plataformas de entretenimiento se asocia de manera negativa con el desempeño, reflejando el impacto del ocio digital en la capacidad de concentración y en la gestión del tiempo de los estudiantes.

Asimismo, se observó que el sueño adecuado, la práctica de actividad física y el mantenimiento de una buena salud mental, si bien tienen un peso relativo menor frente a las horas de estudio, actúan como variables complementarias que inciden de forma favorable en la productividad académica y en el bienestar general del estudiante.

Los resultados del análisis de conglomerados y del ANOVA validaron la existencia de perfiles diferenciados de estudiantes según sus hábitos, destacando la eficacia de las técnicas multivariadas para identificar patrones de comportamiento y establecer segmentaciones significativas. Esta clasificación permite comprender la heterogeneidad de la población estudiantil y ofrece una base empírica para diseñar estrategias educativas más precisas, orientadas al fortalecimiento de los hábitos que favorecen el aprendizaje y al acompañamiento de aquellos grupos que presentan mayores dificultades.

En síntesis, el estudio confirma que la mejora del rendimiento académico requiere un abordaje integral que contemple tanto la disciplina de estudio como el equilibrio entre descanso, bienestar físico y gestión del tiempo. Fomentar hábitos saludables, promover el uso responsable de la tecnología y establecer rutinas de estudio sostenibles se presentan como pilares fundamentales para optimizar los resultados académicos y potenciar el desarrollo personal de los estudiantes.

Tablas y gráficos complementarios

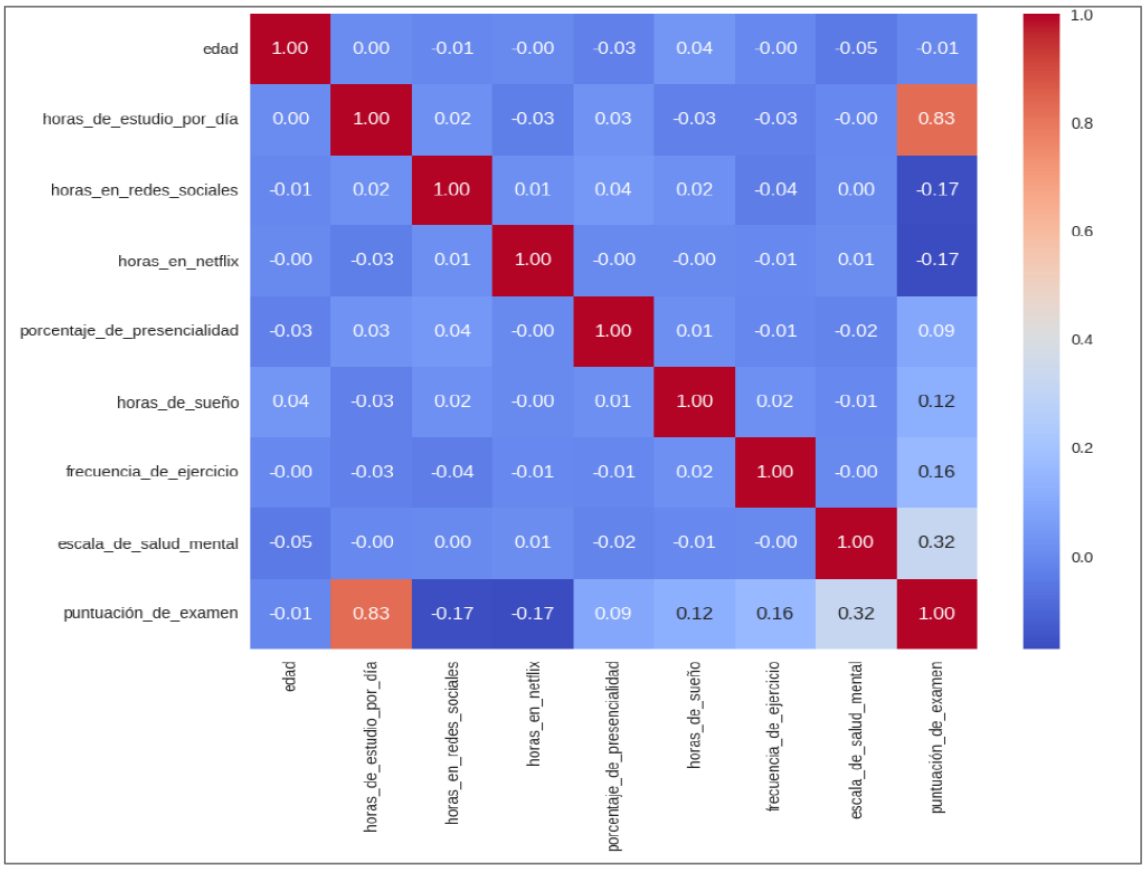

Gráfico 6: Mapa de calor entre matriz de correlación de Pearson con atributos numéricos.

	PA1	PA2	PA3	PA4	PA5
edad	0.068389	-0.427127	0.568128	-0.215535	-0.321703
Horas_de_estudio_por_día	0.929044	0.029627	-0.010222	-0.146741	-0.055987
Horas_en_redes_sociales	-0.100537	0.489144	0.167822	-0.492347	0.108851
Horas_en_Netflix	-0.225452	-0.064284	0.169173	-0.271126	0.457532
porcentaje_de_presencialidad	0.090632	0.761383	0.020366	0.044623	-0.164457
horas_de_sueño	0.013037	0.179551	0.797371	0.179899	0.138746
Frecuencia_de_ejercicio	-0.017586	0.060526	0.142411	0.781184	0.013879
escala_de_salud_mental	0.214578	-0.066757	-0.061911	0.091546	0.808327
Puntuación_de_examen	0.949282	0.022297	0.062080	0.214409	0.151450

Tabla 1: Cargas factoriales de los atributos.

	Variable	Comunalidad	Unicidad	Interpretación
0	edad	0.659832	0.340168	Moderadamente explicada
1	Horas_de_estudio_por_día	0.888772	0.111228	Muy Bien explicada.
2	Horas_en_redes_sociales	0.531788	0.468212	Moderadamente explicada.
3	Horas_en_Netflix	0.366426	0.633574	Mal explicada, carga poco.
4	porcentaje_de_presencialidad	0.617370	0.382630	Moderadamente explicada.
5	horas_de_sueño	0.719823	0.280177	Bien explicada.
6	Frecuencia_de_ejercicio	0.634695	0.365305	Moderadamente explicada.
7	escala_de_salud_mental	0.716107	0.283893	Bien explicada.
8	Puntuación_de_examen	0.974396	0.025604	Muy bien explicada.

Tabla 2: Comunalidades y Unicidades de los atributos.

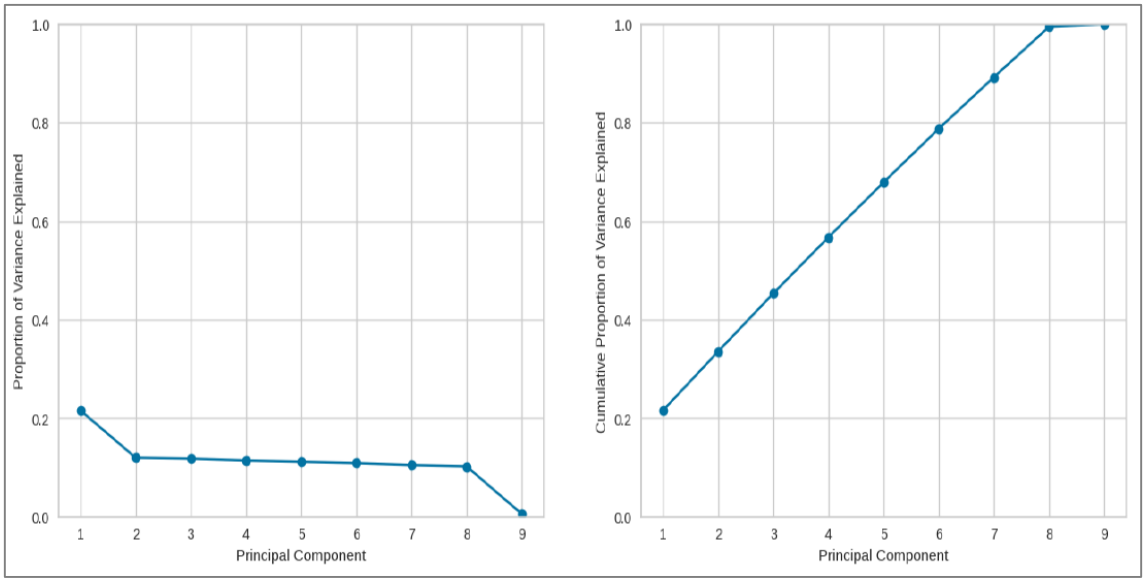

Gráfico 7: Varianza explicada por componente y varianza acumulada.

	PC1	PC2	PC3	PC4	PC5
edad	-0.019269	-0.202180	0.655493	-0.211242	0.380723
Horas_de_estudio_por_día	0.863878	0.244345	0.120679	-0.251984	0.074918
Horas_en_redes_sociales	-0.157649	0.643188	0.097719	0.237224	0.167188
Horas_en_Netflix	-0.209382	0.081701	-0.166298	0.217954	0.491037
porcentaje_de_presencialidad	0.106277	0.505308	0.135122	0.348424	-0.460108
horas_de_sueño	0.101029	-0.114004	0.492797	0.636625	0.221809
Frecuencia_de_ejercicio	0.150597	-0.528972	0.024355	0.419518	-0.395286
escala_de_salud_mental	0.330033	-0.096980	-0.564756	0.335101	0.408968
Puntuación_de_examen	0.986688	-0.027357	0.007640	0.021817	0.023141

Tabla 3: Cargas Factoriales de PCA.

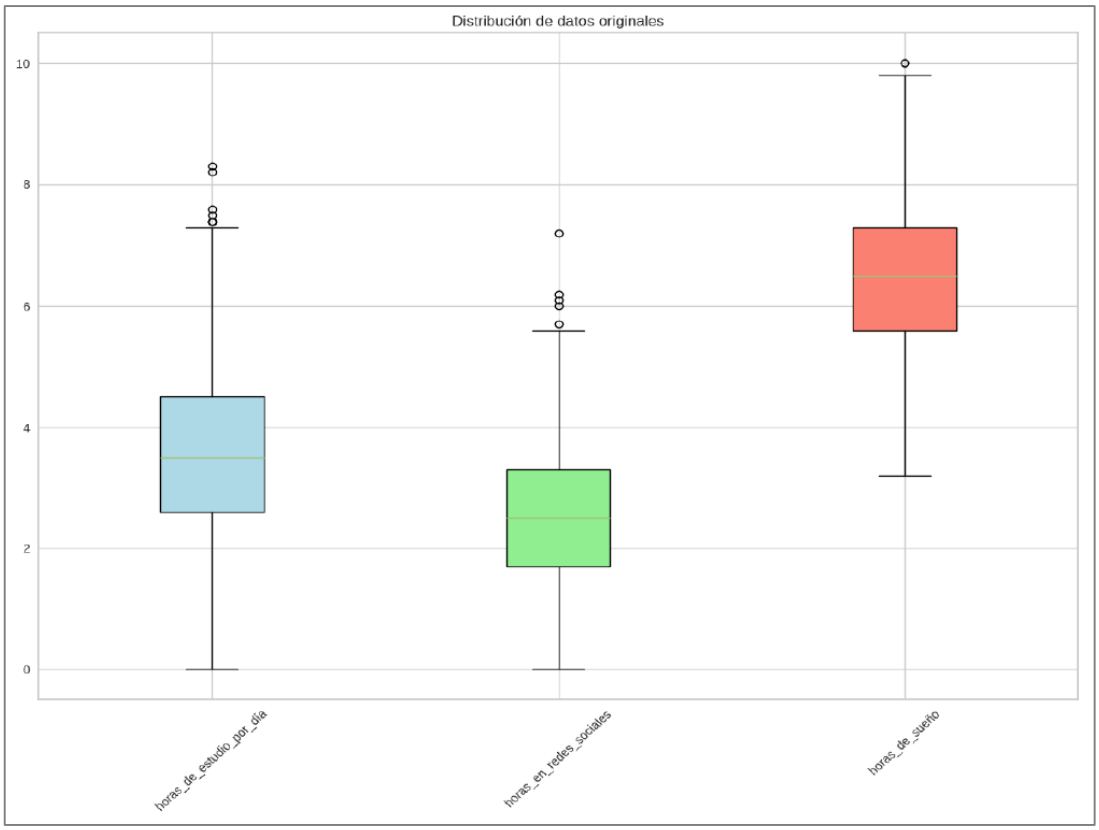

Gráfico 8: Distribución de los datos originales (clúster).

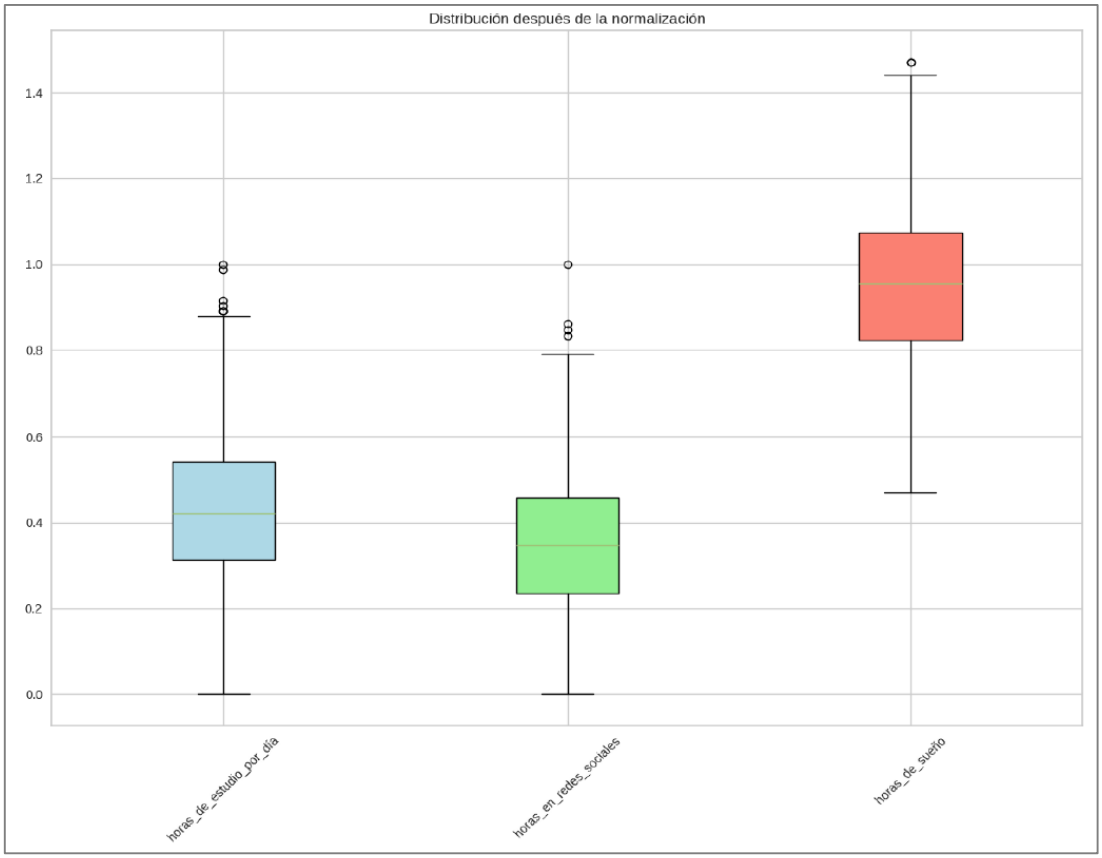

Gráfico 9: Distribución de los datos después de normalizar (clúster).

Código ejecutado en Python

```
**Recopilacion de los datos**

from google.colab import drive
drive.mount('/content/drive')

from google.colab import files

#Enlace drive:https://docs.google.com/spreadsheets/d/1SVzodl0tmF-
ILdbd6l5Q3ifpvh8IvOx4/edit?gid=1015985300#gid=1015985300

import gdown

ID = '1SVzodl0tmF-ILdbd6l5Q3ifpvh8IvOx4'
url = f'https://drive.google.com/uc?id={ID}'
output = 'student_habits_performance_excel_español.xlsx'
gdown.download(url, output, quiet=False)

# Importacion de Librerias
!pip install factor-analyzer
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.decomposition import PCA
from factor_analyzer import FactorAnalyzer
from scipy.linalg import svd
from sklearn.preprocessing import StandardScaler
import scipy.linalg as la

#Modelos estadísticos
import statsmodels.api as sm
import statsmodels.formula.api as smf
from statsmodels.formula.api import ols
from statsmodels.stats.multicomp import pairwise_tukeyhsd
from statsmodels.stats.multitest import multipletests
from statsmodels.multivariate.manova import MANOVA
import scipy.stats as stats
from scipy.stats import f_oneway

#Visualización
import matplotlib.pyplot as plt
import seaborn as sns

#Clustering y distancias
import scipy.cluster.hierarchy as sch
import scipy.spatial.distance as ssd
```

```
from scipy.spatial.distance import pdist
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
from scipy.cluster.hierarchy import fcluster
from yellowbrick.cluster import KElbowVisualizer

data1 = pd.read_excel('student_habits_performance_excel_español.xlsx')

# **Análisis general de los datos**

#Se imprimen las primeras filas para visualizar que se hayan importado correctamente
los datos
pd.set_option('display.max_columns', None)
print(data1.head())

# Información general de las variables
print("\n# Información general:")
print(data1.info())

# Identificacion de columnas categóricas
data_categ_cols = data1.select_dtypes(include='object').columns.tolist()
print("# Columnas categóricas identificadas:")
print(data_categ_cols)

# Identificamos columnas numericas
data_num_cols = data1.select_dtypes(include=['int64','float']).columns.tolist()

# Mostrar columnas que serán tratadas como numericas
print("# Columnas numericas identificadas:")
print(data_num_cols)

# Ver estadísticos descriptivos de todas las variables
pd.set_option('display.max_columns', None)
print("\n# Estadísticas descriptivas:")
print(data1.describe(include='all'))

# Copia del DataFrame original para trabajar de forma segura
data_clean = data1.copy()

# Normalizar los nombres de las columnas: minúsculas, sin espacios
data_clean.columns = data_clean.columns.str.strip().str.lower().str.replace(' ', '_')

# Eliminar columnas que no aportan valor al análisis multivariado
# En este caso: nivel educativo parental (muchos nulos), ID estudiantes (irrelevante
para el caso)
data_clean =
data_clean.drop(columns=['nivel_educativo_de_los_padres','id_estudiante'])

# **APLICACION DE LOS METODOS**
```

```
# **ANÁLISIS FACTORIAL**

# Seleccionamos columnas numericas del dataframe
data_clean.columns = data_clean.columns.str.strip().str.lower().str.replace(' ', '_')
data_num_cols = data_clean.select_dtypes(include=['int64','float']).columns.tolist()
num_data = data_clean[data_num_cols]

# Calculo de Matriz de correlacion de Pearson
r = num_data.corr()

# Vemos la matriz
print(r)

# Se carga la matriz de correlacion en un data frame para mejor visualizacion
data_corr = pd.DataFrame (r, index=data_num_cols, columns=data_num_cols)
data_corr

#Mapa de calor para matriz de correlacion de Pearson
plt.figure(figsize=(10, 8))
sns.heatmap(r, annot=True, cmap='coolwarm', fmt=".2f", square=True)
plt.title('Matriz de correlación de Pearson entre atributos numéricos')
plt.show()

from factor_analyzer.factor_analyzer import calculate_bartlett_sphericity

# Test de esfericidad de Bartlett
chi2, p_value = calculate_bartlett_sphericity(num_data)

print(f"Chi-cuadrado: {chi2:.2f}")
print(f"Valor p: {p_value:.5f}")

if p_value < 0.05:
    print("➤ Rechazamos H₀: existe correlación significativa entre variables. Se
justifica aplicar análisis factorial.")
else:
    print("➤ No se rechaza H₀: no hay suficiente evidencia de correlación. No se
justifica el análisis factorial.")

# Descomposición en valores singulares (SVD)
U, S, Vt = svd(r)

print(S)

suma_total = sum(S)
print("Suma total de los valores singulares:", suma_total)

from IPython.display import display, Math
```

```
# Máxima Verosimilitud: considerar que el número de factores a utilizar (p) debe ser menor o igual a:
display(Math(r"p\leq\left\lVert \frac{1}{2} \cdot \left( 2m + 1 - \sqrt{8m + 1} \right) \right\rVert"))

# Se debe tomar una cantidad de factores menores o iguales a 5
resultado = abs( (1/2) * (2*9 + 1 - np.sqrt(8*9 + 1)) )
resul_redon = round(resultado)

resul_redon

# Análisis factorial (Método de factores principales)
fa2 = FactorAnalyzer(n_factors=5, method='principal', rotation='varimax')

# Ajustar el modelo a los datos, entre columnas numericas.
fa2.fit(num_data)

# Muestra la matriz de cargas factoriales (9 variables × 5 factores).
print(fa2.loadings_)

variables = ['edad', 'Horas_de_estudio_por_día', 'Horas_en_redes_sociales', 'Horas_en_Netflix', 'porcentaje_de_presencialidad', 'horas_de_sueño', 'Frecuencia_de_ejercicio', 'escala_de_salud_mental', 'Puntuación_de_examen']
df_cargas = pd.DataFrame(fa2.loadings_, index=variables, columns=["PA1", "PA2", "PA3","PA4", "PA5"])

# Mostrar resultado
print(df_cargas)

# Calcular comunalidades
comunalidades = fa2.get_communalities()

# Calcular unicidades
unicidades = fa2.get_uniquenesses()

# Crear tabla con ambas
df_varianza = pd.DataFrame({
    "Variable": ['edad', 'Horas_de_estudio_por_día', 'Horas_en_redes_sociales', 'Horas_en_Netflix', 'porcentaje_de_presencialidad', 'horas_de_sueño', 'Frecuencia_de_ejercicio', 'escala_de_salud_mental', 'Puntuación_de_examen'],
    "Comunalidad": comunalidades,
    "Unicidad": unicidades
})

# Mostrar tabla
print(df_varianza)

interpretaciones = [
    "Moderadamente explicada",
```

```
    "Muy Bien explicada.",
    "Moderadamente explicada.",
    "Mal explicada, carga poco.",
    "Moderadamente explicada.",
    "Bien explicada.",
    "Moderadamente explicada.",
    "Bien explicada.",
    "Muy bien explicada."

]

df_varianza["Interpretación"] = interpretaciones
df_varianza

# Se obtienen los factores:
fa2.get_factor_variance()

scores = fa2.transform(num_data)

# Crea un DataFrame con las puntuaciones
df_scores = pd.DataFrame(scores, columns=["PA1", "PA2", "PA3","PA4","PA5"])

# GRAFICO DE DISPERSIÓN - PA1 y PA2
sns.scatterplot(data=df_scores, x='PA1', y='PA2', hue='PA1', palette='viridis')

plt.axhline(0, color='gray', linestyle='--', linewidth=0.5)
plt.axvline(0, color='gray', linestyle='--', linewidth=0.5)
plt.title('Gráfico de dispersión entre PA1 y PA2')
plt.xlabel('Eje X - Rendimiento académico')
plt.ylabel('Eje Y - Asistencia o participación')

plt.grid(False)
plt.show()

# GRAFICO DE DISPERSIÓN - PA1 y PA3
sns.scatterplot(data=df_scores, x='PA1', y='PA3', hue='PA1', palette='viridis')

plt.axhline(0, color='gray', linestyle='--', linewidth=0.5)
plt.axvline(0, color='gray', linestyle='--', linewidth=0.5)
plt.title('Gráfico de dispersión entre PA1 y PA3')
plt.xlabel('Eje X - Rendimiento académico')
plt.ylabel('Eje Y - Sueño y descanso')

plt.grid(False)
plt.show()

# **ANÁLISIS DE COMPONENTES PRINCIPALES**

data_matrix = num_data.to_numpy()
```

```
# Media, matriz de varianza/covarianza y correlaciones
mean_vals = np.mean(data_matrix, axis=0)
cov_matrix = np.cov(data_matrix, rowvar=False)
cor_matrix = np.corrcoef(data_matrix, rowvar=False)

# Convertir a DataFrames
df_medias = pd.DataFrame(mean_vals.reshape(1, -1), columns=variables)
df_cov = pd.DataFrame(cov_matrix, index=variables, columns=variables)
df_cor = pd.DataFrame(cor_matrix, index=variables, columns=variables)

print("Media de cada variable:")
print(df_medias)

print("\nMatriz de covarianza:")
print(df_cov)

print("\nMatriz de correlación:")
print(df_cor)

En este caso se utiliza la matriz de correlación, ya que: Estandariza las variables y
Permite comparar patrones

# PCA con sklearn
scaler = StandardScaler()
data_scaled = scaler.fit_transform(data_matrix)
pca = PCA()
pca.fit(data_scaled)
print("Varianza explicada:", pca.explained_variance_ratio_)
print(sum(pca.explained_variance_ratio_))

# PCA con sklearn
scaler = StandardScaler()
data_scaled = scaler.fit_transform(data_matrix)
pca2 = PCA(n_components=5)
pca2.fit(data_scaled)
print("Varianza explicada:", pca2.explained_variance_ratio_)
print(round(sum(pca2.explained_variance_ratio_),4))

# Devuelve una matriz de forma (n_componentes, n_variables) que contiene los
autovectores del PCA.
pca.components_

# Se crean los Scree Plots
fig, axes = plt.subplots(1, 2, figsize=(15, 6))

ticks = np.arange(pca.n_components_) + 1

# Primer gráfico: varianza explicada por componente
```

```
ax = axes[0]
ax.plot(ticks, pca.explained_variance_ratio_, marker='o')

ax.set_xlabel('Principal Component')  # Etiqueta del eje x
ax.set_ylabel('Proportion of Variance Explained')  # Etiqueta del eje y
ax.set_ylim([0, 1])  # Límite del eje y entre 0 y 1 (porcentaje)
ax.set_xticks(ticks)  # Ubicación de los ticks en el eje x

# Segundo gráfico: varianza acumulada
ax = axes[1]
ax.plot(ticks, pca.explained_variance_ratio_.cumsum(), marker='o')

ax.set_xlabel('Principal Component')  # Etiqueta del eje x
ax.set_ylabel('Cumulative Proportion of Variance Explained')  # Etiqueta del eje y
ax.set_ylim([0, 1])  # Límite del eje y
ax.set_xticks(ticks)  # Ticks del eje x

# Matriz de cargas
loadings = pca2.components_.T * np.sqrt(pca2.explained_variance_)

# Lista con los nombres de las variables originales
variables = ['edad', 'Horas_de_estudio_por_día', 'Horas_en_redes_sociales',
'Horas_en_Netflix', 'porcentaje_de_presencialidad', 'horas_de_sueño',
'Frecuencia_de_ejercicio', 'escala_de_salud_mental', 'Puntuación_de_examen']

# Crea un DataFrame con los loadings y nombres de variables y componentes
df_loadings = pd.DataFrame(loadings, index=variables, columns=["PC1", "PC2",
"PC3","PC4","PC5"])

# Muestra la tabla de cargas factoriales de PCA
print(df_loadings)

# Se convierte la matríz de correlación en un DataFrame
df_cor = pd.DataFrame(cor_matrix, index=variables, columns=variables)

!pip install adjustText #Para graficar el Biplot, hay que ajustar el texto dada la
cantidad de variables

!pip install adjustText

from adjustText import adjust_text
import matplotlib.pyplot as plt

plt.figure()

# Escala para flechas
scale_factor = 2.5

texts = []
```

```
for i, var in enumerate(df_cor.columns):
    x = df_loadings.iloc[i, 0] * scale_factor
    y = df_loadings.iloc[i, 1] * scale_factor

    # Dibujar flechas
    plt.arrow(0, 0, x, y,
              color="blue", alpha=0.3,
              head_width=0.03, head_length=0.04, length_includes_head=True)

    # Agregar etiquetas
    texts.append(plt.text(x * 1.05, y * 1.05, var, color="blue", fontsize=10))

# Evitar superposición de etiquetas
adjust_text(texts, arrowprops=dict(arrowstyle="->", color="gray", alpha=0.5, lw=0.5))

# Líneas guía
plt.axhline(0, color="black", linewidth=0.5, linestyle="--")
plt.axvline(0, color="black", linewidth=0.5, linestyle="--")

# Límites del gráfico
plt.xlim(-scale_factor * 1.3, scale_factor * 1.3)
plt.ylim(-scale_factor * 1.3, scale_factor * 1.3)

# Etiquetas
plt.xlabel("Principal Componente 1")
plt.ylabel("Principal Componente 2")
plt.title("Biplot - Cargas de las Variables en las Componentes Principales")
plt.grid(alpha=0.3)

plt.tight_layout()
plt.show()

# **ANÁLISIS DE CLUSTERS**

data_indep = data_clean[['horas_de_estudio_por_día', 'horas_en_redes_sociales',
'horas_de_sueño']] #defino las variables independientes que vamos a trabajar

# Análisis de dispersión de los datos originales
print("Varianza de cada variable:")
print(data_indep.std())

# Crear figura y ejes
fig, ax = plt.subplots(figsize=(12, 10))

bp = ax.boxplot(data_indep.values, patch_artist=True)

#Eje X
ax.set_xticklabels(data_indep.columns, rotation=45)
```

```
colors = ['lightblue', 'lightgreen', 'salmon', 'plum', 'gold', 'skyblue', 'khaki',
'lightcoral', 'orchid']

for patch, color in zip(bp['boxes'], colors):
    patch.set_facecolor(color)

# Título
ax.set_title("Distribución de datos originales")

plt.tight_layout()
plt.show()

# Normalizar los datos dividiendo por el rango de cada variable

# Se calcula el rango (valor máximo - valor mínimo) para cada variable del DataFrame.
Esto permite escalar todas las variables a una misma base sin modificar su forma
relativa
rge = data_indep.apply(lambda x: np.ptp(x), axis=0)  # np.ptp = "peak to peak",
equivalente a max(x) - min(x)

# Se divide cada valor de cada variable por su respectivo rango. Esto produce una
versión del DataFrame con todas las variables reescaladas entre 0 y 1
(aproximadamente)
data_indep_s = data_indep.div(rge)

# Se imprime la varianza de cada variable una vez normalizada para confirmar
homogeneidad de escalas
print("Varianza después de la normalización:")
print(data_indep_s.var())

# Boxplot después de la normalización
fig, ax = plt.subplots(figsize=(12, 10))

bp = ax.boxplot(data_indep_s.values, patch_artist=True)

colors = ['lightblue', 'lightgreen', 'salmon']

for patch, color in zip(bp['boxes'], colors):
    patch.set_facecolor(color)

ax.set_xticklabels(data_indep_s.columns, rotation=45)

ax.set_title("Distribución después de la normalización")

plt.tight_layout()

plt.show()
```

```
# Escalar los datos
d = StandardScaler().fit_transform(data_indep)

# Método del codo para determinar la cantidad óptima de clusters

model = KMeans()

visualizer = KElbowVisualizer(model, k=(2, 6))

visualizer.fit(data_indep_s)

visualizer.show()

# Aplicar K-Means con 3 clusters
km_3 = KMeans(n_clusters=3, random_state=42).fit(data_indep_s)

# Asignar clusters a los datos
data_indep['cluster_3'] = km_3.labels_

# Visualizar clusters

sns.scatterplot(
    x=data_indep_s.iloc[:, 0],             # Primer eje
    y=data_indep_s.iloc[:, 1],             # Segundo eje
    hue=km_3.labels_,                 # Color según número de cluster asignado
    palette='Set2'                    # Paleta de colores para diferenciar grupos
)

# Se agregan los centroides al gráfico
plt.scatter(
    km_3.cluster_centers_[:, 0],      # Coordenada X de los centroides
    km_3.cluster_centers_[:, 1],      # Coordenada Y de los centroides
    s=200,                            # Tamaño de los marcadores de los centroides
    c='red',                          # Color rojo para destacar los centroides
    marker='X',                       # Forma del marcador: 'X' para que se
diferencie
    label='Centroides'                # Etiqueta de la leyenda
)

# Se añade un título al gráfico
plt.title("Clustering con K-Means")

# Se muestra la leyenda que identifica los centroides
plt.legend()

# Se muestra el gráfico final
plt.show()

# Información de los clusters
```

```
# Se imprime la matriz de centroides del modelo K-Means con 3 clusters. Cada fila representa un cluster, y cada columna representa el valor medio (normalizado) de cada variable en ese cluster
print("Centroides de los clusters:", km_3.cluster_centers_)

# Se imprime la inercia total del modelo, que es la suma de las distancias cuadradas de cada observación a su centroide. Cuanto menor es este valor, más compactos son los clusters.
print("Suma total de cuadrados:", km_3.inertia_)

# **ANOVA**

**Hipótesis del ANOVA para cada variable:**

Hipótesis nula:
$$
H_0: \mu_{\text{Cluster 1}} = \mu_{\text{Cluster 2}} = \mu_{\text{Cluster 3}}
$$

Hipótesis alternativa:
$$
H_1: \exists\ i, j \text{ tal que } \mu_{\text{Cluster } i} \ne \mu_{\text{Cluster } j}
$$

# ANOVA por variable para comparar medias entre clusters

print("\n=== ANOVA por variable ===")

# Se itera sobre cada variable normalizada para realizar una ANOVA unifactorial
for var in data_indep_s.columns:

    grupos = [data_indep_s[var][data_indep['cluster_3'] == g] for g in sorted(data_indep['cluster_3'].unique())]

    f_stat, p_val = f_oneway(*grupos)

    print(f"{var}: F = {f_stat:.2f}, p = {p_val:.4f}")

for var in data_indep_s.columns:
    plt.figure()
    sns.boxplot(x=data_indep['cluster_3'], y=data_indep_s[var], palette='Set2')
    plt.title(f'Boxplot de {var} según Cluster')
    plt.xlabel('Cluster')
    plt.ylabel(var)
    plt.tight_layout()
    plt.show()
```