ESCUELA NACIONAL DE ESTUDIOS SUPERIORES
UNIDAD MORELIA

Universidad Nacional Autónoma de México

Escuela Nacional de Estudios Superiores Unidad Morelia
Tecnologías para la Información en Ciencias

Análisis de Factores Determinantes del Desempeño Académico

Presenta:
Roxana Pérez Medina
Número de cuenta: 424013174

Profesora:
Dra. Haydeé Contreras Peruyero

Asignatura:
Estadística Multivariada

5 de enero de 2025

1. Introducción

El desempeño estudiantil no es el resultado de un único elemento aislado, sino la consecuencia de una compleja interacción de múltiples variables que abarcan aspectos conductuales, psicológicos, ambientales y socioeconómicos. El presente trabajo tiene como objetivo estudiar un conjunto de datos que contiene variables que representan los factores antes mencionados de una muestra de estudiantes. A través de herramientas de estadística multivariada se busca comprender los elementos que influyen en las calificaciones obtenidass, así como identificar perfiles de estudiantes con características similares.

En primer lugar, se realiza un análisis exploratorio de los datos para examinar, de manera superficial, el comportamiento conjunto de las variables y detectar patrones generales. Posteriormente, se ajusta un modelo de Regresión Lineal Múltiple para cuantificaar los predictores significativos del puntaje en exámenes. Es importante destacar que para este proyecto se utiliza una significancia $\alpha = 0{,}05$ Después, se aplica el Análisis de Componentes Principales como técnica de reducción de dimensionalidad. Finalmente, se utilizan técnicas de Clustering para identificar grupos de estudiantes con características y rendimientos similares.

2. Descripción de los datos

El conjunto de datos utilizado en este proyecto, Exam Score Prediction Dataset, fue obtenido de la plataforma Kaggle, publicado en esta por Ayesha Saher. Consiste de 13 variables. Los atributos que conforman el conjunto de datos son los siguientes:

- **user_id**: Identificador único del estudiante.
- **age**: Edad del estudiante.
- **gender**: Identidad de género del estudiante (Male, Female, Other).
- **course**: Curso o programa al que está inscrito el estudiante.
- **study_hours**: Horas de estudio diarias del usuario.
- **class_attendance**: Porcentaje de clases asistidas del usuario.
- **internet_access**:¿El estudiante tiene acceso a internet?.
- **sleep_hours**: Horas de sueño promedio por día del estudiante.
- **sleep_quality**: Calidad de sueño del estudiante (Good, Average, Poor).
- **study_method**: Método de estudio preferido del estudiante (Self-Study, Online Videos, Group Study, Coaching, Mixed).
- **facility_rating**: Calificación de las instalaciones/institución (Low, Medium, High).
- **exam_difficulty**: Dificultad del examen aplicado al estudiante (Easy, Moderate, Hard).
- **exam_score**: Calificación del examen aplicado al estudiante.

Hay 20,000 instancias en el conjunto de datos, de las cuales, todas tienen los registros completos. Es decir, no hay datos faltantes.

3. Análisis Exploratorio de Datos

Se revisaron las distribuciones de las variables numéricas y se evaluaron relaciones preliminares entre las variables.

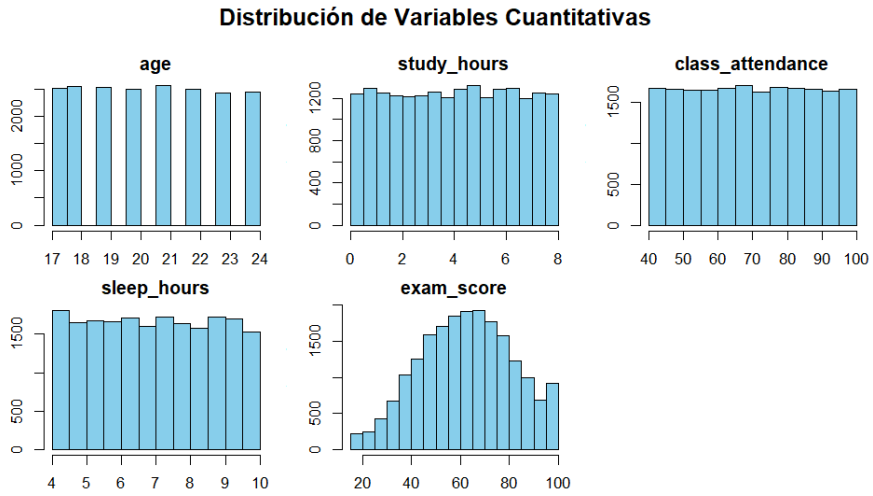

Figura 1: Distribución de las variables cuantitativas.

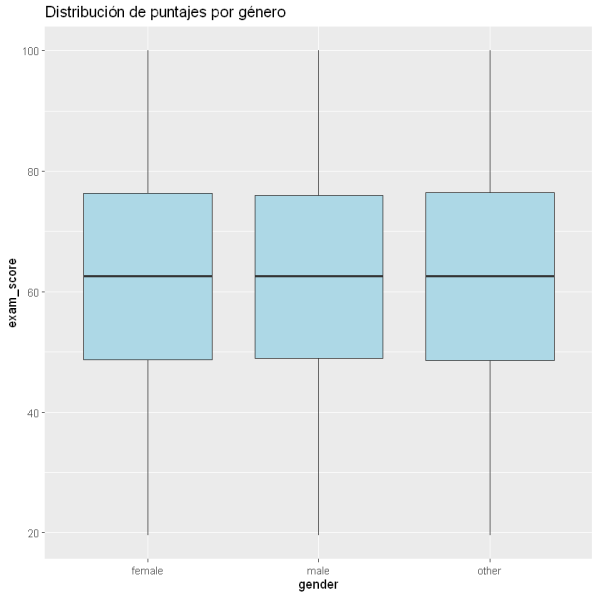

Figura 2: Distribución de calificaciones por género.

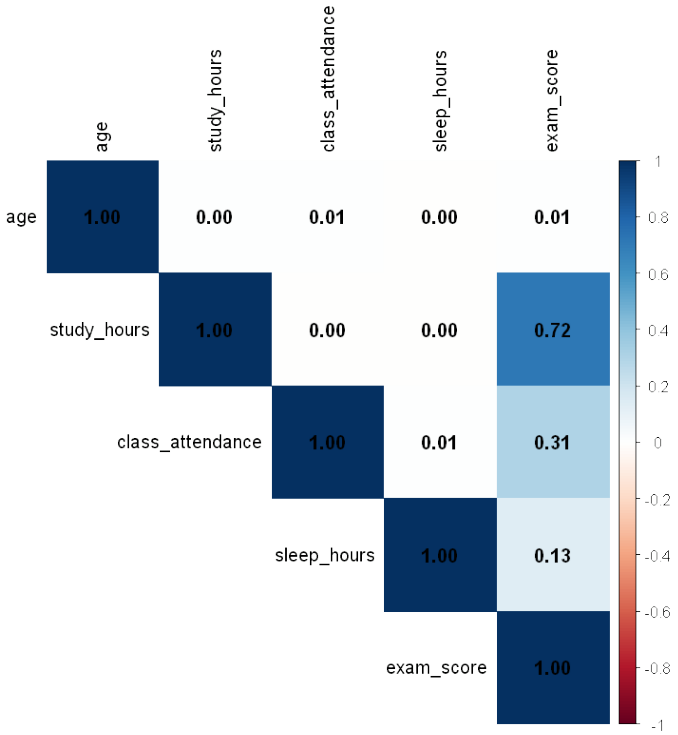

Figura 3: Mapa de correlaciones entre las variables numéricas.

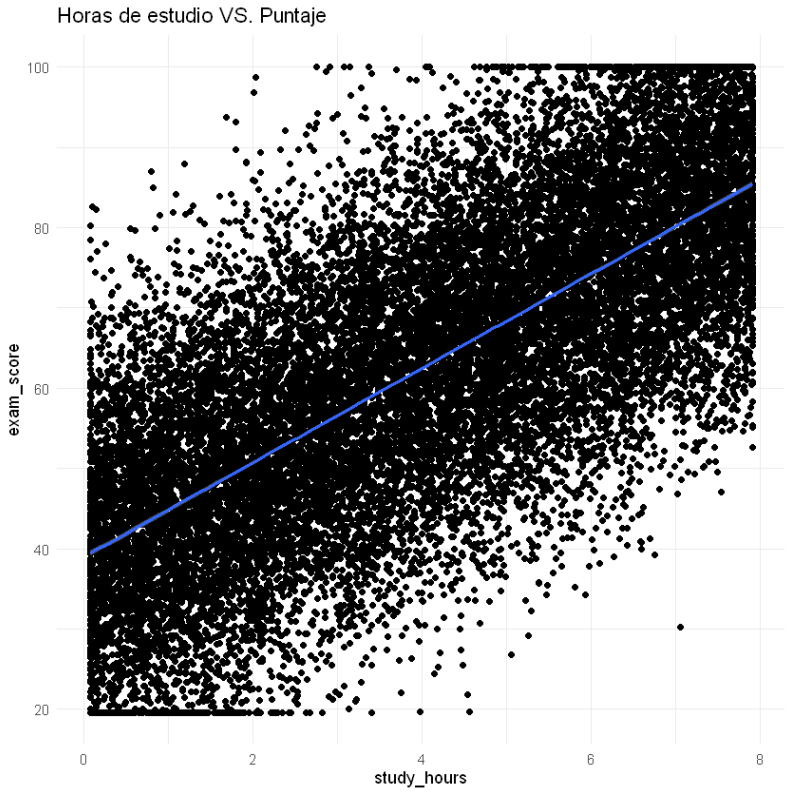

Figura 4: Relación entre horas de estudio y calificación.

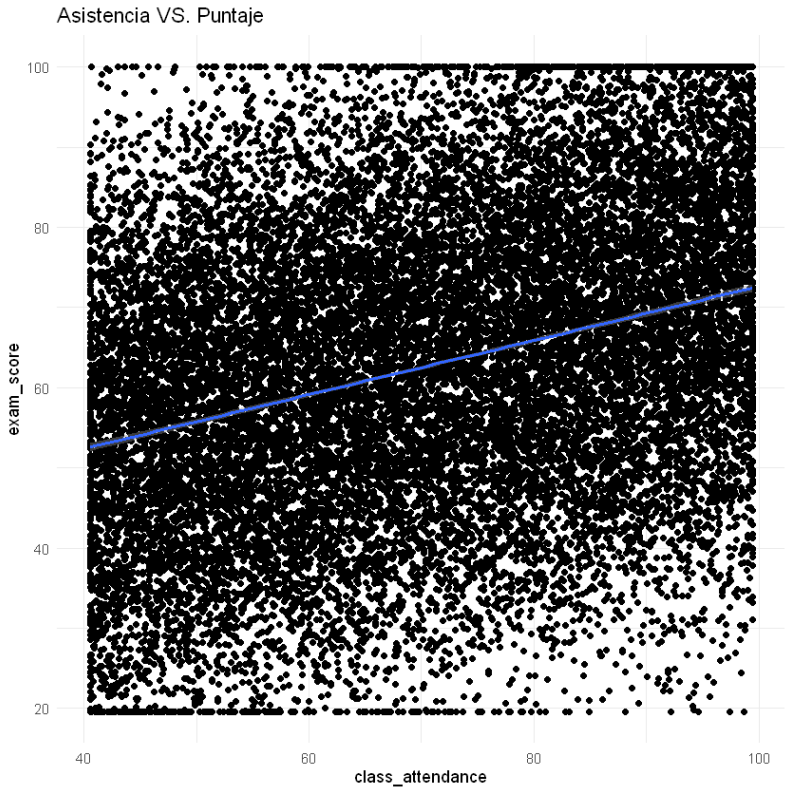

Figura 5: Relación entre porcentaje de asistencia a clases y calificación.

En la Figura 1 se muestran las distribuciones de las variables cuantitativas, se observa que la edad, las horas de estudio, la asistencia a clases y las horas de sueño presentan distribuciones aproximadamente uniformes, sin evidencias claras de asimetrías pronunciadas ni valores atípicos. Por otra parte, la variable de calificación de examen presenta una distribución aproximadamente normal.

La Figura 2 muestra las distribuciones de las calificaciones en los estratos de la variable género mediante una gráfica de boxplots, pero no se puede observar ninguna diferencia clara entre los tres estratos.

En la Figura 3 se muestra un mapa de correlaciones entre las variables cuantitativas, donde se tiene que la mayoría de parejas tienen una correlación de ceo o cercana a cero que muestra que son independientes entre sí. La variable de calificación tiene una correlación alta con horas de estudio, esta relación se presenta más a detalle en la Figura 4 donde se muestra una tendencia positiva con un poco de dispersión. La variable de calificación tiene correlación moderada (0.31) con el porcentaje de asistencia a clases, en la Figura 5 se muestra esta relación con una tendencia positiva, pero con dispersión considerable.

4. Regresión Lineal Múltiple

Se ajustó un modelo de Regresión Lineal Mültiple tomando como variable respuesta la clasificación del examen y el resto de variables como las variables explicativas o regresoras.

$$exam\ score = \beta_0 + \beta_1(gender) + \beta_2(course) + \beta_3(sleep\ hours) + \beta_4(study\ method)$$
$$+\beta_5(study\ hours) + \beta_6(class\ attendance) + \beta 7(sleep\ quality) + \beta_8(internet\ access)$$
$$+\beta_9(facility\ rating) + \beta_{10}(exam\ difficulty) + \beta_{11}(age) + \epsilon$$

4.1. Validación de Supuestos

Antes de interpretar los resultados del modelo, se realizó la validación de supuestos de la regresión lineal mediante en análisis gráfico de residuos y pruebas estadísticas.

4.1.1. Multicolinealidad

Hay que probar que las variables sean independientes, mediante una matriz de correlación:

	Horas Estudio	**Asistencia Clases**	**Horas Sueño**	**Edad**
Horas Estudio	1.000000	-0.001645	-0.004533	0.002955
Asistencia Clases	-0.001645	1.000000	0.007187	0.008449
Horas Sueño	-0.004533	0.007187	1.000000	-0.000385
Edad	0.002955	0.008449	-0.000385	1.000000

Cuadro 1: Matriz de correlaciones entre variables del estudio

Como todas las correlaciones son prácticamente cero, las variables son independientes y no hay multicolinealidad, se cumple este supuesto.

4.1.2. Normalidad en los Residuales

Este supuesto se comprueba con un QQ-plot de los residuales o con una prueba de Shapiro-Wilks, pero como en este caso se tienen 20,000 observaciones no se puede realiza esta última, pues es para muestras más pequeñas.

En esta gráfica se puede observar que los puntos siguen muy bien la línea, sólo se desvían un poco en las orillas, por lo que se puede decir que este supuesto también se cumple.

4.1.3. Homocedasticidad

Este supuesto se revisa con una gráfica de los residuales contra los valores predichos o con una prueba de Breush-Pagan, donde:
H_0 : Los residuos tienen varianza constante (homocedasticidad).
H_1 : Hay heterocedasticidad en los residuos.

En la gráfica se puede un patrón similar a un rombo, los residuos no están tan dispersos así que podría haber heterocedasticidad.

Se obtuvieron los siguientes resultados de la prueba Breush-Pagn:

El p-valor es mayor a la significancia ($p > 0{,}05$), por lo que no se rechaza la hipótesis nula, no hay evidencia suficiente para decir que hay heterocedasticidad. Este supuesto también se cumple.

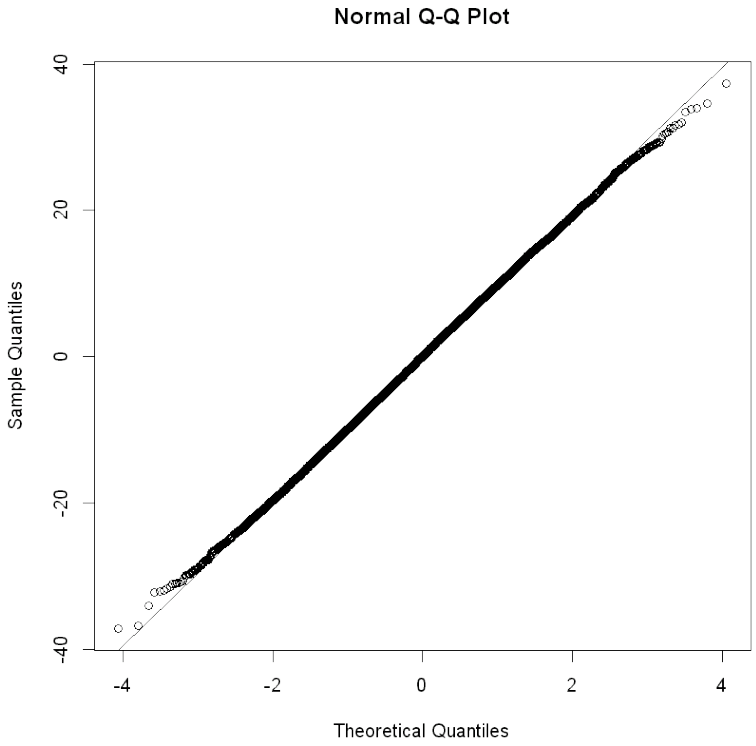

Figura 6: Gráfica de los residuales del modelo.

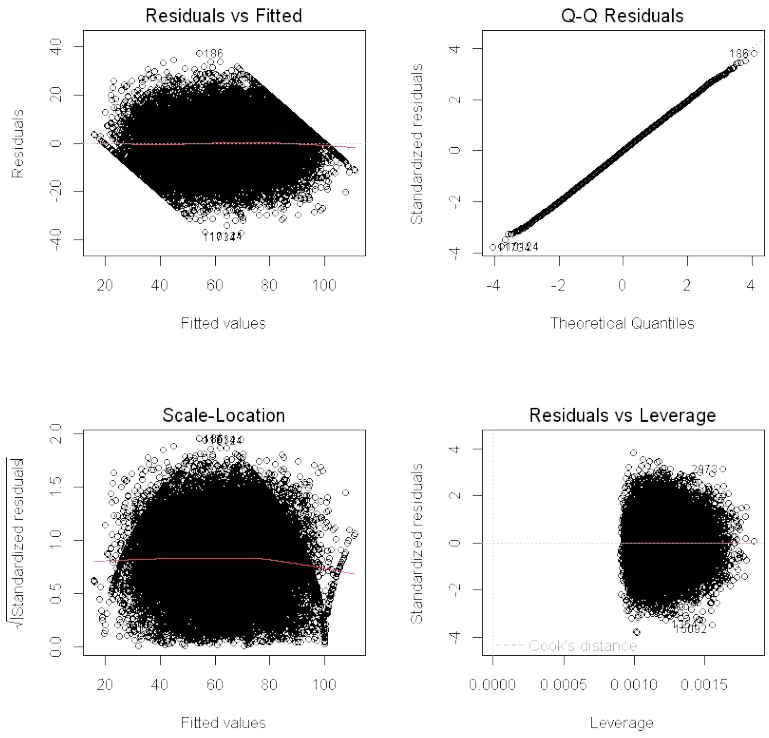

Figura 7: Gráfica de los residuales del modelo VS. los valores predichos.

Parámetro	Valor
Estadístico BP	32.29
Grados de libertad (df)	23
Valor p	0.0943

Cuadro 2: Resultados del test de Breusch-Pagan para homocedasticidad

4.1.4. No autocorrelación

Para evaluar este supuesto se realiza una prueba de Durbin-Watson, con las siguientes hipótesis:
H_0 : No hay autocorrelación en los errores.
H_1 : Hay autocorrelación en los errores.

Se obtuvieron los siguientes resultados:

Parámetro	Valor
Estadístico DW	2.017
Valor p	0.8853

Cuadro 3: Resultados del test de Durbin-Watson para autocorrelación

El p-valor es mayor a la significancia ($p > 0{,}05$), por lo que no se rechaza la hipótesis nula, no hay evidencia suficiente para decir que hay autocorrelación. Y el estadístico DW tiene un valor de aproximadamente 2, por lo que este supuesto también se cumple.

4.2. Pruebas de Hipótesis

Para el modelo de regresión se plantearon las siguientes hipótesis globales:

Hipótesis Nula (H_0 :) Todos los coeficientes asociados a las variables regresoras son iguales a cero:

$$\beta_1 = \beta_2 = ... = \beta_p = 0$$

Hipótesis Alterna (H_1 :) Al menos uno de los coeficientes es distinto de cero.

Estas hipótesis sirven para evaluar si el conjunto de variables regresoras aporta información significativa para explicar la calificación del examen. Se evalua con el estadístico F.

Estos fueron los resultados del modelo completo:

Variables	Coeficiente	Error Est.	Valor t	p
Intercepto	15,113 972	0,819 229	18,449	<,001***
Género (ref: femenino)				
Masculino	0,081 963	0,170 079	0,482	0,630
Otro	0,177 948	0,169 861	1,048	0,295
Curso (ref: no especificado)				
B.Sc	−0,181 423	0,258 515	−0,702	0,483
B.Tech	0,032 088	0,260 375	0,123	0,902
BA	0,053 079	0,258 061	0,206	0,837
BBA	0,080 964	0,259 440	0,312	0,755
BCA	0,096 910	0,257 936	0,376	0,707
Diploma	−0,038 715	0,259 651	−0,149	0,881
Horas de sueño	1,447 385	0,039 985	36,198	<,001***
Método de estudio (ref: coaching)				
Grupo	−7,696 113	0,219 573	−35,050	<,001***
Mixto	−4,850 444	0,219 974	−22,050	<,001***
Videos online	−8,825 206	0,217 562	−40,564	<,001***
Auto-estudio	−9,473 660	0,217 438	−43,570	<,001***
Horas de estudio	5,896 578	0,030 009	196,496	<,001***
Asistencia a clases	0,339 531	0,004 007	84,729	<,001***
Calidad del sueño (ref: regular)				
Buena	4,587 914	0,169 760	27,026	<,001***
Mala	−4,816 629	0,169 324	−28,446	<,001***
Acceso a internet (ref: no)				
Sí	0,085 751	0,193 659	0,443	0,658
Calificación instalaciones (ref: alta)				
Baja	−7,821 262	0,170 203	−45,953	<,001***
Media	−3,767 067	0,169 454	−22,231	<,001***
Dificultad examen (ref: fácil)				
Difícil	0,044 889	0,199 271	0,225	0,822
Moderada	0,062 434	0,159 158	0,392	0,695
Edad	−0,010 385	0,030 316	−0,343	0,732
Métricas del modelo				
Residual Standard Error	9.79 (19976 gl)			
Multiple R^2	0.7322			
Adjusted R^2	0.7319			
F-statistic	2375 (23, 19976 gl), $p < 0{,}001$			
Número de observaciones	20000			

Nota: ***$p < 0{,}001$. Variables categóricas tienen categorías de referencia como se indica.

Cuadro 4: Resultados completos del modelo de regresión lineal múltiple para predictores del puntaje en exámenes

El estadístico F calculado es mucho más alto que el estadístico crítico ($F_{(0,95,23,19976)=1,53}$), y el p-valor es más pequeño que el valor de significancia, por lo que se rechaza la hipótesis nula; al menos uno de los coeficientes de las variables regresoras es distinto de cero.

4.2.1. Pruebas de Hipótesis sobre los Coeficientes Individuales

Para probar cuáles son las variables regresoras que tienen un efecto estadísticamente significativo sobre la calificación, hay que evaluar cada coeficiente, con el t-test, tomando las siguientes hipótesis:

Hipótesis Nula (H_0 :) $\beta_i = 0$

Hipótesis Alterna (H_1 :) $\beta_i \neq 0$

Las variables que se probaron como estadísticamente significativas sobre la calificación, pues su p-valor es menor al nivel de significancia establecido y se rechaza la hipótesis nula, son el intercepto, horas de sueño, método de estudio, horas de estudio, asistencia a clases, calidad del sueño, calificación de las instalaciones.

4.3. ANOVA

El análisis de varianza se utilizó como una herramienta complementaria al resumen del modelo de regresión. En particular, para determinar el efecto de las variables categóricas sobre la calificación de los exámenes.

Variable	**gl**	**Suma Sq**	**Media Sq**	***p*-valor**
Género	2	136.3	68.2	0.491
Curso	6	1327.0	221.2	0.031*
Horas sueño	1	126782.9	126782.9	<0.001***
Método estudio	4	249187.0	62296.7	<0.001***
Horas estudio	1	3678939.0	3678939.0	<0.001***
Asistencia clases	1	683744.3	683744.3	<0.001***
Calidad sueño	2	293124.2	146562.1	<0.001***
Acceso internet	1	10.7	10.7	0.738
Instalaciones	2	202466.8	101233.4	<0.001***
Dificultad examen	2	14.4	7.2	0.927
Edad	1	11.2	11.2	0.732
Residuales	19976	1914519.0	95.8	

Cuadro 5: Tabla ANOVA del modelo de regresión.

Los resultados son consistentes con las pruebas de hipótesis individuales del punto anterior. Las mismas variables mencionadas antes se confirman como significativas con el análisis de varianza. Específicamente, las variables categóricas calidad de sueño, calificación de las instalaciones y métodos de estudio resultan estadísticamente significativas, aun cuando sus categorías de referencia no aparecen explícitamente en el resumen del modelo.

4.4. Selección de Modelo

Con el objetivo de explorar una modelo reducido de regresión, se aplicó selección paso a paso basado en el AIC.

Cuadro 6: Resultados del modelo de regresión lineal múltiple reducido (solo variables significativas)

Variable	**Coeficiente**	**Error Est.**	t	p
Intercepto	15,112 242	0,461 829	32,720	<,001***
Horas de sueño	1,446 783	0,039 966	36,200	<,001***
Método de estudio (ref: coaching)				
Estudio en grupo	−7,696 005	0,219 482	−35,060	<,001***
Método mixto	−4,849 712	0,219 881	−22,060	<,001***
Videos online	−8,825 371	0,217 450	−40,590	<,001***
Auto-estudio	−9,473 642	0,217 319	−43,590	<,001***
Horas de estudio	5,896 161	0,029 992	196,590	<,001***
Asistencia a clases	0,339 524	0,004 005	84,770	<,001***
Calidad del sueño (ref: regular)				
Buena	4,586 318	0,169 671	27,030	<,001***
Mala	−4,815 546	0,169 255	−28,450	<,001***
Calificación instalaciones (ref: alta)				
Baja	−7,819 068	0,170 139	−45,960	<,001***
Media	−3,764 709	0,169 383	−22,230	<,001***
Métricas del modelo				
Error estándar residual	9.788 (19988 gl)			
R^2 múltiple	0.7322			
R^2 ajustado	0.7321			
Estadístico F	4968 (11, 19988 gl), $p < 0{,}001$			
Observaciones	20000			

Nota: ***$p < 0{,}001$. Todas las variables son altamente significativas ($p < 0{,}001$).

El modelo resultante conserva las variables que aportan información relevante para explicar la variabilidad de la calificación de los exámenes, eliminando aquellas cuyo aporte es limitado.

$exam\ score = \beta_0 + \beta_1(sleep\ hours) + \beta_2(study\ method) + \beta_3(study\ hours) + \beta_4(class\ attendance) + \beta_5(sleep\ quality) + \beta_6(facility\ rating) + \epsilon$

Varias variables demográficas contextuales como el género, la edad, el acceso a internet o el curso fueron excluidas dado que su aporte al modelo es limitado. Cabe destacar que el modelo resultante presenta una R^2 prácticamente idéntica a la del modelo completo, lo que indica que la reducción en el número de variables eno implica una pérdida de capacidad explicativa. Por lo tanto, el modelo seleccionado es una alternativa más simple y consistente con los resultados obtenidos anteriormente.

5. Análisis de Componentes Principales (PCA)

Con el fin de explorar la estructura multivariada del conjunto de datos y reducir la dimensionalidad del espacio de variables cuantitativas, se aplicó el Análisis de Componentes Principales (PCA). Esta técnica permite resumir la información contenida en múltiples variables correlacionadas mediante un conjunto reducido de componentes no correlacionados entre sí.

Dado que las variables cuantitativas presentan distintas escalas de medición, se estandarizaron estas variables a que tengan una media de cero y desviación estándar de uno, esto equivale a utilizar PCA a partir de la matriz de correlación.

	Componentes Principales				
	PC1	**PC2**	**PC3**	**PC4**	**PC5**
Estadísticos de los componentes					
Eigenvalor	1,791 988	1,010 838	1,001 437	0,989 227	0,206 510
% Varianza	35,839 77	20,216 75	20,028 74	19,784 54	4,130 20
% Varianza acumulada	35,839 77	56,056 52	76,085 26	95,869 80	100,000 00
Eigenvectores					
Edad	0,011 097	0,377 637	0,788 564	0,485 216	−0,000 550
Horas estudio	0,639 938	−0,381 818	0,067 079	0,172 786	−0,640 578
Asistencia clases	0,275 730	0,644 591	0,088 314	−0,651 822	−0,275 326
Horas sueño	0,117 831	0,544 153	−0,604 833	0,556 626	−0,119 825
Puntaje examen	0,707 421	−0,002 407	−0,006 729	−0,002 569	0,706 752

Nota: Eigenvalores >1 indican componentes que explican más varianza que una variable individual.

Cuadro 7: Resultados completos del Análisis de Componentes Principales (PCA)

Se utilizó un scree plot y la regla de kaiser para determinar cuántos componentes principales usar.

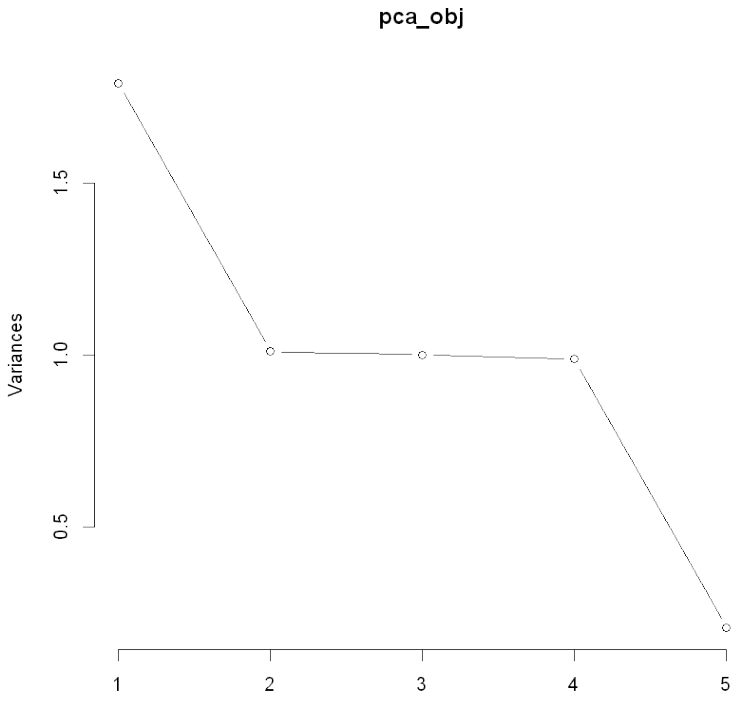

Figura 8: Gráfica del codo para determinar cantidad de componentes principales.

El codo del gráfico está marcado en dos, pero con solo dos componentes se explica únicamente el 56 % de la variabilidad. La regla de Kaiser dice que una forma de escoger la cantidad de componentes principales, es escoger aquellas que tengan un eigenvalor mayor a uno (> 1). Tomando esto en cuenta, podemos usar las primeras tres componentes, llegando a un 76 % de la variabilidad total explicada, que es un valor razonable.

El primer componente principal está fuertemente asociado con el desempeño académico y el nivel de dedicación al estudio. Valores altos en este componente corresponden a estudiantes con mayores horas de estudio y mejores calificaciones. El segundo componente refleja hábitos de organización y asistencia, así como patrones de descanso, diferenciando a estudiantes con mayor regularidad académica. El tercer componente captura variabilidad asociada a características personales, particularmente la edad y los patrones de sueño, y representa información no explicada por los componentes académicos principales.

5.0.1. Visualización

Primero se utilizó un loadingplot para visualizar las relaciones entre las variables. Posteriormente, se creó un biplot para ver tanto las variables como las observaciones. Al tener una cantidad muy grande de observaciones, se complica la interpretación de plots multivariados como este, y de otros tipos.

Se observa que las horas de estudio y la calificación del examen presentan una fuerte contribución al primer componente, el cual puede interpretarse como un eje de desempeño y dedicación académica. Mientras que el segundo componente está asociado principalmente con la asistencia a clase y las horas de sueño, reflejando hábitos de organización y descanso. La variable edad presenta

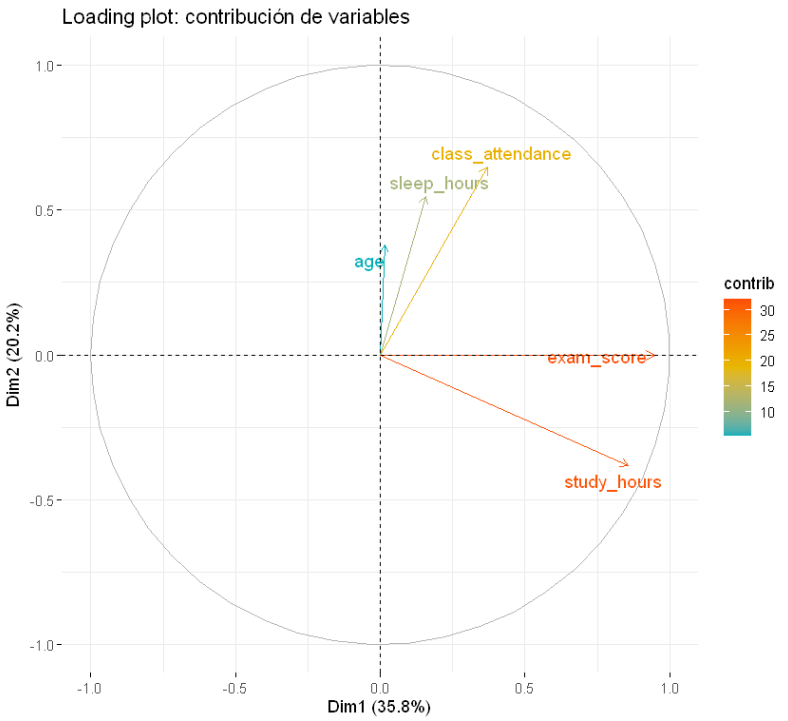

Figura 9: Loading plot para PCA.

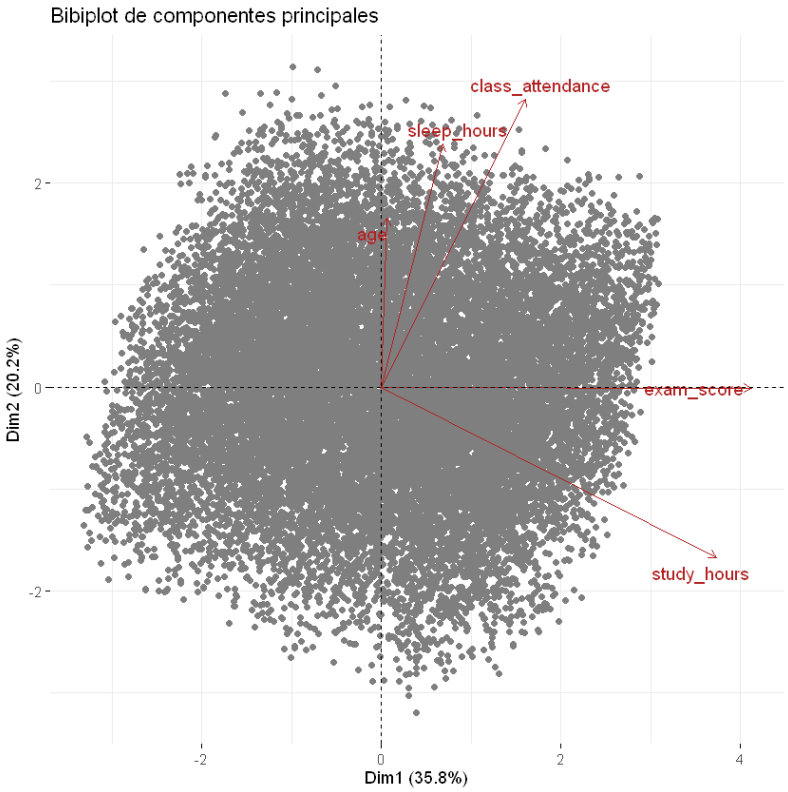

Figura 10: Biplot para PCA.

una contribución menor, al ubicarse cercana al origen del plano factorial.

La concentración de puntos alrededor del origen indica una alta variabilidad individual, sin una separación clara en grupos definidos, lo cual sugiere que los perfiles de estudiantes se distribuyen de manera continua en el espacio de los componentes principales.

6. Clustering

Con el objetivo de identificar patrones y perfiles de estudiantes con características similares, se aplicó un análisis de agrupamiento (clustering). A diferencia de la regresión lineal, cuyo enfoque es inferencial y explicativo, el clustering permite explorar la estructura interna de los datos sin definir una variable respuesta, facilitando la identificación de grupos homogéneos dentro del conjunto de observaciones.

Para este análisis se consideraron únicamente las variables cuantitativas, que fueron estandarizadas desde la sección anterior, para garantizar que todas contribuyeran de maera equilibrada al cálculo de distancias.

Se utillizó clustering jerárquico para crear un dendrograma, con el propósito de facilitar la determinación de un número de clusters (k), para después aplicar el método de KMeans. También se aplicó el método del codo con el mismo propósito, determinar un número adecuado de grupos.

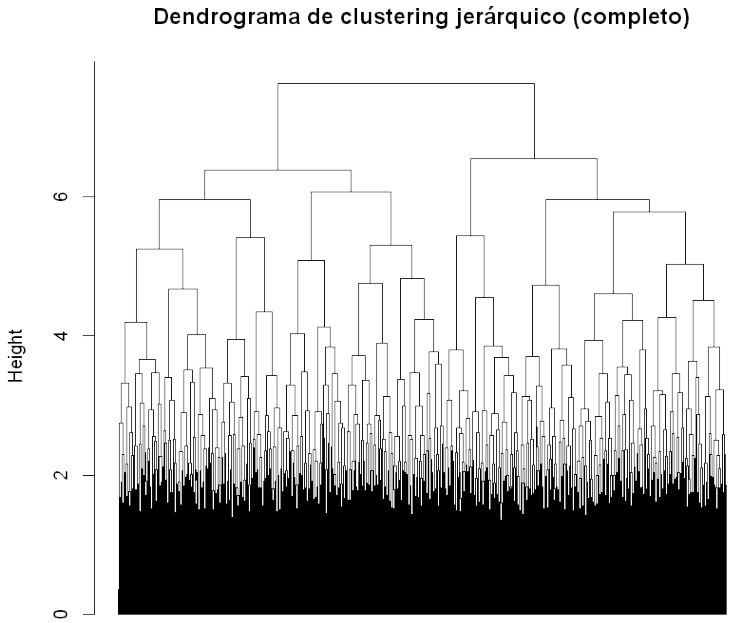

Figura 11: Dendrograma de clustering jerárquico con método de encadenamiento completo.

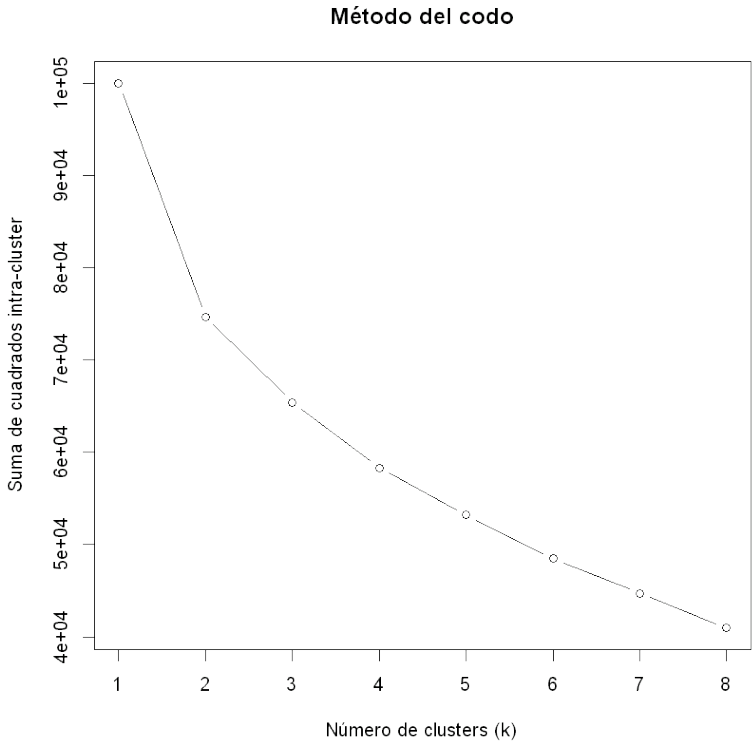

Figura 12: Gráfica de codo para KMeans.

Se identificaron dos clusters. Después de aplicar el algoritmo de KMeans con $k = 2$, se obtuvieron los siguientes centroides (regresados a sus escalas originales):

Cuadro 8: Perfil de clusters: Características promedio por grupo

Variable	**Cluster 1**	**Cluster 2**
Edad	20.50	20.45
Horas de estudio	5.81	2.15
Asistencia a clases (%)	73.16	66.78
Horas de sueño	7.13	6.89
Puntaje examen	76.92	47.68

A partir de estos centroides, el perfil de los estudiantes que forman parte del cluster 1 es de aquello que estudian durante más horas, tienen un porcentaje de asistencias y cantidad de horas de sueño ligeramente más altos y coom consecuencia de esto obtienen calificaciones notablemente más altas. Mientras que los estudiantes dentro del cluster 2 reciben, en promedio, calificaciones reprobatorias. Esto indica que las horas de estudio juegan un papel muy importante al predecir la calificación de un examen, pues las demás variables consideradas tienen cambios leves entre los clusters.

Se realizó la siguiente gráfica, con el propósito de comparar las variables de horas de estudio y asistencia a clases, y cómo afectan a los clusters.

Se observa una separación clara de los estudiantes principalmente a lo largo del eje de las horas de estudio. En particular, uno de los grupos concentra a estudiantes con menores horas de

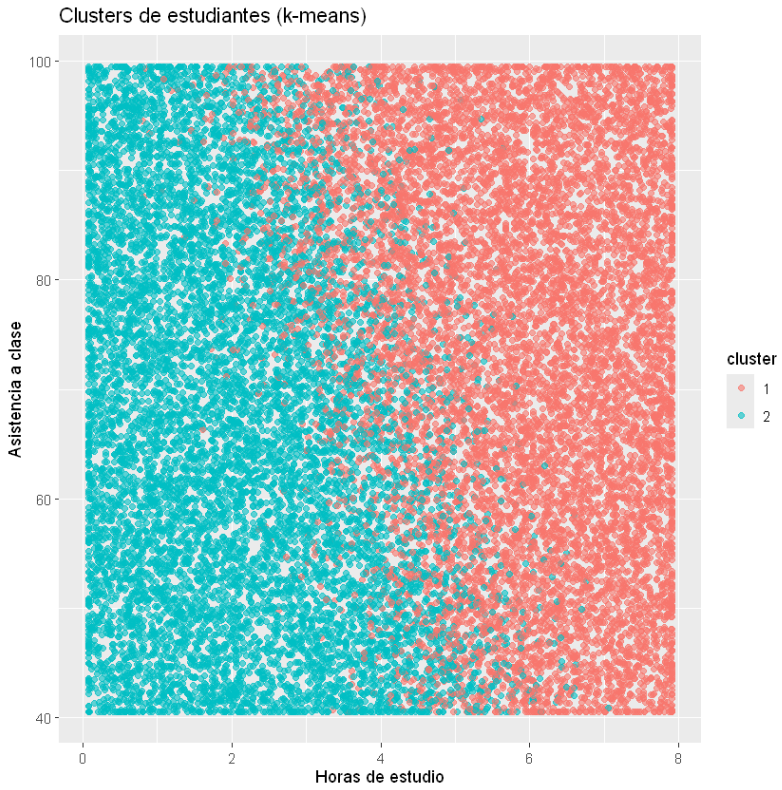

Figura 13: Clusters horas de estudio VS. asistencia a clasees.

estudio, mientras que el otro agrupa a estudiantes con una mayor dedicación al estudio. Por el contrario, la asistencia a clase presenta una distribución similar en ambos clusters, lo que sugiere que su contribución al agrupamiento es secundaria. Estos resultados indican la presencia de perfiles diferenciados de estudiantes en función de sus hábitos de estudio.

7. Conclusiones

En este trabajo se aplicaron diversas técnicas de estadística multivariada con el objetivo de analizar los factores asociados al desempeño académico de los estudiantes. El análisis exploratorio permitió identificar patrones iniciales y relaciones preliminares entre las variables, evidenciando una alta variabilidad individual y destacando la relevancia de los hábitos académicos frente a variables demográficas.

La regresión lineal múltiple mostró que las horas de estudio, la asistencia a clase, las horas de sueño, la calidad del sueño, el método de estudio y la evaluación de las instalaciones presentan un efecto estadísticamente significativo sobre la calificación del examen. En contraste, variables como el género, la edad, el acceso a internet, el curso y la dificultad percibida del examen no mostraron un efecto significativo.

El análisis de componentes principales permitió reducir la dimensionalidad del conjunto de datos y revelar la estructura subyacente de las variables cuantitativas. Los primeros componentes capturaron principalmente información relacionada con el desempeño y la dedicación académica, así como con los hábitos de organización y descanso, lo cual es consistente con los resultados obtenidos en el modelo de regresión.

Por su parte, el análisis de agrupamiento permitió identificar perfiles de estudiantes con características similares, diferenciados principalmente por sus hábitos de estudio. Se observaron diferencias sistemáticas en el desempeño entre los grupos identificados, lo que refuerza la asociación entre los hábitos académicos y el rendimiento. Este enfoque descriptivo complementó los resultados inferenciales y aportó una perspectiva adicional sobre la estructura del conjunto de datos.

En conjunto, los resultados obtenidos mediante las distintas técnicas multivariadas son coherentes entre sí y destacan la importancia de los hábitos académicos y de organización personal como factores clave en el desempeño estudiantil. En este proecto se observó el valor de combinar métodos para obtener una comprensión integral de fenómenos complejos.

8. Bibliografía

Saher, A. (2024). Exam Score Prediction. Kaggle. https://www.kaggle.com/datasets/ayeshaseherr/exame-score-dataset/data

Kaiser Rule - Displayr. (s.f.). https://docs.displayr.com/wiki/Kaiser_Rule

Principal Components (PCA) and Exploratory Factor Analysis (EFA) with SPSS. (s.f.). https://stats.oarc.u spss/

Santibáñez, J. S. (2018, enero). Verificación del supuesto de homocedasticidad. IIMAS UNAM. https://sigma.iimas.unam.mx/jsantibanez/Cursos/Ciencias/2018_1/08_homocedasticidad.html