Análisis de factores exploratorio sobre ejercicio en gimnasios

Sebastián Murillo & Francisco Cordero
ITAM, Inferencia Causal

December 16, 2024

Abstract

Este estudio aplica un análisis de factores exploratorio a datos de rutinas y atributos de usuarios de un gimnasio, con el fin de identificar las dimensiones subyacentes de las características de los usuarios. Partiendo de un conjunto de variables numéricas, e encontraron tres dimensiones principales: intensidad del entrenamiento, atributos físicos y tipo de ejercicio, las cuales explican alrededor del 50% de los datos. Estos resultados aportan claridad sobre cómo se agrupan las variables relevantes del conjunto de datos y el proyecto sirve para entender el tema Análisis de Factores en un contexto aplicado.

1 Introducción

El entrenamiento en el gimnasio es una de las actividades más frecuentes para personas que buscan ejercitarse en la actualidad. La existencia de un lugar donde se puedan realizar distintos tipos de entrenamiento y rutinas da entrada a una pregunta sobre cuál es la estructura que tiene el ejercicio en estas ubicaciones.

Implementando un **análisis de factores exploratorio** se pueden identificar los componentes más importantes que influyen en el ejercicio realizado en el gimnasio, esto con el objetivo de brindar a investigaciones posteriores una descripción general sobre la estructura subyacente en este ámbito.

2 Datos

Se utilizaron datos de distintos usuarios de un gimnasio. Estos datos provienen de la página Kaggle. El conjunto de datos cuenta con 973 observaciones que contienen información sobre las rutinas de ejercicio y atributos físicos de individuos que asisten al gimnasio.

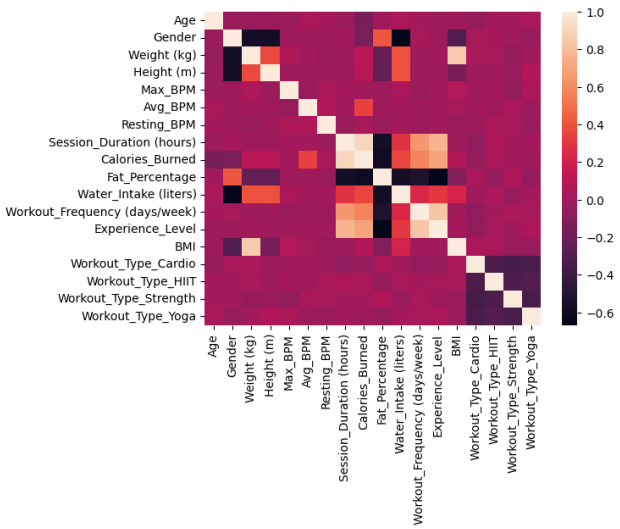
Las variables del conjunto de datos que fueron utilizadas son las siguientes:

- **Age**: Edad del individuo.
- **Gender**: Género del individuo (hombre o mujer).

- **Weight (kg)**: Peso del individuo en kilogramos.
- **Height (m)**: Altura del individuo en metros.
- **Session_Duration (hours)**: Duración de cada sesión de entrenamiento en horas.
- **Calories_Burned**: Total de calorías quemadas durante cada sesión de entrenamiento.
- **Workout_Type**: Tipo de ejercicio realizado. Ej. Cardio, Yoga, etc.
- **Fat_Percentage**: Porcentaje de grasa corporal del individuo.
- **Water_Intake (liters)**: Consumo de agua durante las sesiones de ejercicio. Se mide en litros.
- **Workout_Frequency (days/week)**: Número de sesiones de entrenamiento por semana.
- **Experience_Level**: Nivel de experiencia del individuo. Desde principiante (1) hasta experto (3).

En la Figura 1 se pueden observar las correlaciones de todas las variables presentes en el conjunto de datos. Dada la nula correlación que presentaban las variables **Max_BPM**, **Avg_BPM** y **Resting_BPM**, se decidió no incluirlas para el análisis.

Figure 1: Gráfico de correlación



De igual manera, se decidió eliminar la variable **BMI** dado que se calculaba desde la altura y peso del individuo, por lo que se encontraba altamente correlacionada.

Para poder utilizar la información correctamente del conjunto de datos, se realizaron modificaciones a las variables **Gender** y **Workout_Type**. Se asignaron variables indicadoras a **Gender**, señalizando a los hombres con 0 y a las mujeres con 1. Mientras que para **Workout_Type** se crearon variables indicadoras para tipo de ejercicio entre cardio, fuerza, yoga y alta intensidad (HIIT).

3 Método

A continuación se describe el método de análisis de factores.

El análisis de factores (o FA) es una técnica estadística multivariada que busca describir la variabilidad de un conjunto de variables observadas mediante un número reducido de factores latentes, los cuales no se miden directamente pero resumen la información clave. Esto brinda una representación más simple y ordenada de la estructura subyacente de los datos.

El procedimiento típico del FA implica, en primer lugar, verificar que los datos cumplan dos supuestos importantes: (1) que las variables estén suficientemente correlacionadas (es por eso que decidimos quitar algunas variables como 'MaxBPM', 'Avg BPM' y 'Resting BPM'); y (2) ausencia de multicolinealidad excesiva (por eso quitamos la variable 'BMI'). Además, verificamos que nuestros datos cumplieran con dos pruebas: la de esfericidad de Bartlett, que nos dice si las correlaciones entre las variables no son aleatorias y existe cierta estructura en los datos; y también el índice KMO, que mide qué tan adecuadas son las variables para la factorización. Para ambas pruebas, nuestros datos resultaron ser apropiados para aplicar el método de Análisis de Factores.

Superadas estas condiciones, se debe realizar la extracción de factores y se debe determinar el número de factores a conservar. En nuestro proyecto decidimos realizar la extracción utilizando el método de PCA y determinamos el número de factores consiederando criterios como el diagrama de Scree (buscando el "codo" de la gráfica) y la proporción de varianza explicada.

Después de extraer los factores se puede aplicar una rotación para obtener una estructura más interpretable de los datos. En lo particular de este proyecto, aplicamos la rotación "varimax", que intenta maximizar la varianza de los *factor loadings* dentro de cada factor, brindando una estructura más simple y "limpia". Así podemos producir factores en los que cada variable posee una carga alta en un único factor y cargas cercanas a cero en el resto para identificar cuáles variables afectas más en cada factor.

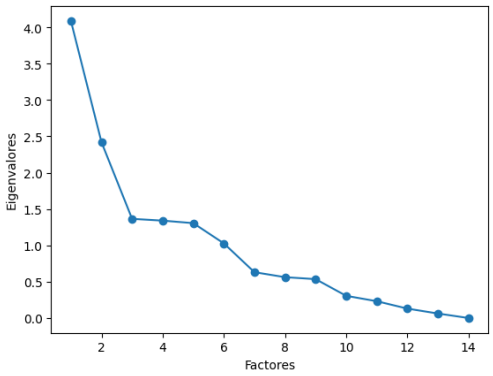
De esta manera, tenemos una mayor interpretación de los loadings y podemos asignar fácilmente un significado conceptual a cada factor, que es el último paso del FA.

En síntesis, el FA permite revelar patrones latentes en datos complejos, ayudando a condensar la información y guiar la generación de hipótesis sobre la estructura interna de los datos estudiados. Esto resulta particularmente útil en el contexto de nuestro análisis, pues permite extraer factores que caractericen las dinámicas del entrenamiento en el gimnasio a partir de las variables originales.

4 Resultados

Para obtener el número adecuado de factores a retener se calcularon sus eigenvalores para poder trazar el gráfico de Scree mostrado en la Figura 2. Este gráfico ayuda a analizar visualmente el aporte de cada factor a la explicación de la varianza en los datos originales.

Figure 2: Gráfico de Scree



En la Figura 2 se puede observar que el número óptimo de factores a conservar son 3, ya que luego de esto al añadir factores adicionales el incremento en ganancia de información deja de ser relevante.

De esta forma, se ajustó el modelo con 3 factores. Las aportaciones de cada variable a cada factor se pueden observar en la siguiente tabla.

Tabla 1

Variable	Factor 1	Factor 2	Factor 3
Age	-0.041643	0.008481	0.014072
Gender	0.025147	**-0.974287**	-0.012231
Weight (kg)	-0.019320	**0.585904**	-0.040423
Height (m)	-0.023094	**0.592848**	-0.020041
Session_Duration (hours)	**0.903517**	0.012460	0.019557
Calories_Burned	**0.828407**	**0.161715**	0.023052
Fat_Percentage	**-0.672982**	**-0.449072**	0.034690
Water_Intake (liters)	**0.317667**	**0.701280**	-0.016464
Workout_Frequency (days/week)	**0.781251**	-0.000885	0.029672
Experience_Level	**0.914076**	0.025945	0.000349
Workout_Type_Cardio	-0.064056	-0.012219	**-0.219027**
Workout_Type_HIIT	0.048285	-0.003031	**-0.178678**
Workout_Type_Strength	-0.011253	-0.013032	**1.035779**
Workout_Type_Yoga	0.024160	0.035058	**-0.194584**

Los factores descubiertos pueden ser descritos como las siguientes características:

- **Intensidad de entrenamiento**: El Factor 1 se ve más afectado por variables como la duración de las sesiones y frecuencia del ejercicio. Por lo que podemos intuir que se debe a que el primer factor es la intensidad del entrenamiento. Este factor por sí mismo acumula un 25.09% de la varianza original de los datos.

- **Atributos físicos**: El Factor 2 tiene mayor relación con los atributos físicos de cada individuo, tales como son el género, altura o peso. Este factor explica 16.90% de la varianza de los datos.

- **Tipo de entrenamiento**: El Factor 3 está relacionado con el tipo de entrenamiento que realiza cada individuo. Es el factor que menor varianza acumula con 8.54%.

Con estos factores, se obtiene una varianza acumulada del 50.53%.

5 Conclusiones

Con los resultados obtenidos, podemos concluir que este análisis de factores puede ayudar a vislumbrar la estructura que siguen los entrenamientos en el gimnasio, no obstante, dado que la varianza acumulada por los 3 factores fue de 50.53% queda claro que el análisis de factores es deficiente a la hora describir por completo la estructura mencionada.

Con los 3 factores descritos se puede explicar la mitad de la varianza de los datos sobre entrenamientos en gimnasios, no obstante, se debe considerar que el recolectar más datos sobre la intensidad de ciertos ejercicios podría incrementar la precisión del análisis, esto ya que para alguien que entrene fuerza puede que pase una hora ejercitándose pero no es lo mismo pasar ese tiempo entrenando con cualquier tipo de pesas.

Otro hallazgo interesante fue el descubrir que la variable de edad solo añade ruido y realmente no afecta a nuestros factores, ni siquiera al Factor 2 que se basa en atributos físicos. Esto se puede deber a que para asistir a un gimnasio normalmente se debe de ser mayor de edad, por lo que podemos deducir que para cualquier adulto las características físicas relevantes trascienden la edad.

Este análisis puede ser mejorado en un futuro, pero es un ejercicio muy instructivo sobre cómo funciona el análisis de factores.

6 Referencias

1. Khorasani, V. (Octubre de 2024). *Gym Members Exercise Dataset*. Kaggle. https://www.kaggle.com/datasets/valakhorasani/gym-members-exercise-dataset/data