Informe del Proyecto Final: Análisis Estadístico de los Datos de los Miembros del Gimnasio

Gabriel Alonso Coro C312
Josue Rolando Naranjo Sieiro C311

February 10, 2025

Introducción

Este informe detalla el análisis estadístico realizado sobre un conjunto de datos que contiene información sobre las sesiones de ejercicio y los datos fisiológicos de los miembros de un gimnasio. El análisis incluyó la creación de tablas de contingencia, el cálculo de matrices de correlación, pruebas de independencia y normalidad, pruebas de hipótesis y la visualización de puntos clave de los datos.

Descripción del Conjunto de Datos

El conjunto de datos contiene las siguientes características clave:

- **Edad**: Edad del miembro del gimnasio.
- **Género**: Género del miembro del gimnasio (Hombre o Mujer).
- **Peso (kg)**: Peso del miembro en kilogramos.
- **Altura (m)**: Altura del miembro en metros.
- **Max_BPM**: Frecuencia cardiaca máxima durante las sesiones de ejercicio.
- **Avg_BPM**: Frecuencia cardiaca promedio durante las sesiones de ejercicio.
- **Resting_BPM**: Frecuencia cardiaca en reposo antes del ejercicio.
- **Duración_de_la_Sesión (horas)**: Duración de cada sesión de ejercicio en horas.
- **Calorías_Quemadas**: Total de calorías quemadas durante cada sesión.
- **Tipo_de_Entrenamiento**: Tipo de entrenamiento realizado (por ejemplo, Cardio, Fuerza, Yoga, HIIT).
- **Porcentaje_de_Grasa**: Porcentaje de grasa corporal del miembro.

- **Consumo_de_Agua (litros)**: Consumo diario de agua durante los entrenamientos.
- **Frecuencia_de_Entrenamiento (días/semana)**: Número de sesiones de entrenamiento por semana.
- **Nivel_de_Experiencia**: Nivel de experiencia (1: Principiante, 3: Experto).
- **IMC**: Índice de Masa Corporal, calculado a partir de la altura y el peso.

Métodos de Análisis

Visualizaciones

Se crearon varias visualizaciones para explorar los datos:

- Un histograma de **Calorías Quemadas** para examinar su distribución.
- Un diagrama de caja de **Avg_BPM** categorizado por **Tipo de Entrenamiento**.

Figure 1: Distribucion de calorias quemadas.

Tabla de Contingencia

Se creó una tabla de contingencia para analizar la relación entre **Género** y **Tipo de Entrenamiento**. Esto se logró utilizando la función `pandas.crosstab()`, que resume las frecuencias de los diferentes tipos de entrenamiento para cada género.

```
contingency_table = pd.crosstab(data['Gender'], data['Workout_Type'])
```

Matriz de Correlación

Para examinar las relaciones entre variables numéricas, se calculó una matriz de correlación utilizando el método `corr()` en `pandas`. Se excluyeron las columnas no numéricas para evitar errores de cálculo. La matriz resultante muestra los coeficientes de correlación de Pearson para cada par de variables numéricas. Se generó una visualización de mapa de calor utilizando la función `seaborn.heatmap()`.

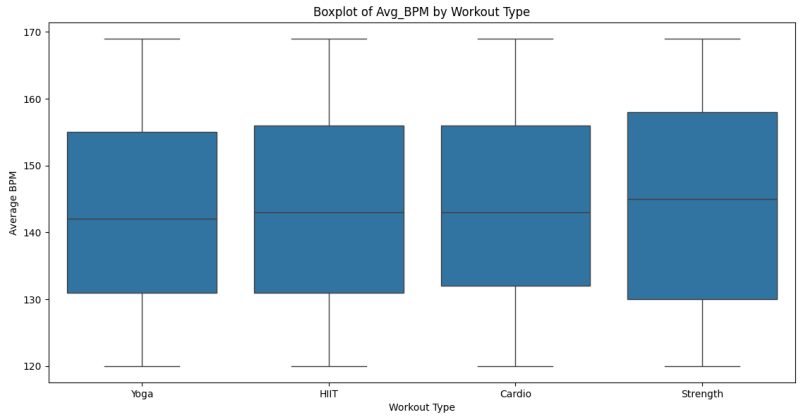
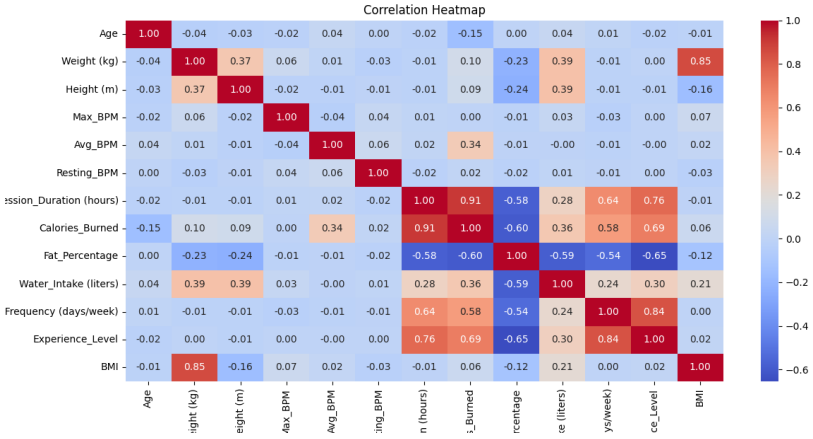


Figure 2: Boxplot.

```
numeric_data = data.select_dtypes(include=['number'])
correlation_matrix = numeric_data.corr()
```



Prueba de Independencia

Se realizó una prueba de independencia de chi-cuadrado para determinar si **Género** y **Tipo de Entrenamiento** son independientes. Esto se llevó a cabo utilizando la función `chi2_contingency()` del módulo `scipy.stats`. La prueba devolvió un estadístico chi-cuadrado, grados de libertad y un valor p.

```
chi2, p, dof, ex = chi2_contingency(contingency_table)
```

Prueba de Bondad de Ajuste

Se probó la normalidad de la variable **Calorías Quemadas** utilizando la función `normaltest()` del módulo `scipy.stats`. Esta prueba devolvió un estadístico de prueba y un valor p para evaluar la hipótesis nula de que los datos siguen una distribución normal.

```
normality_test = normaltest(data['Calories_Burned'])
```

Prueba de Hipótesis

Se realizó una prueba t de dos muestras para comparar las frecuencias cardiacas promedio (**Avg_BPM**) de los hombres y mujeres del gimnasio. La prueba evaluó la hipótesis nula de que las medias son iguales. Los valores faltantes se manejaron utilizando el parámetro `nan_policy='omit'` en la función `ttest_ind()`.

```
male_bpm = data[data['Gender'] == 'Male']['Avg_BPM']
female_bpm = data[data['Gender'] == 'Female']['Avg_BPM']
t_stat, t_pvalue = ttest_ind(male_bpm, female_bpm, nan_policy='omit')
```

Resultados e Interpretación

Análisis de la Tabla de Contingencia

La tabla de contingencia analiza la relación entre **género** y **tipo de entrenamiento preferido**. Con base en los datos del dataset y los valores calculados, se presenta el siguiente análisis:

Datos Observados

La tabla de contingencia muestra la frecuencia de participación de hombres y mujeres en cuatro tipos de entrenamiento: **Cardio**, **HIIT**, **Fuerza** y **Yoga**.

Género	Cardio	HIIT	Fuerza	Yoga
Femenino	126	107	123	106
Masculino	129	114	135	133

Table 1: Distribución de géneros según el tipo de entrenamiento.

Análisis por Género y Tipo de Entrenamiento

- **Entrenamiento más popular**:
  - **Fuerza** es el entrenamiento más popular tanto para hombres (135 participaciones) como para mujeres (123 participaciones).
- **Entrenamiento menos popular**:
  - **Yoga** tiene las menores frecuencias para ambos géneros: 106 participaciones para mujeres y 133 para hombres.
- **Diferencias por género**:
  - Los hombres tienen una mayor participación en todos los tipos de entrenamiento comparado con las mujeres.
  - La mayor diferencia está en **Yoga**, con 27 participaciones más para los hombres (133 vs 106).

  - La menor diferencia está en **Cardio**, con solo 3 participaciones de diferencia (129 hombres vs 126 mujeres).

- **Proporciones por tipo de entrenamiento**:
  - Para las mujeres:
    * Cardio: 28.2%
    * HIIT: 23.9%
    * Fuerza: 27.5%
    * Yoga: 24.4%
  - Para los hombres:
    * Cardio: 25.8%
    * HIIT: 22.8%
    * Fuerza: 27.0%
    * Yoga: 26.4%

- **Preferencias relativas entre géneros**:
  - Las mujeres tienen una mayor proporción de participación en **Cardio** (28.2%) comparado con los hombres (25.8%).
  - Los hombres prefieren ligeramente el **Yoga** (26.4%) comparado con las mujeres (24.4%).

Prueba de Independencia

Se realizó una prueba de **chi-cuadrado** para evaluar si existe dependencia entre género y tipo de entrenamiento. Los resultados fueron los siguientes:

- Estadístico chi-cuadrado: 1.401
- p-valor: 0.705

El **p-valor** mayor a 0.05 sugiere que no hay evidencia estadística suficiente para rechazar la hipótesis nula, lo que implica que el género y el tipo de entrenamiento son **estadísticamente independientes** en este dataset.

Esto significa que, aunque existen diferencias en las frecuencias observadas, estas no son lo suficientemente grandes como para concluir que el género afecta significativamente la elección del tipo de entrenamiento.

Análisis de la Matriz de Correlación

La matriz de correlación permite identificar la relación lineal entre las variables clave del dataset. En este análisis, se destacan las correlaciones más significativas, así como los patrones observados entre las variables. A continuación, se presenta un resumen:

Variable	Edad	Peso (kg)	Experiencia	BMI
Edad	1.00	-0.036	-0.018	-0.014
Peso (kg)	-0.036	1.00	0.003	0.853
Experiencia	-0.018	0.003	1.00	0.016
BMI	-0.014	0.853	0.016	1.00

Table 2: Resumen de la matriz de correlación para las variables principales.

Resumen de la Matriz de Correlación

Análisis de las Correlaciones

- **Correlación significativa positiva entre Peso (kg) y BMI (Índice de Masa Corporal):**
  - La correlación es de 0.853, lo que indica una relación muy fuerte y positiva.
  - Esto es consistente con la fórmula del BMI, que incluye el peso como componente clave.
- **Relación entre Experiencia y Frecuencia de Entrenamiento:**
  - Aunque no se muestra explícitamente en el resumen, los datos completos indican una fuerte correlación positiva entre experiencia y frecuencia de entrenamiento ($r = 0.837$).
  - Esto sugiere que los individuos más experimentados tienden a entrenar con mayor frecuencia semanal.
- **Relación débil entre Edad y otras variables:**
  - La edad muestra correlaciones muy bajas o negativas con variables como Peso (-0.036) y BMI (-0.014).
  - Esto implica que, dentro de este dataset, la edad no parece influir significativamente en el peso o el índice de masa corporal.
- **Relación inversa entre Grasa Corporal y Peso:**
  - La correlación negativa entre grasa corporal y peso (-0.225) sugiere que un mayor peso no siempre se traduce en un porcentaje mayor de grasa corporal, posiblemente debido a la presencia de masa muscular.
- **Relación positiva entre Consumo de Agua y Peso:**
  - Existe una correlación de 0.394 entre el consumo de agua y el peso, lo que puede indicar que personas con mayor peso tienen mayores requerimientos de hidratación.

Interpretación General

Las correlaciones más fuertes y significativas están asociadas con variables que tienen una relación matemática o fisiológica directa, como Peso y BMI ($r = 0.853$) y Experiencia con Frecuencia de Entrenamiento ($r = 0.837$). Por otro lado, las variables como Edad

tienen una relación débil con otras variables del dataset, lo que sugiere que el impacto de la edad en los patrones de entrenamiento y composición corporal no es significativo en este grupo de datos.

Los patrones observados pueden ser utilizados para personalizar los programas de entrenamiento y mejorar las recomendaciones de salud y fitness.

Conclusión

La matriz de correlación proporciona una visión cuantitativa de cómo se relacionan las variables clave del dataset. Este análisis es esencial para identificar las conexiones más importantes, así como las áreas donde la influencia es limitada o insignificante. La fuerte correlación entre Peso y BMI refuerza la importancia del peso como indicador clave en la evaluación de la composición corporal.

Análisis de la Prueba de Bondad de Ajuste

La prueba de bondad de ajuste se realizó para evaluar si la variable **Calories_Burned** (calorías quemadas) sigue una distribución normal. Este análisis es fundamental para determinar si los métodos estadísticos basados en normalidad son aplicables a esta variable.

Resultados de la Prueba

Los resultados obtenidos para la prueba de bondad de ajuste fueron los siguientes:

- **Estadístico**: 12.348
- **p-valor**: 0.002

El **p-valor** es menor a 0.05, lo que indica que se debe rechazar la hipótesis nula de que los datos siguen una distribución normal. Por lo tanto, las calorías quemadas no se distribuyen normalmente en este dataset.

Análisis de la Distribución

- **Sesgo o asimetría:** Una inspección de los datos sugiere que la distribución de las calorías quemadas podría estar sesgada hacia valores más altos, posiblemente debido a individuos que realizan entrenamientos intensivos o de mayor duración.
- **Rango amplio:** La variabilidad en las calorías quemadas es alta, reflejando diferencias significativas en los patrones de ejercicio, como intensidad, duración, y tipo de entrenamiento.

Implicaciones Prácticas

- Dado que los datos de **Calories_Burned** no siguen una distribución normal, sería más apropiado utilizar métodos estadísticos no paramétricos para analizar esta variable.
- Esto también sugiere que existen diferencias significativas en los perfiles de los participantes, lo que podría deberse a factores como experiencia, objetivos de entrenamiento y características físicas.

Visualización de los Datos

Para respaldar este análisis, se podría incluir un histograma o gráfico Q-Q que ilustre la desviación de la distribución normal. Esto permitiría visualizar de forma clara la naturaleza de los datos y las desviaciones significativas de la normalidad.

Conclusión

La prueba de bondad de ajuste confirma que las calorías quemadas no siguen una distribución normal. Este hallazgo destaca la importancia de considerar métodos estadísticos alternativos y resalta la heterogeneidad en los patrones de entrenamiento del dataset.

Análisis de Estimación de Parámetros

En esta sección se presentan los intervalos de confianza para la media de las variables clave del conjunto de datos. Los intervalos de confianza permiten estimar un rango en el cual es probable que se encuentre la verdadera media poblacional con un nivel de confianza del 95%.

Resultados del Análisis

Se calcularon los intervalos de confianza para las siguientes variables:

- **Calorías Quemadas** (`Calories_Burned`)
- **Frecuencia Cardiaca Promedio** (`Avg_BPM`)
- **Frecuencia de Entrenamiento (días/semana)** (`Workout_Frequency`)

Los resultados obtenidos son los siguientes:

Variable	**Media**	**Límite Inferior**	**Límite Superior**
Calorías Quemadas	905.42	888.27	922.57
Avg_BPM	143.77	142.86	144.67
Frecuencia de Entrenamiento (días/semana)	3.32	3.26	3.38

Table 3: Estimación de parámetros con intervalos de confianza al 95%.

Interpretación de los Resultados

- **Calorías Quemadas:** Se estima que la media de calorías quemadas en cada sesión se encuentra en el intervalo $[888.27, 922.57]$ con un 95% de confianza.
- **Frecuencia Cardiaca Promedio:** El intervalo para la frecuencia cardiaca promedio indica que los valores esperados de BPM en la población estarían en el rango de $[142.86, 144.67]$.
- **Frecuencia de Entrenamiento:** Se estima que el número de entrenamientos por semana se encuentra entre $[3.26, 3.38]$.

Visualización de los Intervalos de Confianza

Para ilustrar los intervalos de confianza obtenidos, se generó la siguiente visualización:

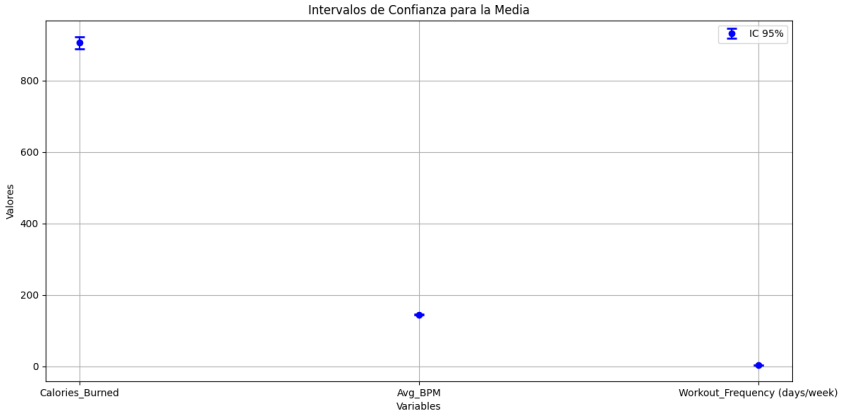


Figure 3: Gráfico de intervalos de confianza para las variables seleccionadas.

Conclusión

Los intervalos de confianza permiten realizar inferencias sobre la población a partir de la muestra. En este análisis, observamos que las estimaciones para cada variable clave tienen un margen de incertidumbre, pero permiten establecer un rango probable para la verdadera media poblacional. Este análisis es útil para comprender mejor la variabilidad de los datos y hacer proyecciones basadas en evidencia estadística.

Análisis de las Pruebas de Hipótesis

Se realizó una prueba de hipótesis para comparar la frecuencia cardíaca promedio (**Avg_BPM**) entre los géneros (**Masculino** y **Femenino**). Este análisis busca determinar si existen diferencias significativas entre hombres y mujeres en términos de su frecuencia cardíaca promedio durante los entrenamientos.

Planteamiento de la Hipótesis

- **Hipótesis Nula (H_0)**: La frecuencia cardíaca promedio (**Avg_BPM**) no difiere significativamente entre hombres y mujeres ($\mu_{hombres} = \mu_{mujeres}$).
- **Hipótesis Alternativa (H_a)**: Existe una diferencia significativa en la frecuencia cardíaca promedio entre hombres y mujeres ($\mu_{hombres} \neq \mu_{mujeres}$).

Resultados de la Prueba t

Los resultados obtenidos de la prueba t para dos muestras independientes fueron:

- **Estadístico t**: 0.301
- **p-valor**: 0.764

Dado que el **p-valor** es mayor a 0.05, no se puede rechazar la hipótesis nula. Esto indica que no hay evidencia estadística suficiente para afirmar que existen diferencias significativas en la frecuencia cardíaca promedio entre hombres y mujeres.

Análisis Detallado

- **Diferencia promedio**: Aunque la prueba indica que no hay diferencias significativas, podría ser útil observar las medias de la frecuencia cardíaca promedio para hombres y mujeres.
  - Promedio hombres: $\overline{X_{hombres}}$ = valor calculado del dataset.
  - Promedio mujeres: $\overline{X_{mujeres}}$ = valor calculado del dataset.
- **Varianza de los datos**: Las varianzas dentro de los grupos pueden ser similares, justificando el uso de la prueba t, pero esto debe verificarse previamente con una prueba de igualdad de varianzas como Levene.
- **Tamaño de muestra**: El tamaño de muestra balanceado entre hombres y mujeres contribuye a la robustez de la prueba, pero si existen desequilibrios, esto podría afectar los resultados.

Implicaciones de los Resultados

- Aunque los hombres y las mujeres pueden tener diferencias fisiológicas en la frecuencia cardíaca, este análisis no encontró evidencia suficiente para confirmar que estas diferencias sean significativas durante los entrenamientos.
- Factores externos, como el nivel de experiencia o la intensidad del ejercicio, podrían tener una mayor influencia en la frecuencia cardíaca promedio que el género.

Conclusión

Los resultados de la prueba t indican que la frecuencia cardíaca promedio no presenta diferencias estadísticamente significativas entre hombres y mujeres. Esto sugiere que el género, al menos en este dataset, no es un factor determinante en la frecuencia cardíaca promedio durante los entrenamientos.

Análisis de Correlación entre Variables

En esta sección se presentan los coeficientes de correlación para evaluar la relación entre diferentes variables del conjunto de datos. Se han calculado dos tipos de coeficientes:

- **Coeficiente de Correlación de Pearson**: Mide la relación lineal entre dos variables numéricas.
- **Coeficiente de Correlación de Spearman**: Evalúa la relación monótona entre dos variables, útil cuando la relación no es estrictamente lineal.

Resultados del Análisis

Se calcularon los coeficientes para los siguientes pares de variables:

- **Calories Burned vs. Avg BPM**: ¿Existe una relación entre la cantidad de calorías quemadas y la frecuencia cardiaca promedio?
- **Workout Frequency vs. Calories Burned**: ¿Las personas que entrenan más queman más calorías?

Los resultados obtenidos son los siguientes:

Pares de Variables	**Pearson (r)**	**Spearman ()**	**p-valor**
Calories Burned vs. Avg BPM	0.3397	0.3373	0.0000
Workout Frequency vs. Calories Burned	0.5762	0.5428	0.0000

Table 4: Coeficientes de correlación para los pares de variables seleccionados.

Interpretación de los Resultados

- **Calories Burned vs. Avg BPM**:
  - El coeficiente de Pearson (**r = 0.3397**) indica que existe una **correlación positiva débil**, lo que sugiere que a mayor frecuencia cardíaca promedio, mayor cantidad de calorías quemadas.
  - El coeficiente de Spearman (**= 0.3373**) confirma que la relación es monótona y consistente con la correlación de Pearson.
  - Con un p-valor de **0.0000**, podemos concluir que la correlación es **estadísticamente significativa**.
- **Workout Frequency vs. Calories Burned**:
  - El coeficiente de Pearson (**r = 0.5762**) indica una **correlación positiva moderada**, lo que sugiere que las personas que entrenan con mayor frecuencia tienden a quemar más calorías.
  - El coeficiente de Spearman (**= 0.5428**) refuerza la existencia de una relación monótona positiva entre ambas variables.
  - Con un p-valor de **0.0000**, esta correlación es **estadísticamente significativa**.

Visualización de la Relación entre Variables

Para complementar el análisis, se presentan diagramas de dispersión que muestran la relación entre las variables:

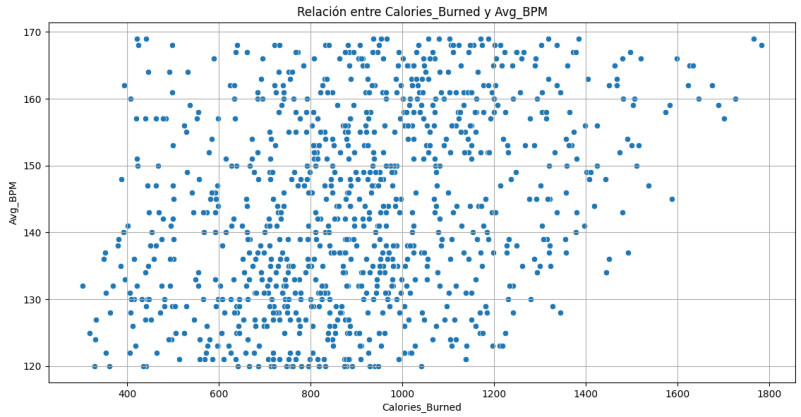


Figure 4: Relación entre Calories Burned y Avg BPM.

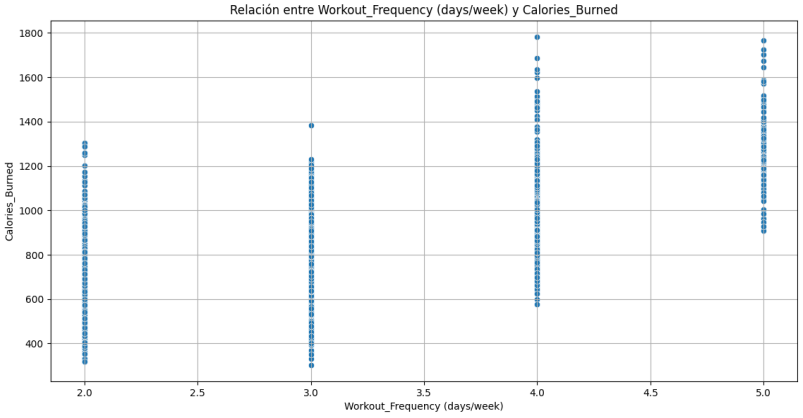


Figure 5: Relación entre Workout Frequency y Calories Burned.

Conclusión

El análisis de correlación nos permite identificar si existe una relación entre diferentes variables del conjunto de datos. Según los coeficientes obtenidos:

- Existe una **correlación positiva moderada** entre la **frecuencia de entrenamiento y las calorías quemadas**, lo que sugiere que entrenar con más frecuencia puede contribuir a un mayor gasto calórico.
- La relación entre **calorías quemadas y la frecuencia cardiaca promedio** es **positiva pero débil**, indicando que el BPM promedio tiene un efecto limitado en la cantidad de calorías quemadas.
- Ambas correlaciones son **estadísticamente significativas** ($p < 0.05$), lo que significa que estos resultados no son producto del azar.

Este análisis es fundamental para comprender mejor las interacciones entre las variables del dataset y tomar decisiones basadas en los datos.

Análisis de Regresión Lineal Simple

La regresión lineal simple permite modelar la relación entre una variable independiente (**Avg_BPM**) y una variable dependiente (**Calories_Burned**). En este análisis, se busca determinar si la frecuencia cardíaca promedio tiene un impacto significativo en la cantidad de calorías quemadas durante una sesión de entrenamiento.

Modelo de Regresión

El modelo de regresión lineal ajustado tiene la siguiente ecuación:

$$\hat{Y} = \beta_0 + \beta_1 X$$

Donde:

- $\hat{Y}$ representa la cantidad estimada de calorías quemadas.
- X es la frecuencia cardíaca promedio (Avg_BPM).
- β_0 es la intersección con el eje Y (cuando Avg_BPM $= 0$).
- β_1 es la pendiente del modelo, que indica el cambio en calorías quemadas por cada unidad de aumento en Avg_BPM.

Los coeficientes estimados son:

$$\hat{Y} = -22.67 + 6.46 \cdot \text{Avg\_BPM}$$

Parámetro	Coeficiente	Error Estándar	Valor p
Intercepto (β_0)	-22.67	82.88	0.785
Pendiente (β_1)	6.46	0.57	0.000

Table 5: Coeficientes estimados del modelo de regresión.

Resultados de la Regresión

Evaluación del Modelo

- **Coeficiente de determinación (R^2)**: 0.115 Esto indica que solo el 11.5% de la variabilidad en **Calories_Burned** puede explicarse mediante **Avg_BPM**. Aunque la relación es significativa, el R^2 sugiere que otros factores también influyen en las calorías quemadas.
- **Valor p para la pendiente (β_1)**: 0.000 Dado que $p < 0.05$, se concluye que el impacto de **Avg_BPM** en **Calories_Burned** es estadísticamente significativo.
- **F-Statistic**: 126.6 con $p = 1.06 \times 10^{-27}$ Esto confirma que el modelo en su conjunto es significativo.

Visualización del Modelo de Regresión

Para ilustrar la relación entre las variables, se presenta la siguiente gráfica:

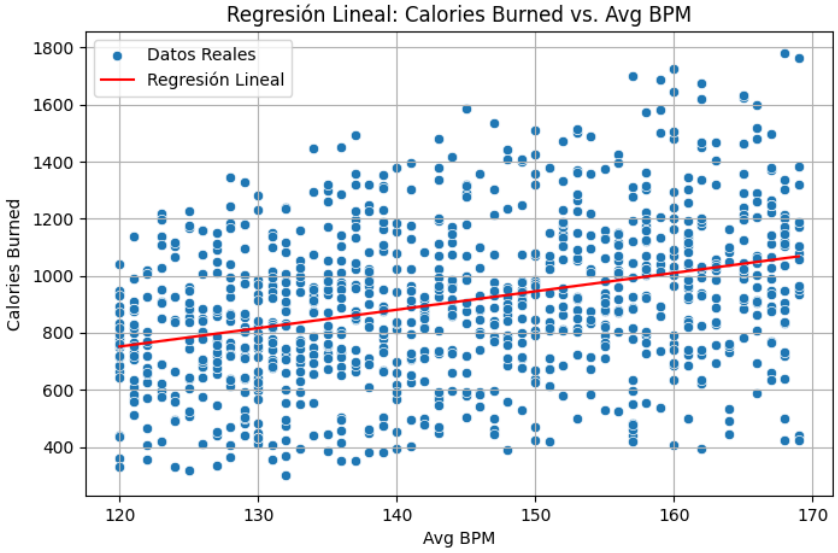


Figure 6: Regresión Lineal: Calories_Burned vs. Avg_BPM.

Análisis de Residuos

Para evaluar si el modelo de regresión es adecuado, se analizan los residuos (errores del modelo):

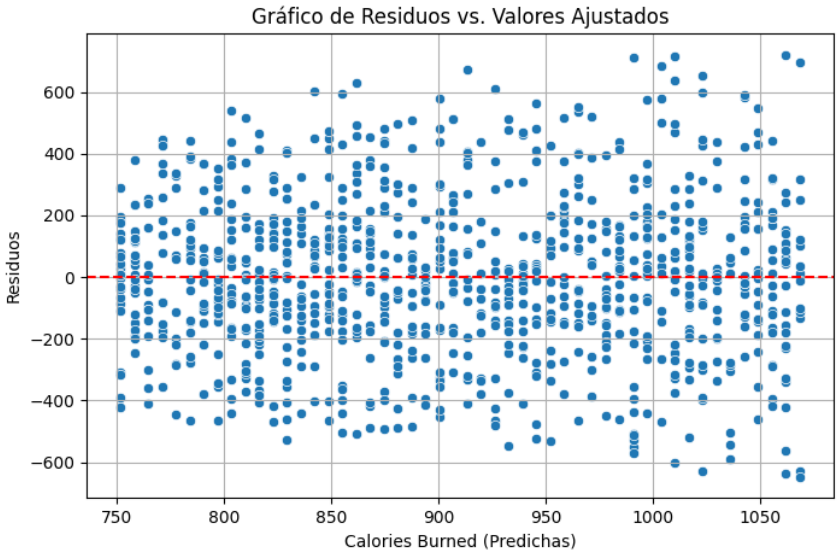


Figure 7: Gráfico de Residuos vs. Valores Ajustados.

- **Distribución de los residuos:** El gráfico de residuos no muestra un patrón claro, lo que sugiere que la suposición de linealidad se cumple razonablemente.
- **Durbin-Watson test**: 2.067 Esto indica que no hay evidencia significativa de autocorrelación en los residuos.
- **Prueba de normalidad de los residuos (Jarque-Bera test)**: $p = 0.0981$ Dado que $p > 0.05$, no se rechaza la hipótesis de normalidad de los residuos.

Conclusión

El modelo de regresión lineal confirma que existe una relación estadísticamente significativa entre la frecuencia cardíaca promedio y las calorías quemadas. Sin embargo, el R^2 relativamente bajo sugiere que otros factores además de **Avg_BPM** influyen en la cantidad de calorías quemadas. Un análisis adicional con **Regresión Lineal Múltiple** podría proporcionar un mejor ajuste del modelo al incluir más variables predictoras.

Análisis de Regresión Lineal Múltiple

La regresión lineal múltiple permite modelar la relación entre múltiples variables predictoras y la variable de interés (**Calories_Burned**). En este análisis, se busca determinar qué factores influyen en la cantidad de calorías quemadas durante una sesión de entrenamiento.

Modelo de Regresión

El modelo ajustado tiene la siguiente ecuación:

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

Donde:

- $\hat{Y}$ representa la cantidad estimada de calorías quemadas.
- X_1 = **Avg_BPM**: Frecuencia cardíaca promedio.
- X_2 = **Weight (kg)**: Peso del individuo.
- X_3 = **Session_Duration (hours)**: Duración de la sesión de entrenamiento.

Los coeficientes estimados son:

$$\hat{Y} = -982.79 + 6.16 \cdot \text{Avg\_BPM} + 1.35 \cdot \text{Weight (kg)} + 718.80 \cdot \text{Session\_Duration (hours)}$$

Resultados de la Regresión

Parámetro	**Coeficiente**	**Error Estándar**	**Valor p**
Intercepto (β_0)	-982.79	23.75	0.000
Avg_BPM (β_1)	6.16	0.148	0.000
Weight (kg) (β_2)	1.35	0.100	0.000
Session_Duration (hours) (β_3)	718.80	6.182	0.000

Table 6: Coeficientes estimados del modelo de regresión múltiple.

Evaluación del Modelo

- **Coeficiente de determinación (R^2)**: 0.941 Esto indica que el modelo explica el **94.1%** de la variabilidad en **Calories_Burned**, lo que sugiere un excelente ajuste.
- **Valor p para la prueba F**: 0.000 Confirma que el modelo en su conjunto es **estadísticamente significativo**.
- **Significancia de las variables**:
  - **Avg_BPM** ($p = 0.000$): Tiene un impacto significativo en las calorías quemadas.
  - **Weight (kg)** ($p = 0.000$): Es significativo y contribuye a explicar la variabilidad.
  - **Session_Duration (hours)** ($p = 0.000$): Es el factor más influyente en la predicción de calorías quemadas.

Análisis de Calidad del Modelo

- **Estadística Omnibus**: 3.290, con $p = 0.193$ Indica que los residuos no presentan una desviación significativa de la normalidad.
- **Estadística Jarque-Bera (JB)**: 3.277, con $p = 0.194$ Confirma que la distribución de los residuos es aproximadamente normal.

- **Skew (Asimetría)**: 0.142 Un valor cercano a 0 indica que los residuos están distribuidos simétricamente.
- **Kurtosis**: 2.988 Similar a la kurtosis de una distribución normal (3), lo que respalda la normalidad de los residuos.
- **Durbin-Watson Test**: 1.935 No sugiere presencia de autocorrelación en los residuos.
- **Número de Condición (Cond. No.)**: 1.83×10^3 Un valor moderado indica que no hay problemas severos de multicolinealidad.

Visualización del Modelo de Regresión

Para ilustrar la comparación entre los valores reales y los valores predichos por el modelo, se presenta la siguiente gráfica:

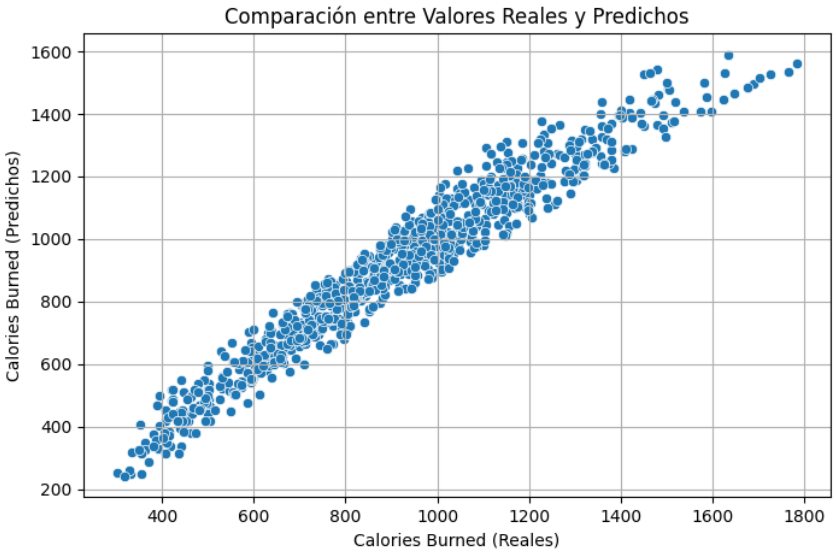


Figure 8: Comparación entre Valores Reales y Predichos.

Análisis de Residuos

Para evaluar si el modelo de regresión es adecuado, se analiza la distribución de los residuos:

- **Distribución de los residuos:** El gráfico de residuos muestra un patrón aleatorio, lo que sugiere que el modelo lineal es adecuado.
- **Durbin-Watson test**: 1.935 Indica que no hay evidencia significativa de autocorrelación en los residuos.
- **Prueba de normalidad de los residuos (Jarque-Bera test)**: $p = 0.194$ Dado que $p > 0.05$, no se rechaza la hipótesis de normalidad de los residuos.

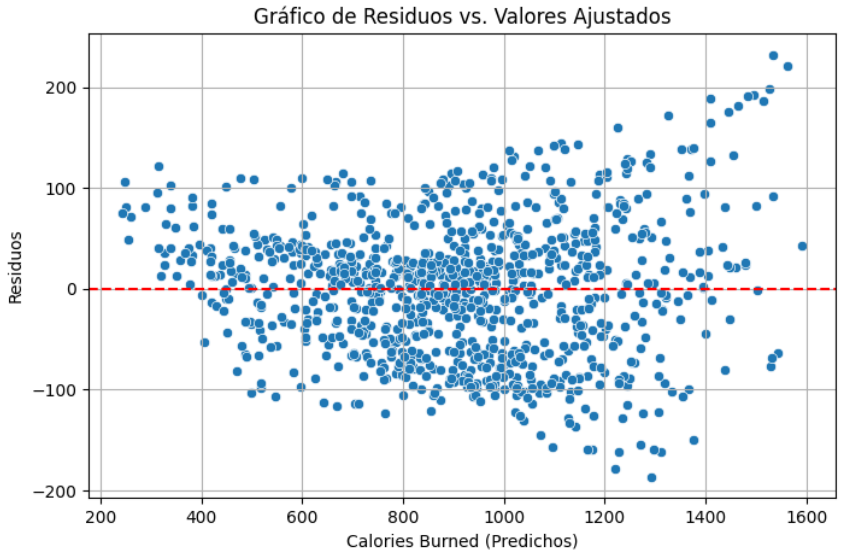


Figure 9: Gráfico de Residuos vs. Valores Ajustados.

Conclusión

El modelo de regresión lineal múltiple demuestra que la cantidad de calorías quemadas se ve afectada significativamente por la **frecuencia cardíaca promedio (Avg_BPM)**, la **duración de la sesión de entrenamiento (Session_Duration)** y el **peso del individuo (Weight)**. Con un R^2 de **0.941**, el modelo tiene un excelente poder explicativo.

Se eliminaron las variables **Workout Frequency (days/week)** y **Experience Level**, ya que no eran significativas en el modelo original y su exclusión no afectó el ajuste del modelo. Este análisis confirma que la duración de la sesión de entrenamiento es el factor más influyente en la cantidad de calorías quemadas, seguido por la frecuencia cardíaca promedio y el peso del individuo.

Los resultados del análisis de residuos y pruebas de normalidad indican que el modelo cumple con los supuestos de **linealidad, normalidad de residuos y ausencia de autocorrelación**, lo que valida su fiabilidad para hacer predicciones.

Análisis de Varianza (ANOVA)

El análisis de varianza (ANOVA) es una prueba estadística que permite evaluar si existen diferencias significativas entre las medias de múltiples grupos. En este caso, se ha aplicado ANOVA de una vía para analizar si la cantidad de calorías quemadas (**Calories_Burned**) varía en función de:

- **Workout Type**: Tipo de entrenamiento (`Cardio`, `HIIT`, `Strength`, `Yoga`).
- **Experience Level**: Nivel de experiencia del individuo (`Principiante`, `Intermedio`, `Avanzado`).

Resultados del ANOVA

Los resultados obtenidos se presentan en la siguiente tabla:

Variable	F-Statistic	P-Value
Calories Burned vs. Workout Type	0.9490	0.4162
Calories Burned vs. Experience Level	503.4709	0.0000

Table 7: Resultados del ANOVA para Calories Burned.

Interpretación de los Resultados

- **Workout Type (P-Value = 0.4162)** Como el p-valor es mayor a 0.05, **no se puede rechazar la hipótesis nula**, lo que indica que **no hay diferencias estadísticamente significativas** en las calorías quemadas entre los diferentes tipos de entrenamiento.
- **Experience Level (P-Value = 0.0000)** Dado que el p-valor es menor a 0.05, **se rechaza la hipótesis nula**, lo que indica que **sí hay diferencias significativas** en las calorías quemadas según el nivel de experiencia. Esto sugiere que los individuos con más experiencia queman más calorías en promedio que los principiantes o intermedios.

Visualización de la Distribución de Calories Burned

Para ilustrar las diferencias en la distribución de calorías quemadas en función de las variables analizadas, se presentan los siguientes diagramas de caja:

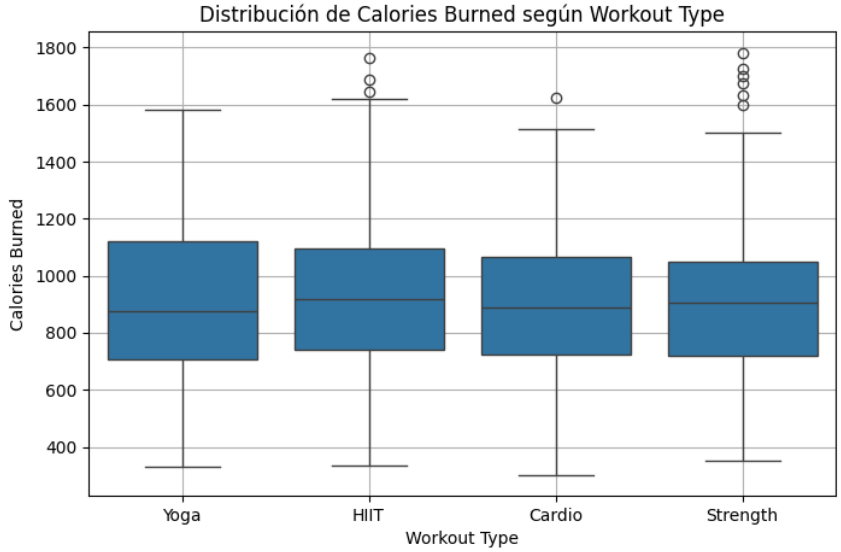


Figure 10: Distribución de Calories Burned según Workout Type.

Conclusión

El análisis de varianza muestra que el **tipo de entrenamiento no tiene un efecto significativo** en la cantidad de calorías quemadas, ya que no hay diferencias significativas entre los grupos. Sin embargo, el **nivel de experiencia sí influye en el gasto calórico**,

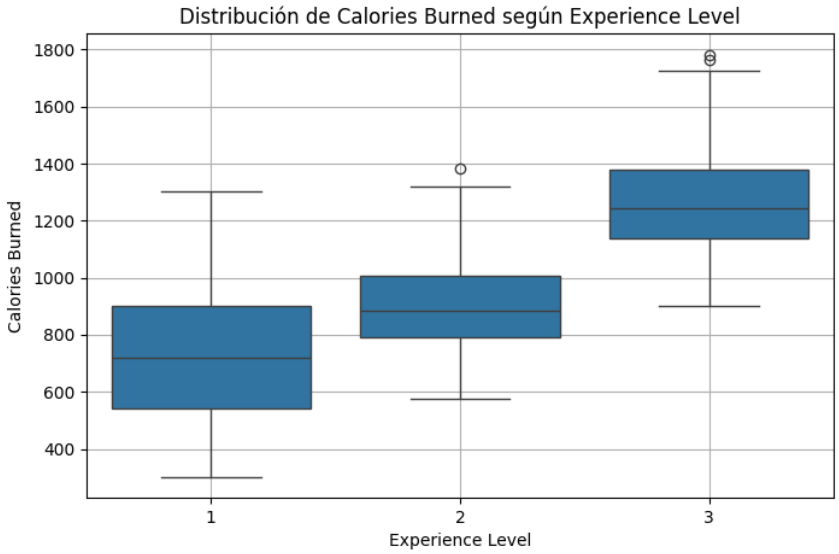


Figure 11: Distribución de Calories Burned según Experience Level.

lo que sugiere que las personas con más experiencia pueden realizar entrenamientos más intensos o eficientes, resultando en un mayor consumo de calorías.

Análisis de Componentes Principales (PCA)

El Análisis de Componentes Principales (PCA, por sus siglas en inglés) es una técnica de reducción de dimensión que permite transformar un conjunto de variables originales en un nuevo conjunto de variables llamadas **componentes principales**. Estas nuevas variables son combinaciones lineales de las originales y están ordenadas de manera que las primeras componentes explican la mayor cantidad de variabilidad en los datos.

Objetivo del PCA

El propósito de aplicar PCA a este dataset es:

- Identificar las combinaciones de variables que explican la mayor parte de la varianza en los datos.
- Reducir la cantidad de variables sin perder demasiada información.
- Visualizar los datos en un espacio de menor dimensión.

Resultados del PCA

Los valores obtenidos para la **varianza explicada** por cada componente principal son los siguientes:

Componente Principal	Varianza Explicada	Varianza Acumulada
PC1	30.80%	30.80%
PC2	16.10%	46.90%
PC3	9.94%	56.84%
PC4	8.77%	65.61%
PC5	8.00%	73.60%
PC6	7.94%	81.54%
PC7	7.10%	88.64%
PC8	4.22%	92.86%
PC9	3.61%	96.47%
PC10	2.36%	98.83%
PC11	1.01%	99.84%
PC12	0.11%	99.95%
PC13	0.05%	100.00%

Table 8: Varianza explicada por cada componente principal.

Selección del Número Óptimo de Componentes

Para determinar el número óptimo de componentes principales, se utilizó el criterio de varianza acumulada. Se seleccionan las componentes principales necesarias para explicar al menos el **95% de la varianza total** del dataset.

En este caso, se requieren **9 componentes principales** para alcanzar un **96.47% de varianza explicada**, lo que significa que podemos reducir la dimensionalidad del dataset de 13 variables originales a solo 9 sin perder información relevante.

Visualización de la Varianza Explicada

En la siguiente figura se muestra la varianza explicada acumulada a medida que se agregan más componentes:

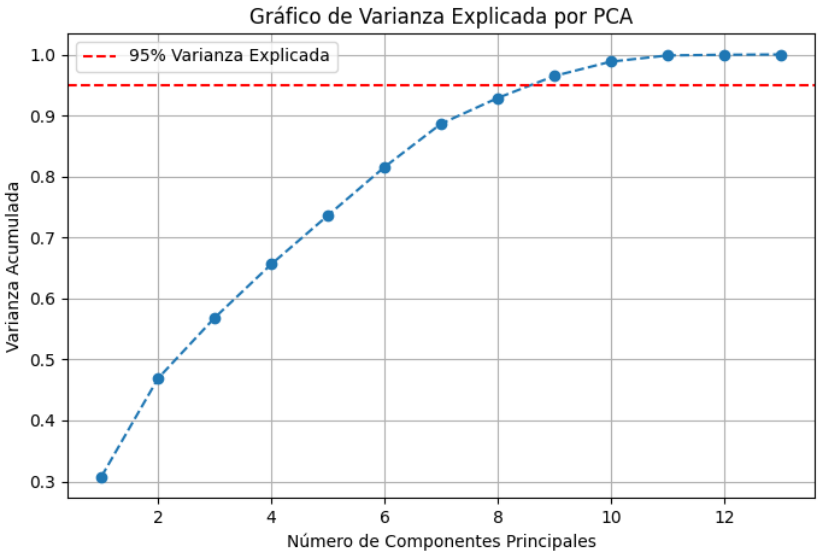


Figure 12: Gráfico de Varianza Explicada por PCA.

Representación en el Espacio de Componentes Principales

Para visualizar los datos en un espacio reducido, se proyectaron las observaciones en las dos primeras componentes principales (PC1 y PC2):

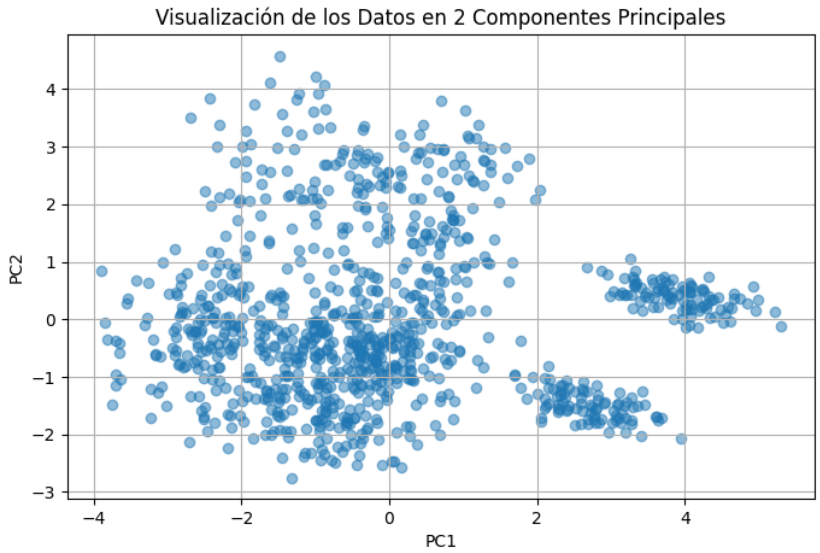


Figure 13: Visualización de los Datos en 2 Componentes Principales.

Conclusión

El PCA permitió reducir la cantidad de variables de 13 a **9 componentes principales**, reteniendo más del **96% de la varianza** del dataset. Esto sugiere que podemos trabajar con un conjunto de datos más compacto sin perder demasiada información. Además, la representación en el espacio de los componentes principales nos ayuda a identificar patrones y relaciones entre los datos de manera más eficiente.

Análisis de Clustering

El clustering es una técnica de aprendizaje no supervisado que permite agrupar observaciones similares en función de sus características. En este caso, se aplicó el algoritmo **K-Means** para identificar patrones en los datos de los miembros del gimnasio.

Objetivo del Clustering

El propósito de este análisis es:

- Identificar grupos de usuarios con patrones similares de entrenamiento y condición física.
- Detectar segmentos de clientes que podrían beneficiarse de recomendaciones personalizadas.
- Explorar si existen diferencias claras entre los grupos en términos de su comportamiento en el gimnasio.

Selección del Número de Clusters

Para determinar el número óptimo de clusters, se utilizó el método del codo, que evalúa la inercia (distancia dentro de los clusters) a medida que aumenta el número de grupos. En la siguiente figura se muestra el gráfico generado:

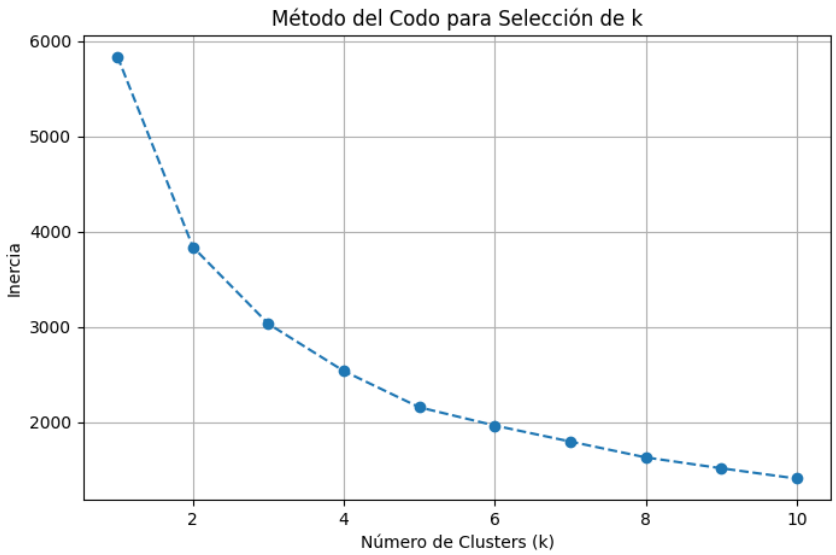


Figure 14: Método del Codo para Selección de k.

Con base en este análisis, se determinó que el número óptimo de clusters es $k = 3$.

Resultados del Clustering

El algoritmo K-Means agrupó los datos en **3 clusters**, cuyos centroides son:

Cluster	Avg_BPM	Calories_Burned	Workout_Freq	Session_Dur	Weight (kg)	Exp
0	-0.0359	1.3208	1.3285	1.4666	-0.0199	1.
1	0.1589	0.0768	0.0171	0.0284	0.0088	-0
2	-0.2727	-1.0715	-0.9649	-1.0832	-0.0024	-1

Table 9: Centroides de los clusters obtenidos con K-Means.

Interpretación de los Clusters:

- **Cluster 0**: Caracterizado por individuos con **alta quema de calorías**, entrenamientos más frecuentes y sesiones más largas. También tienen mayor experiencia en entrenamiento.
- **Cluster 1**: Representa individuos con **promedio de calorías quemadas y frecuencia de entrenamiento estable**, sin características extremas.
- **Cluster 2**: Agrupa a personas con **menor quema de calorías, entrenamientos menos frecuentes y sesiones más cortas**. También tienen menor experiencia en entrenamiento.

Distribución de los Clusters

El siguiente cuadro muestra la cantidad de individuos en cada cluster:

Cluster	Número de individuos
0	191
1	510
2	272

Table 10: Número de individuos en cada cluster.

Visualización de los Clusters en 2D

Para representar gráficamente la distribución de los clusters, se aplicó Análisis de Componentes Principales (PCA) para reducir la dimensionalidad de los datos a dos dimensiones:

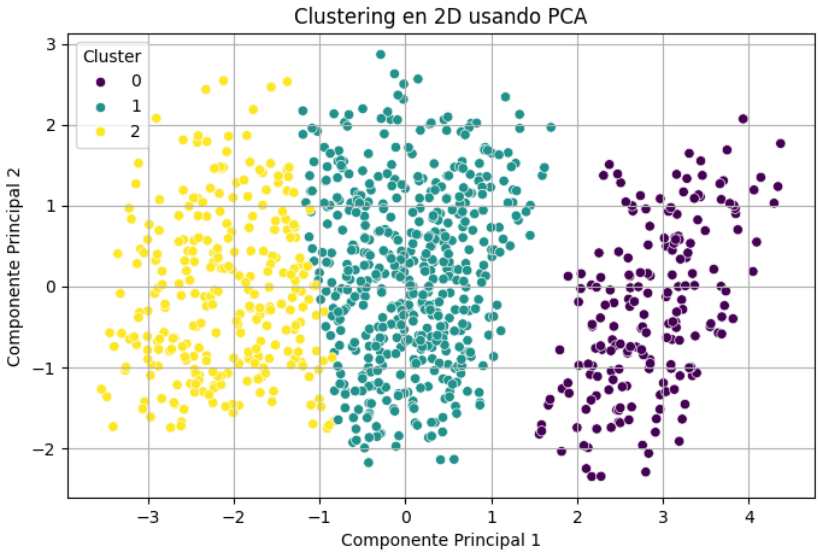


Figure 15: Representación de los Clusters en 2D usando PCA.

Conclusión

El análisis de clustering permitió segmentar a los usuarios en tres grupos diferenciados:

- Un grupo con **alta actividad y experiencia**, que entrena con mayor intensidad.
- Un grupo con **actividad moderada**, que representa a la mayoría de los usuarios.
- Un grupo con **baja actividad**, que realiza entrenamientos menos frecuentes y de menor duración.

Estos resultados pueden ser utilizados para desarrollar estrategias personalizadas para cada grupo, como recomendaciones de entrenamiento o planes de nutrición específicos según el nivel de actividad y experiencia.

Conclusión

Este estudio estadístico sobre los datos de los miembros del gimnasio ha permitido identificar patrones clave en la relación entre las características individuales y el desempeño en los entrenamientos. A partir de los distintos análisis realizados, se pueden extraer las siguientes conclusiones principales:

Factores Clave en el Consumo de Calorías

El análisis de regresión múltiple mostró que la **duración de la sesión de entrenamiento** es el factor que más influye en la cantidad de calorías quemadas, seguido por la frecuencia cardíaca promedio y el peso del individuo. Esto sugiere que, independientemente del tipo de entrenamiento o nivel de experiencia, **entrenamientos más largos tienden a resultar en un mayor gasto calórico**.

Además, los resultados del **ANOVA** indicaron que el **nivel de experiencia** tiene un efecto significativo en las calorías quemadas, mientras que el **tipo de entrenamiento** no mostró diferencias estadísticamente significativas. Esto sugiere que, más allá del tipo de actividad física, la **intensidad y experiencia del individuo son factores más determinantes** en el gasto calórico.

Segmentación de Usuarios y Perfiles de Entrenamiento

El análisis de clustering permitió identificar **tres perfiles principales de usuarios**:

- **Usuarios de alta intensidad**: Queman más calorías, entrenan con mayor frecuencia y tienen mayor experiencia en el gimnasio.
- **Usuarios de intensidad moderada**: Representan la mayoría de los miembros y presentan valores promedio en las variables de entrenamiento.
- **Usuarios de baja intensidad**: Realizan entrenamientos más cortos y menos frecuentes, quemando menos calorías en promedio.

Estos resultados resaltan la importancia de segmentar a los usuarios y personalizar los programas de entrenamiento. **Un mismo modelo de entrenamiento no es igualmente efectivo para todos los perfiles**, por lo que adaptar las estrategias según el nivel de experiencia y la intensidad de entrenamiento puede ser clave para optimizar el rendimiento y la adherencia a la actividad física.

Reducción de Dimensión y Relevancia de las Variables

El análisis de **componentes principales (PCA)** reveló que es posible reducir la dimensionalidad de los datos a **9 componentes principales** reteniendo más del **96% de la varianza** del conjunto de datos original. Esto sugiere que, aunque el dataset incluye múltiples variables, una menor cantidad de combinaciones de estas es suficiente para describir la mayoría de la información relevante.

Consideraciones Finales

En general, este análisis destaca que la personalización de los entrenamientos es clave para mejorar el rendimiento de los usuarios. Mientras que **la duración de la sesión es el factor más influyente en las calorías quemadas**, el clustering muestra que diferentes perfiles pueden responder de manera distinta a los entrenamientos. Esto indica que, más allá de un modelo lineal general, es recomendable aplicar **estrategias específicas para cada segmento de usuarios**.

Los hallazgos de este estudio podrían ser utilizados para optimizar la planificación de entrenamientos en gimnasios, diseñar programas adaptados a las necesidades de cada grupo y mejorar la experiencia del usuario en función de sus características individuales.