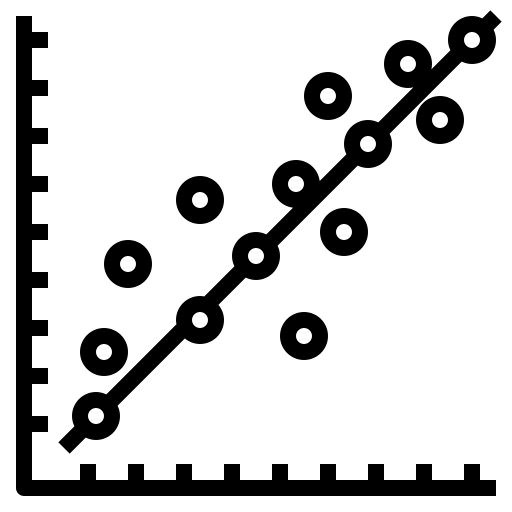
**UNIVERSIDAD MAYOR DE SAN ANDRÉS FACULTAD DE CIENCIAS PURAS Y NATURALES**

**CARRERA DE INFORMÁTICA**



****

**REGRESIÓN LINEAL**

**PREDICTOR DE CALORIAS QUEMADAS**

**Realizado por:**

UNIV. RONALDO ELVIN YUPANQUI SALINAS

La Paz – Bolivia

2024

**Análisis, Modelado y Optimización de Datos del Gym Members Exercise Dataset mediante Técnicas de Machine Learning y Estadística**

**1. Introducción**

En este estudio se aborda el análisis del **Gym Members Exercise Dataset**, un conjunto de datos que contiene información detallada sobre atributos físicos, rutinas de ejercicio y métricas de rendimiento de miembros de un gimnasio. Consta de **973 registros** con variables clave como edad, peso, frecuencia cardíaca, duración del ejercicio, calorías quemadas y tipo de actividad física.

El principal objetivo es **desarrollar un modelo predictivo de calorías quemadas** utilizando técnicas de **Machine Learning**, como:

1. **Descripción General Del Dataset**
2. **Preprocesamiento de datos**.
3. **Regresión lineal múltiple**.
4. **Análisis de Componentes Principales (PCA)**.
5. **Agrupamiento con K-means**.
6. **Optimización mediante Recocido Simulado en Las N - Reinas**.

Finalmente, el estudio culmina con la implementación del modelo de regresión lineal para predecir con alta precisión las calorías quemadas durante sesiones de ejercicio, facilitando así la personalización de programas de entrenamiento.

**3. Descripción General Del Dataset**

Este conjunto de datos ofrece una visión detallada sobre las rutinas de ejercicio, los atributos físicos y las métricas de rendimiento de los miembros de un gimnasio. Contiene un total de **973 muestras de datos**, abarcando indicadores clave como la frecuencia cardíaca, las calorías quemadas, la duración del ejercicio, el tipo de actividad realizada y otras métricas relacionadas con la salud y el fitness. Además, incluye información demográfica y el nivel de experiencia de los miembros, lo que permite realizar un análisis integral sobre los patrones de ejercicio y la evolución del rendimiento físico en función de variables como la edad, el género y la experiencia.

**Descripción de Variables**

El **Gym Members Exercise Dataset** está compuesto por una serie de variables clave que permiten un análisis profundo y multidimensional. A continuación, se describen las principales características del dataset:

1. **Edad (Age)**: Edad del miembro, expresada en años. Un factor crucial para entender cómo las distintas generaciones abordan su entrenamiento y recuperan su estado físico.
2. **Género (Gender)**: Género del miembro (Masculino o Femenino). Esta variable facilita la comparación de hábitos y resultados entre diferentes géneros.
3. **Peso (Weight)**: Peso del miembro en kilogramos. Este dato es esencial para calcular otros indicadores de salud, como el Índice de Masa Corporal (BMI).
4. **Altura (Height)**: Altura en metros, utilizada para calcular el BMI y comprender las correlaciones entre estatura y rendimiento físico.
5. **Frecuencia Cardíaca Máxima (Max\_BPM)**: El valor máximo de la frecuencia cardíaca durante el ejercicio, en latidos por minuto (BPM). Indica la intensidad de la actividad cardiovascular durante las sesiones.
6. **Frecuencia Cardíaca Promedio (Avg\_BPM)**: El promedio de latidos por minuto a lo largo de la sesión. Un indicador clave de la intensidad sostenida del ejercicio.
7. **Frecuencia Cardíaca en Reposo (Resting\_BPM)**: La frecuencia cardíaca del miembro antes de comenzar cualquier actividad física. Utilizada para evaluar el estado de salud general y la eficiencia del sistema cardiovascular.
8. **Duración de la Sesión (Session\_Duration)**: Duración total de la sesión de ejercicio en horas, proporcionando una medida de la dedicación del miembro a su rutina de entrenamiento.
9. **Calorías Quemadas (Calories\_Burned)**: Total de calorías quemadas durante la sesión de ejercicio. Es un indicador de la intensidad y el tipo de actividad realizada.
10. **Tipo de Ejercicio (Workout\_Type)**: El tipo de actividad realizada (Cardio, Fuerza, Yoga, HIIT, etc.), ofreciendo una visión completa sobre las preferencias de entrenamiento de los miembros.
11. **Porcentaje de Grasa Corporal (Fat\_Percentage)**: El porcentaje de grasa corporal, un indicador clave de la salud física y un factor determinante en los objetivos de entrenamiento.
12. **Ingesta de Agua (Water\_Intake)**: Cantidad de agua consumida en litros, esencial para evaluar la hidratación y su relación con el rendimiento.
13. **Frecuencia de Ejercicio (Workout\_Frequency)**: El número de días por semana que el miembro se dedica al ejercicio, un dato crucial para entender los hábitos y la disciplina de los participantes.
14. **Nivel de Experiencia (Experience\_Level)**: El nivel de experiencia del miembro, clasificado en tres categorías: Principiante (1), Intermedio (2), y Experto (3). Esta variable es fundamental para analizar cómo la experiencia influye en el rendimiento y la elección del tipo de ejercicio.
15. **Índice de Masa Corporal (BMI)**: Calculado a partir de la altura y el peso del miembro. Un indicador estándar para evaluar la relación entre el peso y la estatura, ayudando a determinar la clasificación en términos de salud física.

**Objetivo de Investigación:**

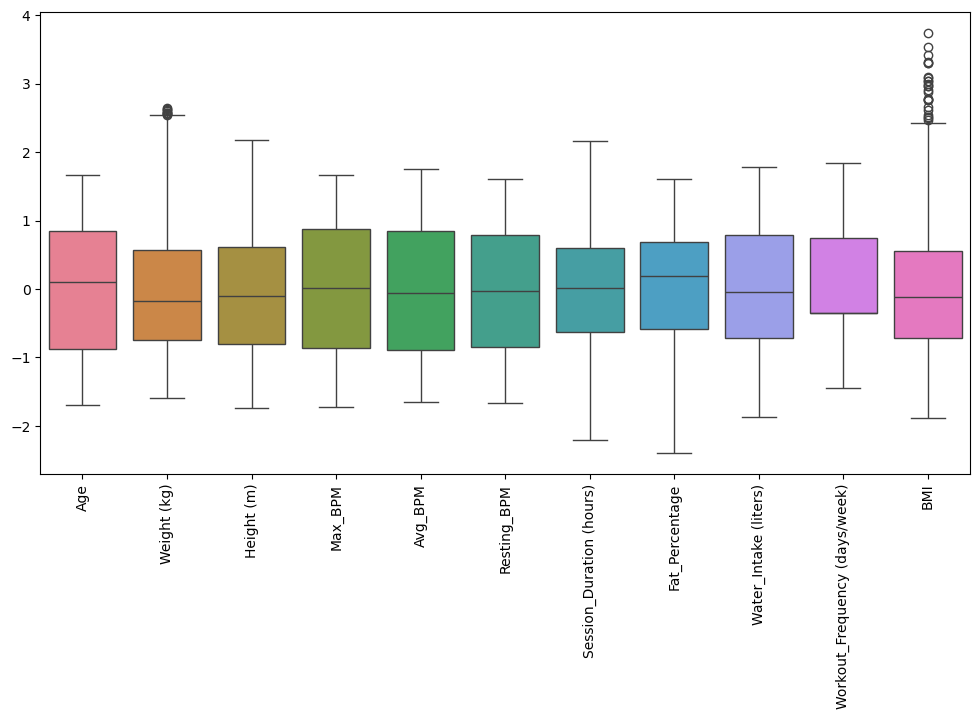
El objetivo de esta investigación es **desarrollar una regresión lineal multiple para predecir las calorías quemadas** durante las sesiones de ejercicio de los miembros de un gimnasio, utilizando el **Gym Members Exercise Dataset**. Este modelo predictivo buscará identificar patrones complejos y relaciones no lineales entre diversas características físicas, demográficas y de entrenamiento (como edad, peso, frecuencia cardíaca, tipo de ejercicio, duración de la sesión, entre otras) y el gasto calórico.

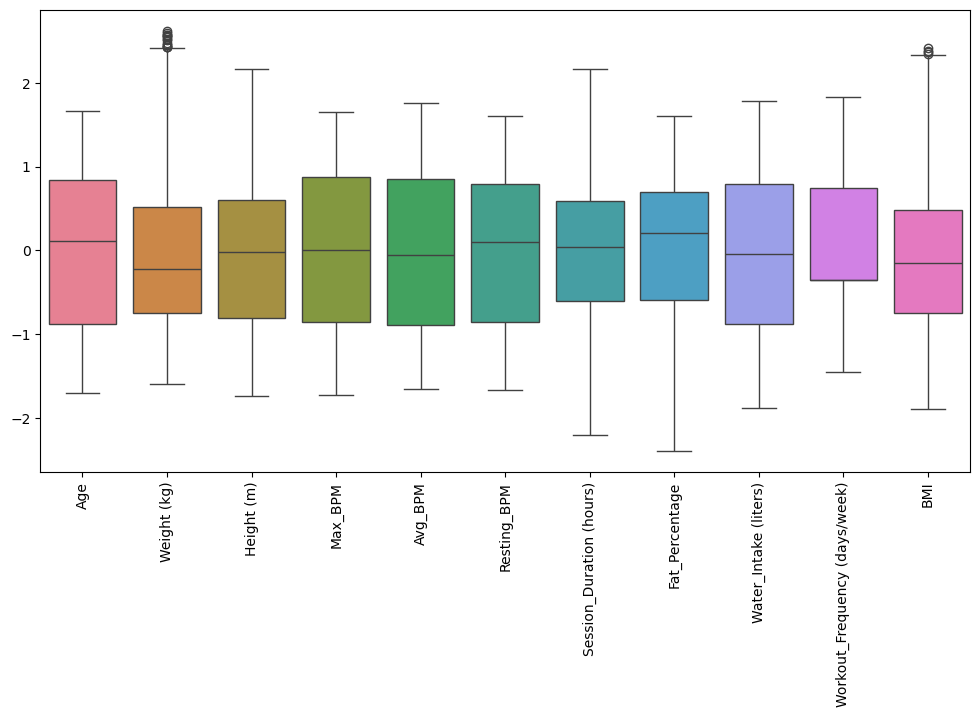
El objetivo específico es generar un modelo de regresión lineal que pueda predecir con alta precisión la cantidad de calorías quemadas en función de estas variables, permitiendo así la personalización de las rutinas de ejercicio y la mejora en la planificación de programas de entrenamiento, basados en un análisis más preciso del rendimiento físico de los miembros del gimnasio.

**Preprocesamiento de Datos**

En el código proporcionado, se llevaron a cabo diversas etapas de preprocesamiento para preparar los datos antes de entrenar el modelo de Regresión Lineal:

1. **Lectura de los datos**: El conjunto de datos se cargó desde un archivo CSV y se inspeccionaron las primeras filas, las estadísticas descriptivas y la información general sobre el dataset utilizando las funciones head(), info() y describe().
2. **Revisión de valores faltantes**: Se realizó un análisis de valores nulos con isnull().sum() para identificar si había columnas con valores faltantes que necesitaran ser eliminados o imputados.
3. **Transformación de variables categóricas**: Las columnas Gender y Workout\_Type se transformaron en variables binarias utilizando get\_dummies(). Esto es necesario para que los modelos de Machine Learning puedan trabajar con datos numéricos en lugar de categóricos.
4. **Escalado de características numéricas**: Se aplicó StandardScaler a las columnas numéricas para asegurar que todas las características estén en la misma escala (media 0 y desviación estándar 1), lo cual es importante para modelos que son sensibles a la escala, como los basados en distancias.
5. **Detección y manejo de valores atípicos (outliers)**: Se identificaron y eliminaron los valores atípicos en la columna BMI utilizando el rango intercuartílico (IQR). Esta técnica garantiza que los datos no contengan valores extremos que puedan afectar el rendimiento del modelo.





1. **División de datos en conjuntos de entrenamiento y prueba**: Los datos fueron divididos en un conjunto de características (X) y la variable objetivo (y), que es Calories\_Burned. A continuación, se dividió el dataset en entrenamiento y prueba usando train\_test\_split(), con un 80% de los datos para entrenamiento y un 20% para prueba, además también se dividió el dataset en entrenamiento y prueba usando 50% de los datos para entrenamiento y un 50% para prueba esto para un enfoque académico e investigación.
2. **Exportación de los conjuntos procesados**: Los conjuntos de entrenamiento y prueba se exportaron a archivos CSV para su posterior uso en la fase de modelado.

**Balanceo de Datos**

En este caso, **no se utilizó ninguna técnica de balanceo de datos** porque el objetivo es predecir una variable continua, Calories\_Burned, lo que implica que estamos trabajando con un problema de **regresión**. El balanceo de datos es comúnmente utilizado en problemas de **clasificación**, donde las clases pueden estar desbalanceadas (por ejemplo, en problemas de clasificación binaria o multiclase). Sin embargo, para un problema de regresión, donde la variable objetivo es continua, las técnicas de balanceo como **oversampling** o **undersampling** no son aplicables ni necesarias.

La variable Calories\_Burned no está dividida en clases discretas, sino que es un valor continuo que puede tomar una amplia gama de valores. Por lo tanto, las técnicas de balanceo de clases no son apropiadas en este caso, ya que no existe una categoría desbalanceada que corregir. En su lugar, el modelo de regresión se ajustará a los valores continuos de la variable objetivo sin necesidad de aplicar balanceo.

**3. Regresión Lineal Múltiple**

La regresión lineal es un modelo supervisado de aprendizaje automático ideal para **predicción de valores continuos**, como en este caso, donde el objetivo es predecir **"Calories\_Burned"** (calorías quemadas). Esta variable dependiente es continua y cuantitativa, lo cual hace que la regresión lineal sea un enfoque natural para resolver el problema planteado.

El dataset proporciona múltiples variables independientes (predictoras) como **"Age"**, **"Weight"**, **"Max\_BPM"**, **"Session\_Duration"**, entre otras. Estas características tienen una relación plausible con la cantidad de calorías quemadas, ya que factores fisiológicos (como el peso, altura y frecuencia cardiaca) y patrones de ejercicio (duración y tipo de entrenamiento) afectan directamente el gasto energético de una persona.

**Ventajas de la Regresión Lineal**

La regresión lineal presenta varias ventajas que la convierten en una herramienta adecuada para el problema propuesto:

* **Interpretabilidad:** La regresión lineal es uno de los algoritmos más interpretables. Los coeficientes obtenidos en el modelo pueden mostrar la influencia de cada variable en la predicción de calorías.
* **Simplicidad:** Es un método sencillo, eficiente y rápido, especialmente con un dataset de tamaño moderado como el proporcionado (973 muestras).
* **Relaciones lineales:** Aunque el ejercicio y el gasto calórico pueden tener cierta no linealidad, una regresión lineal es una excelente aproximación inicial. Las características como **"Session\_Duration"** y **"Max\_BPM"** probablemente tengan una relación lineal significativa con **"Calories\_Burned"**.
* **Requisitos de preprocesamiento:** La regresión lineal es robusta siempre que las variables estén correctamente escaladas y no exista multicolinealidad severa.

**3.1 Análisis: Split 80/20 vs Split 50/50 en Regresión Lineal Múltiple**

En este apartado se comparan los resultados obtenidos al aplicar **Regresión Lineal Múltiple** con dos particiones del dataset:

1. **Split Académico (80/20)**: Se utiliza el **80%** de los datos para entrenamiento y el **20%** para validación.
2. **Split Investigativo (50/50)**: Se divide el dataset en partes iguales, usando el **50%** para entrenamiento y el **50%** para validación.

**Resultados con Split Académico (80/20)​**

1. **Métricas de Desempeño**:
   * **Error Absoluto Medio (MAE)**: Menor debido a una mayor cantidad de datos de entrenamiento, permitiendo al modelo ajustarse.

**En el conjunto de prueba: 30.2850**

* + **Error Cuadrático Medio (MSE)**: Más bajo al capturar mejor las tendencias y reducir sobreajustes.

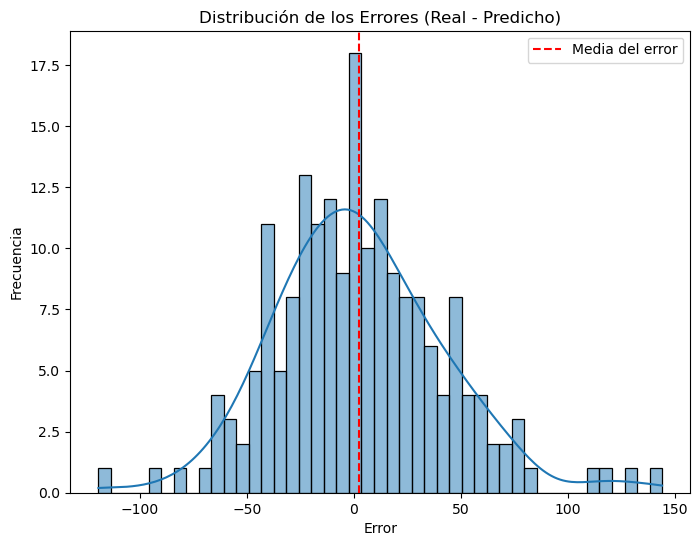
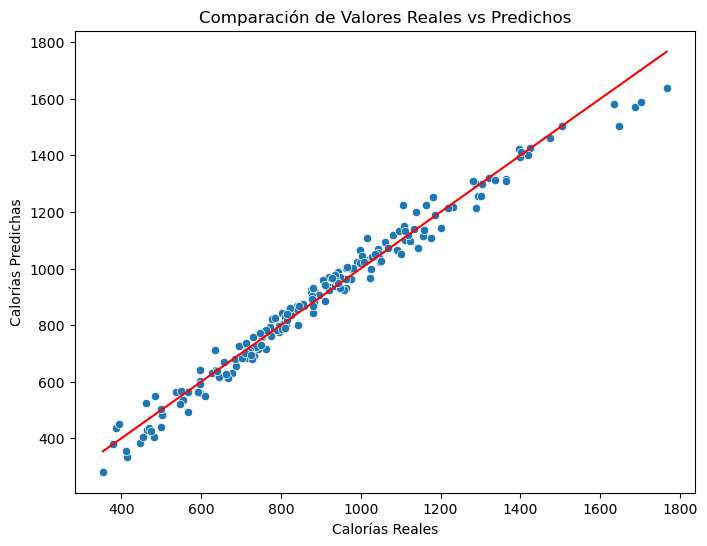
**En el conjunto de prueba: 1570.4539**

* + **Coeficiente de Determinación (R2)**: Alto valor, indicando un ajuste más preciso a los datos de entrenamiento.

**En el conjunto de prueba: 0.9801**

**RESULTADOS:**

|  |  |
| --- | --- |
| Real | Predicho |
| 354.0 | 281.977976 |
| 636.0 | 641.409146 |
| 678.0 | 632.156034 |
| 1688.0 | 1572.094533 |
| 962.0 | 1000.919072 |
| 695.0 | 726.413194 |
| 813.0 | 800.421805 |
| 804.0 | 805.278947 |
| 598.0 | 559.843159 |
| 1163.0 | 1224.155322 |



**Resultados con Split Investigativo (50/50)​**

1. **Métricas de Desempeño**:
   * **Error Absoluto Medio (MAE)**: Ligeramente más alto que en el split 80/20 debido a la reducción de datos en el entrenamiento.

**En el conjunto de prueba: 31.0224**

* + **Error Cuadrático Medio (MSE)**: Aumenta debido a un modelo menos ajustado.

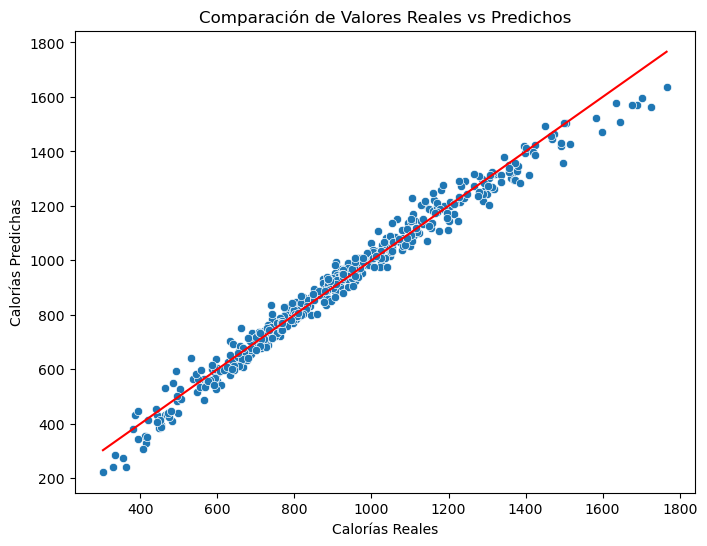
**En el conjunto de prueba: 1676.5007**

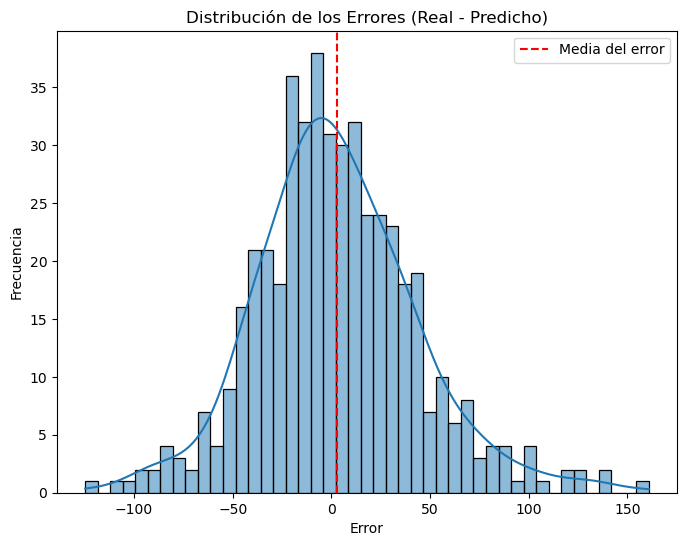
* + **Coeficiente de Determinación (R2)**: Ligeramente menor al no capturar completamente las relaciones

**En el conjunto de prueba: 0.9787**

**RESULTADOS:**

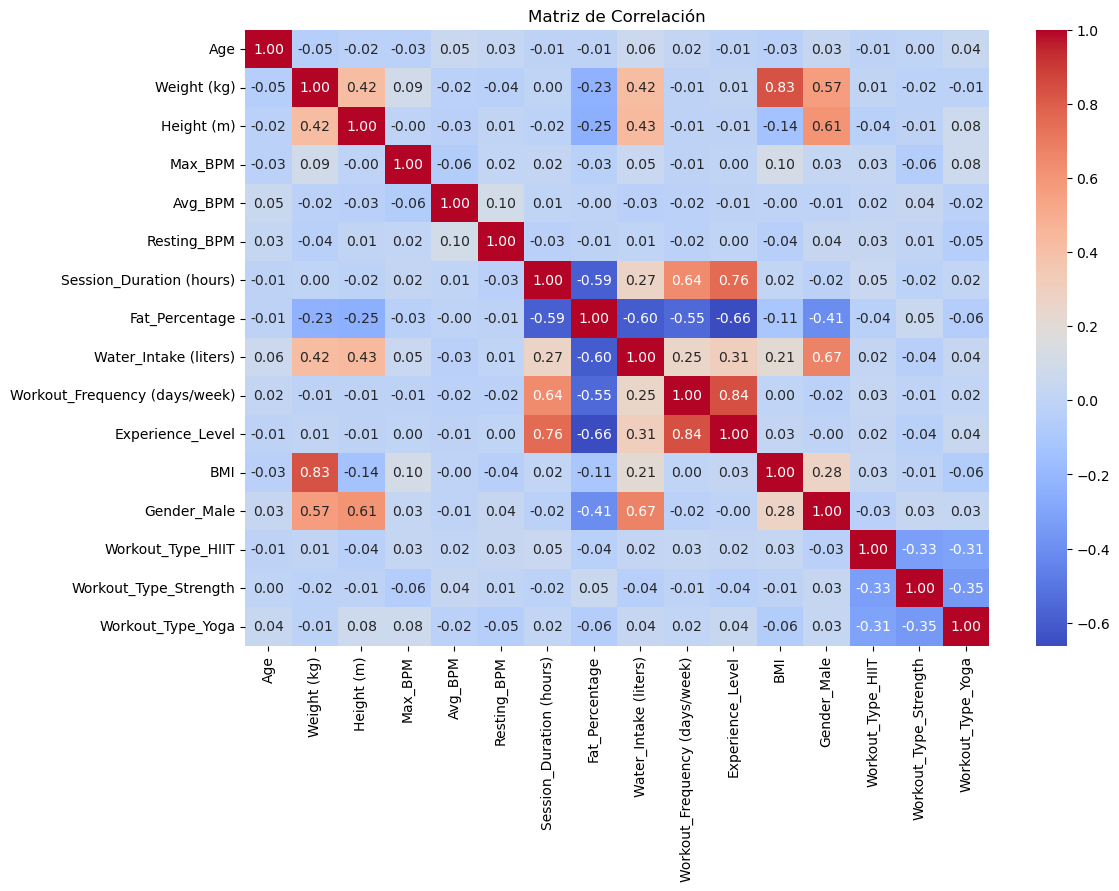
|  |  |
| --- | --- |
| Real | Predicho |
| 354.0 | 273.304617 |
| 636.0 | 641.286085 |
| 678.0 | 639.029023 |
| 1688.0 | 1569.810788 |
| 962.0 | 990.980706 |
| 695.0 | 725.564258 |
| 813.0 | 803.308708 |
| 804.0 | 803.657081 |
| 598.0 | 558.281987 |
| 1163.0 | 1221.106496 |

****

****

**Comparación General**

|  |  |  |
| --- | --- | --- |
| **Métrica** | **Split 80/20** | **Split 50/50** |
| **Error Absoluto Medio** | **30.2850** | **31.0224** |
| **Error Cuadrático Medio** | **1570.4539** | **1676.5007** |
| **R2R^2R2 Coeficiente** | **0.9801** | **0.9787** |

**MATRIZ DE CORRELACIÓN PARA AMBOS CASOS**

**ANALISIS MEDIANTE LA METRICA VIF**

El **VIF** (**Variance Inflation Factor** o Factor de Inflación de la Varianza) es una métrica utilizada en la regresión lineal para **detectar multicolinealidad** entre las variables independientes (predictoras).

**Definición Matemática**

El VIF para una variable independiente Xi​ se calcula como:

Donde:

* Ri2​: Es el coeficiente de determinación obtenido al ajustar una **regresión lineal** donde la variable Xi es explicada por las demás variables independientes.

**Interpretación del VIF**

* **VIF = 1**: No hay multicolinealidad entre la variable y las demás.
* **VIF entre 1 y 5**: La multicolinealidad es moderada y generalmente aceptable.
* **VIF > 5**: Existe una alta **multicolinealidad**. La variable está altamente correlacionada con otras, lo cual puede distorsionar los coeficientes de la regresión y afectar la estabilidad del modelo.
* **VIF > 10**: La multicolinealidad es extrema y debe abordarse.

**PARA MABAS SITUCIONES ESTA METRICA RESULTO EN LO SIGUIENTE**

|  |  |
| --- | --- |
| Variable | VIF |
| Weight (kg) | 74.303605 |
| BMI | 65.322876 |
| Height (m) | 23.607291 |
| Experience\_Level | 5.813555 |
| Gender\_Male | 5.173890 |
| Fat\_Percentage | 2.412912 |
| Water\_Intake (liters) | 2.398359 |
| Workout\_Frequency (days/week) | 2.219180 |
| Session\_Duration (hours) | 2.146551 |
| Workout\_Type\_Yoga | 2.034781 |
| Workout\_Type\_Strength | 1.929031 |
| Workout\_Type\_HIIT | 1.785387 |
| Age | 1.063607 |
| Max\_BPM | 1.038119 |
| Avg\_BPM | 1.026254 |
| Resting\_BPM | 1.023643 |

**4. Análisis de Componentes Principales (PCA)​**

El **PCA** significa **Análisis de Componentes Principales** es una técnica ampliamente utilizada para la **reducción de dimensionalidad**. El objetivo principal del PCA es reducir un conjunto de datos con un gran número de variables correlacionadas en un conjunto más pequeño de variables **no correlacionadas**, conocidas como **componentes principales**. Estos componentes retienen la mayor cantidad posible de **varianza** presente en el conjunto de datos original.

La importancia del PCA radica en su capacidad para:

1. Reducir la **complejidad computacional** al disminuir la dimensionalidad.
2. Mitigar la **multicolinealidad** entre variables eliminando la redundancia.
3. Mejorar el rendimiento del modelo en conjuntos de datos de **alta dimensionalidad**.

**Matemáticas detrás de PCA**

El PCA es una aplicación de conceptos de **álgebra lineal** en el aprendizaje automático. El proceso para derivar los componentes principales a partir de un conjunto de datos puede dividirse en **seis pasos sistemáticos**:

**4.1. Preparar el conjunto de datos (d-dimensional)**

Un conjunto de datos puede representarse con d+1 dimensiones, donde d son las características (variables predictoras) y 1 corresponde a la variable objetivo (etiquetas). En el PCA, se **descartan las etiquetas**, dejando únicamente el espacio de características (d-dimensiones).

Por ejemplo, si comenzamos con un conjunto de **12 columnas**, el PCA analizará estas 12 variables para identificar nuevos ejes (componentes principales) que maximicen la varianza.

**4.2. Calcular la media de cada dimensión**

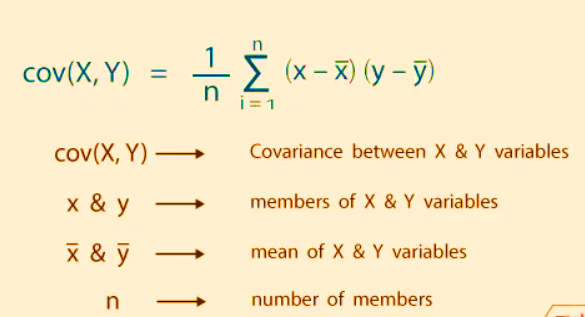
Para cada característica en el conjunto de datos, se calcula la **media**. Luego, los datos se **centran**, restando la media a cada valor para que el conjunto de datos tenga una media igual a cero. Matemáticamente:

donde:

* **A** es la matriz original de datos.
* **μ** es el vector de medias de cada columna.

**4.3. Calcular la matriz de covarianza**

La **matriz de covarianza** mide la relación entre las variables. Se calcula con la siguiente fórmula:



La matriz resultante es **simétrica**, donde:

* Los **elementos diagonales** representan la **varianza** de cada característica.
* Los **elementos fuera de la diagonal** representan la **covarianza** entre pares de características.

**4.4. Calcular los autovalores y autovectores**

Los **autovectores** representan las **direcciones** principales de los datos, mientras que los **autovalores** indican la **magnitud de la varianza** en esas direcciones. Matemáticamente:

donde:

* A es la matriz de covarianza.
* λ es el autovalor.
* ν es el autovector asociado a λ.

Se resuelve la ecuación característica:

para obtener los autovalores y sus autovectores correspondientes.

**4.5. Ordenar los autovectores y seleccionar los k componentes principales**

Los autovectores se ordenan en función de sus **autovalores** de mayor a menor. Los autovectores con los **mayores autovalores** son seleccionados para formar la matriz de transformación W, que tendrá dimensiones d×k, donde k es el número de componentes principales seleccionados.

El criterio para elegir k se basa en retener la **mayor cantidad de varianza** del conjunto de datos original.

**4.6. Transformar los datos al nuevo subespacio**

Finalmente, el conjunto de datos original se proyecta en el nuevo subespacio utilizando la matriz de transformación W:

donde:

* **X** es la matriz de datos centrada.
* WT es la **transpuesta** de la matriz de autovectores.

El resultado es un conjunto de datos transformado Y con **menor dimensionalidad**, donde los componentes principales retienen la mayor parte de la información original.

**Resultados para 12 Componentes**

|  |  |  |
| --- | --- | --- |
| Ejecución | MSE | R² |
| 1 | 2627.9878700 | 0.9613650 |
| 2 | 2840.2010350 | 0.9628510 |
| 3 | 2779.8562140 | 0.9598010 |
| 4 | 2628.8821770 | 0.9695440 |
| 5 | 2610.5887030 | 0.9631940 |

**MSE: 2697.5032, Promedio R²: 0.9634**

**Mediana MSE: 2628.8822, Mediana R²: 0.9629**

**Resultados para 10 Componentes**

|  |  |  |
| --- | --- | --- |
| Ejecución | MSE | R² |
| 1 | 2714.5621650 | 0.9600920 |
| 2 | 2882.9514810 | 0.9622920 |
| 3 | 3020.9875310 | 0.9563150 |
| 4 | 2654.3322720 | 0.9692490 |
| 5 | 2631.8056170 | 0.9628950 |

**Promedio MSE: 2780.9278, Promedio R²: 0.9622**

**Mediana MSE: 2714.5622, Mediana R²: 0.9623**

**Resultados para 11 Componentes**

|  |  |  |
| --- | --- | --- |
| Ejecución | MSE | R² |
| 1 | 2697.0956940 | 0.9603490 |
| 2 | 2844.2845620 | 0.9627980 |
| 3 | 2820.0706830 | 0.9592200 |
| 4 | 2626.7843290 | 0.9695680 |
| 5 | 2625.4130380 | 0.9629850 |

**Promedio MSE: 2722.7297, Promedio R²: 0.9630**

**Mediana MSE: 2697.0957, Mediana R²: 0.9628**

**Resultados para 9 Componentes**

|  |  |  |
| --- | --- | --- |
| Ejecución | MSE | R² |
| 1 | 5991.0373970 | 0.9119230 |
| 2 | 4694.7391840 | 0.9385940 |
| 3 | 7565.1225880 | 0.8906030 |
| 4 | 6895.9877320 | 0.9201090 |
| 5 | 3642.1150010 | 0.9486510 |

**Promedio MSE: 5757.8004, Promedio R²: 0.9220**

**Mediana MSE: 5991.0374, Mediana R²: 0.9201**

**Resultados para 5 Componentes**

|  |  |  |
| --- | --- | --- |
| Ejecución | MSE | R² |
| 1 | 23522.1398490 | 0.6541890 |
| 2 | 25365.8325450 | 0.6682230 |
| 3 | 25376.6022360 | 0.6330370 |
| 4 | 26717.2803400 | 0.6904780 |
| 5 | 19235.9117730 | 0.7287990 |

**Promedio MSE: 24043.5533, Promedio R²: 0.6749**

**Mediana MSE: 25365.8325, Mediana R²: 0.6682**

**Resultados para 3 Componentes**

|  |  |  |
| --- | --- | --- |
| Ejecución | MSE | R² |
| 1 | 24187.4226580 | 0.6444080 |
| 2 | 27290.9590160 | 0.6430420 |
| 3 | 26864.2861980 | 0.6115250 |
| 4 | 29420.2841550 | 0.6591630 |
| 5 | 24092.5916690 | 0.6603260 |

**Promedio MSE: 26371.1087, Promedio R²: 0.6437**

**Mediana MSE: 26864.2862, Mediana R²: 0.6444**

**5. Análisis de Agrupamiento K-Means en el Dataset de Entrenamiento Físico**

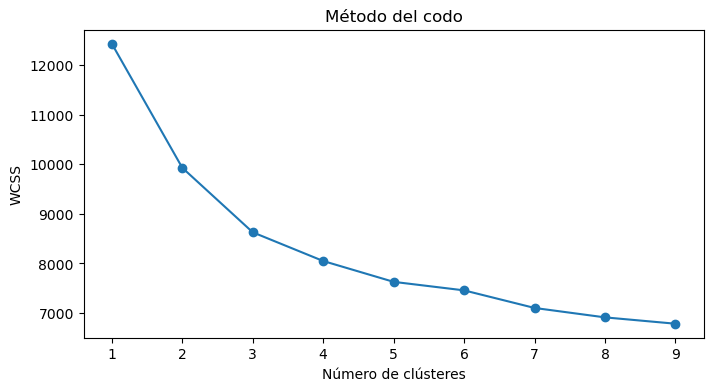
El método **K-Means**, una técnica de **aprendizaje no supervisado**, fue aplicado al **Gym Members Dataset** para segmentar a los miembros del gimnasio en grupos homogéneos según variables clave relacionadas con sus características físicas, hábitos de entrenamiento y rendimiento calórico.

**1. Preparación y Preprocesamiento de Datos**

* Se integraron y escalaron las variables mediante **StandardScaler** para garantizar que todas las características tuvieran igual peso en el modelo.
* Se incluyeron atributos como:
  + Datos demográficos: **Edad, Peso, Altura**.
  + Indicadores fisiológicos: **Max\_BPM, Avg\_BPM, Resting\_BPM**.
  + Hábitos de ejercicio: **Duración de sesión, Frecuencia, Tipo de entrenamiento**.
  + Métricas de rendimiento: **Calorías quemadas**.

**2. Método del Codo para la Selección de Clústeres**

Se aplicó el **método del codo** analizando la **suma de distancias al cuadrado intra-clúster (WCSS)**. La gráfica mostró un **punto óptimo en k=4**, lo que sugiere 4 agrupaciones distintas dentro del dataset.



**3. Análisis de los Clústeres**

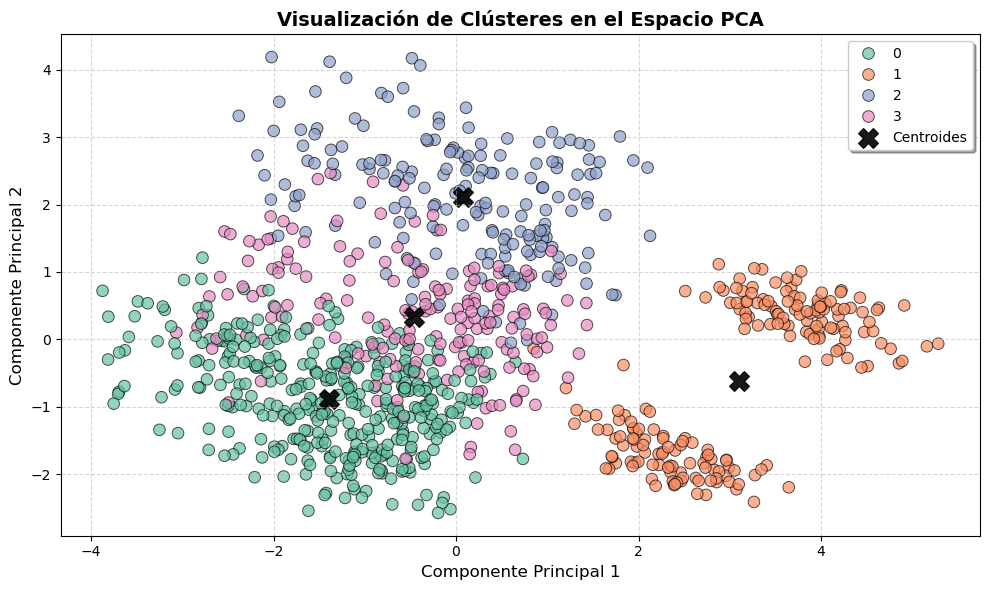
Los resultados del modelo k=4 segmentaron a los miembros del gimnasio en cuatro grupos con características particulares:

|  |  |
| --- | --- |
| **Clúster** | **Descripción General** |
| **Clúster 0** | Mujeres jóvenes, de menor talla y peso, con mayor porcentaje de grasa. Entrenan con **baja frecuencia** y duración, aunque tienen **experiencia moderada**. Su hidratación es deficiente y presentan **baja quema calórica**. |
| **Clúster 1** | El grupo más **fit**: entrenan **más tiempo** y con **alta frecuencia** semanal. Tienen un porcentaje de grasa muy bajo, excelente hidratación, altísima experiencia y registran la **mayor quema calórica**. Contiene una mezcla equilibrada de géneros. |
| **Clúster 2** | Predominantemente **hombres altos y pesados**, bien hidratados y con buena capacidad cardiovascular (**alto Max\_BPM**). A pesar de su **alta experiencia**, entrenan menos días y duran menos tiempo, lo que resulta en una **quema calórica moderada**. |
| **Clúster 3** | Hombres **mayores**, delgados (**bajo BMI**), con menos frecuencia y duración en el entrenamiento. Aunque su hidratación es adecuada y poseen experiencia, su intensidad de entrenamiento es menor, resultando en **baja quema calórica**. |

**4. Visualización con PCA**

Se utilizó **PCA (Análisis de Componentes Principales)** para reducir la dimensionalidad de los datos a **2 componentes principales**, facilitando así la visualización de los clústeres:

* Los datos fueron proyectados en un plano bidimensional (PCA1 y PCA2).
* Los centroides de los clústeres fueron transformados al mismo espacio PCA y representados con un marcador **X**.
* La visualización mostró una clara separación entre los clústeres, validando la segmentación obtenida.



**5. Interpretación de Resultados**

* El modelo K-Means fue efectivo para identificar grupos con **patrones de entrenamiento y características físicas diferenciadas**.
* El **Clúster 1** destaca por su alto rendimiento calórico, mientras que **Clústeres 0 y 3** muestran oportunidades para mejorar hábitos de entrenamiento.
* La visualización en el espacio PCA confirmó que las características seleccionadas son relevantes para la segmentación, mostrando grupos bien definidos.

**6. Simulado Recocido Aplicado al Problema de las N-Reinas**

El **Simulado Recocido** es una técnica de optimización basada en métodos probabilísticos que permite escapar de mínimos locales explorando soluciones vecinas y aceptando, en ocasiones, soluciones menos óptimas con el fin de alcanzar un mínimo global. Esta técnica fue implementada para resolver el **problema de las N-Reinas**, un desafío clásico en teoría de algoritmos y optimización combinatoria.

**Formulación del Problema**

El problema consiste en colocar N reinas en un tablero de ajedrez N×N de tal manera que **ninguna reina ataque a otra**. En términos formales, se deben evitar:

1. **Ataques en filas**: Dos reinas no deben compartir la misma fila.
2. **Ataques en columnas**: Dos reinas no deben compartir la misma columna (resuelto automáticamente en esta implementación).
3. **Ataques en diagonales**: La distancia entre dos reinas en filas y columnas debe diferir.

**Implementación del Simulado Recocido**

El método se estructura en varias fases:

**1. Solución Inicial Aleatoria**

Se genera una solución aleatoria colocando una reina en cada columna, pero en filas seleccionadas al azar.

**2. Función de Costo**

La función de costo evalúa cuántos pares de reinas se atacan entre sí, calculando las colisiones en filas y diagonales.

Costo = Número total de ataques entre reinas

**3. Generación de Soluciones Vecinas**

Para explorar el espacio de búsqueda, se genera una solución vecina modificando **al azar la posición de una reina** en una columna específica.

**4. Probabilidad de Aceptación**

La función de aceptación probabilística decide si se acepta una solución peor en función de la diferencia de costos y la **temperatura actual** T:

**5. Bucle del Simulado Recocido**

El algoritmo reduce gradualmente la temperatura, permitiendo explorar más al inicio y enfocarse en soluciones cercanas al final.

**Resultados Obtenidos**

Para N=6, se aplicó el **Simulado Recocido** con los siguientes parámetros:

* **Temperatura inicial**: 1000.
* **Factor de enfriamiento**: 0.995.
* **Temperatura mínima**: 1e-3.
* **Máximas iteraciones**: 100000.

**Mejor solución encontrada: [1, 3, 5, 0, 2, 4]**

**Ataques: 0**

**Iteraciones realizadas: 1641**

**Se encontró una solución óptima (sin ataques).**

**Tablero:**

**. . . Q . .**

**Q . . . . .**

**. . . . Q .**

**. Q . . . .**

**. . . . . Q**

**. . Q . . .**

**7. Conclusiones**

Este estudio integró múltiples técnicas de análisis de datos y Machine Learning:

* **Preprocesamiento**: Garantizó datos limpios y consistentes.
* **Regresión lineal**: Proporcionó una base estadística para las predicciones.
* **PCA**: Redujo la complejidad y dimensionalidad de los datos.
* **K-means**: Segmentó a los usuarios en grupos relevantes.
* **Recocido simulado**: Optimización eficiente del problema de las N-Reynas.

**Referencias**

* **Raschka, S., & Mirjalili, V. (2019).** *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2.* Packt Publishing. ISBN: 978-1789955750.
* **Seber, G. A. F., & Lee, A. J. (2012).** *Linear Regression Analysis.* Wiley. DOI: 10.1002/9781118274446.
* **Ruppert, D. (2004).** *Statistics and Data Analysis for Financial Engineering.* Springer. DOI: 10.1007/978-1-4757-5215-7.