Trabajo Fin de Máster

Diseño y desarrollo de una aplicación web interactiva para el análisis de datos de entrenamientos de alta intensidad

Cristian Marques Corrales

Máster en Técnicas Estadísticas

Curso 2023-2024

Propuesta de Trabajo Fin de Máster

Título en galego: Deseño e desenvolvemento dunha aplicación web interactiva para a análise de datos de adestramento de alta intensidade
Título en español: Diseño y desarrollo de una aplicación web interactiva para el análisis de datos de entrenamientos de alta intensidad
English title: Design and development of an interactive web application for the analysis of high intensity training data
Modalidad: B
Autor: Cristian Marques Corrales, Universidade de Vigo
Directores: Marta Sestelo Pérez, Universidade de Vigo
Tutores: Jonathan Riveiro Álvarez, Centro Suma
Recomendaciones:
Otras observaciones:

Firmado por
RIVEIRO ALVAREZ
JONATHAN -

Doña Marta Sestelo Pérez, Profesora Titular de la Universidade de Vigo y Don Jonathan Riveiro Álvarez, Responsable del Área de Entrenamiento de Centro Suma, informan que el Trabajo Fin de Máster titulado

Diseño y desarrollo de una aplicación web interactiva para el análisis de datos de entrenamientos de alta intensidad

fue realizado bajo su dirección por Don Cristian Marques Corrales para el Máster en Técnicas Estadísticas. Estimando que el trabajo está terminado, da su conformidad para su presentación y defensa ante un tribunal.

En Vigo, a 22 de Julio de 2024.

La directora:
Doña Marta Sestelo Pérez

SESTELO PEREZ MARTA - 76998300G
Firmado digitalmente por SESTELO PEREZ MARTA - 76998300G
Fecha: 2024.07.22 14:24:18 +02'00'

El tutor:
Don Jonathan Riveiro Álvarez

Firmado por
RIVEIRO ALVAREZ
JONATHAN -

El autor:
Don Cristian Marques Corrales

Firmado por
MARQUES CORRALES
CRISTIAN -
***8645** el día
22/07/2024 con un
certificado

Índice general

Capítulo 1

Introducción

El entrenamiento con resistencias es el principal medio por el cual podemos incrementar la hipertrofia muscular (Schoenfeld, 2010). Este tipo de entrenamiento tiene diversos beneficios para la salud como un aumento en la fuerza y tamaño muscular, una mejora en la salud metabólica y un menor riesgo de lesiones (Hutchins, 1992; Little y McGuff, 2009). Dentro de la categoría de entrenamiento con resistencia, existen diversas metodologías y estrategias. Una de las más conocidas es el entrenamiento de fuerza de alta intensidad (HIST). Las características de este tipo de entrenamiento se analizarán en detalle en el Capítulo 2 de estado del arte. De manera resumida, indicar que se trata de un entrenamiento de corta duración con restricciones de descanso entre la ejecución de ejercicios, llegando hasta el fallo muscular, es decir, hasta la incapacidad de contracción en la fase concéntrica, generando así intensidad en los entrenamientos. Como se comentó anteriormente, presenta una serie de beneficios para la salud y eficiencia en el tiempo, ya que son considerablemente más cortos que los entrenamientos de fuerza tradicionales y potencialmente menos lesivos que otras tendencias presentes en el mundo fitness como los entrenamientos Hiit o CrossFit.

Este trabajo se centra en el diseño, desarrollo y despliegue de una aplicación web para el Centro Suma[1] que permite al usuario final la visualización de los resultados de su entrenamiento y la estimación de su fuerza máxima. El Centro Suma, es un grupo de profesionales con formación en diferentes especialidades que concibe al ser humano como un ser complejo, que necesita un abordaje integral. Su misión es la de aportar una respuesta a las necesidades de las personas, sin ser entidades independientes sin ninguna relación entre ellas, sino ofreciendo un enfoque interdisciplinar con una actitud constructiva y abierta, para dar la mejor solución y conseguir los mejores resultados posibles. La empresa cuenta con distintos departamentos como el de psicología, fisioterapia, entrenamiento y nutrición. En este trabajo, hemos colaborado con el departamento de entrenamiento.

En general, nuestro trabajo persigue dos objetivos:

- Crear una plataforma capaz de enseñar de forma interactiva la evolución de la condición física, aportando un gran beneficio a la empresa y a sus usuarios.
- Ajustar modelos de aprendizaje automático que permiten una correcta estimación de la fuerza máxima (RM).

Para llevar a cabo el primer objetivo, hemos utilizado el software estadístico R (R Core Team, 2021) y su popular librería de visualización de datos Shiny (Chang et al., 2023). La plataforma desarrollada permite, a través de otras librerías, tener un sistema de acceso por usuario y contraseña, un modo administrador, así como conexión a la nube (Google Drive). Para el alcanzar el segundo objetivo, hemos optado por el ajuste de un bosque aleatorio y un modelo de regresión aditiva generalizada (Breiman, 2001; Wood, 2017). La fuerza máxima o repetición máxima (RM) es una métrica que

[1] Para más información consultar Centro Suma.

estima la capacidad de una persona de levantar el máximo peso en un solo movimiento o repetición. Existen dos formas de estimarlo: el método directo y el indirecto. El primero, implica realizar una prueba donde el usuario ejecuta el ejercicio con pesos exigentes, hasta encontrar uno que no pueda levantar más de una vez. El segundo, estima la fuerza máxima a través de fórmulas (Epley, 1985; Lombardi, 1989; O'Connor y Simmons, 1989; Brzycki, 1993; González-Badillo y Gorostiaga-Ayestarán, 1997), ejecutando el ejercicio con un peso menos exigente pero realizando el movimiento hasta el fallo muscular. Estas fórmulas, por lo tanto, estiman la fuerza máxima en función del número de repeticiones y de la carga. En consecuencia, una de las aportaciones de este trabajo es poder estimar el RM mediante métodos de aprendizaje automático sin necesidad de usar el método directo, el cual presenta problemas de seguridad para los usuarios (Niewiadomski et al., 2008), y sin necesidad de que el centro utilice una de las sesiones de entrenamiento para estimarlo por el método indirecto. Diversas técnicas estadísticas han sido utilizadas en el preprocesado de los datos, entre ellas, métodos de recorte ("winsorización"), imputaciones de media, selección e importancia de variables, entre otras.

A continuación, se detalla la estructura de este trabajo. El Capítulo 2 presenta el estado del arte, que analiza los trabajos científicos en el ámbito del entrenamiento de fuerza de alta intensidad, así como los de ámbito estadístico que predicen el RM u otras métricas de composición y rendimiento físico. El Capítulo 3 detalla la metodología utilizada, incluyendo el diseño de la investigación, los datos, las herramientas de software utilizadas, el funcionamiento general de los modelos ajustados, las técnicas de validación cruzada y las métricas de error utilizadas. El Capítulo 4 resume el diseño y funcionamiento de la plataforma, ofreciendo una versión impresa de la plataforma desarrollada. El análisis exploratorio de datos, los resultados y la discusión de los hallazgos se incluyen en el Capítulo 5 y, finalmente, el Capítulo 6 incorpora las conclusiones finales y posibles líneas de investigación futuras.

Capítulo 2

Estado del arte

Este capítulo proporciona una revisión de la literatura relacionada con el entrenamiento de fuerza de alta intensidad y con las técnicas estadísticas aplicadas en este ámbito. El objetivo es situar el trabajo en el contexto de los desarrollos recientes, identificando las tendencias, metodologías y herramientas más relevantes utilizadas en el campo. Esta revisión destaca las fortalezas y limitaciones de los enfoques existentes, justificando el uso de esta modalidad de entrenamiento y la necesidad de técnicas estadísticas capaces de predecir correctamente métricas de rendimiento físico.

2.1. Visión general del entrenamiento de fuerza de alta intensidad

Jones (1970) es ampliamente reconocido como una figura pionera en el desarrollo de las metodologías de entrenamiento de alta intensidad (HIST). Su trabajo en la década de 1970 sentó las bases para un nuevo enfoque del entrenamiento con fuerza, enfatizando la importancia de entrenamientos breves, intensos e infrecuentes. Jones también fue responsable de la la invención de las máquinas de ejercicio *Nautilus*, las cuales revolucionaron la industria del *fitness* al proporcionar un medio para lograr una estimulación muscular máxima con un riesgo mínimo de lesiones (Jones, 1970). Esta metodología se centró en los principios de sobrecarga, especificidad y resistencia progresiva. Se aboga por realizar ejercicios hasta el punto de fallo muscular momentáneo, donde no se pueda completar más repeticiones con una forma adecuada. Este enfoque está diseñado para maximizar el reclutamiento de fibras musculares y estimular la hipertrofia de manera más efectiva que los entrenamientos tradicionales de menor intensidad (Jones, 1970). Los beneficios de la metodología HIST incluyen un aumento de la fuerza y tamaño muscular, una mejora de la resistencia muscular y una mayor eficiencia metabólica. Además, el uso de máquinas minimiza el riesgo de lesiones al proporcionar una resistencia controlada y constante a lo largo del rango de movimiento (Jones, 1970).

Los postulados más relevantes del trabajo de Jones se pueden resumir de la siguiente manera:

- Una sola serie por ejercicio y grupo muscular.
- Un sólo día de entrenamiento semanal por grupo muscular.
- Cada serie debe realizarse hasta el fallo muscular (en la fase concéntrica) con 8-12 repeticiones por serie.
- Tiempo del ejercicio. Este concepto implica prestar atención al ritmo (velocidad y pausas) al que se realiza cada repetición y se expresa en segundos de duración en cada fase del movimiento.

Ken Hutchins es reconocido por el desarrollo del protocolo de ejercicio *SuperSlow* (Hutchins, 1992), una extensión de los principios HIST de Arthur Jones. Su metodología implica realizar cada repetición de un ejercicio a un ritmo muy lento, típicamente 10 segundos para la fase concéntrica y 10 segundos para la fase excéntrica. Este ritmo lento está diseñado para eliminar el impulso, aumentando así el tiempo bajo tensión y maximizando el reclutamiento de fibras musculares. El protocolo *SuperSlow* ofrece varios beneficios, incluyendo una mejora en la fuerza y resistencia muscular, una mayor salud de las articulaciones y tejidos conectivos, y un menor riesgo de lesiones debido a la naturaleza controlada de los movimientos (Hutchins, 1992). El trabajo de Hutchins ha sido influyente en la promoción de un enfoque más seguro y efectivo del entrenamiento de fuerza, particularmente para adultos mayores y personas con limitaciones físicas. Como filosofía de ejercicio, *SuperSlow* abarca un amplio espectro de consideraciones que se exponen a continuación.

- **Frecuencia:** la actividad de alta intensidad requiere un mayor intervalo de recuperación. La experiencia de la empresa con *SuperSlow* muestra que la mayoría de los sujetos requieren de tres a siete días de descanso entre entrenamientos. Esto no significa que deban estar totalmente inactivos. Sin embargo, es común que muchos usuarios crean erróneamente que otra actividad como correr, nadar, o andar en bicicleta ayuda o no perjudica a su progreso. Para obtener resultados óptimos, la actividad de esfuerzo debe limitarse durante el intervalo de descanso o este se puede ver comprometido.

- **Duración:** si una actividad es suficientemente intensa, entonces sólo puede continuarse durante un breve período de tiempo. En la experiencia de la empresa, realizar un entrenamiento que exceda los treinta minutos es una indicación de que la intensidad fue demasiado baja. Por supuesto, esto varía de forma individual. Algunos sujetos avanzados de *SuperSlow* entrenan a tal intensidad que su entrenamiento dura menos de diez minutos.

- **Intensidad:** la intensidad es el grado de esfuerzo momentáneo aplicado en el ejercicio. Actividades que se pueden continuar indefinidamente, como caminar o trotar, se consideran de intensidad baja. La actividad breve que causa un fallo muscular en solo uno a cuatro minutos se considera de alta intensidad. Hay que tener en cuenta que intensidad puede referirse en un sentido limitado a un ejercicio en particular, o puede referirse en un sentido más amplio a una serie de ejercicios que comprenden el entrenamiento general. Es posible que un ejercicio sea de alta intensidad, mientras que la intensidad media de todo el entrenamiento sea moderada o baja.

- **Forma:** la forma adecuada, ante todo, significa una velocidad de movimiento lenta. Esto es deseable desde tres perspectivas. En primer lugar, una velocidad lenta minimiza la aceleración (arranque y parada repentina), la fuente de fuerza excesiva y, por lo tanto, las lesiones. En segundo lugar, la velocidad lenta minimiza el impulso que descarga indeseablemente los músculos. La carga muscular eficiente es de lo que se trata el ejercicio de alta intensidad. Y tercero, la velocidad lenta permite la concentración y el estudio durante el movimiento. Cuando te mueves rápidamente, no hay tiempo para pensar en el movimiento. Además, se debe evitar hacer muecas, contonearse, retorcerse, agitarse, contener la respiración, agarrar excesivamente las manos y tensiones innecesarias en áreas del cuerpo que no están destinadas a participar en ejercicios particulares.

- **Seguridad:** la seguridad también significa una velocidad de movimiento lenta, como ya se ha comentado. Sin embargo, existen otras consideraciones de seguridad. Mantenerse fresco es una de ellas, ya que el sobrecalentamiento es una amenaza para el bienestar del usuario. La técnica de respiración adecuada minimiza los peligros de un derrame cerebral. La posición y el soporte

adecuados de la cabeza y el cuello protegen al sujeto de la tensión del cuello y el dolor de cabeza. La ejecución adecuada del ejercicio está asociada a que el cuerpo permanezca alineado y protegido. La forma de moverse dentro y fuera del equipo y entre ejercicios también está coreografiada para maximizar la seguridad y la eficiencia del efecto del ejercicio.

- **Objetivo:** otro aspecto importante de la filosofía de la metodología *SuperSlow* es la razón por la que se realiza. En general, hacemos ejercicio, no para disfrutar del ejercicio, sino para que podamos aplicar los beneficios obtenidos en todas las demás actividades (o inactividad) en nuestras vidas. El ejercicio no es un lujo, sino un requisito básico para una vida normal y saludable. El ejercicio a menudo se aplica para quemar grasa corporal y para mejorar ciertas áreas del cuerpo. La reducción puntual, aunque es la justificación más común para muchos productos de éxito comercial, a menudo es un engaño. Además, el ejercicio es un método muy ineficaz para quemar calorías adicionales más allá de las que se consumen normalmente en las actividades diarias típicas. Ciertamente, el ejercicio quemará algunas calorías adicionales, pero más importante es el aumento del metabolismo basal que resulta de poseer una mayor musculatura. El ejercicio de alto volumen y baja intensidad conduce al desgaste muscular, por lo que lo más importante para el control de la grasa corporal son las calorías que no se consumen. El control calórico es el factor más importante para abordar la delgadez corporal.

Doug McGuff avanzó aún más los principios del HIST a través de su libro "*Body by Science*", coescrito con John Little (Little & McGuff, 2009). Su enfoque de enfatiza la base fisiológica para el entrenamiento de fuerza de alta intensidad y baja frecuencia. Según su estudio, el objetivo principal del ejercicio es el de desencadenar una respuesta adaptativa en el cuerpo, lo cual se puede lograr de manera más efectiva a través de entrenamientos breves e intensos seguidos de períodos adecuados de recuperación. Esta metodología implica realizar una serie de ejercicios compuestos utilizando un ritmo lento y controlado para minimizar el impulso y maximizar la tensión muscular (*SuperSlow*) y se ha demostrado que mejora la hipertrofia muscular y las ganancias de fuerza al tiempo que reduce el riesgo de lesiones (Little & McGuff, 2009). Este estudio, destaca que los beneficios del enfoque HIST inducen una mejora en la masa muscular, un aumento en la tasa metabólica, una mejora en la salud cardiovascular y una mejor función física general. Además, enfatizan la importancia de prescripciones de ejercicio individualizadas basadas en las características fisiológicas únicas y los objetivos de acondicionamiento físico de cada persona.

Las contribuciones de Arthur Jones, Doug McGuff y Ken Hutchins han desarrollado significativamente el campo del entrenamiento de fuerza de alta intensidad, proporcionando metodologías basadas en evidencia que enfatizan la importancia de sesiones de ejercicio breves, intensas y controladas. Estos enfoques ofrecen numerosos beneficios, incluyendo un aumento en la fuerza y tamaño muscular, una mejora en la salud metabólica y un menor riesgo de lesiones. Su trabajo continúa influyendo las prácticas modernas de acondicionamiento físico y proporciona un marco valioso para el entrenamiento de fuerza seguro y efectivo.

En concreto, la empresa con la que que se ha colaborado en este trabajo (Centro Suma), sigue una metodología de entrenamiento de fuerza de alta intensidad (HIST), influenciada por las metodologías analizadas anteriormente. En general, se siguen los postulados de Jones (1970) con alguna modificación. En este caso, la empresa no contabiliza las repeticiones sino el tiempo bajo carga (TUL), buscando ejecutar los movimientos entre 1.5 y 3 minutos. Todo el gimnasio está equipado con las máquinas *Nautilus*, también creadas por Jones (1970). Su principal ventaja es que adapta la carga a cada una de las zonas del movimiento (de todo el rango articular), lo cual ayuda en las zonas más comprometidas o vulnerables y añade resistencia en las zonas o puntos del movimiento en las que los músculos sean más fuertes. Gracias a esto, no sólo se puede entrenar un músculo de un modo mucho más eficiente y completo, sino que también se reduce notablemente el desgaste musculo-articular. Por otra parte, el

protocolo de entrenamiento utilizado es el *SuperSlow* (Jones, 1970; Hutchins, 1992), que como se ha comentado anteriormente, consiste en trabajar a una cadencia de 10 segundos de contracción en cada una de las fases del movimiento. Controlar las aceleraciones resultan un requisito indispensable para disminuir el daño articular y para aumentar el reclutamiento muscular, mejorando de esa forma los resultados.

En conclusión a esta revisión de estado del arte, parece existir suficiente evidencia científica, aparte de la propia experiencia de la empresa, para apoyar las múltiples ventajas del entrenamiento de fuerza de alta intensidad (HIST) y, en particular, de la metodología *SuperSlow*. A continuación, nos centraremos en estudiar aquellos trabajos que han utilizado modelos estadísticos para predecir variables de rendimiento físico.

2.2. Modelos existentes para el análisis de datos de entrenamiento

El RM es un parámetro fundamental en el ámbito de la ciencia del deporte y, en particular, del entrenamiento de fuerza. Se define como la cantidad máxima de peso que un individuo puede levantar en una sola repetición para un ejercicio dado, manteniendo una técnica correcta. Este valor es utilizado para evaluar la fuerza máxima de una persona y es esencial en la planificación de programas de entrenamiento. Como hemos referido anteriormente, existen dos métodos para estimar este valor: directo e indirecto. El primero, implica la evaluación de la fuerza en una prueba de rendimiento, donde el individuo ejecute los ejercicios hasta el fallo muscular. El segundo, se estima mediante fórmulas que aproximan su valor (Epley, 1985; Lombardi, 1989; O'Connor y Simmons, 1989; Brzycki, 1993; González-Badillo y Gorostiaga-Ayestarán, 1997). La evaluación de la fuerza muscular, particularmente la estimación de la repetición máxima (RM), ha sido un punto central en la investigación en ciencias del deporte debido a sus implicaciones para el entrenamiento y la seguridad. Esta revisión examina estudios clave que abordan la determinación, estimación y precisión del RM, destacando varios métodos y factores que influyen en estas estimaciones.

Niewiadomski et al. (2008) exploran las consideraciones de seguridad en la determinación y predicción del RM. Su estudio enfatiza la importancia de asegurar prácticas seguras al evaluar la fuerza máxima. Resaltan los riesgos potenciales y proponen métodos para mitigarlos, los cuales son cruciales tanto para los practicantes como para los investigadores. Este estudio subraya que una predicción precisa del RM puede ayudar a evitar tensiones y lesiones innecesarias durante las pruebas.

Stanelle et al. (2021) ajustan un modelo de regresión para la predicción de la fuerza muscular utilizando demografía, dimensiones esqueléticas y medidas de composición corporal. Su investigación proporciona un enfoque integral al considerar múltiples predictores, lo que puede mejorar la precisión de las predicciones del RM. El estudio encontró que los factores demográficos (como la edad y el sexo), las dimensiones esqueléticas y la composición corporal contribuyen significativamente a los modelos de predicción, ofreciendo una visión más holística de la evaluación de la fuerza. En particular, han encontrado significación en los siguientes predictores: altura, peso corporal, BMI (*Body Mass Index*), edad, género, masa magra, masa grasa, masa libre de grasa, porcentaje de grasa y dimensiones músculoesqueléticas.

LeSuer et al. (1997) evaluaron la precisión de varias fórmulas de predicción para estimar el rendimiento del RM en el *press* de banca, la sentadilla y el peso muerto. Sus hallazgos revelaron una alta correlación entre los coeficientes predichos y los obtenidos con el RM. En general, el número de repeticiones al fallo muscular (diez o menos) es un buen predictor del RM. Aunque existen varias fórmulas de predicción, su precisión puede variar significativamente. Este estudio sugiere que las fórmulas evaluadas predicen con mayor precisión el RM en el *press* de banca, seguido por las sentadillas y el peso

muerto. Las ecuaciones tienen una mayor imprecisión para estimar el RM en el peso muerto, por lo que los investigadores sugieren que es necesario un estudio adicional para encontrar una fórmula específica para este ejercicio. En definitiva, los resultados de esta investigación parecen sugerir que la ecuación de Wathan es más precisa en la estimación del RM, reportando un pvalor en el test de comparación de medias (estimado y real) mayor que el nivel de significación (contraste no significativo) para el *press* de banca y sentadilla.

El estudio de Sayers et al. (2018) examina cómo diferentes cargas submáximas afectan la precisión de la predicción del RM. En este estudio, se utiliza una regresión lineal para la estimación del RM con un tamaño muestral de doce usuarios que ejecutan un movimiento de *press* banca en una máquina *Smith*. Los investigadores midieron la velocidad de la barra a varias cargas (30 %, 40 %, 50 %, 60 % y 70 % del RM estimado). Encontraron que las cargas de rango medio (30 %-50 % del RM) proporcionan predicciones más precisas, ofreciendo un enfoque práctico para la evaluación de la fuerza. De forma similar al estudio de Niewiadomski et al. (2008), estos autores enfatizan la importancia de la seguridad, indicando que el uso de cargas más ligeras en las mediciones del RM, no solo pueden tener capacidad predictiva sino que exponen en menor medida el usuario a potenciales lesiones.

Colectivamente, estos estudios contribuyen a la comprensión de la predicción del RM al abordar diferentes aspectos, incluyendo la seguridad, influencias demográficas, métricas corporales y la precisión de las fórmulas de predicción. En conclusión, la predicción del RM es un tema multifactorial que se beneficia de un enfoque integral que incorpora factores demográficos y de composición corporal, así como la adecuada selección de ecuaciones de predicción. Estos estudios mejoran la fiabilidad y aplicabilidad de las predicciones del RM, contribuyendo en última instancia a prácticas de entrenamiento de fuerza más seguras y efectivas.

Capítulo 3

Metodología

En este capítulo, se describe la metodología empleada para llevar a cabo la presente investigación. El enfoque metodológico adoptado está estructurado en varias secciones, cada una de las cuales aborda aspectos importantes del proceso de investigación, desde el diseño inicial hasta las técnicas utilizadas. A continuación, se proporciona una visión general de las secciones incluidas en este capítulo.

En primer lugar, en la Sección 3.1, se presenta el diseño de la investigación, donde se detalla el marco conceptual que guía el estudio, así como las hipótesis planteadas y los objetivos específicos que se buscan alcanzar. Esta sección establece la base sobre la cual se desarrollan las demás etapas del proceso de investigación. También se describen las fuentes de datos utilizadas, el proceso de recolección y las características principales de los datos obtenidos. Se destaca la importancia de la calidad y relevancia de los datos para garantizar la validez de los resultados del estudio.

A continuación, en la Sección 3.2, se aborda el uso de software y otras herramientas empleadas en el estudio. Aquí, se especifican las herramientas tecnológicas y los entornos de desarrollo utilizados para la implementación de los modelos y el análisis de los datos, proporcionando una justificación para su elección.

La Sección 3.3, técnicas de preprocesado, describe los métodos aplicados para preparar los datos antes de su uso en los modelos de aprendizaje automático. Esto incluye técnicas de limpieza de datos y transformación, con el objetivo de mejorar la calidad y la eficacia del proceso de modelado.

La Sección 3.4 se centra en la metodología de los modelos de aprendizaje automático y constituye una parte central de este capítulo. Es en ella donde se detallan los algoritmos y enfoques específicos utilizados para construir los modelos utilizados.

Para garantizar la robustez y la generalización de los modelos, se implementó un procedimiento de validación cruzada. La Sección 3.5 explica alguno de los métodos de validación cruzada utilizados para evaluar el rendimiento de los modelos y prevenir el sobreajuste, asegurando resultados confiables y replicables.

Finalmente, en la Sección 3.6, se presentan las métricas de error utilizadas para medir la precisión y efectividad de los modelos desarrollados, se describen las métricas específicas seleccionadas y se justifica su relevancia en el contexto del estudio.

3.1. Diseño de la investigación

Tal y como ha sido mencionado en secciones anteriores, el primer objetivo de este trabajo es desarrollar una plataforma web que permita a los clientes de la empresa visualizar sus datos de entrenamiento, tanto por medio de gráficas como mediantes el uso de métricas y resúmenes numéricos, de forma que puedan controlar la evolución de su rendimiento deportivo.

Adicionalmente, la metodología que utiliza la empresa en la actualidad implica la pérdida de una sesión de entrenamiento cada trimestre, para efectuar una prueba de rendimiento deportivo. En esta

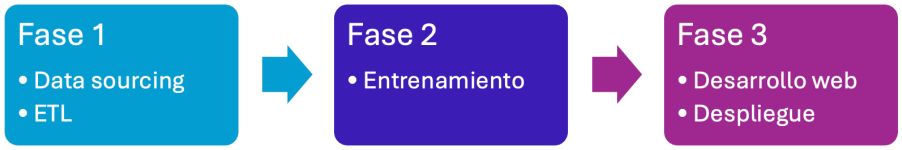

Figura 3.1: Esquema de la investigación.

prueba, se estima de forma indirecta y para cada una de las máquinas o tipo de ejercicio el RM o repetición máxima del cliente (máxima carga que el usuario es capaz de levantar en una repetición). En base a esto, el segundo objetivo planteado en este trabajo es encontrar el mejor modelo que permitan estimar esta variable con precisión, sin necesidad de perder sesiones de entrenamiento destinadas a realizar la prueba y sin necesidad de incurrir en prácticas inseguras (estimación directa).

Para ello, el presente trabajo de investigación se ha llevado a cabo en tres fases (Figura 3.1), las cuales incluyen distintas etapas: 1) obtención (*data sourcing*), manejo y limpieza de los datos (ETL); 2) entrenamiento de los modelos de aprendizaje automático y evaluación del rendimiento (entrenamiento); y 3) desarrollo de la plataforma web (desarrollo web) y despliegue. A continuación se resume el orden de las etapas.

La recopilación de los datos proviene de la anotación manual, por parte del monitor del gimnasio, de diversas variables de rendimiento deportivo y condición física de los usuarios. Estas anotaciones se realizan por tipo de movimiento (ejercicio). Como se ha comentado anteriormente, el centro cuenta con diversas máquinas *Nautilus* (Jones, 1970) para un entrenamiento integral y efectivo del cuerpo[1]:

- *pulldown*: movimiento de tracción vertical con foco en los dorsales y trabajo secundario de los deltoides posteriores y bíceps.
- *chest press*: movimiento de empuje horizontal con foco en los pectorales y trabajo secundario en los deltoides anteriores y tríceps.
- *leg press*: movimiento de empuje de piernas con foco en los cuádriceps, glúteos e isquiotibiales.
- *hip thrust*: movimiento de extensión de cadera con foco en los glúteos y trabajo secundario en los isquiotibiales, erectores espinales y cuádriceps.
- *row rear*: movimiento de tracción horizontal con foco en los deltoides posteriores y trabajo secundario en los romboides, trapecio medio e inferior, y músculos de la parte superior de la espalda.
- *shoulder press*: movimiento de empuje vertical con foco en los deltoides posteriores, laterales y anteriroes y trabajo secundario en el tríceps braquial, trapecio superior, y músculos de la parte superior del pectoral.

Estos registros son posteriormente trasladados a formato digital, usando plantillas de cálculo y unidades de almacenamiento de Google. Esta tarea influye significativamente en la calidad y cantidad de los datos. Consecuentemente, la precisión de las estimaciones de los modelos de aprendizaje automático dependen, en gran medida, del diseño y de la calidad de esta recopilación. Los datos provienen de la recompilación de las variables mencionadas en las distintas máquinas.

Aunque inicialmente el Centro Suma iba a proporcionar datos para el trabajo de un considerable número de usuarios, el tamaño muestral final con el que contamos es de $n = 30$ usuarios, lo que entendemos bastante limitante a la hora de entrenar modelos que den lugar a estimaciones con un bajo error.

Las variables recopiladas para cada usuario y su significado se resumen en la Tabla 3.1:

[1] Para más información consultar Centro Suma.

Variable	Descripción
num_sesiones	Número de entrenamientos llevados a cabo desde el inicio.
fecha	Fecha del entrenamiento.
ejercicio	Tipo de movimiento ejecutado.
peso_semana (carga)	Peso (kg) levantado en un determinado ejercicio en entrenamiento regular (semanal).
peso_test	Peso (kg) levantado en un determinado ejercicio en prueba periódica (descartada).
minutos / TUL (*Time Under Load*)	Duración (minutos) del ejercicio.
minutos_totales / tiempo total	Duración (minutos) del entrenamiento completo.
repeticiones_test	Número de veces que se repite el movimiento (fase de contracción y fase de relajación) en un ejercicio (descartada).
repeticiones	Estimación del número de repeticiones considerando que una repetición media se ejecuta en 20 segundos (solo se usa en la aplicación).
RM	Peso máximo (kg) que el usuario es capaz de levantar en un único movimiento (repetición máxima) obtenido a partir de (González-Badillo & Gorostiaga-Ayestarán, 1997): $$RM = \frac{peso\_test}{1{,}0278 - (0{,}0278 \times repeticiones\_test)}$$
altura	Altura (cm).
peso_corporal	Peso (kg).
imc	Índice de masa corporal.
edad_metabólica	Diferencia entre el índice metabólico basal del individuo y el índice metabólico basal de un grupo de individuos de la misma edad.
agua_percen	% de agua en el cuerpo.
grasa_visceral	% de grasa visceral.
masa_osea	% de masa osea.

grasa_total_percen	% de grasa total.
grasa_brazo_izq_percen	% de grasa en brazo izquierdo.
grasa_brazo_dcho_percen	% de grasa en brazo derecho.
grasa_pierna_izq_percen	% de grasa en pierna izquierda.
grasa_pierna_dcha_percen	% de grasa en pierna derecha.
grasa_tronco_percen	% de grasa en tronco.
músculo_total	% de músculo total.
músculo_brazo_izq	% de músculo en brazo izquierdo.
músculo_brazo_dcho	% de músculo en brazo derecho.
músculo_pierna_izq	% de músculo en pierna izquierda.
músculo_pierna_dcha	% de músculo en pierna derecha.
músculo_tronco	% de músculo en tronco.

Tabla 3.1: Descripción de las variables utilizadas en el trabajo.

La recopilación de los datos se realiza en dos etapas. Por una parte, se registran las variables de rendimiento deportivo en casa sesión de entrenamiento (ejercicio, peso_semana, minutos y minutos_totales). Estos registros tienen una periodicidad semanal en usuarios regulares. Por otra parte, cada trimestre se realiza una prueba de rendimiento para analizar los avances obtenidos. Esta prueba registra el peso y las repeticiones ejecutadas en cada movimiento en las condiciones de la prueba, facilitando así el cálculo del RM. Como comentamos anteriormente, esta última será la variable respuesta de nuestro modelo y es utilizada para comparar el rendimiento deportivo de los usuarios a lo largo del tiempo. Un incremento en esta variable significa necesariamente un incremento de las repeticiones con el mismo peso o un incremento del peso con las mismas repeticiones. Por lo tanto, indica una mejora en el rendimiento deportivo siempre que se incrementa. A su vez, una mejora en esta métrica debería ir asociada a una mejora tanto a nivel de salud como física, donde gradualmente se mejoran las métricas de composición corporal. Estas últimas, son registradas en el momento de inscripción en el gimnasio y en cada prueba trimestral.

Anotar que las variables de peso_test y repeticiones_test no se han incluído en el entrenamiento de los modelos, ya que se obtienen a través de un cálculo directo con la variable objetivo.

Llegado a este punto resaltar que, en esta investigación, hemos considerado exclusivamente la última observación disponible por usuario anterior a la última prueba de rendimiento realizada. Este hecho es fruto de la falta de información, ya que la empresa está en proceso de trasladar información pasada a formato digital. Además, al considerar una observación por usuario, evitamos el incumplimiento de independencia entre observaciones.

3.2. Selección de software y herramientas

En este trabajo hemos utilizado R Shiny (Chang et al., 2023) y otras librerías de soporte para el desarrollo íntegro de la plataforma web (Wickham, 2016; Auguie, 2017; Wickham et al., 2019; Müller, 2020; Sievert, 2020; Chang y Ribeiro, 2021; R Special Interest Group on Databases (R-SIG-DB) et al., 2022; Thieurmel y Perrier, 2022; Bryan, 2023; Chang y Cheng, 2023; McGowan y Bryan, 2023; Müller, 2023; Perrier et al., 2023; Xie et al., 2023; Signorell, 2024). Shiny se ha convertido en una herramienta innovadora y flexible que permite construir aplicaciones web con una interfaz interactiva en el entorno del lenguaje de programación estadístico R (R Core Team, 2021). Shiny es un marco de aplicación para desarrollar software utilizado en la visualización de datos para análisis estadístico en tiempo real, sin que el desarrollador tenga conocimientos avanzados en programación web (Chang et al., 2023). Destaca por su capacidad de transformar análisis realizados en R en aplicaciones web interactivas. Esta característica lo convierte en una herramienta muy valiosa, no solo para la investigación académica sino también para el fomento de la colaboración, donde los análisis y visualizaciones muy complejos podrían ser compartidos de manera accesible y comprensible (Sievert, 2020). Es una herramienta que facilita la colaboración entre compañeros y mejora la comunicación de resultados estadísticos, mejorando así la reproducibilidad y transparencia de la investigación.

La estructura de una aplicación Shiny se basa en dos componentes principales: una interfaz de usuario (UI) y un servidor (*backend*). En términos más generales, la UI define el aspecto y la disposición de la aplicación, mientras que el servidor es responsable de la lógica y el control del procesamiento de datos. Esta separación facilita el desarrollo de la aplicación, permitiendo al desarrollador centrarse en el aspecto técnico y al mismo tiempo en el diseño sin gran tecnicismo (Beeley & Sukhdeve, 2018). En resumen, Shiny ha demostrado ser una herramienta importante para la visualización y análisis estadístico. Proporciona una base muy sólida para el desarrollo de aplicaciones web, sin necesitar de muchas habilidades en desarrollo web. No obstante, Shiny puede presentar algunas limitaciones cuando buscamos un alto nivel de personalización. Por este motivo, sus desarrolladores han permitido la integración en su entorno de otros lenguajes de programación como HTML y CSS, los cuales han sido utilizados en este trabajo. Estos lenguajes son la base del desarrollo *frontend* (interfaz) y permiten una personalización completa de la aplicación web.

Por último, cabe mencionar que la plataforma ha sido hospedada en shinyapps.io. Se trata de un servidor exclusivo para aplicaciones Shiny, desarrollado por RStudio. Es un servicio que ofrece alta escalabilidad, ya que cuenta con una tarifa en función del uso. Además, cuenta con un accesso SSL encriptado por definición, lo cual lo hace un entorno seguro.

3.3. Técnicas de preprocesado

La estadística es una disciplina fundamental en la investigación científica y el análisis de datos. Dos técnicas importantes en el tratamiento de los datos son la *winsorización* y las imputaciones. Estas metodologías permiten manejar valores extremos y datos faltantes, mejorando la calidad y la validez de los análisis estadísticos (Tukey, 1962; Enders, 2022).

La *winsorización* es una técnica utilizada para limitar el impacto de los valores atípicos en los análisis estadísticos. Se trata de una técnica que lleva el nombre del estadístico Charles P. Winsor, en honor a algunos de sus trabajos que la influenciaron como el de Hastings et al. (1947) y que consiste en reemplazar los valores extremos de una distribución por valores más cercanos a la media o mediana de la distribución. La *winsorización* se basa en la idea de que los valores extremos pueden distorsionar las medidas de tendencia central y dispersión. Al sustituir estos valores por los percentiles más altos y más bajos dentro de un rango especificado (generalmente el 5 % y 95 %), se reduce su influencia, permitiendo una representación más generalista de los datos. Una de las principales limitaciones de la *winsorización* es que, aunque mitiga el impacto de los valores atípicos, también puede eliminar información valiosa contenida en esos extremos. Además, la elección del umbral de *winsorización* es subjetiva y puede variar dependiendo del investigador y del contexto del estudio. En este trabajo se han

entrenado modelos con *winsorización* en las colas (5 % y 95 %). Como hemos referido en el apartado de base de datos, la recopilación de los datos por parte de la empresa se realiza de forma manual y, por lo tanto, está sujeta a errores tipográficos. Esto, en conjunto con la natural distribución de los datos, genera datos atípicos que dificultan el correcto entrenamiento de los modelos de aprendizaje automático. En las Figuras 5.1 y 5.2 (Capítulo 5) se aprecian las variables que están sujetas a este problema.

La imputación es otra técnica crucial en el tratamiento de los datos, especialmente para abordar el problema de datos faltantes. Esta metodología consiste en reemplazar los valores ausentes con estimaciones razonables basadas en la información disponible. Las técnicas de imputación se fundamentan en la premisa de que la ausencia de datos no debe llevar a la pérdida de información valiosa. Las metodologías varían desde imputaciones simples, como la media o la mediana, hasta métodos más complejos, como la imputación múltiple y el algoritmo *Expectation-Maximization* (EM) (Enders, 2022). Aunque las imputaciones mejoran la calidad de los datos, también introducen suposiciones que pueden no ser siempre válidas. Además, los métodos de imputación pueden ser inadecuados si los datos faltantes no son al azar (MAR - *Missing At Random*), lo que puede sesgar los resultados (Rubin, 1987; Morita, 2021). Como fue referido anteriormente, consideramos el tamaño muestral pequeño y limitante, ya que las técnicas de aprendizaje automático se benefician de una gran cantidad de datos donde se puedan descubrir patrones y tendencias no visibles al ojo humano. Considerando este aspecto, es imperativo realizar imputaciones que permitan utilizar observaciones prácticamente completas que, en otro caso, serían descartadas. En este trabajo hemos optado por realizar imputaciones de media a las covariables, dejando la variable objetivo intacta.

La *winsorización* y las imputaciones son técnicas fundamentales en el tratamiento de datos atípicos y faltantes en la estadística. Aunque cada metodología tiene sus ventajas y limitaciones, su uso adecuado puede mejorar significativamente la calidad de los análisis estadísticos y la validez de las conclusiones. En general, la utilización de estas dos técnicas han incrementado considerablemente el rendimiento de los modelos desarrollados. Cabe mencionar que tanto la *winsorización* como la imputación han sido efectuadas únicamente a las covariables, mientras que la variable respuesta se mantuvo en su estado original. Los casos con datos faltantes en la variable respuesta han sido excluidos.

3.4. Modelos de aprendizaje automático

En el ámbito de la inteligencia artificial y la ciencia de datos, los modelos de aprendizaje automático se erigen como herramientas fundamentales para la creación de sistemas capaces de aprender y adaptarse a partir de datos. Este capítulo se centra en la metodología detrás del bosque aleatorio y de los modelos de regresión aditiva generalizada, los cuales han sido elegidos para este trabajo por su gran aplicabilidad a diversos problemas.

3.4.1. Bosque aleatorio

El bosque aleatorio (Breiman, 2001; James et al., 2021) es una técnica que se utiliza tanto para la clasificación como para la regresión. Para entender el funcionamiento de este método, es necesario introducir primero el concepto de árbol de decisión (Quinlan, 1986; Hastie et al., 2009). Un árbol de decisión se construye mediante una serie de divisiones binarias basadas en los valores de las características (variables) de los datos. Al dividir un conjunto de datos en subconjuntos más pequeños basados en las características de entrada, se forma una estructura similar a un árbol. Cada nodo representa una regla de decisión y cada rama representa el resultado de esa regla específica. El nodo raíz es el nodo superior del árbol que representa todo el conjunto de datos. Se divide en dos o más conjuntos homogéneos. Los nodos internos representan una característica y una regla de decisión (por ejemplo, $(X_i \leq \text{valor})$ en base a la cual se divide el conjunto de datos. Las ramas o aristas conectan los nodos y representan el resultado de una regla de decisión. El nodo terminal representa una predicción final (etiqueta de clase o valor continuo). Existen distintos algoritmos de árboles de decisión como el ID3 y el

C4.5 (Quinlan, 1986) o el CART (Breiman et al., 1984). No obstante, todos ellos siguen una estrategia similar. En el Algoritmo 1 se resumen los pasos del funcionamiento general de un árbol de decisión.

Algoritmo 1 Árbol de decisión

1. Para cada nodo, se selecciona la característica que mejor separa los datos en diferentes clases utilizando métricas de error. Dividimos el espacio de predictores, es decir, el conjunto de posibles valores para $X_1, X_2, \ldots, X_p$, en J regiones distintas y no superpuestas, $R_1, R_2, \ldots, R_J$. Para cada observación que cae en la región R_j ($j = 1, \ldots, J$), hacemos la misma predicción, que es simplemente la media de los valores de la respuesta para las observaciones de entrenamiento en R_j. Como métricas de error, es habitual utilizar la entropía, la impureza de Gini o la ganancia de información (IG) para los problemas de clasificación, así como el error cuadrático medio (MSE) para la regresión. No obstante, existe un amplio número de métricas disponibles.

   - Entropía:
   $$H(S) = -\sum_{i=1}^{c} p_i \log_2(p_i),$$
   donde p_i es la proporción de elementos de la clase i ($i = 1, \ldots, c$ en el conjunto de datos S y c el número de clases.

   - Impureza de Gini:
   $$G(S) = 1 - \sum_{i=1}^{c} p_i^2.$$

   - IG:
   $$IG(S, A) = H(S) - H(S, A)$$
   donde $H(S)$ es la entropía del conjunto de datos antes de la división y $H(S, A)$ la entropía después de la división del atributo A.

   - Error cuadrático medio (MSE):
   $$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

2. De manera general, el objetivo será encontrar las regiones $R_1, \ldots, R_J$ que minimicen el RSS, por ejemplo, dado por:
   $$\sum_{j=1}^{J}\sum_{i \in R_j}(y_i - \hat{y}_{R_j})^2,$$
   donde $\hat{y}_{R_j}$ es la media de la variable respuesta de las observaciones de entrenamiento contenidas en la región R_j. La característica seleccionada se utiliza para dividir el nodo en dos nodos hijos, de manera que se maximice la ganancia de información (o se minimice la impureza de Gini / error de regresión).

3. Este proceso se repite recursivamente para cada nodo hijo hasta que se cumpla un criterio de parada, como la profundidad máxima del árbol o el número mínimo de observaciones en un nodo.

4. Para evitar el sobreajuste, se puede aplicar una técnica de poda, eliminando nodos que no proporcionen una mejora significativa en la precisión del modelo.

El bosque aleatorio es una extensión del concepto de árboles de decisión, introducido por Breiman (2001), que mejora la precisión del modelo mediante la agregación de múltiples árboles de decisión.

La construcción de un bosque aleatorio se describe de manera resumida en el Algoritmo 2.

Algoritmo 2 Bosque aleatorio

1. *Bootstrap Aggregating* (Bagging): se generan múltiples subconjuntos de datos a partir del conjunto de datos original mediante muestreo con reemplazo. Cada subconjunto se utiliza para entrenar un árbol de decisión independiente.

2. En cada división de un árbol, se selecciona un subconjunto aleatorio de características en lugar de todas las características disponibles. Esto reduce la correlación entre los árboles y mejora la diversidad del bosque.

3. Cada árbol de decisión se entrena de manera independiente utilizando su subconjunto de datos y características.

4. Las predicciones de todos los árboles se combinan para obtener la predicción final del bosque aleatorio. Para clasificación se utiliza el voto mayoritario

$$\hat{y} = \text{moda}\{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_B\},$$

mientras que para regresión se utiliza el promedio de las predicciones individuales

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} \hat{y}_b,$$

con $b = 1, \ldots, B$ conjuntos de entrenamiento.

En el bosque aleatorio, cada árbol se entrena con una muestra *bootstrap* del conjunto de datos, dejando una porción de los datos fuera de la muestra (*out-of-bag*). Esto resulta interesante porque será posible hacer predicciones para estas observaciones usando los árboles donde no se han considerado las observaciones OOB durante el ajuste, generando así una única predicción para cada observación. Si repetimos este proceso para cada una de las observaciones de nuestro conjunto de datos, podremos obtener medidas de error globales (error OOB) de manera directa, sin necesidad de utilizar una muestra test.

Adicionalmente, los bosques aleatorios también proporcionan una medida de la importancia de las variables, basada en la reducción de la impureza (Gini) o en el incremento del error OOB al permutar aleatoriamente los valores de la variable en cuestión.

En resumen, los árboles aleatorios y los bosques aleatorios son poderosas técnicas de aprendizaje supervisado que combinan la simplicidad de los árboles de decisión con el poder de la agregación y la selección aleatoria de características, ofreciendo modelos robustos y precisos para una variedad de tareas de clasificación y regresión. La ventaja del bosque aleatorio incluye su capacidad para manejar grandes conjuntos de datos con mayor dimensionalidad y proporcionar estimaciones de la importancia de variables (Liaw & Wiener, 2002). También es menos propenso al sobreajuste en comparación con los árboles de decisión individuales. Sin embargo, son computacionalmente más exigentes y menos interpretables que los árboles de decisión individuales.

3.4.2. Modelos de regresión aditiva generalizada

Para entender la metodología de los Modelos de regresión Aditiva Generalizada (GAM), es necesario introducir el funcionamiento general de la regresión lineal y de los modelos lineales generalizados (Nelder & Wedderburn, 1972). La regresión es una técnica fundamental en el análisis estadístico que

permite modelar la relación entre una variable dependiente Y y un conjunto de p variables independientes $\mathbf{X} = X_1, \ldots, X_p$ por medio de la siguiente expresión

$$Y = \mu(\mathbf{x}) + \varepsilon$$

donde $\mu(X) = E[Y \mid X]$ es el valor esperado condicional de la respuesta Y dado X, y ε es una medida del error.

El modelo de regresión más básico es el modelo de regresión lineal, que asume una relación lineal entre la variable de respuesta y los predictores, definida como

$$\mu(X) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon$$

donde $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ son los parámetros del modelo lineal a ajustar, y ahora se asume que el error ε está distribuido normalmente con media cero y varianza constante σ^2 (Montgomery et al., 2021).

Cabe mencionar que el modelo de regresión linear requiere el cumplimiento de los siguientes supuestos:

- Linealidad. La variable respuesta debe estar linealmente relacionada con el vector de variables independientes $\mathbf{X}$.
- Independencia. Las observaciones deben ser independientes entre sí.
- Homocedasticidad. Los residuos deben tener varianza constante para cualquier nivel de $\mathbf{X}$.
- Normalidad. Los residuos deben seguir una distribución Normal.
- Multicolinealidad. En el caso de la regresión linear múltiple, las variables independientes no deben estar muy correlacionadas entre sí.

A medida que las aplicaciones de la regresión lineal se ampliaron, surgió la necesidad de modelos más flexibles que pudieran manejar tipos de datos más variados y distribuciones de error diferentes. Los Modelos Lineales Generalizados (GLM) fueron introducidos para abordar estas limitaciones. Los GLM se componen de tres componentes (Nelder y Wedderburn, 1972):

1. Componente aleatorio: especifica una distribución de probabilidad de la familia exponencial para la variable respuesta
$$f(y; \theta, \phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\},$$
donde y es la variable respuesta, θ es el parámetro de la media, ϕ es el parámetro de dispersión, $a(.)$, $b(.)$ y $c(.)$ son funciones específicas de la distribución.
2. Componente sistemático: describe el predictor lineal, el cual es una combinación lineal de las variables explicativas
$$\eta = X\beta,$$
donde η es el predictor lineal, X la matriz de variables explicativas y β el vector de coeficientes.
3. Función *link*: una función $g(.)$ que conecta la media de la variable respuesta $\mu(X) = E[Y \mid X]$ con el predictor lineal
$$g(\mu) = X\beta.$$

Algunas funciones *link* comunes incluyen:

- *Link* identidad: $g(\mu) = \mu$
- *Link* logit: $g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$

- *Link* logarítmico: $g(\mu) = \log(\mu)$

La estimación de los parámetros del GLM se realiza generalmente mediante el uso del método de máxima verosimilitud (MLE) pudiendo utilizar para ello el algoritmo IRLS (*Iteratively Reweighted Least Squares*) (Nelder y Wedderburn, 1972; Björck, 1996).

Los GLM asumen los siguientes supuestos:

- Linealidad de los predictores con la variable respuesta.
- Independencia de las observaciones entre sí.
- La variable respuesta sigue una distribución de la familia exponencial.
- $g(.)$ relaciona correctamente $E(Y \mid X)$ a η.

La necesidad de flexibilizar los anteriores supuestos condujeron a los modelos de regresión aditiva generalizada (GAM), que representan una extensión flexible de los modelos lineales generalizados, permitiendo la inclusión de relaciones no lineales entre las variables predictoras y la variable respuesta. Los GAM fueron introducidos por Hastie y Tibshirani (1986) como una forma de suavizar las relaciones entre variables mediante la suma de funciones no paramétricas. Esta metodología permite modelar efectos no lineales de covariables de una manera que es a la vez interpretable y computacionalmente eficiente.

La estructura básica de un GAM se puede definir de la siguiente forma:

$$g(\mu) = \beta_0 + f_1(X_1) + \cdots + f_p(X_p),$$

donde f_j son funciones suaves y desconocidas. La estimación de las funciones f_j se realiza de manera no paramétrica usando técnicas de suavizado como splines (Wood, 2017) o regresión tipo *kernel* (Wand & Jones, 1994), entre otras. Dentro de las más comúnmente utilizadas nos encontramos con los splines cúbicos penalizados.

Los splines cúbicos se definen por una serie de puntos de control (*knots*) y una función cúbica suave entre estos puntos. El criterio de suavizado incluye una penalización para evitar el sobreajuste, balanceando entre la fidelidad a los datos y la suavidad de la curva ajustada. Este balance se controla típicamente mediante un parámetro de suavizado λ, que se selecciona usando métodos como la validación cruzada (Wood, 2017). En este caso, el ajuste de un GAM se realiza generalmente a través del método de máxima verosimilitud penalizada. En este enfoque, se maximiza una función de verosimilitud que incluye un término de penalización para las funciones de suavizado. Este proceso se puede llevar a cabo mediante diversos algoritmos, entre ellos el método de *Fisher scoring* modificado o el algoritmo de *backfitting* (Jennrich y Sampson, 1976; Wood, 2017).

La validación de los modelos GAM implica evaluar la bondad del ajuste, la significación de los términos suavizados y la capacidad predictiva del modelo. Herramientas comunes para el diagnóstico incluyen gráficos de residuos, gráficos de función parcial y criterios de información como AIC (*Akaike Information Criterion*) y BIC (*Bayesian Information Criterion*). La significación de los términos suavizados se puede evaluar utilizando un test de razón de verosimilitud o criterios basados en test de hipótesis (Wood, 2017).

Los modelos GAM asumen algunos de los supuestos anteriores y además los siguientes:

- Aditividad. El efecto de los predictores en la variable respuesta se puede sumar, lo que permite el suavizado del efecto con funciones no-lineales.
- Suavizado. La relación entre los predictores y la variable respuesta es suave.
- Independencia de las observaciones entre sí.
- La variable respuesta sigue una distribución de la familia exponencial.
- $g(.)$ relaciona correctamente $E(Y \mid X)$ a η.

3.5. Validación cruzada

La validación cruzada es una técnica fundamental en el análisis estadístico y el aprendizaje automático para evaluar la capacidad de generalización de un modelo. Esta metodología es esencial para evitar el sobreajuste y asegurar que el modelo seleccionado tenga un buen desempeño en datos no vistos previamente.

El método de validación cruzada *k-fold* es uno de los más utilizados debido a su balance entre sesgo y varianza. En este método, el conjunto de datos se divide en k subconjuntos o *folds* de tamaño aproximadamente igual. El proceso de validación cruzada se realiza en k iteraciones, donde en cada iteración, uno de los k subconjuntos se utiliza como conjunto de validación y los restantes $k-1$ subconjuntos se utilizan como conjunto de entrenamiento (James et al., 2021). La estimación del error se obtiene promediando los errores de validación obtenidos en cada iteración. La formulación matemática del error promedio es:

$$\text{CV}_{kfold} = \frac{1}{k}\sum_{i=1}^{k}\text{Error}^{(i)},$$

donde $\text{Error}^{(i)}$ representa el error de validación en la i-ésima iteración.

El método *Leave-One-Out* (LOO) es un caso especial de la validación cruzada *k-fold* donde k es igual al número total de observaciones en el conjunto de datos (n) (James et al., 2021). En este método, se entrena el modelo n veces, cada vez usando $n-1$ observaciones como conjunto de entrenamiento y una sola observación como conjunto de validación. La fórmula del error promedio para LOO es:

$$\text{CV}_{LOO} = \frac{1}{n}\sum_{i=1}^{n}\text{Error}_i,$$

donde Error_i es el error de predicción para la i-ésima observación cuando se utiliza como conjunto de validación.

Dado que en cada iteración se utiliza $n-1$ observaciones para entrenar el modelo, este método aprovecha al máximo el conjunto de datos disponible, lo cual es particularmente beneficioso en conjuntos de datos pequeños (Arlot & Celisse, 2010). También proporciona una estimación no sesgada del error de generalización, ya que cada observación se valida de manera independiente. Sin embargo, entrenar el modelo n veces puede ser computacionalmente ineficiente, especialmente para modelos complejos o conjuntos de datos grandes. Además, las estimaciones del error tienden a tener una varianza alta, ya que cada conjunto de entrenamiento en LOO es muy similar al conjunto completo, pero se diferencia solo en una observación. Esto puede llevar a resultados inestables (James et al., 2021).

En conclusión, la validación cruzada *k-fold* ofrece un buen equilibrio entre el coste computacional y la precisión de la estimación del error, mientras que el LOO maximiza el uso de los datos disponibles a cambio de un mayor coste computacional y con una mayor varianza en sus estimaciones. La elección del método adecuado depende del tamaño muestral y de los recursos computacionales disponibles. Debido a la limitación del tamaño muestral, en este trabajo hemos adoptado una validación cruzada LOO para el modelo GAM. Sin embargo, para el bosque aleatorio hemos utilizado una validación cruzada *10-fold*, ya que con este método el modelo reporta mejor rendimiento.

3.6. Métricas de error

Para evaluar la precisión de los modelos predictivos empleados en este estudio, se utilizaron las siguientes métricas estadísticas de precisión: error absoluto medio (MAE), error cuadrático medio (MSE), raíz del error cuadrático medio (RMSE) y el coeficiente de determinación (R^2). A continuación, se define su formulación matemática (Willmott y Matsuura, 2005; Hyndman y Koehler, 2006; Chai y Draxler, 2014; James et al., 2021):

- Error absoluto medio (MAE): se calcula como el promedio de los valores absolutos de los errores entre las predicciones y los valores reales. Su fórmula es:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|,$$

donde y_i es el valor observado, $\hat{y}_i$ es el valor predicho y n es el número total de observaciones.

- Error cuadrático medio (MSE): el MSE es el promedio de los cuadrados de los errores, es decir, la diferencia cuadrática media entre los valores estimados y el valor real (en este caso, norma L_2). Su fórmula es:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

- Raíz del error cuadrático medio (RMSE): el RMSE es simplemente la raíz cuadrada del MSE. Proporciona una medida de la magnitud del error en las mismas unidades que la variable de interés, lo que facilita su interpretación. Se calcula mediante la fórmula:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}.$$

- Coeficiente de determinación (R^2): el R^2, también conocido como coeficiente de determinación, es una medida estadística que representa la proporción de la variabilidad de la variable respuesta que es explicada por la/s covariable/s a través del modelo de regresión. Un R^2 de 1 indica que el modelo predice perfectamente la variable dependiente, mientras que un R^2 de 0 indica que el modelo no mejora la predicción sobre el promedio simple. Se define como:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

donde $\bar{y}$ es el promedio de los valores observados.

Capítulo 4

Diseño del sistema

En el desarrollo de aplicaciones modernas, la arquitectura y el diseño del sistema son componentes cruciales que determinan no solo la funcionalidad, sino también la eficiencia y la seguridad de la aplicación. Este capítulo presenta una visión integral del diseño del sistema de nuestra aplicación Shiny, destacando los aspectos más importantes de su construcción y funcionamiento. A través de estas secciones, el capítulo ofrece una descripción del diseño del sistema, proporcionando una base sólida para comprender cómo se construye y opera la aplicación Shiny. Cada sección está diseñada para resaltar los aspectos técnicos y de diseño que contribuyen a la efectividad general de la aplicación.

La aplicación Shiny desarrollada combina una interfaz de usuario intuitiva con la arquitectura *backend*, asegurando una experiencia de usuario fluida y segura. Este capítulo se organiza en varias secciones para proporcionar una comprensión detallada de cada componente del sistema.

En la Sección 4.1 de diseño de la interfaz de usuario, abordamos los principios y decisiones de diseño que guiaron la creación de la interfaz. Se exploran los elementos visuales y funcionales que facilitan la interacción del usuario con la aplicación.

La Sección 4.2 de arquitectura del *backend* se divide en dos subsecciones donde se describe la estructura técnica subyacente que soporta la aplicación, asegurando su rendimiento y escalabilidad. La primera Subsección (4.2.1) incluye detalles sobre el mecanismo de autenticación y seguridad, la cual es una preocupación primordial en cualquier aplicación web. Aquí se describen en detalle en los mecanismos de autenticación implementados para proteger los datos de los usuarios y garantizar su privacidad. La segunda (Subsección 4.2.2), aborda la integración de los datos de usuario con la aplicación. En concreto, en este apartado se examina cómo los datos de usuario se recopilan, procesan y se integran, permitiendo funcionalidades avanzadas y una mayor usabilidad.

4.1. Diseño de la interfaz de usuario

A petición de la empresa, la aplicación se ha diseñado con foco en los dispositivos móviles, caracterizados por una pantalla de aproximadamente cinco pulgadas. No obstante, se ha programado la aplicación de forma responsiva, ajustándose de forma automática a la pantalla del usuario, pudiendo este acceder a la aplicación desde un móvil, tableta o ordenador. Parte del diseño y elección de gráficos son fruto de las peticiones de la empresa. No obstante, ha habido una gran flexibilidad y espacio para la creatividad del alumno. A continuación se muestra una colección de imágenes que resumen la aplicación (Figuras 4.1, 4.2, 4.3, 4.4, 4.5, 4.6, 4.7, 4.8, 4.9, 4.10, 4.11 y 4.12).

Una vez se accede a la dirección web de la plataforma, la primera interacción que tenemos con la aplicación es con la página de *login* (Figura 4.1). Esta página presenta un diseño básico pero funcional, donde simplemente hay que introducir nombre de usuario y contraseña, previamente configurados por el administrador.

Una vez realizado el *login* de forma exitosa, se muestra la página principal (Figura 4.2). La zona

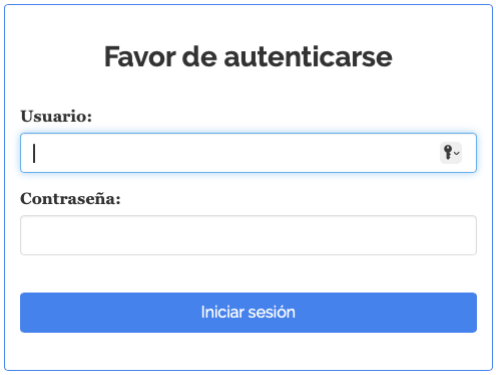

Figura 4.1: Página de *login*.

de menú se mantiene fija independientemente de la pestaña seleccionada. Se compone del logotipo de la empresa en la zona superior, seguido de una imagen del gimnasio (*banner*) y de un menú con cinco pestañas seleccionables. Cada una de estas pestañas nos lleva a una página de interacción distinta, justo por debajo de la página principal. La pestaña de Fuerza Máxima resume los datos de las pruebas trimestrales que se realizan, mientras que la de Registro de Sesiones se centra en los entrenamientos regulares (semanales). La pestaña de Composición Corporal reporta algunas métricas importantes de condición física que se detallarán más adelante, la de Análisis Comparativo compara el rendimiento deportivo y físico del usuario con la media de usuarios del centro y, por último, la pestaña de Contacto muestra un mapa con la localización del centro junto con información de contacto, como correo electrónico y dirección de la página web.

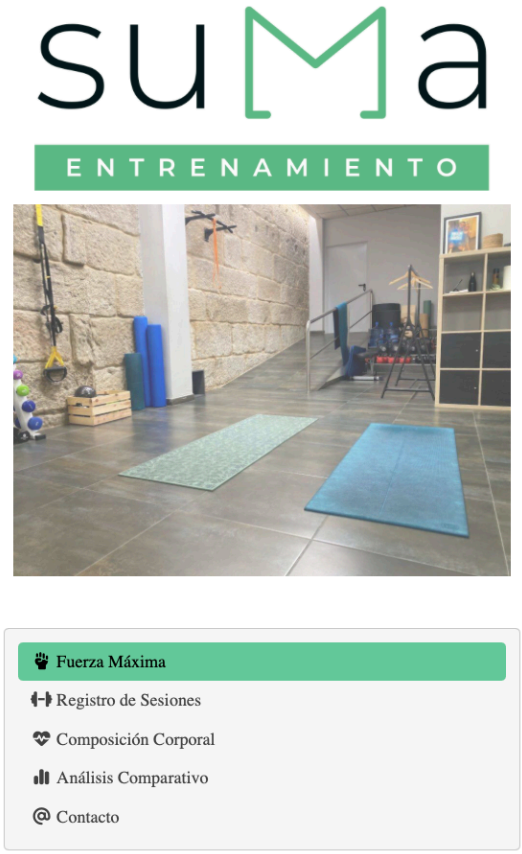

Figura 4.2: Menú principal de la aplicación web.

Dentro de la zona interactiva (debajo del menú principal) y para las pestañas Fuerza Máxima y Registo de Sesiones, existen dos bloques diferenciados. El bloque superior (Figura 4.3), contiene un filtro de fecha, donde el usuario puede filtrar un rango de fechas específico para analizar. A continuación, en el bloque inferior, aparece un menú seleccionable con el nombre del ejercicio físico e imagen de la máquina donde realizar dicho ejercicio. Al cambiar el tipo de ejercicio, se cambia de forma unísona la imagen de la máquina utilizada.

Indicar que, a lo largo del desarrollo, se han probado varias configuraciones de gráficos y gif's. Sin embargo, se ha tratado de construir una interfaz de usuario atractiva, de bajo coste computacional y técnico, premiando así la usabilidad, escalabilidad e interpretabilidad de la aplicación.

En la zona inferior de la zona interactiva, se muestran los resultados de cada pestaña. Empezando por la pestaña de Fuerza Máxima (Figura 4.4), en esta sección es posible constatar la evolución del RM (repetición máxima) en las pruebas trimestrales realizadas hasta la fecha actual. Cambios en el filtro de fecha o en el tipo de ejercicio, generan un cambio automático en los gráficos que se muestran.

En la pestaña de Registro de Sesiones (Figura 4.5) obtenemos una vista similar a la anterior. En este caso, se valoran los datos de los entrenamientos semanales, teniendo en cuenta el peso levantado (carga), las repeticiones y el tiempo bajo carga (TUL) por tipo de ejercicio.

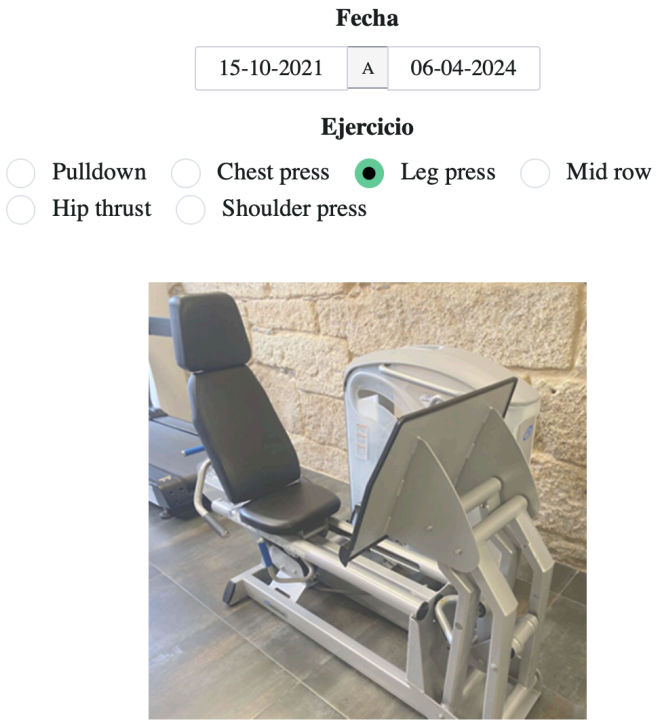

Figura 4.3: Filtros de fecha y ejercicio presentes en las pestañas Fuerza Máxima y Registro de Sesiones.

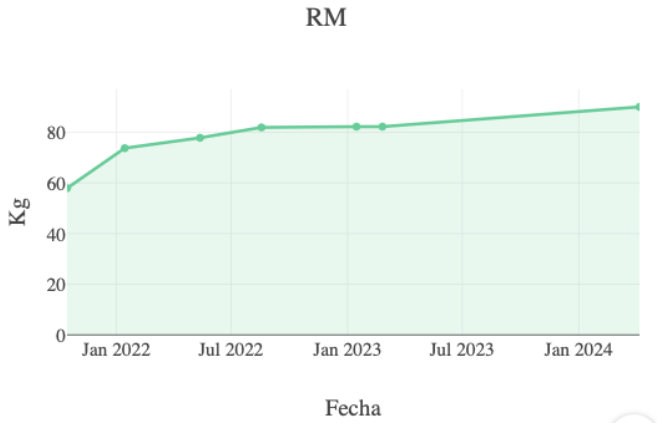

Figura 4.4: Evolución de la RM para un usuario dentro de la pestaña Fuerza Máxima.

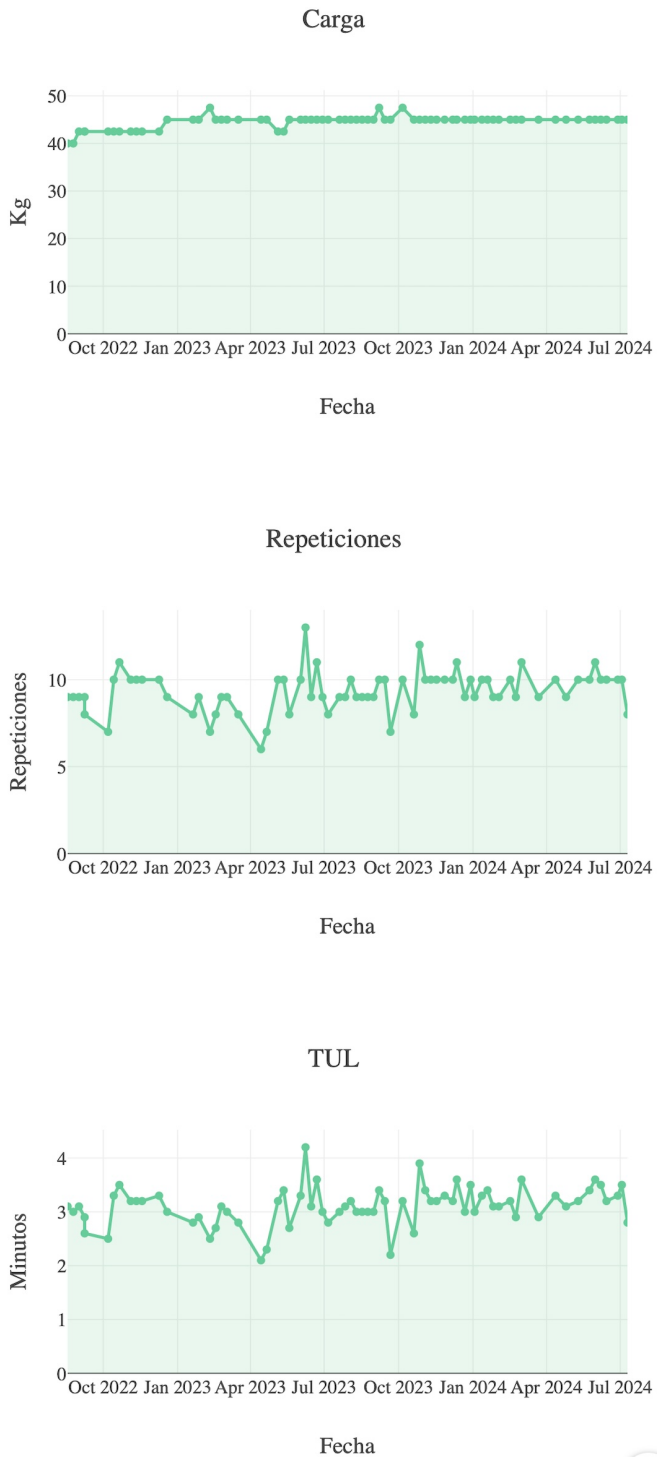

Figura 4.5: Evolución de la carga, número de repeticiones y TUL dentro de la pestaña Registro de Sesiones.

La pestaña de Composición Corporal tiene un menú adicional en la zona superior con tres pestañas seleccionables (Figura 4.6). En la pestaña General (Figura 4.7, panel izquierda) se muestran algunas

métricas como peso, altura, IMC, masa muscular (%), grasa (%), edad metabólica, densidad ósea y porcentaje de agua corporal. En la pestaña Segmentos Corporales (Figura 4.7, panel derecha) se analizan posibles diferencias de musculatura y grasa entre los miembros izquierdo y derecho del cuerpo (pierna y brazo). Tanto la pestaña General como la de Segmentos Corporales, proporcionan información de la última medición corporal disponible. Por último, la pestaña Histórico (Figura 4.8) nos reporta información en formato tabla de mediciones pasadas a efectos comparativos.

Figura 4.6: Menú dentro de la pestaña Composición Corporal.

La pestaña de Análisis Comparativo compara la media del RM (kg), Carga (kg) y TUL (minutos) por tipo de ejercicio de todos los usuarios del centro con el último registro del usuario autenticado (Figura 4.9). Además, se incorporan filtros de género y edad (Figura 4.10) para poder realizar comparaciones en función del interés del usuario.

Por último, la pestaña de Contacto (Figura 4.11) incorpora un mapa de Google con la localización, página web y correo electrónico del centro.

Cabe destacar que la plataforma cuenta con un modo administrador (Figura 4.12), donde se puede gestionar la incorporación o eliminación de usuarios, así como crear o modificar contraseñas. Esta página solo está disponible cuando nos autenticamos como administrador y dispone de dos botones (Actualizar estadísticas y Entrenar modelos) los cuales ejecutan dos ficheros distintos de R, desde Google Drive, que permiten actualizar los datos de la pestaña de Análisis Comparativo y entrenar los modelos de aprendizaje automático. Por debajo de estos botones, se ha incorporado una barra de progreso que le indica al administrador si el estado de la ejecución de los ficheros está inactivo, en ejecución o concluido.

4.2. Arquitectura del backend

La arquitectura del *backend* en Shiny juega un papel crucial en la gestión y procesamiento eficiente de datos, así como en la entrega de contenido dinámico y personalizado a los usuarios finales. Esta sección se enfoca en los principios fundamentales para diseñar y construir la infraestructura *backend* de la aplicación Shiny.

4.2.1. Autenticación de usuarios y seguridad

Para la autenticación de los usuarios se ha utilizado el paquete shinymanager (Thieurmel & Perrier, 2022). Este paquete permite una gestión integral de los usuarios y su autenticación. Como se puede ver en la Figura 4.13, una de las funciones que permite este paquete es agregar nuevos usuarios o editar los existentes. Para un nuevo registro, es necesario aportar un nombre de usuario. Además, se pueden configurar parámetros adicionales como fecha de inicio o expiración del acceso, así como especificar si el usuario tiene poder de administrador. En un nuevo registro, el paquete se encarga de sugerir una contraseña segura inicial, la cual puede ser cambiada por el usuario en su primer acceso a la aplicación.

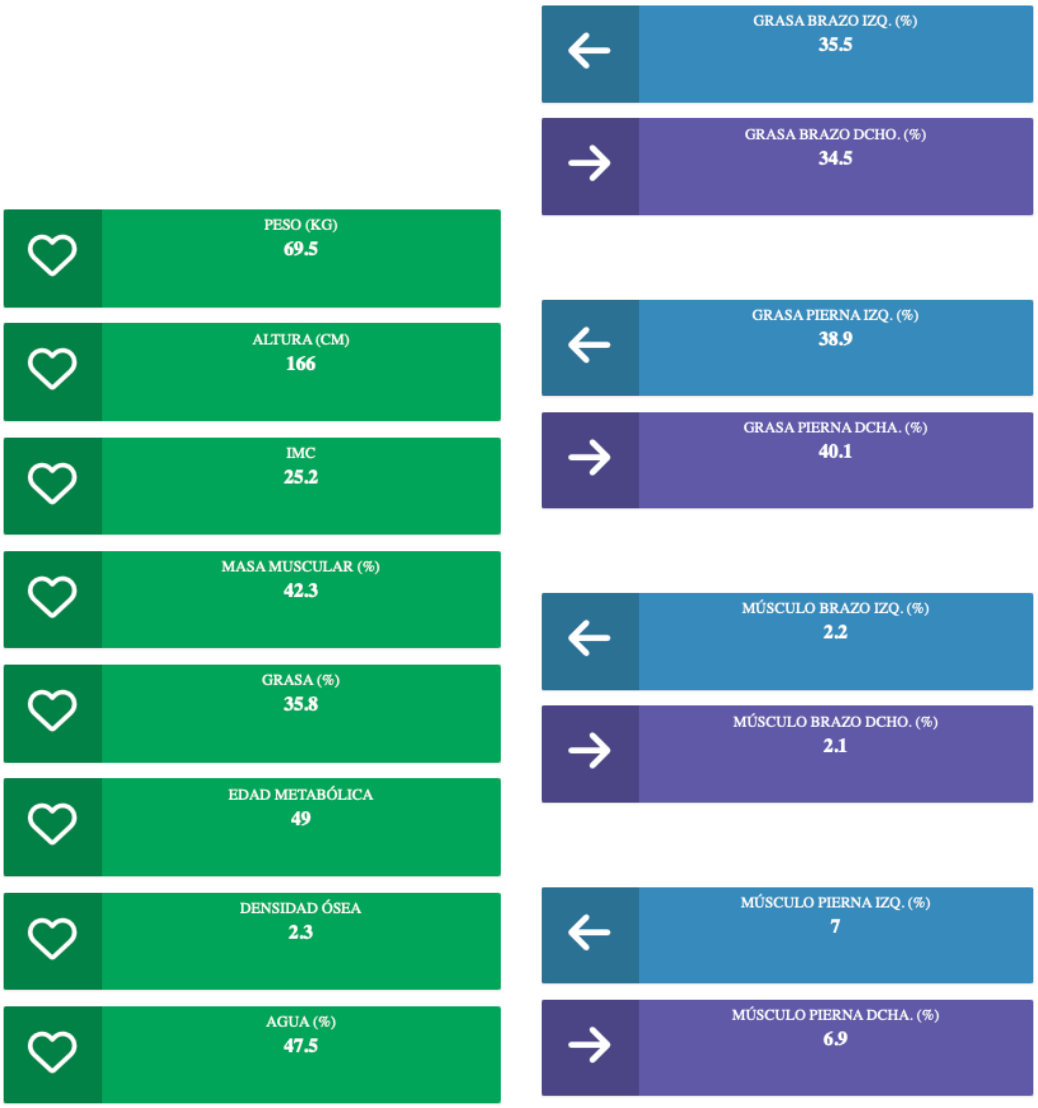

Figura 4.7: Pestaña General (panel de la izquierda) y pestaña Segmentos Corporales (panel de la derecha) dentro del menú Composición Corporal.

	fecha	altura	peso	imc	fecha_nacimiento	sexo	edad_metabolica	agua_percen	grasa_visceral	masa_osea	grasa_total_percen	grasa_brazo_izq_percen	grasa_brazo_dcho_percen
1	8/01/2021	173	77,3	25,8		masculino	40	55,80%	7	3	21,70%	15,60%	14,80%
2	26/03/2021	173	75	24,5		masculino	30	58,40%	6	3,1	18,3	13,6	12,9
3	7/05/2021	173	74,2	25,4		masculino	34	58,60%	6	3	19,7	15,2	14,3
4	10/08/2021	173	73,2	24,5		masculino	30	58,60%	6	3	18,3	13,8	12,5

Figura 4.8: Pestaña Histórico dentro del menú Composición Corporal.

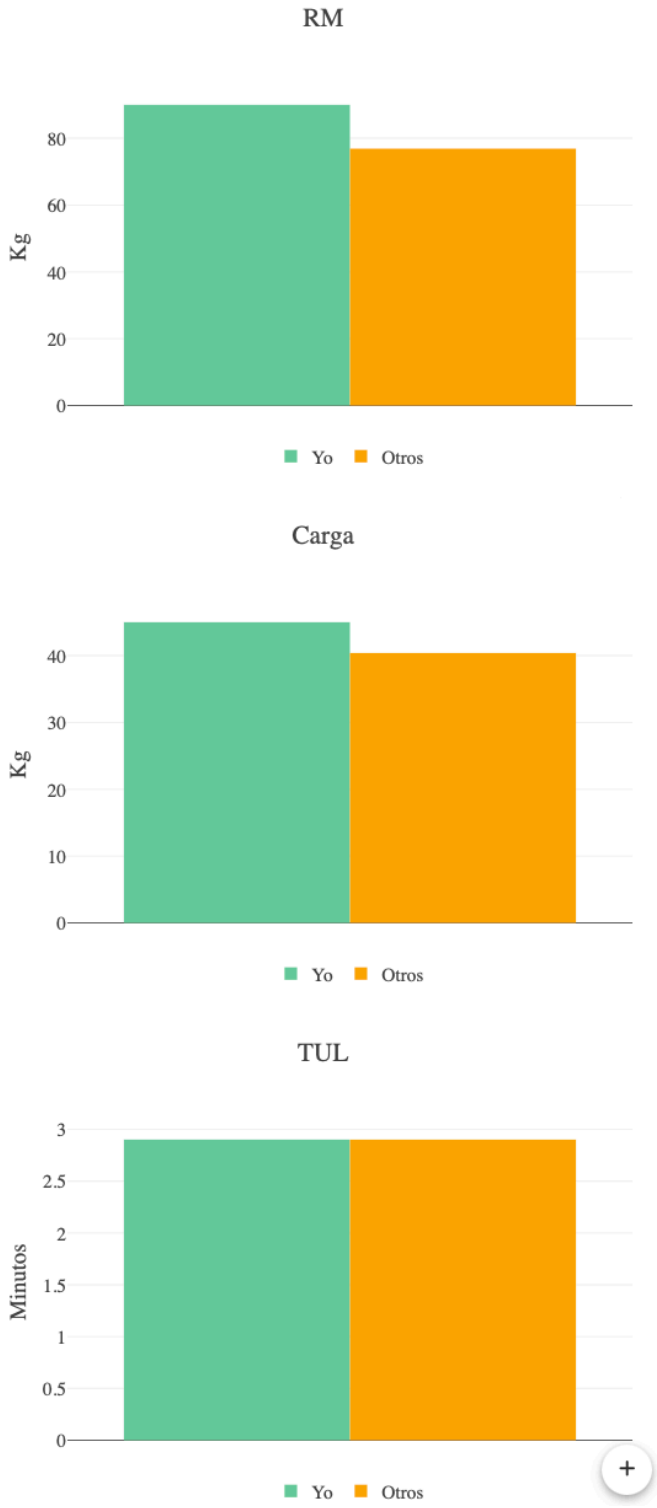

Figura 4.9: Pestaña de Análisis Comparativo.

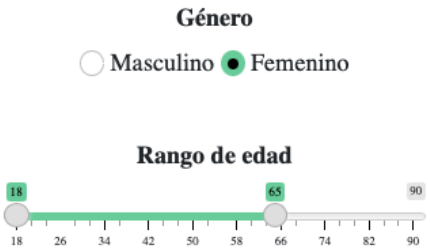

Figura 4.10: Filtro de género y edad dentro de la pestaña Análisis Comparativo.

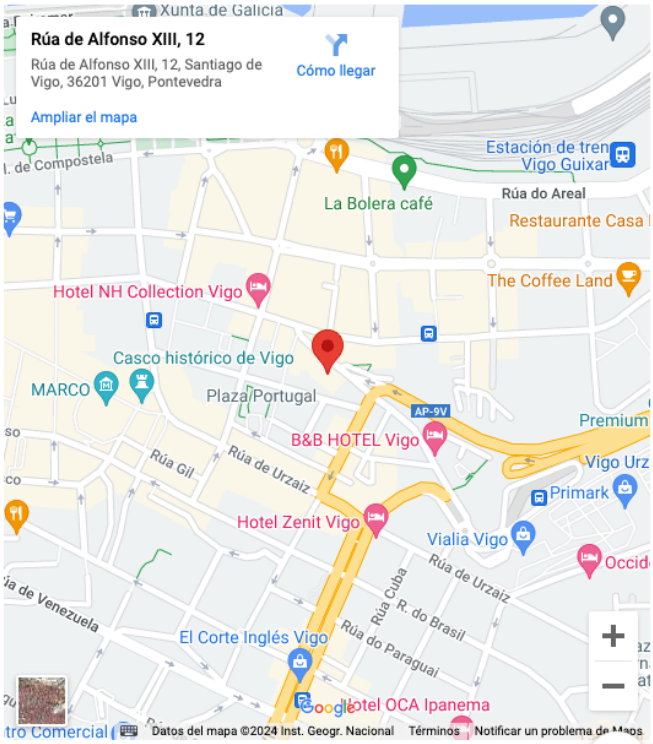

Figura 4.11: Vista de la pestaña de Contacto.

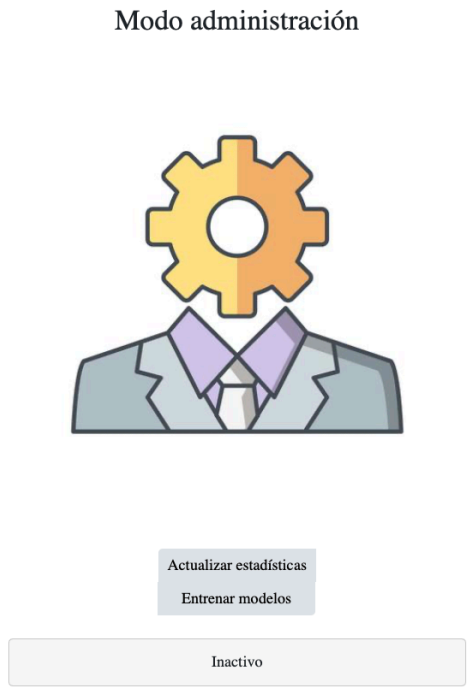

Figura 4.12: Vista del Modo Administrador.

Figura 4.13: Gestión de usuarios dentro del Modo Administrador.

Por otra parte, es posible cambiar la contraseña de los usuarios o pedir que se cambie de forma automática en el siguiente acceso (Figura 4.14).

Figura 4.14: Gestión de contraseñas dentro del Modo Administrador.

Para almacenar las credenciales se ha utilizado SQLite (Hipp, 2020), un sistema de gestión de bases de datos SQL, multiplataforma y de código abierto. Los nombres de usuario y contraseñas son guardados en un documento sqlite encriptado para mayor seguridad. La encriptación es un sistema donde una función matemática transforma una cadena de caracteres (usuario y contraseña) en una secuencia alfanumérica distinta pero única para posteriormente descifrarla mediante un algoritmo de desencriptación. Esto implica que en el caso de una brecha de información, no se trasmite la información en su estado de origen sino en una forma encriptada, y sin saber el algoritmo de desencriptación, dificulta su descifrado. Una vez se accede a la aplicación, la primera iteración del programa es la lectura de esta base de datos, verificando que el usuario y contraseña introducidos se encuentran dentro de la información registrada. Este documento es de vital importancia, siendo necesario efectuar copias de seguridad regulares. Como la aplicación es hospedada en shinyapps.io, el usuario no es capaz de acceder al código fuente o ficheros de la aplicación. Por lo tanto, la única forma de hacerlo sería a través de una brecha en el propio servidor de shinyapps.

4.2.2. Integración de los datos de usuario

Una vez el sistema reconoce al usuario, se le permite el acceso a la aplicación. La siguiente iteración del programa es acceder a Google Drive mediante los paquetes de R googledrive (McGowan & Bryan, 2023) y googlesheets4 (Bryan, 2023). En este proceso, el programa accede a un documento con el mismo nombre que el del usuario identificado. De esta forma, se garantiza que cada usuario solo accede a su información personal. Este documento contiene los datos de entrenamiento del usuario y es la base para la mayoría de las representaciones gráficas que se muestran en la aplicación.

Las estadísticas generales del centro son cargadas a través de un fichero común a todos los usuarios que contiene las medias de las variables previamente calculadas. Este documento es actualizado a través del botón Actualizar estadísticas (Figura 4.12). La elección de este procedimiento se ha llevado a cabo en base al tiempo de ejecución de la aplicación, que se considera elevado en el caso de calcular estas estadísticas generales cada vez que un usuario accede al sistema.

Capítulo 5

Estudio empírico

El presente capítulo ofrece un análisis detallado del proceso de modelado estadístico llevado a cabo en nuestro estudio. Este capítulo se estructura en tres secciones principales que abarcan desde el análisis inicial de los datos hasta la interpretación y discusión de los resultados obtenidos.

La Sección 5.1 se centra en el análisis exploratorio de datos, fase preliminar del análisis, donde se exploran y visualizan los datos para identificar patrones, tendencias y posibles anomalías. Este tipo de análisis es importante, ya que permite comprender mejor la estructura y las características de los datos, facilitando así la formulación de hipótesis y la selección de modelos adecuados.

La Sección 5.2 presenta los hallazgos del modelado estadístico. Aquí, se detalla la selección de variables y el rendimiento general del modelado estadístico. Se incluyen tablas, gráficos y métricas de desempeño que ilustran los ajustes y la precisión de los modelos empleados.

Por último, la Sección 5.3 se centra en la discusión, ofreciendo una interpretación crítica de los resultados obtenidos, relacionándolos con los objetivos del estudio y con la literatura existente. Se analizan las implicaciones de los hallazgos y se discuten las limitaciones del estudio.

5.1. Análisis exploratorio de datos

La parte empírica de este trabajo comienza con el análisis exploratorio de datos, mostrando distintas representaciones gráficas que incorporan tanto medidas de posición como de dispersión, distribuciones de frecuencias para variables categóricas, y análisis de correlación.

Las Figuras 5.1 y 5.2 contienen diagramas de cajas para algunas de las variables consideradas. La Figura 5.1 muestra las variables RM, carga y TUL por tipo de ejercicio. Se observa que tanto el RM como la carga tienen una distribución de peso similar, donde el tipo de ejercicio de *leg press*, de manera general, presenta mayores valores en todo su rango intercuartílico y en su mínimo y máximo. De forma opuesta, el ejercicio *shoulder press* presenta menores valores que todos los demás ejercicios, posiblemente porque se trata de un movimiento aislado (engloba un músculo en vez de varios). Considerando la variable TUL (*Time Under Load*), se observa una menor disparidad entre ejercicios, donde los movimientos generalmente se ejecutan entre 2 y 3.5 minutos, tiempos esperados si el ejercicio es ejecutado con la carga correcta. Las tres variables mencionadas presentan algunos valores atípicos.

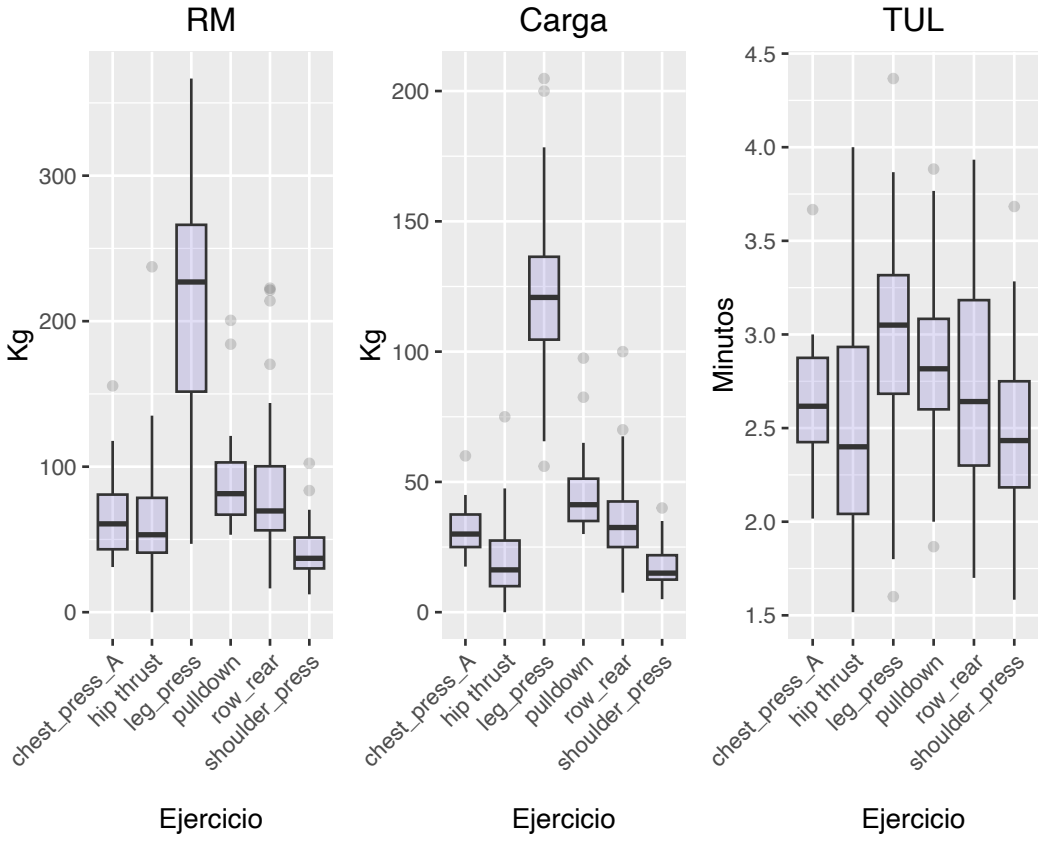

Figura 5.1: Diagramas de cajas para las variables RM, Carga y TUL considerando los distintos tipos de ejercicios.

La Figura 5.2 muestra algunas de las variables que se consideran relevantes para el estudio. Su distribución nos indica que la mayoría de usuarios tiene entre 40 y 55 años y más de 100 sesiones de trabajo. El músculo total es una variable complementaria de la grasa total y se observa que la mediana de la población se encuentra en torno al 40 %. El tiempo total de las sesiones se encuentra entre los 15 y los 17.5 minutos, lo que refleja la eficiencia en tiempo de esta metodología de entrenamiento, como se ha comentado en la revisión literaria. Por último, cabe indicar que tanto la altura como la densidad osea toman valores esperados para el tipo de usuario de este centro, con un bajo número de datos atípicos.

La Figura 5.3 representa la distribución morfológica y de género de los usuarios de la muestra. La morfología humana se refiere al estudio de la forma y estructura del cuerpo humano. Existen distintas formas de cuerpo como el ectomorfo, el mesomorfo y el endomorfo. Cada tipo se caracteriza por una serie de atributos físicos específicos: los ectomorfos suelen ser delgados y con poco tejido adiposo; los mesomorfos, musculosos y atléticos; y los endomorfos, con mayor tendencia a acumular grasa corporal. La morfología humana no solo se centra en la apariencia externa, sino también en la comprensión de cómo estas características pueden influir en aspectos como la salud, el rendimiento físico y las predisposiciones a ciertas enfermedades (Sheldon et al., 1940). En la Figura 5.3, es posible apreciar que la mayoría de usuarios son mesomorfos (56.7 %), seguidos por los endomorfos (23.3 %) y, por último, los ectomorfos (20 %). En cuanto al género, sobre tres cuartas partes son mujeres (76.7 %).

A continuación, se muestra un estudio de correlación por tipo de ejercicio considerando únicamente las variables continuas (Figuras 5.4, 5.5, 5.6, 5.7, 5.8 y 5.9). En general, se observa una alta correlación

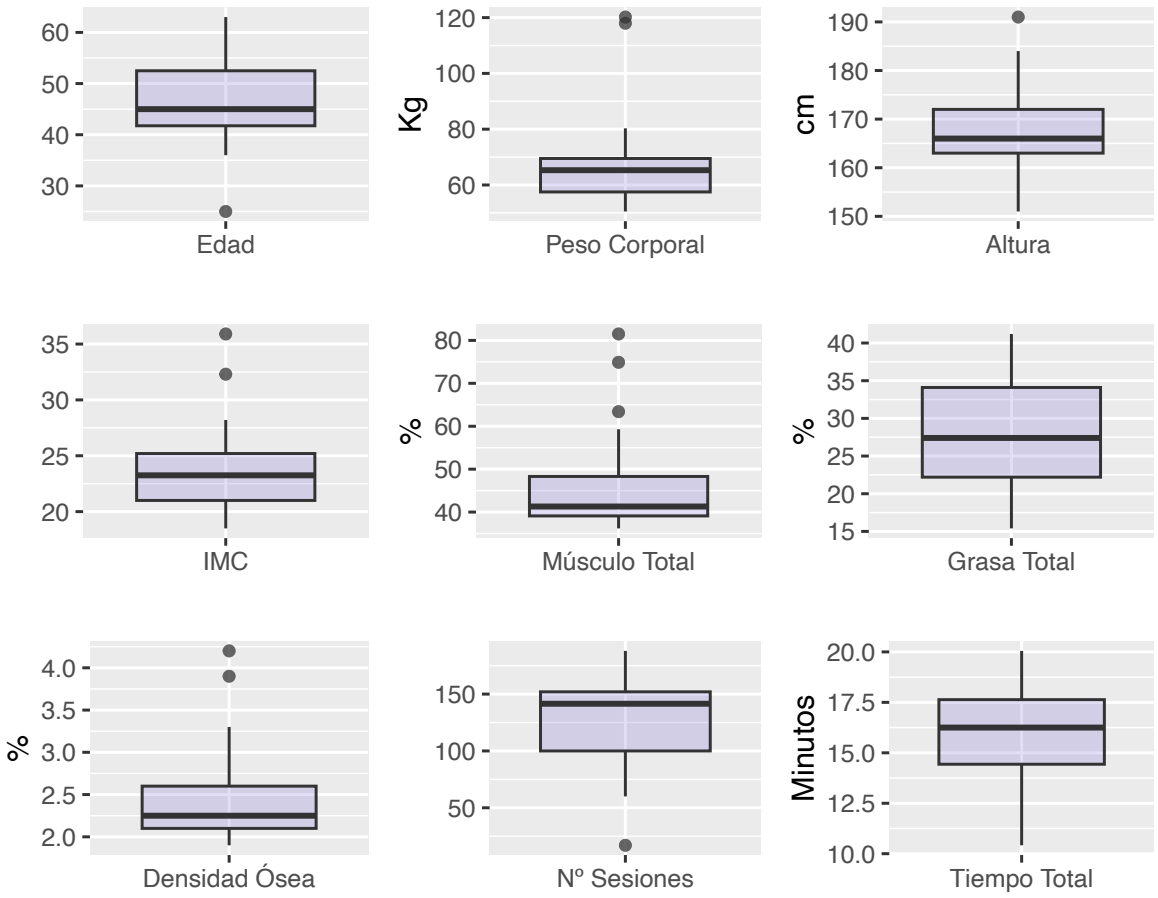

Figura 5.2: Diagramas de cajas para la variable Edad, Peso Corporal, Altura, IMC, Músculo Total, Grasa Total, Densidad Ósea, Nº Sesiones y Tiempo Total.

(mayor al 0.9) de la variable RM con el peso levantado en las pruebas (peso_test) y con el peso levantado en los entrenamientos semanales (peso_semana). Por otra parte, el peso corporal, el IMC y el músculo total también presentan una alta correlación con la variable objetivo, de entre 0.6 a 0.8, en la mayor parte de los ejercicios. La altura y el número de sesiones generalmente presenta una correlación moderada que puede oscilar entre el 0.3 y el 0.5. Las demás combinaciones de variables no presentan un coeficiente de correlación relevante. Este conjunto de figuras también permite inspeccionar la relación de las variables explicativas con la variable respuesta. Una simple visualización de los *scatterplots* permite identificar que algunos predictores, en general, tienen una relación relativamente lineal (peso_test, peso_semana, reps, peso_corporal) con el RM, mientras que las demás suelen presentar curvaturas más acentuadas, dependiendo del ejercicio en cuestión. Recordar que tanto la variable peso_test como repeticiones no han sido incorporadas al modelado estadístico.

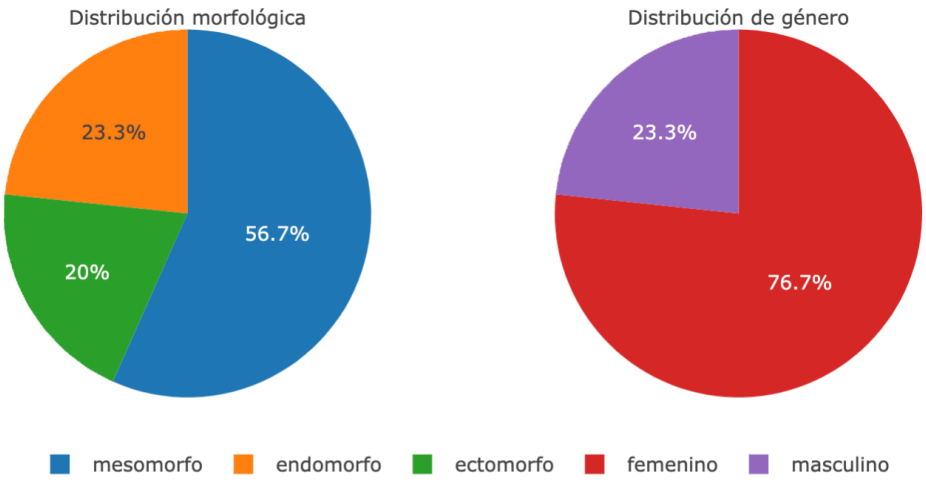

Figura 5.3: Distribución morfológica y de género de los usuarios de la muestra de estudio.

5.2. Resultados

Nuestro objetivo en esta sección es el de predecir la fuerza máxima (RM) mediante las covariables disponibles (Tabla 3.1) y los métodos seleccionados. Se muestran los modelos finales ajustados utilizando las dos técnicas propuestas: el Bosque aleatorio (Liaw & Wiener, 2002) y los modelos GAM (Hastie et al., 2009; Wood, 2017). Cabe destacar, que los paquetes Kuhn y Max (2008), Wickham et al. (2019) y Peterson y Carl (2020) también han sido cruciales en el desarrollo de esta sección. Adicionalmente, se resumen las métricas obtenidas para cada uno de los modelos. Es importante resaltar que se ha ajustado un modelo para cada tipo de ejercicio considerando las dos metodologías con el fin de decidir, en una fase posterior, el mejor modelo para cada tipo de ejercicio. Cabe destacar que hemos optimizado el ajuste de los modelos en función del coeficiente de determinación en la validación cruzada, seleccionando aquellos ajustes que proporcionan un mayor valor de esta métrica.

En primer lugar, a través del bosque aleatorio, hemos estudiado la importancia de las variables (Figura 5.10). Este estudio nos permite tener una idea inicial de cuáles son las variables más relevantes para predecir el RM. Además, nos permite un análisis comparativo con el modelo GAM, observando si existen discrepancias importantes en la selección de variables. No obstante, no se ha utilizado este método para discriminar variables en el ajuste del modelo. Los hiperparámetros finales del bosque aleatorio se especifican en la Tabla 5.1. Se realizó una búsqueda aleatoria de hiperparámetros con una ventana de búsqueda que se resume de la siguiente forma: *mtry*[1]: {1, 2, 3, 4, 5, 10, 15, 20}; *ntree*[2]: {50, 100, 200, 300}; *nodes*[3]: {10, 20, 30}. Se ha limitado la optimización a un máximo de 100 combinaciones distintas de hiperparámetros para el bosque aleatorio.

[1]Número de variables seleccionadas al azar como candidatos en cada división.
[2]Número de árboles.
[3]Número máximo de nodos terminales permitidos en los árboles del bosque aleatorio.

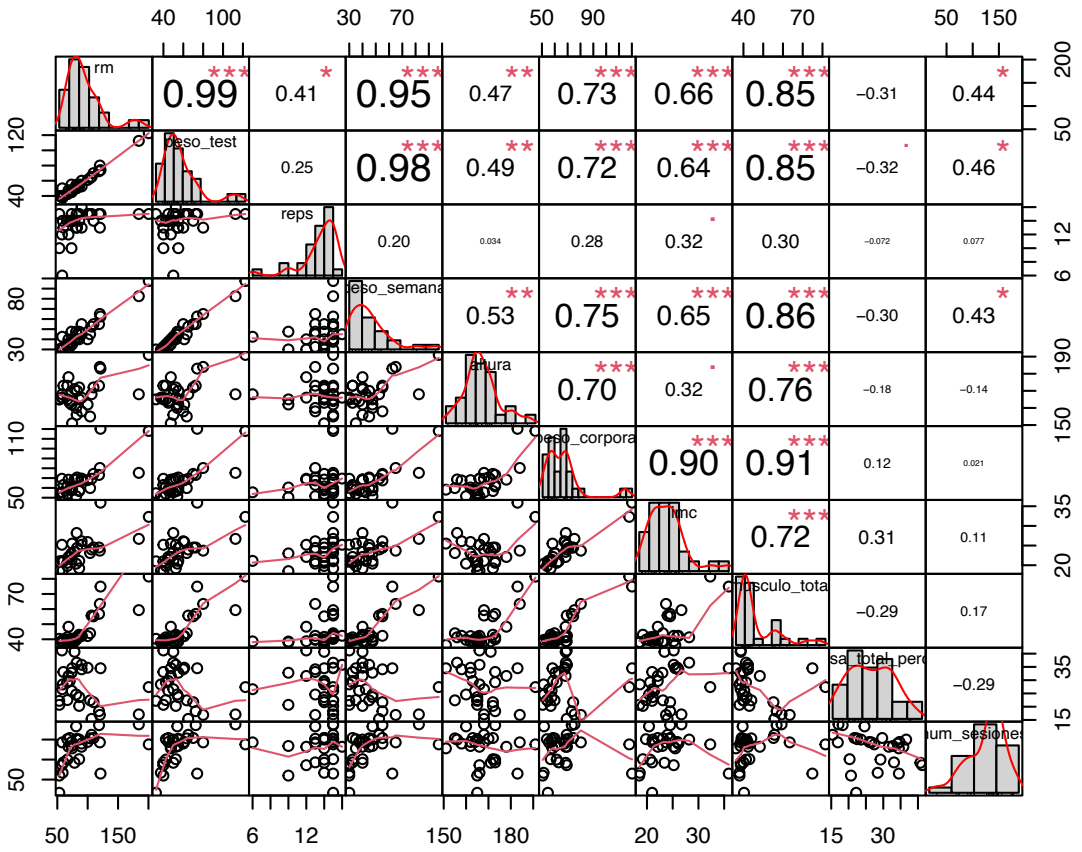

Figura 5.4: Gráfico de correlaciones entre variables continuas para el ejercicio *pulldown*.

Ejercicio	mtry	ntree	nodes
Pulldown	10	50	25
Chest Press	1	50	19
Leg Press	4	200	25
Hip Thrust	20	100	23
Row Rear	3	50	19
Shoulder Press	20	50	19

Tabla 5.1: Hiperparámetros utilizados en cada una de los modelos por tipo de ejercicio utilizando el Bosque aleatorio.

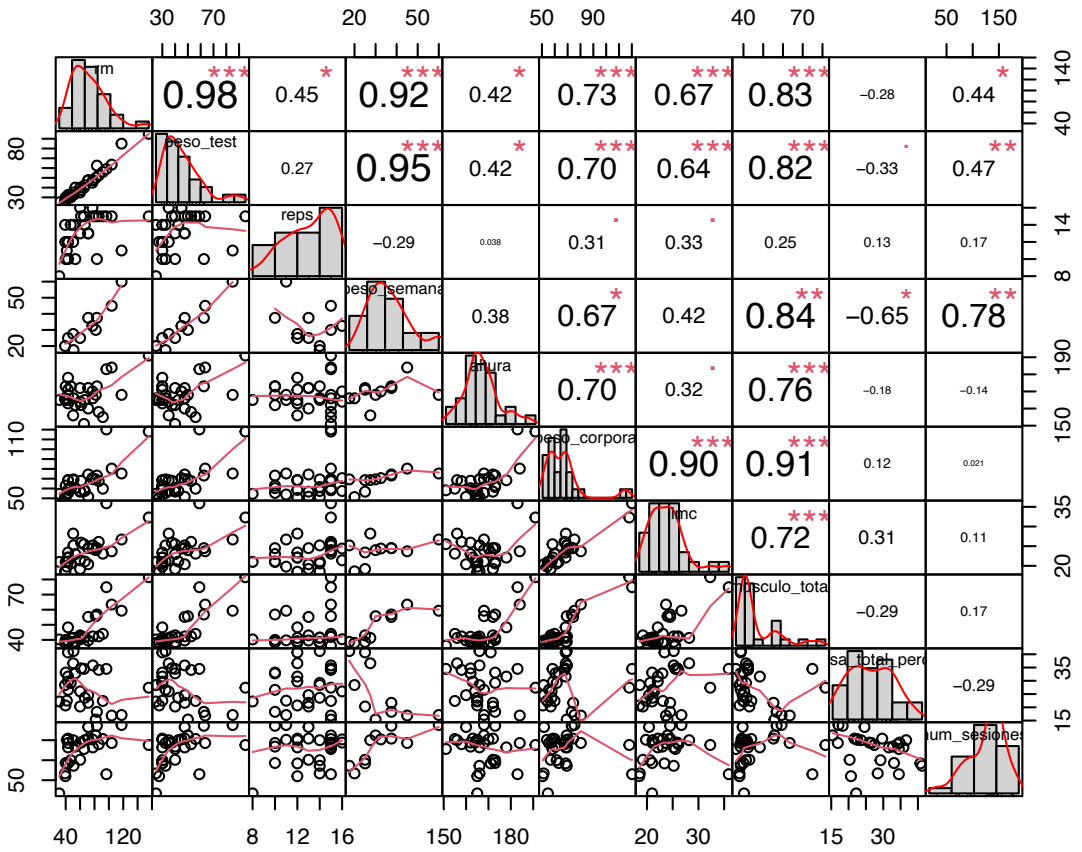

Figura 5.5: Gráfico de correlaciones entre variables continuas para el ejercicio *chest press*.

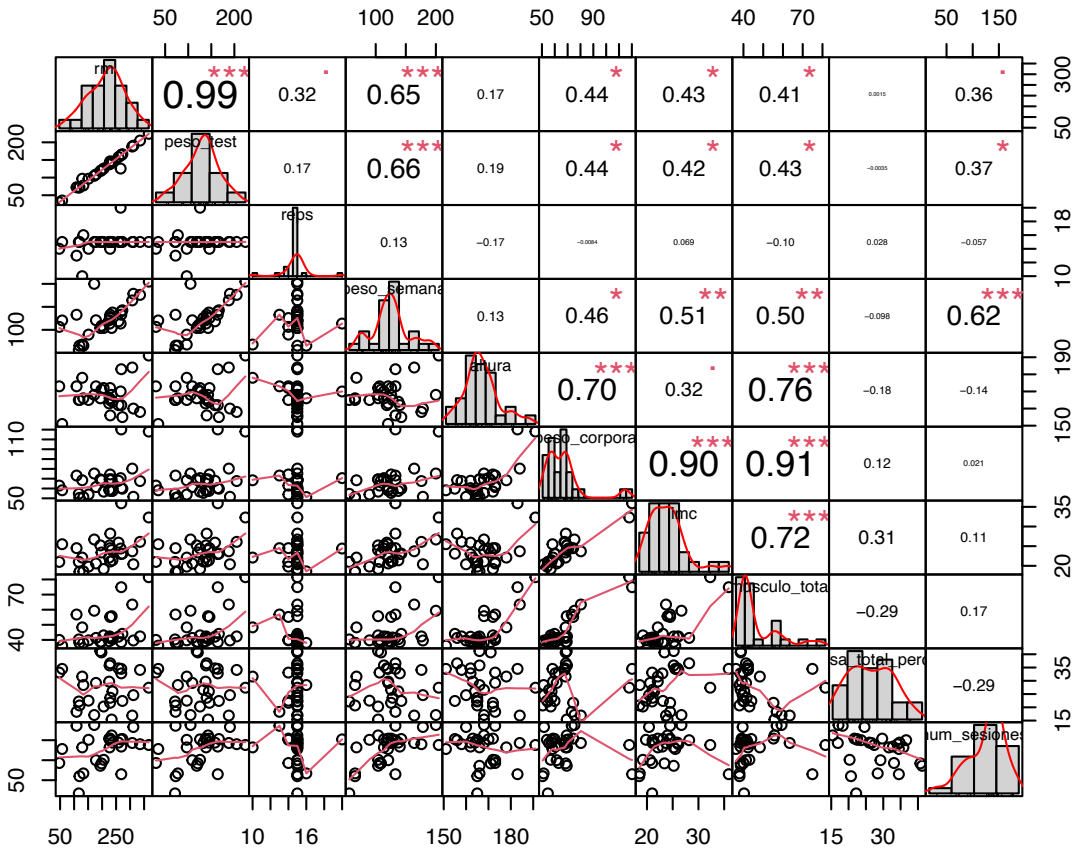

Figura 5.6: Gráfico de correlaciones entre variables continuas para el ejercicio *leg press*.

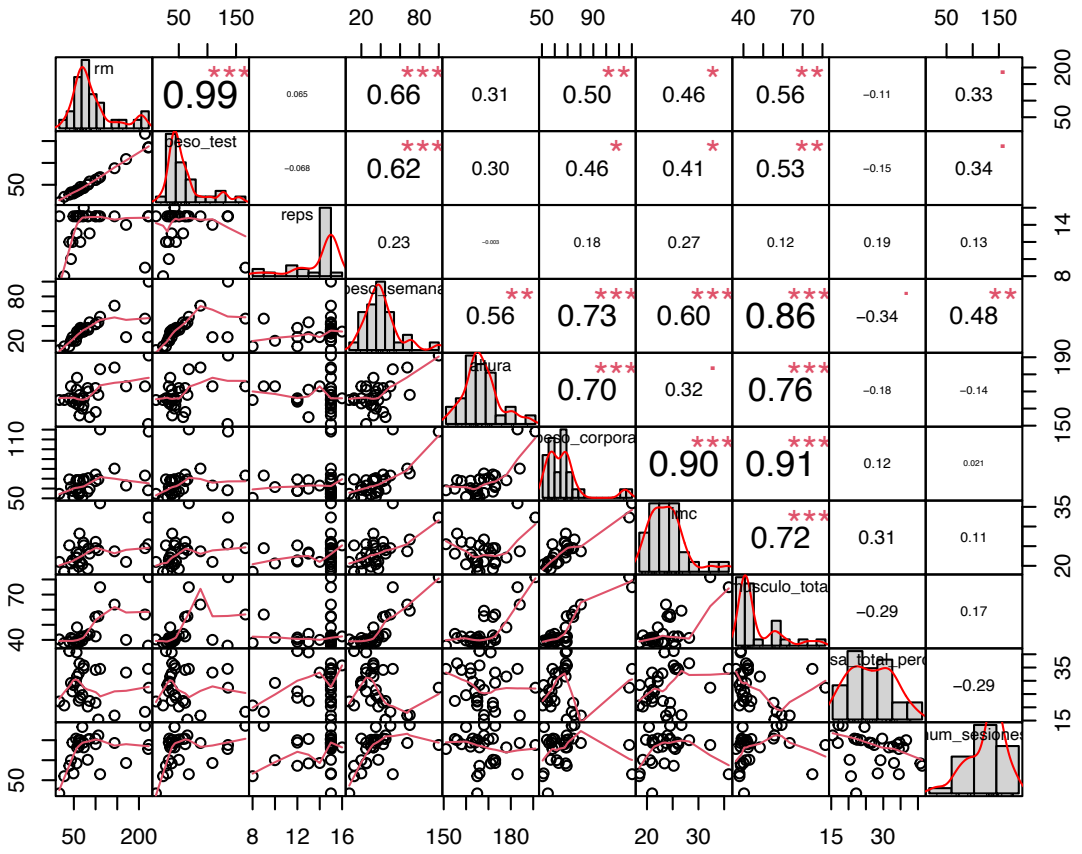

Figura 5.7: Gráfico de correlaciones entre variables continuas para el ejercicio *hip thrust*.

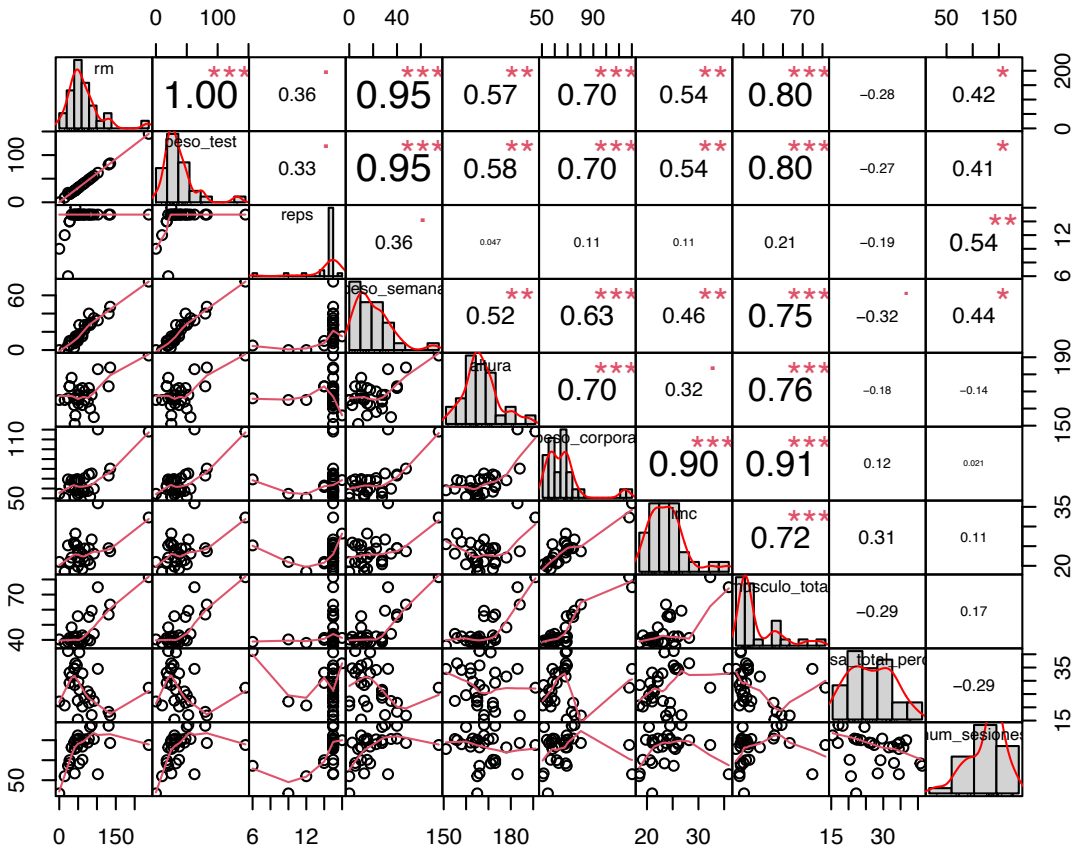

Figura 5.8: Gráfico de correlaciones entre variables continuas para el ejercicio *row rear*.

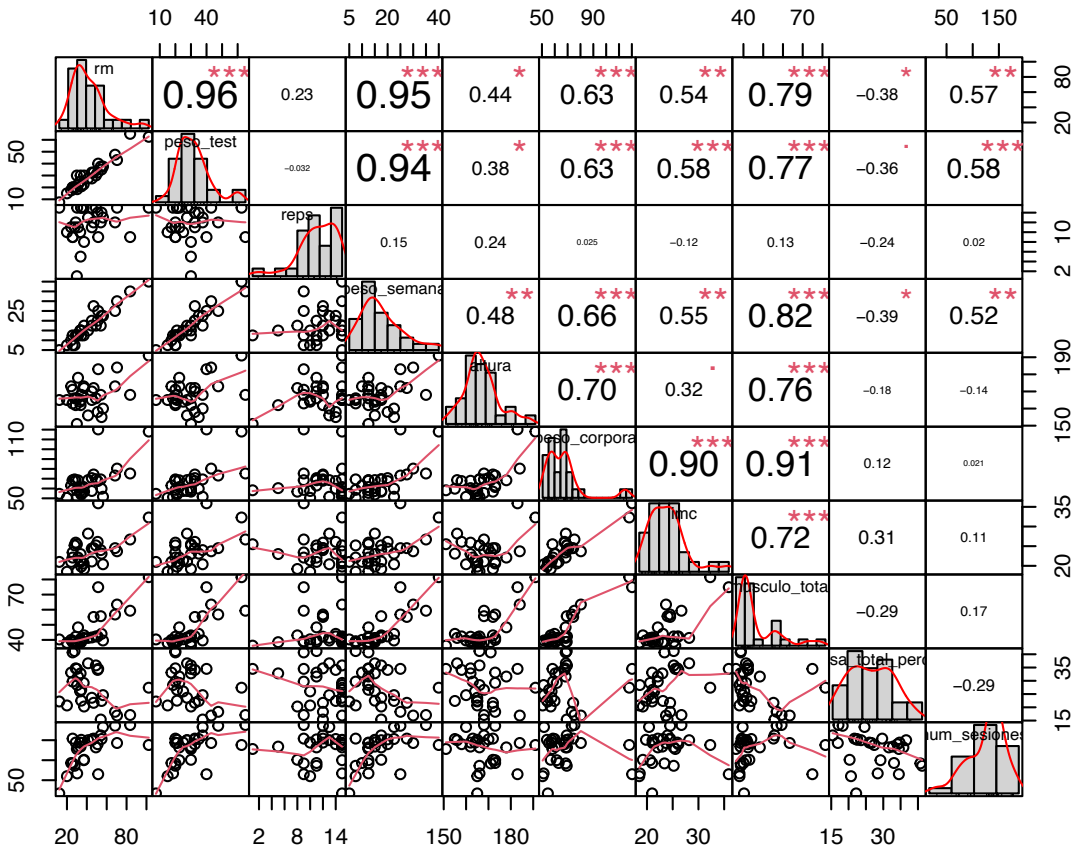

Figura 5.9: Gráfico de correlaciones entre variables continuas para el ejercicio *shoulder press*.

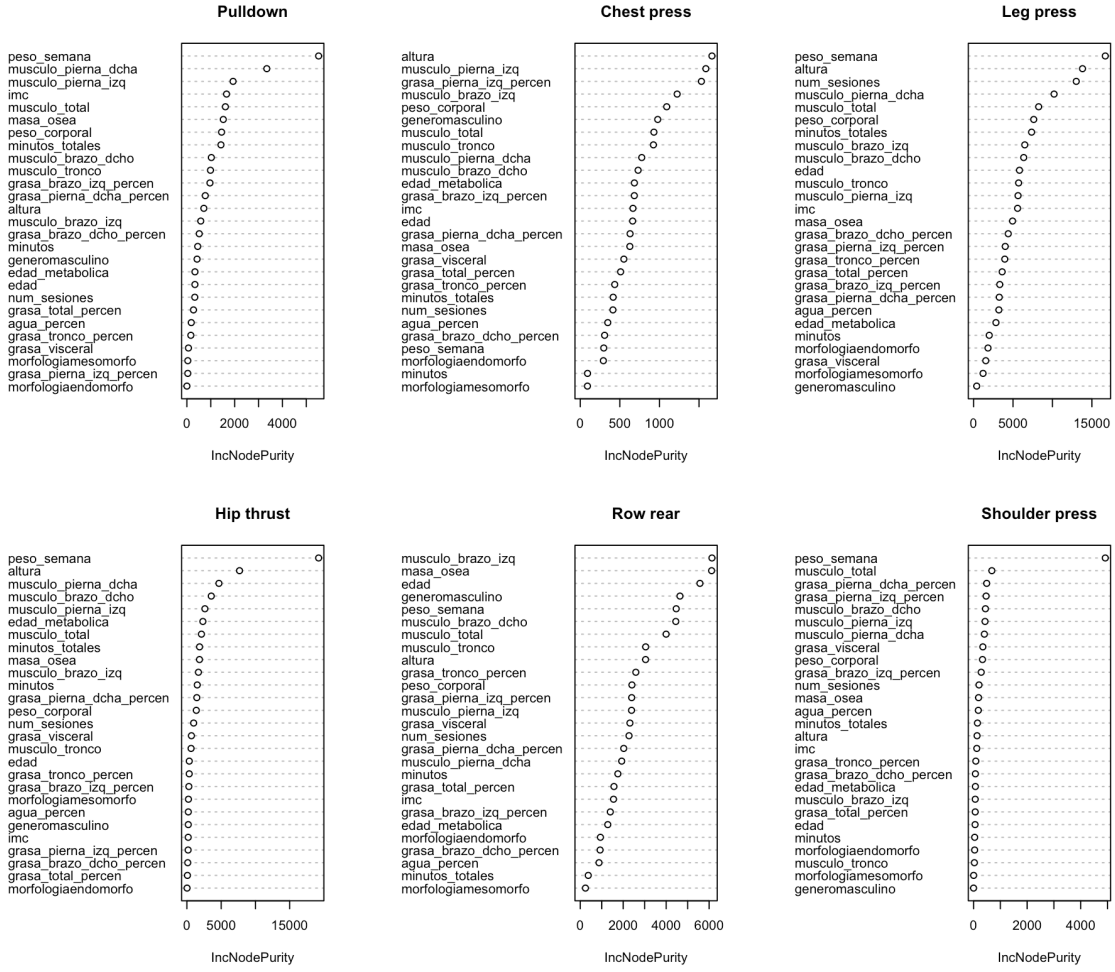

Figura 5.10: Variables utilizadas en el Bosque aleatorio ordenadas por su importancia. Se muestra un gráfico para cada uno de los modelos ajustados según el tipo de ejercicio. *mtry*: Número de variables seleccionadas al azar como candidatos en cada división. *ntree*: Número de árboles. *nodes*: Número máximo de nodos terminales permitidos en los árboles del bosque aleatorio.

En segundo lugar, a través del modelo GAM, hicimos inicialmente una selección de variables utilizando el método propuesto por Marra y Wood (2011) y la función gam de la librería mgcv de Wood (2017) (mediante el argumento `select = TRUE`). En este trabajo, se aborda el desafío de seleccionar variables relevantes en el contexto de los modelos aditivos generalizados (GAM). Su método se basa en un marco de penalización que integra la selección de suavidad y la selección de variables simultáneamente. Al emplear un enfoque de doble penalización, permiten la reducción de los términos suaves a cero, excluyendo efectivamente las variables no influyentes del modelo. Este enfoque aprovecha tanto las propiedades de inducción de esparsidad de la penalización L1 como las restricciones de suavidad de las penalizaciones tradicionales de *splines*, resultando en un procedimiento robusto y eficiente para identificar predictores significativos mientras se mantiene la flexibilidad del modelo aditivo. Teniendo en cuenta esto, a continuación, se ajusta un nuevo modelo que únicamente incluye las variables con efecto (efecto parcial distinto de cero). Por último, se comprueba el ajuste de este modelo, eliminando aquellas variables sin efecto, dando como resultado un modelo reducido final. En algunos casos, el descarte de variables no significativas condujo a un modelo con mayor error en la validación cruzada. En estos casos, dichas variables se mantuvieron como predictoras reflejando nuestro foco por el rendimiento y precisión en las estimaciones, en detrimento de la interpretabilidad de los modelos y sus variables. Los listados 5.1, 5.2, 5.3, 5.4, 5.5 y 5.6 muestran los resúmenes de los distintos modelos GAM ajustados para cada tipo de ejercicio.

La métricas de los modelos se puede observar en la Tabla 5.2, que indica el error obtenido en la validación cruzada. La Tabla 5.3 reporta algunos ejemplos de las predicciones que se obtienen con ambos modelos. En el siguiente sección, se discutirá en mayor detalle los hallazgos de esta sección de resultados.

Listado 5.1: Resumen del modelo GAM (*pulldown*).

```
Family: gaussian
Link function: identity

Formula:
rm ~ -1 + s(peso_semana) + (grasa_visceral) + (musculo_total) +
    s(edad) + (minutos) + morfologia

Parametric coefficients:
                    Estimate Std. Error t value Pr(>|t|)
grasa_visceral       -1.0352     1.0991  -0.942  0.36004
musculo_total         1.5863     0.4912   3.230  0.00514 **
minutos               6.4783     3.1981   2.026  0.05950 .
morfologiaectomorfo  -5.6734    20.8225  -0.272  0.78869
morfologiaendomorfo  12.2620    18.6576   0.657  0.52022
morfologiamesomorfo   5.0318    19.1232   0.263  0.79575
---
Signif. codes:  0 '*** 0.001 '** 0.01 '* 0.05 '. 0.1 ' ' 1

Approximate significance of smooth terms:
                edf Ref.df      F p-value
s(peso_semana) 5.557  6.490 22.058  <2e-16 ***
s(edad)        2.150  2.593  1.608   0.155
---
Signif. codes:  0 '*** 0.001 '** 0.01 '* 0.05 '. 0.1 ' ' 1

R-sq.(adj) =  0.966   Deviance explained = 98.1%
-REML = 92.403  Scale est. = 40.25      n = 30
```

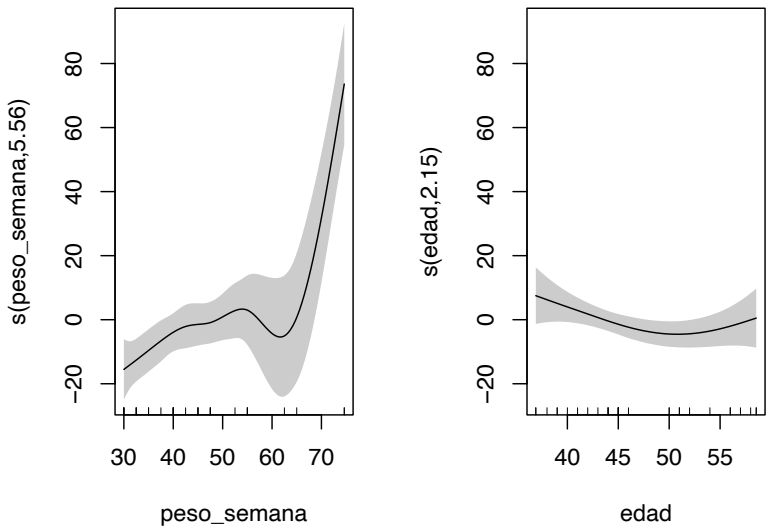

Figura 5.11: Efectos parciales suaves ajustados utilizando el modelo GAM para el ejercicio *pulldown*.

Listado 5.2: Resumen del modelo GAM (*chest press*).

```
Family: gaussian
Link function: identity

Formula:
rm ~ s(num_sesiones) + s(edad_metabolica) + (agua_percen) + (grasa_visceral) +
    (grasa_brazo_izq_percen) + s(musculo_pierna_dcha) + +(minutos) +
    s(minutos_totales) + genero

Parametric coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)              -186.1233    58.9106  -3.159  0.01404 *
agua_percen                 2.9981     0.7717   3.885  0.00497 **
grasa_visceral             -3.1941     1.4696  -2.173  0.06271 .
grasa_brazo_izq_percen      1.5873     0.4758   3.336  0.01084 *
minutos                    24.3922     8.7625   2.784  0.02463 *
generomasculino            33.9402    15.0489   2.255  0.05529 .
---
Signif. codes:  0 '*** 0.001 '** 0.01 '* 0.05 '. 0.1 ' ' 1

Approximate significance of smooth terms:
                          edf Ref.df      F  p-value
s(num_sesiones)         5.068  5.765 13.850 0.001104 **
s(edad_metabolica)      5.019  5.855  8.931 0.003930 **
s(musculo_pierna_dcha) 3.077  3.559  5.761 0.024152 *
s(minutos_totales)      3.116  3.564 27.089 0.000107 ***
---
Signif. codes:  0 '*** 0.001 '** 0.01 '* 0.05 '. 0.1 ' ' 1

R-sq.(adj) =  0.982   Deviance explained = 99.5%
-REML = 86.999  Scale est. = 13.473     n = 30
```

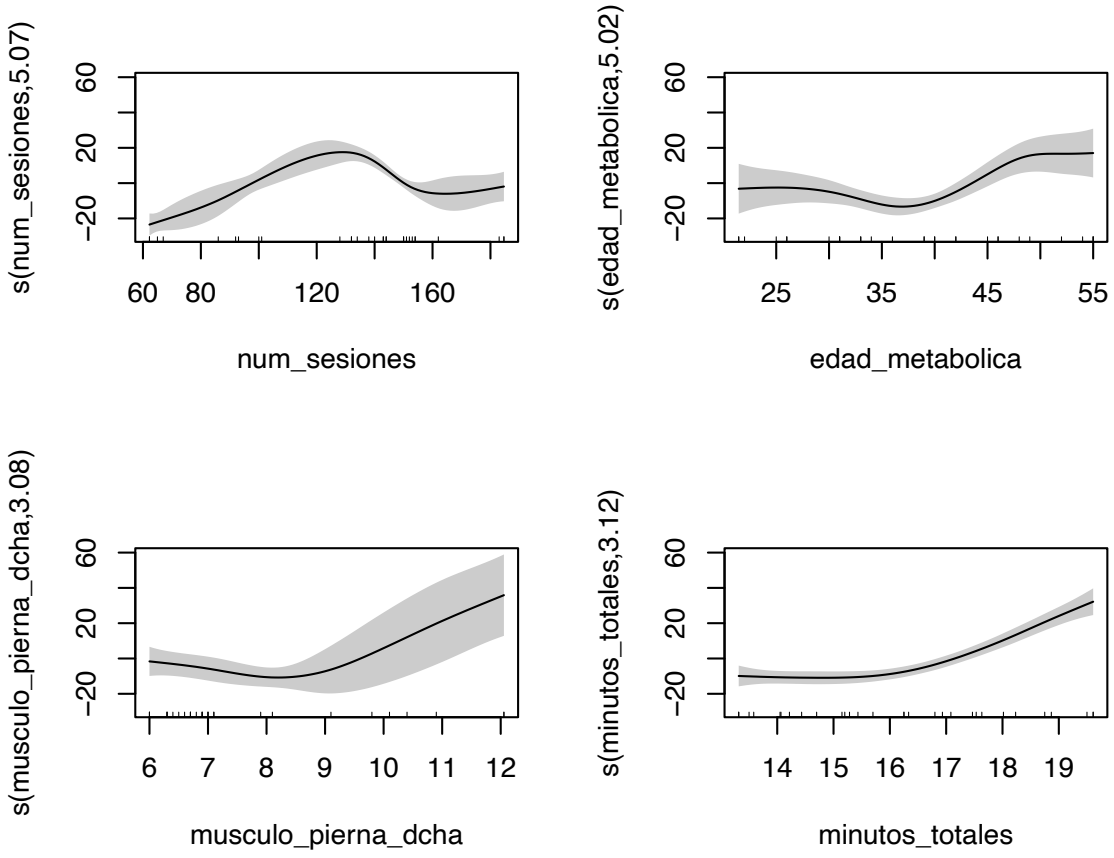

Figura 5.12: Efectos parciales suaves ajustados utilizando el modelo GAM para el ejercicio *chest press*.

Listado 5.3: Resumen del modelo GAM (*leg press*).

```
Family: gaussian
Link function: identity

Formula:
rm ~ -1 + peso_semana + s(num_sesiones) + s(minutos) + s(minutos_totales) +
    s(peso_corporal)

Parametric coefficients:
             Estimate Std. Error t value Pr(>|t|)
peso_semana   1.70261    0.05251   32.42 8.72e-16 ***
---
Signif. codes:  0 '*** 0.001 '** 0.01 '* 0.05 '. 0.1 ' ' 1

Approximate significance of smooth terms:
                     edf Ref.df     F p-value
s(num_sesiones)    3.997  4.637 1.279  0.2151
s(minutos)         2.853  3.431 3.041  0.0488 *
s(minutos_totales) 3.806  4.531 2.222  0.1378
s(peso_corporal)   1.678  1.964 0.538  0.5686
---
Signif. codes:  0 '*** 0.001 '** 0.01 '* 0.05 '. 0.1 ' ' 1

R-sq.(adj) =  0.804   Deviance explained =   89%
-REML = 142.85  Scale est. = 1229.8     n = 29
```

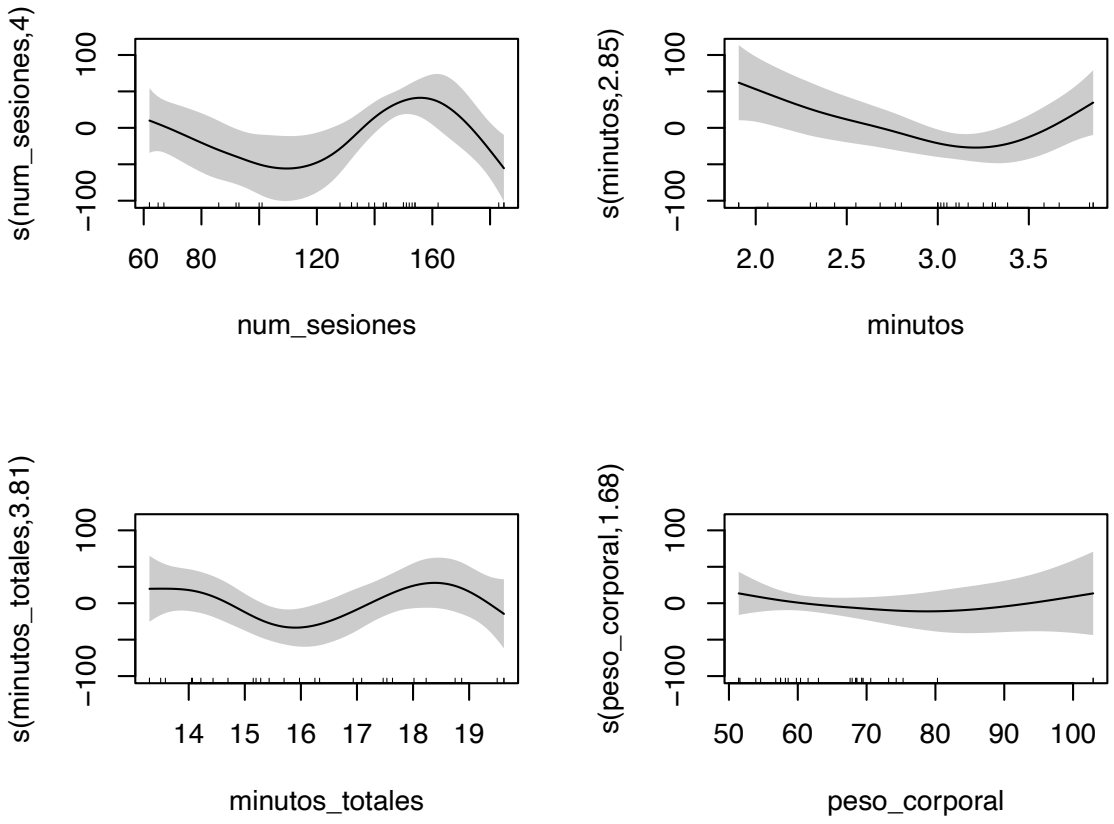

Figura 5.13: Efectos parciales suaves ajustados utilizando el modelo GAM para el ejercicio *leg press*.

Listado 5.4: Resumen del modelo GAM (*hip thrust*).

```
Family: gaussian
Link function: identity

Formula:
rm ~ s(peso_semana) + (num_sesiones) + s(altura) + (musculo_brazo_izq) +
    s(musculo_pierna_dcha) + genero

Parametric coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       128.4600    41.2391   3.115  0.00735 **
num_sesiones        0.3349     0.1046   3.203  0.00615 **
musculo_brazo_izq -38.9865    18.5576  -2.101  0.05362 .
generomasculino   -61.1602    26.6534  -2.295  0.03717 *
---
Signif. codes:  0 '*** 0.001 '** 0.01 '* 0.05 '. 0.1 ' ' 1

Approximate significance of smooth terms:
                        edf Ref.df      F  p-value
s(peso_semana)        4.374  5.233 14.588 4.26e-05 ***
s(altura)             2.550  3.095  1.987 0.161065
s(musculo_pierna_dcha) 3.600  4.285 11.436 0.000211 ***
---
Signif. codes:  0 '*** 0.001 '** 0.01 '* 0.05 '. 0.1 ' ' 1

R-sq.(adj) =  0.954   Deviance explained = 97.6%
-REML = 101.07  Scale est. = 95.704     n = 29
```

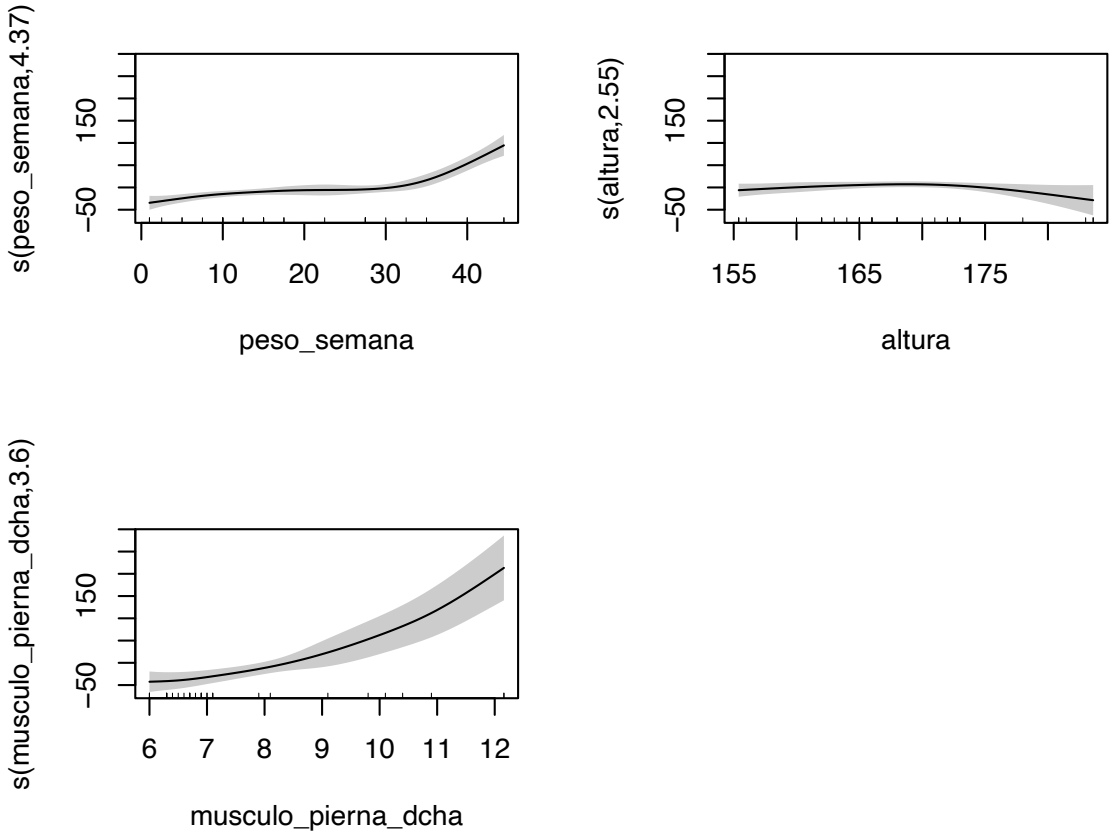

Figura 5.14: Efectos parciales suaves ajustados utilizando el modelo GAM para el ejercicio *hip thrust*.

Listado 5.5: Resumen del modelo GAM (*row rear*).

```
Family: gaussian
Link function: identity

Formula:
rm ~ s(altura) + (grasa_total_percen) + s(grasa_brazo_izq_percen) +
    s(grasa_tronco_percen) + s(minutos) + s(minutos_totales)

Parametric coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         565.544     77.748   7.274 5.35e-05 ***
grasa_total_percen  -17.039      2.775  -6.141 0.000189 ***
---
Signif. codes:  0 '*** 0.001 '** 0.01 '* 0.05 '. 0.1 ' ' 1

Approximate significance of smooth terms:
                           edf Ref.df     F  p-value
s(altura)                3.999  4.680 2.652 0.176276
s(grasa_brazo_izq_percen) 2.921  3.514 1.936 0.209628
s(grasa_tronco_percen)   3.983  4.619 5.260 0.010278 *
s(minutos)               6.197  7.003 14.818 0.000278 ***
s(minutos_totales)       1.122  1.201 11.825 0.006784 **
---
Signif. codes:  0 '*** 0.001 '** 0.01 '* 0.05 '. 0.1 ' ' 1

R-sq.(adj) =  0.948   Deviance explained = 98.4%
-REML = 117.33  Scale est. = 161.55     n = 29
```

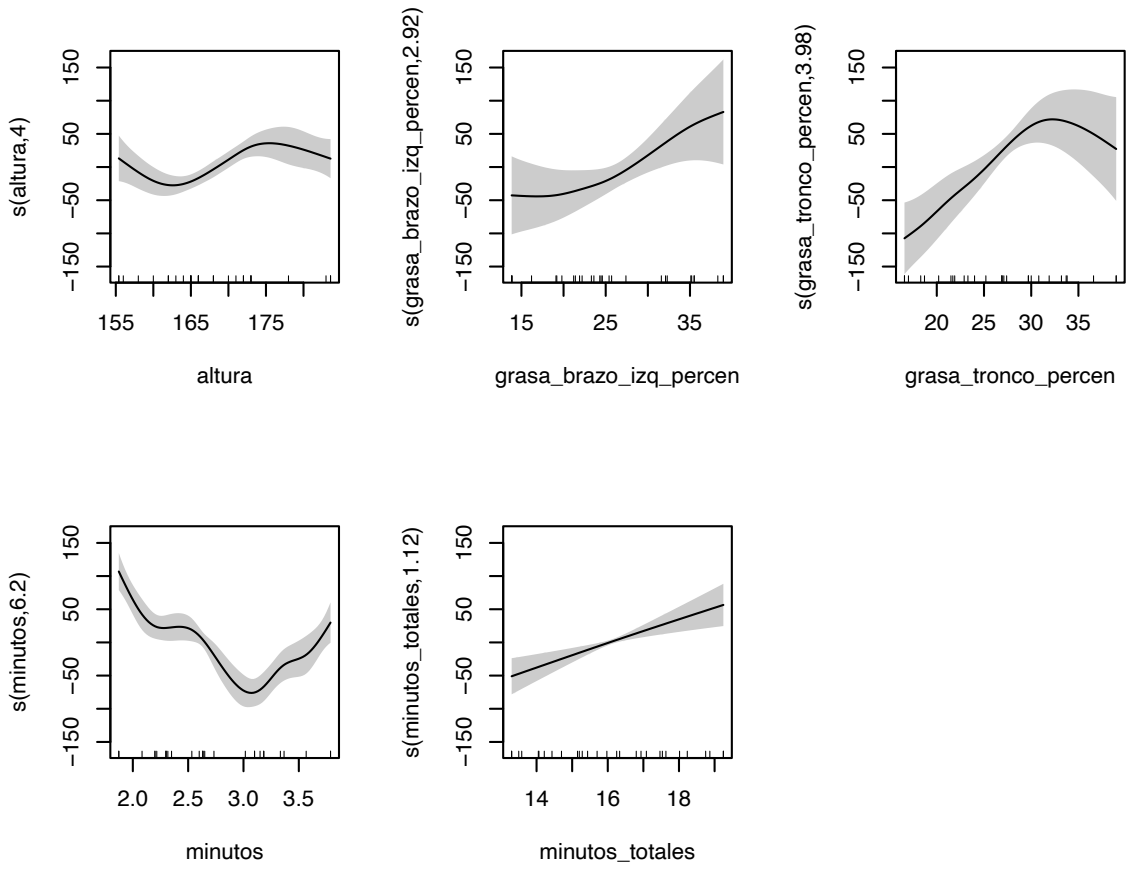

Figura 5.15: Efectos parciales suaves ajustados utilizando el modelo GAM para el ejercicio *row rear*.

Listado 5.6: Resumen del modelo GAM (*shoulder press*).

```
Family: gaussian
Link function: identity

Formula:
rm ~ s(peso_semana) + (num_sesiones) + (masa_osea) + s(musculo_pierna_izq)

Parametric coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   91.66602   34.53472   2.654   0.0149 *
num_sesiones   0.08187    0.04005   2.044   0.0537 .
masa_osea    -24.40150   13.49163  -1.809   0.0849 .
---
Signif. codes:  0 '*** 0.001 '** 0.01 '* 0.05 '. 0.1 ' ' 1

Approximate significance of smooth terms:
                         edf Ref.df     F  p-value
s(peso_semana)         4.449  5.311 9.054 0.000105 ***
s(musculo_pierna_izq)  1.669  2.081 3.412 0.049383 *
---
Signif. codes:  0 '*** 0.001 '** 0.01 '* 0.05 '. 0.1 ' ' 1

R-sq.(adj) =  0.923   Deviance explained = 94.4%
-REML = 91.333  Scale est. = 29.179     n = 30
```

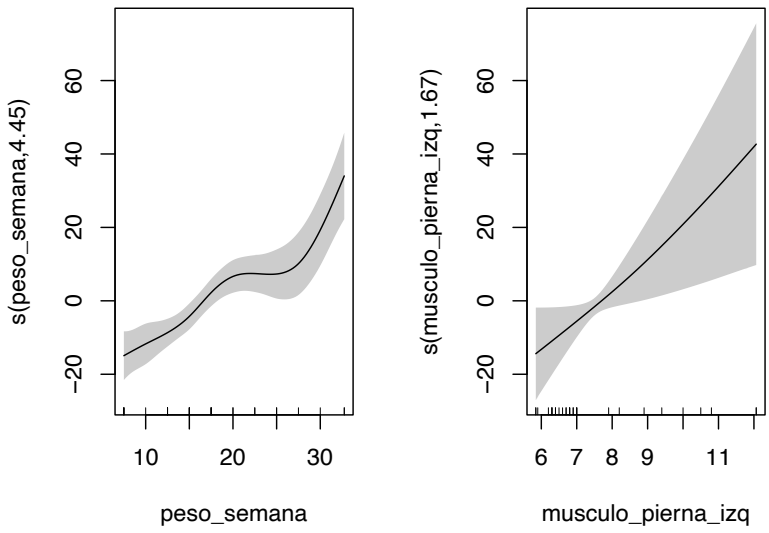

Figura 5.16: Efectos parciales suaves ajustados utilizando el modelo GAM para el ejercicio *shoulder press*.

	Bosque aleatorio				GAM			
Ejercicio	MSE	RMSE	MAE	R^2	MSE	RMSE	MAE	R^2
Pulldown	497.5	19.0	15.2	0.92	21.8	4.6	3.6	0.98
Chest Press	376.3	16.0	13.9	0.76	5.1	2.2	1.7	0.99
Leg Press	5552.1	66.8	57.1	0.77	664.4	25.7	19.8	0.89
Hip Thrust	1074.7	26.5	21.4	0.87	47.7	6.9	5.4	0.97
Row Rear	2936.0	46.5	38.3	0.69	48.8	6.9	6.0	0.98
Shoulder Press	154.6	10.0	8.4	0.86	20.3	4.5	3.6	0.94

Tabla 5.2: Métricas de error obtenidas por CV para los distintos modelos considerando el tipo de ejercicio realizado. Bosque aleatorio: 10-fold CV. GAM: LOO CV.

	Pulldown			Chest press			Leg press		
Usuario	RM	$\widehat{RM}_{ba}$	$\widehat{RM}_{gam}$	RM	$\widehat{RM}_{ba}$	$\widehat{RM}_{gam}$	RM	$\widehat{RM}_{ba}$	$\widehat{RM}_{gam}$
1	90.0	85.0	90.4	77.7	68.3	74.3	226.9	223.6	250.7
2	81.0	78.8	78.5	71.2	58.7	68.3	311.7	172.2	276.9
3	184.1	125.7	183.8	117.7	82.9	117.7	171.9	223.1	223.3
4	53.2	66.7	51.2	31.0	50.5	30.7	116.9	161.4	126.3
	Hip thrust			Row rear			Shoulder press		
Usuario	RM	$\widehat{RM}_{ba}$	$\widehat{RM}_{gam}$	RM	$\widehat{RM}_{ba}$	$\widehat{RM}_{gam}$	RM	$\widehat{RM}_{ba}$	$\widehat{RM}_{gam}$
1	46.6	42.1	44.4	72.9	88.2	77.4	49.1	41.1	43.0
2	81.8	67.3	78.4	69.5	70.6	75.4	56.2	44.1	54.3
3	85.9	129.4	84.7	16.3	72.6	2.4	83.5	68.9	91.1
4	49.1	40.8	59.0	61.3	56.5	67.5	12.2	27.4	13.7

Tabla 5.3: Predicciones de RM utilizando el Bosque aleatorio ($\widehat{RM}_{ba}$) y el modelo GAM ($\widehat{RM}_{gam}$) para los primeros cuatro usuarios en diferentes ejercicios.

5.3. Discusión

En esta sección de discusión vamos a analizar detenidamente los hallazgos de la sección de resultados. En este análisis se han considerado dos cuestiones importantes. Por un lado, la selección de variables, su significación e interpretación. Por otro, el rendimiento y precisión de los modelos ajustados, así como su adecuación al objetivo propuesto.

Teniendo en cuenta que la principal métrica de comparación en este trabajo es el coeficiente de determinación, en la Tabla 5.2 se puede observar que el modelo GAM obtiene mejores resultados en todos los tipos de ejercicio. Por lo tanto, este será el modelo seleccionado y foco de nuestra sección de discusión.

Analizando el estudio de variables por tipo de ejercicio, podemos ver que en el ejercicio *pulldown*, el peso semanal tiene un efecto muy significativo y suave sobre el RM, mientras que el contraste sobre el coeficiente asociado a la variable músculo total y a la variable minutos (TUL) resulta significativo, dando resultado un efecto lineal de estas variables sobre la respuesta (Figura 5.11 y listado 5.1). En general, a medida que se incrementa el peso levantado en los entrenamientos semanales se incrementa la repetición máxima alcanzada en la prueba. El músculo total y los minutos también están positivamente asociados con la variable objetivo, donde un incremento de una unidad en estas variables supone un incremento en media de 1.58 y 6.47 kg, respectivamente, en el RM. Ajustando mediante el bosque aleatorio (Figura 5.10), las variables con mayor importancia en este movimiento son el peso semanal y la musculatura de las piernas (izquierda y derecha). En este sentido, obtenemos una selección similar al GAM, donde el peso semanal es la variable más importante y el músculo total también sale bien posicionado, aunque se da preferencia a la musculatura de las piernas. Los minutos no salen como una variable muy importante en este caso.

Pasando el foco al movimiento *chest press*, el modelo GAM reporta un efecto claro en la respuesta (RM) de las variables minutos totales (tiempo total de entrenamiento), número de sesiones, edad metabólica y músculo de la pierna derecha, todos ellos con un efecto suave sobre el RM (listado 5.2). En general, un incremento del músculo de la pierna derecha o de los minutos totales se asocia con un incremento del RM (Figura 5.12). El número de sesiones da lugar a un incremento de la variable respuesta inicialmente, pero se observa un descenso a partir de las 130 sesiones. Quizás una explicación a este evento pueda ser que existe un punto de inflexión, donde el rendimiento del usuario empieza a decrecer por desmotivación, cansancio de la actividad u otros factores.

Por otra parte, existe un efecto significativo del porcentaje de agua, los minutos (TUL), la grasa del brazo izquierdo, la grasa visceral y el género. Todos ellos con una relación lineal sobre el RM. Se destaca la grasa visceral con un efecto negativo, donde un incremento en esta variable supone un descenso de 3.1 kg en media en la variable objetivo.

Observando la importancia de variables del bosque aleatorio (Figura 5.10), podemos ver que las variables más relevantes son: altura, músculo del brazo y pierna izquierda, grasa de la pierna izquierda, peso corporal, género masculino, músculo total y del tronco. Existe algún solapamiento con el GAM relativamente a musculatura por miembro, grasa y género, no obstante, la altura y el peso corporal solamente salen como importantes en el bosque aleatorio.

En el ejercicio *leg press*, las únicas variables con un efecto significativo son el peso semanal, con un efecto lineal positivo de 1.7 kg en media por cada kg de carga semanal, y los minutos con un efecto suave (listado 5.3) que es positivo sobre el RM a partir de los 3 minutos (TUL) (Figura 5.13).

El bosque aleatorio reporta un mayor número de variables como la altura, número de sesiones, musculatura total y por miembro, aunque el peso semanal es también la variable más importante.

Si nos centramos ahora en el ejercicio *hip thrust*, el peso semanal y el músculo de la pierna derecha tienen un efecto claro y no paramétrico en la respuesta, donde incrementos en estas variables conducen a incrementos en el RM (Figura 5.14 y listado 5.4). Por otra parte, el contraste asociado al número de sesiones es significativo, donde un incremento unitario en el número de sesiones conlleva un incremento del RM en 0.33 kg. Destacamos también el efecto de la variable género, dando como resultado mejores marcas (RM) en el caso de las mujeres que lo hombres.

Analizando la importancia de variables del bosque aleatorio (Figura 5.10), el peso semanal sale

como la variable más relevante, destacándose también la altura, aunque en menor medida.

En el ejercicio *row rear*, los minutos (TUL), los minutos totales (tiempo total de entrenamiento) y la grasa del tronco tienen un efecto no paramétrico significativo (listado 5.5). Al igual que en el ejercicio *leg press*, se establece un punto de inflexión en movimientos con una duración de 3 minutos (TUL), donde el RM decrece para valores inferiores y crece para los superiores (Figura 5.15). Realizar un movimiento demasiado rápido puede significar que el peso o las repeticiones del ejercicio no son suficientemente exigentes, conduciendo así a peores marcas de RM. Por otra parte, la duración total del entrenamiento está positivamente asociada con la variable respuesta en todo su rango, con un efecto casi lineal. También se observa que la grasa del tronco tiene una forma cóncava, sugiriendo que los valores en las colas (delgadez o obesidad) están asociados con una menor repetición máxima. Por otra parte, el contraste asociado a la variable grasa total es significativo con un efecto lineal de -17 kg por unidad de grasa.

El bosque aleatorio reporta un conjunto de variables distinto, dando mayor importancia a la musculatura total y por miembro, la masa ósea, edad, género masculino y peso semanal (Figura 5.10).

Por último, en el ejercicio *shoulder press*, destacar que el peso semanal tiene un efecto no paramétrico significativo, mientras que el músculo de la pierna izquierda tiene un efecto prácticamente lineal (listado 5.6). Ambos tienen un efecto positivo sobre la variable objetivo (Figura 5.16). Por otra parte, el contraste de significación asociado al número de sesiones es ligeramente significativo con un efecto lineal y positivo sobre el RM.

Al igual que en el ejercicio *hip thrust*, el bosque aleatorio considera exclusivamente el peso semanal como predictor (Figura 5.10).

La segunda parte de nuestro análisis, se ciñe a la capacidad predictiva de los modelos ajustados (Tabla 5.2). Se observa que ambos modelos tienen elevados coeficientes de determinación, en especial el modelo GAM. En ningún caso se observa que el bosque aleatorio supere al modelo GAM, tanto en términos de coeficiente de determinación como de las demás métricas de error, por lo que se recomienda el uso del ajuste GAM para la predicción del RM. Esta mayor precisión se ve reflejada en la Tabla 5.3, la cuál contiene algunos ejemplos de predicción por modelo y tipo de ejercicio. Aunque el bosque aleatorio tiene métodos para evitar el sobreajuste, como el número o profundidad de los árboles, en nuestra opinión, el modelo GAM puede tener una mayor capacidad de evitar el sobreajuste. El GAM permite el control del suavizado con técnicas de regresión penalizada, evitando así complejidad en el modelado al mismo tiempo que se ajusta de forma suave a los datos, dotándose así de una mayor capacidad de generalización. Quizás con un mayor tamaño muestral es posible observar un mayor rendimiento en el bosque aleatorio, ya que este tipo de modelo se beneficia por una gran cantidad de datos.

Capítulo 6

Conclusiones

El desarrollo de esta aplicación web interactiva para el análisis de datos de entrenamientos de fuerza de alta intensidad ha sido todo un reto al que se le ha dedicado un gran un esfuerzo integral. A través de este proyecto, se han logrado los dos objetivos planteados por la empresa.

Por un lado, la aplicación permite a los usuarios ingresar, procesar y visualizar sus datos de entrenamiento de manera eficiente. Esto facilita la comprensión y monitorización del progreso del entrenamiento, resultando más fácil para los usuarios tomar decisiones informadas sobre sus rutinas de ejercicio. El diseño de la interfaz de usuario, enfocado en la compatibilidad móvil, asegura la accesibilidad y facilidad de uso de la aplicación, mientras que la implementación de medidas de seguridad, como la autenticación de usuarios y el almacenamiento de datos cifrados, asegura que los datos de los usuarios estén protegidos. Esto es crucial para mantener la confianza del usuario así como cumplir con la ley de protección de datos. Hospedar la aplicación en shinyapps.io también ha permitido tener una solución escalable, donde es posible ajustarse a una tarifa que varía en función del número de conexiones mensuales.

Al incorporar modelos de aprendizaje automático como el bosque aleatorio y los modelos aditivos generalizados (GAM), la aplicación es capaz de estimar con precisión la fuerza o repetición máxima (RM) de un posible usuario para cada uno de los ejercicios. Esta característica elimina la necesidad de estimar el RM mediante pruebas directas, que suponen un riesgo de lesión, así como perder una sesión de entrenamiento para estimarla por el método indirecto. De esta forma, se garantiza la seguridad del usuario y la eficiencia en tiempo mientras se mantiene la precisión en la evaluación de la fuerza.

El estudio empírico realizado confirma que existe un número de variables significativas para predecir el RM. Considerando el mejor modelo (modelo GAM), el peso semanal, los minutos (TUL) y los minutos totales (tiempo total), suelen ser las variables con mayor efecto en la mayoría de ejercicios. En menor medida, algunas variables de composición corporal como la musculatura y grasa total o por miembro, también muestran tener una importante capacidad predictiva. Adicionalmente, en este estudio se ha obtenido un buen rendimiento tanto con el bosque aleatorio como con el GAM, aunque el segundo reporta una precisión considerablemente mayor al obtener coeficientes de determinación superiores al 0.9 en su mayoría. Por lo tanto, el modelo GAM será el que se despliegue en la aplicación para predecir el RM de los usuarios. Los resultados de este trabajo nos llevan a considerar que el GAM tiene una mayor habilidad de generalización que el bosque aleatorio, debido a su capacidad para ajustarse de manera no paramétrica, controlando el suavizado y el sobreajuste con técnicas de regresión penalizadas.

Por último, cabe mencionar algunas limitaciones del estudio. Por un lado, consideramos limitante el tamaño muestral de treinta individuos. A medida que crezca el número de usuarios de la empresa (o que los registros de los usuarios actuales se actualicen de manera digital), se hace imperativo revisar los modelos propuestos en este trabajo. Por otra parte, la recopilación de los datos, por parte de la empresa, supone la anotación manual de las métricas de trabajo y composición corporal en distintas fechas. Esto significa que para cada usuario existen varias observaciones. Debido a que la empresa está en proceso de trasladar esta información a formato digital, no hemos tenido acceso a todas las observaciones por

usuario. Esto nos ha llevado a tomar la decisión de considerar, para cada usuario, únicamente el RM registrado en la última prueba y las observaciones registradas en el registro (entrenamiento) semanal anterior a dicha prueba. De esta manera, se ha trabajado con un registro por usuario manteniendo el supuesto de independencia entre observaciones. Quedaría pendiente, como trabajo futuro y cuando culmine el proceso de digitalización de la empresa, la exploración de un ajuste mediante modelos de regresión mixtos (Duda, Hart et al., 1973; Lindsay, 1995) que permitan el uso de varios registros por usuario.

Apéndice A

Código de R

El código R del trabajo está compuesto por seis ficheros independientes para mayor organización. Existen cuatro ficheros que son la base de la aplicación: el fichero **ui.R** define la interfaz gráfica de la aplicación, que a su vez es apoyado por el fichero **style.CSS**; **server.R** define la lógica del servidor (*backend*); **global.R** carga las librerías necesarias y algunos ajustes globales como la conexión con SQLite. Por otra parte, el fichero **statistics.R** recompila todos los datos de usuario y calcula la media de las variables objetivo. Se trata de un fichero que sirve de soporte a la aplicación, ya que calcula las métricas usadas en la pestaña de Análisis Comparativo y recompila los datos para el modelado estadístico (**models.R**).

Este código no ha sido incluido en el presente documento, sin embrago podría ser proporcionado bajo solicitud previa.

Bibliografía

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection.

Auguie, B. (2017). *gridExtra: Miscellaneous Functions for "Grid"Graphics* [R package version 2.3]. https://CRAN.R-project.org/package=gridExtra

Beeley, C., & Sukhdeve, S. (2018). *Web Application Development with R using Shiny* (Vol. 2).

Björck, Å. (1996). *Numerical Methods for Least Squares Problems*. Society for Industrial; Applied Mathematics. https://doi.org/10.1137/1.9781611971484

Breiman, L., Friedman, J., Stone, C., & Olshen, R. (1984). *Classification and Regression Trees*. Taylor & Francis. https://books.google.es/books?id=JwQx-WOmSyQC

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.

Bryan, J. (2023). *googlesheets4: Access Google Sheets using the Sheets API V4* [R package version 1.1.1]. https://CRAN.R-project.org/package=googlesheets4

Brzycki, M. (1993). Strength testing—predicting a one-rep max from reps-to-fatigue. *Journal of physical education, recreation & dance*, *64*(1), 88-90.

Chai, T., & Draxler, R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, *7*. https://doi.org/10.5194/gmd-7-1247-2014

Chang, W., & Cheng, J. (2023). *later: Utilities for Scheduling Functions to Execute Later with Event Loops* [R package version 1.3.2]. https://CRAN.R-project.org/package=later

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., & Borges, B. (2023). *shiny: Web Application Framework for R* [R package version 1.7.5]. https://CRAN.R-project.org/package=shiny

Chang, W., & Ribeiro, B. (2021). *shinydashboard: Create Dashboards with 'Shiny'* [R package version 0.7.2]. https://CRAN.R-project.org/package=shinydashboard

Duda, R., Hart, P., et al. (1973). *Pattern classification and scene analysis* (Vol. 3). Wiley New York.

Enders, C. (2022). *Applied missing data analysis*. Guilford Publications.

Epley, B. (1985). Poundage chart. *Boyd Epley Workout. Lincoln, NE: Body Enterprises*, *86*.

González-Badillo, J., & Gorostiaga-Ayestarán, E. (1997). Foundations of strength training. *Application to Sport Performance; Inde: Barcelona, Spain*.

Hastie, T., & Tibshirani, R. (1986). Generalized additive models, cubic splines and penalized likelihood. *Submitted for publication*.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2). Springer.

Hastings, C., Mosteller, F., Tukey, J., & Winsor, C. (1947). Low moments for small samples: a comparative study of order statistics. *The Annals of Mathematical Statistics*, *18*(3), 413-426.

Hipp, R. (2020). SQLite. https://www.sqlite.org/index.html

Hutchins, K. (1992). *Super Slow: The Ultimate Exercise Protocol*. Ken Hutchins. https://books.google.es/books?id=5njapwAACAAJ

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*. https://doi.org/10.1016/j.ijforecast.2006.03.001

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning: with Applications in R*. https://doi.org/10.1007/978-1-0716-1418-1

Jennrich, R. I., & Sampson, P. (1976). Newton-Raphson and related algorithms for maximum likelihood variance component estimation. *Technometrics*, *18*(1), 11-17.

Jones, A. (1970). *Nautilus training principles*. Arthur Jones Productions.

Kuhn & Max. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, *28*(5), 1-26. https://doi.org/10.18637/jss.v028.i05

LeSuer, D., McCormick, J., Mayhew, J., Wasserstein, R., & Arnold, M. (1997). The accuracy of prediction equations for estimating 1-RM performance in the bench press, squat, and deadlift. *Journal of Strength and Conditioning Research*, *11*. https://doi.org/10.1519/00124278-199711000-00001

Liaw, A., & Wiener, M. (2002). Classification and Regression with Random Forest. *R News*, *2*.

Lindsay, B. (1995). Mixture models: theory, geometry, and applications.

Little, J., & McGuff, D. (2009). *Body by Science: A Research Based Program to Get the Results You Want in 12 Minutes a Week*. McGraw Hill Professional.

Lombardi, P. (1989). Beginning weight training: the safe and effective way. *(No Title)*.

Marra, G., & Wood, S. N. (2011). Practical variable selection for generalized additive models. *Computational Statistics & Data Analysis*, *55*(7), 2372-2387.

McGowan, L., & Bryan, J. (2023). *googledrive: An Interface to Google Drive* [R package version 2.1.1]. https://CRAN.R-project.org/package=googledrive

Montgomery, D., Peck, E., & Vining, G. (2021). *Introduction to linear regression analysis*. John Wiley & Sons.

Morita, K. (2021). Introduction to Multiple Imputation. *Annals of Clinical Epidemiology*, *3*. https://doi.org/10.37737/ace.3.1_1

Müller, K. (2020). *here: A Simpler Way to Find Your Files* [R package version 1.0.1]. https://CRAN.R-project.org/package=here

Müller, K. (2023). *hms: Pretty Time of Day* [R package version 1.1.3]. https://CRAN.R-project.org/package=hms

Nelder, A., & Wedderburn, R. (1972). Generalized linear models. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *135*(3), 370-384.

Niewiadomski, W., Gąsiorowska, A., Cybulski, G., Laskowska, D., & Langfort, J. (2008). Determination and Prediction of One Repetition Maximum (1RM): Safety Considerations. *Journal of Human Kinetics*, *19*. https://doi.org/10.2478/v10078-008-0008-8

O'Connor, R., & Simmons, J. (1989). *Weight training today*. Brooks Cole.

Perrier, V., Meyer, F., & Granjon, D. (2023). *shinyWidgets: Custom Inputs Widgets for Shiny* [R package version 0.8.0]. https://CRAN.R-project.org/package=shinyWidgets

Peterson, B., & Carl, P. (2020). *PerformanceAnalytics: Econometric Tools for Performance and Risk Analysis* [R package version 2.0.4]. https://CRAN.R-project.org/package=PerformanceAnalytics

Quinlan, R. (1986). Induction of decision trees. *Machine learning*, *1*, 81-106.

R Core Team. (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

R Special Interest Group on Databases (R-SIG-DB), Wickham, H., & Müller, K. (2022). *DBI: R Database Interface* [R package version 1.1.3]. https://CRAN.R-project.org/package=DBI

Rubin, D. (1987). The calculation of posterior distributions by data augmentation: Comment: A non-iterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: The SIR algorithm. *Journal of the American Statistical Association*, *82*(398), 543-546.

Sayers, M., Schlaeppi, M., Hitz, M., & Lorenzetti, S. (2018). The impact of test loads on the accuracy of 1RM prediction using the load-velocity relationship. *BMC Sports Science, Medicine and Rehabilitation*, *10*. https://doi.org/10.1186/s13102-018-0099-z

Schoenfeld, B. (2010). The mechanisms of muscle hypertrophy and their application to resistance training. https://doi.org/10.1519/JSC.0b013e3181e840f3

Sheldon, W., Stevens, S., & Tucker, W. (1940). The varieties of human physique.

Sievert, C. (2020). *Interactive Web-Based Data Visualization with R, plotly, and shiny*. Chapman; Hall/CRC. https://doi.org/10.1201/9780429447273

Signorell, A. (2024). *DescTools: Tools for Descriptive Statistics* [R package version 0.99.54]. https://CRAN.R-project.org/package=DescTools

Stanelle, S., Crouse, S., Heimdal, T., Riechman, S., Remy, A., & Lambert, B. (2021). Predicting muscular strength using demographics, skeletal dimensions, and body composition measures. *Sports Medicine and Health Science*, *3*. https://doi.org/10.1016/j.smhs.2021.02.001

Thieurmel, B., & Perrier, V. (2022). *shinymanager: Authentication Management for 'Shiny' Applications* [R package version 1.0.410]. https://CRAN.R-project.org/package=shinymanager

Tukey, J. (1962). The future of data analysis. En *Breakthroughs in Statistics: Methodology and Distribution* (pp. 408-452). Springer.

Wand, M. P., & Jones, M. C. (1994). *Kernel smoothing*. CRC press.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Willmott, C., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, *30*. https://doi.org/10.3354/cr030079

Wood, S. (2017). *Generalized Additive Models: An Introduction with R* (2.ª ed.). Chapman; Hall/CRC.

Xie, Y., Cheng, J., & Tan, X. (2023). *DT: A Wrapper of the JavaScript Library 'DataTables'* [R package version 0.29]. https://CRAN.R-project.org/package=DT