

UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE INFORMÁTICA
HUALPÉN - CHILE



“DESARROLLO DE UN MODELO DE APRENDIZAJE AUTOMÁTICO PARA CLASIFICAR EL NIVEL DE HABILIDAD DE CICLISTAS”

YERKO TOMÁS MUÑOZ ESCOBAR

MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERO EN INFORMÁTICA

Profesor Guía: Javier Maldonado Carmona
Profesor Correferente: Jorge Ortiz Pino

Agosto - 2024

DEDICATORIA

Dedico esta memoria a todas las personas que formaron parte de este proceso. En especial,
a mi familia.

AGRADECIMIENTOS

Agradezco enormemente a los profesores que fueron de gran ayuda en estos últimos años.

Cristian, Gastón y Javier fueron una pieza fundamental en el constante desarrollo de mi carrera, siempre estuvieron allí para solucionar cualquier inconveniente. Además, los académicos quienes logré conocer un poco más este último semestre: José y Jorge, muchas gracias por el apoyo, consejos y enseñanzas.

A mis compañeros, que terminaron siendo grandes amigos y colegas, Rocio, Thomas, Abner y Pablo. Sin ustedes, hubiera sido complicado superar los distintos problemas que enfrenté durante mi etapa universitaria.

A mis abuelos, tíos y primos por apoyarme en la loca idea de irme de intercambio al otro lado del mundo.

A los amigos que me hice en Finlandia. Mitchell, Kevin, Daria y Philipp, son las mejores personas que pude haber conocido.

A Dios, por jamás abandonarme.

Gracias... totales.

RESUMEN

Resumen— La cantidad de ciclistas ha ido en aumento a través de los años, especialmente la cantidad de usuarios registrados en aplicaciones como Strava. Esto ha puesto en marcha panoramas cada vez más complejos de interpretar por herramientas digitales, ya que, al ser un deporte tan versátil, es difícil generalizar cuando se trata de rendimiento. En esta memoria, se ha llevado a cabo una revisión literaria respecto a que variables influyen más en el rendimiento del ciclista, con el fin de comprender que aspectos tienen mayor relevancia a la hora de clasificar su rendimiento. Esto se ha dividido en dos capacidades fisiológicas: capacidad aeróbica y capacidad anaeróbica. Para construir el modelo de clasificación se combinaron estimaciones de estudios ya hechos, aplicados en un dataset open source del proyecto GoldenCheetah. Las categorías fueron establecidas como: Casual, Principiante, Experimentado y Élite. Los resultados de la clasificación, utilizando XGBoost, entregan una precisión máxima de 89 % aproximadamente respecto a la capacidad aeróbica y 91 % respecto a la capacidad anaeróbica, utilizando variables como la potencia promedio, edad, distancia total, frecuencia cardíaca promedio, frecuencia de entrenamiento semanal, velocidad promedio, ganancia de elevación, total de actividades, duración del entrenamiento y trabajo total.

Palabras Clave— Aprendizaje Automático, XGBoost, Ciclistas, GoldenCheetah, Rendimiento.

ABSTRACT

Abstract— The number of cyclists has been increasing over the years, especially the number of registered users on apps such as Strava. This has set in motion increasingly complex scenarios to be interpreted by digital tools, since, being such a versatile sport, it is difficult to generalize when it comes to performance. In this report, a literature review has been conducted regarding which variables most influence cyclist performance, in order to understand which aspects have the most relevance when classifying performance. This has been divided into two physiological capacities: aerobic capacity and anaerobic capacity. To build the classification model, estimates from previous studies were combined and applied in an open source dataset from the GoldenCheetah project. The categories were established as: Casual, Beginner, Experienced and Elite. The classification results, using XGBoost, deliver a maximum accuracy of approximately 89 % for aerobic capacity and 91 % for anaerobic capacity, using variables such as average power, age, total distance, average heart rate, weekly training frequency, average speed, elevation gain, total activities, training duration and total work.

Keywords— Machine Learning, XGBoost, Cyclists, GoldenCheetah, Performance.

GLOSARIO

CP: Critical Power.

ATP: Adenosín Trifosfato.

VO₂max: Consumo máximo de oxígeno.

FTP: Umbral funcional de Potencia o Functional Threshold Power.

WPK: Watts Per Kilogram o Watts (Potencia) por kilogramos.

RPE: Rate of Perceived Exertion o Índice de Esfuerzo Percibido.

VT: Ventilatory Threshold o Umbral ventilatorio.

LT: Lactate Threshold o Umbral de Lactato.

ML: Machine Learning o aprendizaje automático.

TT: Time-trials o pruebas contra-reloj.

SNP: Polimorfismo de nucleótido único.

IMC: Índice de masa corporal.

KNN: K-nearest neighbors.

SHAP: SHapley Additive exPlanations.

API: Application programming interface.

Dataset: Conjunto de datos.

UTFSM: Universidad Técnica Federico Santa María.

FWS: Forward Wrapper Selection.

ÍNDICE DE CONTENIDOS

RESUMEN	IV
ABSTRACT	V
GLOSARIO	VI
ÍNDICE DE FIGURAS	IX
ÍNDICE DE TABLAS	X
INTRODUCCIÓN	1
CAPÍTULO 1: DEFINICIÓN DEL PROBLEMA	2
1.1 Contexto de la memoria	2
1.2 Problema	2
1.3 Objetivos	3
1.3.1 Objetivo General	3
1.3.2 Alcance	4
1.3.3 Objetivos específicos	4
CAPÍTULO 2: MARCO CONCEPTUAL	5
2.1 Capacidades fisiológicas	5
2.2 Capacidad aeróbica	6
2.2.1 Volumen máximo de oxígeno	7
2.3 Capacidad anaeróbica	9
2.4 Cadencia de pedaleo	11
2.5 Distintas potencias en el ciclismo	12
2.5.1 Potencia absoluta	12
2.5.2 Potencia relativa	12
2.6 Intensidad	14
2.6.1 Frecuencia cardíaca	15
2.7 Potencia crítica	16
2.8 Características antropométricas	18
2.8.1 ¿Como se desarrollan las características antropométricas en distintos escenarios?	19
2.9 Clasificación de ciclistas	19
2.10 Análisis de datos	21
2.11 Selección de características	23
2.11.1 Método wrapper	24
2.12 Aprendizaje automático	25
2.12.1 Aprendizaje Supervisado	27
2.12.2 Decision Tree - Árboles de decisiones	29
2.12.3 Gradient Boosting	29

2.12.4	Extreme Gradient Boosting	29
2.12.5	Características del modelo	30
2.12.6	Aprendizaje por conjuntos - Ensemble Learning	32
2.13	Evaluación del rendimiento en modelos de machine learning	32
2.13.1	Métricas de evaluación	32
2.13.2	Validación cruzada	33
2.13.3	Curva ROC y Área Bajo la Curva (AUC-ROC)	33
2.13.4	SHapley Additive exPlanations	34
2.14	Metodología	34
CAPÍTULO 3: PROPUESTA DE SOLUCIÓN		36
3.1	Fuente de datos	36
3.2	Análisis exploratorio - Parte I	37
3.3	Análisis exploratorio - Parte II	42
3.4	Pre-procesamiento	46
3.5	Entrenamiento XGBoost	50
3.5.1	Aplicación del modelo seleccionado	51
3.6	F1-Score y Matriz de Confusión	53
3.6.1	SHAP Values	56
3.6.2	Selección de características	57
CAPÍTULO 4: VALIDACIÓN DE LA SOLUCIÓN		60
4.1	Preparación de los datos	60
4.2	Resultados finales	61
4.3	Validación página web	63
CAPÍTULO 5: CONCLUSIONES		67
5.1	Desempeño de modelos iniciales	67
5.2	Selección de características y resultados finales	67
5.3	Página web y trabajos futuros	68
REFERENCIAS BIBLIOGRÁFICAS		69

ÍNDICE DE FIGURAS

1	Distintos sensores en bicicletas.	3
2	Relación entre características fisiológicas.	6
3	Producción aeróbica.	8
4	Representación de umbrales - Adaptación	10
5	Representación del pedaleo	11
6	Greipel (1,84 y 75 kilos) versus Cavendish (1,75 y 70 kilos).	13
7	Etapas planas.	14
8	Pruebas contrarreloj y fases de alta montaña.	14
9	Relación entre HR y Power (W').	16
10	Ejemplo del critical power.	17
11	Criterio para clasificación de ciclistas.	20
12	Etapas del análisis de datos.	22
13	Representación selección de características.	23
14	Adaptación de como funciona el método wrapper	24
15	Machine Learning como sub-campo de IA	26
16	Categorías en Machine Learning.	26
17	Ejemplo distintas categorías ML.	28
18	Ejemplo de arboles de decisiones	30
19	Representación metodología CRISP-DM.	35
20	Output GoldenCheetah-OpenData.	36
21	Columnas atletas dataset.	38
22	Explorando datos de atletas.	38
23	Explorando datos de actividades.	38
24	Actividades mmp.	39
25	Actividades.	39
26	Actividades.	39
27	Flujo de trabajo 3 datasets.	40
28	Valores únicos de las edades.	41
29	Filtro top 10 edades negativas con más actividades.	42
30	Filtro top 10 edades positivas con más actividades.	42
31	Comparación picos de potencia promedio entre edades.	43
32	Hombres y mujeres.	43
33	Diferencias de potencias.	43
34	Distribución de actividades por atleta.	44
35	Matriz de correlación en distintas potencias.	44
36	Relación entre capacidad anaeróbica y aeróbica.	45
37	Número de actividades y potencia absoluta.	45
38	Relación entre edad y rendimiento (potencia relativa).	46
39	Diferencias entre edades válidas y potencialmente válidas.	47
40	Diferencias entre edades válidas y outliers.	48
41	Potencia relativa en distintas categorías.	49
42	Comparación de modelos.	51

43	Resultados - Capacidad anaeróbica.	52
44	Resultados - Capacidad aeróbica.	53
45	Matriz de confusión de ambas predicciones.	54
46	Feature importance - Rosa Aeróbico y Azul Anaeróbico.	54
47	Gain y cover - Aeróbico.	55
48	Gain y cover - Anaeróbico.	56
49	SHAP Values - Anaeróbico.	57
50	SHAP Values - Aeróbico.	57
51	Forward Wrapper Selection.	58
52	Falsos y verdaderos positivos.	59
53	Curva ROC-AOC - Ambos modelos.	63
54	Página principal donde se sube el csv o se ingresan los datos manualmente. .	65
55	Carga del conjunto de datos - Aplica el modelo XGBoost entrenado.	65
56	Resultados predicción en la web.	66
57	Visualización de resultados.	66

ÍNDICE DE TABLAS

1	Valores fisiológicos de ciclistas profesionales	16
2	Clasificación de deportistas según nivel de actividad física y rendimiento. . .	21
3	Tipos de análisis de datos. Fuente: [Zapier, 2023].	22
4	Métodos de selección de características. Fuente: Elaboración propia.	25
5	Descripción de características	37
6	Resumen estadístico del dataset de atletas (n = 1400). Fuente: Elaboración propia.	41
7	Categorización de la capacidad anaeróbica. Fuente: Elaboración propia. . . .	48
8	Categorización de la capacidad aeróbica. Fuente: Elaboración propia.	49
9	Distribución de Atletas por Categoría y Capacidad. Fuente: Elaboración propia.	50
10	Resumen de Resultados de los Modelos. Fuente: Elaboración propia.	50
11	Características utilizadas en el modelo XGBoost. Fuente: Elaboración propia. .	52
12	Clasificación de la frecuencia y duración del entrenamiento.	61
13	Capacidad Anaeróbica. Fuente: Elaboración propia.	62
14	Capacidad Aeróbica. Fuente: Elaboración propia.	62
15	Categorización de la capacidad anaeróbica (1min). Fuente: Elaboración propia.	64
16	Categorización de la capacidad aeróbica (12min). Fuente: Elaboración propia.	64
17	Características más importantes para la predicción de capacidades.	68

INTRODUCCIÓN

El crecimiento constante del número de ciclistas se ha convertido en una tendencia notable en la sociedad contemporánea. A nivel mundial, se estima que hay 1.000 millones de bicicletas en uso, lo que demuestra la magnitud de este fenómeno [Smith, 2023]. En China, la bicicleta sigue siendo un medio de transporte esencial, con una flota estimada de 400 millones de bicicletas en 2021 [Statista, 2023]. Este aumento en la popularidad del ciclismo a nivel global ha generado un interés creciente en el tema, como lo demuestra el incremento en el uso de aplicaciones móviles especializadas, entre las que destaca Strava.

Según el informe más reciente, Strava cuenta con más de 10 billones de actividades compartidas y tiene miembros en la comunidad de 190 países, con un promedio de 40 millones de actividades subidas por semana [Strava, 2020]. Este crecimiento refleja el impacto significativo del registro de actividades y el interés constante de los usuarios en su rendimiento.

Aunque Strava es una opción popular para registrar el rendimiento de las actividades ciclistas, existen otras aplicaciones similares como MapMyRide, Endomondo y Runtastic, que cumplen una función similar. Estas aplicaciones permiten a los usuarios registrar de manera resumida y amigable aspectos clave de sus actividades físicas, como la duración, la distancia recorrida y las calorías quemadas, etc.

Sin embargo, más allá de las variables tradicionalmente medidas por estas aplicaciones, existen otros indicadores de rendimiento que pueden ser relevantes para los ciclistas y que podrían contribuir significativamente a su mejora. Estos incluyen mediciones fisiológicas como el consumo máximo de oxígeno ($VO_{2\text{máx}}$), potencia crítica, potencia absoluta, frecuencia cardíaca, velocidad, entre otros.

Estos indicadores, que evalúan diversas características del cuerpo humano, pueden resultar difíciles de interpretar y no siempre proporcionan una comprensión inmediata de su significado en términos de rendimiento ciclista. Lo que nos lleva a cuestionarnos si existe una forma sencilla de cuantificar estos índices.

Desarrollar un modelo simple para clasificar los niveles de rendimiento de los ciclistas podría ser clave para diferenciar entre principiantes, experimentados y élite según sus índices, lo que proporcionaría una base para comprender los límites individuales y buscar mejoras continuas sin exceder las capacidades físicas. Y así también, contribuir a que los ciclistas se mantengan motivados cada vez más respecto a la mejora continua.

CAPÍTULO 1

DEFINICIÓN DEL PROBLEMA

1.1. Contexto de la memoria

Esta memoria se inscribe en el proceso de obtención del título de pregrado del Departamento de Electrónica e Informática de la Universidad Técnica Federico Santa María, sede Concepción, en la carrera de Ingeniería en Informática. Se centra en el campo de la ciencia de datos, con un enfoque específico en la clasificación y creación de modelos para ciclistas a través de Machine Learning.

El proyecto se enfoca en el desarrollo de un modelo para clasificar el nivel de habilidad de los ciclistas. Para ello, se empleará como fuente de datos el proyecto GoldenCheetah [Liversedge, 2021], puesto que contiene una amplia cantidad de datos valiosos para el fin de este estudio. Además, el dataset cuenta con distintas métricas comúnmente utilizadas en el mundo de los ciclistas para evaluar y clasificar. Cabe mencionar que los datos están completamente anonimizados, no hay nombres, ni ningún dato que pueda identificar la identidad del atleta.

1.2. Problema

Aunque hoy en día contamos con tecnología para registrar actividades físicas como el ciclismo, los datos obtenidos (calorías, distancia, velocidad, cadencia, etc.) no son suficientes para evaluar y mejorar el rendimiento de manera precisa. Dispositivos como bike computers y sensores de cadencia (figura 1), que miden las revoluciones por minuto de nuestro pedaleo, ofrecen información valiosa sobre aspectos como la velocidad y la intensidad del ejercicio. Sin embargo, sin un análisis adecuado que tenga en cuenta otros factores, su utilidad es limitada a la hora de comprender el panorama completo del rendimiento del ciclista.

Incluso los medidores de potencia, que miden la potencia generada por el ciclista en vatios, a pesar de su larga trayectoria en el mercado «con pioneros como SRM desde hace más de 40 años» presentan problemas de precisión y no ofrecen una visión completa del rendimiento. Aumentar la velocidad o la cadencia, por ejemplo, no siempre refleja una mejora en la condición física, ya que puede estar influenciada por factores externos como el terreno o el viento.

De hecho, como se discute en [Passfield *et al.*, 2017], persisten deficiencias en cuanto a la precisión, validez y fiabilidad de estas herramientas. Este panorama plantea interrogantes sobre qué variables son las más indicativas para cuantificar el rendimiento, cómo podemos ajustar la precisión de los dispositivos existentes y de qué manera, ciertos valores se traducen



Figura 1: Distintos sensores en bicicletas.
Fuente: iGPSPORT.

en una mejora real para el ciclista.

Si bien los medidores de potencia y otros sensores han sido una herramienta valiosa para los ciclistas, sus limitaciones han dejado un vacío en la comprensión integral del rendimiento. Esta tesis propone una solución innovadora para abordar este problema: un modelo que, al considerar no solo los datos de entrenamiento, sino también variables fisiológicas como la edad, el peso y la frecuencia cardíaca, puede clasificar el rendimiento del ciclista en aspectos clave como la capacidad aeróbica y anaeróbica. Al integrar estos datos, el modelo podrá ofrecer una evaluación más precisa y personalizada, permitiendo a los ciclistas entender mejor su rendimiento y tomar decisiones informadas para mejorarlo.

1.3. Objetivos

1.3.1. Objetivo General

Este estudio tiene como objetivo desarrollar un modelo capaz de categorizar el rendimiento general de ciclistas, con el fin de proporcionar una herramienta de evaluación más precisa y personalizada que permita a los ciclistas y entrenadores optimizar sus planes de entrenamiento y mejorar su rendimiento. Para ello, se analizarán en profundidad las relaciones

entre variables fisiológicas, antropométricas y de entrenamiento, con especial énfasis en la capacidad aeróbica y anaeróbica. A través de técnicas de machine learning, se construirá un modelo que permita descubrir patrones y relaciones complejas en los datos para clasificar a los ciclistas en diferentes categorías. El modelo será rigurosamente evaluado en dos conjuntos de datos independientes para asegurar su precisión y capacidad de generalización a nuevas situaciones.

1.3.2. Alcance

Debido a la gran cantidad de datos en el archivo base (102GB), se ha reducido significativamente mediante la creación de tres dataframes más manejables. Además de esto, cabe mencionar que se busca crear una propuesta únicamente en base a los datos recolectados por GoldenCheetah, el modelo considerará solo algunas variables (que han demostrado ser indicativas) por lo cual, puesto que el mundo del ciclismo es demasiado ambiguo, puede que los resultados no sean tan precisos como evaluaciones hechas en laboratorios.

1.3.3. Objetivos específicos

1. Realizar una revisión bibliográfica para comprender los diferentes perfiles de ciclistas y su relación con el rendimiento.
2. Analizar los datos de GoldenCheetah para obtener una visión general de su estructura y contenido.
3. Preparar el conjunto de datos para su análisis mediante técnicas de pre-procesamiento.
4. Analizar la relación entre las variables y el rendimiento del ciclista utilizando técnicas de modelamiento estadístico.
5. Evaluar la capacidad predictiva de los modelos desarrollados e identificar las características más relevantes.
6. Construir un modelo de clasificación que permita categorizar el rendimiento del ciclista. Validar este mismo en otro conjunto de datos.

CAPÍTULO 2

MARCO CONCEPTUAL

En el mundo del ciclismo, la medición del rendimiento va más allá de simples variables como la distancia, el tiempo y la potencia. La medicina deportiva, o "Sports Medicine", juega un papel crucial en la evaluación de las capacidades físicas individuales de los ciclistas [Faria *et al.*, 2005b]. Para entender cómo estas capacidades se traducen en rendimiento, es fundamental distinguir entre las características físicas, que se refieren al cuerpo en sí, y las fisiológicas, que describen las funciones del organismo [Dieguez *et al.*, 2016]. Ambas son esenciales para comprender el desarrollo, los efectos y las diferencias entre individuos en este deporte.

A continuación, se explorará en detalle el perfil teórico del ciclista, analizando todas las variables y conceptos relevantes para cuantificar su rendimiento, desde las capacidades aeróbicas y anaeróbicas hasta aspectos como la intensidad, antropometría, potencia crítica y la cadencia de pedaleo. Este análisis sentará las bases para comprender cómo estas características interactúan y contribuyen al rendimiento general del ciclista. Así también, se revisarán los conceptos informáticos y la metodología a utilizar.

2.1. Capacidades fisiológicas

Desde una perspectiva biológica, la fisiología se define como «la rama de la biología que se ocupa de las funciones y actividades de la vida o de la materia viva (como órganos, tejidos o células) y de los fenómenos físicos y químicos que intervienen en ella» [Merriam-Webster, nd]. En el ámbito del ciclismo, estas capacidades fisiológicas son de suma importancia para evaluar el rendimiento, dado que las competiciones abarcan desde sprints de unos pocos segundos hasta eventos de resistencia como el Tour de France, que recorre aproximadamente 5,000 km en 23 días [Faria, 1984]. Este amplio espectro de exigencias competitivas requiere que los ciclistas desarrollen y mantengan tanto su capacidad aeróbica como anaeróbica.

La potencia necesaria para alcanzar una velocidad determinada depende de una compleja interacción entre variables fisiológicas, como el consumo máximo de oxígeno $VO_{2\text{máx}}$, el umbral anaeróbico, la economía y eficacia del movimiento [Jeukendrup *et al.*, 2000, Lucia *et al.*, 1999a, Lucia *et al.*, 2000, Balmer *et al.*, 2000]. Así también, se ven influenciadas variables mecánicas (tipo de bicicleta, neumáticos, frenos, etc.) y ambientales (velocidad del viento, temperatura, humedad, altitud).

De hecho, en estudios como [Paul William Macdermid y Stannard, 2015] y [Silveira y Rama, 2020] han demostrado diferencias significativas en el rendimiento ciclista según el tipo de neumático y la superficie, tanto en desaceleración ($P = 0.014$) como

en escalada ($P = 0.032$). Estos estudios revelan, por ejemplo, que neumáticos más anchos pueden mejorar el agarre en curvas, mientras que neumáticos más lisos pueden reducir la resistencia al rodamiento.

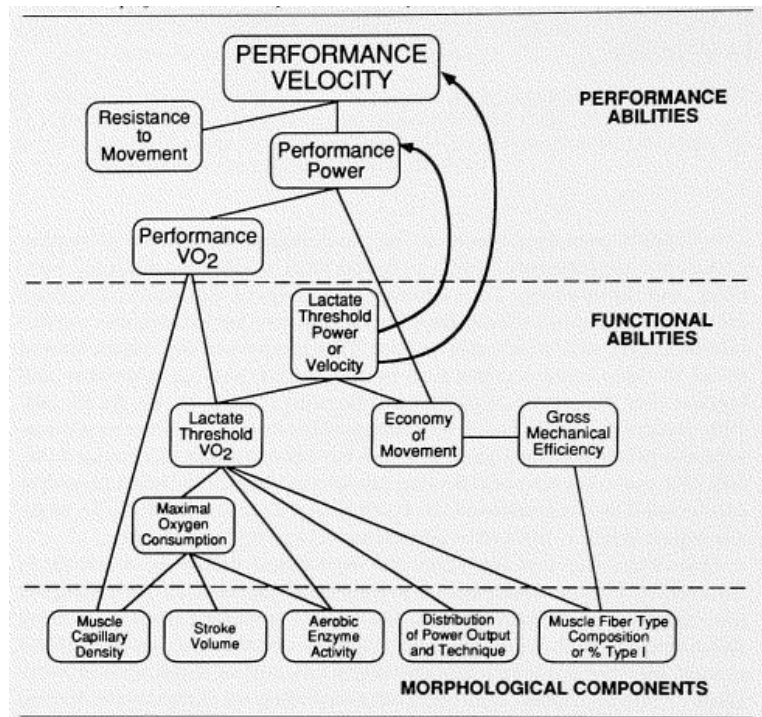


Figura 2: Relación entre características fisiológicas.
Fuente: [Coyle, 1999a].

Con el fin de no perder el foco, se estudiarán principalmente las características mencionadas al comienzo de este capítulo, omitiendo detallar en conceptos relevantes como la economía y eficacia del ciclista, aspectos nutricionales/ambientales y propiedades mecánicas como los tipos de bicicleta y su composición. Aún así, se mencionarán algunos aspectos complementarios que tienen una pequeña relación y contribución al propósito de esta memoria. La figura 2 detalla más sobre la relación entre distintos factores fisiológicos.

2.2. Capacidad aeróbica

Si vamos a la RAE, la definición que nos entrega para aeróbico, se define como «aquella que se produce con la utilización de oxígeno libre» [Real Academia Española, 2024]. Ahora, desde una perspectiva más científica, en la literatura sobre fisiología en ciclistas, la capacidad aeróbica se define como «la máxima capacidad del organismo para transportar y utilizar oxígeno durante el ejercicio físico» [Faria *et al.*, 2005b].

Esta última definición se basa en la comprensión de que el oxígeno es esencial para la producción de energía durante el ejercicio aeróbico. Cuanto mayor sea la capacidad aeróbica de

un individuo ($\text{VO}_2 \text{ max}$), más eficientemente podrá su cuerpo utilizar el oxígeno para generar energía, lo que se traduce en un mejor rendimiento en actividades de resistencia, puesto que existen competencias, como se mencionó anteriormente, donde los ciclistas profesionales deben ser capaces de tolerar altas cargas de trabajo. Algunos ejemplos de competencias aparte del Tour de France, son el Giro de Italia y Vuelta a España.

2.2.1. Volumen máximo de oxígeno

El consumo máximo de oxígeno ($\text{VO}_2 \text{ máx}$), también conocido como volumen máximo de oxígeno, es una variable fundamental en la evaluación del rendimiento deportivo. Representa la cantidad máxima de oxígeno que el organismo puede utilizar durante el ejercicio y está determinado tanto por la capacidad del sistema cardiovascular para suministrar oxígeno como por la demanda de los tejidos musculares [Bassett y Howley, 2000].

El ejercicio se suele clasificar en dos categorías principales: aeróbico y anaeróbico. El ejercicio anaeróbico se caracteriza por su alta intensidad y corta duración, lo que obliga al cuerpo a generar energía rápidamente sin utilizar oxígeno, limitando así la duración de este tipo de esfuerzo.

Por otro lado, el ejercicio aeróbico, como el ciclismo de resistencia, es de menor intensidad y mayor duración. En este caso, la energía se produce en presencia de oxígeno, lo que permite mantener la actividad durante períodos más prolongados.

Es en este contexto, donde el $\text{VO}_2 \text{ máx}$ adquiere una importancia crucial, ya que este indicador nos proporciona datos concretos para evaluar tanto la capacidad como la eficiencia aeróbica de un individuo. En otras palabras, el $\text{VO}_2 \text{ máx}$ nos permite cuantificar la habilidad del cuerpo para consumir y utilizar oxígeno durante el ejercicio, lo cual es un factor determinante en el rendimiento de los ciclistas de resistencia. El $\text{VO}_2 \text{ máx}$ se expresa comúnmente en mL/kg/min o L/min .

Este se puede estimar mediante diversas ecuaciones. Una de ellas es la siguiente:

$$\text{VO}_2 (\text{mL/min}) = \frac{\text{kg}}{\text{min}} \times 1,9 \frac{\text{mL}}{\text{min}} + \left[3,5 \frac{\text{mL}}{\text{kg} \cdot \text{min}} \times \text{kg peso corporal} \right] + 260 \frac{\text{mL}}{\text{min}} \quad (1)$$

Donde:

- 1.9 mL/min = Relación entre VO_2 y potencia. 3.5 mL/kg/min = VO_2 en reposo.
- 260 mL/min = Costo de oxígeno del ciclismo sin carga.

Aun así, esta ecuación no se recomienda para cargas de potencia superiores al umbral de lactato [Latin *et al.*, 1993]. Además, la precisión de la estimación del VO_2 máx a partir de ecuaciones de regresión, al menos en ciclistas de alto nivel, puede ser inexacta debido a una atenuación de la relación VO_2 / W a altas cargas de trabajo, una característica distintiva de los mejores ciclistas [Lucía *et al.*, 2002].

En entornos reales, el gasto energético en términos de VO_2 para el ciclismo al aire libre puede estimarse utilizando la ecuación presentada por [McCole *et al.*, 1990]:

$$\text{VO}_2 (\text{L/min}) = -4,5 + 0,17 \cdot C_1 + 0,052 \cdot C_2 + 0,022 \cdot C_3 \quad (2)$$

Donde:

- C_1 = Velocidad del ciclista (km/h).
- C_2 = Velocidad del viento (km/h).
- C_3 = Peso del ciclista (kg).

La siguiente imagen (figura 3) muestra un esquema simplificado de la producción de energía aeróbica durante el ejercicio. Bajo el contexto de esta memoria se considera relevante conocer las bases del funcionamiento aeróbico, para más detalles, revisar [Coyle, 1999b].

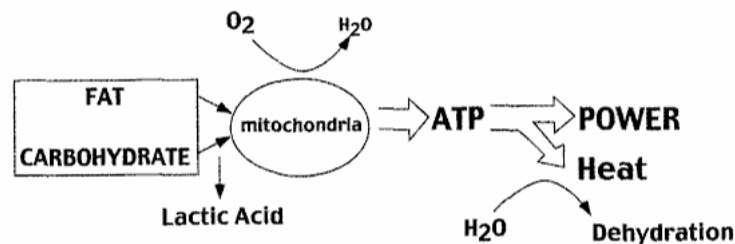


Figura 3: Producción aeróbica.

Fuente: [Coyle, 1999b].

Se observa cómo la mayor parte de la energía utilizada por los ciclistas proviene de la resíntesis de ATP^1 en las mitocondrias², donde se oxidan carbohidratos y grasas, [Coyle, 1995] destaca su importancia. Estos ciclistas, al enfrentarse a diferentes condiciones de disponibilidad de oxígeno, carbohidratos y grasas, pueden experimentar variaciones significativas

¹Molécula que actúa como la principal fuente de energía para la mayoría de los procesos celulares.

²La resíntesis de ATP transforma la energía de los alimentos en energía utilizable.

en su rendimiento. Por ejemplo, cuando los niveles de oxígeno son bajos o la disponibilidad de carbohidratos es limitada, el cuerpo puede acumular iones de hidrógeno, lo que conduce a la acidosis y afecta negativamente el rendimiento [Robergs *et al.*, 2004]. Además, la generación de calor como subproducto de estas reacciones químicas puede resultar en hipertermia³ si no se disipa eficientemente, lo que puede limitar aún más el rendimiento de los ciclistas, especialmente en entornos desafiantes [Febbraio *et al.*, 1994]. Por ejemplo, en condiciones de alta temperatura y humedad. La deshidratación, causada por la pérdida de líquidos debido a la sudoración durante el ejercicio prolongado, puede agravar estos efectos al contribuir a la hipertermia y afectar la función cardiovascular de los ciclistas. Asimismo, estrategias como la ingesta adecuada de líquidos y carbohidratos pueden ayudar a los atletas de resistencia a mitigar estos efectos y mantener un suministro óptimo de energía durante el ejercicio [González-Alonso *et al.*, 1995]. Cabe recalcar que no se entrará más en detalle respecto a conceptos médicos y nutricionales puesto que se aleja del propósito de esta memoria.

2.3. Capacidad anaeróbica

El término “anaeróbico” describe los procesos metabólicos que ocurren en ausencia de oxígeno. A diferencia del metabolismo aeróbico, que es más eficiente pero requiere oxígeno, el metabolismo anaeróbico permite al organismo generar energía rápidamente, aunque de forma menos eficiente, cuando la demanda de oxígeno supera la disponibilidad [Wagner, 2000]. Este proceso depende de la concentración y actividad de enzimas específicas.

Se puede clasificar de dos formas:

1. **Capacidad anaeróbica aláctica:** Se basa en la utilización de reservas de fosfocreatina (PCr) para resintetizar adenosín trifosfato (ATP), la moneda energética del cuerpo. Esta vía es extremadamente rápida pero de corta duración, permitiendo esfuerzos máximos de hasta 10 segundos.
2. **Capacidad anaeróbica láctica:** Implica la degradación de glucosa⁴ (o glucógeno muscular) a través de la glucólisis⁵, generando ATP y ácido láctico⁶ como subproducto. Según [Burke, 1997], la glicólisis anaeróbica es el proceso mediante el cual el cuerpo descompone la glucosa en ausencia de oxígeno para producir ATP, esencial para el rendimiento en actividades de alta intensidad. Esta vía es más lenta que la aláctica, pero puede sostener esfuerzos de mayor duración (entre 30 segundos y 2 minutos).

Un ejemplo del mundo real se puede observar en el Tour de France, donde en las etapas de

³La hipertermia es un aumento de la temperatura corporal por encima de lo normal.

⁴Tipo de azúcar simple. Principal fuente de energía para las células del cuerpo.

⁵Permite a los ciclistas descomponer rápidamente la glucosa para obtener energía durante esfuerzos intensos.

⁶Se acumula en los músculos de los ciclistas durante el ejercicio intenso y puede contribuir a la fatiga muscular.

montaña, los ciclistas realizan esfuerzos anaeróbicos cuando atacan en las subidas empinadas o responden a los ataques (aceleraciones) de otros ciclistas. Un ejemplo notable fue el ataque de Chris Froome en la etapa 19 del Tour de France 2013, donde mostró una impresionante capacidad anaeróbica al responder y lanzar ataques repetidos en una pendiente pronunciada (Tour de France: 29 de junio al 21 de julio de 2013).

Por último, también existen formas de cuantificar la capacidad anaeróbica fuera de entornos de laboratorio. Una de ellas es:

$$W' = (\text{Potencia máxima} - \text{Potencia crítica}) \times \text{Tiempo hasta la extenuación} \quad (3)$$

Esta formula fue propuesta en un principio por [Monod y Scherrer, 1965] con el fin de calcular la capacidad anaeróbica total sin la necesidad de un equipamiento especializado. Existe una revisión un poco más actualizada de esta formula en [Morton, 1996].

Las otras han sido descartadas puesto que abarcan otras variables que están fuera del alcance y explicación de esta memoria, sin embargo se pueden encontrar en las siguientes investigaciones: Margaria-Kalamen [Kalamen, 1968], Test de Máxima Velocidad [Vandewalle et al., 1987] y Prueba de Sprints Repetidos [Bangsbo, 1996].

A continuación, la figura (4) adjunta ilustra la relación entre la intensidad del ejercicio, la concentración de lactato en sangre y la transición del metabolismo aeróbico (ya revisado) al anaeróbico.

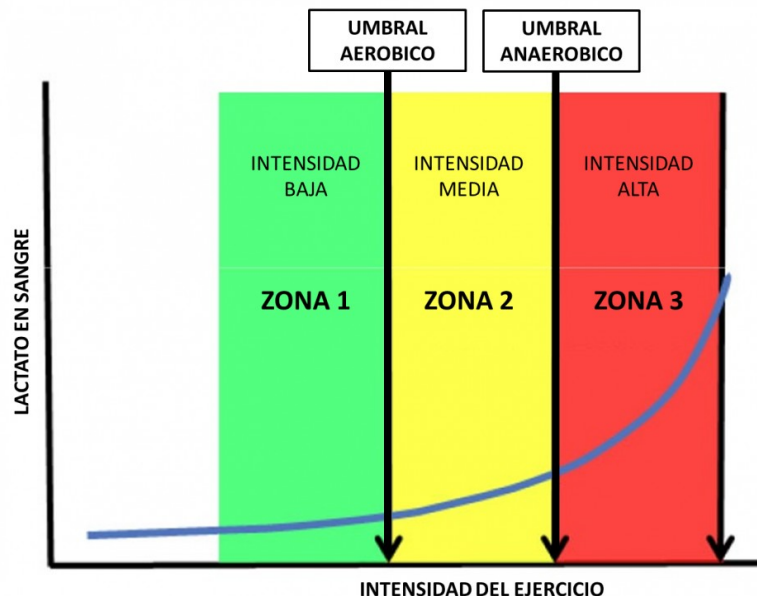


Figura 4: Representación de umbrales - Adaptación
Fuente: [Seiler y Tønnessen, 2009]

2.4. Cadencia de pedaleo

La cadencia de pedaleo es un factor determinante en el rendimiento ciclista, aunque no existe un consenso sobre los criterios óptimos para su selección [Takaishi *et al.*, 1994, Takaishi *et al.*, 1996]. Investigaciones previas sugieren que la cadencia influye en la fatiga neuromuscular y en el patrón de reclutamiento de fibras musculares [Coggan, 2003]. A medida que aumenta la cadencia, se reclutan menos fibras musculares de contracción rápida (tipo II) y más fibras de contracción lenta (tipo I), en respuesta a la menor fuerza requerida por revolución del pedal [Ahlquist y Bassett, 1992]. Usualmente, la cadencia se mide en RPM (Revoluciones por minuto) bajo la variable “CAD”.

Es importante destacar que la eficiencia de la cadencia de pedaleo se ve influenciada por diversos factores, como la longitud de la biela, la posición del cuerpo, los desplazamientos lineales y angulares, las velocidades y aceleraciones de los segmentos corporales, y las fuerzas en las articulaciones y músculos [Faria y Cavanagh, 1978]. Además, la experiencia en ciclismo también juega un papel importante en la eficiencia metabólica a diferentes cadencias [Lucía *et al.*, 2001a]. Hasta la fecha de un estudio hecho por [Faria *et al.*, 2005a] no se ha encontrado una cadencia de pedaleo óptima en la literatura. La imagen 5 representa como funciona el pedaleo.

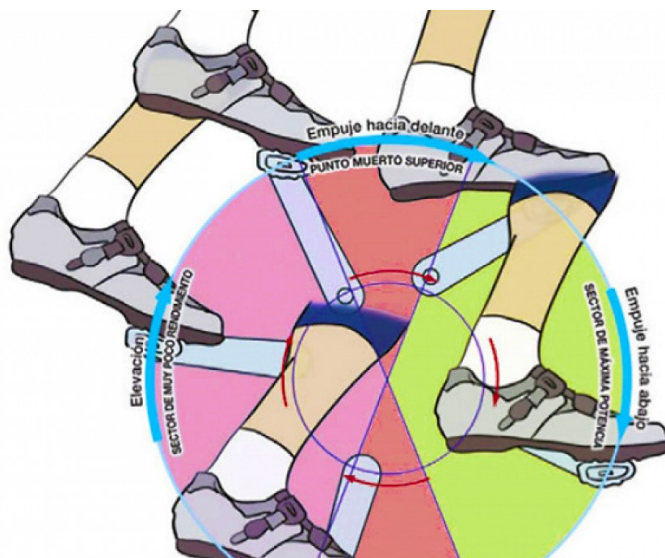


Figura 5: Representación del pedaleo
Fuente: [Viivo, 2017]

Estudios en ciclistas profesionales han demostrado que la eficiencia bruta (GE)⁷ aumenta con la cadencia de pedaleo, alcanzando valores máximos a 100 rpm, junto con una disminución del consumo de oxígeno, frecuencia cardíaca, percepción subjetiva de esfuerzo y actividad muscular [Lucía *et al.*, 2004]. Estos hallazgos respaldan la noción de que una cadencia alta

⁷La eficiencia bruta (GE) se define como la relación entre la potencia mecánica producida y el gasto energético total.

puede ser beneficiosa para ciclistas bien entrenados, como lo demuestra el caso de Lance Armstrong, quien adoptó una cadencia inusualmente alta con resultados exitosos.

2.5. Distintas potencias en el ciclismo

Se ha hecho mención sobre la potencia en esta memoria en varias ocasiones, sin embargo, una definición clara y detallada es importante, puesto que la potencia en el contexto de ciclistas es una de las métricas más importantes para cuantificar el rendimiento. En comparación con medir la velocidad o la frecuencia cardíaca (se detallará en la siguiente sección) medir la potencia en el ciclismo ofrece una respuesta más directa e inmediata sobre la intensidad del esfuerzo. La potencia no solo determina la velocidad, sino que también influye en las respuestas cardiovascular, metabólica y perceptiva del ciclista. Además, los cambios en la potencia se detectan rápidamente, permitiendo un mejor control de la intensidad del entrenamiento. Medir regularmente la potencia durante el entrenamiento y las carreras proporciona un indicador directo de la eficacia del entrenamiento, facilitando ajustes para optimizar los resultados [Coggan, 2003].

2.5.1. Potencia absoluta

Es la cantidad total de energía que un ciclista genera por unidad de tiempo, medida en vatios (W). Representa la fuerza bruta aplicada a los pedales y es un indicador crucial del rendimiento general del ciclista. Una mayor potencia absoluta implica una mayor capacidad para superar la resistencia del aire y la gravedad, lo que se traduce en una mayor velocidad en diferentes terrenos.

La potencia absoluta es más relevante para el rendimiento en líneas rectas, aunque también influye el coeficiente aerodinámico del ciclista. A partir de 15 km/h, la mayor parte de la potencia se utiliza para superar la resistencia del aire, y la potencia requerida aumenta exponencialmente con la velocidad. Por lo tanto, un ciclista con mayor potencia absoluta puede mantener velocidades más altas al vencer la resistencia del aire de manera más eficiente.

2.5.2. Potencia relativa

Mientras que la potencia absoluta nos indica nuestro rendimiento en terrenos planos, existe otra potencia mucho más precisa y comúnmente utilizada para distintos escenarios denominada potencia relativa.

La potencia relativa utiliza la potencia absoluta (W) y el peso del ciclista para realizar un simple cálculo. Normalmente se verá de esta forma: $W \text{ kg}^{-1}$ en valores como, por ejemplo,

5,5 wpk o $W\text{ kg}^{-1}$, que sería el equivalente a ir a 400 W y tener un peso de 72kg, $400 / 72 = 5,5\text{ wpk}$.

Esta no solo es crucial en ascensos, sino también en situaciones que demandan esfuerzos breves e intensos, como sprints⁸, adelantamientos y pruebas contrarreloj. Esta capacidad de generar una alta potencia en cortos períodos de tiempo, conocida como potencia anaeróbica, puede evaluarse mediante pruebas específicas, como el protocolo de Baron [Baron, 2001], que consiste en series de 10 segundos a cadencias variables tras un calentamiento controlado.



Figura 6: Greipel (1,84 y 75 kilos) versus Cavendish (1,75 y 70 kilos).
Fuente: [Arberas, 2017].

La potencia relativa (W/kg) emerge como un indicador clave del rendimiento en ciclismo, especialmente relevante en situaciones que demandan esfuerzos breves e intensos como las ya mencionadas. Esta métrica, que relaciona la potencia generada con el peso corporal del ciclista, no solo es crucial en ascensos, sino también en pruebas de potencia anaeróbica (figura 6).

Además, la evidencia científica respalda la utilidad de la potencia pico ($W\text{ peak}$) como predictor del rendimiento en ciclismo [Bentley *et al.*, 1998, Hawley y Noakes, 1992]. La relación potencia/peso superior a 5.5 W/kg en 4 minutos se ha identificado como un requisito para ciclistas competitivos de alto nivel [Palmer *et al.*, 1994], lo que refuerza la importancia de esta métrica en la evaluación y clasificación del rendimiento ciclista.

Ya que se han definido las diferencias entre potencia relativa, absoluta, y así también, se han mencionado anteriormente escenarios de prueba como “*time-trials*” (pruebas contrarreloj) o “*sprints*”, las imágenes propuestas ejemplifican de mejor manera las distintas etapas que

⁸Un sprint se refiere a un esfuerzo máximo y explosivo de corta duración, generalmente en terreno llano, con el objetivo de alcanzar la máxima velocidad posible.

pueden enfrentar los ciclistas (figuras 7, 8a y 8b).

F (flat stages)

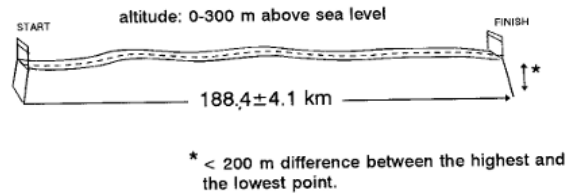
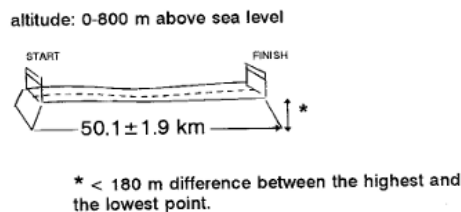
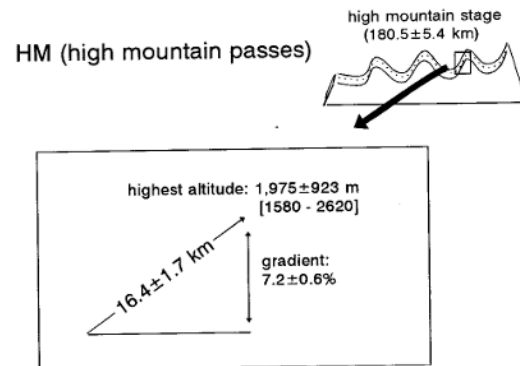


Figura 7: Etapas planas.
Fuente: [Lucía *et al.*, 2001b].

TT (time trials)



(a) Pruebas contrarreloj.



(b) Fases de alta montaña.

Figura 8: Pruebas contrarreloj y fases de alta montaña.
Fuente: [Lucía *et al.*, 2001b].

2.6. Intensidad

La intensidad en el ciclismo es un concepto clave para optimizar el entrenamiento y el rendimiento. Se refiere al nivel de esfuerzo físico requerido para realizar una actividad y puede ser medida a través de diversas variables, siendo las más comunes la frecuencia cardíaca (FC), la potencia (W) y la percepción subjetiva del esfuerzo (PSE) [Coggan, 2003].

El entrenamiento ideal para cualquier atleta es aquel que presenta el desafío suficiente para generar una mejora continua, pero sin llegar a ser tan exigente que provoque enfermedades, lesiones o sobre entrenamiento [Coggan, 2003]. Lograr este equilibrio puede ser difícil en cualquier deporte, pero especialmente en el ciclismo, debido a las variaciones en la resistencia al movimiento causadas por factores externos como la altitud, el clima, el terreno y la superficie de la carretera o pista, entre otros.

No obstante, la velocidad actual o incluso la velocidad promedio a menudo son indicadores deficientes de la intensidad del entrenamiento, lo que dificulta la regulación de la carga total de entrenamiento. Por lo tanto, se requiere una medida más precisa y fiable para cuantificar la intensidad del esfuerzo en ciclismo.

2.6.1. Frecuencia cardíaca

La frecuencia cardíaca, medida en latidos por minuto (ppm), es un indicador fisiológico ampliamente utilizado para medir la intensidad del ejercicio. Esta aumenta a medida que se incrementa el esfuerzo físico. La FC máxima (FC_{máx}) es el número máximo de latidos que el corazón puede alcanzar durante un esfuerzo máximo y varía según la edad y el estado físico del individuo [Bassett y Howley, 2000].

Para determinar la intensidad del ejercicio, se pueden utilizar porcentajes de la FC_{máx} o la FC de reserva (FC_{máx} - FC en reposo) [Bassett y Howley, 2000]. Aunque la potencia puede ser un indicador más directo de la intensidad del ejercicio, la FC sigue siendo un parámetro valioso para evaluar el nivel de esfuerzo alcanzado durante el entrenamiento y la competición en ciclismo profesional.

Cabe mencionar que, actualmente es bastante popular y utilizada la formula propuesta por [Fox *et al.*, 1971]

$$FoxFC_{max} = 220 - edad \quad (4)$$

lo cual no es ampliamente recomendada debido a su inexactitud. Al contrario, la formula propuesta por [Tanaka *et al.*, 2001] si queremos seguir utilizando la edad para predecir nuestro FC_{max}, obtiene mejores resultados.

$$TanakaFC_{max} = 208 - 0,7 \times edad \quad (5)$$

Así también, en el contexto específico de ciclistas, [Barboza *et al.*, 2017] ha recomendado una ecuación que incluye tanto la edad como la frecuencia cardíaca a 150 W obtenida durante una prueba de ejercicio graduada (GXT) en cicloergómetro en hombres jóvenes y sanos. Esta ecuación podría ofrecer una predicción más precisa de la FC_{max} al considerar un parámetro adicional relacionado con el rendimiento físico del individuo.

No obstante, la aplicabilidad de esta fórmula a ciclistas en general puede ser limitada. El estudio de [Barboza *et al.*, 2017] se centró en hombres jóvenes y sanos, por lo que sus resultados podrían no ser generalizables a otras poblaciones, como mujeres, personas mayores o personas con condiciones de salud específicas.

Como se puede apreciar en el siguiente gráfico (figura 9), la línea de tendencia superior tiene un R² de 0.824, lo que significa que explica el 82.4 % de la variabilidad en la frecuencia cardíaca. La línea de tendencia inferior tiene un R² de 0.991, lo que significa que explica el 99.1 %

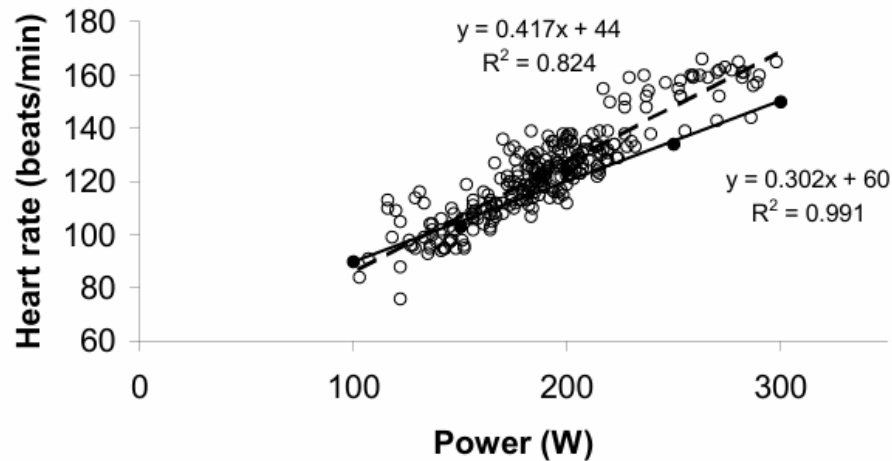


Figura 9: Relación entre HR y Power (W').

Fuente: [Coggan, 2003].

de la variabilidad. Esto sugiere que la relación entre la frecuencia cardíaca y la potencia es más fuerte a niveles de potencia más altos.

Por último, una tabla de mediciones (1) en ocho ciclistas de élite que fueron sometidos a una prueba de esfuerzo incremental en bicicleta estática un mes antes del Tour de France para determinar sus umbrales ventilatorios (VT1 y VT2) y las frecuencias cardíacas correspondientes. [Lucia *et al.*, 1999b].

Tabla 1: Valores fisiológicos de ciclistas profesionales

Fuente: [Lucia *et al.*, 1999b].

Parámetro	Valor
HRmax (latidos/min)	191 ± 8
VO ₂ max (L/min)	5.1 ± 0.4
VO ₂ max (ml/kg/min)	74.0 ± 5.8
Wmax (W)	501.2 ± 24.8

Los resultados de la tabla muestran las medias y desviaciones estándar (SD) de la frecuencia cardíaca máxima (HRmax), el consumo máximo de oxígeno (VO₂max), la potencia máxima (Wmax) y la frecuencia cardíaca en los umbrales. Los umbrales ventilatorios están omitidos dentro del fin de esta memoria puesto que son variables netamente medidas en laboratorios.

2.7. Potencia crítica

El Critical Power (CP) es un concepto fundamental en la fisiología del ejercicio que se define como la máxima potencia de salida que un individuo puede mantener de manera sostenible

sin agotar sus reservas de energía anaeróbica [Moritani *et al.*, 1981]. Esencialmente, es el umbral que separa los esfuerzos en los que el cuerpo puede mantener un equilibrio metabólico estable de aquellos en los que se produce un desequilibrio progresivo que conduce a la fatiga. El CP se sitúa en un rango del 70-90 % del consumo máximo de oxígeno ($\text{VO}_2\text{máx}$), dependiendo del estado de entrenamiento del individuo [Wilkerson *et al.*, 2004]. Esto lo hace relevante para una amplia gama de eventos deportivos, desde carreras de 800 metros hasta 10.000 metros [Wilkerson *et al.*, 2004].

En corredores de maratón de élite, por ejemplo, se ha observado que el ritmo promedio de carrera durante un maratón es cercano al 96 % de su velocidad crítica (CS, equivalente al CP en carrera) [Hill, 1925]. Esto sugiere que estos atletas regulan su ritmo para mantenerse cerca de su umbral de esfuerzo sostenible.

La siguiente gráfica (figura 10) muestra la relación entre la potencia desarrollada por un ciclista y el tiempo durante un ejercicio de alta intensidad. La curva descendente indica que a medida que aumenta el tiempo de ejercicio, la potencia que el ciclista puede mantener disminuye. La potencia crítica (CP) representa la máxima potencia que el ciclista puede sostener indefinidamente. El punto W' corresponde a una potencia inicial alta, que el ciclista solo puede mantener durante un tiempo muy corto. El umbral de lactato (LT) marca el punto a partir del cual la producción de lactato aumenta significativamente, lo que limita la capacidad de continuar ejercitándose a alta intensidad

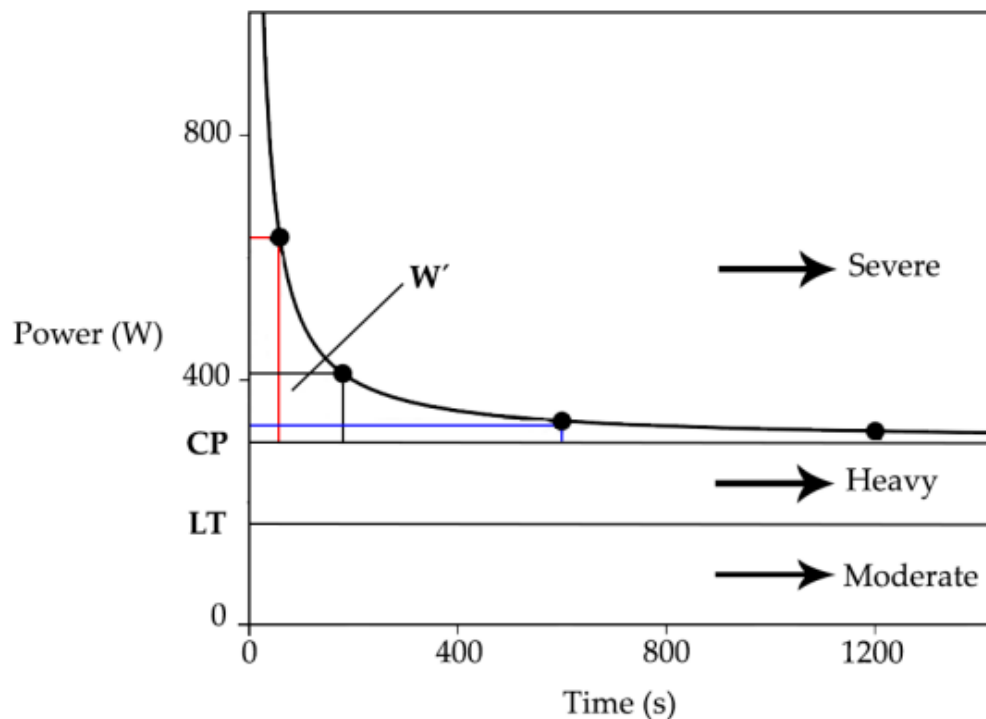


Figura 10: Ejemplo del critical power.
Fuente: [Jones y Vanhatalo, 2017].

Adicionalmente, se presentan algunas formulas para calcular el critical power. Una de ellas, es la siguiente:

$$P = \frac{W'}{t} + CP \quad (6)$$

Donde:

- (P) es la potencia media ejercida durante un intervalo de tiempo determinado.
- (W') es la cantidad de trabajo (energía) que el ciclista puede realizar por encima de su CP, normalmente en kJ.
- (t) es la duración del intervalo de tiempo en el que se ejerce la potencia (P).
- (CP) es el critical power, el valor que buscamos estimar

Alternativamente, se puede linealizar la ecuación mediante una transformación matemática. Al restar CP de ambos lados y multiplicar por (t), obtenemos:

$$(P - CP) \cdot t = W' \quad (7)$$

Esta ecuación representa una línea recta donde la pendiente es W' y el intercepto en el eje y es -CP. Al realizar una regresión lineal sobre los datos transformados, se pueden estimar los valores de CP y W' a partir de la pendiente y el intercepto de la línea de mejor ajuste.

2.8. Características antropométricas

El ciclismo de ruta, caracterizado por la diversidad de terrenos y exigencias competitivas, ha llevado a la especialización de los ciclistas en diferentes modalidades, como contrarreloj (TT), pruebas en terrenos llanos y escalada (ver imágenes 8a). Estas especialidades presentan diferencias significativas en las variables antropométricas, es decir, en las características morfológicas de los ciclistas que pueden influir en su rendimiento (masa corporal, altura, superficie corporal y áreas frontales, índice de masa corporal (IMC), peso (kg), etc.) [Padilla *et al.*, 1999]. Para más detalle, la imagen citada en el capítulo 2 [2.1] demuestra la relación entre componentes morfológicos, habilidades funcionales y habilidades de rendimiento. Imagen - [2].

2.8.1. ¿Como se desarrollan las características antropométricas en distintos escenarios?

Como ya se ha revisado en la sección 2.5 el peso afecta considerablemente a la hora de medir la potencia relativa, puesto que en competencias donde el ciclista debe enfrentar cuestas hacia arriba, tener un peso elevado juega en contra debido a principios físicos lógicos. A continuación, se hará énfasis en las distintas especialidades que enfrentan los ciclistas y su relación con características antropométricas.

1. **Ciclistas de TT (Time-trial) y terrenos planos:** En esta categoría se destacan por ser más altos y pesados, con una altura promedio de 180-185 cm y un peso de 70-75 kg, lo que resulta en un índice de masa corporal (IMC) cercano a 22 [Padilla *et al.*, 1999, Lucia *et al.*, 2000]. Esta constitución física favorece la aerodinámica y la potencia en terrenos planos y carreras contra el reloj.
2. **Escaladas:** En contraste, los especialistas en escalada tienden a ser más bajos y ligeros, con una altura promedio de 175-180 cm y un peso de 60-66 kg, presentando un IMC de 19-20 [Padilla *et al.*, 1999, Lucia *et al.*, 2000]. Esta morfología facilita el ascenso en pendientes pronunciadas al reducir la masa corporal que debe ser desplazada
3. **Ciclistas versátiles:** Aquellos capaces de destacar tanto en TT como en escalada, presentan características antropométricas intermedias, más cercanas a las de los especialistas en TT [Padilla *et al.*, 1999]. Esto sugiere que un equilibrio entre altura, peso y masa muscular puede ser beneficioso para un rendimiento óptimo en diferentes terrenos.

Como conclusión de esta sección, comprender estas diferencias y su relación con el rendimiento puede ser fundamental para optimizar el entrenamiento, la nutrición y la estrategia competitiva de cada ciclista. Es primordial destacar que, si bien las variables antropométricas son relevantes, no son el único factor determinante para el rendimiento (como se ha revisado a lo largo de esta memoria). Otros aspectos, como la ya revisada capacidad fisiológica, la técnica, la táctica y el equipamiento también juegan un papel crucial en el éxito de un ciclista. Además, si bien no se detallará sobre nutricionismo en el ciclismo en este estudio, existen distintas investigaciones donde se abordan planes específicos de alimentación para ciclistas como en [Fernández Rodríguez *et al.*, 2021], donde se lleva a cabo un estudio de un ciclista de 27 años que busca mejorar su rendimiento.

2.9. Clasificación de ciclistas

Dada la amplia cantidad de variables que influyen en el rendimiento del ciclista, como se ha revisado y profundizado anteriormente, es difícil establecer un perfil único que aborde todos los aspectos en conjunto. A pesar de esto, luego de

una extensa revisión en la literatura, se han encontrado significativas contribuciones al propósito de esta memoria. Por ejemplo, existen parámetros fisiológicos que han sido considerados características importantes para determinar el nivel de rendimiento de un ciclista, como la potencia máxima y los umbrales ventilatorio/láctico [Menaspà *et al.*, 2012, Menaspà *et al.*, 2010, Impellizzeri y Marcora, 2007, Louis *et al.*, 2013, Novak y Dascombe, 2014, Padilla *et al.*, 2001]. Si bien estas características son de suma importancia, **en el propósito de esta investigación y dada las limitaciones del dataset, se utilizarán solo la potencia (como capacidad aeróbica y anaeróbica), características antropométricas (edad, peso, género), frecuencia de entrenamiento y otros índices tales como frecuencia cardiaca, distancia y velocidad.**

Respecto a la clasificación, existen múltiples versiones, por ejemplo en [Jeukendrup *et al.*, 2000] se han definido perfiles de ciclistas carretera de élite, bien entrenados y de categoría mundial (figura 11).

Category	Trained cyclists	Well-trained	Élite	World Class
Training and race status				
Training frequency	2-3 times a week	3-7 times a week	5-8 times a week	5-8 times a week
Training duration	30-60 min	60-240 min	60-360 min	60-360 min
Training background	1 year	3-5 years	5-15 years	5-30 years
Race days per year	0-10	0-20	50-100	90-110
UCI ranking	-	-	first 2000	first 200
Physiological variables				
W_{max} (W)	250-400	300-450	350-500	400-600
W_{max} (W/kg)	4.0-5.0	5.0-6.0	6.0-7.0	6.5-8.0
VO_{2max} (L/min)	4.5-5.0	5.0-5.3	5.2-6.0	5.4-7.0
VO_{2max} (ml/kg/min)	64-70	70-75	72-80	75-90
Economy (W/L/min)	72-74	74-75	76-77	>78

Figura 11: Criterio para clasificación de ciclistas.

Fuente: [Jeukendrup *et al.*, 2000].

Notese como se define la frecuencia de entrenamiento para el ciclista en distintas categorías y así también, se define su experiencia, potencia máxima absoluta-relativa, economía y VO_{2max} . Esta tabla afirma la idea del enfoque al momento de clasificar los ciclistas, lo cual es un índice positivo.

Finalmente, existe una propuesta interesante hecha por [McKay *et al.*, 2022] (tabla 2) donde se establecen ciertos criterios a la hora de clasificar ciclistas basado en características mas generales, como por ejemplo, pautas mínimas propuestas por la organización mundial de la salud, índices de competencia, premios obtenidos, etc.

Tabla 2: Clasificación de deportistas según nivel de actividad física y rendimiento.

Fuente: [McKay *et al.*, 2022].

Descripción	Criterios
Sedentarios	No cumplen con las pautas mínimas de actividad física.
Recreacionalmente activos	Cumplen con las pautas mínimas de actividad física de la OMS.
Entrenados/En desarrollo	Participación a nivel local, entrenamiento regular, identificación con un deporte específico.
Altamente entrenados/Nivel nacional	Compiten a nivel nacional, entrenan de forma estructurada y periodizada.
Élite/Nivel internacional	Compiten a nivel internacional, se encuentran entre los mejores del mundo en sus disciplinas.
Clase mundial	Medallistas olímpicos o mundiales, poseedores de récords mundiales, atletas excepcionales.

2.10. Análisis de datos

El análisis de datos, en su esencia, es la transformación de información en conocimiento accionable [Tukey, 1962]. Este proceso sistemático, que implica la recopilación, organización, interpretación y visualización de datos, se ha convertido en una herramienta indispensable en diversos campos, desde la ciencia y la tecnología hasta la economía y las ciencias sociales [Shmueli, 2010].

En el contexto del ciclismo, el análisis de datos ha emergido como un recurso invaluable para optimizar el rendimiento y tomar decisiones informadas [Kohavi y Provost, 2001]. Al igual que en otros ámbitos, el análisis de datos en el ciclismo se basa en diversas técnicas, desde la estadística descriptiva que permite resumir y describir las características de un conjunto de datos, hasta el aprendizaje automático que posibilita la construcción de modelos predictivos y la identificación de patrones complejos [Breiman, 2001b].

La aplicación de estas técnicas en el ciclismo abarca una amplia gama de datos, algunos de ellos son métricas de rendimiento en la potencia, la velocidad y la frecuencia cardíaca. Así también variables ambientales como la temperatura, la humedad y la altitud [Provo *et al.*, 2003]. Al analizar estos datos, los ciclistas y entrenadores pueden identificar áreas de mejora, ajustar planes de entrenamiento, optimizar estrategias de carrera e incluso tomar decisiones sobre el equipo y la tecnología a utilizar.

Por ejemplo, mediante el uso de modelos de regresión, se puede evaluar cómo un cambio en el entrenamiento (variable independiente X) afecta el rendimiento en una carrera (variable dependiente Y) [James *et al.*, 2013]. Esto se puede representar matemáticamente como:

$$Y = a \times X + b + error() \tag{8}$$

Donde 'a' y 'b' son coeficientes que minimizan el error de predicción del modelo.

También existen distintos tipos de análisis, en la siguiente tabla (3) se resumen 5 de los más relevantes.

Tabla 3: Tipos de análisis de datos.
Fuente: [Zapier, 2023].

Tipo de análisis	Propósito
Análisis de texto	Extraer información de grandes cantidades de texto.
Análisis estadístico	Resumir y describir las características de un conjunto de datos.
Análisis de diagnóstico	Identificar las causas de un problema o situación.
Análisis predictivo	Predecir el futuro.
Análisis prescriptivo	Recomendar acciones o decisiones.

En resumen, el análisis de datos en el ciclismo, al igual que en otros campos, se ha convertido en una herramienta esencial para la toma de decisiones basadas en evidencia. Aporta un enorme conocimiento al caso de estudio de esta memoria, puesto que muchas técnicas de análisis de datos se utilizarán para la primera fase de la construcción del modelo. La figura 12 demuestra el proceso paso simplificado.

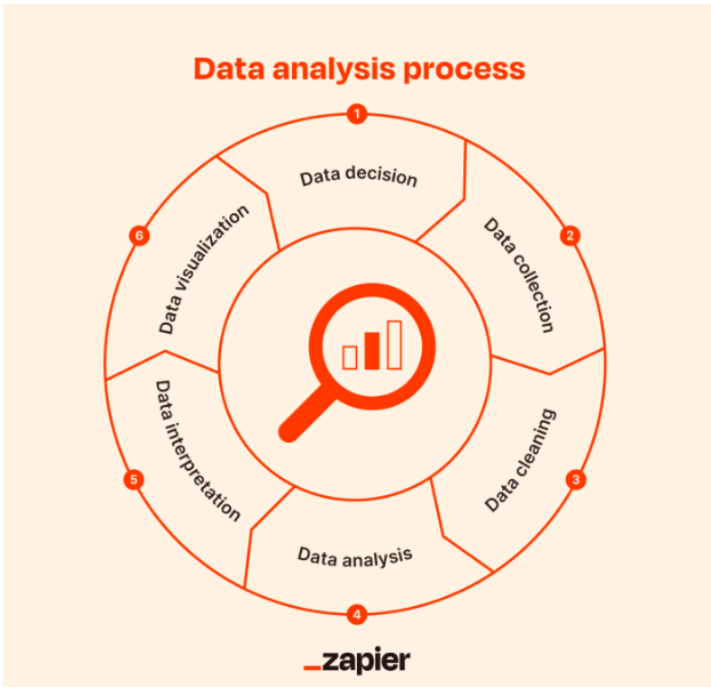


Figura 12: Etapas del análisis de datos.
Fuente: [Zapier, 2023].

2.11. Selección de características

La selección de características (feature selection) es un paso crucial en el aprendizaje automático que busca identificar y seleccionar el subconjunto más informativo de características de un conjunto de datos. El objetivo principal es mejorar el rendimiento del modelo, reducir su complejidad y aumentar su interpretabilidad. Aunque normalmente se hace en el pre-procesamiento, es importante profundizar en este paso dada su relevancia en el contexto de esta memoria de título.

En el contexto de la predicción del riesgo de enfermedades, la selección de características se utiliza para identificar los SNPs (variantes genéticas) más relevantes asociados con una enfermedad en particular. Esto ayuda a comprender los procesos biológicos subyacentes y a desarrollar modelos predictivos más precisos. Ahora, si aplicamos esto en el contexto de ciclistas, ayudaría enormemente a escoger el mejor intervalo de tiempos o el mejor indicador fisiológico para determinar la característica más relevante y así, entrenar un modelo de manera óptima. De hecho, se ha comparado que la selección de características «puede aumentar la eficiencia del aprendizaje, la precisión predictiva y reducir la complejidad de los resultados aprendidos» [Koller y Sahami, 1996, Kohavi y John, 1997, Hall, 2000]. A continuación se verá un ejemplo que demuestra como funciona el feature selection (figura 13).



Figura 13: Representación selección de características.

Fuente: [Pudjihartono *et al.*, 2022].

2.11.1. Método wrapper

El método wrapper (método envoltorio) es una categoría de técnicas de selección de características que evalúa subconjuntos de características basándose en el rendimiento de un algoritmo de aprendizaje específico. A diferencia de los métodos de filtro, que evalúan las características de forma independiente del modelo, los métodos wrapper utilizan el propio modelo para guiar el proceso de selección.

Los métodos wrapper utilizan el rendimiento del algoritmo clasificador elegido como métrica para ayudar a seleccionar el mejor subconjunto de características [Guyon y Elisseeff, 2003, Remeseiro y Bol'on-Canedo, 2019]. Esto significa que los métodos wrapper pueden identificar el conjunto de características que mejor se adapta al algoritmo de aprendizaje específico que se utilizará para construir el modelo predictivo.

Aunque los métodos wrapper pueden producir modelos con un rendimiento predictivo más alto, también tienen algunas desventajas. Son computacionalmente más intensivos que los métodos de filtro, especialmente cuando se trabaja con conjuntos de datos de alta dimensión como los datos genómicos. Además, los métodos wrapper son más propensos al overfitting, ya que optimizan la selección de características para un algoritmo específico, lo que puede no generalizarse bien a otros modelos o conjuntos de datos.

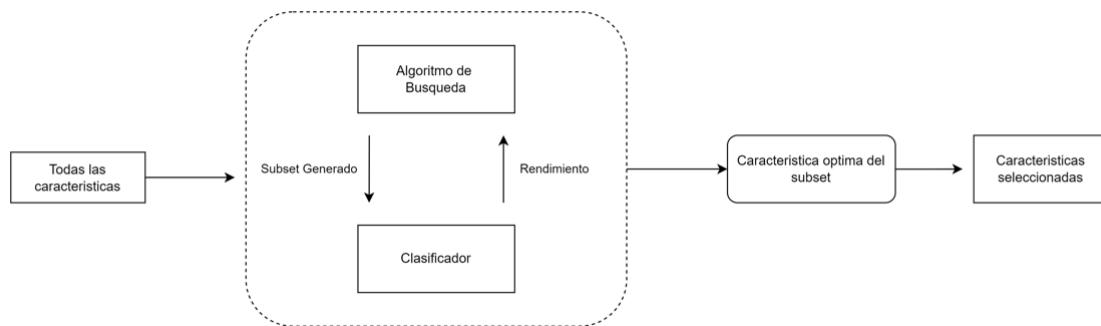


Figura 14: Adaptación de como funciona el método wrapper

Fuente: [Pudjihartono *et al.*, 2022]

En la figura 14, se ve como el proceso comienza con un conjunto completo de características y, a través de algoritmos de búsqueda, explora diferentes combinaciones de estas. Cada subconjunto generado se utiliza para entrenar un modelo de clasificación, cuyo rendimiento se evalúa. El objetivo es encontrar el subconjunto óptimo de características que permita al modelo generalizar mejor a nuevos datos. Por último, elimina características redundantes o irrelevantes, reduciendo así la complejidad del modelo y agilizando el proceso de entrenamiento.

Así también, existen otros métodos, que no se revisarán detalladamente pero en su lugar, se construyó una tabla (4) que los resume de mejor manera.

Métodos de filtro	Métodos envolventes	Métodos embebidos
Conjunto genérico de métodos que no incorporan un algoritmo de aprendizaje automático específico.	Evalúa un algoritmo de aprendizaje automático específico para encontrar las características óptimas.	Incorpora (fija) características durante el proceso de construcción del modelo. La selección de características se realiza en cada iteración de la fase de entrenamiento del modelo.
Mucho más rápido en comparación con los métodos envolventes en términos de complejidad temporal.	Tiempo de computación elevado para un conjunto de datos con muchas características.	Se sitúa entre los métodos de filtro y los métodos envolventes en términos de complejidad temporal.
Menos propenso al sobreajuste.	Alta probabilidad de sobreajuste debido a que implica el entrenamiento de modelos de aprendizaje automático con diferentes combinaciones de características.	Generalmente utilizado para reducir el sobreajuste penalizando los coeficientes de un modelo que son demasiado grandes.
Ejemplos: Correlación, prueba de Chi-cuadrado, ANOVA, ganancia de información, etc.	Ejemplos: Selección hacia adelante, eliminación hacia atrás, selección paso a paso, etc.	Ejemplos: LASSO, Elastic Net, Ridge Regression, etc.

Tabla 4: Métodos de selección de características. Fuente: Elaboración propia.

2.12. Aprendizaje automático

Como se mencionó en los objetivos [1.3.3] el propósito de esta memoria es crear un modelo capaz de clasificar el nivel de un ciclista en una escala de: casual, principiante, experimentado y élite. Para esto, no podemos aplicar directamente un modelo en específico, sin antes conocer aspectos relevantes de lo que es uno de los pilares fundamentales de la ciencia de datos, machine learning.

El machine learning (aprendizaje automático) es un campo de estudio de la inteligencia artificial (véase la imagen 15) que se centra en el desarrollo y análisis de algoritmos capaces de aprender de los datos y mejorar su rendimiento en una tarea específica a través de la experiencia, sin necesidad de instrucciones explícitas para cada posible escenario. Esto se logra mediante la identificación de patrones y relaciones en los datos, lo que permite a los sistemas realizar predicciones, tomar decisiones o generar nuevos conocimientos.

Se utiliza en diversos campos, como el procesamiento del lenguaje natural, la visión por

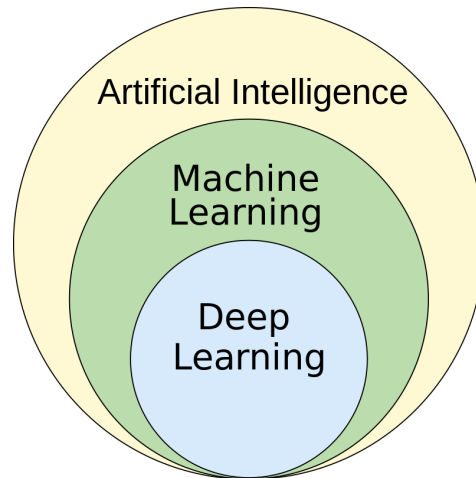


Figura 15: Machine Learning como sub-campo de IA
Fuente: [V *et al.*, 2020]

computadora, el reconocimiento de voz, el filtrado de correo electrónico, la agricultura, la medicina, los deportes, entre otros. En el ámbito empresarial, su capacidad predictiva lo convierte en una herramienta esencial para el análisis de datos y la toma de decisiones. Aunque no se limita exclusivamente a la estadística, esta disciplina, junto con otras áreas matemáticas como la optimización y el álgebra lineal, proporciona las bases para el desarrollo y la comprensión de los algoritmos de aprendizaje automático [Hu *et al.*, 2020, Yoosefzadeh-Najafabadi *et al.*, 2021]. La capacidad de construir modelos predictivos y sistemas de toma de decisiones basados en datos es lo que hace del machine learning una herramienta tan poderosa y transformadora en la actualidad.

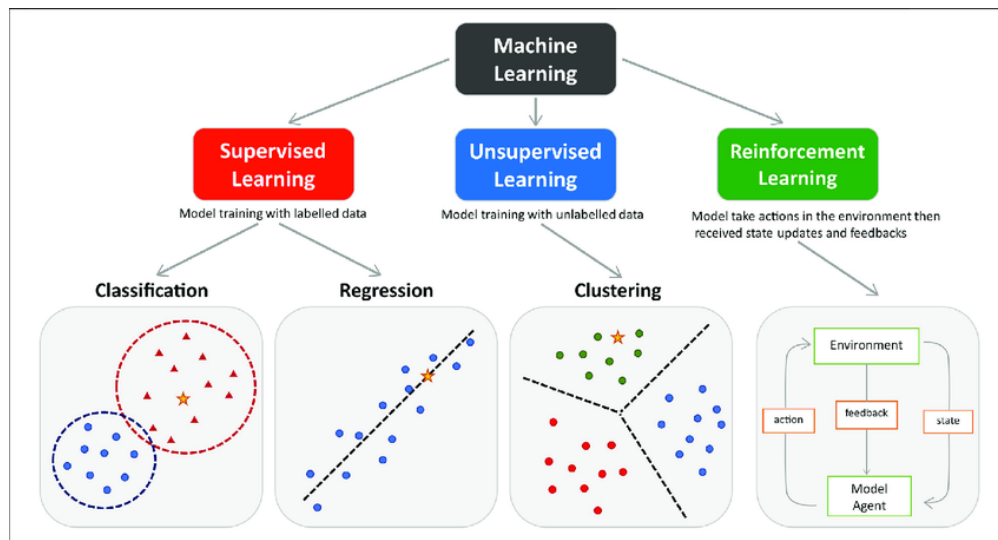


Figura 16: Categorías en Machine Learning.
Fuente: [V *et al.*, 2020].

Los enfoques de machine learning se clasifican en tres categorías principales (figura 16), ba-

sadas en el tipo de “señal” o “retroalimentación” que recibe el sistema de aprendizaje. Ahora, se hará énfasis solo en el “aprendizaje supervisado”, utilizado para la construcción del modelo.

2.12.1. Aprendizaje Supervisado

El aprendizaje supervisado es una técnica que se utiliza para deducir una función a partir de datos de entrenamiento. Estos datos de entrenamiento consisten en pares de objetos (generalmente vectores), donde una componente del par representa los datos de entrada y la otra, los resultados deseados. La salida de la función puede ser un valor numérico, como en los problemas de regresión, o una etiqueta de clase, como en los problemas de clasificación.

El objetivo del aprendizaje supervisado es crear una función que pueda predecir el valor correspondiente a cualquier objeto de entrada válida después de haber sido entrenada con una serie de ejemplos. Para lograr esto, la función debe ser capaz de generalizar a partir de los datos presentados a nuevas situaciones no vistas previamente.

El aprendizaje supervisado puede generar dos tipos de modelos, generalmente una función que transforma los datos de entrada en los resultados deseados. Para resolver un problema de aprendizaje supervisado, se deben seguir varios pasos:

1. **Determinar el tipo de ejemplos de entrenamiento:** Decidir qué tipo de datos se utilizarán para formar el modelo, como caracteres escritos a mano, palabras completas o líneas de escritura.
2. **Reunir un conjunto de entrenamiento:** Recopilar un conjunto de datos de entrada y sus salidas correspondientes, ya sea a partir de mediciones o mediante expertos.
3. **Representación de la función de entrada:** Transformar el objeto de entrada en un vector de características descriptivas del objeto. Este vector debe ser lo suficientemente grande para predecir con precisión la salida, pero no tan grande como para causar la maldición de la dimensionalidad.
4. **Determinar la estructura de la función y la técnica de aprendizaje:** Seleccionar una estructura adecuada, como una red neuronal artificial o un árbol de decisión, y la técnica de aprendizaje correspondiente.
5. **Completar el diseño:** Ejecutar el algoritmo de aprendizaje en el conjunto de entrenamiento y ajustar los parámetros mediante validación cruzada. Luego, medir el rendimiento del algoritmo con un conjunto de pruebas independiente.

En el ámbito del aprendizaje supervisado, también conocido como clasificación, existen diversos modelos, cada uno con sus fortalezas y debilidades. La elección del clasificador adecuado depende en gran medida de las características de los datos a clasificar, ya que no

existe un clasificador único que funcione mejor en todos los problemas (teorema del «No hay almuerzo gratis»⁹).

Los clasificadores más utilizados incluyen:

- Redes neuronales (como el perceptrón multicapa).
- Máquinas de vectores de soporte (SVM).
- Algoritmo de los K vecinos más cercanos (KNN).
- Modelos de mixturas.
- Clasificador bayesiano ingenuo.
- Árboles de decisión.
- Funciones de base radial.
- XGBoost.

Puesto que existen varios clasificadores, no se entrará en detalle en cada uno de ellos. Por otra parte, en 2.12.4 se definirá el modelo que será utilizado para el caso de estudio. A continuación, se muestra una imagen (figura 17) que ejemplifica bien las 3 categorías mencionadas, donde, en simples palabras, podemos ver como el supervisado es «*task driven*» donde se necesita enseñarle a la maquina que queremos, mientras que en el no-supervisado es todo lo contrario, la maquina se «guía por si sola» y en el último, aprende de sus errores.

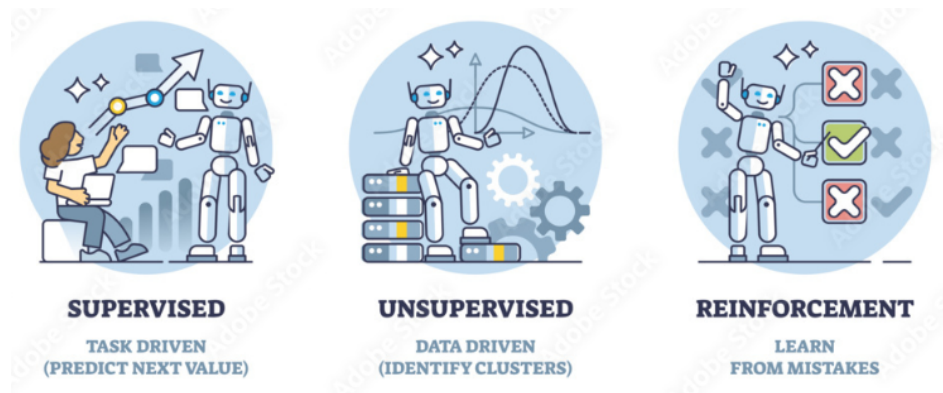


Figura 17: Ejemplo distintas categorías ML.

Fuente: [Adobe Stock, 2024].

⁹El teorema del “No hay almuerzo gratis” establece que, promediado en todos los posibles problemas, ningún algoritmo de aprendizaje automático tiene una ventaja intrínseca sobre otro.

2.12.2. Decision Tree - Árboles de decisiones

Los árboles de decisión son modelos de aprendizaje supervisado utilizados tanto para tareas de clasificación como de regresión. Representan una serie de decisiones secuenciales basadas en las características de los datos, organizadas en una estructura similar a un árbol (véase la figura 18). Cada nodo interno del árbol representa una prueba en un atributo, cada rama representa el resultado de la prueba y cada nodo hoja (o terminal) representa una etiqueta de clase (en clasificación) o un valor numérico (en regresión) [Breiman *et al.*, 1984].

2.12.3. Gradient Boosting

Gradient Boosting es una técnica específica de boosting que se basa en el uso de gradientes para optimizar la función de pérdida. En cada iteración, se calcula el gradiente de la función de pérdida con respecto a la predicción actual del modelo. Este gradiente indica la dirección en la que se debe ajustar el modelo para reducir el error. Luego, se entrena un nuevo modelo que se enfoca en predecir este gradiente, y se añade al conjunto con un factor de aprendizaje que controla su contribución.

La fórmula clave del Gradient Boosting es:

$$h_m(x) = -\frac{\partial L(y_i, F_{m-1}(x_i))}{\partial F_{m-1}(x_i)} \quad (9)$$

donde $h_m(x)$ es el nuevo modelo a entrenar en la iteración m , y el lado derecho de la ecuación representa el gradiente negativo de la función de pérdida L con respecto a la predicción actual $F_{m-1}(x_i)$.

Este enfoque permite que Gradient Boosting se adapte a diferentes tipos de funciones de pérdida y problemas de aprendizaje, ya sean de clasificación o regresión. XGBoost, como una implementación avanzada de Gradient Boosting, aprovecha esta flexibilidad y añade mejoras adicionales en términos de regularización y eficiencia computacional.

2.12.4. Extreme Gradient Boosting

XGBoost (Extreme Gradient Boosting) es un algoritmo de aprendizaje supervisado que se basa en la potenciación por gradiente, un método que combina múltiples modelos débiles (árboles de decisión en este caso) para crear uno más robusto y preciso. Además, XGBoost utiliza técnicas de regularización para evitar el sobreajuste (overfitting¹⁰) y mejorar la capa-

¹⁰El overfitting ocurre cuando un algoritmo se ajusta demasiado o incluso exactamente a sus datos de entrenamiento.

idad del modelo para generalizar a nuevos datos. Se ha destacado por su eficiencia, flexibilidad y precisión en tareas de clasificación y regresión. Desarrollado por Tianqi Chen y colaboradores, XGBoost ha ganado popularidad en competencias de ciencia de datos, como las organizadas por Kaggle, debido a su rendimiento superior y capacidad para manejar grandes conjuntos de datos [Chen y Guestrin, 2016].

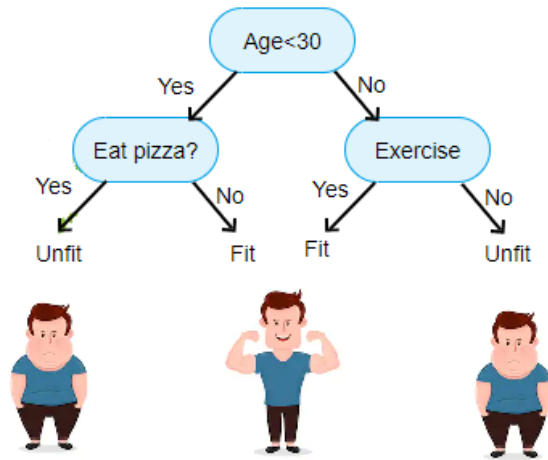


Figura 18: Ejemplo de árboles de decisiones
Fuente: [Kholkin et al., 2020]

Este algoritmo se ha consolidado como una herramienta líder en machine learning, gracias a su rendimiento superior en tareas de clasificación y regresión. Su éxito radica en una combinación de características que optimizan tanto la precisión como la eficiencia computacional.

2.12.5. Características del modelo

XGBoost se distingue de otros algoritmos de potenciación de gradiente por:

- **Regularización Mejorada:** Emplea técnicas sofisticadas para penalizar la complejidad del modelo, evitando el sobreajuste y mejorando la generalización.
- **Poda Estratégica:** Reduce el tamaño de los árboles de decisión de manera inteligente, eliminando ramas innecesarias y mejorando la eficiencia.
- **Optimización de Segundo Orden:** Utiliza información de segundo orden (hessianos) para acelerar la convergencia y encontrar soluciones más precisas.
- **Aleatoriedad Controlada:** Introduce elementos aleatorios en el proceso de construcción de árboles, mejorando la diversidad y evitando mínimos locales.
- **Escalabilidad:** Puede ejecutarse tanto en sistemas individuales como distribuidos, lo que permite abordar problemas a gran escala.

- **Selección Automática de Características:** Identifica automáticamente las características más relevantes, simplificando el proceso de modelado.

En esencia, XGBoost construye un modelo aditivo, donde cada término es un árbol de decisión. Estos árboles se entrenan de forma secuencial, y cada uno se enfoca en corregir los errores del modelo anterior. Para optimizar este proceso, XGBoost utiliza una aproximación de Taylor de segundo orden en la función de pérdida, lo que le permite aprovechar el método de Newton-Raphson para una convergencia más rápida.

Un algoritmo XGBoost genérico, sin regularización, puede describirse de la siguiente manera:

Entradas:

- Conjunto de entrenamiento: $\{(x_i, y_i)\}_{i=1}^N$
- Función de pérdida diferenciable: $L(y, F(x))$
- Número de iteraciones (árboles): M
- Tasa de aprendizaje: α

Algoritmo:

1. Inicializar: $f^{(0)}(x) = \arg \min_{\gamma} \sum_{i=1}^N L(y_i, \gamma)$

2. Para $m = 1$ hasta M :

a) Calcular gradientes y hessianos:

$$g_m(x_i) = \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)} \right]_{f(x)=f^{(m-1)}(x)}$$
$$h_m(x_i) = \left[\frac{\partial^2 L(y_i, f(x_i))}{\partial f(x_i)^2} \right]_{f(x)=f^{(m-1)}(x)}$$

b) Ajustar un árbol $f_m(x)$ a los datos $\{(x_i, g_m(x_i)/h_m(x_i))\}_{i=1}^N$

c) Actualizar el modelo: $f^{(m)}(x) = f^{(m-1)}(x) + \alpha f_m(x)$

3. Salida: $f(x) = f^{(M)}(x) = \sum_{m=0}^M f_m(x)$

En la práctica, XGBoost incorpora términos de regularización para controlar la complejidad del modelo y evitar el sobreajuste.

2.12.6. Aprendizaje por conjuntos - Ensemble Learning

El aprendizaje por conjuntos, o ensemble learning, es una poderosa estrategia en el machine learning que se basa en la combinación de múltiples modelos de aprendizaje para obtener un rendimiento predictivo superior al que se lograría con un solo modelo. A diferencia de un enfoque que busca el “mejor” modelo individual, el ensemble learning aprovecha la diversidad de un conjunto de modelos, que pueden ser del mismo tipo (por ejemplo, múltiples árboles de decisión) o de diferentes tipos (por ejemplo, combinar árboles de decisión con redes neuronales).

La intuición detrás de esta estrategia es que cada modelo individual puede tener fortalezas y debilidades en diferentes aspectos del problema. Al combinar sus predicciones, se busca que los errores de un modelo se compensen con los aciertos de otros, resultando en un modelo global más robusto y preciso.

La diversidad entre los modelos es clave para el éxito del ensemble learning. Esta diversidad puede lograrse de varias maneras, como utilizar diferentes algoritmos de aprendizaje, entrenar los modelos con subconjuntos distintos de los datos, o introducir aleatoriedad en el proceso de entrenamiento.

En el contexto de XGBoost, el ensemble learning se manifiesta en la combinación de múltiples árboles de decisión. Cada árbol se entrena de manera secuencial, enfocándose en corregir los errores de los árboles anteriores. Esta estrategia, conocida como boosting, permite que el conjunto de árboles aprenda de forma iterativa y mejore su capacidad predictiva.

2.13. Evaluación del rendimiento en modelos de machine learning

La etapa de evaluación y validación es crucial para garantizar la robustez y fiabilidad de cualquier modelo de aprendizaje automático, especialmente en el contexto de la clasificación de ciclistas, donde la precisión y la interpretabilidad de las predicciones son fundamentales para la toma de decisiones informadas.

2.13.1. Métricas de evaluación

En la clasificación de ciclistas, donde a menudo nos encontramos con clases desbalanceadas (por ejemplo, más ciclistas aficionados que profesionales), el F1-score se destaca como una métrica de evaluación clave [Powers, 2020]. Este equilibra la precisión (la proporción de predicciones positivas correctas) y el recall (la proporción de verdaderos positivos identificados correctamente) proporcionando una medida más completa del rendimiento del modelo que la precisión sola.

La matriz de confusión complementa el F1-score al ofrecer una visión detallada de los aciertos y errores del modelo, distinguiendo entre verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos [Tharwat, 2020]. Esta información es valiosa para identificar los tipos de errores que el modelo comete con mayor frecuencia y ajustar su entrenamiento en consecuencia.

Si bien el F1-score y la matriz de confusión son esenciales, otras métricas como la precisión, el recall y la exactitud pueden ser útiles en contextos específicos. Por ejemplo, la precisión podría ser más relevante si el costo de los falsos positivos es alto (por ejemplo, clasificar erróneamente a un ciclista aficionado como profesional), mientras que el recall podría ser más importante si el costo de los falsos negativos es alto (por ejemplo, no identificar a un ciclista con alto potencial).

Por otro lado, en el contexto de los árboles de decisión, métricas como Cover, Gain y Weight se utilizan para evaluar la importancia de las características, estas se definen como:

- **Cover:** Indica la proporción de datos que una característica afecta en el árbol de decisión. Un valor alto de cover sugiere que una característica se utiliza en muchas divisiones del árbol, influyendo en una gran parte de las predicciones.
- **Gain:** Mide la reducción en la impureza (por ejemplo, la entropía) que se logra al dividir los datos utilizando una característica en particular. Un valor alto de gain indica que una característica proporciona una división informativa que mejora significativamente la capacidad del modelo para separar las clases.
- **Weight:** Representa la importancia relativa de una característica en un árbol de decisión. Generalmente, se calcula como la frecuencia con la que una característica se utiliza como punto de división para crear nuevas ramas. Un weight alto indica que la característica es muy relevante para las predicciones del modelo. Cuando se habla de feature importance o f1-score sin especificar, se suele referir al weight.

2.13.2. Validación cruzada

La validación cruzada es una técnica esencial para estimar el rendimiento de un modelo en datos nuevos y evitar el sobre-ajuste [Arlot y Celisse, 2010]. Consiste en dividir el conjunto de datos en k subconjuntos (folds) y entrenar el modelo k veces, utilizando cada vez un fold diferente como conjunto de prueba y los restantes como conjunto de entrenamiento.

2.13.3. Curva ROC y Área Bajo la Curva (AUC-ROC)

La curva ROC (Receiver Operating Characteristic) es una herramienta gráfica que permite visualizar el rendimiento de un clasificador binario en función de su umbral de decisión

[Fawcett, 2006]. El área bajo la curva ROC (AUC-ROC) proporciona una medida única del rendimiento general del modelo, independientemente del umbral de decisión elegido.

El AUC-ROC varía entre 0.5 (clasificador aleatorio) y 1 (clasificador perfecto), y cuanto mayor sea su valor, mejor será la capacidad del modelo para distinguir entre las clases. En el contexto de la clasificación de ciclistas, un AUC-ROC alto indica que el modelo es capaz de diferenciar con precisión entre ciclistas de diferentes niveles.

2.13.4. SHapley Additive exPlanations

Los valores SHAP (SHapley Additive exPlanations) son un método que permite interpretar las predicciones de modelos de aprendizaje automático, asignando valores de importancia a cada característica para cada instancia individual [Lundberg y Lee, 2017]. Estos valores SHAP proporcionan una visión detallada de cómo cada característica contribuye a la predicción final, lo que es especialmente útil para comprender las decisiones del modelo en el contexto de la clasificación de ciclistas.

En este estudio, se utilizarán los valores SHAP para analizar las predicciones del modelo XG-Boost. Esto permitirá identificar las características más importantes para la clasificación de ciclistas y comprender cómo estas características influyen en el rendimiento.

2.14. Metodología

Para abordar el desafío de clasificar el rendimiento de los ciclistas, se adopta la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) [Chapman *et al.*, 2000]. Esta metodología, ampliamente utilizada en proyectos de ciencia de datos, proporciona una estructura clara y sistemática para guiar el proceso desde la comprensión del problema hasta la implementación y el mantenimiento del modelo (figura 19). Bajo el contexto de este estudio, el CRISP-DM (adaptado) se compone de seis fases iterativas e interconectadas:

1. **Comprensión del negocio:** En esta fase, se define el problema de clasificación de ciclistas, se identifican los objetivos del proyecto y se establecen los criterios de éxito. El objetivo, tal como se define en 1.3.3 es clasificar a los ciclistas en diferentes niveles de rendimiento (casual, principiante, experimentado y profesional) con el fin de aportar conocimiento y propuesta de mejoras al ciclista o entrenador.
2. **Comprensión de los datos:** Se recopilan y exploran los datos relevantes para el problema. En el caso de la clasificación de ciclistas, incluye datos antropométricos (edad, peso, género) y datos de rendimiento (potencia, velocidad, cadencia, critical power). Se analizan las características de los datos, se identifican valores atípicos y se evalúa la calidad de los datos.

3. **Preparación de los datos:** Los datos se limpian, transforman y preparan para el modelo. Esto incluye la imputación de valores faltantes, la normalización de variables, la creación de nuevas características y la división del conjunto de datos en entrenamiento, validación y prueba.
4. **Modelado:** Se seleccionan y aplican algoritmos de aprendizaje automático adecuados para el problema de clasificación. En este caso, se utiliza XGBoost, Random Forest y KNN, dada su precisión y excelentes resultados en clasificación [Chen y Guestrin, 2016, Shwartz-Ziv y Armon, 2023, Breiman, 2001a, Fernández-Delgado *et al.*, 2014, Zhang *et al.*, 2017]. Se entrenan los modelos con los datos de entrenamiento y se ajustan los hiper-parámetros para optimizar el rendimiento.
5. **Evaluación:** Se evalúa el rendimiento de los modelos utilizando métricas como el F1-score, la matriz de confusión y la curva ROC. Se comparan los resultados de diferentes modelos y se selecciona el que mejor se adapte a los criterios de éxito establecidos en la fase de comprensión del negocio. También se aplica validación cruzada y SHAP.
6. **Implementación:** El modelo seleccionado se implementa en un entorno de producción y se monitorea su rendimiento a lo largo del tiempo. Se pueden realizar ajustes y actualizaciones según sea necesario para garantizar que el modelo siga siendo relevante y preciso.

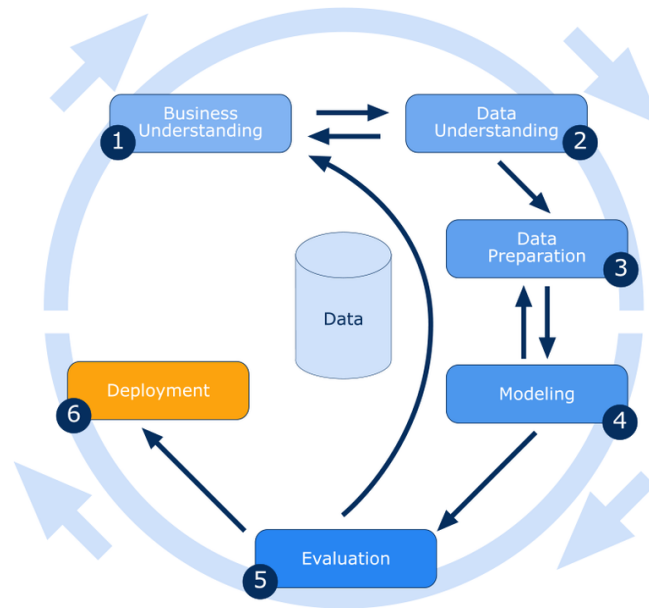


Figura 19: Representación metodología CRISP-DM.

Fuente: [Roy, 2020].

CAPÍTULO 3

PROPUESTA DE SOLUCIÓN

Para abordar el caso de estudio, proponemos una solución basada en la metodología CRISP-DM, que abarca desde la obtención de los datos hasta la construcción del modelo.

En lugar de Strava o Proccyclingstats, utilizaremos el proyecto de código abierto GoldenCheetah como fuente de datos. Esta elección garantiza una mayor diversidad, incluyendo ciclistas de diferentes niveles, lo que permitirá una clasificación más precisa y evita el sesgo hacia ciclistas profesionales. Además, GoldenCheetah ofrece datasets pre-procesados y simplificados, facilitando el análisis y reduciendo la complejidad computacional en comparación con la fuente original de 102GB.

En las siguientes secciones se hará énfasis en la descripción de los datos, el análisis exploratorio, el pre-procesamiento que se realizó y las técnicas que se emplearon para obtener insights¹¹.

3.1. Fuente de datos

Tal como se mencionó en 1.1, se escogió GoldenCheetah Open-Data [Liversedge, 2021] como fuente de datos. En la siguiente figura 20 se muestra como se dividieron los datasets¹².

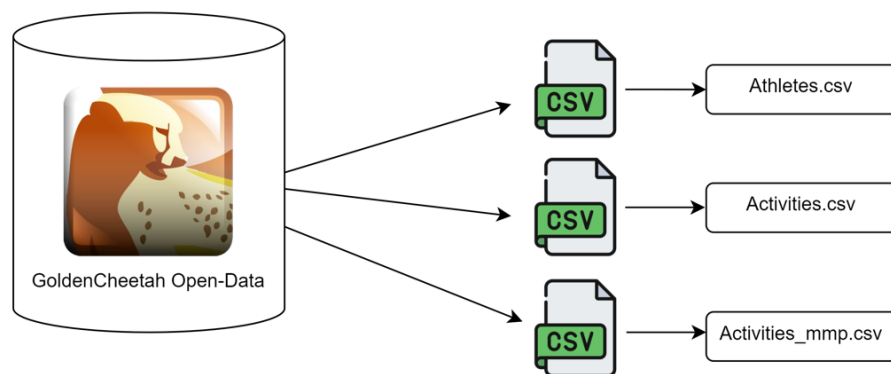


Figura 20: Output GoldenCheetah-OpenData.
Fuente: Elaboración propia.

¹¹Información y resultados que aporten al caso de estudio.

¹²Conjunto de datos.

A primera vista, los tres conjuntos de datos parecen contener información sobre atletas, actividades y métricas de rendimiento. Para confirmar esto y comprender mejor su contenido, se cargaron en Google Colab, aprovechando su capacidad de procesamiento en la nube.

Un análisis inicial de las dimensiones de cada conjunto de datos reveló lo siguiente:

- **Athletes.csv:** 1400 filas y 36 columnas.
- **Activities.csv:** 2,398,134 filas y 34 columnas.
- **Activities_mmp.csv:** 1,341,824 filas y 105 columnas.

Con los datos cargados y una comprensión inicial de su estructura, se procede a realizar un análisis exploratorio de datos (EDA) para profundizar en su contenido y relaciones.

3.2. Análisis exploratorio - Parte I

El análisis exploratorio es una parte importante a la hora de comenzar la asimilación de los datos. Anteriormente, se utilizó un «df.shape» para conocer el espacio muestral de los datasets, sin embargo, es necesario detallar más respecto a cada uno, utilizando funciones como «df.describe, df.columns (figura 21) o df.head».

Tabla 5: Descripción de características

Característica	Descripción
id	Identificador único para cada atleta o registro.
age	Edad del atleta.
gender	Género del atleta.
activities, bike, run, swim, other	Activities es el recuento total, bike, run, swim y other son la cantidad únicas de actividades en cada una de las disciplinas.
weightkg, weightstd	Peso del atleta en kilogramos y su desviación estándar.
nTiempo_critical_power	Potencia crítica sostenida durante una duración determinada. (n)
nTiempo_peak_wpk	Pico de potencia relativa en un tiempo determinado.
nTiempo_peak	Pico de potencia absoluta en un tiempo determinado.

En athletes.csv, tenemos columnas (tabla 5) que han sido definidas como importantes a la hora de cuantificar el rendimiento de un ciclista, tal como se detalló en el marco teórico (2.1).

```
Index(['id', 'age', 'gender', 'activities', 'bike', 'run', 'swim', 'other',
      '1s_critical_power', '15s_critical_power', '2m_critical_power',
      '3m_critical_power', '5m_critical_power', '8m_critical_power',
      '10m_critical_power', '20m_critical_power', '30m_critical_power',
      '1m_peak_wpk', '5m_peak_wpk', '10m_peak_wpk', '20m_peak_wpk',
      '30m_peak_wpk', '20s_peak', '20s_peak_wpk', '60s_peak', '60s_peak_wpk',
      '180s_peak', '180s_peak_wpk', '240s_peak', '240s_peak_wpk', '420s_peak',
      '420s_peak_wpk', '720s_peak', '720s_peak_wpk', 'weightkg',
      'weightstd'],
      dtype='object')
```

Figura 21: Columnas atletas dataset.
Fuente: Elaboración propia.

	id	age	gender	activities	bike	run	swim	other	1s_critical_power	15s_critical_power	...	180s_peak	180s_peak_wpk	240s_peak	240s_peak_wpk	420s_peak	420s_peak_wpk
1	9bdd724b-4148-4bee-9344-9e7a526c71c1	36	M	21	21	0	0	0	681.00	500.13333	...	294.0	4.388060	289.0	4.313433	270.0	4.240000
2	f43673fa-600b-4b0b-8c3a-e1627347ebe1	46	M	227	3	78	14	132	599.00	562.60000	...	259.0	3.741679	260.0	3.714551	252.0	3.687500
3	bcf7da24-5c1b-4793-85c5-306f00878d2a	18	M	207	203	0	0	4	1339.00	817.53333	...	373.0	4.782051	355.0	4.551282	338.0	4.320000
4	18f038c9-9eef-47c7-8bcc-69287a903a1e	31	M	62	0	0	0	62	994.00	843.86667	...	397.0	4.841463	392.0	4.780488	384.0	4.700000

Figura 22: Explorando datos de atletas.
Fuente: Elaboración propia.

En la figura 22, se confirma la hipótesis inicial, sin embargo, es necesario profundizar más. Como por ejemplo, en la columna «activities» suma la cantidad total de los distintos deportes realizados por el ciclista, clasificados en «bike», «run», «swim» y «other». Así también, se destaca la amplitud de variables numéricas por sobre las categóricas. Existe una variable «id» que representa el atleta, pero anonimizado. Ahora se procede a explorar los otros datasets.

```
[12] df_act.tail(3)
```

	id	file	date	age	gender	sport	data	1s_critical_power	15s_critical_power	2m_critical_power	...	total_work	average_speed	...
2398131	310b9b37-1c08-4c7c-9db3-e7853f3db400	2019_04_18_10_02_50.csv	2019/04/18 14:02:50 UTC	47	M	Bike	TDSPHC-AGL-E---	689.0	309.80000	251.73333	...	1238.880	27.88145	...
2398132	310b9b37-1c08-4c7c-9db3-e7853f3db400	2019_04_19_11_15_47.csv	2019/04/19 09:15:47 UTC	47	M	Bike	TDSPHC-AGL-E---	269.0	198.86667	135.27500	...	333.026	20.98630	...
2398133	310b9b37-1c08-4c7c-9db3-e7853f3db400	2019_04_20_17_19_39.csv	2019/04/20 15:19:39 UTC	47	M	Bike	TDSPHC-AGL-E---	677.0	422.20000	228.26667	...	1586.688	29.39440	...

3 rows x 34 columns

Figura 23: Explorando datos de actividades.
Fuente: Elaboración propia.

En la figura 23, se tuvo que realizar un `df.tail(3)` en vez de un `df.head(n)` ya que las primeras columnas de `activities.csv` tenían muchos nulos. Esto es un punto negativo para este dataset. Además, como el anterior, tiene muchas columnas, por eso las imágenes han sido recortadas con el fin de mostrar correctamente la composición de los datasets.

Algo bueno que destacar de `activities.csv` es que tiene más variables útiles que `athletes.csv`, como por ejemplo, en este existen índices de variabilidad y de percepción del esfuerzo. De

igual forma, se registra la distancia, velocidad promedio, cadencia, elevación y fecha. Sin embargo, dada la enorme cantidad de nulos y baja calidad de datos (por ejemplo, en sport existen nombres aleatorios con caracteres difíciles de filtrar) este dataset es potencial a ser descartado para construir el modelo en si, pero se conserva como candidato para evaluar y validar el modelo una vez funcional.

Otro aspecto sumamente importante que se descubrió durante el EDA, es que si se hace un filtro por el 'id' en los distintos dataset, se pueden sacar varias conclusiones iniciales respecto a la composición de cada uno. Luego de filtrar, se obtiene lo siguiente en las figuras 24 y 25.

	id	date	weight	duration	distance	average_power	1	2	3	4
107152	000c6417-e1e4-497e-89e6-bb21e17ec355	2017/01/31 17:36:44 UTC	71.0	3948.0	0.00000	196.76570	783	57.0	737.0	724.0
107153	000c6417-e1e4-497e-89e6-bb21e17ec355	2017/02/01 13:48:14 UTC	71.0	12753.0	84.50926	193.48016	1016	110.0	978.0	928.0

2 rows x 105 columns

Figura 24: Actividades mmp.
Fuente: Elaboración propia.

	id	file	date	age	gender	sport	data	1s_critical_power	15s_critical_power	2m_critical_power	...
17863	000c6417-e1e4-497e-89e6-bb21e17ec355	2017_01_31_18_36_44.csv	2017/01/31 17:36:44 UTC	22	M	Bike	T-PHC-A-E---	783.0	384.46667	233.70833	...
17863	000c6417-e1e4-497e-89e6-bb21e17ec355	2017_02_01_14_48_14.csv	2017/02/01 13:48:14 UTC	22	M	Bike	TDSPHC-AGL-E---	1016.0	586.93333	377.90000	...

2 rows x 24 columns

Figura 25: Actividades.
Fuente: Elaboración propia.

Notese como en el «activities_mmp», la columna 1 coincide con el valor de «1s_critical_power». Esto ocurre dado que en un dataset, como lo es actividades mmp, se registra la potencia en cada intervalo de tiempo específico (va de 1 segundo hasta 7200) y en «activities», se utilizan todos los registros del «activities_mmp» ya sea para la potencia relativa, potencia promedio, o en este caso, para el critical power en 1 segundo. Como es el puntapié inicial, no varía ni decrece. Ahora, realizando el filtro en el dataset de los atletas:

	id	age	gender	activities	bike	run	swim	other	1s_critical_power	15s_critical_power	...	180s_peak
86	000c6417-e1e4-497e-89e6-bb21e17ec355	20	M	399	394	4	0	1	1978.0	1045.8	...	479.0

1 rows x 36 columns

Figura 26: Actividades.
Fuente: Elaboración propia.

En la figura 26 los datos parecieran ser completamente distintos, pero no es así, ya que están directamente relacionado con los que se mostraron anteriormente. En el dataset de «activities_mmp» se calcula la potencia desde 1 segundo hasta 7200, pero también incluye otras

variables como el calculo de potencia promedio, duración, distancia, id y peso. En el de «activities» se registra cada actividad completa por si sola, por ejemplo, si el atleta realizó una actividad el 03/11/2018, esa actividad por si sola se va a «activities», sin embargo, el cálculo de todas las métricas que requieren potencia, se almacenan en «activities_mmp» y por último, el «athletes» simplifica y calcula los promedios de todas las actividades. La siguiente figura (27) ejemplifica de mejor manera lo que ocurre con estos 3 conjuntos de datos:

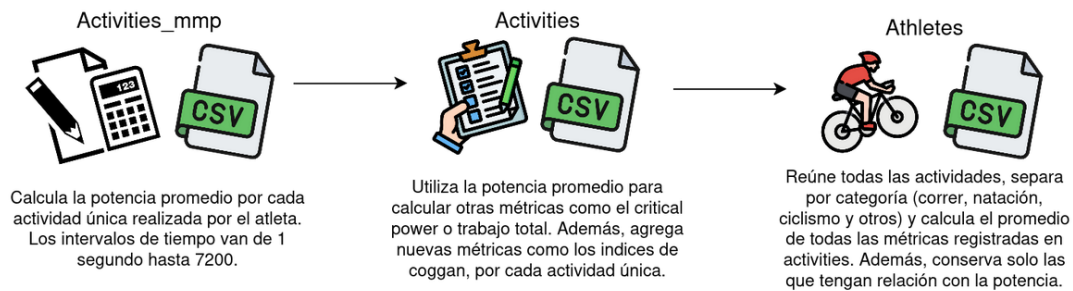


Figura 27: Flujo de trabajo 3 datasets.

Fuente: Elaboración propia.

Puesto que el athletes.csv almacena el promedio de todas las actividades, resulta un excelente candidato para desarrollar el modelo. Además, tiene una baja cantidad de nulos (21 en weightkg y 25 en weightstd) a diferencia de los otros.

Luego de esto, se realizó un filtro por atletas que no tuvieran actividades ni critical power registrado. El por qué de esto, se justifica con el simple hecho de que atletas que no registren actividades ni critical power, no aportan al propósito de este estudio (de la misma forma, los nulos). La cantidad de atletas que que representan dicha cantidad es 92 (6.57 %).

Además, se construye una tabla con las desviaciones estándar, medias y valores estadísticos que arroja un df.describe (tabla 6), con el fin de ejemplificar parte del espacio muestral en el dataset de atletas.

La edad, tiene como mínimo -965 y como máximo 2018. Si nos vamos a la terminología, estos serían considerados valores atípicos u outliers. Se hizo un count para corroborar cuantos valores únicos se encontraban fuera de un rango normal y establecido por competencias [AusCycling, 2024] y lo que arrojó fueron 48. Vale decir, que 48 valores únicos de los 110 en la columna edad, pertenecen a valores atípicos. Haciendo el conteo por el rango de edades (13-90) solo son 169 atletas, lo que equivale a un 12.1 % del total. Si lo vemos de otra forma, es decir, a cuantas actividades estaríamos perdiendo con esos valores, equivale a: 114.480. Esto último demuestra que si elimináramos valores atípicos, estaríamos perdiendo una cantidad importante de actividades. Puesto que las actividades contribuyen netamente al entrenamiento del modelo, se debe considerar esto a la hora de la limpieza y normalización. La figura 28 muestra los valores únicos de la columna edad.

Tabla 6: Resumen estadístico del dataset de atletas (n = 1400).
Fuente: Elaboración propia.

Variable	Media $\pm$ STD	Mínimo	Máximo
Edad (años)	34.80 $\pm$ 114.32	-965.00	2018.00
Actividades	525.55 $\pm$ 690.01	0.00	10476.00
Ciclismo	342.82 $\pm$ 464.07	0.00	4680.00
Carrera	51.80 $\pm$ 145.99	0.00	1375.00
Natación	12.98 $\pm$ 63.34	0.00	1236.00
Otras	117.95 $\pm$ 386.15	0.00	10474.00
1s CP	1184.76 $\pm$ 645.56	0.00	2999.00
15s CP	769.09 $\pm$ 389.27	0.00	2982.80
2m CP	394.56 $\pm$ 213.88	0.00	2897.03
3m CP	362.81 $\pm$ 202.49	0.00	2851.04
180s Peak	363.82 $\pm$ 335.43	0.00	8764.00
180s Pk wpk	78.80 $\pm$ 1197.44	0.00	24500.00
240s Peak	344.41 $\pm$ 325.65	0.00	8659.00
240s Pk wpk	75.05 $\pm$ 1137.33	0.00	22050.00
Peso (kg)	74.28 $\pm$ 10.57	45.59	152.91
Peso (STD)	2.71 $\pm$ 64.95	0.00	2399.14

En resumen:

- Se debe hacer énfasis en la columna de edad debido a sus valores atípicos.
- Los atletas con 1s_critical_power en 0 no contribuyen al caso de estudio, así también los que no registran actividades.
- Existe una buena calidad de datos, en comparación con los otros datasets, lo cual da un visto bueno para comenzar a trabajar en la construcción del modelo.

```
110
[ 37  36  46  18  31  47  55  26  42  43  38  61  58 -1
 -39  45  50  86  44  33  52  39  62  25  68  57  23  49
 34  29  64  41  51  67 -53  48 -44 2018 35 -50 30 22
 40 -69 20 54 63 -58 -67 -54 21 -49 -72 32 -51 77
 53 -73 27 16 66 -62 24 -48 -63 -66 65 60 -60 28
 69 14 -56 -47 59 -42 -55 -79 -64 71 19 -71 -40 -59
 -52 -33 56 -80 -43 -57 -70 -65 -68 -61 -45 4 125 -75
 -36 17 15 -46 0 -965 72 -41 2 -26 78 13]
```

Figura 28: Valores únicos de las edades.
Fuente: Elaboración propia.

3.3. Análisis exploratorio - Parte II

Luego de la primera parte del EDA, donde ya se abordaron distintos aspectos del dataset que son sumamente importante para comprenderlo en su totalidad antes de si quiera manipular o construir el modelo de clasificación, existen más aspectos relevantes.

Por ejemplo, en ultima instancia se mencionó que la edad contiene distintos valores atípicos, ahora, si bien como se estableció, podrían ser eliminados, a continuación, se comparan las diferencias.

Luego de hacer un filtro tanto en las edades positivas, como las negativas, que tenían más actividades (top 10), se obtuvo:

Los 10 atletas con más actividades en el rango de edad de -10 a -90 son:

	id	age	activities
754	7cd74375-d7ce-483f-98c8-10d34764757f	-45	5451
1167	b02ffb08-ec17-4e02-878e-71f03d87e763	-26	3737
648	6454a948-9715-49bd-8db8-2c7d5fa20b3f	-71	3135
106	bc3dd711-a083-4f7f-8943-d5924858bba6	-67	3021
1055	72hhah9a-d866-4bca-a69b-89a93484d305	-50	2484
408	38ec277a-e504-483c-9d59-6af0ff6bccab	-50	2484
113	004578c5-0911-4760-d2f2-d1/cad0000051	-54	2299

Figura 29: Filtro top 10 edades negativas con más actividades.

Fuente: Elaboración propia.

Los 10 atletas con más actividades en el rango de edad de 10 a 90 son:

	id	age	activities
1339	bb697b5c-8906-4343-ad73-288e50fc4390	64	10476
1290	2bf4beac-396c-4b3b-9c59-556a8754491a	40	5582
1239	7905b2a2-427a-482a-b3e0-f1f46c999755	53	4811
405	347c8f92-0469-413a-903a-9c05fa96e52d	53	4807
030	4177a04-cc30-4d0a-b0bf-eb4a7a0ad7c8	46	4001
251	3f42c575-a7ff-4b5c-99dc-5d49de8451b2	50	3958
200	0c737210-0103-4002-9030-03c47f007c11	00	3097
373	14af4992-a4f3-45f7-9e24-de753af98c72	49	3712

Figura 30: Filtro top 10 edades positivas con más actividades.

Fuente: Elaboración propia.

Estos dos IDs (figura 29 y 30) tienen algo peculiar, y es que, ambos deberían ser 50 años. Sin embargo, como los números negativos parecieran simplemente ser la contra-parte pero mal tipeada, se debe justificar esto. Mediante el siguiente gráfico de líneas (figura 31), es fácil contemplar que las diferencias no son tan grandes. A pesar de que el atleta de 50 años comparado con el negativo tengan una diferencia de casi 2.000 actividades, nos sirve para explorar y demostrar que los negativos o datos mal tipeados no necesariamente son datos inservibles.

Analizando otras variables, como por ejemplo, el género, nos encontramos con que los hombres predominan en cantidad respecto a este dataset (1364 hombres y 36 mujeres, véase la figura 32). También es importante detallar que las variables son categóricas, pero se encuentran distribuidas en M para hombre y F para mujer, lo que podría una variable potencial para aplicar one-hot-encoding o la función .cat para asignarlos como 0 y 1 respectivamente. Otro aspecto importante, es que los hombres suman un total de 723.954 actividades y las mujeres 11.817. La figura 33 compara las potencias promedio entre hombres y mujeres.

La actividad «bike» se alza como favorita con 479.948 de actividades en total (para hombres y mujeres) lo cual reafirma nuevamente la positividad de este dataset, puesto que al tener más datos de actividades de ciclistas, ayuda a construir un modelo más preciso en lo que se busca trabajar. Véase la figura 34.

Para evitar la multicolinealidad y, de igual manera, justificar el uso de variables predictoras, se presenta la matriz de correlación (fig: 35) entre peak-wpk y peak. Las peak-wpk presentan

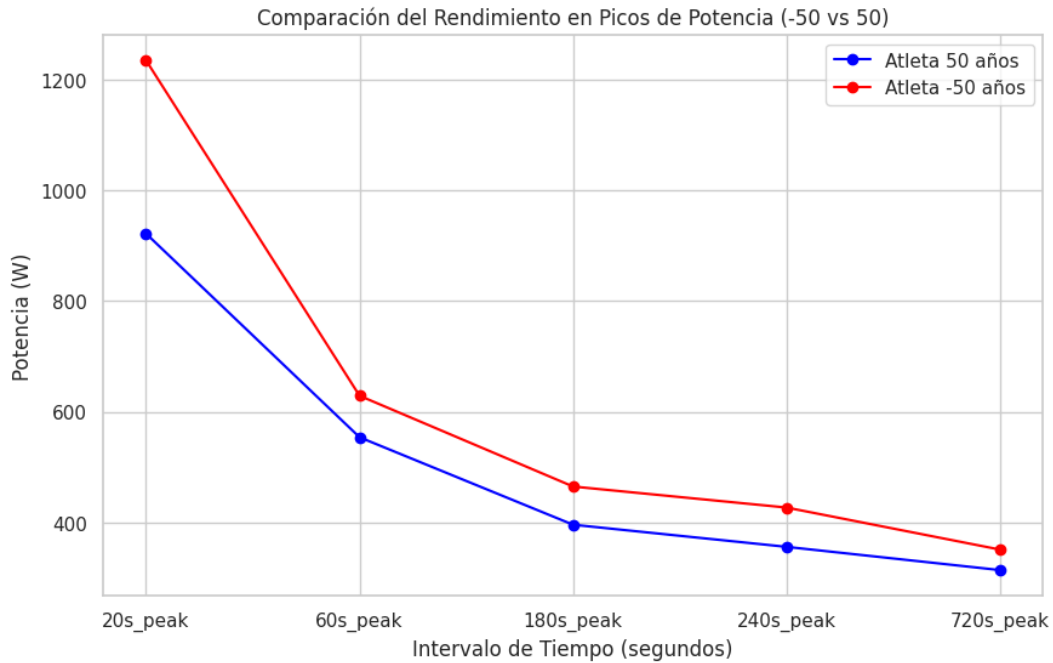


Figura 31: Comparación picos de potencia promedio entre edades.
Fuente: Elaboración propia.

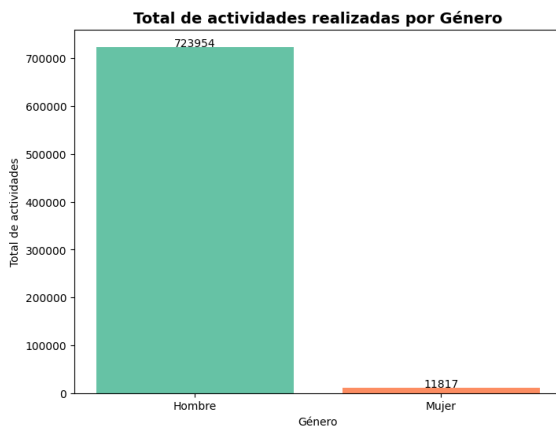


Figura 32: Hombres y mujeres.
Fuente: Elaboración propia.

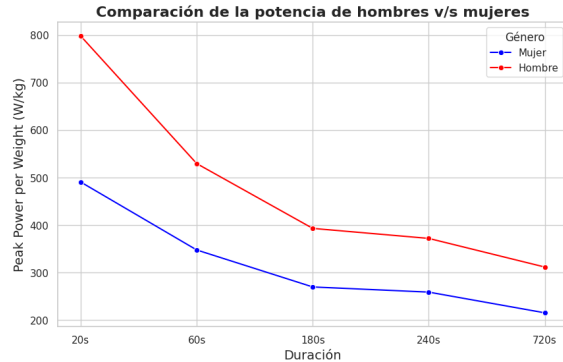


Figura 33: Diferencias de potencias.
Fuente: Elaboración propia.

la potencia relativa respecto al peso, tal como se explicó en la sección 2.6. Los que solamente tienen peak, representan la potencia absoluta (2.5), vale decir, que esta se puede obtener directamente de un sensor de potencia sin realizar ningún calculo y significa un buen indicador de cuanta potencia en Watts ejerce sobre el pedal el ciclista.

Respecto a relaciones lineales, validan parte de la literatura descrita en esta memoria (2.2) donde se habla de la variación en la potencia respecto a la capacidad aeróbica como anaeróbica. En este dataset, hay indicios de que a medida que aumenta la potencia máxima de-

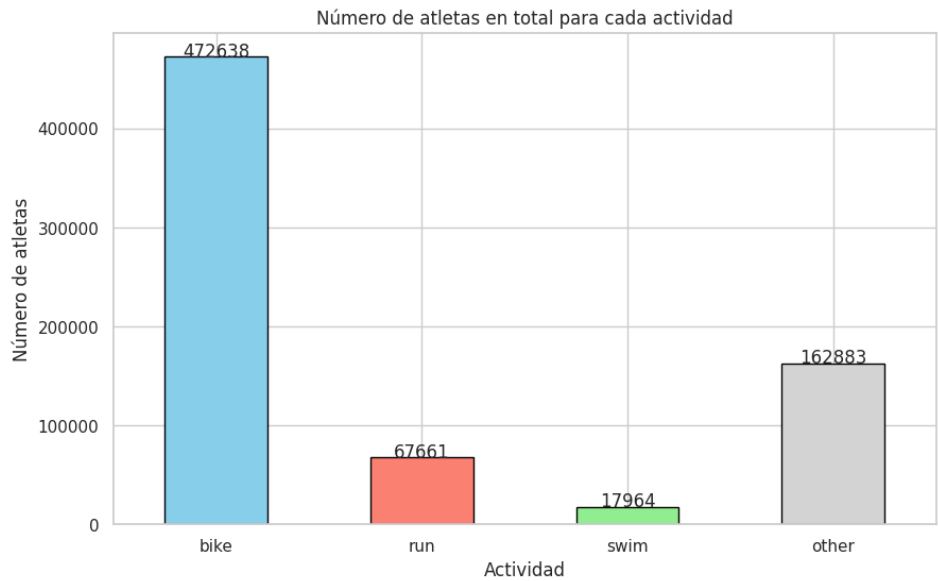


Figura 34: Distribución de actividades por atleta.
Fuente: Elaboración propia.

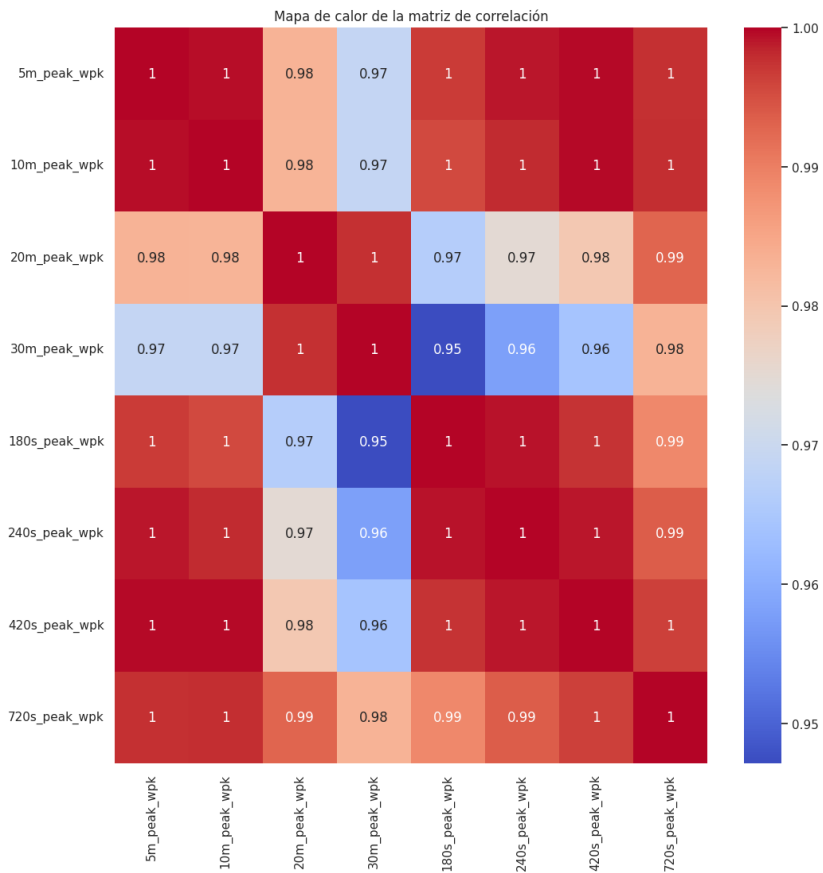


Figura 35: Matriz de correlación en distintas potencias.
Fuente: Elaboración propia.

sarrollada en un esfuerzo de 240 segundos, también tiende a aumentar la potencia máxima sostenida durante 30 minutos (44 % variabilidad y $p > 0.05$, figura 36). En simples palabras, los atletas con una mayor capacidad aeróbica (reflejada en la potencia máxima sostenida durante 30 minutos) también tienden a tener una mayor capacidad anaeróbica (reflejada en la potencia máxima en esfuerzos cortos). También se presentan otras relaciones lineales, como por ejemplo, el aumento de potencia en relación a su número de actividades (figura 37) y la relación entre la edad y potencia (a medida que aumenta la edad disminuye la potencia, figura 38).

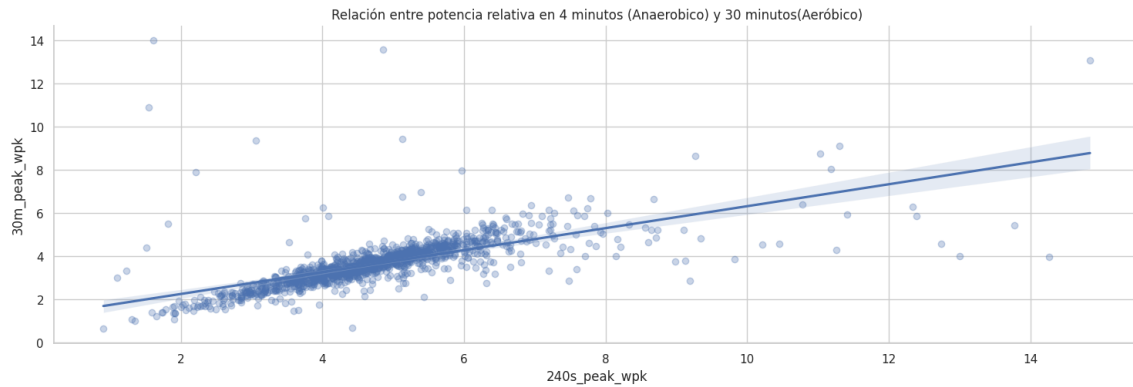


Figura 36: Relación entre capacidad anaeróbica y aeróbica.
Fuente: Elaboración propia.

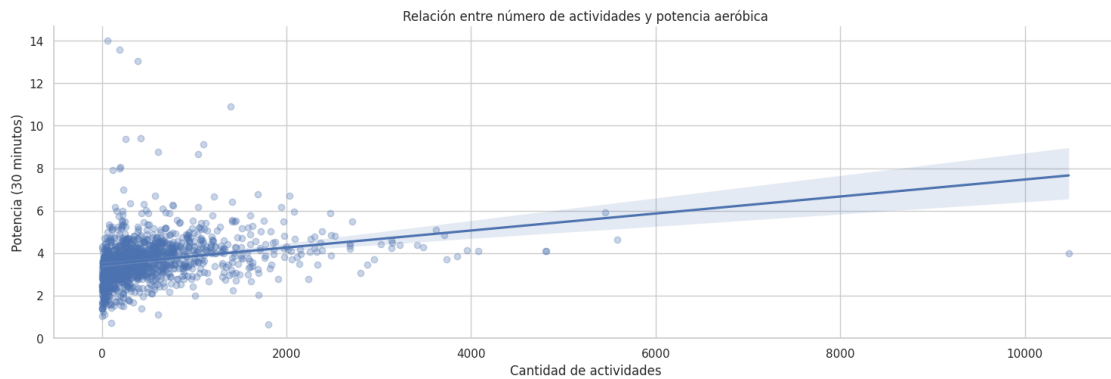


Figura 37: Número de actividades y potencia absoluta.
Fuente: Elaboración propia.

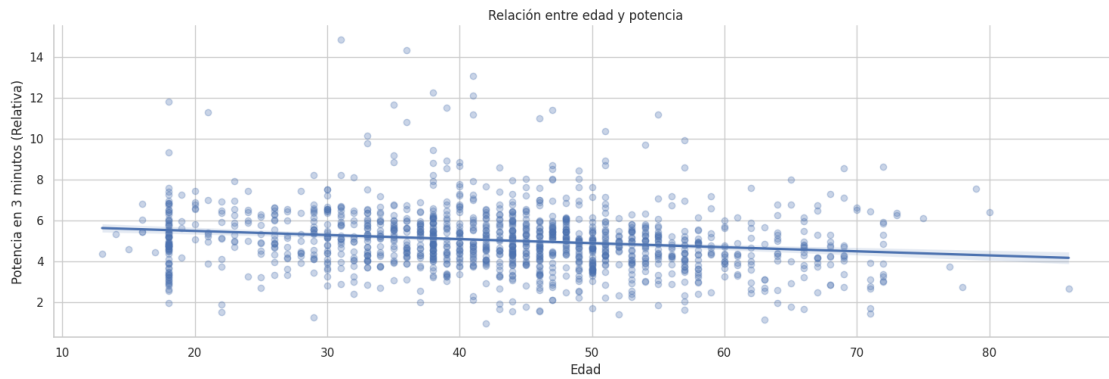


Figura 38: Relación entre edad y rendimiento (potencia relativa).
Fuente: Elaboración propia.

3.4. Pre-procesamiento

Luego de explorar y analizar el dataset de los atletas, se logró cumplir con el primer paso de la metodología propuesta. Ahora, el siguiente paso, según la metodología propuesta, corresponde a «Preparación de los datos» (Sección: 2.14).

En este paso, se tomó en cuenta el análisis exploratorio para saber que variables necesitaban ser modificadas. Es por esto, que en el primer paso, se ignoraron todos los datos que fueran menor a 0 en las métricas de potencia. Dado que todos los atletas que no registraban valores en métricas como «1s_critical_power» o «5m_peak_wpk» no calificaban para el propósito de este estudio. Así también, se eliminaron los nulos, que, como se mencionó en la sección anterior, oscilaban entre los 20-25 datos que solamente pertenecían a la categoría «weightkg» y «weightstd».

Al tener el conjunto de datos sin nulos ni atletas que no representaran actividades con sensores de potencia, la calidad incrementó considerablemente. No obstante, existían datos imprecisos o fuera de lo normal, tal como se vio en 28 habían atletas con edades negativas o valores fuera de lo normal, como 2018. Para solucionar esto, se hizo lo siguiente:

Primero, se ploteo un gráfico (fig: 31) donde se pudieran apreciar las diferencias de potencias entre 2 grupos por un rango de edad. Los rangos están inspirados en la escala de edades para ciclista de [AusCycling, 2024].

1. **Potencialmente válido:** todos los valores entre -10 y -90.
2. **Válido:** todos los valores entre 10 y 90.
3. **Outliers:** todos los valores fuera del rango etario definido.

Luego de aplicar este filtro, se realizaron comparaciones visuales con el fin de comprobar

si los dos grupos que conformaban valores atípicos, se diferenciaban enormemente de los valores válidos o no.

Se estableció un umbral inicial de 16 % de desviación respecto a la potencia promedio observada en los datos considerados válidos. Este umbral se basa en la suposición de que la mayoría de los ciclistas presentan valores de potencia dentro de un rango razonable alrededor de la media (figura 39 & 40).

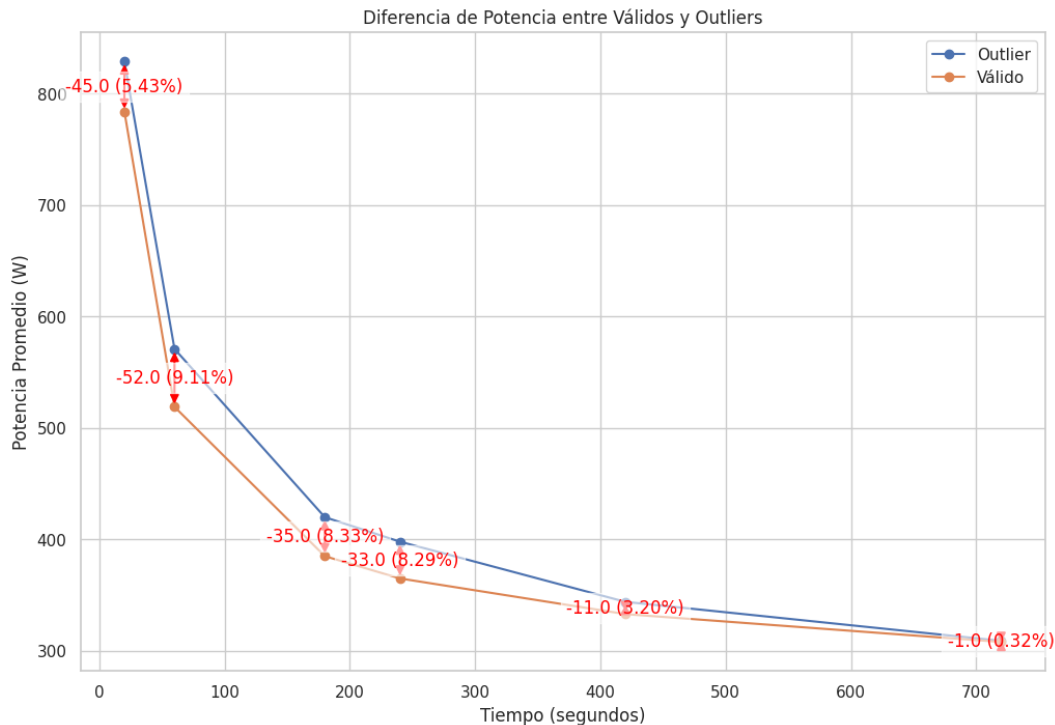


Figura 39: Diferencias entre edades válidas y potencialmente válidas.

Fuente: Elaboración propia.

Se observa como la diferencia de potencia entre las distintas categorías creadas para las edades, no es significativamente alta, por lo cual, se asume que en los 3 grupos se encuentran edades completamente reales y solo se trató de un error de tipeo¹³ o del programa mismo en registrar este dato. Dicho esto, se aplica normalización para el grupo etario de «potencialmente válido» dejando así todos los números negativos en su contra parte positiva (ejemplo 50 y -50 ya no existe, ahora ambos son 50 como valor único). Para los outliers, los valores se agruparon a la media, ya que, por ejemplo, para 2018 no hay como interpretar su edad real.

Avanzando con el pre-procesamiento, se aplicó la categorización para los ciclistas basado en la potencia relativa, distribuido en capacidad aeróbica y anaeróbica. La construcción de las siguientes tablas está inspirada en la propuesta hecha por [Jeukendrup *et al.*, 2000] (Véase la figura 11). Sin embargo, ya que no se dispone de todas las variables que se presentan en

¹³Error al escribir una variable esperada, por ejemplo, agregar un carácter de más.

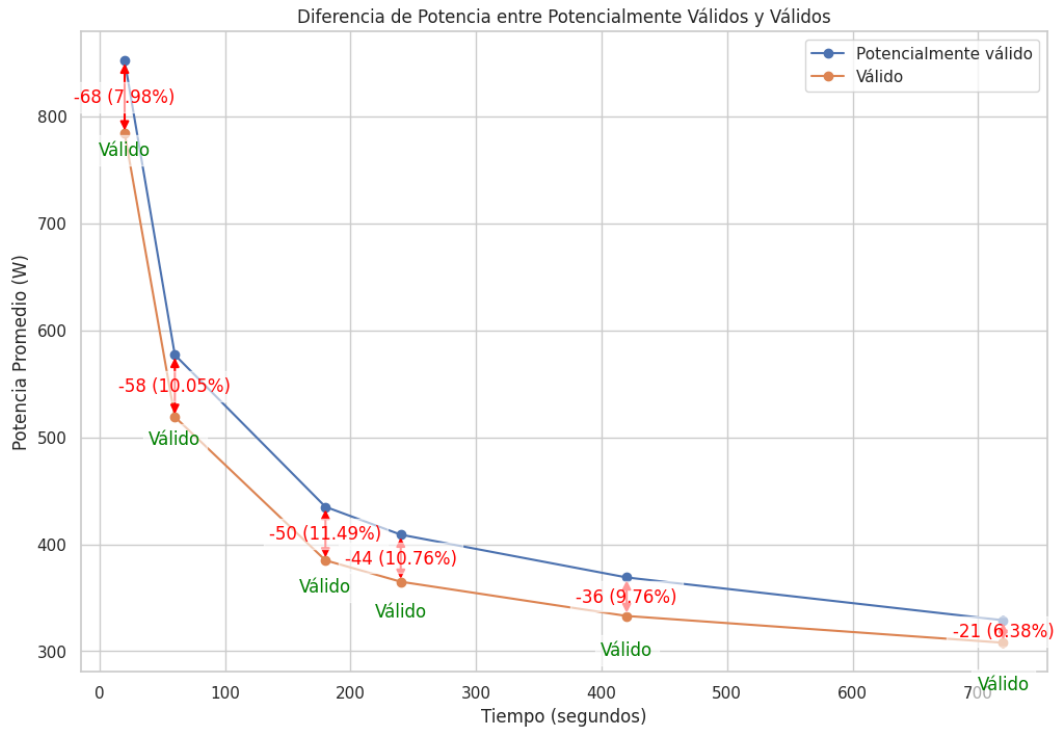


Figura 40: Diferencias entre edades válidas y outliers.

Fuente: Elaboración propia.

dicha tabla, se simplificó en capacidad anaeróbica y aeróbica (tablas 7 y 8), puesto que son dos aspectos sumamente importantes para clasificar el rendimiento de un ciclista y, además, ambos conjuntos de datos cuentan con mediciones en tiempos cortos (anaeróbico) y rutinas extendidas (aeróbico).

Tabla 7: Categorización de la capacidad anaeróbica.

Fuente: Elaboración propia.

Capacidad anaeróbica	Rango de potencia (5m_peak_wpk)
Casual	<3.5 WPK
Principiante	3.5 - 4.5 WPK
Experimentado	4.5 - 6.0 WPK
Élite	>6.0 WPK

El gráfico de violines en la figura 41 muestra una clara tendencia hacia una mayor potencia promedio a medida que avanzamos de las categorías más bajas a las más altas, con una variación estimada entre el 100 % y 150 % entre los ciclistas casuales y de élite. Además, es interesante observar que la dispersión de los datos también varía entre las categorías. Los ciclistas de élite, por ejemplo, tienden a tener una distribución más estrecha de la potencia, lo que sugiere una mayor consistencia en su rendimiento.

La presencia de outliers, particularmente en la categoría «Casual», indica que hay una varia-

Tabla 8: Categorización de la capacidad aeróbica.
Fuente: Elaboración propia.

Capacidad aeróbica	Rango de potencia (30m_peak_wpk)
Casual	<2.5 WPK
Principiante	2.5 - 3.5 WPK
Experimentado	3.5 - 4.5 WPK
Élite	>4.5 WPK

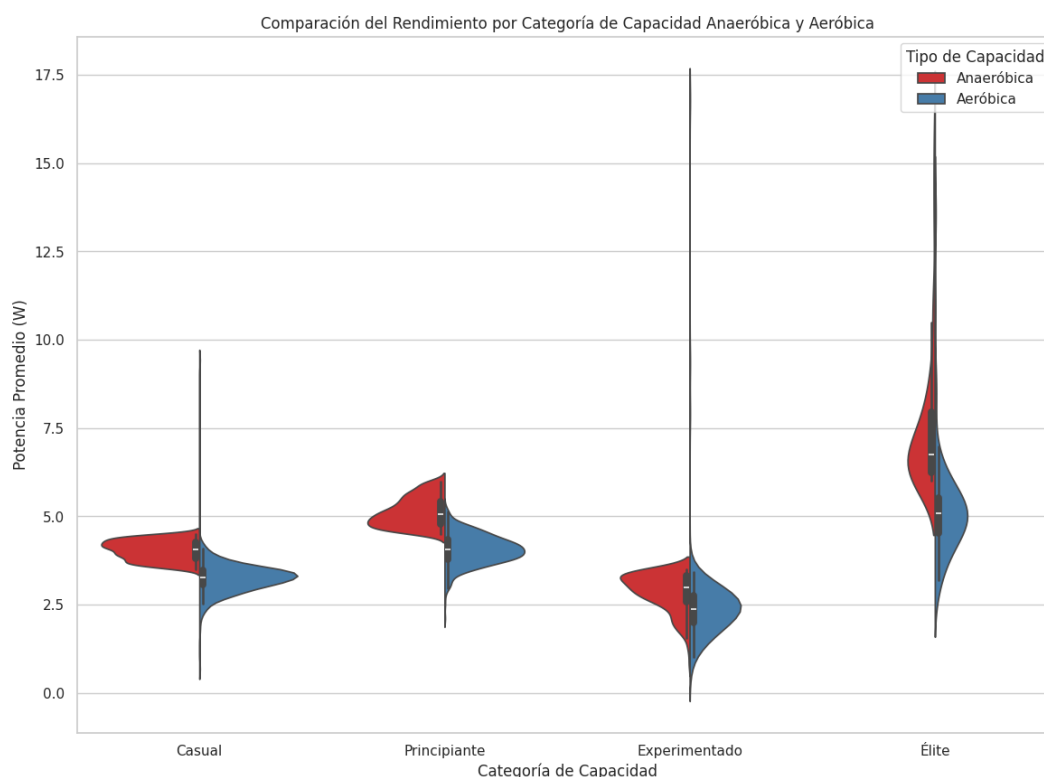


Figura 41: Potencia relativa en distintas categorías.
Fuente: Elaboración propia.

bilidad significativa dentro de cada grupo. Estos valores atípicos podrían deberse a factores como errores de medición, actividades inusuales o características individuales de los ciclistas.

Al comparar las diferentes categorías, podemos observar que algunos violines se superponen, lo que sugiere que la clasificación no siempre es rígida y que algunos ciclistas «Casual» pueden tener un rendimiento comparable al de ciclistas «Principiantes» en ciertas métricas.

Es importante destacar que este análisis se basa en la potencia promedio y no considera otros factores como la duración de la actividad, el tipo de terreno o las condiciones climáticas.

Por último, se dropean columnas no relevantes para el propósito de este estudio como «id» y

Tabla 9: Distribución de Atletas por Categoría y Capacidad.
Fuente: Elaboración propia.

	Casual	Principiante	Experimentado	Élite
Capacidad Anaeróbica	208	505	426	159
Capacidad Aeróbica	138	514	444	202

«weightstd». Además, se convierten las variables para clasificar en números del 0 al 3 puesto que los modelos de clasificación solo aceptan este formato. Lo mismo con la variable «gender» utilizando .map. (Tabla 9).

3.5. Entrenamiento XGBoost

Con una visión clara del problema y datos de alta calidad, el siguiente paso es escoger el mejor modelo que se ajuste al problema que estamos abordando. Puesto que los arboles de decisiones son extremadamente poderosos, se comparan XGBoost y Random Forest. Además, se agrega KNN tal como se mencionó en el paso 4 de la metodología CRISP-DM 2.14.

Tabla 10: Resumen de Resultados de los Modelos.
Fuente: Elaboración propia.

Modelo	Media del F1-Score	Desviación Estándar
XGBoost	0.8998	0.0178
Random Forest	0.8819	0.0157
KNN	0.5026	0.0310

Los resultados obtenidos (tabla 10) demuestran que XGBoost logra la mayor precisión en la tarea de clasificación, superando en un considerable margen a los otros modelos. Random Forest también obtiene resultados sobresalientes, consolidando la idea de que los modelos basados en árboles de decisión son altamente efectivos para este tipo de problemas multi-clase. La alta dimensionalidad de los datos y la posible existencia de interacciones no lineales entre las variables podrían explicar el buen desempeño de estos modelos.

Sin embargo, KNN muestra un rendimiento significativamente inferior, aproximadamente un 40 % por debajo de XGBoost (véase otra representación en figura 42). Esto sugiere que la naturaleza de los datos y la complejidad de las relaciones entre las variables no se ajustan bien a este algoritmo basado en la distancia.

Si bien los árboles de decisión han demostrado ser una excelente opción para este problema en particular, es importante considerar que su rendimiento puede variar dependiendo de las características específicas de los datos. En futuros trabajos, sería interesante explorar

técnicas de ensemble más avanzadas, como LightGBM o CatBoost, para evaluar si pueden ofrecer mejoras adicionales.

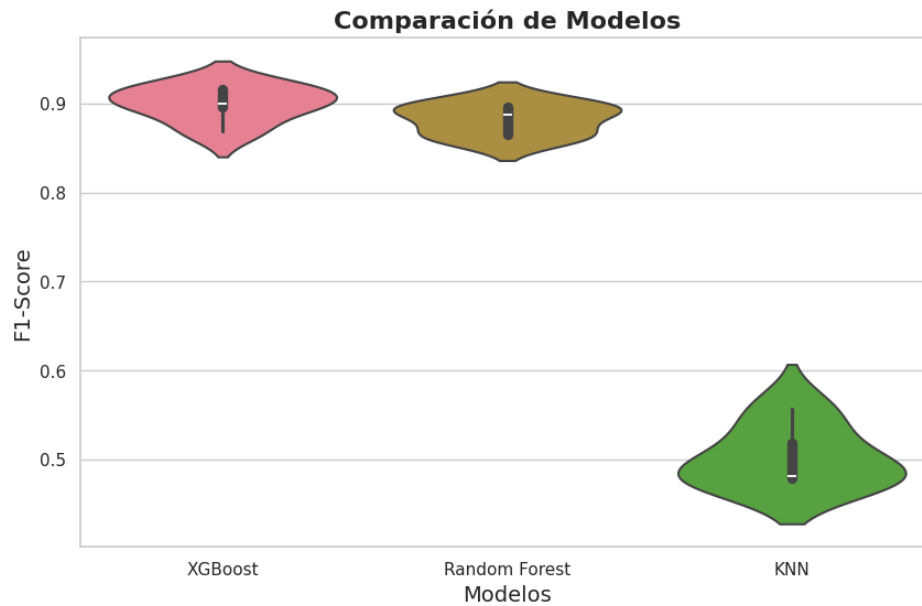


Figura 42: Comparación de modelos.
Fuente: Elaboración propia.

3.5.1. Aplicación del modelo seleccionado

Dada la buena precisión y resultados previos, se entrenó un modelo de XGBoost utilizando un conjunto de características cuidadosamente seleccionadas para evitar alta correlación con la variable objetivo. Así también, este modelo XGBoost servirá en la explicación y selección de características (además de las previamente seleccionadas). XGBoost asigna una métrica de importancia a cada característica, cuantificando su contribución a la precisión de las predicciones finales. Esta métrica se basa en la reducción de la impureza que cada característica induce en el modelo. La figura 14 ejemplifica mejor este proceso.

Antes de ejecutar el código, es importante preparar los datos para que puedan ser utilizados por el modelo de machine learning. Esto implica escalar las características de los ciclistas (tabla 11) para que estén en una escala similar y dividir los datos en un conjunto de entrenamiento y otro de prueba (80-20). Además, se utiliza una técnica llamada SMOTE para equilibrar las clases de la variable objetivo (capacidad anaeróbica o capacidad aeróbica), asegurando que el modelo no se sesgue hacia las clases más frecuentes.

Una vez ejecutado, este se entrena utilizando los datos de entrenamiento, aprendiendo a identificar patrones entre las características de los ciclistas y su capacidad anaeróbica o aeróbica. El objetivo es que el modelo sea capaz de predecir cualquiera de las dos clases (por

Tabla 11: Características utilizadas en el modelo XGBoost.
Fuente: Elaboración propia.

Característica	Descripción
age	Edad del atleta (en años).
gender	Sexo del atleta (masculino/femenino).
activities	Actividades físicas realizadas.
Potencia máxima	20s_peak, 60s_peak, 180s_peak, 420s_peak, 720s_peak (en watts). Cada intervalo de tiempo en segundos es la potencia máxima alcanzada en ese tiempo.
weightkg	Peso del atleta (en kilogramos).

separado, ya que abordan distintos espectros) de un nuevo ciclista en función de sus características.

Finalmente, el modelo entrenado se evalúa utilizando los datos de prueba. Se realizan predicciones sobre estos datos y se comparan con los valores reales de ambas capacidades. A través de métricas como la precisión, la matriz de confusión y el informe de clasificación, se evalúa qué tan bien el modelo puede generalizar a nuevos datos y realizar predicciones precisas. Los resultados son los siguientes:

Precisión capacidad anaeróbica: 0.8836633663366337.

Precisión capacidad aeróbica: 0.8592233009708737.

Con un resultado bastante prometedor, que se aprecia en la figura 43 y 44, el modelo es capaz de predecir con un 88 % de precisión la categoría del ciclista respecto a su capacidad anaeróbica y 85 % respecto a su capacidad aeróbica con las mismas características. A continuación, se muestra el reporte para comprender los porcentajes por cada una de ellas (0 - Casual, 1 - Principiante, 2 - Experimentado y 3 - Élite).

Classification Report:				
	precision	recall	f1-score	support
0	1.00	0.94	0.97	107
1	0.82	0.81	0.82	99
2	0.83	0.94	0.88	97
3	0.89	0.84	0.86	101
accuracy			0.88	404
macro avg	0.88	0.88	0.88	404
weighted avg	0.89	0.88	0.88	404

Figura 43: Resultados - Capacidad anaeróbica.
Fuente: Elaboración propia.

Classification Report:					
	precision	recall	f1-score	support	
0	0.96	0.94	0.95	98	
1	0.75	0.83	0.78	99	
2	0.87	0.81	0.84	118	
3	0.88	0.87	0.87	97	
accuracy			0.86	412	
macro avg	0.86	0.86	0.86	412	
weighted avg	0.86	0.86	0.86	412	

Figura 44: Resultados - Capacidad aeróbica.

Fuente: Elaboración propia.

Al comparar los resultados de ambos modelos, podemos observar algunas diferencias interesantes:

- **Precisión Global:** Ambos modelos muestran una precisión global similar (alrededor del 86-88 %), lo que indica un buen desempeño general.
- **Clase 0 (Casual):** Ambos modelos obtienen una alta precisión y recall, lo que sugiere que son muy buenos para identificar a los ciclistas casuales.
- **Clase 1 (Principiante):** El modelo de capacidad anaeróbica parece tener un ligero mejor desempeño en términos de precisión.
- **Clase 2 (Experimentado) y 3 (Élite):** En estas clases, las diferencias son menos marcadas, aunque el modelo de capacidad anaeróbica tiende a tener un F1-score ligeramente más alto.

En general, el modelo se adapta y logra predecir bastante bien las relaciones no lineales existentes, lo que potencia la tabla y valida la afirmación propuesta por [Jeukendrup *et al.*, 2000]. Ahora, es necesario evaluar y comprender el modelo más en profundidad, ya que esto puede ayudar a buscar mejoras y ajustes en base al F1-Score o la matriz de confusión.

3.6. F1-Score y Matriz de Confusión

Las siguientes imágenes demuestran la variación respecto a los falsos y verdaderos positivos y la contribución general al modelo en conjunto utilizando el F1-Score (figura 45 y 46).

El análisis de las matrices de confusión revela que los modelos de clasificación propuestos exhiben un desempeño prometedor en la tarea de categorizar a los ciclistas según su nivel de rendimiento (casual, principiante, experimentado y élite). Sin embargo, se observan algunas limitaciones, particularmente en la diferenciación entre las categorías de «experimentado» y «élite». La presencia de falsos positivos y falsos negativos en estas clases sugiere que el

DESARROLLO DE UN MODELO DE APRENDIZAJE AUTOMÁTICO PARA CLASIFICAR EL NIVEL DE HABILIDAD DE CICLISTAS

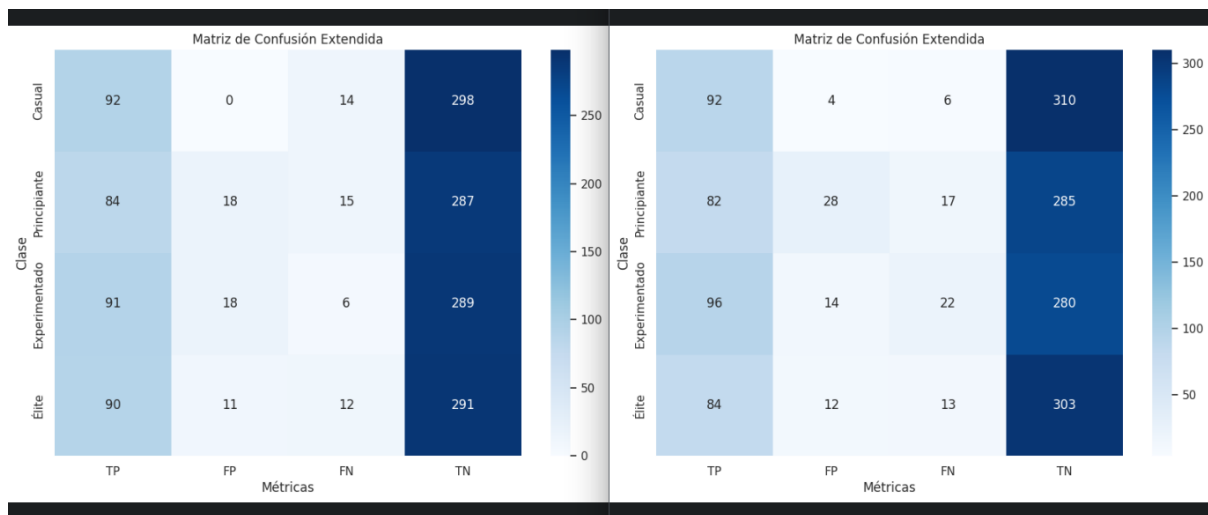


Figura 45: Matriz de confusión de ambas predicciones.
Fuente: Elaboración propia.

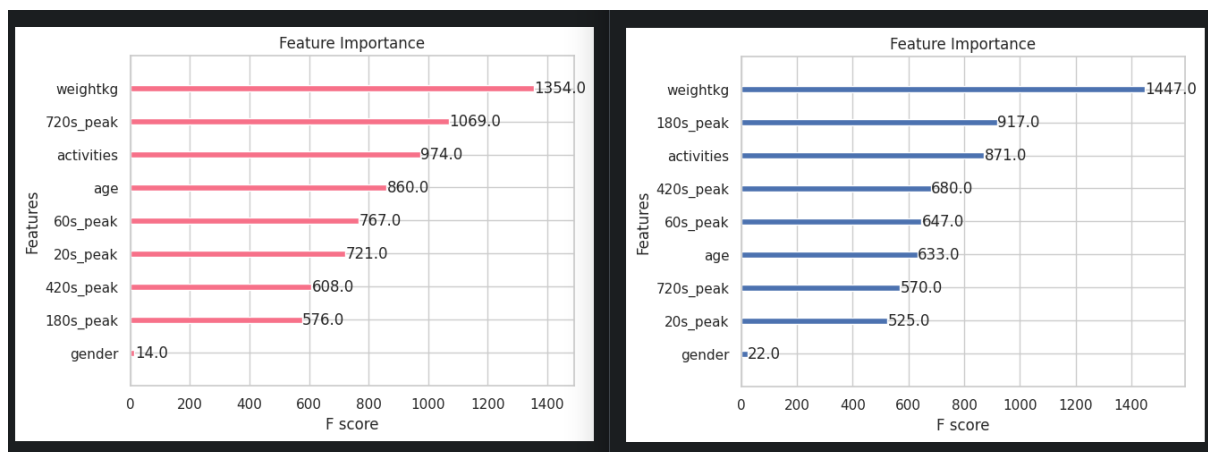


Figura 46: Feature importance - Rosa Aeróbico y Azul Anaeróbico.
Fuente: Elaboración propia.

modelo podría beneficiarse de la incorporación de datos adicionales, como la experiencia competitiva y la frecuencia de entrenamiento específico, para mejorar la precisión en la discriminación entre niveles de rendimiento muy cercanos.

Al analizar la importancia de las características en cada modelo, se puede deducir que:

- **Capacidad Aeróbica:** La edad y el volumen total de entrenamiento se alzan como los predictores más relevantes, lo que corrobora la hipótesis de que la experiencia acumulada y el entrenamiento a largo plazo son fundamentales para desarrollar una sólida capacidad aeróbica. Asimismo, la capacidad de mantener un esfuerzo sostenido durante intervalos de duración moderada (60s, 180s y 420s) demuestra ser un factor determinante.
- **Capacidad Anaeróbica:** Los intervalos de alta intensidad de corta duración (60s y 180s) emergen como los principales predictores, lo que concuerda con la naturaleza explosiva de la capacidad anaeróbica. Además, el peso corporal se revela como una variable significativa, influyendo en ambas capacidades de manera compleja. Si bien un peso mayor puede ser una desventaja en términos de potencia relativa, puede conferir ventajas en ciertos tipos de esfuerzos.

Por último, respecto a feature importance, se presentan otras formas de medir esta característica, como son «cover» y «gain», definidas en 2.13.1. Figuras 47 & 48.

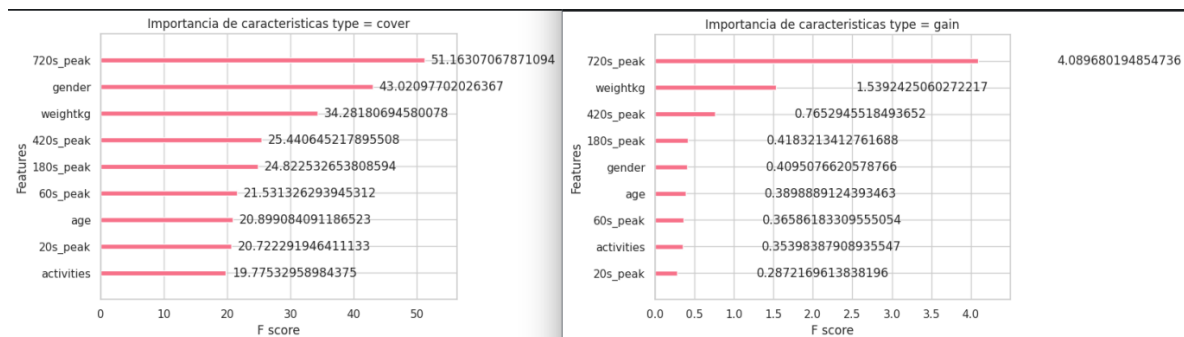


Figura 47: Gain y cover - Aeróbico.

Fuente: Elaboración propia.

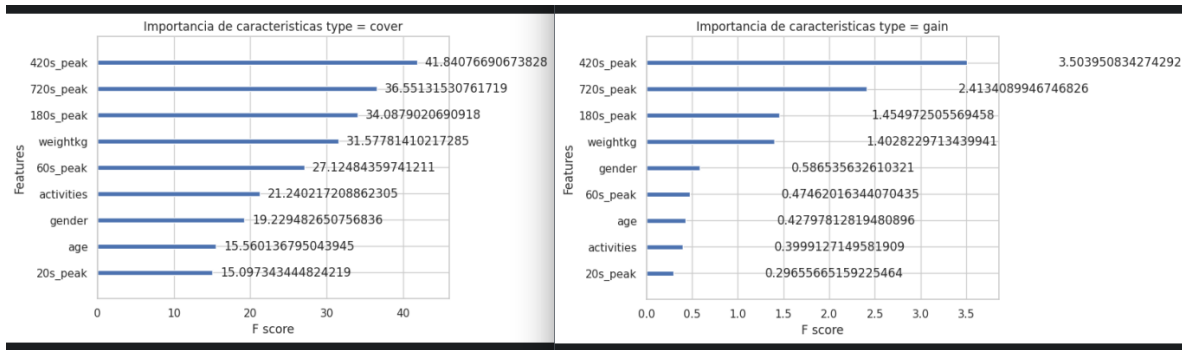


Figura 48: Gain y cover - Anaeróbico.

Fuente: Elaboración propia.

3.6.1. SHAP Values

Luego de un detallado análisis mediante feature importance, es importante también comprender este modelo mediante los SHAP Values, que demuestran de manera individual la contribución de cada característica, a diferencia de los F1-Score que es más general y en conjunto.

En el gráfico tipo «beeswarm» (figura 49) se pueden percatar ciertas influencias (anaeróbica). Por ejemplo, el peso tiene un alto impacto (como se ha visto en resultados anteriores) pero su efecto es más disperso y depende de otros factores. Podría tener un efecto positivo o negativo dependiendo de la composición corporal y otros factores fisiológicos.

La edad y la cantidad total de actividad tienen un impacto moderado, lo que sugiere que la experiencia y el entrenamiento regular también contribuyen a la capacidad anaeróbica. Esto es muy importante para el próximo conjunto de datos, ya que se pueden incluir estas métricas para aumentar la precisión del modelo. El género parece no tener mayor impacto, pero esto se debe netamente a que la cantidad de hombres es superior a las mujeres por mucha diferencia (1320 vs 80 aprox). No se consideró hacer un balanceo de atletas ya que no es el caso de estudio.

En gráfico (figura 50) generado con el entrenamiento respecto a la capacidad aeróbica, a diferencia de la capacidad anaeróbica, donde los intervalos más cortos eran cruciales, en la capacidad aeróbica observamos que los intervalos de mayor duración (720s y 420s) tienen un impacto más pronunciado. Esto es coherente con la naturaleza de la capacidad aeróbica, que se desarrolla a través de esfuerzos sostenidos y de larga duración. Además, la característica «activities» tiene un impacto positivo significativo, lo que sugiere que el volumen total de entrenamiento es un factor clave para desarrollar la capacidad aeróbica.

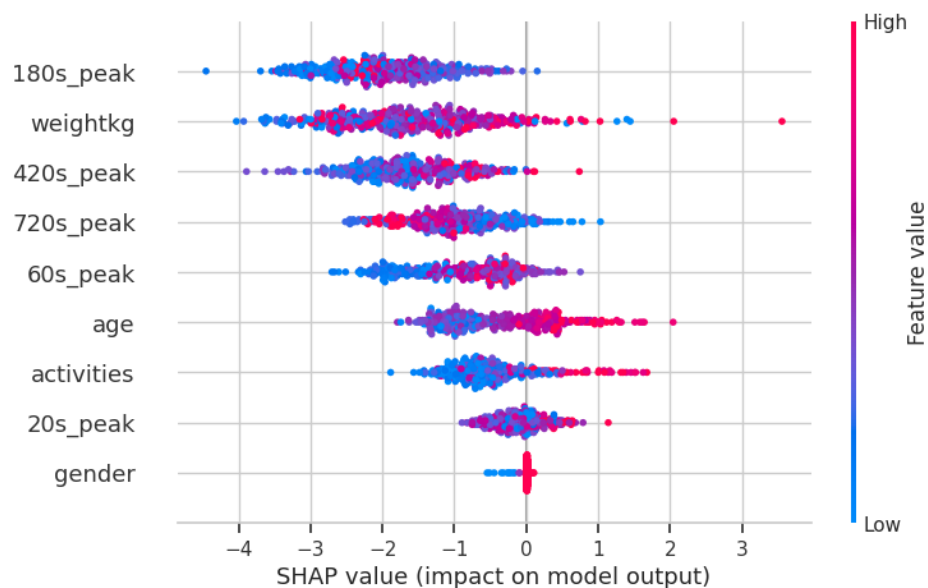


Figura 49: SHAP Values - Anaeróbico.
Fuente: Elaboración propia.

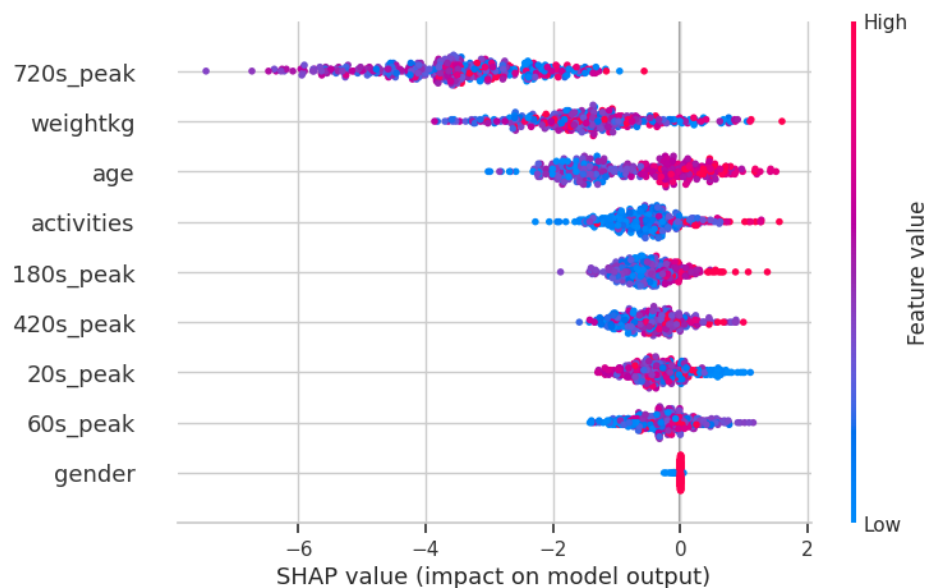


Figura 50: SHAP Values - Aeróbico.
Fuente: Elaboración propia.

3.6.2. Selección de características

Teniendo una comprensión de los datos, alineado netamente con la literatura revisada, como último paso de este capítulo, se lleva a cabo el forward wrapper selection para selección de características y mejorar la precisión de ambos modelos, utilizando XGBoost nuevamente.

Primero, se ordenaron las características en una variable en relación a su importancia (feature importance). Luego se definen dos funciones:

- **evaluate_model:** Esta función toma un subconjunto de características y las utiliza para entrenar un modelo (XGBoost en este caso). Luego evalúa la precisión del modelo en un conjunto de prueba independiente.
- **forward_selection:** Esta función principal implementa el algoritmo hacia adelante. Parte con un conjunto vacío de características seleccionadas y el conjunto completo como características restantes. Itera evaluando cada característica restante agregándola al conjunto seleccionado y entrenando nuevamente el modelo. Si la precisión mejora, se agrega la característica al conjunto seleccionado y se actualiza el mejor conjunto encontrado hasta el momento. El proceso continúa hasta que no se encuentre ninguna característica que mejore la precisión.

```

Evaluando conjunto: ['180s_peak', 'weightkg', '420s_peak', 'activities'], Precisión: 0.8685388538853885
Evaluando conjunto: ['180s_peak', 'weightkg', '420s_peak', '20s_peak'], Precisión: 0.8564356435643564
Evaluando conjunto: ['180s_peak', 'weightkg', '420s_peak', '60s_peak', '720s_peak'], Precisión: 0.8663366336633663
Evaluando conjunto: ['180s_peak', 'weightkg', '420s_peak', '60s_peak', 'gender'], Precisión: 0.8737623762376238
Evaluando conjunto: ['180s_peak', 'weightkg', '420s_peak', '60s_peak', 'age'], Precisión: 0.8688118811881188
Evaluando conjunto: ['180s_peak', 'weightkg', '420s_peak', '60s_peak', 'activities'], Precisión: 0.8836633663366337
Evaluando conjunto: ['180s_peak', 'weightkg', '420s_peak', '60s_peak', '20s_peak'], Precisión: 0.8787128712871287
Evaluando conjunto: ['180s_peak', 'weightkg', '420s_peak', '60s_peak', 'activities', '720s_peak'], Precisión: 0.888
Evaluando conjunto: ['180s_peak', 'weightkg', '420s_peak', '60s_peak', 'activities', 'gender'], Precisión: 0.881188
Evaluando conjunto: ['180s_peak', 'weightkg', '420s_peak', '60s_peak', 'activities', 'age'], Precisión: 0.868811881
Evaluando conjunto: ['180s_peak', 'weightkg', '420s_peak', '60s_peak', 'activities', '20s_peak'], Precisión: 0.8712
Evaluando conjunto: ['180s_peak', 'weightkg', '420s_peak', '60s_peak', 'activities', '720s_peak', 'gender'], Precisión: 0.8886138613861386
Evaluando conjunto: ['180s_peak', 'weightkg', '420s_peak', '60s_peak', 'activities', '720s_peak', 'age'], Precisión: 0.8886138613861386
Evaluando conjunto: ['180s_peak', 'weightkg', '420s_peak', '60s_peak', 'activities', '720s_peak', '20s_peak'], Precisión: 0.8886138613861386
Evaluando conjunto: ['180s_peak', 'weightkg', '420s_peak', '60s_peak', 'activities', '720s_peak', '20s_peak', 'gender'], Precisión: 0.8886138613861386
Mejores características: ['180s_peak', 'weightkg', '420s_peak', '60s_peak', 'activities', '720s_peak']
Mejor precisión: 0.8886138613861386

```

Figura 51: Forward Wrapper Selection.
Fuente: Elaboración propia.

Como se aprecia en la figura 51, se va evaluando el conjunto con cada combinación posible, utilizando las características (o features) ordenadas del F1-score, iterando sin parar hasta recorrer toda la lista y quedarse con que obtienen mejor precisión. En este caso, las mejores features son los intervalos de tiempo mas cercanos a 5m (la variable target) y además, el peso y cantidad de actividades. Todo esto para la capacidad anaeróbica, que en general, se mejora la precisión con esta técnica solo en un 1 % puesto que no se cuenta con una gran cantidad de variables. Sin embargo, se demuestra como mejora de todas maneras.

Para la capacidad aeróbica, los resultados en la selección de features arrojaron lo siguiente:

- **Mejores características:** ['180s_peak', 'weightkg', '420s_peak', '60s_peak', 'activities', '720s_peak'].
- **Mejor precisión:** 0.8886138613861386.

Resultados que, de igual manera, se complementan bajo todo el estudio hecho en esta memoria y aportan hallazgos importantes para considerar en otros modelos, donde las actividades claramente presentan una importancia en capacidades aeróbicas.

Con el fin de demostrar la variación respecto a los falsos y verdaderos positivos, se hizo un gráfico de líneas que demuestra como va cambiando a medida que se agregan características, todo esto con las «best features» que entregó el método envoltorio o wrapper anterior (figura 52).

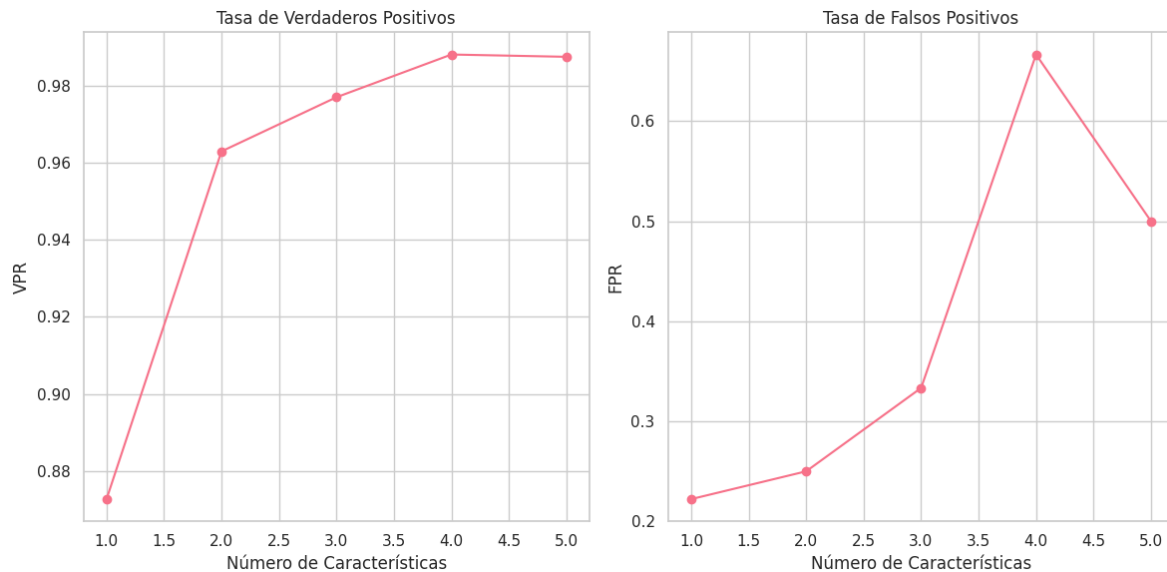


Figura 52: Falsos y verdaderos positivos.

Fuente: Elaboración propia.

El siguiente paso es validar el modelo en el conjunto de datos «activities.csv» donde, si bien la calidad como se mencionó en un principio es baja, se realizarán técnicas de preprocesamiento e ingeniería de características similares a las aplicadas en este capítulo. La idea es validar si las técnicas y modelo aplicado en este dataset aplican para otro y que podemos rescatar cuando se agregan mas features de otro tipo, como cualidades de entrenamiento, velocidad, distancia, etc.

CAPÍTULO 4

VALIDACIÓN DE LA SOLUCIÓN

El primer paso una vez cargado el conjunto de datos en Google Colab, fue realizar un análisis exploratorio más allá de lo que se conocía. Lo primero, además de que las dimensiones son 2.398.134 filas y 34 columnas, es que tiene el mismo problema con las edades que el dataset de «athletes.csv». Además, un punto a favor, es que contiene una amplia variedad de actividades que solo son de ciclismo. A diferencia del otro, que fue más generalizado, se conservaron estas features para filtrar el tipo de deporte:

```
sports_features = ['Bike', 'Cycling', 'Mountain Bike', 'MTB', 'Rennrad', 'BDC', 'Cycle', 'Ciclismo', 'Cykling', 'Crossrad', 'Rennvelo', 'EBikeRide']
```

Luego de aplicar el filtro, la variable 'sport' quedó con features relacionadas solamente al ciclismo. Además de esto, se observaron variables no relevantes para el caso de estudio, como 'data' que corresponde al sensor o dispositivo de donde se obtuvo la información, 'file' que es el nombre del .csv que se generó, 'id' que es el identificador único del atleta y Unnamed, una variable aleatoria sin importancia.

4.1. Preparación de los datos

Dado que este conjunto de datos contenía muchísima información, era necesario hacer los filtros correspondientes antes de aplicar y poner en práctica el modelo. Para esto, se eliminaron nulos, se ajustaron las fechas en el formato 'datetime' y se modificaron los tiempos de la duración en las rutinas, que estaban en segundos y se pasaron a minutos. De igual manera, se promediaron todos los valores y en vez de tener 2.398.134 filas y 34 columnas ahora hay un atleta único con el resumen de todas sus actividades únicas, teniendo así 4695 filas (atletas) y 32 columnas (características).

Puesto que era necesario agregar más features al modelo, se calcularon promedios de frecuencia semanal utilizando la columna 'date' (que se realizaron entre el 2017 y 2019) y también se aplicó para la duración de entrenamiento. Esto con el fin de seguir la tabla de [Jeukendrup *et al.*, 2000].

Luego de tener los datos pre-procesados correctamente, características nuevas para contribuir al modelo en base a los resultados del capítulo 2.14 y una comprensión total del problema, se llevan a cabo los entrenamientos y ajustes finales al modelo utilizando XGBoost. La siguiente tabla (12) simplifica las categorizaciones nuevas que se agregaron a este conjunto de datos.

Tabla 12: Clasificación de la frecuencia y duración del entrenamiento.

Fuente: Elaboración propia.

Categoría	Frecuencia (sesiones/semana)	Duración (minutos/sesión)
Baja	< 2	< 30
Moderada	2 – 3	30 – 60
Intermedia	3 – 6	60 – 150
Avanzada	> 6	> 150

4.2. Resultados finales

Se entrenó un nuevo modelo XGBoost utilizando las mismas técnicas de escalamiento y balanceo que en el capítulo anterior. Para evaluar el desempeño del modelo, se calculó el F1-score. Además, se emplearon los valores SHAP con el fin de obtener una comprensión profunda de la contribución individual en las predicciones hechas. A partir de estos resultados, se llevó a cabo un proceso de selección de características utilizando la técnica de forward wrapper selection, con el fin de identificar el subconjunto de características que maximiza la precisión del modelo para ambas capacidades. Por último, se realizó un análisis detallado de la matriz de confusión para evaluar la capacidad del modelo de clasificar correctamente la clase positiva, prestando especial atención a las tasas de verdaderos positivos y falsos negativos. Las features generales antes de aplicar el forward wrapper selection, fueron las siguientes:

features = ['gender', 'age', 'workout_time', 'total_distance', 'elevation_gain', 'duracion_entrenamiento'¹⁴, 'average_speed', 'total_actividades', 'frecuencia_semanal', 'average_power', 'average_hr', 'total_work']

Sin embargo, con el fin de maximizar la precisión y descartar las características no relevantes en la capacidad a predecir (utilizando forward wrapper selection) los resultados de features óptimos son:

Mejores características para predecir la capacidad aeróbica: ['average_power', 'age', 'total_distance', 'average_hr', 'frecuencia_semanal', 'average_speed', 'elevation_gain', 'total_actividades', 'duracion_entrenamiento', 'workout_time', 'total_work'] Mejor precisión: 0.881282910483485.

Mejores características para predecir la capacidad anaeróbica: ['average_power', 'elevation_gain', 'age', 'total_distance', 'average_hr', 'total_actividades', 'average_speed', 'workout_time', 'frecuencia_semanal', 'duracion_entrenamiento'] Mejor precisión: 0.8811708169506335.

Si bien los resultados parecieran ser similares, abordan distintos espectros de entrenamiento

¹⁴Duración de entrenamiento es la clasificación promediada utilizando la tabla 12.

y no necesariamente indica una relación lineal. Esto se puede comprender de mejor forma con las siguientes tablas 13 y 14 que sintetizan los resultados de precision, recall, f1-score y soporte.

Tabla 13: Capacidad Anaeróbica.
Fuente: Elaboración propia.

Clase	Precisión	Recall	F1-Score	Soporte
Casual	0.82	0.79	0.80	590
Principiante	0.93	0.97	0.95	558
Experimentado	0.78	0.77	0.77	583
Élite	1.00	1.00	1.00	558

Tabla 14: Capacidad Aeróbica.
Fuente: Elaboración propia.

Clase	Precisión	Recall	F1-Score	Soporte
Casual	0.83	0.81	0.82	529
Principiante	0.90	0.98	0.93	512
Experimentado	0.79	0.73	0.74	520
Élite	0.99	1.00	1.00	528

Además, se ve un impacto positivo respecto al rate de verdaderos positivos, ya que con cada característica mejora más la cantidad de ciclistas correctamente clasificado. Esto valida y potencia la hipótesis planteada anteriormente, que el hecho de tener datos sobre el entrenamiento y frecuencia del ciclista, es más fácil para el modelo clasificarlo respecto a su capacidad aeróbica.

La validación cruzada, entregó los siguientes resultados:

- **Capacidad anaeróbica:** Cross-Validation Scores: [0.84753167 0.86937527 0.85932722 0.88859764 0.90865385]. Mean Cross-Validation Score: 0.8746971300870383
- **Capacidad aeróbica:** Cross-Validation Scores: [0.86404978 0.87123025 0.87745333 0.89277166 0.89655172] Mean Cross-Validation Score: 0.8804113500932637.

Tanto la capacidad aeróbica como la anaeróbica muestran valores medios de cross-validation score relativamente altos (alrededor de 0.87-0.88), lo que indica un buen rendimiento general del modelo. La variabilidad de los scores de cross-validation es relativamente baja, lo que sugiere que el modelo es estable y no está sobre-ajustado. La siguiente curva ROC (figura 53) nos muestra qué tan bien el modelo puede distinguir entre diferentes niveles de experiencia en ciclismo (0-casual, 1-principiante, 2-experimentado y 3-élite). Un área bajo la curva (AUC)

cercana a 1 indica que el modelo es excelente en clasificar correctamente a los ciclistas en su nivel correspondiente, mientras que un AUC cercano a 0.5 sugiere que el modelo es prácticamente aleatorio en sus predicciones. En este caso, un AUC promedio de 0.96 sugiere que el modelo es muy preciso en la clasificación de los ciclistas.

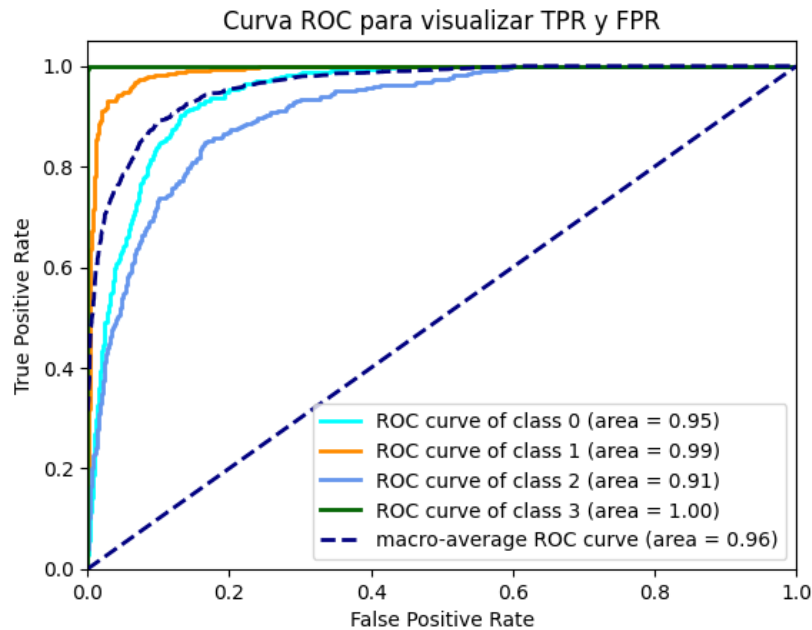


Figura 53: Curva ROC-AOC - Ambos modelos.
Fuente: Elaboración propia.

4.3. Validación página web

Con el objetivo de evaluar el desempeño del modelo en un entorno real, se desarrolló una sencilla aplicación web utilizando Streamlit, un framework de Python que facilita la creación de interfaces web sin la necesidad de conocimientos profundos en HTML, CSS y JavaScript. La idea detrás de Streamlit es permitir desplegar rápidamente una aplicación básica para probar la eficiencia del modelo y compartirlo con otros usuarios, gracias a su opción de despliegue gratuito en la nube.

Antes de construir la aplicación, se combinaron los conjuntos de datos `athletes.csv` y `activities.csv` en un único dataset. Con base en los resultados del algoritmo de selección de características `forward wrapper` aplicado a ambos conjuntos, seleccionamos las siguientes variables para nuestro modelo:

```
features_web = ['age', 'gender', 'activities', 'weightkg', 'frecuencia_semanal', 'workout_time',  
'total_distance', 'elevation_gain', 'average_speed', 'average_power', 'average_hr']
```

Para definir las variables objetivo, se utilizó la potencia máxima en 60 segundos (`60s_peak`)

como indicador de la capacidad anaeróbica y la potencia máxima en 12 minutos (720s_peak) para la capacidad aeróbica. Debido a la combinación de los datasets, los rangos de valores de estas variables cambiaron, por lo que se ajustó la clasificación de manera acorde (véase las tablas 15 y 16). Además, para normalizar los resultados y hacerlos comparables entre diferentes atletas, se calculó la potencia relativa dividiendo la potencia absoluta por el peso del atleta (por ejemplo, 5m peak wpk, 30min peak wpk, etc.). **Al aplicar estas categorizaciones al dataset combinado, obtuvimos una precisión del 94 % para la clasificación de la capacidad anaeróbica y del 86 % para la capacidad aeróbica.**

Tabla 15: Categorización de la capacidad anaeróbica (1min).

Fuente: Elaboración propia.

Capacidad anaeróbica	Rango de potencia (1m_peak_wpk)
Casual	<3.5 WPK
Principiante	3.5 - 4.5 WPK
Experimentado	4.5 - 5.5 WPK
Élite	>5.5 WPK

Tabla 16: Categorización de la capacidad aeróbica (12min).

Fuente: Elaboración propia.

Capacidad aeróbica	Rango de potencia (12m_peak_wpk)
Casual	<2.5 WPK
Principiante	2.5 - 3.5 WPK
Experimentado	3.5 - 4.0 WPK
Élite	>4.0 WPK

Luego de ajustar los datos en este último dataset, se guardaron los modelos utilizando joblib. Lo que hace joblib es guardar el modelo entrenado en un archivo específico con todas las relaciones y patrones que detectó una vez entrenado. Esto es muy útil, ya que simplifica el re-entrenamiento en otros entornos y sirve para validar de igual manera que tan acertada es la predicción si lleváramos el modelo a un entorno de producción.

Es importante mencionar que, si solamente guardamos la predicción del modelo pero aplicamos escalamiento de features, no funcionará, ya que debe ser consistente. Es por esto, que entre el escalamiento y asignación de las variables de prueba y entrenamiento (X e Y) se debe utilizar 'joblib.dump(scaler, 'scaler_anaerobico.joblib')' (por ejemplo) y luego, aplicando XGBoost, se asigna de esta manera: dump(model, 'model_anaerobico.joblib').

- scaler_anaerobico.joblib: Se utiliza para transformar los datos de entrada de un nuevo ciclista para que estén en la misma escala que los datos de entrenamiento.
- model_anaerobico.joblib: Se utiliza para hacer predicciones sobre la capacidad anaeróbica de un nuevo ciclista, utilizando los datos transformados por el scaler.

Teniendo todo esto listo, se disponibiliza el proyecto en github [Yerk0v, 2024]. Aquí se encuentran todos los requerimientos y pasos a seguir, ya que hay que tener mucho cuidado con las versiones que se utilizan. Una vez desplegada la aplicación, se visualiza de la siguiente forma (figuras 54, 55, 56, 57):



Figura 54: Página principal donde se sube el csv o se ingresan los datos manualmente.
Fuente: Elaboración propia.



Figura 55: Carga del conjunto de datos - Aplica el modelo XGBoost entrenado.
Fuente: Elaboración propia.



Figura 56: Resultados predicción en la web.
Fuente: Elaboración propia.

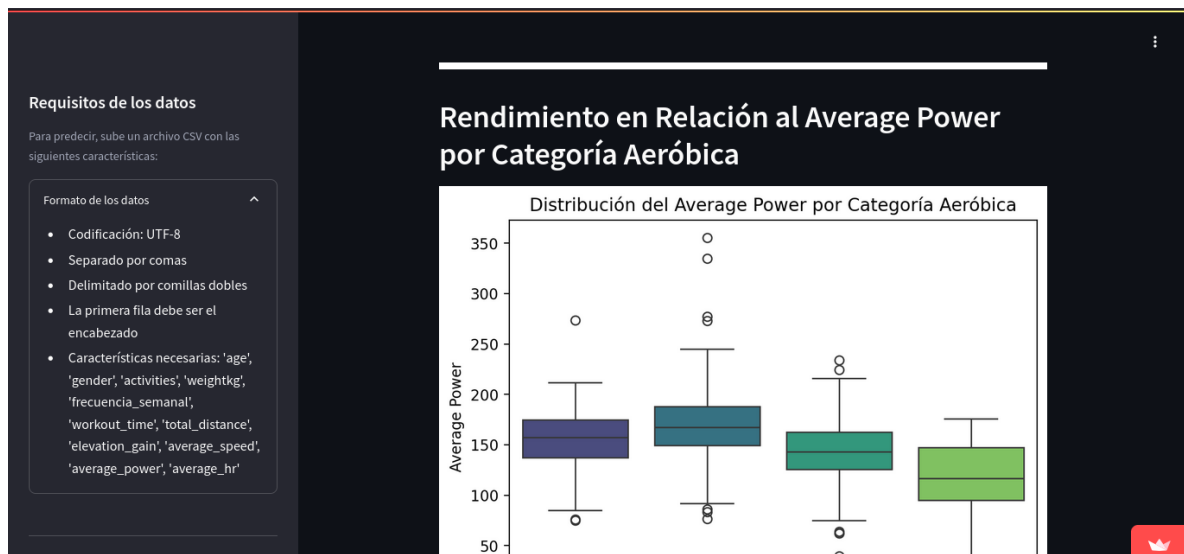


Figura 57: Visualización de resultados.
Fuente: Elaboración propia.

CAPÍTULO 5

CONCLUSIONES

A lo largo de esta memoria, se han explicado los distintos factores que influyen en el rendimiento del ciclista. Desde un principio, se definieron características claves como el Critical Power 2.7, aunque no se incluyó en el aprendizaje del modelo debido a su alta correlación con los intervalos de tiempo seleccionados. Esto se debe a que la Critical Power utiliza la potencia absoluta para calcular su valor. No obstante, se recomienda considerar la integración de la Critical Power cuando se incorporen variables relacionadas con el volumen máximo de oxígeno (VO₂max), ya que podría mejorar significativamente la precisión en la predicción de la capacidad aeróbica y permitir nuevos descubrimientos si se entrena con otras características usando XGBoost. Esto ayudaría a entender mejor el perfil del ciclista en escenarios de resistencia.

5.1. Desempeño de modelos iniciales

En los primeros modelos, donde se utilizaron menos variables para predecir ambas capacidades, XGBoost mostró ser eficaz en la identificación de relaciones no lineales. El análisis de la importancia de las características (feature importance) reveló que el modelo se basaba en gran medida en 'weightkg' y en el intervalo de tiempo más cercano al objetivo, ya que estas eran las variables más influyentes para predecir ambos objetivos.

En particular, en el primer modelo basado en el dataset `athletes.csv`, no solo se validaron y corroboraron los valores establecidos por la tabla de [Jeukendrup *et al.*, 2000], sino que también **se evidenció cómo variables como la edad, la cantidad de actividades y el género contribuyen al rendimiento del ciclista**. Los valores SHAP para ambas capacidades (50) reflejan que la edad y la cantidad de actividades tienen un impacto positivo significativo en las predicciones del modelo.

5.2. Selección de características y resultados finales

La selección de características, realizada mediante la técnica de forward wrapper selection, simplificó la cantidad de variables utilizadas y demostró que la inclusión de información adicional, más allá de la potencia más cercana al intervalo de tiempo objetivo, mejora la clasificación de los ciclistas. Esto se refleja en la tasa VPR-FPR (52), donde los ciclistas fueron correctamente ubicados en sus respectivas categorías.

En la validación del modelo con el dataset `activities.csv`, el proceso de limpieza fue complejo debido a la variedad de nombres en la columna 'sport', lo que dificultó la comprensión

y filtrado de las actividades relevantes. Sin embargo, después de la limpieza y el resumen de los datos de los atletas, el panorama resultó favorable. Contar con un conjunto de datos más rico en características del entrenamiento realizado por el ciclista aportó información valiosa, alineada con los objetivos específicos planteados (1.3.3).

Las mejores características identificadas para predecir la capacidad aeróbica y anaeróbica se visualizan en la siguiente tabla 17:

Tabla 17: Características más importantes para la predicción de capacidades.

Fuente: Elaboración propia.

Capacidad	Características	Precisión
Aeróbica	average_power, age, total_distance, average_hr, frecuencia_semanal, average_speed, elevation_gain, total_actividades, duracion_entrenamiento, workout_time, total_work	87-89 %
Anaeróbica	average_power, elevation_gain, age, total_distance, average_hr, total_actividades, average_speed, workout_time, frecuencia_semanal, duracion_entrenamiento	86-91 %

El análisis del «feature importance» para ambos modelos refuerza la relevancia de las variables identificadas, alineándose con lo documentado en la literatura científica. **Las variables como la edad, potencia media y ganancia de elevación resultaron ser esenciales para ambas capacidades, destacando la influencia de la madurez física, la capacidad para generar potencia explosiva, y la exigencia de entrenamientos en subidas.**

5.3. Página web y trabajos futuros

Esta investigación ha sentado las bases para un sistema de evaluación del rendimiento ciclista altamente preciso. Al combinar técnicas de selección de características y aprendizaje automático, se ha desarrollado un modelo capaz de predecir con alta fiabilidad las capacidades aeróbicas y anaeróbicas de un ciclista. La implementación de una aplicación web demuestra el potencial de esta herramienta para ser utilizada por entrenadores y atletas. Futuras investigaciones podrían explorar la incorporación de datos más detallados sobre la experiencia del ciclista y su fisiología para optimizar aún más el modelo y ofrecer recomendaciones personalizadas de entrenamiento.

REFERENCIAS BIBLIOGRÁFICAS

- [Adobe Stock, 2024] Adobe Stock (2024). Types of machine learning with algorithms, classification outline diagram. Online Image. Accessed: 2024-06-28.
- [Ahlquist y Bassett, 1992] Ahlquist, L. y Bassett, D.R. and Sufit, R. e. a. (1992). The effect of pedaling frequency on glycogen depletion rates in type i and type ii quadriceps muscle fibers during submaximal cycling exercise. *European Journal of Applied Physiology and Occupational Physiology*, 65(5):360–364.
- [Arberas, 2017] Arberas, A. (2017). La potencia absoluta: Más allá del kilovatio. <https://www.zikloland.com/la-potencia-absoluta-mas-alla-del-kilovatio/>. Una colaboración de Aritz Arberas, Preparador físico del Astana, en Preparación Física.
- [Arlot y Celisse, 2010] Arlot, S. y Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics surveys*, 4:40–79.
- [AusCycling, 2024] AusCycling (2024). Age categories racing. *AusCycling News*. Accedido el 2 de julio de 2024.
- [Balmer et al., 2000] Balmer, J., Davison, R. C., y Bird, S. R. (2000). Peak power predicts performance power during an outdoor 16.1-km cycling time trial. *Medicine and science in sports and exercise*, 32(8):1485–1490.
- [Bangsbo, 1996] Bangsbo, J. (1996). Quantification of anaerobic energy production during intense exercise. *Medicine and Science in Sports and Exercise*, 28:116–123.
- [Barboza et al., 2017] Barboza, J. A., Nogueira, F. S., y Pompeu, F. A. (2017). A new accurate model to predict maximal heart rate. *Journal of Exercise Physiology Online*, 20:23–28.
- [Baron, 2001] Baron, R. (2001). Aerobic and anaerobic power characteristics of off-road cyclists. *Medicine and science in sports and exercise*, 33(8):1387–1393.
- [Bassett y Howley, 2000] Bassett, D. R., J. y Howley, E. T. (2000). Limiting factors for maximum oxygen uptake and determinants of endurance performance. *Medicine and Science in Sports and Exercise*, 32(1):70–84.
- [Bentley et al., 1998] Bentley, D. J., Wilson, G. J., Davie, A. J., y Zhou, S. (1998). Correlations between peak power output, muscular strength and cycle time trial performance in triathletes. *The Journal of sports medicine and physical fitness*, 38(3):201–207.
- [Breiman, 2001a] Breiman, L. (2001a). Random forests. *Machine learning*, 45(1):5–32.
- [Breiman, 2001b] Breiman, L. (2001b). *Statistical modeling: The two cultures (with comments and a rejoinder by the author)*, volumen 16. JSTOR.

- [Breiman *et al.*, 1984] Breiman, L., Friedman, J., Stone, C., y Olshen, R. (1984). *Classification and Regression Trees*. Taylor & Francis.
- [Burke, 1997] Burke, L. (1997). Nutrition for post-exercise recovery. *Australian journal of science and medicine in sport*, 29(1):3–10.
- [Chapman *et al.*, 2000] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., y Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. *CRISP-DM Consortium*.
- [Chen y Guestrin, 2016] Chen, T. y Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *CoRR*, abs/1603.02754.
- [Coggan, 2003] Coggan, A. R. (2003). Training and racing using a power meter: an introduction. En *Training and racing using a power meter*.
- [Coyle, 1995] Coyle, E. F. (1995). Integration of sport and exercise training, nutrition, and recovery. *Exercise and Sport Sciences Reviews*, 23:165–181.
- [Coyle, 1999a] Coyle, E. F. (1999a). Physiological determinants of endurance exercise performance. *Journal of science and medicine in sport*, 2(3):181–189.
- [Coyle, 1999b] Coyle, E. F. (1999b). Physiological determinants of endurance exercise performance. *Journal of Science and Medicine in Sport*, 2(3):181–189.
- [Dieguez *et al.*, 2016] Dieguez, T., Conceição, O., y Fernandes, A. (2016). *The Challenges of Managing the Private Institutions of Social Solidarity: Experiences from Portugal*, pp. 23–54. IGI Global.
- [Faria *et al.*, 2005a] Faria, E., Parker, D., y Faria, I. (2005a). The science of cycling: Factors affecting performance part 2. *Sports medicine (Auckland, N.Z.)*, 35:313–37.
- [Faria *et al.*, 2005b] Faria, E., Parker, D., y Faria, I. (2005b). The science of cycling: physiology and training - part 1. *Sports medicine (Auckland, N.Z.)*, 35:285–312.
- [Faria, 1984] Faria, I. (1984). Applied physiology of cycling. *Sports medicine (Auckland, N.Z.)*, 1:187–204.
- [Faria y Cavanagh, 1978] Faria, I. y Cavanagh, P. (1978). *The physiology and biomechanics of cycling*. John Wiley and Sons, New York.
- [Fawcett, 2006] Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874.
- [Febbraio *et al.*, 1994] Febbraio, M. A., Snow, R. J., Hargreaves, M., Stathis, C. G., Martin, I. K., y Carey, M. F. (1994). Title: Effect of carbohydrate ingestion on metabolism during running and cycling. *Journal of Applied Physiology*, 76:2559–2563.

- [Fernández Rodríguez *et al.*, 2021] Fernández Rodríguez, L., Olaya-Cuartero, J., y Martínez-Sanz, J. M. (2021). Planificación dietético-nutricional para una prueba de mountain bike de xco: estudio de caso. *Retos*, 44:209–220.
- [Fernández-Delgado *et al.*, 2014] Fernández-Delgado, M., Cernadas, E., Barro, S., y Amorim, D. (2014). Weka: A machine learning workbench. *Journal of Machine Learning Research*, 15:3233–3234.
- [Fox *et al.*, 1971] Fox, S. M. I., Naughton, J. P., y Haskell, W. L. (1971). Physical activity and the prevention of coronary heart disease. *Annals of Clinical Research*, 3:404–432.
- [González-Alonso *et al.*, 1995] González-Alonso, J., Mora-Rodríguez, R., Below, P. R., Coyle, E. F., y Wenger, C. B. (1995). Title: The effects of carbohydrate ingestion on the glycemic response to a high-intensity exercise. *Journal of Applied Physiology*, 80:849–855.
- [Guyon y Elisseeff, 2003] Guyon, I. y Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182.
- [Hall, 2000] Hall, M. A. (2000). Correlation-based feature selection of discrete and numeric class machine learning. En *ICML*.
- [Hawley y Noakes, 1992] Hawley, J. A. y Noakes, T. D. (1992). Peak power output predicts maximal oxygen uptake and performance time in trained cyclists. *European journal of applied physiology and occupational physiology*, 65(1):79–83.
- [Hill, 1925] Hill, A. (1925). The physiological basis of athletic records. *Nature*, 116:544–548.
- [Hu *et al.*, 2020] Hu, J., Niu, H., Carrasco, J., Lennox, B., y Arvin, F. (2020). Voronoi-based multi-robot autonomous exploration in unknown environments via deep reinforcement learning. *IEEE Transactions on Vehicular Technology*, 69(12):14413–14423.
- [Impellizzeri y Marcora, 2007] Impellizzeri, F. y Marcora, S. (2007). The physiology of mountain biking. *Sports Medicine*, 37(1):59–71.
- [James *et al.*, 2013] James, G., Witten, D., Hastie, T., y Tibshirani, R. (2013). *An introduction to statistical learning*, volumen 112. Springer.
- [Jeukendrup *et al.*, 2000] Jeukendrup, A. E., Craig, N. P., y Hawley, J. A. (2000). The bioenergetics of world class cycling. *Journal of Science and Medicine in Sport*, 3(4):414–433.
- [Jones y Vanhatalo, 2017] Jones, A. M. y Vanhatalo, A. (2017). The 'critical power' concept: Applications to sports performance with a focus on intermittent high-intensity exercise. *Sports medicine (Auckland, N.Z.)*, 47(Suppl 1):65–78.
- [Kalamen, 1968] Kalamen, J. (1968). *Measurement of Maximum Muscle Power in Man*. Tesis doctoral, The Ohio State University, Columbus, Ohio.

- [Kholkine *et al.*, 2020] Kholkine, L., De Schepper, T., Verdonck, T., y Latré, S. (2020). A machine learning approach for road cycling race performance prediction. En Brefeld, U., Davis, J., Van Haaren, J., y Zimmermann, A., editores, *Machine Learning and Data Mining for Sports Analytics*, pp. 103–112, Cham. Springer International Publishing.
- [Kohavi y John, 1997] Kohavi, R. y John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, 97:273–324.
- [Kohavi y Provost, 2001] Kohavi, R. y Provost, F. (2001). Data mining and knowledge discovery handbook. En *Data mining and knowledge discovery handbook*. Springer.
- [Koller y Sahami, 1996] Koller, D. y Sahami, M. (1996). Toward optimal feature selection. En *ICML*, pp. 284–292.
- [Latin *et al.*, 1993] Latin, R. W., Berg, K. E., Smith, P., y Others (1993). Validation of a cycle ergometry equation for predicting steady-rate vo₂. *Med Sci Sports Exerc*, 25:970–974.
- [Liversedge, 2021] Liversedge, M. (2021). Goldencheetah opendata project. without notes.
- [Louis *et al.*, 2013] Louis, J., Billaut, F., Bernad, T., Vettoretti, F., Hausswirth, C., y Brisswalter, J. (2013). Physiological demands of a simulated bmx competition. *International journal of sports medicine*, 34(6):491–496.
- [Lucia *et al.*, 1999a] Lucia, A., Hoyos, J., Carvajal, A., y Others (1999a). Heart rate response to professional road cycling: The tour de france. *International Journal of Sports Medicine*, 20:167–172.
- [Lucia *et al.*, 1999b] Lucia, A., Hoyos, J., y Chicharro, J. L. (1999b). Preferred pedaling cadence in professional cycling. *Medicine & Science in Sports & Exercise*, 31:442–447.
- [Lucia *et al.*, 2000] Lucia, A., Joyos, H., y Chicharro, J. L. (2000). Physiological response to professional road cycling: climbers vs. time trialists. *International Journal of Sports Medicine*, 21:505–512.
- [Lucia *et al.*, 2004] Lucia, A., San Juan, A. F., Montilla, M., CaNete, S., Santalla, A., Earnest, C., y Pérez, M. (2004). In professional road cyclists, low pedaling cadences are less efficient. *Medicine and Science in Sports and Exercise*, 36(6):1048–1054.
- [Lucía *et al.*, 2001a] Lucía, A., Hoyos, J., y Chicharro, J. L. (2001a). Preferred pedalling cadence in professional cycling. *Medicine and Science in Sports and Exercise*, 33(8):1361–1366.
- [Lucía *et al.*, 2001b] Lucía, A., Hoyos, J., y Chicharro, J. L. (2001b). Preferred pedalling cadence in professional cycling. *Medicine and science in sports and exercise*, 33(8):1361–1366.
- [Lucía *et al.*, 2002] Lucía, A., Hoyos, J., Santalla, A., y Others (2002). Kinetics of vo₂ in professional cyclists. *Med Sci Sports Exerc*, 34:320–325.
- [Lundberg y Lee, 2017] Lundberg, S. M. y Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

- [McCole *et al.*, 1990] McCole, S. D., Claney, K., Conte, J. C., y Others (1990). Energy expenditure during bicycling. *J Appl Physiol*, 68:748–753.
- [McKay *et al.*, 2022] McKay, A., Stellingwerff, T., Smith, E., Martin, D., Mujika, I., Goosey-Tolfrey, V., Sheppard, J., y Burke, L. (2022). Defining training and performance caliber: A participant classification framework. *International journal of sports physiology and performance*, 17(2):317–331.
- [Menaspà *et al.*, 2012] Menaspà, P., Rampinini, E., Bosio, A., Carlomagno, D., Riggio, M., y Sassi, A. (2012). Physiological and anthropometric characteristics of junior cyclists of different specialties and performance levels. *Scandinavian journal of medicine & science in sports*, 22(3):392–398.
- [Menaspà *et al.*, 2010] Menaspà, P., Sassi, A., e Impellizzeri, F. M. (2010). Aerobic fitness variables do not predict the professional career of young cyclists. *Medicine and science in sports and exercise*, 42(4):805–812.
- [Merriam-Webster, nd] Merriam-Webster (n.d.). Physiology. Retrieved May 14, 2024 from <https://www.merriam-webster.com/dictionary/physiology>.
- [Monod y Scherrer, 1965] Monod, H. y Scherrer, J. (1965). The work capacity of a synergic muscular group. *Ergonomics*, 8:329–338.
- [Moritani *et al.*, 1981] Moritani, T., Nagata, A., deVries, H., y Muro, M. (1981). Critical power as a measure of physical work capacity and anaerobic threshold. *Ergonomics*, 24(5):339–350.
- [Morton, 1996] Morton, R. H. (1996). A 3-parameter critical power model. *Ergonomics*, 39:611–619.
- [Novak y Dascombe, 2014] Novak, A. y Dascombe, B. (2014). Physiological and performance characteristics of road, mountain bike and bmx cyclists. *Journal of Science and Cycling*, 3(3).
- [Padilla *et al.*, 1999] Padilla, S., Mujika, I., Cuesta, G., Goiriena, J., y Guevara, J. (1999). Level ground and uphill cycling ability in professional road cycling. *Medicine & Science in Sports & Exercise*, 31(6):878–885.
- [Padilla *et al.*, 2001] Padilla, S., Mujika, I., Orbañanos, J., Santisteban, J., Angulo, F., y José Goiriena, J. (2001). Exercise intensity and load during mass-start stage races in professional road cycling. *Medicine and science in sports and exercise*, 33(5):796–802.
- [Palmer *et al.*, 1994] Palmer, G. S., Hawley, J. A., Dennis, S. C., y Noakes, T. D. (1994). Heart rate responses during a 4-d cycle stage race. *Medicine and science in sports and exercise*, 26(10):1278–1283.
- [Passfield *et al.*, 2017] Passfield, L., Hopker, J., Jobson, S., Friel, D., y Zabala, M. (2017). Knowledge is power: Issues of measuring training and performance in cycling. *Journal of Sports Sciences*, 35(14):1426–1434.

- [Paul William Macdermid y Stannard, 2015] Paul William Macdermid, P. W. F. y Stannard, S. R. (2015). The influence of tyre characteristics on measures of rolling performance during cross-country mountain biking. *Journal of Sports Sciences*, 33(3):277–285.
- [Powers, 2020] Powers, D. M. (2020). Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*.
- [Provo et al., 2003] Provo, F., Fawcett, T., y Kohavi, R. (2003). Hypothesis testing: A primer. *Data Mining and Knowledge Discovery Handbook*, pp. 845–870.
- [Pudjihartono et al., 2022] Pudjihartono, N., Fadason, T., Kempa-Liehr, A. W., y O’Sullivan, J. M. (2022). A review of feature selection methods for machine learning-based disease risk prediction. *Frontiers in Bioinformatics*, 2.
- [Real Academia Española, 2024] Real Academia Española (2024). Diccionario de la lengua española. Versión 23.7 en línea.
- [Remeseiro y Bol’on-Canedo, 2019] Remeseiro, B. y Bol’on-Canedo, V. (2019). A review of feature selection methods in medical applications. *Computers in Biology and Medicine*, 112:103375.
- [Robergs et al., 2004] Robergs, R. A., Ghiasvand, F., y Parker, D. (2004). Biochemistry of exercise-induced metabolic acidosis. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, 287(3):R502–R516.
- [Roy, 2020] Roy, A. (2020). Chapter 1: Introduction to crisp-dm framework in data science. *LinkedIn*. [Artículo en línea].
- [Seiler y Tønnessen, 2009] Seiler, S. y Tønnessen, E. (2009). Intervals, thresholds, and long slow distance: the role of intensity and duration in endurance training. *SPORTSCIENCE · sportsci.org*, 13:32–53.
- [Shmueli, 2010] Shmueli, G. (2010). *Explain data analysis*. John Wiley & Sons.
- [Shwartz-Ziv y Armon, 2023] Shwartz-Ziv, R. y Armon, A. (2023). Tabular data: Deep learning is not all you need. *Information Fusion*, 81:84–90.
- [Silveira y Rama, 2020] Silveira, A. y Rama, L. (2020). The influence of different track surfaces and bicycles on mountain biking performance. *International Journal of Performance Analysis in Sport*, 20(6):1052–1060.
- [Smith, 2023] Smith, R. (2023). Cycling statistics.
- [Statista, 2023] Statista (2023). Number of bicycles in china from 2014 to 2021.
- [Strava, 2020] Strava (2020). Strava releases year in sport trend report. <https://press.strava.com/articles/strava-releases-year-in-sport-trend-report>.

- [Takaishi *et al.*, 1994] Takaishi, T., Yasuda, Y., y Moritani, T. (1994). Neuromuscular fatigue during prolonged pedaling rates. *European Journal of Applied Physiology and Occupational Physiology*, 69:154–158.
- [Takaishi *et al.*, 1996] Takaishi, T., Yasuda, Y., Ono, T., y Moritani, T. (1996). Optimal pedaling rate estimated from neuromuscular fatigue for cyclists. *Medicine and Science in Sports and Exercise*, 28(12):1492–1497.
- [Tanaka *et al.*, 2001] Tanaka, H., Monahan, K. D., y Seals, D. R. (2001). Age-predicted maximal heart rate revisited. *Journal of the American College of Cardiology*, 37(1):153–156.
- [Tharwat, 2020] Tharwat, A. (2020). Classification assessment methods. *Applied Computing and Informatics*, 16(1):168–192.
- [Tukey, 1962] Tukey, J. W. (1962). The future of data analysis. *The annals of mathematical statistics*, pp. 1–67.
- [V *et al.*, 2020] V, S., S, N., y M, P. (2020). An empirical science research on bioinformatics in machine learning. *Journal of Mechanics of Continua and Mathematical Sciences*, 7.
- [Vandewalle *et al.*, 1987] Vandewalle, H., Peres, G., Heller, J., Panel, J., y Monod, H. (1987). Force-velocity relationship and maximal power on a cycle ergometer. correlation with the height of a vertical jump. *European Journal of Applied Physiology and Occupational Physiology*, 56(6):650–656.
- [Viivo, 2017] Viivo, J. (2017). Informe: Técnica de pedaleo, mejora tu técnica mientras entrenas en el gym. <https://cicloturismovivo.com/2017/09/14/informe-tecnica-de-pedaleo-mejora-tu-tecnica-mientras-entrenas-en-el-gym/>. CicloTurismoVivo.
- [Wagner, 2000] Wagner, P. D. (2000). New ideas on limitations to vo2max. *Exercise and Sport Sciences Reviews*, 28(1):10–14.
- [Wilkerson *et al.*, 2004] Wilkerson, D., Koppo, K., Barstow, T., Kim, C., Clanton, T., y Gurd, B. (2004). Effect of work rate on the functional 'gain' of phase ii pulmonary o2 uptake response to exercise. *Respiratory physiology & neurobiology*, 142(2):211–223.
- [Yerk0v, 2024] Yerk0v (2024). cyclists-clasification. <https://github.com/Yerk0v/cyclists-clasification>.
- [Yoosefzadeh-Najafabadi *et al.*, 2021] Yoosefzadeh-Najafabadi, M., Hugh, E., Tulpan, D., Sulik, J., y Eskandari, M. (2021). Application of machine learning algorithms in plant breeding: Predicting yield from hyperspectral reflectance in soybean? *Front. Plant Sci.*, 11:624273.
- [Zapier, 2023] Zapier (2023). 5 data analysis examples to help you make better decisions. <https://zapier.com/blog/data-analysis-example/>. Accedido el 30 de junio de 2024.

[Zhang *et al.*, 2017] Zhang, S., Li, X., Zong, M., Zhu, X., y Wang, R. (2017). An improved k-nearest neighbor algorithm for text categorization. *arXiv preprint arXiv:1701.08366*.