



TRABAJO DE FIN DE GRADO

“Análisis de datos en ciclismo para el control del
rendimiento durante la temporada.”

AUTOR: AARON MULVEY IZQUIERDO

DIRECTOR: JAVIER ORBAÑANOS PALACIOS

GRADO EN CIENCIAS DE LA ACTIVIDAD FÍSICA Y DEL DEPORTE

CURSO ACADÉMICO: 2020 - 2021

CONVOCATORIA: JUNIO 2021

INDICE

1	Resumen.	7
1.1	Objetivos.	7
1.2	Problemas asociados y metodología del trabajo desarrollada.	7
2	Marco teórico.	8
2.1	Ciclismo en ruta.	8
2.2	Factores limitantes del rendimiento en ciclismo.	8
2.3	Control de la evolución del ciclista durante la temporada.	11
2.3.1	Test de potencia (Power Profile Test).	12
2.3.2	Test de lactato.	13
2.3.3	Test de Resistencia a la fatiga	14
2.3.4	Eficiencia.	14
2.3.5	VO ₂ Score.	15
3	Big data.	16
3.1	El Big data en el deporte.	16
3.2	El Big data en el ciclismo.	17
3.3	Ciencia de datos.	17
3.3.1	Limpieza de datos.	18
3.3.2	Análisis exploratorio.	18
4	Diseño del estudio y resultados	20
4.1	Obtención de la información.	20

4.2	Preprocesado de los datos.	20
4.2.1	Dispositivos para la obtención de los datos.	20
4.2.2	Formato de los datos.	20
4.2.3	Lenguajes de programación Python y R.	21
4.2.4	Conversión del formato de los archivos.	21
4.3	Limpieza de datos.	22
4.3.1	Análisis de las variables.	22
4.3.2	Análisis Outliers de la Frecuencia Cardiaca (FC).	23
4.3.3	Análisis Outliers de la Potencia.	27
4.3.4	Análisis de la Missing data.	28
4.3.5	Transformación de datos, creación de nuevas variables.	33
4.4	Evolución de los indicadores de rendimiento durante la temporada.	40
4.4.1	Perfil de potencia.	40
4.4.2	Test de lactato.	44
4.4.3	Test de resistencia a la fatiga	45
4.4.4	Control del entrenamiento mediante la eficiencia y el $VO_{2\text{ Score}}$	46
4.5	Modelos para la valoración y predicción de rendimiento.	46
4.5.1	Modelos de influencia del volumen y la intensidad en el rendimiento.	47
4.5.2	Relación de la eficiencia y del $VO_{2\text{ Score}}$ con el FTP.	49
5	Conclusiones.	52
6	Bibliografía	53

TABLA DE ILUSTRACIONES:

<i>Ilustración 1. Esquema general de los factores limitantes del rendimiento. (Joyner & Coyle, 2008)</i>	<i>11</i>
<i>Ilustración 2 Tabla test de potencia. (Coggan, 2006).</i>	<i>13</i>
<i>Ilustración 3 Wickham, H., & Grolemund, G. (2017). R for Data Science</i>	<i>18</i>
<i>Ilustración 4 Tabla resúmenes estadísticos. Elaboración propia.</i>	<i>23</i>
<i>Ilustración 5 Boxplot Outliers Frecuencia Cardiaca. Elaboración propia.</i>	<i>24</i>
<i>Ilustración 6 Distribución de frecuencias de la FC, ciclista 1. Elaboración propia.</i>	<i>25</i>
<i>Ilustración 7 Distribucion de frecuencias de la FC, ciclista 2. Elaboración propia.</i>	<i>25</i>
<i>Ilustración 8 Boxplot distribución de la FC sin Outliers. Elaboración propia.</i>	<i>26</i>
<i>Ilustración 9 Frecuencia de la FC sin Outliers. Elaboración propia.</i>	<i>27</i>
<i>Ilustración 10 Ilustración 6 Boxplot distribución Potencia sin Outliers. Elaboración propia.</i>	<i>28</i>
<i>Ilustración 11 Combinaciones de patrones de Missing data. Elaboración propia.</i>	<i>29</i>
<i>Ilustración 12 Entrenamiento sin potenciómetro. Fuente: Training Peaks.</i>	<i>30</i>
<i>Ilustración 13 Frecuencia cardiaca exportada de Training Peaks. Elaboración propia.</i>	<i>31</i>
<i>Ilustración 14 Potencia Imputada con suavizado $k = 50$. Elaboración propia.</i>	<i>32</i>
<i>Ilustración 15 Potencia Imputada real. Elaboración propia.</i>	<i>32</i>
<i>Ilustración 16 Visualización de la potencia normalizada. Elaboración propia.</i>	<i>34</i>
<i>Ilustración 17 Comportamiento de variables durante el entrenamiento. Elaboración propia.</i>	<i>36</i>
<i>Ilustración 18 Comportamiento de las variables durante una serie de velocidad. Elaboración propia.</i>	<i>36</i>
<i>Ilustración 19 Control de carga Training Peaks. Fuente Training Peaks.</i>	<i>39</i>
<i>Ilustración 20 Control de carga con datos de potencia diarios. Elaboración propia.</i>	<i>39</i>
<i>Ilustración 21 Test de 5" desarrollo de potencia y torque. Elaboración propia.</i>	<i>40</i>
<i>Ilustración 22 Test de potencia 1' . Elaboración propia.</i>	<i>41</i>

<i>Ilustración 23 Test de 5'. Elaboración Propia.....</i>	<i>42</i>
<i>Ilustración 24 Test de potencia de 20'. Elaboración propia.....</i>	<i>42</i>
<i>Ilustración 25 Resultados relativos y absolutos test de potencia. Elaboración propia.....</i>	<i>43</i>
<i>Ilustración 26. Perfil de potencia de las ciclistas. Elaboración propia.</i>	<i>44</i>
<i>Ilustración 27 Desarrollo de potencia durante Test de Lactato. Elaboración propia.....</i>	<i>45</i>
<i>Ilustración 30 Evolución diaria de la eficiencia y el VO2 Score. Elaboración propia.....</i>	<i>46</i>
<i>Ilustración 31 Relación de variables con el FTP. Elaboración propia.....</i>	<i>49</i>
<i>Ilustración 32 Arquitectura de la red neuronal. Elaboración propia.</i>	<i>50</i>
<i>Ilustración 33 Predicciones de la eficiencia. Elaboración propia.....</i>	<i>51</i>
<i>Ilustración 34 Predicciones del VO₂ Score. Elaboración propia.....</i>	<i>52</i>

1 Resumen.

El presente trabajo se trata de un estudio hecho con ciclistas femeninas, realizando un análisis de grandes conjuntos de datos de estas. El trabajo inicia con una exploración de datos desestructurados de las ciclistas, y mediante el proceso de la ciencia de datos se estructuran los datos para obtener un control del entrenamiento de las ciclistas. Para finalizar, se evalúa el rendimiento y se crean modelos de predicción de este.

1.1 Objetivos.

El objetivo principal de este trabajo se basa en la explotación de grandes conjuntos de datos para el control y seguimiento de la temporada de un grupo de ciclistas femeninas. Para ello, se recopilaron datos durante un periodo de la temporada y se observó la evolución de los parámetros claves para el rendimiento en el ciclismo y se establecen predicciones de rendimiento sobre los mismos.

1.2 Problemas asociados y metodología del trabajo desarrollada.

La recopilación de datos de la actividad de las deportistas se realizó mediante ciclo computadores y potenciómetros. Además, se hicieron varios test a las ciclistas durante la temporada para evaluar su rendimiento.

El principal problema de la recopilación de datos durante los entrenamientos fue que los archivos no se podían leer con un software común por lo que para poder realizar un análisis sobre estos, fue necesario un tratamiento de grandes conjuntos de datos en un lenguaje de programación usado para la ciencia de datos.

2 Marco teórico.

2.1 Ciclismo en ruta.

El ciclismo de competición en ruta es un deporte de imputación individual realizado por equipos con el fin de cruzar la meta del recorrido establecido en primer lugar. El rendimiento en el ciclismo depende de factores fisiológicos, que influyen en la producción de energía mecánica, factores ambientales, que afectan las demandas energéticas de los ciclistas, diseño de la bicicleta, factores nutricionales, estrategias de carrera y psicología de las deportistas.

Las pruebas de esta disciplina tienen una duración que oscila entre 1 y 2 horas por lo que la cualidad que más importancia cobra en este deporte es la resistencia, más específicamente, la resistencia aeróbica (Mujica y Padilla, 2001).

2.2 Factores limitantes del rendimiento en ciclismo.

Para llegar al máximo rendimiento deportivo, todos los deportes tienen diferentes variables que pueden marcar la diferencia entre una medalla de oro y una medalla de plata. A estas variables se les llaman factores limitantes del rendimiento, y cobran importancia en los deportistas de elite.

La antropometría de las personas varía mucho entre los diferentes deportes, pudiendo otorgar a los deportistas una gran ventaja competitiva frente al resto en caso de tener unos datos antropométricos óptimos para su deporte. En ciclismo, un estudio realizado por Impellizzeri et al. (2008) con ciclistas profesionales femeninas con una edad comprendida entre los 28 ± 5 años, se observó que la estatura media fue de 167 ± 4 centímetros, el peso medio se situó entre los $51,8 \pm 3,4$ kilogramos y el índice de masa corporal (IMC) rondó los $18,7 \text{ kg/m}^2$. Se puede concluir, que las ciclistas profesionales son de estatura media y

de bajo peso debido a un bajo porcentaje graso, esto hace que la masa que tienen que movilizar sobre la bicicleta sea menor, y por ello, a misma potencia en términos absolutos, la potencia relativa sea mayor, haciendo que las deportistas sean más eficientes y eficaces en carrera.

El ciclismo, al ser un deporte que requiere un perfil fisiológico mayoritariamente de resistencia, la principal fuente de energía se obtiene de la ruta de oxidación de ácidos grasos y la glucólisis aeróbica.

Por lo que uno de los principales factores que limita el rendimiento en ciclismo en cuanto a los parámetros fisiológicos es el consumo máximo de oxígeno, que es la variable más empleada para clasificar a los ciclistas.

En un estudio de Chicharro (1999), se evaluó el consumo máximo de oxígeno de ciclistas de elite varones y se obtuvo un valor de 72 ml/min/kg. Dos años después en una investigación similar Lucia (2001) obtuvo como resultado un valor de 72 ml/min/kg en ciclistas de elite.

Otras variables que afectan al rendimiento de los ciclistas son la potencia máxima y la potencia máxima relativa que son capaces de generar durante el pedaleo. En una investigación de Lucia (2002), se observaron datos de potencia máxima en ciclistas de elite de 522 Watts y una potencia relativa de 7,2 Watt/kg. Un año antes, Mujica (2001) también en un estudio con ciclistas de elite obtuvo valores de potencia máxima de 434 vatios y una potencia relativa de 6,4 w/kg.

En cuanto al porcentaje del consumo máximo de oxígeno, en el umbral de lactato, Astrand (1969) en un estudio con 24 ciclistas profesionales observó que el LT se situaba en una media del 77% del VO_{2max} . En cuanto al el segundo umbral respiratorio o el OBLA se

situó en una media del 88% del $\text{VO}_{2\text{max}}$. En la misma investigación se observaron valores máximos de lactato en sangre de 9,9 mmol/l y frecuencias cardíacas máximas de 194 rpm.

En ciclismo el uso de sistemas anaeróbicos para la producción de energía es mínimo, se puede observar cómo los cambios drásticos de velocidad de corta duración en carrera son alimentados por sistemas de producción de energía aláctica gracias a la división de las moléculas de PCr para la obtención de ATP. Mientras se sostienen estos altos esfuerzos en el tiempo, la provisión de ATP es formada a través de la formación de lactato. Este proceso genera un cambio de acidez en el cuerpo por lo que la capacidad de amortiguación del cuerpo a altos niveles de acidez juega un papel crucial en la fatiga resultante del ciclista (Hopker & Jobson, 2012, p.5).

Por último, es remarcable que el $\text{VO}_{2\text{max}}$, el lactato u otras respuestas fisiológicas no son las que ganan las carreras, sin embargo, la aplicación de trabajo mecánico o potencia aplicada juega un papel fundamental en el rendimiento de los ciclistas. Esta, es la capacidad de convertir la energía disponible en trabajo mecánico (Hopker & Jobson, 2012, p.17).

El siguiente gráfico de Coyle (2008), resume los factores limitantes del rendimiento fisiológicos mencionados para los deportes cíclicos de resistencia.

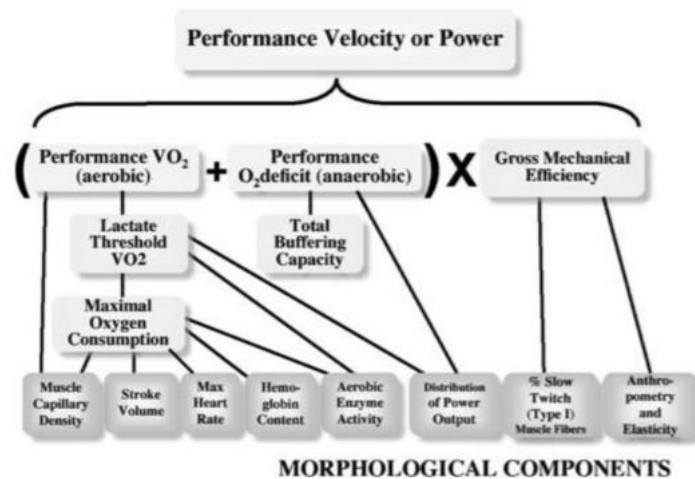


Ilustración 1. Esquema general de los factores limitantes del rendimiento. (Joyner & Coyle, 2008)

2.3 Control de la evolución del ciclista durante la temporada.

El entrenamiento en ciclismo exige la realización de grandes cargas de entrenamiento y una óptima distribución de las diferentes intensidades de trabajo para la mejor preparación de los ciclistas. Todo proceso de entrenamiento es una sucesión de cargas de entrenamiento y períodos de recuperación, que, si presentan una buena coordinación, dan como resultado una mejora progresiva del rendimiento.

La determinación de una serie de variables que permitan valorar la evolución del estado de las ciclistas a través de datos extraídos del entrenamiento y a través de diferentes tests permiten establecer un diagnóstico sobre el estado de forma, y aplicar las cargas de trabajo adecuadas para mejorar el rendimiento, además de observar el rendimiento actual del atleta.

En la realización del presente trabajo, tras la revisión bibliográfica realizada sobre indicadores de rendimiento en ciclismo, y en colaboración con el preparador de las ciclistas participantes en el mismo se establecieron los siguientes indicadores extraídos

tanto de los datos del entrenamiento (eficiencia, Vo2score) como de diferentes pruebas de valoración a lo largo del periodo de la temporada analizado:

2.3.1 Test de potencia (Power Profile Test).

Durante la realización de las carreras de ciclismo existe una amplia variedad de situaciones que requieren demandas energéticas y neuromusculares diferentes. Una alta potencia de 5 segundos tiene una gran importancia en un sprint exitoso. En el otro extremo un umbral de potencia funcional (FTP) alto, es un requisito indispensable en las carreras que requieren largos esfuerzos sostenidos (Hopker & Jobson, 2012, p.55).

En el trabajo de Coyle et al. A finales de 1980 y principios de 1990, se demostró que una prueba en el laboratorio de 1 hora estaba altamente correlacionada con la prueba contrarreloj de 40 km en carretera, también se demostró que la potencia máxima que se podía mantener durante 1 hora estaba correlacionada con los diferentes umbrales, a esto Andrew Coggan lo nombró potencia de umbral funcional (FTP).

Para la medición del FTP, se realiza un test de potencia de 20 minutos contrarreloj, tras ello se saca el promedio de potencia y se resta el 5% de la potencia obtenida, ya que el resultado del test de 20 minutos es el 105% del FTP.

Allen y Coggan describieron la prueba del perfil de potencia como la evaluación del rendimiento máximo en 5 segundos 1 minuto y 5 minutos, tras un calentamiento de 45 minutos. Para comparar los resultados obtenidos se divide el resultado de la potencia entre el peso, obteniendo así el perfil de potencia relativa. Andrew Coggan (2006) elaboró una tabla orientativa para observar el resultado de potencia obtenidos y evaluar el rendimiento comparándolo con los de otros ciclistas.

En nuestro estudio hemos determinado la potencia producida en 5 y 12 segundos, como parámetros asociados a factores neuromusculares, 30 segundos y 1 minuto, como parámetros asociados a la producción glucolítica, 5 minutos, como parámetro asociado a la potencia aeróbica y 20 minutos como parámetro asociado a la capacidad aeróbica.

Maximal power output (in W/kg)								
	Men				Women			
	5 s	1 min	5 min	FT	5 s	1 min	5 min	FT
World class (e.g., international pro)	24.04	11.50	7.60	6.40	19.42	9.29	6.61	5.69
	23.77	11.39	7.50	6.31	19.20	9.20	6.52	5.61
	23.50	11.27	7.39	6.22	18.99	9.11	6.42	5.53
	23.22	11.16	7.29	6.13	18.77	9.02	6.33	5.44
	22.95	11.04	7.19	6.04	18.56	8.93	6.24	5.36
	22.68	10.93	7.08	5.96	18.34	8.84	6.15	5.28
	22.41	10.81	6.98	5.87	18.13	8.75	6.05	5.20
	22.14	10.70	6.88	5.78	17.91	8.66	5.96	5.12
Exceptional (e.g., domestic pro)	21.86	10.58	6.77	5.69	17.70	8.56	5.87	5.03
	21.59	10.47	6.67	5.60	17.48	8.47	5.78	4.95
	21.32	10.35	6.57	5.51	17.26	8.38	5.68	4.87
	21.05	10.24	6.46	5.42	17.05	8.29	5.59	4.79
	20.78	10.12	6.36	5.33	16.83	8.20	5.50	4.70
	20.51	10.01	6.26	5.24	16.62	8.11	5.41	4.62
	20.23	9.89	6.15	5.15	16.40	8.02	5.31	4.54
	19.96	9.78	6.05	5.07	16.19	7.93	5.22	4.46
Excellent (e.g., cat. 1)	19.69	9.66	5.95	4.98	15.97	7.84	5.13	4.38
	19.42	9.55	5.84	4.89	15.76	7.75	5.04	4.29
	19.15	9.43	5.74	4.80	15.54	7.66	4.94	4.21
	18.87	9.32	5.64	4.71	15.32	7.57	4.85	4.13
	18.60	9.20	5.53	4.62	15.11	7.48	4.76	4.05
	18.33	9.09	5.43	4.53	14.89	7.39	4.67	3.97
	18.06	8.97	5.33	4.44	14.68	7.30	4.57	3.88
	17.79	8.86	5.22	4.35	14.46	7.21	4.48	3.80
Very good (e.g., cat. 2)	17.51	8.74	5.12	4.27	14.25	7.11	4.39	3.72
	17.24	8.63	5.01	4.18	14.03	7.02	4.30	3.64
	16.97	8.51	4.91	4.09	13.82	6.93	4.20	3.55
	16.70	8.40	4.81	4.00	13.60	6.84	4.11	3.47
	16.43	8.28	4.70	3.91	13.39	6.75	4.02	3.39
	16.15	8.17	4.60	3.82	13.17	6.66	3.93	3.31
	15.88	8.05	4.50	3.73	12.95	6.57	3.83	3.23
	15.61	7.94	4.39	3.64	12.74	6.48	3.74	3.14
Good (e.g., cat. 3)	15.34	7.82	4.29	3.55	12.52	6.39	3.65	3.06
	15.07	7.71	4.19	3.47	12.31	6.30	3.56	2.98
	14.79	7.59	4.08	3.38	12.09	6.21	3.46	2.90
	14.52	7.48	3.98	3.29	11.88	6.12	3.37	2.82
	14.25	7.36	3.88	3.20	11.66	6.03	3.28	2.73
	13.98	7.25	3.77	3.11	11.45	5.94	3.19	2.65
	13.71	7.13	3.67	3.02	11.23	5.85	3.09	2.57
	13.44	7.02	3.57	2.93	11.01	5.76	3.00	2.49
Fair (e.g., cat. 5)	13.16	6.90	3.46	2.84	10.80	5.66	2.91	2.40
	12.89	6.79	3.36	2.75	10.58	5.57	2.82	2.32
	12.62	6.67	3.26	2.66	10.37	5.48	2.72	2.24
	12.35	6.56	3.15	2.58	10.15	5.39	2.63	2.16
	12.08	6.44	3.05	2.49	9.94	5.30	2.54	2.08
	11.80	6.33	2.95	2.40	9.72	5.21	2.45	1.99
	11.53	6.21	2.84	2.31	9.51	5.12	2.35	1.91
	11.26	6.10	2.74	2.22	9.29	5.03	2.26	1.83
Untrained (e.g., non-racer)	10.99	5.99	2.64	2.13	9.07	4.94	2.17	1.75
	10.72	5.87	2.53	2.04	8.86	4.85	2.07	1.67
	10.44	5.76	2.43	1.95	8.64	4.76	1.98	1.58
	10.17	5.64	2.33	1.86	8.43	4.67	1.89	1.50

Ilustración 2 Tabla test de potencia. (Coggan, 2006).

2.3.2 Test de lactato.

El porcentaje del valor de VO_{2max} que se puede mantener en el tiempo es un indicador de la capacidad de rendimiento de resistencia de un ciclista. El umbral de lactato a menudo es usado como un indicador de este porcentaje, ya que denota cuando el cuerpo está bajo un estrés elevado y con un uso acelerado de las reservas de carbohidratos. Esto es denominado como umbral de lactato o ventilatorio, que, a pesar de las discordancias entre los términos por la comunidad científica, en el ciclismo ambos términos significan lo mismo.

El umbral de lactato representa la intensidad de ejercicio en un estado estable más alta que un atleta puede mantener durante periodos prolongados, pasados los 30 minutos.

Para la medición del lactato en sangre se usa un test de lactato, en este caso las ciclistas realizan un recorrido circular en un puerto 8 veces, la intensidad en cada repetición es aumentada (basada en incrementos de 0,5 w/kg). Tras realizar una vuelta se extrae una pequeña muestra de sangre y es analizada para ver el lactato acumulado en esa serie.

2.3.3 Test de Resistencia a la fatiga

Este test se realiza en un estado de fatiga, dado que recientemente se ha publicado que uno de los factores que diferencia más claramente a diferentes niveles de ciclistas se basa en la capacidad de producir potencia en el tiempo tras diferentes niveles de carga precedente acumulada (meter referencia y añadirla a la bibliografía). En nuestro diseño para este indicador la ciclista, tras realizar 1500 kilojulios de trabajo, completa un test de 15 minutos de potencia.

2.3.4 Eficiencia.

La eficiencia se ha valorado a través de todos los datos diarios del entrenamiento y su evolución a lo largo de la temporada. La eficiencia se define como la relación entre la

potencia normalizada y la frecuencia cardiaca media. Es una métrica de Training Peaks que evalúa la capacidad de los ciclistas de realizar una producción de potencia (W) con un menor gasto energético (medido, en este caso, a través de la frecuencia cardiaca). Teniendo en cuenta que importantes incrementos de potencia no van, proporcionalmente, acompañados de grandes aumentos de la frecuencia cardiaca es previsible observar aumentos del parámetro de eficiencia cuando las ciclistas entrenen en zonas de entrenamiento más intensas, lo cual va sucediendo a lo largo de las semanas como consecuencia del diseño del entrenamiento.

2.3.5 VO_2 Score.

El VO_2 Score, es una métrica similar a la eficiencia creada por Alan Couzens, que tiene en cuenta la potencia normalizada, el peso, la FC basal, la FC máxima y la FC media del entrenamiento. Al contar con más variables de entrada, la convierte en una métrica más precisa que el parámetro de eficiencia anteriormente descrito.

3 Big data.

Con el avance de la era digital en pleno siglo XXI, hoy en día nuevas tecnologías se abren paso, ganando una popularidad creciente, este es el caso del Big data o también conocido como macrodatos en su traducción al español.

El termino Big data refiere a la manipulación, administración y al análisis de grandes conjuntos de datos. Estos conjuntos de datos son tan masivos que es necesario la utilización de Softwares específicos para su tratamiento.

La primera vez que el termino Big data fue utilizado fue en el MGI (McKinsey Global Insitute) en junio de 2011.

Desde entonces diversos autores han tratado de acotar su significado. (Manyika, J. et al., 2011) definen Big Data como el conjunto de datos cuyo tamaño va más allá de la capacidad de captura, almacenado, gestión y análisis de las herramientas de base de datos.

Por otro lado, según Gartner (2012): “Son activos de información caracterizados por su alto volumen, velocidad y variedad, que demandan soluciones innovadoras y eficientes de procesado para la mejora del conocimiento y toma de decisiones en las organizaciones.”

3.1 El Big data en el deporte.

En el mundo del deporte en 1972 Hamilton lanzo su primer reloj *weareable* el cual era capaz de registrar información. Desde entonces no se ha hecho más que recopilar información y datos. Pero no fue hasta 1985, tras el lanzamiento de Microsoft Excel, cuando se publicó el libro de The Bill James Historical Baseball Abstract, en el cual se analizaron datos en data bases de Excel.

Gracias al avance de las tecnologías en el año 2000 Catapult Sports desarrolla su primer prototipo experimental de chaleco wearable. Con el paso del tiempo y tras el éxito de los estadísticos del equipo de Beisbol Oakland Athletic en el 2002, muchas organizaciones y clubes deportivos empezaron a interesarse por estas nuevas tecnologías de la información. En el 2006 Nike y Apple sacan su nueva aplicación de entrenamiento personal Nike+.

En la segunda década del siglo XXI el auge de esta tecnología fue en aumento, el FC Barcelona implementaba el sistema COR para el análisis de rendimiento de los jugadores, en 2014 la federación alemana de futbol contrata SAP y gana el mundial de futbol y en 2018 los Philadelphia Eagles ganan la Super Bowl usando un nuevo sistema de la información,

3.2 El Big data en el ciclismo.

El mundo del ciclismo sufrió una evolución similar al resto de los deportes entre 1983 y 1980 Greg Lemond se alzó con tres Tours de Francia y dos campeonatos del mundo siendo el primer ciclista en utilizar los medidores de potencia en su bicicleta. Desde entonces en ciclismo se empezaron a recopilar datos con grandes avances en los dispositivos gracias a mediciones más sofisticadas. Un gran avance en el ciclismo pudo verse en 2015 cuando en el Tour de Francia se obtienen datos en tiempo real de las bicicletas, sincronizando toda la información.

3.3 Ciencia de datos.

En 1962 años, John Tukey llamó a una reforma académica en estadística, a través de uno de los más importantes papers de esa época, llamado “The Future of Data Analysis”, donde señalaba la necesidad futura de una ciencia cuyo interés sea aprender de la data o análisis de datos. (Tukey, 1962, citado por Chian, 2019)

Por ello, la ciencia de datos se define como la disciplina que convierte los datos en bruto o “*Raw Data*” en conocimiento, entendimiento y herramientas para su comunicación.

Wickham y Grolemund (2017) señalan que en general el proceso seguido en la ciencia de datos se divide en 4 fases, la obtención e importación de la “*Raw Data*”, la limpieza de datos, el análisis exploratorio y por último la presentación o comunicación de la información obtenida.

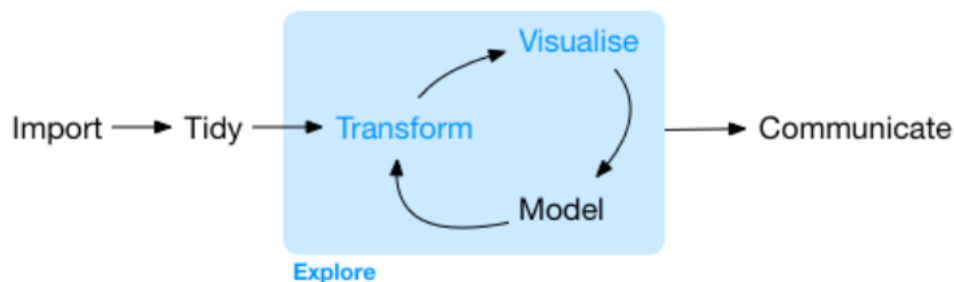


Ilustración 3 Wickham, H., & Grolemund, G. (2017). *R for Data Science*

3.3.1 Limpieza de datos.

Las primeras personas en hacer referencia a la limpieza de datos fueron Dasu y Johnson (2003) en su libro “*Exploratory Data Mining and Data Cleaning*” definían la limpieza o “*Data Scrubbing*” como el proceso mediante el cual se corrigen o eliminan datos incorrectos, corruptos, formateados incorrectamente, duplicados o incompletos dentro de un conjunto de datos.

Los principios de la limpieza de datos siguen un consenso, para que pueda ser entendida sencillamente, estos principios son los siguientes:

- En una base de datos cada columna es una variable.
- Cada fila de una base de datos es llamada observación.

3.3.2 Análisis exploratorio.

El análisis exploratorio de datos también conocido como EDA (por sus siglas en inglés exploratory data analysis), es la tercera fase del proceso de la ciencia de datos. El análisis exploratorio trata de sacar conclusiones a partir de unos datos mediante la transformación, visualización y modalización de un conjunto de datos previamente limpiados o “*Tidy Data*”. Este ciclo se repite iterativamente.

Entre las técnicas de EDA se encuentran los resúmenes estadísticos, las tablas de correlación o los gráficos de percentiles.

Para el inicio del EDA se analizan las variaciones y distribuciones de las variables, gracias a ello se pueden clasificar las observaciones de una variable en valores típicos y valores inusuales o “*Outliers*”. También es común encontrar en un “*dataset*” valores faltantes o “*Missing Data*”.

Cuando en un dataset se encuentran valores atípicos o faltantes hay que transformarlos o eliminarlos para no encontrar futuros inconvenientes en la última fase del proceso de la ciencia de datos.

Para el tratamiento de los datos es común hacer un análisis de covariación de las diferentes variables. Para encontrar relación entre ellas. En caso de existir relación entre las variables, se procede con la búsqueda de patrones y la modelización.

Tras finalizar la parte de transformación de los datos y modelización, se procede con la visualización de los datos mediante resúmenes estadísticos y gráficas.

4 Diseño del estudio y resultados

4.1 Obtención de la información.

Lo principal fue la recopilación de la información para ello fue usada la aplicación de Training Peaks. Los datos comenzaron a recabarse desde el 2 de octubre de 2020, en la actualidad se han conseguido datos hasta el 28 de abril de 2021. Los datos se obtuvieron de 3 atletas del equipo Fundación Euskadi femenino.

4.2 Preprocesado de los datos.

4.2.1 Dispositivos para la obtención de los datos.

Los datos fueron obtenidos a través de ciclo computadores Garmin (diferentes modelos dependiendo de la ciclista) y potenciómetros Power2max proporcionados por el equipo.

4.2.2 Formato de los datos.

Los dispositivos Garmin, generan un archivo en formato FIT (Flexible and Interoperable Data Transfer). Este tipo de archivos pertenecen a la categoría de archivos GIS (Global Information System) y son comúnmente utilizados para registrar rutas.

La mayor desventaja de este tipo de formatos radica en la dificultad de leer la información contenida con cualquier Software de lectura de datos. Por esta misma razón se necesitan programas especiales para poder leer este tipo de archivos, sin usar aplicaciones como la de Garmin Connect (Windows y Android) o Garmin Training Center (IOS).

Garmin en su página web proporciona un conversor de archivos llamado FIT JDK, el cual convierte los archivos FIT en archivos CSV (Coma Separated Values), pero el código está escrito en Java y los archivos pierden información.

Por ello, la mejor idea fue desarrollar un conversor en un lenguaje de programación de ciencia de datos para tener un entorno de desarrollo unificado.

4.2.3 Lenguajes de programación Python y R.

Debido a la problemática de la lectura de datos es necesario el uso de un lenguaje de programación. En la actualidad los lenguajes de programación más populares entre los “*Data Scientist*” son R y Python.

Python es un lenguaje de alto nivel dedicado a la programación en general. Por otro lado, R es un lenguaje de programación con un enfoque a la estadística. Ambos lenguajes están orientados a objetos.

Tras probar ambos lenguajes, se observó una clara curva de aprendizaje más sencilla en el lenguaje R (Como valoración personal) por lo que, este fue el escogido para el análisis de datos.

El entorno de desarrollo (IDE) que se utilizó con R fue RStudio.

4.2.4 Conversión del formato de los archivos.

Lo primero que se hizo fue importar todos los archivos que fueron previamente descargados de la página Web de Training Peaks, cada archivo representa un entrenamiento de un atleta.

Los archivos de los diferentes atletas fueron importados a la vez en forma de lista. El software de R carece de funciones que puedan leer archivos con extensión FIT por lo que fue necesario descargar un paquete creado por Mike Smith llamado “FitFileR”.

Este paquete lee los archivos y mediante funciones es capaz de separar los registros del entrenamiento segundo a segundo, los registros tomados manualmente y los resúmenes de las sesiones.

El único inconveniente es que este paquete fue creado para leer un único archivo. Cada archivo tiene una gran cantidad de información por lo que es si se tratase de leer una cantidad de archivos a la vez, R se demoraría mucho tiempo en leerlos todos, por lo que, fue necesario crear una función que leyese y extrajese la información requerida para cada programa, esto redujo en gran medida la complejidad computacional y por lo tanto el tiempo de lectura.

4.3 Limpieza de datos.

Tras la lectura de los datos se obtuvo un dataframe en el cual las filas representaban observaciones, y las columnas variables. La distancia entre las observaciones era de 1 segundo. Las variables con mayor importancia fueron: Tiempo, potencia, cadencia, velocidad, altura y frecuencia cardiaca (FC).

4.3.1 Análisis de las variables.

El primer paso el paso es crear resúmenes estadísticos de las variables para encontrar para ver cómo están distribuidas y conocer el número exacto de datos faltantes. Los dataframes analizados tenían 3,5 millones de observaciones y 20 variables, 70 millones de datos en total.

Las variables que fueron analizadas en profundidad fueron, la potencia, la cadencia, la frecuencia cardiaca y la velocidad. En el siguiente resumen se puede observar la distribución de estas variables:

	potencia	frecuencia	cadencia	velocidad
Min.	0.0000	32.0000	0.0000	0.000000
1st Qu.	111.0000	138.0000	71.0000	5.524000
Median	146.0000	151.0000	80.0000	7.493000
Mean	200.0196	160.8365	74.8263	7.310057
3rd Qu.	179.0000	170.0000	87.0000	9.013000
Max.	65535.0000	255.0000	149.0000	21.880000
NA's	98101.0000	11430.0000	100238.0000	68774.000000

Ilustración 4 Tabla resúmenes estadísticos. Elaboración propia.

Tras observar el resumen se observa que hay una gran cantidad de datos faltantes para la variable velocidad, y relativamente pocos datos para el resto de las variables, teniendo en cuenta la cantidad de observaciones por variable el porcentaje de “*Missing Data*” de cada variable es: 0,30% en la variable de potencia, 0,002% de datos faltantes en la frecuencia cardiaca, 0,35% para la cadencia y 3,71% de datos faltantes de velocidad.

Por otro lado, se observa que los valores máximos de la frecuencia cardiaca y de la potencia son desmesurados, por lo que será necesario un análisis de variación y de distribución de frecuencias para detectar posibles “*Outliers*” en estas variables.

4.3.2 Análisis Outliers de la Frecuencia Cardiaca (FC)

El valor máximo y mínimo de la frecuencia cardiaca eran excesivamente elevados por lo que se procedió a hacer un análisis de variación y distribución de los datos, mediante dos gráficos se pudieron representar los Outliers de la FC correctamente.

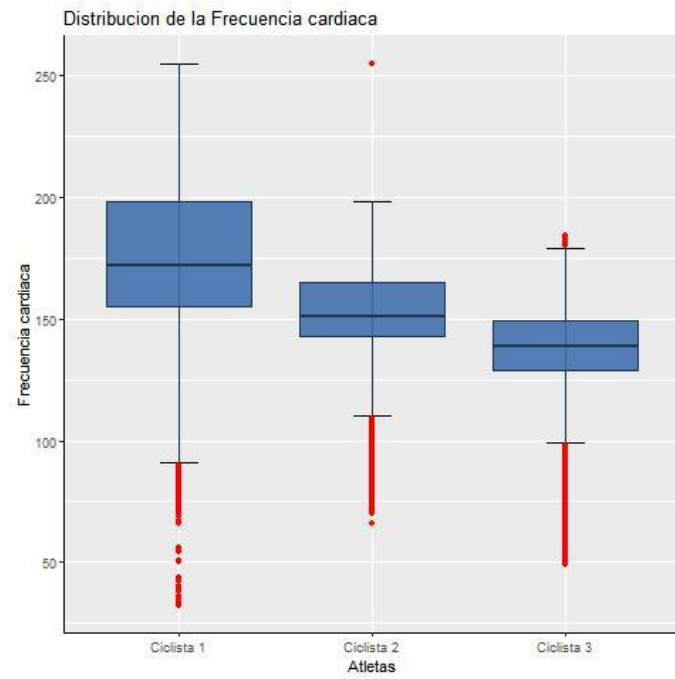


Ilustración 5 Boxplot Outliers Frecuencia Cardiaca. Elaboración propia.

Las ciclistas 1 y 2 reportaron valores de frecuencia cardiaca excesivamente elevados, por lo que fueron tomados como Outliers. El problema principal de los gráficos boxplot, es que no son capaces de detectar multimodalidad, por lo que, en el caso de ambas ciclistas, se realizaron gráficos de histogramas para ver la densidad de los datos.

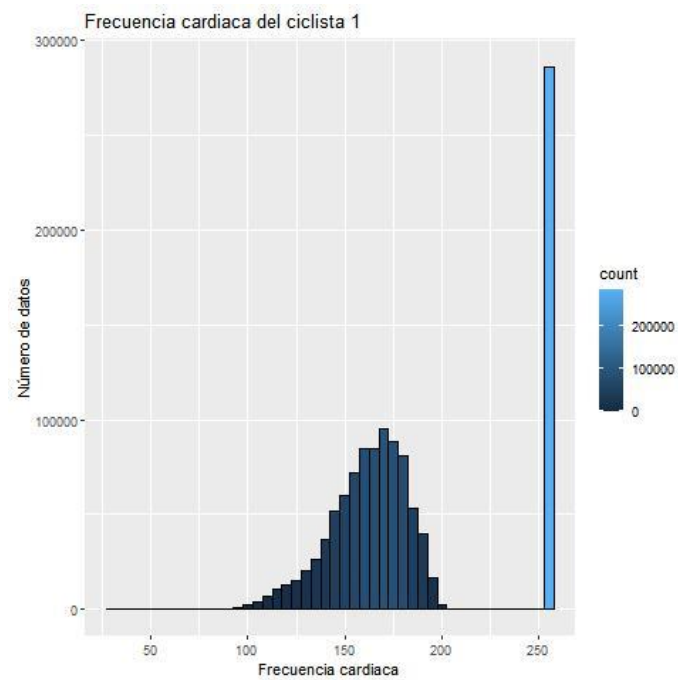


Ilustración 6 Distribución de frecuencias de la FC, ciclista 1. Elaboración propia.

En el caso de la ciclista 1 se observó que la mayoría de los datos recayeron en una FC de 255 rpm, por lo que se dedujo en una primera hipótesis que cuando el pulsómetro no registraba correctamente los datos, se obtiene ese valor.

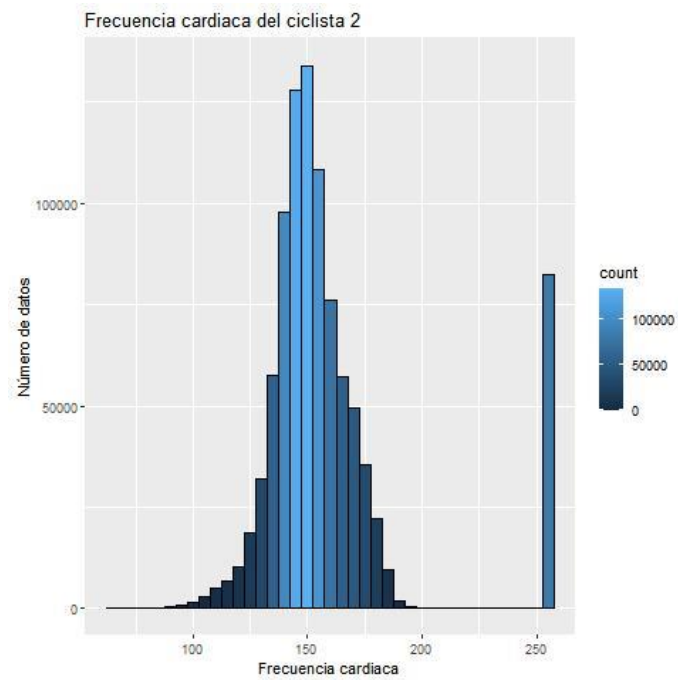


Ilustración 7 Distribucion de frecuencias de la FC, ciclista 2. Elaboración propia.

Por otra parte, la ciclista 2 obtuvo resultados similares, con la diferencia de que los Outliers aparecieron con menor frecuencia que los datos típicos.

Los Outliers fueron transformados y se convirtieron en Missing data, el resultado de la eliminación de los Outliers fue el siguiente:

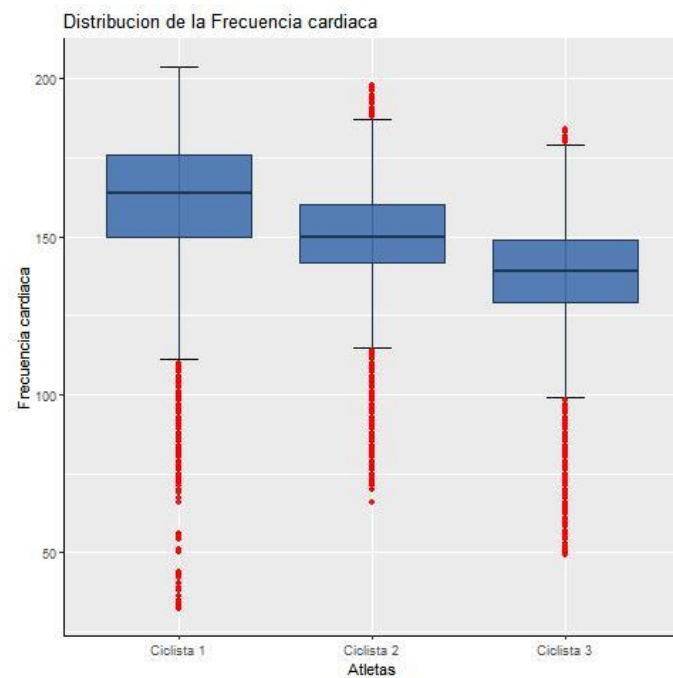


Ilustración 8 Boxplot distribución de la FC sin Outliers. Elaboración propia.

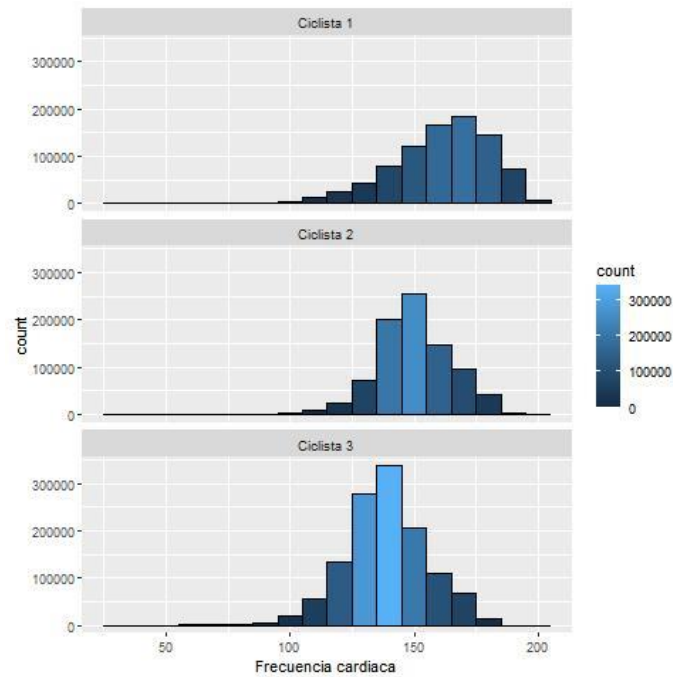


Ilustración 9 Frecuencia de la FC sin Outliers. Elaboración propia.

4.3.3 Análisis Outliers de la Potencia.

Tras analizar la frecuencia cardiaca otro variable que parecía tener Outliers fue la potencia. Los máximos de esta fueron de 65536 W, y el valor mínimo fue de 0 W, este último no se contó como posible Outlier ya que cuando las ciclistas dejaban de pedalear el dato que se registraba fue de 0 W.

Se observó que la potencia máxima era de 65536 W, los datos en este valor eran muy poco frecuentes por lo que se volvió tomarlos como valores faltantes y se volvió a analizar la potencia.

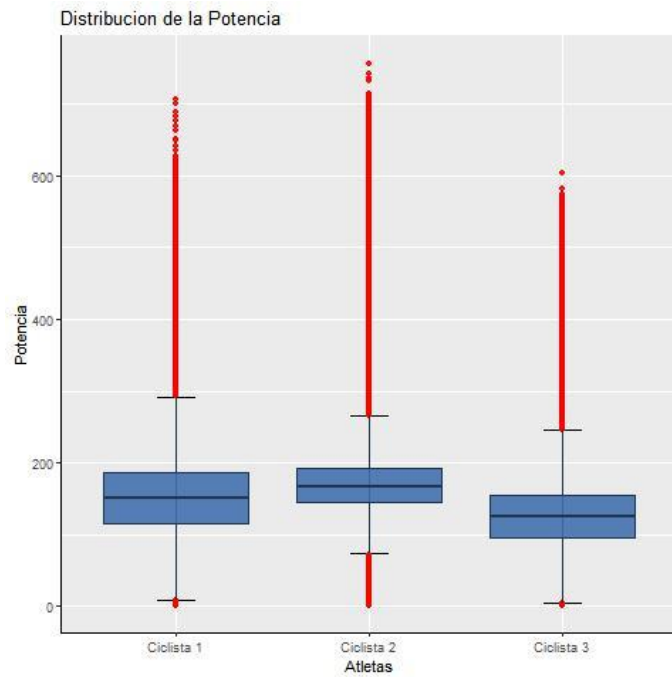


Ilustración 10 Ilustración 6 Boxplot distribución Potencia sin Outliers. Elaboración propia.

En este boxplot la potencia se representó mucho mejor, los puntos rojos indican datos alejados de los datos típicos por lo que en otro caso probablemente no habría que tenerlos en cuenta, pero en ciclismo, al hacer velocidad el valor de la potencia se eleva mucho por lo que, en este caso estos datos no fueron tomados como Outliers.

4.3.4 Análisis de la Missing data.

Una vez convertidos los Outliers en Missing data, el siguiente paso fue encontrar patrones entre las variables con observaciones faltantes. Para ello se utilizó la librería VIM de R que mediante un gráfico que, mediante múltiples combinaciones, relaciona los valores faltantes de cada variable respecto a las demás, analizando la frecuencia de aparición de cada combinación.

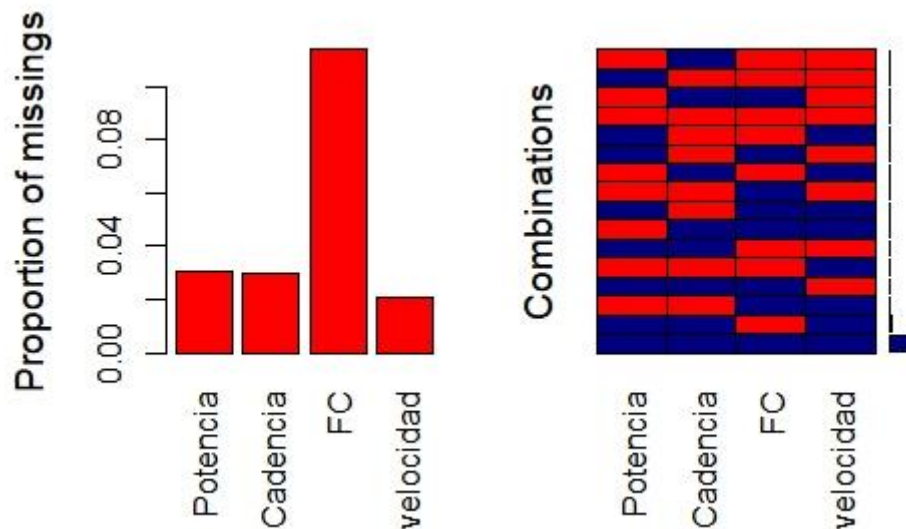


Ilustración 11 Combinaciones de patrones de Missing data. Elaboración propia.

Tras la manipulación de los Outliers la variable que más Missing data tenía paso a ser la FC. Se obtuvieron 16 combinaciones de Missing data posibles, la combinación que se dio con más frecuencia fue la inexistencia de Missing data entre las observaciones, seguida por la existencia de Missing data en la FC.

Al tener alrededor del 1% de Missing data se decidió hacer una imputación de esta. Las observaciones eran continuas, esto es no eran individuos aislados, mantenían una relación temporal. Por otro lado, la volatilidad de variación de algunas variables era muy alta y existía covariación entre ellas, por lo que el modelo de imputación debía de ser multivariante.

Para la imputación se utilizó la librería MICE de R, esta librería hace una imputación múltiple utilizando el algoritmo “*Multiple Imputation by Chained Equations*”. Al ser datos continuos el mejor método para hacer la imputación era el “*Predictive mean matching*” (PMM).

Esta forma de rellenar la información faltante se completa mediante un proceso iterativo, esto es, que cada valor se calcula continuamente hasta tener un valor ajustado. El número de iteraciones para cada valor fue de 50.

Una vez acabada la imputación había que analizar el error, el mayor problema en este apartado fue que los ciclistas a veces entrenaban sin potenciómetro por lo que no se podía saber que potencia habían desarrollado durante los entrenamientos.

Para analizar que tal había trabajado el modelo lo que se hizo fue coger un entrenamiento desde Training Peaks en el que no se tenían registros de potencia, y comparar con la potencia predicha.

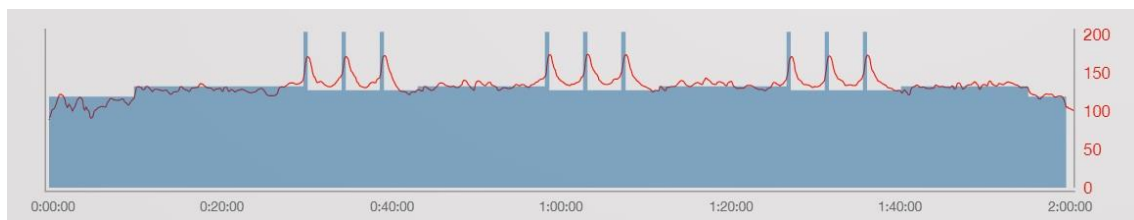


Ilustración 12 Entrenamiento sin potenciómetro. Fuente: Training Peaks.

La imagen anterior esta sacada de la página Web de Training Peaks, es un gráfico que representa un entrenamiento sin el uso de potenciómetro. Los bloques azules representan la intensidad y la duración de cada serie o parte del entrenamiento, estos bloques son creados por el entrenador. Tras subirlo a la plataforma el atleta ve el entrenamiento y lo realiza, los efectos producidos por el entrenamiento en cuanto a la frecuencia cardiaca del atleta son representados con la línea roja, el valor de esta está reflejado en el eje izquierdo.

Es observable que la frecuencia cardiaca tiene un cierto retraso en cuanto a su covariación con la hipotética potencia, esto se debe a que la respuesta del sistema aeróbico (¿o cardio respiratorio igual mejor no?) no es inmediata a las variaciones de potencia.

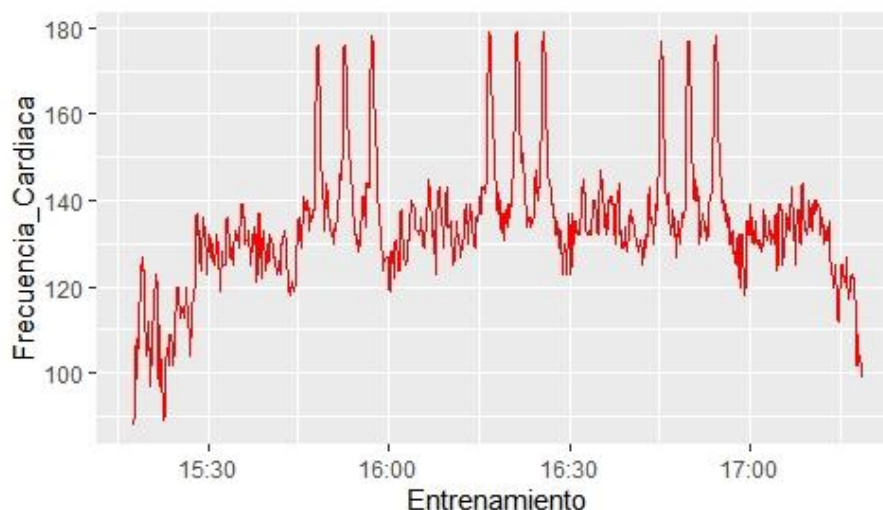


Ilustración 13 Frecuencia cardiaca exportada de Training Peaks. Elaboración propia.

En la siguiente ilustración se ve representados los mismos valores anteriores de frecuencia cardiaca, esta vez, los valores son extraídos de los pulsómetros Garmin, por lo que es necesario un análisis en R para poder visualizar el entrenamiento.

A diferencia del gráfico anterior en este gráfico no viene representado el entrenamiento, Training Peaks no proporciona esa información. El mayor problema de no tener datos de potencia es que las ciclistas de entrena Javier Orbañanos, llevan un control de la carga del entrenamiento basado en el modelo del “*Training Stress Score*” (TSS), modelo propuesto por Allen y Coggan (2010). Este modelo de cuantificación de la carga esta basado en la potencia que registran los pulsómetros por lo que tener datos de potencia diarios y fiables es fundamental para estos ciclistas.

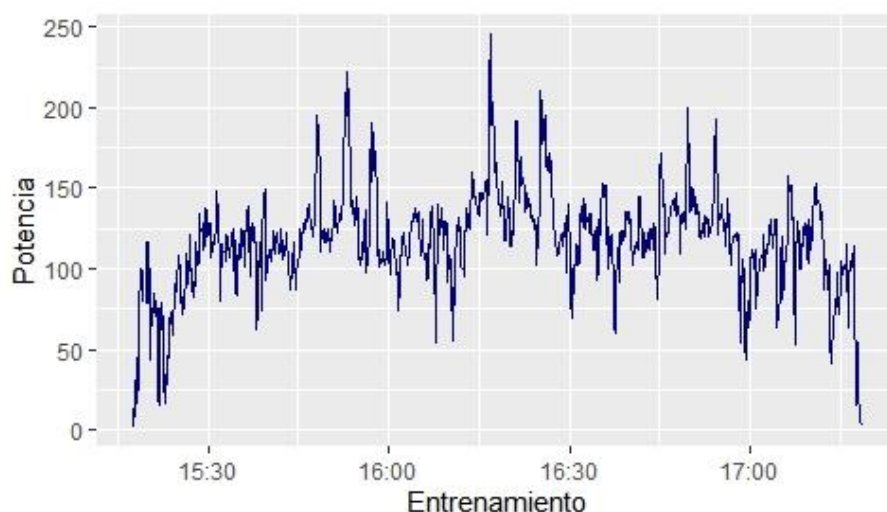


Ilustración 14 Potencia Imputada con suavizado $k = 50$. Elaboración propia.

Tras la predicción de la potencia de este entrenamiento teniendo en cuenta como variables predictoras la velocidad, FC y el desnivel, se obtuvieron los datos de potencia que se pueden observar en la ilustración anterior. Estos datos fueron suavizados con una media móvil de 50 periodos para poder ver las similitudes entre los gráficos visualmente.

Los registros de potencia suelen tener más ruido y se encuentra más variabilidad entre las observaciones.

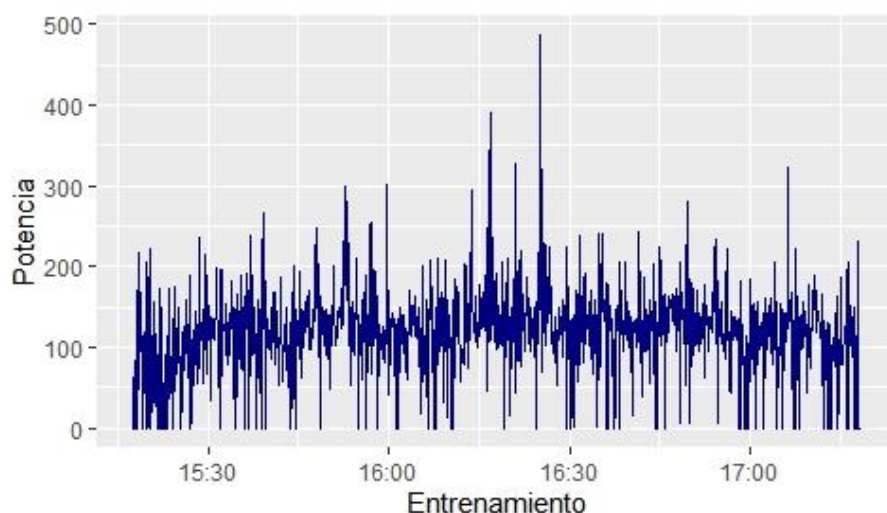


Ilustración 15 Potencia Imputada real. Elaboración propia.

Como se observa en este gráfico, esta es la potencia real que predijo el modelo. Es observable que muchos datos tienen un valor de 0 W, esto se debe a que el atleta en diferentes puntos deja de pedalear y no se registra ningún valor de potencia ni de cadencia.

4.3.5 Transformación de datos, creación de nuevas variables.

Una vez habiendo limpiado el data set y sus variables brutas, se procedió con la creación de nuevas variables y métricas para la evaluación y control de la performance en el ciclismo.

4.3.5.1 Variables de monitorización del entrenamiento.

- Potencia Normalizada.

La primera variable creada fue la potencia normalizada, esta es definida por Allen & Coggan (2010) como “una estimación de la potencia que un deportista podría haber mantenido, con un mismo coste fisiológico, si su producción de potencia hubiera sido perfectamente constante”. La ecuación de esta métrica es la siguiente, para $n = 30$ periodos.

$$Pn = \left(\frac{\sum_{i=1}^n x_i \left(\frac{1}{n} \sum_{i=1}^n x_{i-1} \right)^4}{n} \right)^{\frac{1}{4}}$$

Tras realizar la ecuación en R, se transformó el resultado en una serie temporal en la cual la potencia estaba representada en color rojo y la potencia normalizada por su parte en azul.

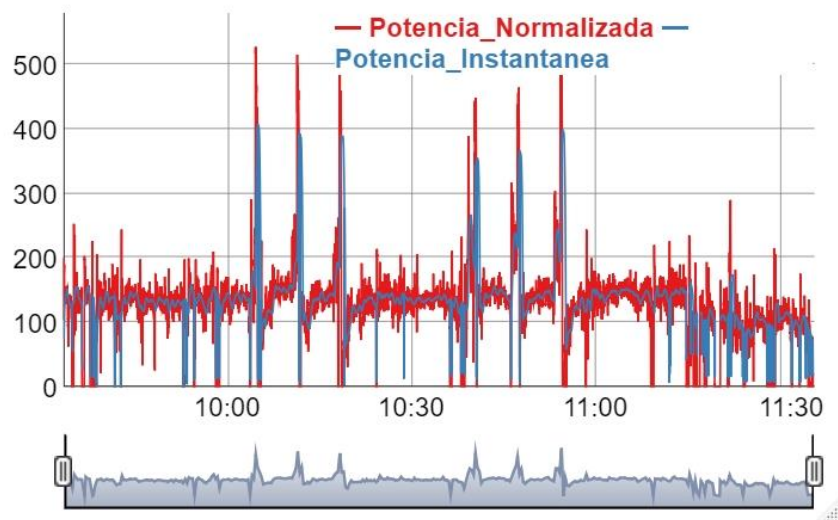


Ilustración 16 Visualización de la potencia normalizada. Elaboración propia.

El único problema de la potencia normalizada fue que, al realizarse con medias móviles, dependiendo de los periodos asignados, al principio de cada subconjunto temporal se creaban 30 observaciones de Missing values. Por lo que para solucionar este problema se optó por predecir los valores faltantes mediante un modelo de predicción univariante autorregresivo de medias móviles. El modelo utilizado fue ARIMA “*Autoregressive integrated moving average*”.

4.3.5.2 Variables de rendimiento residual del entrenamiento.

- Eficiencia.

Esta métrica, es una métrica que evalúa el rendimiento de una manera residual durante las sesiones de entrenamiento. Es una variable de carácter fisiológico que relaciona la frecuencia cardiaca y la potencia normalizada. El mayor inconveniente de esta métrica es que pierde validez en esfuerzos sub máximos, ya que el sistema aeróbico pierde relevancia en esfuerzos sub máximos de corta duración y la potencia puede llegar a tener valores muy altos. Su fórmula es la siguiente:

$$Eficiencia = \frac{Potencia\ normalizada\ (Pn)}{Frecuencia\ Cardiaca\ (FC)}$$

- $\text{VO}_2 \text{ Score}$.

El $\text{VO}_2 \text{ Score}$ es una métrica similar a la eficiencia, pero tiene en cuenta mas variables por lo que la hace más completa. Esta métrica tiene en cuenta el peso del atleta, la frecuencia cardiaca basal y la frecuencia cardiaca máxima. El resultado esta supuesto a ser una aproximación indirecta del consumo máximo de oxígeno ($\text{VO}_{2\text{max}}$),

Pero como en el anterior caso las zonas de entrenamiento pueden repercutir en el resultado. Su fórmula es la siguiente:

$$\text{VO}_2 \text{ Score} = \frac{\left(\frac{Pn}{75}\right) \cdot \left(\frac{1000}{\text{Peso}}\right)}{\left(\frac{FC - FC_{\text{basal}}}{FC_{\text{max}} - FC_{\text{basal}}}\right)}$$

La eficiencia y el $\text{VO}_2 \text{ Score}$ son variables similares su distribución durante un entrenamiento se puede ver en la imagen a continuación.

- Torque.

El torque en ciclismo se define como la fuerza que se aplica a los pedales para moverse. Para hacer una estimación indirecta de este se calculó mediante la relación entre la potencia instantánea y la cadencia, teniendo en cuenta el área de la circunferencia $2\pi R$, y el sistema métrico de la cadencia, revoluciones por minuto (rpm).

$$\text{Torque} = \frac{\text{Potencia}}{\left(\frac{(\text{Cadencia} \cdot 2\pi)}{60}\right)}$$

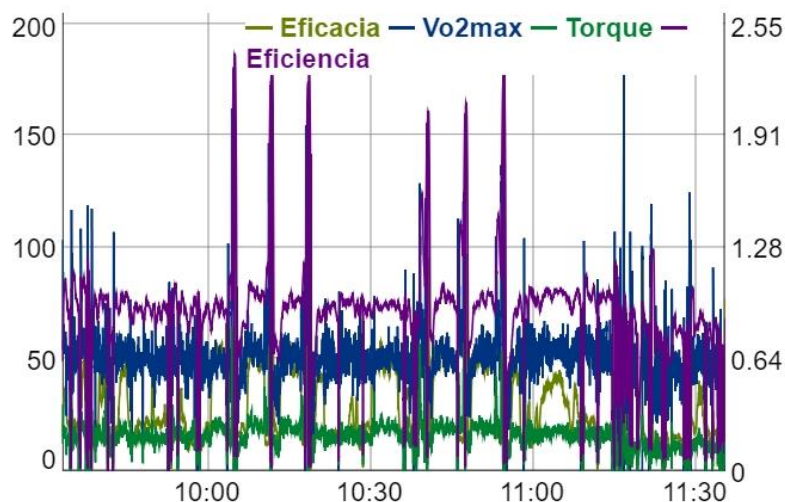


Ilustración 17 Comportamiento de variables durante el entrenamiento. Elaboración propia.

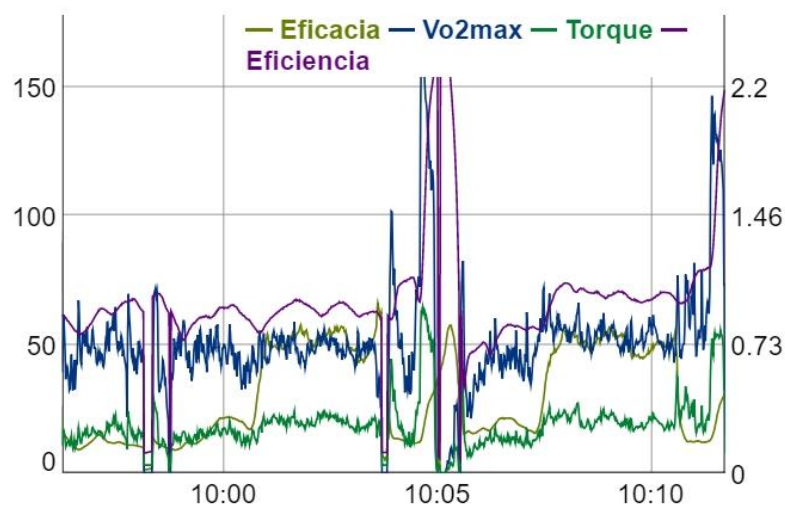


Ilustración 18 Comportamiento de las variables durante una serie de velocidad. Elaboración propia.

4.3.5.3 Variables de cuantificación de la carga del entrenamiento.

- Training Stress Score (TSS).

La TSS determina la carga de la sesión de entrenamiento, aunque Training Peaks entre los datos que deja exportar este sea uno de estos, al haber hecho una limpieza exhaustiva de los datos y tener más datos de potencia que la propia aplicación, se decidió recalcular la carga de entrenamiento para hacer una comparativa con los datos que ofrece la

aplicación. La fórmula de la TSS relaciona la intensidad y el volumen de entrenamiento

La fórmula de la TSS es la siguiente:

$$TSS = \left(\frac{\text{Tiempo}(s) \cdot \text{Potencia norm} \cdot \text{Factor Intenidad}}{\text{Umbral anaerobico de potencia} \cdot 3600} \right) 100$$

Siendo el factor de intensidad:

$$\text{Factor de Intensidad} = \frac{\text{Potencia Normal}}{\text{Potencia Umbral}}$$

- Carga aguda (ATL).

La carga aguda es el promedio de las cargas durante un microciclo o periodo en este caso un periodo de 7 días. Para la obtención de este dato se creó una fórmula de media móvil exponencial de 7 periodos, ya que al estar ponderada los entrenamientos más cercanos al presente tienen una mayor ponderación ya que su carga tiene más influencia en el estado de forma presente.

$$ATL = \frac{TSS_1 + (1 - \alpha)^1 TSS_2 \dots + (1 - \alpha)^n TSS_n}{1 + (1 - \alpha)^1 \dots + (1 - \alpha)^n}$$

Siendo α una constante que varía de 0 a 1 en base al número de periodos n:

$$\alpha = \frac{2}{n - 1}$$

- Carga crónica (CTL).

La carga crónica fue calculada de la misma manera que la carga aguda, pero siendo el periodo mayor en este caso de 42 periodos, ya que la carga crónica se define como la condición física actual o el “*Fitness*”.

$$CTL = \frac{TSS_1 + (1 - \alpha)^1 TSS_2 \dots + (1 - \alpha)^{42} TSS_{42}}{1 + (1 - \alpha)^1 \dots + (1 - \alpha)^{42}}$$

Y por su parte α sería:

$$\alpha_{42} = \frac{2}{42 - 1}$$

- Estado de forma (TSB).

La TSB se define como el nivel de rendimiento actual en base a la carga, se calcula mediante la resta de la carga crónica menos la carga aguda. Tal que, si la carga aguda disminuye y la carga crónica es alta, el deportista reportara altos niveles de forma física para tener buen rendimiento.

$$TSB = CTL - ATL$$

- Diferencias entre Training Peaks y datos calculados:

Para la comparación de los resultados obtenidos se compararon los gráficos de control de la carga de Training Peaks con los resultados obtenidos. En el eje x se representaba la línea temporal con una periodicidad diaria, el eje y era doble, en el lado izquierdo están los valores de CTL y ATL, en el lado derecho se encuentran los valores de TSB.

En el gráfico de Training Peaks se representa la TSB con el área anaranjada, la CTL con el área azul y la ATL con la línea rosa.

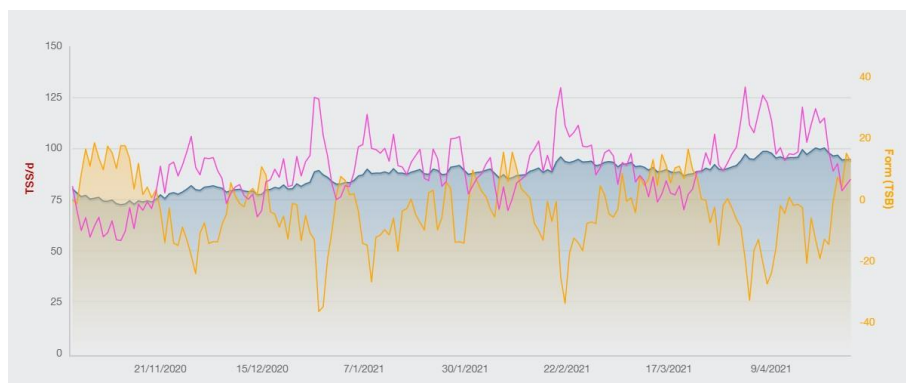


Ilustración 19 Control de carga Training Peaks. Fuente Training Peaks

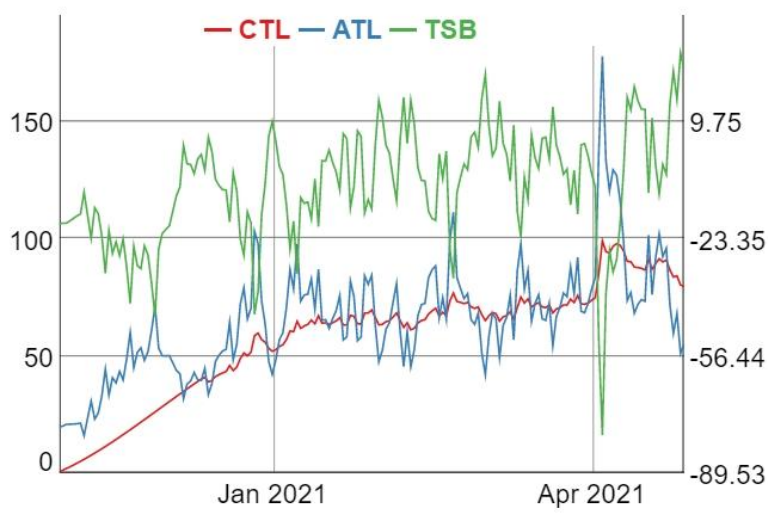


Ilustración 20 Control de carga con datos de potencia diarios. Elaboración propia.

Las principales diferencias entre los gráficos es que el gráfico de Training Peaks reportaba máximos de TSB cada vez más bajos, mientras que los altos de ATL aumentaban con el tiempo.

En los resultados que se obtuvieron los máximos de TSB y ATL siguieron una tendencia positiva durante la temporada.

Esta diferencia puede deberse principalmente a la diferencia de CTL en ambos gráficos. Mientras que en el gráfico de Training Peaks se pudo ver una tendencia casi lateral, en el otro gráfico se observa una clara tendencia alcista de la CTL. Los resultados tenían

sentido ya que la primera observación del eje x fue el inicio de la temporada, por lo que tras meses de entrenamiento la CTL debería verse incrementada sustancialmente.

El mayor problema que se obtuvo en los resultados es que la CTL en el día 1 tendía a 0 lo cual es un dato irreal ya que el estado de forma inicial no puede ser igual a 0. Esto se debe a que, al calcularse mediante medias móviles exponenciales de 42 periodos, los primeros datos no tenían valor, por lo que se calcularon con la tendencia. En la variable de ATL se resolvió de la misma manera el problema, en este caso fueron 7 periodos por lo que la diferencia con la realidad fue menor.

4.4 Evolución de los indicadores de rendimiento durante la temporada.

4.4.1 Perfil de potencia.

Se realizaron test de 5, 12 y 30 segundos y test de 1, 5 y 20 minutos. El resultado de los test se calculó haciendo el promedio de producción de potencia en ese intervalo de tiempo.

4.4.1.1 Análisis del perfil de potencia.

- Test de 5 segundos.



Ilustración 21 Test de 5" desarrollo de potencia y torque. Elaboración propia.

El test de 5 segundos se realizó con una velocidad inicial de 0 m/s. Después se realizó un sprint durante cinco segundos. Después de acabarlo, se efectuó un descanso activo de 3 minutos y se repitió la prueba.

En el gráfico se observan dos variables, la línea morada representa la fuerza de torque ejercida y la línea verde representa la potencia desarrollada. Al ser un test de corta duración y iniciarse desde parado, los valores iniciales de torque superan a los de potencia ya que el torque está relacionado con la cadencia y esta tarda unos segundos a llegar a su valor crítico.

- Test de 1 minuto.

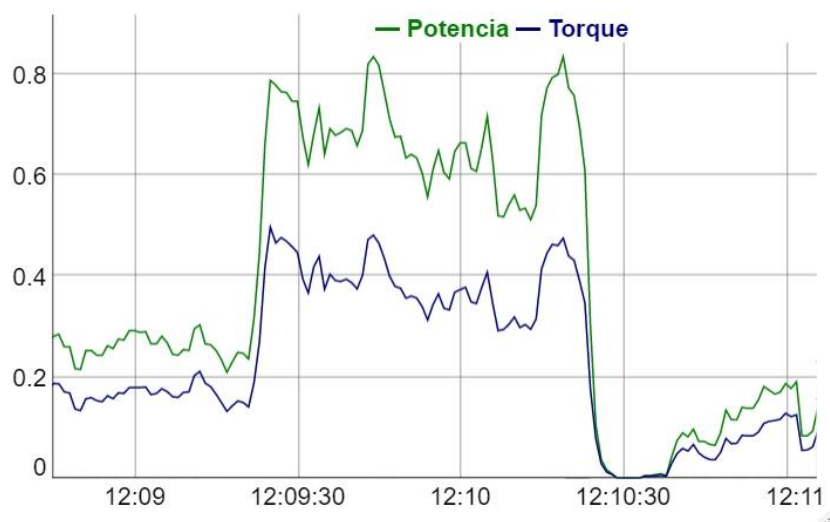


Ilustración 22 Test de potencia 1'. Elaboración propia.

El test de un minuto no se realizó desde parado, por lo que al tener un valor de cadencia inicia la potencia siempre se mantuvo por encima del torque. Desde el inicio del test se observa una caída leve de la potencia que, en los últimos diez segundos es incrementada hasta llegar a su máximo valor.

- Test de Potencia de 5 minutos.

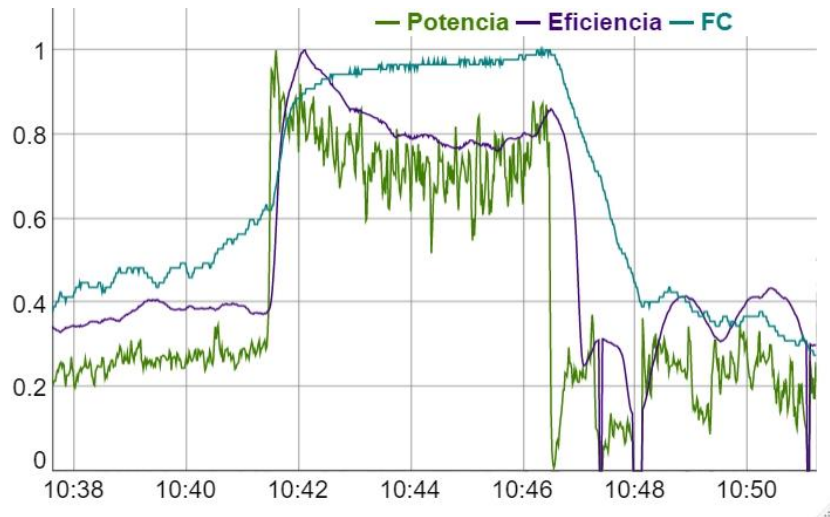


Ilustración 23 Test de 5'. Elaboración Propia.

En el test de 5 minutos no se tuvo en cuenta el torque, en su caso se analizaron la frecuencia cardiaca y la eficiencia. Se observa como la eficiencia y la potencia están altamente correlacionadas, con la diferencia de que la potencia normalizada cuenta con cierto retardo respecto a la potencia.

- Test de potencia de 20 minutos.

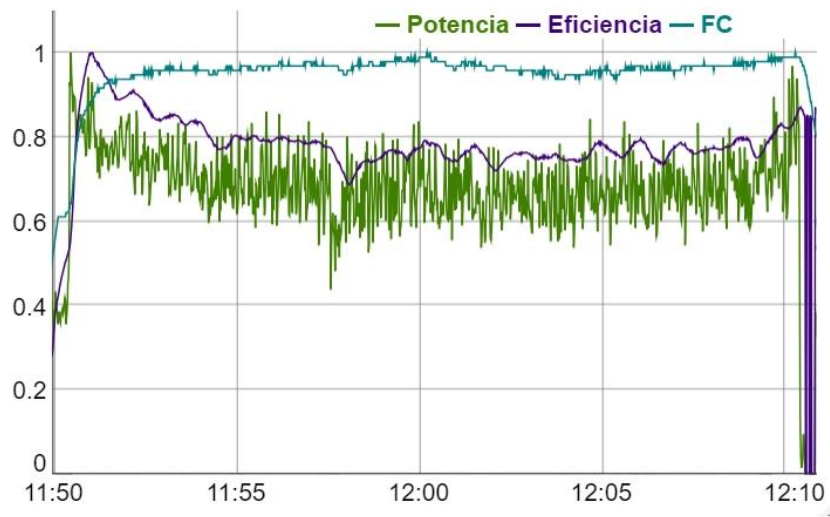


Ilustración 24 Test de potencia de 20'. Elaboración propia.

Por último, en el test de 20 minutos, se analizaron las mismas variables que en el test de 5 minutos. Es remarcable que, en los primeros 10 minutos hay una caída de potencia y de eficiencia y esta potencia es mantenida hasta el final de la prueba.

4.4.1.2 Evolución de la potencia durante la temporada.

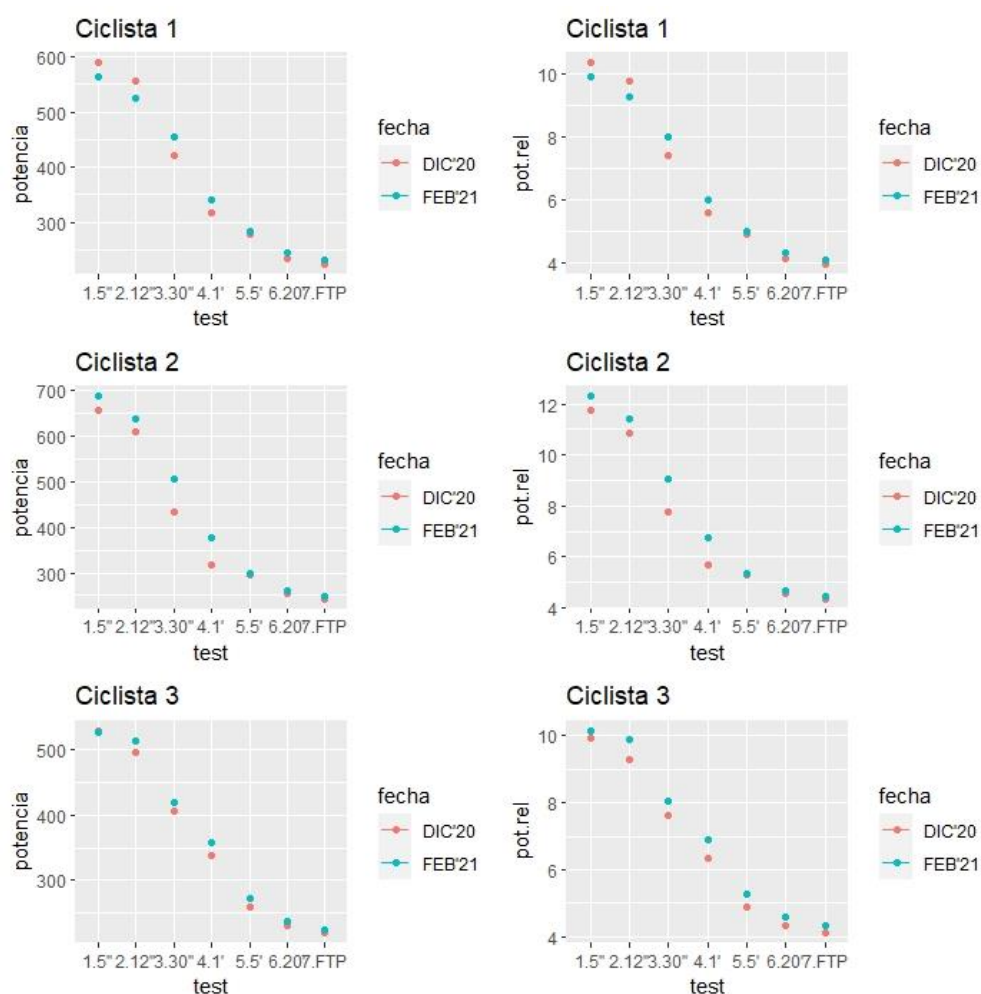


Ilustración 25 Resultados relativos y absolutos test de potencia. Elaboración propia.

En el siguiente gráfico se observa los resultados de los test de potencia. En el eje x están representados los diferentes test de potencia que se realizaron durante la temporada, ordenados por duración de menor a mayor. En el eje y están representados los valores medios de cada test de potencia. Los gráficos de la columna izquierda representan la potencia absoluta (w) de las ciclistas y los gráficos de la columna derecha representan la potencia relativa de las ciclistas (w/kg). En cuanto a los puntos, como pone en la leyenda, los puntos rojos representan los resultados de los test realizados en diciembre y los puntos azules representan los test realizados en febrero del siguiente año. La ciclista 1 empeoró los resultados de los test de potencia de carácter anaeróbico y mejoró sus resultados en

los test de potencia de carácter aeróbico en términos de potencia relativa y absoluta. La ciclista 2 obtuvo mejores resultados en el febrero que en diciembre. La mejora de potencia relativa fue mayor ya que mantuvo su peso y aumento la potencia absoluta. Por último los resultados de la ciclista 3 también fueron mejores en todos los test que se realizaron, la potencia relativa aumento significativamente, esto fue debido a que entre los test hubo, no sólo una mejora de la potencia absoluta sino también una diferencia de peso de dos kilogramos.

En la siguiente ilustración se observan tres Radial Plot del Perfil de potencia de las ciclistas dependiendo el resultado obtenido en los test de potencia. Para crear los gráficos se tuvo en cuenta la tabla de Coggan (2006) y se crearon rangos de potencia relativa entre los máximos y los mínimos de la tabla, después, se obtuvo la distancia entre el mínimo y el resultado obtenido en el test y se compararon los resultados obteniendo un porcentaje.

Las tres ciclista obtuvieron resultados similares, obteniendo mejores resultados en los test de carácter aeróbico y peores resultados en los test de carácter anaeróbico. La principal diferencia entre las ciclistas fue que la ciclista 2 obtuvo mejores resultados en los test de corta duración. En general las tres ciclistas tienen un perfil de potencia similar a de las ciclistas escaladoras.

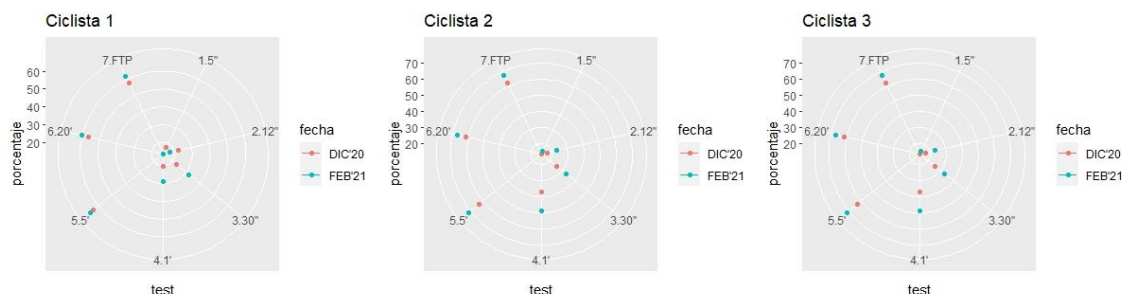


Ilustración 26. Perfil de potencia de las ciclistas. Elaboración propia.

4.4.2 Test de lactato.

- Análisis del Test de Lactato.

En el grafico a continuación se observa la potencia durante un test de lactato, las líneas verticales dividen las series que se realizaron y el lactato obtenido en la serie, Es observable como la potencia media de cada serie es mayor a la anterior.

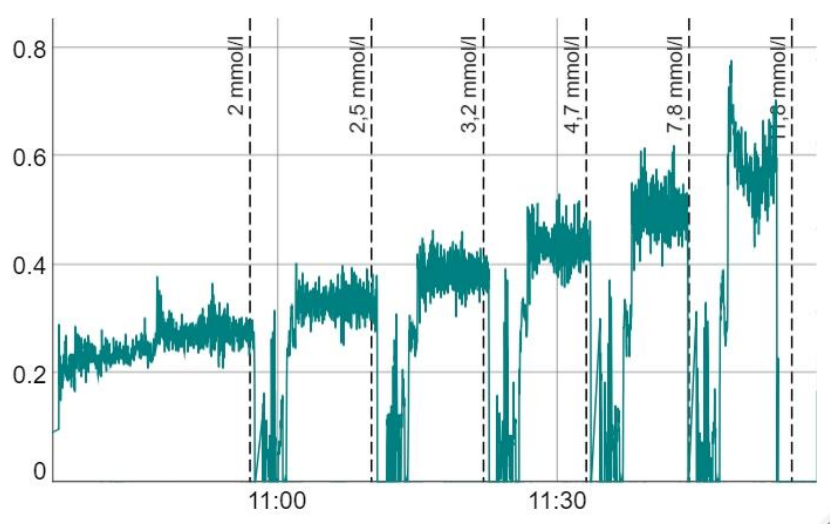


Ilustración 27 Desarrollo de potencia durante Test de Lactato. Elaboración propia.

- Evolución de los test.

Como se observa en la tabla siguiente los resultados del test fueron positivos en cuanto a la máxima potencia relativa que se desarrollo en el test. La ciclista 1 fue la única en mejorar su potencia en el LT y ninguna ciclista mejoró su potencia en el OBLA.

	Ciclista 1		Ciclista 2		Ciclista 3	
	Test 1	Test 2	Test 1	Test 2	Test 1	Test 2
W/kg Max	4,39	4,49	5,35	4,78	4,97	5,36
W/kg OBLA	4,24	4,38	5,17	5,05	4,42	4,26
W/kg LT	4,09	3,9	4,65	4,28	4,16	4,06

4.4.3 Test de resistencia a la fatiga

- Análisis del Test

En la siguiente tabal se observan las diferencias de los dos test de resistencia a la fatiga respecto al test de potencia de 20 minutos de las ciclistas. La ciclista 2 no pudo realizar el test por lo que no se tuvo en cuenta.

Ambas ciclistas obtienen mejores resultados en el primer test que en el segundo siendo destacable la poca caída de vatios de la ciclista 3 en el primer test.

	Ciclista 3		Ciclista 1	
	W	w/kg	W	w/kg
Test 1	2	0,03	23	0,38
Test 2	18	0,31	31	0,59

4.4.4 Control del entrenamiento mediante la eficiencia y el VO_2 Score.

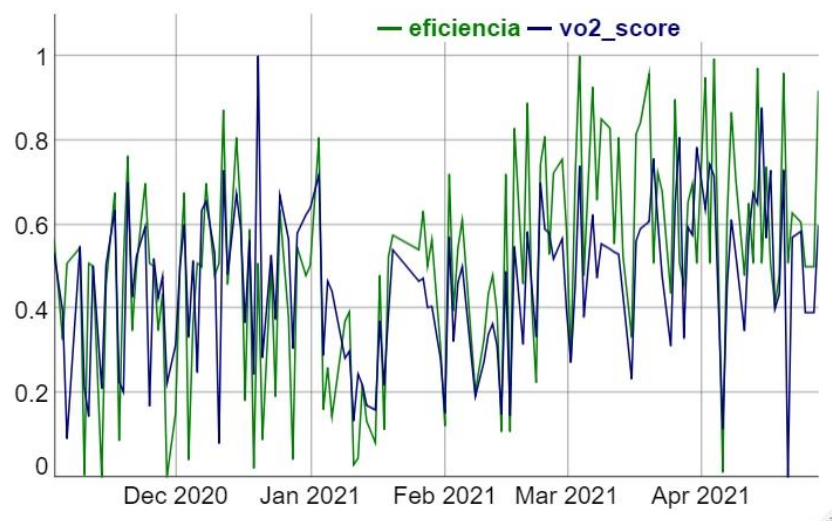


Ilustración 28 Evolución diaria de la eficiencia y el VO_2 Score. Elaboración propia.

Se observó que ambas variables estaban altamente correlacionadas, la eficiencia tenía más variación que el VO_2 Score. Estas dos variables de control del rendimiento en los entrenamientos estuvieron altamente correlacionadas con el volumen y la intensidad. En febrero se redujo la carga de entrenamiento y esto se vio visiblemente en la eficiencia y en el VO_2 Score.

4.5 Modelos para la valoración y predicción de rendimiento.

4.5.1 Modelos de influencia del volumen y la intensidad en el rendimiento.

Uno de los principios generales del entrenamiento es la individualización, cada atleta tiene una respuesta fisiológica distinta ante un mismo estímulo por lo que, es interesante saber si los atletas responden mejor al volumen o a la intensidad en el entrenamiento.

Para ello se recopilaron datos semanales de volumen e intensidad y medias semanales de las variables de rendimiento. La primera hipótesis fue que las variables que eran más sensibles a ser manipuladas en altas intensidades tendrán cierta correlación con el entrenamiento intenso. Por ello, se analizaron el máximo número de variables de rendimiento posibles.

Una vez recopilados todos los datos se procedió a hacer una regresión lineal para evaluar que atletas respondieron mejor a altas intensidades y cuales a altos volúmenes.

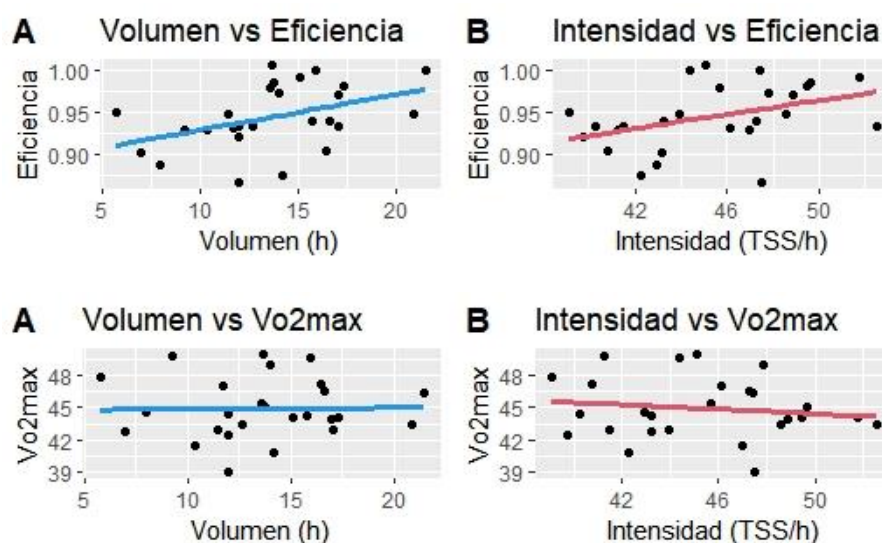


Ilustración 36 Respuesta volumen e intensidad de la ciclista 1. Elaboración propia.

Los siguientes 8 gráficos representan la respuesta semanal de rendimiento al volumen en azul, y a la intensidad en rojo.

La ciclista 1 no parece una buena respondedora al entrenamiento, ya que ambas variables no estaban muy correlacionadas a la carga del entrenamiento.

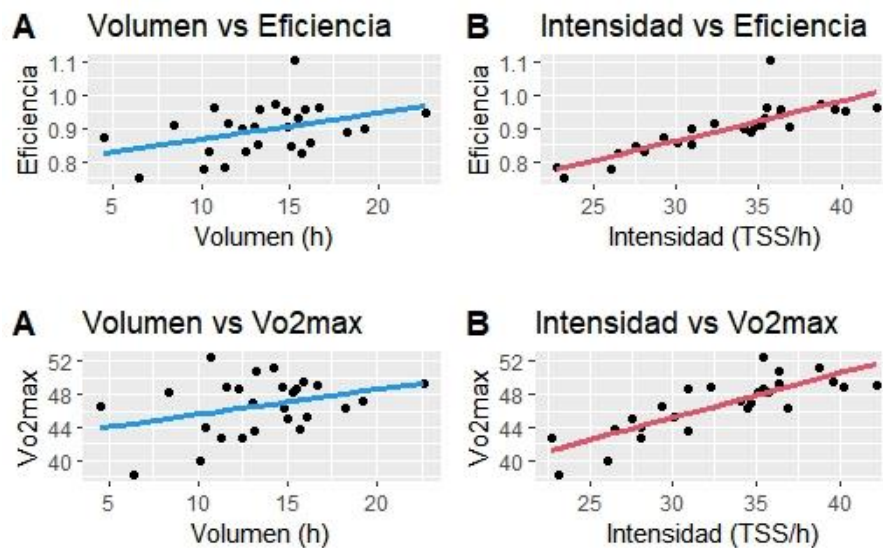


Ilustración 21 Respuesta volumen e intensidad de la ciclista 3. Elaboración propia.

En cuanto a la ciclista 2, se ve una respuesta al entrenamiento positiva teniendo alto nivel de respuesta a entrenamientos de intensidades máximas y sub máximas en las variables de eficiencia y VO_2 Score. La ciclista respondió con altos niveles de rendimiento en volúmenes de entre 10 a 17 horas semanales, por su parte el máximo rendimiento lo obtuvo en intensidades alrededor de 35 TSS/h.

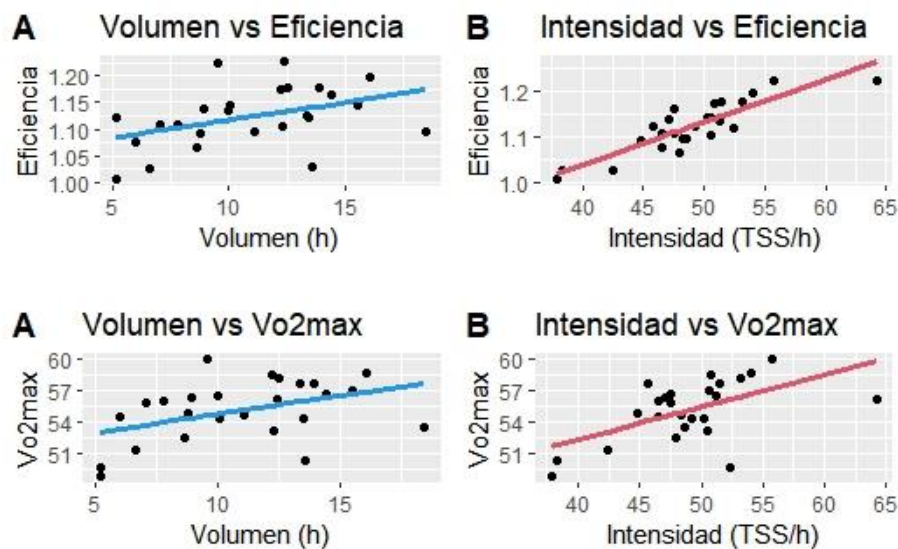


Ilustración 22 Respuesta volumen e intensidad de la ciclista 2. Elaboración propia.

Por último, la respuesta de la ciclista 3 fue positiva a altas intensidades alrededor de 50 TSS/h. No se observaron mejoras de rendimiento significativas por parte del volumen.

4.5.2 Relación de la eficiencia y del VO_2 Score con el FTP.

Se realizó una regresión lineal de la influencia del VO_2 Score y la eficiencia con el FTP. Para la regresión se escogieron los resultados de 12 test de las ciclistas. El error cuadrático medio de la regresión fue de 4,34 vatios. La eficiencia tenía mayor correlación que el VO_2 Score, esto es debido a que la regresión se realizó en términos absolutos y el VO_2 Score tiene en cuenta el peso de las ciclistas.

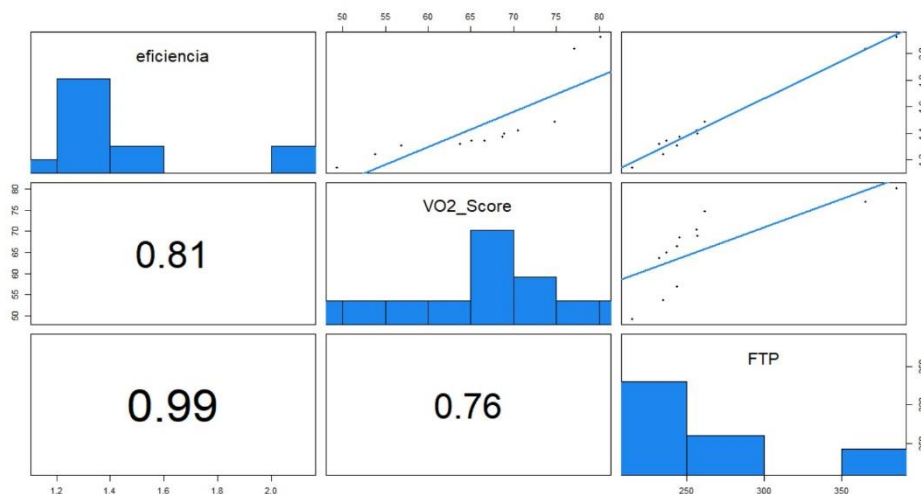


Ilustración 29 Relación de variables con el FTP. Elaboración propia.

4.5.2.1 Predicciones de rendimiento en el entrenamiento con redes neuronales.

Una vez vista la respuesta al entrenamiento de las ciclistas y la alta correlación de las variables de eficiencia y VO_2 Score con el FTP, se creó un modelo para el control futuro de la evolución de estas variables. El modelo escogido para las predicciones de rendimiento fue una red neuronal que funcionaba mediante el algoritmo de “Back Propagation”.

La red neuronal se creó con arquitectura de dos capas con 7 y 5 nodos en estas. Los datos que se introdujeron fueron el volumen de entrenamiento en segundos diario por cada zona de entrenamiento, la potencia relativa en el umbral y la forma física. Los resultados que

devolvió la red fueron, la eficiencia, la eficacia, el VO_2 Score y el torque. Para el entrenamiento de la red se usaron 500 entrenamientos de las ciclistas.

Los datos no fueron reescalados por lo que el error cuadrático medio de cada variable estará supeditados al valor de estas.

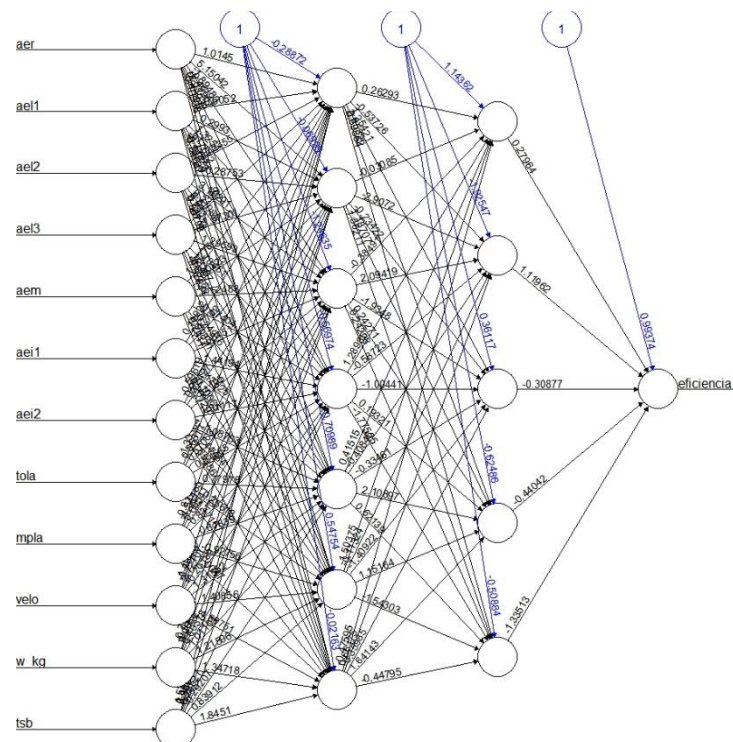


Ilustración 30 Arquitectura de la red neuronal. Elaboración propia.

- Predicciones de eficiencia.

Los resultados de eficiencia fueron muy buenos teniendo tan solo un error de 0.001, la eficiencia de las atletas oscilaba entre 1 y 2 por lo que suponía un error considerablemente bajo.

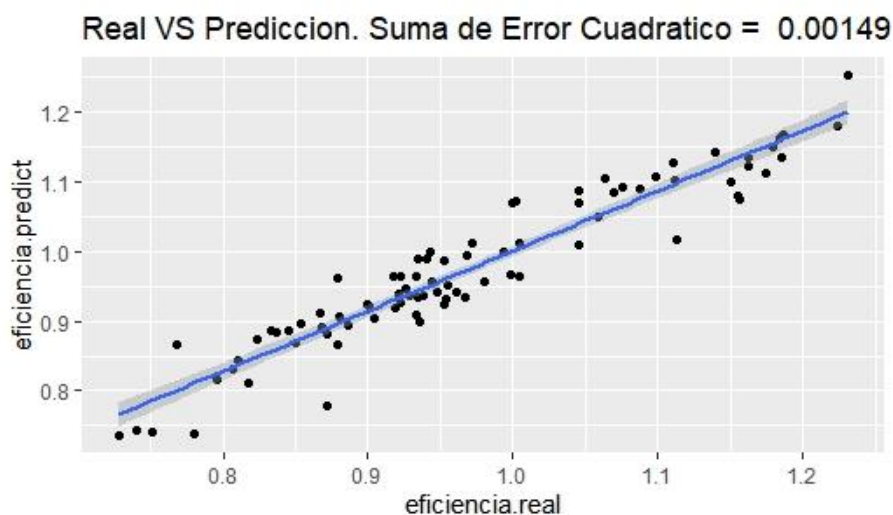


Ilustración 31 Predicciones de la eficiencia. Elaboración propia

- Predicciones de VO_2 Score.

La variable VO_2 Score fue la que mejor comportamiento tuvo en cuanto a predicciones de rendimiento, el error cuadrático medio fue de 0.516. Teniendo en cuenta que los valores de VO_2 Score de las ciclistas oscilaron entre 40 y 60, se obtuvo un resultado bastante positivo.

A pesar de que el VO_2 Score y la eficiencia sean variables bastante similares, la red se comportó mejor prediciendo la primera, esto puede deberse a que una de las variables de entrada es la potencia relativa en el umbral, y esta como el VO_2 Score, tienen en cuenta el peso de la ciclista, por lo que, esta variable de entrada redujo mucho el error en el resultado como se puede observar en el gráfico anterior.

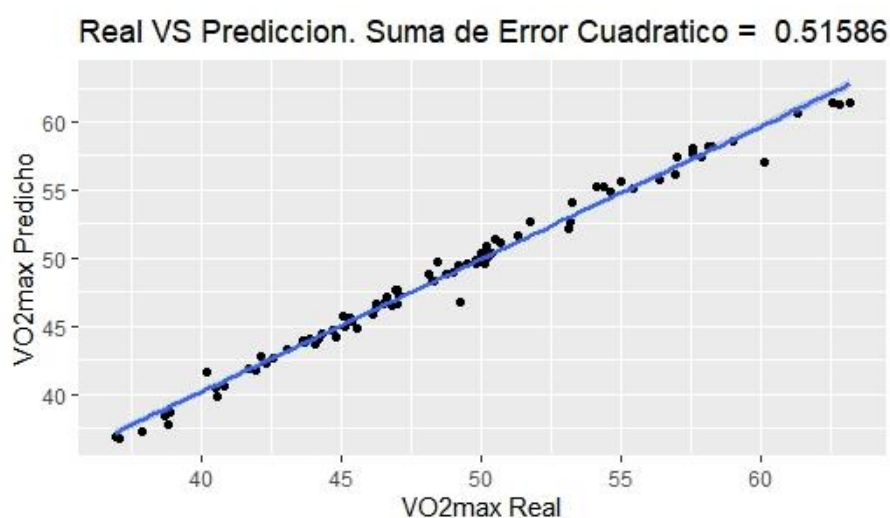


Ilustración 32 Predicciones del VO₂ score. Elaboración propia.

5 Conclusiones

Con el auge de las tecnologías es observable como cada vez más deportes están teniendo una transición a la digitalización. Esto supone una gran ventaja para los entrenadores ya que pueden realizar una valoración más objetiva de sus ciclistas y contribuir a la mejora del rendimiento en los deportes.

Este trabajo, nos ha abierto muchas posibilidades diferentes de tener un control de las ciclistas, además de ello, gracias a la estructuración de los datos se crean nuevos frentes a abordar en el mundo del entrenamiento. Por último, hemos observado que, al realizar un programa como este, el tiempo de trabajo del entrenador se reduce en gran manera ya que, el programa está automatizado, por lo que se puede realizar con datos de otros ciclistas.

6 Bibliografía

- Algarra, J., & Gorrotxategi, A. (2012). *Entrenamiento de ciclistas de ruta*. Biocorp Europa.
- Allen, H., & Coggan, A. (2010). Training and Racing With a Power Meter. *VeloPress*.
- Asker, E., & Jeukendrup. (2002). *High-Performance Cycling*. United Kingdom: Human Kinetics.
- Astrand, P. O. (1970). *Work test with bicycle ergometer*. Vanberg, Sweden: Monark-Crescent AB.
- Chian, A. (2019). *Data Science con R: Fundamentos y Aplicaciones*. Obtenido de <https://bookdown.org/BEST/DSFA/>.
- Chicharro, J., Carvajal, A., Pardo, J., Pérez, M., & Lucía, A. (1999). Physiological parameters determined at OBLA vs. a fixed heart rate of 175 beats x min⁻¹ in an incremental test performed by amateur and professional cyclists. *Japanese Journal of Physiology*.
- Elizondo, S. C., & Álvarez, J. L. (s.f.). Big data en futbol, el nuevo juego. *Istmo*, 24-25.
- Gartner. (2012). The Importance of Big Data. *Top Technology Predictions for 2013 and Beyond*.
- Hopker, J., & jobson, S. (2012). *Performance Cycling*. Londres: Bloomsdury.
- Impellizzeri, F. M., T. Ebert, A. S., Menaspà, P., Rampinini, E., & Martin, D. T. (2008). *Level ground and uphill cycling ability in elite female mountain bikers and road cyclists*. *Eur J Appl Physiol*.

- Johnson, T. D. (2003). *Exploratory Data Mining and Data Cleaning*. John Wiley & Sons, Inc.
- Lucia, A., Hoyos, J., Santalla, A., Earnest, C., & Chicharro, J. (2004). Which laboratory variable is related with time trial performance time in the Tour de France? *Br J Sports Med*, 636-640.
- Lucía, A., Hoyos, M. P., & Chicharro, J. L. (2000). Heart rate and performance parameters in elite cyclists: a longitudinal study. *Medi. Sci. Sports Exerc.*, 1777-1782.
- Manyica, J., & Brown, M. (2011). *Big Data: The next frontier for innovation competition and opportunity*. Mckinsey Global Institute.
- Mujika, I. (2012). *Endurance Training - Science and Practice*.
- Tukey, J. (1962). *The Future of Data Analysis*.
- Wickham, H., & Grolemund, G. (2017). *R for Data Science*.