Trabajo de Fin de Máster
para acceder al

MÁSTER UNIVERSITARIO EN DATA SCIENCE

Curso académico 2018 - 2019

ANÁLISIS DE DATOS DE DISPOSITIVOS DEPORTIVOS

(Data analysis of sports devices)

Trabajo realizado por **Elsa Cerezo Fernández**
Dirigido por **Francisco Matorras Weinig**
[Defensa julio - 2019]

Índice de contenido

Índice de figuras

Índice de tablas

Agradecimientos

Con este proyecto pongo fin a una larga etapa de compromiso con el estudio que da paso a una nueva etapa de crecimiento personal y profesional. Por ello, quiero agradecer, en primer lugar, a mi familia todo el apoyo que me han brindado durante toda mi formación personal y académica hasta el día de hoy.

En segundo lugar, quiero dar las gracias a mis amigos y amigas por comprenderme, animarme y ayudarme. En especial, le doy las gracias a Álvaro por su apoyo incondicional desde el primer día.

También quería dedicar unas palabras a los deportistas que me facilitaron sus datos deportivos desinteresadamente, porque sin ellos este trabajo no hubiera sido posible, gracias.

Por último, quisiera agradecer a Francisco Matorras su dedicación, constancia e interés en la dirección de mi Trabajo de Fin de Máster.

Resumen

En este trabajo se busca interpretar los datos registrados por dispositivos sobre actividad física con el objetivo de probar si es viable reproducir la frecuencia cardiaca de un individuo a partir de una serie de variables derivadas de estos datos.

Para alcanzar esta meta, se han seguido distintos pasos comunes en cualquier proyecto de análisis de datos. Primero, se han reunido los datos de distintas personas que hayan registrado su actividad física mediante diferentes dispositivos y se han procesado cambiando la estructura y formato de los ficheros que los almacenan. Después, se han liberado de datos anómalos y se les ha aplicado correcciones necesarias. El penúltimo paso ha sido la extracción de variables derivadas que aporten información relacionada con el valor de la frecuencia cardiaca y, para acabar, se ha procedido el análisis e interpretación de los resultados.

Los resultados obtenidos muestran que los modelos propuestos consiguen reproducir correctamente la variable objetivo cuando se entrenan con la misma carrera. Sin embargo, la generalización a otras carreras o a otros atletas se ha conseguido de forma parcial reproduciendo suficientemente bien las tendencias, pero faltando mayor precisión en los valores absolutos.

Palabras claves: análisis de datos, redes neuronales, predicción de la frecuencia cardiaca, inteligencia artificial.

Abstract

This work seeks to interpret the data recorded by physical activity devices in order to test the feasibility reproducing the heart rate of an individual from a set of variables.

To achieve this goal, several common steps are followed in any data analysis project. First, data is collected from different people who have registered their physical activity through different devices and are processed by changing the structure and format of the files that store them. Afterwards, anomalous data will be released, and necessary corrections will be applied. The penultimate step is the extraction of derived variables that provide information related to the value of the heart rate and, finally, we proceed to the analysis and interpretation of results.

The results obtained show that the proposed models are able to reproduce correctly the target variable when they train with the same race. However, the generalization to other races or other athletes has been partially achieved by reproducing the trends well enough, but lacking precision in the absolute values.

Keywords: data analysis, neural networks, heart rate prediction, artificial intelligence.

Capítulo 1
Introducción

En este capítulo se explica el contexto y objetivo del presente proyecto de fin de máster, así como definir su alcance.

1.1. Antecedentes y objetivo

Hoy en día se ofertan en el mercado una gran cantidad de *wearables* con un gran abanico de precios, marcas y modelos, dedicados a la medición de características relacionadas con la actividad física del usuario, tales como el geoposicionamiento, la frecuencia cardiaca, la velocidad a la que se encuentra, las calorías que gasta, etc.

Sin embargo, desde hace ya unos años esta categoría de dispositivos electrónicos ha creado una fuerte simbiosis con el mundo deportivo, ya que registrar todas estas características del deportista en un momento concreto puede ayudar a mejorar su rendimiento y detectar en qué puntos y qué factores son desfavorecedores y también puede servir para buscar perfiles de deportistas similares en redes sociales o incluso se podría detectar y/o prevenir algún problema de salud.

Con todo lo anterior, este proyecto se va a enfocar en realizar una prueba de viabilidad para la aplicación de la predicción de la frecuencia cardiaca, respecto a una serie de variables previamente seleccionadas para el análisis de distintos casos de deportistas en los que se identifiquen los factores que afectan mayormente a su rendimiento [1][2].

1.2. Alcance

El proyecto ha trabajado con ficheros dados de forma voluntaria por distintos deportistas anónimos de distintos rangos de experiencia y edad.

Para llevar a cabo el estudio, se han realizado diversos pasos que se resumen en la *Figura 1.1* y se comentan a continuación:

1. Transformar los ficheros a un formato accesible y homogéneo para todos los casos.
2. Eliminar los registros anómalos y ajustar las mediciones con correcciones.
3. En base a los datos, calcular nuevas variables derivadas entre las que se realizará una selección para el cuarto y último paso.
4. Entrenar un modelo de predicción y realizar la etapa de análisis.

Figura 1.1. Esquema de los pasos que se llevan a cabo durante el proyecto

Por otro lado, decir que los datos recogidos y tratados son datos personales que contienen fechas y datos de geoposicionamiento con los que se podría llegar a identificar a los individuos y, por lo tanto, los datos no serán publicados en acceso abierto. Sin embargo, todas las funciones y código desarrollado para el tratamiento y análisis de los datos permanecerán en un repositorio de GitHub[3].

Finalmente, hay que comentar que, dado el peso del proyecto y su duración, éste se ha planteado como un ejemplo de viabilidad y, por lo tanto, las funciones implementadas durante el desarrollo pueden ser optimizadas y deberían generalizarse de forma más sistemática. Algunas de esas mejoras se comentan en el séptimo capítulo del documento.

Capítulo 2
Recursos utilizados

A continuación, se comentan las librerías, estándares y herramientas utilizadas durante la ejecución del proyecto:

2.1. Tecnologías, librerías y herramientas utilizadas

1. **Formato GPX** [4][5]

El Sistema de Posicionamiento Global (GPS sus siglas en inglés) es un sistema que permite determinar la posición de un objeto en la Tierra definiendo un valor de latitud y otro de longitud mediante la recepción de una señal de una serie de satélites que orbitan entorno a la tierra.

Este sistema se utiliza tanto en *smartphones*, como en otros dispositivos para registrar los datos de geoposicionamiento, y para poder acceder a estos datos es necesario almacenarlos en ficheros con formatos determinados cómo el que se explica en este apartado.

GPX (*GPS eXchange Format*) es un formato de datos basado en XML para el intercambio de datos GPS entre aplicaciones y servicios web que describe puntos, recorridos y rutas.

Entre los ficheros GPX que se poseen para el proyecto, todos ellos recogen la latitud, longitud, elevación y tiempo del punto en el que se realizó el registro. Sin embargo, según el esquema, existe un apartado de extensiones cuyo contenido depende del dispositivo que registre la información. Es en este último apartado donde se puede encontrar información sobre la frecuencia cardiaca, la cadencia u otros datos de interés que dependen de la marca y/o modelo del dispositivo.

El código XML que aparece a continuación, es un ejemplo del contenido que se puede encontrar en cada uno de los ficheros de datos y en el que podemos ver que, en este caso, en el apartado de extensiones se ha recogido la frecuencia cardiaca del usuario:

```
<gpx creator="StravaGPX" xmlns:xsi="..."
                         xsi:schemaLocation="..."
                         version="1.1" xmlns="..."
                         xmlns:gpxtpx="..."
                         xmlns:gpxx="...">
   <metadata>
     <time>2018-10-04T18:20:45Z</time>
   </metadata>
   <trk>
     <name>Carrera de noche</name>
     <type>9</type>
     <trkseg>
        <trkpt lat="43.4761380" lon="-3.7989940">
           <ele>4.5</ele>
           <time>2018-10-04T18:20:45Z</time>
           <extensions>
                <gpxtpx:TrackPointExtension>
                     <gpxtpx:hr>168</gpxtpx:hr>
                </gpxtpx:TrackPointExtension>
           </extensions>
        </trkpt>
        ...
     </trkseg>
      ...
```

Figura 2.1. Ejemplo del contenido de un fichero GPX

2. **Formato TCX** [6][7][8]

Como GPX, es también un formato de datos basado en XML, que utiliza el sistema GPS y que se utiliza para la transferencia de datos referentes a la actividad física de un usuario. Además, en este caso es un formato introducido por la compañía Garmin.

Por otro lado, se podría decir que este tipo de ficheros, por lo general y como indica su esquema, incluye más información que los GPX. Además, pueden incluir cálculos promedios sobre la actividad realizada. A continuación, se muestra una parte de un fichero TCX como ejemplo del contenido que guardan:

```
<?xml version="1.0" encoding="UTF-8"?>
<TrainingCenterDatabase
  xsi:schemaLocation="..." xmlns:ns5="..." xmlns:ns3="..." xmlns:ns2="..." xmlns="..." xmlns:xsi="..." xmln
s:ns4="...">
   <Activities>
      <Activity Sport="Running">
         <Id>2016-01-03T08:18:23.000Z</Id>
         <Lap StartTime="2016-01-03T08:18:23.000Z">
           <TotalTimeSeconds>497.272</TotalTimeSeconds>
           <DistanceMeters>1000.0</DistanceMeters>
           <MaximumSpeed>2.921000003814697</MaximumSpeed>
           <Calories>110</Calories>
           <AverageHeartRateBpm>
             <Value>160</Value>
           </AverageHeartRateBpm>
           <MaximumHeartRateBpm>
             <Value>172</Value>
           </MaximumHeartRateBpm>
           <Intensity>Active</Intensity>
           <TriggerMethod>Manual</TriggerMethod>
           <Track>
             <Trackpoint>
               <Time>2016-01-03T08:18:23.000Z</Time>
               <Position>
                 <LatitudeDegrees>43.39021844789386</LatitudeDegrees>
                 <LongitudeDegrees>-3.786661634221673</LongitudeDegrees>
               </Position>
               <AltitudeMeters>113.5999984741211</AltitudeMeters>
               <DistanceMeters>0.75</DistanceMeters>
               <HeartRateBpm>
                 <Value>90</Value>
               </HeartRateBpm>
               <Extensions>
                 <TPX xmlns="...">
                   <RunCadence>42</RunCadence>
                 </TPX>
               </Extensions>
             </Trackpoint>
             ...
           </Track>
           <Extensions>
             <LX xmlns="http://www.garmin.com/xmlschemas/ActivityExtension/v2">
               <MaxRunCadence>92</MaxRunCadence>
             </LX>
             <LX xmlns="http://www.garmin.com/xmlschemas/ActivityExtension/v2">
               <AvgRunCadence>78</AvgRunCadence>
             </LX>
             <LX xmlns="http://www.garmin.com/xmlschemas/ActivityExtension/v2">
               <AvgSpeed>2.010999917984009</AvgSpeed>
             </LX>
             <LX xmlns="http://www.garmin.com/xmlschemas/ActivityExtension/v2">
               <Steps>1298</Steps>
             </LX>
           </Extensions>
         </Lap>
         ...
      </Activity> </Activities> </TrainingCenterDatabase>
```

Figura 2.2. Ejemplo del contenido de un fichero TCX

3. **Formato FIT** [9]

Es un formato y protocolo propuesto por Garmin para la transmisión de datos sobre a actividad física del usuario.

En comparación con los otros formatos que se manejan en este proyecto (GPX y TCX) es el fichero que más información recoge. Sin embargo, en este caso no es posible mostrar un ejemplo, ya que se necesitan herramientas específicas para la interpretación del fichero.

4. **Formato CSV** [10][11]

El formato CSV (*Comma-Separated Values*) contiene en su interior valores separados por comas, como su propio nombre indica.

Este formato permite almacenar datos de forma tabular, de tal forma que cada fila del documento corresponde a una fila de la tabla en cuestión. Por el contrario, cada columna está definida por los valores separados por comas y, generalmente, el fichero contiene una primera fila llamada *header* en la que se incluyen los nombres de cada columna.

5. **STRAVA** [12]

Es una red social basada en Internet y GPS, enfocada a *runners* y ciclistas principalmente, en la que utilizando un *smartphone* o un dispositivo GPS compatible, STRAVA está capacitada para visualizar y compartir estos datos, aunque también puede realizar un cierto tipo de análisis.

Una vez subes los datos a la plataforma, cabe la posibilidad de descargar los archivos en ficheros TCX.

6. **GoldenCheetah** [13][14]

GoldenCheetah es una herramienta de análisis escrita en C++ y que pertenece a un proyecto en abierto colaborativo.

Esta aplicación, ofrece muchas posibilidades de análisis de la sesión deportiva de un usuario dado un fichero de registros. También permite manipular estos datos, en particular ofrece la posibilidad de convertir archivos de un formato a otro.

7. **Python** [15]

Es un lenguaje de programación multiparadigma, que usa tipado dinámico, es multiplataforma y sigue los principios de legibilidad y transparencia.

a. Numpy [16]
   Es una extensión de Python que constituye una biblioteca de funciones matemáticas para operar con vectores y matrices.

b. Pandas [17]
   Biblioteca que sirve para la manipulación y análisis de datos en Python. Ofrece estructuras de datos y operaciones para manipular tablas y series temporales.

c. XML.etree.ElementTree [18]

Es un paquete de Python que permite la lectura de ficheros con estructura XML.

d. Matplotlib [19]

Esta biblioteca ofrece una gran variedad de posibilidades para la representación de datos.

8. **Git** [20]

Git es un software utilizado para el control de versiones, gratuito y de código abierto. A diferencia de otros sistemas de control de versiones centralizados como Apache Subversion (SVN), Git es un sistema distribuido en el que cada desarrollador tiene el histórico de versiones completo en su repositorio local.

Para la creación del repositorio remoto se ha utilizado la plataforma GitHub.[21]

9. **Trello** [22]

Es un software para administración de proyectos, utilizado para la gestión de tareas durante el desarrollo del proyecto.

10. **Jupyter** [23]

Es una aplicación de código abierto que permite crear documentos que contienen código, ecuaciones, visualizaciones y texto narrativo y que sirve para la limpieza y transformación de datos, visualización de datos y aprendizaje automático, entre otras utilidades.

11. **R** [24]

R es un lenguaje y entorno de software libre para computación estadística y gráficos. En este proyecto se ha utilizado para entrenar los modelos de redes neuronales con la librería Keras.

12. **Keras** [25][26]

Se trata de una biblioteca de redes neuronales de código abierto de alto nivel escrita en Python y que funciona por encima de TensorFlow (plataforma de desarrollo de soluciones con machine learning).

2.2. Búsqueda y selección de herramientas

Desde el comienzo del desarrollo del proyecto, no se ha discriminado ninguno de los dos lenguajes que actualmente aparecen entre los más utilizados en el ámbito del análisis de datos, R y Python. [27]

El primer paso que se ha realizado ha sido la búsqueda de librerías que permitieran leer los ficheros y poder almacenar los datos con un formato homogéneo y con una fácil accesibilidad a los datos, como es el formato CSV.

Para ello, se ha llevado a cabo un breve proceso de investigación sobre las distintas librerías que permiten la lectura de los distintos formatos de los que se disponen y, así, poder escoger la herramienta que más facilite este proceso de lectura y formateo de ficheros. Las librerías encontradas se presentan en la tabla (*Tabla 2.1*) que aparece a continuación:

	Python	**R**
GPX	- Gpxpy - Lxml - Xml.etree.ElementTree	- XML - plotKML
TCX	- Xml.etree.ElementTree - Tcxparser	- TrackeR
FIT	- Fitparse	- Fitdc

Tabla 2.1. Herramientas encontradas para la lectura de los ficheros GPX, TCX y FIT

En primer lugar, se han investigado las librerías disponibles para Python y se han encontrado algunas como Gpxpy [28], Lxml [29] y Xml.etree.ElementTree [30] que nos pueden servir para la lectura de GPX y de TCX (solo la tercera). Sin embargo, también se han descartado algunas como Tcxparser [31] y Fitparse [32] por no obtener una salida de datos fácilmente manipulable.

Por otro lado, en R se han encontrado las librerías como XML [33], plotKML [34], TrackeR [35] y Fitdc [36], pero no han dado buenos resultados, ya que las dos primeras dan problemas al leer el apartado de extensiones del fichero, puesto que concatena los valores de los subapartados de extensiones en una sola cadena de texto y, TrackeR y Fitdc que, aunque funcionan, el resultado que se obtiene con ambas tampoco facilita la manipulación de los datos.

Finalmente, dado que Xml.etree.ElementTree funciona correctamente y permite extraer los datos estructurados en XML y pasarlos a un formato tabular muy cómodamente, es la librería escogida para convertir los ficheros GPX y TCX.

Como no se ha encontrado ninguna librería que sirviese para la lectura de los ficheros FIT, se ha tomado la decisión de convertir los archivos a TCX con la aplicación GoldenCheetah y posteriormente convertir los TCX a CSV con la librería indicada en el párrafo anterior.

Capítulo 3
Exploración de los datos

En el presente apartado se comenta tanto el contenido de los datos recogidos, como el proceso de preparación que se ha seguido para poder acceder a ellos.

3.1. Adquisición y origen de los datos

En este proyecto se han utilizado datos facilitados de forma voluntaria por siete individuos y que han sido medidos por distintos dispositivos dedicados al registro de datos deportivos como, por ejemplo, *smartphones* y *smartwatches*.

Desde los distintos dispositivos se han subido los datos a la página del fabricante y desde ahí se han exportado a STRAVA. Esta plataforma ha permitido descargar los datos en forma de más de 3000 documentos (400 MB aproximadamente) y entre los que se pueden encontrar tres formatos diferentes: GPX, TCX y FIT.

3.2. Transformación a formato tabular

Originalmente los ficheros de datos que se han obtenido abarcan tres formatos distintos, con diferencias en su estructura y su contenido. Esto ha llevado a tomar la decisión de convertir todos los datos en ficheros con formato CSV para que resulte más cómodo tratar con los conjuntos de datos.

El primer paso que se ha llevado a cabo ha sido implementar dos funciones que recorriesen los ficheros GPX y TCX, leyesen los datos y los guardasen en ficheros CSV. Para ello, se ha utilizado la librería Xml.etree.ElementTree y, aunque ha funcionado correctamente para la mayoría de ficheros, en los TCX se ha tenido que detectar y eliminar un espacio que presentaban al inicio, que corrompía el formato XML y, por ello, no podían ser leídos por la librería.

Para el caso de los ficheros en formato FIT, éstos se han convertido a formato TCX, ya que era la opción en la que se perdía la menor cantidad de datos posible, y se les ha aplicado la misma función que al resto de ficheros con ese formato.

Por otro lado, en la conversión a CSV no se ha querido perder ningún dato y, por lo tanto, la mayoría de los archivos siguen difiriendo en su contenido. Por esta razón, se han seleccionado las variables que comparten y que pueden ser necesarias para el objetivo del proyecto y se han generado copias de los ficheros en formato tabular que finalmente comparten tanto estructura, como tipo de contenido. A continuación, se muestra el contenido de los ficheros CSV de los que se parte:

- sport: Este campo se representa con una cadena de texto que etiqueta el tipo de actividad a la que corresponde el registro.
- time: Fecha y hora en la que se ha guardado el registro de la actividad.

- latitud: Latitud geográfica del *trackpoint* registrado.
- longitud: Longitud geográfica del *trackpoint* registrado.
- elevacion: Altitud respecto al nivel del mar del momento en el que se registró el *trackpoint*.
- hr: Frecuencia cardíaca medida en pulsaciones por minuto (*ppm*).
- cadencia: Cantidad de pasos por minuto que se dan al correr.

Por último, en esta fase de preprocesamiento de los datos, a partir de la última estructura de CSV que se ha comentado, se han calculado una serie de variables más con el objetivo de analizarlas y detectar anomalías y aplicar correcciones, como se explica más adelante en el *Capítulo 4*. Las variables derivadas que se añaden a la estructura anterior de CSV son las siguientes:

- distancia: Distancia entre dos puntos geográficos. Para ello, se ha utilizado la fórmula de Haversine. [37]

$$2 \cdot r \cdot \arcsin\left(\sqrt{\sin^2\left(\frac{lat2-lat1}{2}\right) + \cos(lat1) \cdot \cos(lat2) \cdot \sin^2\left(\frac{lon2-lon1}{2}\right)}\right)$$

- velocidad: Velocidad en km/h que lleva el usuario entre un punto y el anterior.
- inc_ele: Diferencia de elevación entre un punto y el anterior.
- pendiente: Pendiente entre un punto y el anterior.
- velocidad_v: Velocidad vertical en km/h que lleva el usuario entre un punto y el anterior. En este caso, la distancia que se cuenta al realizar el cálculo es la diferencia de la elevación entre los dos puntos.
- direccion: Dirección del vector formado entre un punto y el anterior en grados. El cálculo se ha realizado con la siguiente fórmula:

$$\tan^{-1}\left(\frac{lat2-lat1}{lon2-lon1}\right) \cdot \frac{180}{\pi}$$

- cambio_dirr: Diferencia en grados entre la dirección de un punto y el anterior.

3.3. Análisis preliminar de los datos

Tras transformar los ficheros a CSV se han agrupado los registros según su tipo de actividad y se han representado los datos en escala logarítmica (*Figura 3.1*) para obtener una previsualización de los tipos de actividades que hay entre los datos obtenidos y el número de registros de *trackpoints* de los que se dispone entre todos los conjuntos de datos.

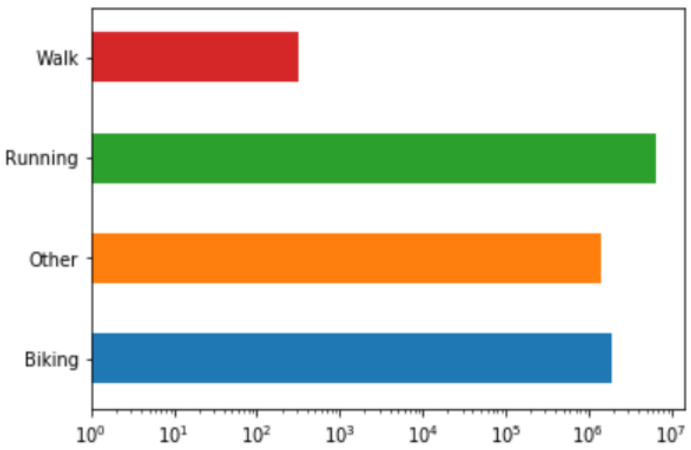

Figura 3.1. Número de registros en forma logarítmica y cómo se reparten entre los distintos tipos de actividades

Inicialmente, algunas actividades estaban etiquetadas con números, sin embargo, se disponía de información sobre las rutinas deportivas de los individuos a los que pertenecen los datos y, por lo tanto, se ha supuesto que algunas de esas etiquetas también se referían a la actividad de *Running* y el resto, las etiquetadas como *Other,* pertenecen a la actividad *Swimming*.

Finalmente, como se puede observar en el anterior gráfico, *Running* es la actividad predominante entre los datos recogidos con más de seis millones de registros, seguida de *Biking* y *Other,* con casi dos millones y millón y medio de registros, respectivamente. Por esta razón, el proyecto se centrará en la actividad de *Running*.

Capítulo 4
Curación de los datos

En este apartado se intentan localizar los problemas de medida que tienen los dispositivos para mitigar la repercusión que pueden tener en el resultado del análisis puesto que pueden darse valores muy alejados de la realidad y pueden confundir al modelo.

Este tipo de errores son en cierta manera imprevisibles porque no solo dependen del dispositivo [38], la precisión en la medida está también condicionada por el lugar en el que se encuentre el usuario. Por ejemplo, si el individuo se encuentra en una ciudad, rodeado de edificios muy altos, no tiene la misma calidad de acceso a los satélites y eso repercute en la calidad de la medida.

Como ejemplo, se muestran las siguientes imágenes (*Figura 4.1*) de una parte del recorrido de un maratón. En el primer caso, lo que se puede ver es el mapa del lugar, en el que se identifica la Avenida Diagonal de Barcelona que describe una recta. Sin embargo, en el segundo caso, podemos ver el mismo recorrido trazado por los registros del dispositivo de uno de los participantes y, como se puede observar, la recta que debería haberse descrito tiene bastantes irregularidades.

Figura 4.1. Parte del recorrido de un maratón en el que se puede ver el error en la precisión de la medida de la posición

Por otro lado, gran parte de las medidas que se utilizan en el proyecto dependen de la posición del usuario o, en su defecto, de la distancia que recorre y, por lo tanto, este error hay que tenerlo en cuenta y corregirlo. Por esta razón, se ha procedido a analizar la

distribución y desviación de los datos para poder definir tanto el umbral a partir del cual los valores se consideran anómalos, como el filtro que se ha de aplicar a las correcciones.

Por otra parte, se ha hecho un estudio superficial sobre la variable objetivo, la frecuencia cardiaca, y parece indicar la ausencia de muchas anomalías, aunque sería conveniente hacer un estudio más complejo y detallado.

A continuación, se describe el proceso que se ha llevado a cabo en la detección de anomalías y en la corrección de medidas.

Lo primero para tener en cuenta es que, aunque las funciones que se desarrollen para que se realice de forma sistemática todo el proceso de curación sobre los conjuntos de datos almacenados, las pruebas y el proceso de desarrollo estas funciones se han realizado sobre dos conjuntos de datos de los que se sabe que hay datos anómalos. Uno de ellos, el que más se ha utilizado durante las pruebas por ser del que más información se tiene, es de un maratón de 42 kilómetros de Barcelona, parte del cuál se ha mostrado en las dos imágenes anteriores. El otro *dataset* se conoce que es una media maratón y, por lo tanto, se supone que su distancia es de aproximadamente 21 kilómetros.

Continuando el proceso de curación, para definir los umbrales a partir de los cuales se considerarán que un valor es anómalo o si hay que aplicar alguna corrección, se analiza la distribución de los datos del maratón, cuyos histogramas se muestran a continuación (*Figuras 4.2, 4.3, 4.4, 4.5, 4.6 y 4.7)*. De esta forma, se podrán capturar los datos que aparezcan más a los extremos de la campana de Gauss formada en los histogramas y decidir si eliminarlos o corregirlos. Así, desaparecerán valores podrían dañar el aprendizaje del futuro modelo.

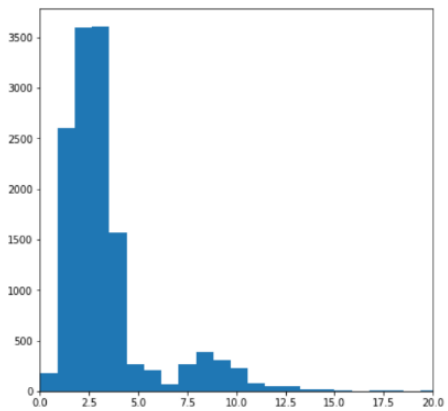

Figura 4.2. Histograma de los valores de la distancia recorrida entre los trackpoints del maratón

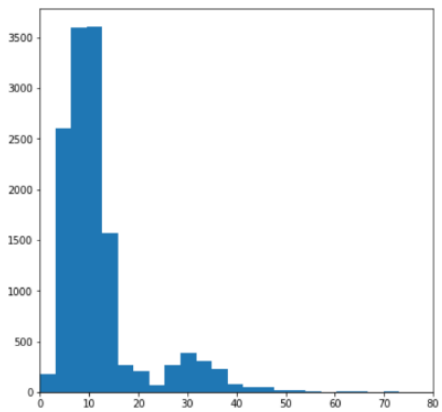

Figura 4.3. Histograma de los valores de la velocidad en cada trackpoint del maratón

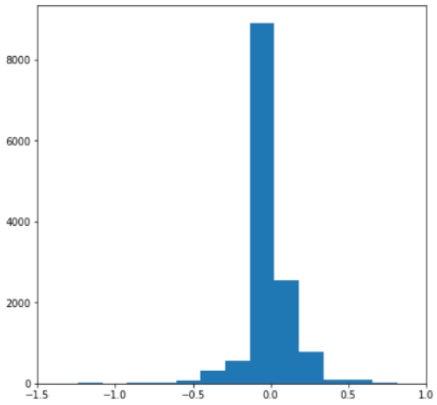

Figura 4.4. Histograma del incremento de elevación entre los trackpoints del maratón

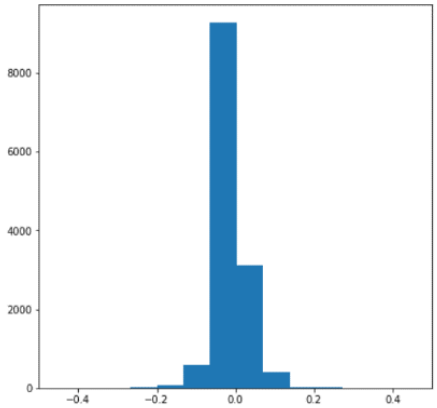

Figura 4.5. Histograma de los valores de la pendiente entre los trackpoints del maratón

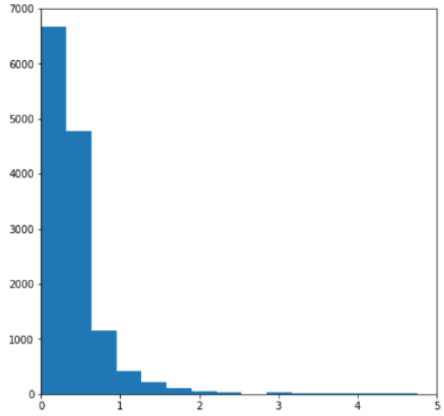

Figura 4.6. Histograma de los valores de la velocidad vertical de cada trackpoint del maratón

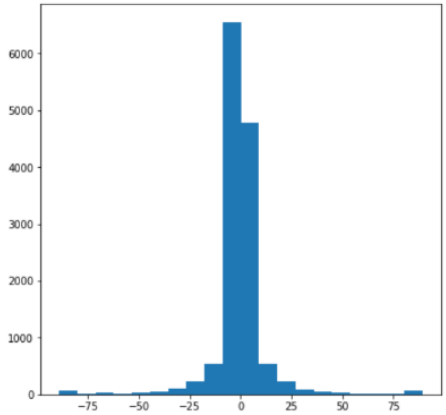

Figura 4.7. Histograma de los valores del cambio de dirección entre los trackpoints del maratón

Si comprobamos el total de distancia recorrida, su suma resulta un total de 45.279 metros, lo que quiere decir que hay un error varios kilómetros en el registro del posicionamiento del usuario. Sin embargo, se puede ver claramente en el histograma de la velocidad que hay algo que falla, es imposible que este deportista haya alcanzado velocidades de más de 70 km/h.

En los siguientes apartados se profundiza en cada una de las tareas que completan la curación de los datos.

4.1. Eliminación de anomalías

En este primer paso se eliminan directamente aquellos registros cuyos valores se alejan de la realidad de forma obvia. Por ejemplo, los valores de velocidad superiores a 45 km/h que es el récord mundial. [39]

En este segundo paso, se buscan los datos anómalos menos obvios. Por ello, después de analizar los histogramas anteriores, se define como primer filtro para detectar los valores anómalos en potencia, la suma entre el valor más común dentro de la campana de Gauss representada en los histogramas y 1’5 veces el valor de la desviación estándar del conjunto

de datos (σ), que corresponde con las zonas de la distribución de Gauss fuera de la zona amarilla que aparece en la *Figura 4.8.*

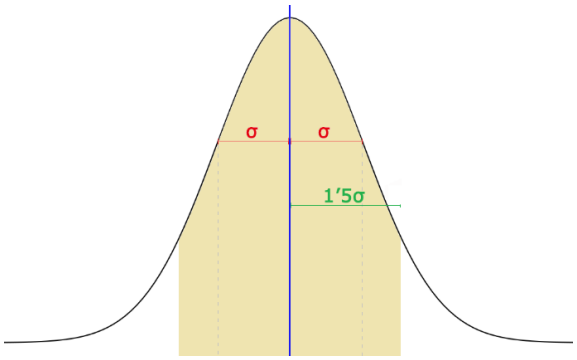

Figura 4.8. Ejemplo de campana de Gauss en la que se marca en amarillo la parte la parte en la que se considera a un valor dentro de lo posible (valor con mayor frecuencia + 1'5 sigma)

Después, de los valores que pasaron el primer filtro, se calcula la media con los tres registros anteriores y tres registros posteriores al registro que se está evaluando y si su valor supera el doble de la media calculada, se elimina.

Las variables que primero se han liberado de errores han sido la velocidad y la velocidad vertical, ya que son las variables más directas para detectar un cambio anormal en la distancia o elevación del usuario.

Por otro lado, las variables dirección y de cambio de dirección no se contemplan ya que necesitan un sistema de detección de anomalías más complejo que el utilizado en las otras métricas.

4.2. Aplicación de correcciones

Después de eliminar los datos anómalos se procede a aplicar correcciones a las variables distancia, velocidad, velocidad vertical e incremento de elevación. Algunas de ellas no reciben ninguna corrección directa, ya que las correcciones que se realizan en otras variables se ven reflejadas en ellas, como es el ejemplo de la distancia y la velocidad o el incremento de elevación y la velocidad vertical.

En primer lugar, aunque no se puede saber la distancia exacta entre dos puntos, sí que se puede estimar estadísticamente una corrección promedio que tenga en cuenta estos problemas. Para ilustrarlo se usa la distancia cartesiana.

Para esta primera corrección, se añade un ruido gaussiano de anchura σ y en promedio se debe de corregir la distancia dividida por un factor que más adelante se explicará. Este factor se puede hallar promediando l' con aproximación de δ, que es como se resuelve a continuación, o por simulación.

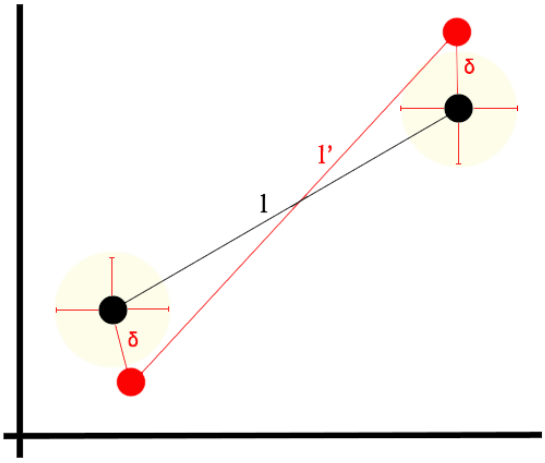

Figura 4.9. Representación de la distancia medida (l') frente a la distancia real (l) entre dos puntos

Teóricamente, si se midiesen dos puntos, como los dos puntos negros de la *Figura 4.9*, se obtendría la distancia entre ellos aplicando la fórmula $l = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$. Sin embargo, en la realidad cada coordenada de cada punto tiene un error en su medida al que se llamará δ (resultando los puntos rojos de la *Figura 4.9*), que incluso ante un error simétrico en los dos puntos, existe un efecto neto a aumentar la distancia y es eso lo que se quiere corregir. Por lo tanto, la distancia con error (l') que se obtendría resultaría de la siguiente fórmula:

$$x_1' = x_1 + \delta_1;\ y_1' = y_1 + \delta_2;\ x_2' = x_2 + \delta_3;\ y_2' = y_2 + \delta_4$$

$$l' = \sqrt{(x_2' - x_1')^2 + (y_2' - y_1')^2}$$

Si se desarrolla esta ecuación, realizando el promedio de las distancias medidas ($\overline{l'}$), se puede observar que el error que se comete en la distancia medida puede generalizarse a un factor, que es el que se aplicará para la corrección:

$\overline{l'} \cong l \cdot \left(1 + 2 \cdot \left(\frac{\sigma}{l}\right)^2\right)$, donde σ es la desviación estándar de las δ.

Por lo tanto, para hallar cada uno de los nuevos valores de distancia (l_i) se aplica el siguiente calculo:

$l_i = \frac{l_i'}{1+2\cdot\left(\frac{\sigma}{l_i'}\right)^2}$, donde l_i' es el valor de la distancia medida.

Al ser la velocidad dependiente de la distancia recorrida, solo hace falta recalcular los valores de la velocidad con los nuevos valores de la distancia para corregir la primera.

Por otro lado, se ha planteado también la corrección del incremente de elevación y de la velocidad vertical, aunque finalmente no se ha realizado por presentar más complejidad que la corrección anterior y no ser muy relevante en el análisis posterior que se ha hecho.

Después de eliminar los datos anómalos, los valores del incremento de elevación varían entre $[-1{,}1]$, y los valores de la velocidad vertical entre $[0, 5]$ m/s, con esto se quiere decir que al haber eliminado los datos que más destacaban por su anomalía los valores resultantes tienen una variación pequeña y su rango de valores es totalmente posible. Esto se ha probado en más de un conjunto de datos y es un escenario común, por esta razón, se considera que se necesitaría un sistema de corrección, en lo referente a la elevación, más complejo y preciso.

Durante el proceso de curación se ha dado más importancia al conjunto de valores que forman una carrera que a cada valor individual y, por esta razón, se ha dado más importancia a que la información general de la carrera, como la distancia total de la misma, se acercara al valor real. Aunque, por supuesto, para llegar a este fin se han tenido que eliminar los valores alejados de la realidad, aunque muchos de ellos siguen conteniendo un porcentaje de error.

Finalmente, se ha comprobado el rendimiento de los algoritmos de eliminación de anomalías y aplicación de correcciones en tres carreras con referencia exacta, un maratón y dos medias maratones. El primero se trata de un caso particular en el que se había detectado valores muy alejados de la realidad y al aplicar los métodos se redujo los 3’3 kilómetros de diferencia entre la distancia medida y la real, a tan solo 23 metros. En las dos medias maratones se aplicaron también las técnicas de curación y se redujeron los errores de 500 y 200 metros, respectivamente, a 112 y 60 metros.

Capítulo 5
Análisis

Como se ha indicado al principio del documento, el objetivo es crear un modelo que aprenda la frecuencia cardiaca de un individuo que practique running en base a una serie de características de su entrenamiento.

El modelo se va a entrenar mediante una red neuronal y las variables que se introduzcan como entrada a la red serán seleccionadas según los resultados que se vayan obteniendo en las distintas pruebas.

5.1. ¿Qué es una Red Neuronal?

Una red neuronal se trata de un modelo matemático que simula a su homólogo biológico y que está formado por capas interconectadas que, a su vez, se constituyen de nodos a los que se les llama neuronas. Estas capas son de tres tipos: de entrada (*input layer* en la *Figura 5.1*), de salida (*layer* en la *Figura 5.1*) y ocultas (*hidden* layer en la *Figura 5.1*). La capa de entrada contiene tantos nodos como variables predictoras contenga el problema, la capa de salida tiene tantos nodos como salidas se esperen dependiendo del tipo de problema a resolver y, por último, las capas ocultas o intermedias que, a diferencia de los otros dos tipos, su número es variable igual que el número de nodos que contienen.

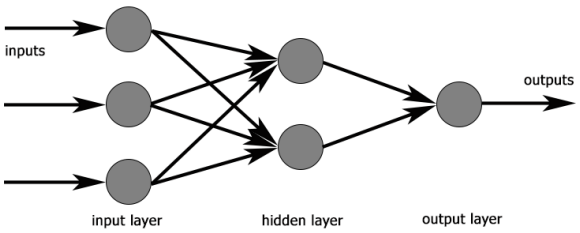

Figura 5.1. Red neuronal sencilla de tres capas

Estos sistemas aprenden y se forman a sí mismos, en lugar de ser programados de forma completa, y destacan en áreas donde la resolución de ciertos problemas es difícil de expresar con la programación convencional. Para realizar este aprendizaje automático, el objetivo es minimizar una función de pérdida o de *loss*, que evalúa la red en su totalidad. Cada una de las interconexiones entre neuronas tiene un peso distinto que se multiplica por los datos de salida de las neuronas de la capa anterior y forman la entrada de la capa siguiente. Entonces, por cada salida, la red va cambiando los pesos de sus interconexiones modificando la influencia de cada neurona para conseguir el menor *loss* posible.

Por otro lado, también existen otras características que ayudan a configurar la red neuronal. Entre ellas se pueden encontrar las funciones de activación, el *learning rate* y la función de optimización. Las primeras funciones convierten los datos que llegan a cada neurona a un nuevo rango de valores. La selección de este tipo de funciones en cada capa es importante

porque varía de un problema a otro, sobre todo en la capa de salida. El *learning rate*, en cambio, determina lo rápido que entrena la red. Una tasa demasiado alta hará que se superen los mínimos y una tasa demasiado baja tomará demasiado tiempo para converger o quedará en un mínimo local. En redes más complejas que las utilizadas en este proyecto, para lograr una convergencia más rápida y evitar los problemas anteriores, la tasa de aprendizaje varía durante el entrenamiento mediante un programa de tasas o utilizando una tasa adaptativa.

Por último, hay que comentar que para obtener el mínimo resultado en la función de *loss* se deben encontrar los valores óptimos de los pesos y, para ello, se hace uso de la función de optimización.

La correcta configuración de estos parámetros de entrenamiento es básica para mejorar el aprendizaje de las redes neuronales. Sin embargo, hay otro tipo de parámetros para redes más complejas que deben usarse cuando no se consiguen buenos resultados con los parámetros básicos.

5.2. Extracción de características

Antes de pasar directamente a crear la red neuronal y a entrenar el modelo, se han definido una serie de variables calculadas a partir de las que ya se tenían en los conjuntos de datos.

Para empezar, se han seleccionado tres características que son las que afectan más directamente al rendimiento de una persona. Éstas se pueden responden en las siguientes tres preguntas:

1. ¿Cuánto ha corrido el usuario?
2. ¿Cómo ha variado la elevación del recorrido durante la sesión?
3. ¿Qué velocidad ha llevado durante la sesión?

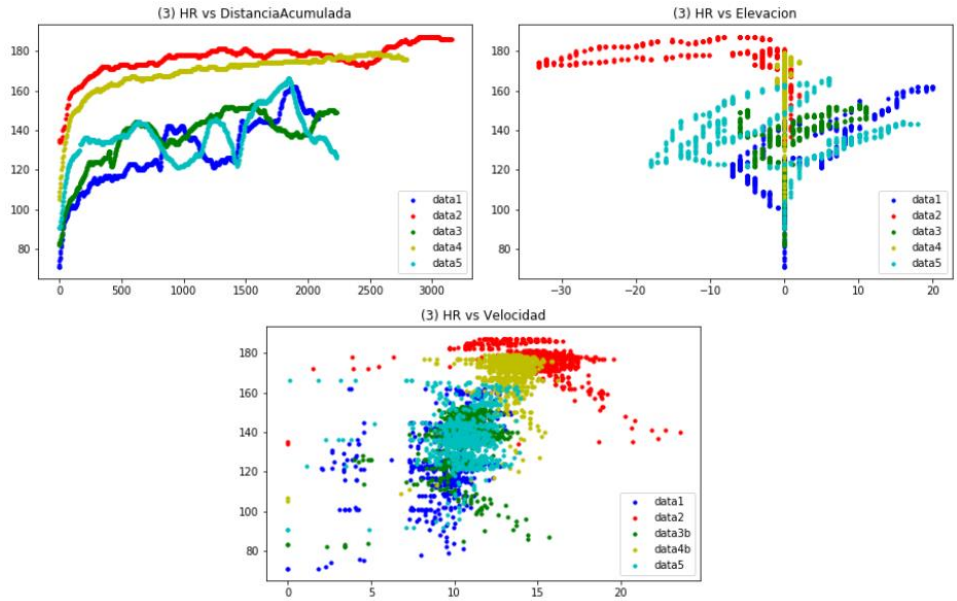

Figura 5.2. Correlación entre la frecuencia cardiaca (HR) y otras variables como la distancia acumulada (arriba izquierda), el incremento de elevación (arriba derecha) y la velocidad (abajo)

Como se muestra en la anterior figura (*Figura 5.2*), la distancia que se lleva recorrida es la que más relación tiene con la frecuencia cardiaca del usuario. Por otro lado, con la velocidad se puede observar también una fuerte dependencia, aunque probablemente se consiga una gráfica de correlación más clara representando el promedio de la velocidad en un intervalo de tiempo. Por último, el incremento de elevación durante el recorrido no muestra una clara relación, pero no se descarta, ya que se cree que también es una variable muy influyente en el rendimiento del deportista.

A partir de las tres características iniciales se definen una serie de variables que formarán parte de la estructura del input de la red de entrenamiento y entre las que se pueden encontrar cuatro grupos:

- En primer lugar, se encuentran las variables instantáneas cuyo valor es específico de cada registro, no tienen en cuenta el resto. En este grupo se encuentran la frecuencia cardiaca y la velocidad.
- En segundo lugar, se definen las variables acumuladoras, donde hay dos variables que hacen el sumatorio de los valores de otras variables durante todo el entrenamiento, el incremento de elevación acumulado y la distancia acumulada.
- El tercer tipo que se ha declarado es para las variables que acumulan el tiempo durante el que se cumplen unas condiciones como ir a cierta velocidad, pendiente o velocidad vertical.
- Por último, se pueden encontrar las variables que tienen en cuenta el esfuerzo dedicado en un intervalo de tiempo inmediatamente anterior. Este es el caso del promedio de la velocidad, el de la pendiente y el recuento de la distancia recorrida.

En un principio se definieron 30 variables, sin embargo, después de representar sus correlaciones respecto a la frecuencia cardiaca, se redefinieron los intervalos en los que se habían divido las variables porque sino algunas variables no aportaban nada. Por lo tanto, tras este primer filtro realizado, finalmente se han definido las 23 variables que pasarán al proceso de análisis:

	Nombre	Id	Tipo
X	Frecuencia cardiaca	hr	Objetivo
1	Velocidad	velocidad	Predictora
2	Incremento de elevación acumulado	incEleAcu	Predictora
3	Distancia recorrida	distanciaAcu	Predictora
	Tiempo a una velocidad vertical:		
4	- Menor que 1 m/s	tvv_menor_1	Predictora
5	- Mayor que 1 m/s	tvv_mayor_1	Predictora
	Tiempo a una pendiente:		
6	- Menor que 0	tp_menor_0	Predictora
7	- Mayor que 0	tp_mayor_0	Predictora
	Tiempo a una velocidad:		
8	- Menor que 7 km/h	tv_menor_7	Predictora
9	- Entre 7 y 11 km/h	tv_7_11	Predictora
10	- Entre 11 y 16 km/h	tv_11_16	Predictora
11	- Mayor que 16 km/h	tv_mayor_16	Predictora
	Promedio de velocidad en un intervalo de tiempo de:		
12	- 1 minuto	ventAvgV_1_min	Predictora
13	- 5 minutos	ventAvgV_5_min	Predictora
14	- 10 minutos	ventAvgV_10_min	Predictora
15	- 15 minutos	ventAvgV_15_min	Predictora
	Promedio de la pendiente en un intervalo de tiempo de:		
16	- 1 minuto	ventAvgP_1_min	Predictora
17	- 5 minutos	ventAvgP_5_min	Predictora
18	- 10 minutos	ventAvgP_10_min	Predictora
19	- 15 minutos	ventAvgP_15_min	Predictora
	Distancia recorrida en un intervalo de tiempo de:		
20	- 5 minutos	ventDAcu_5_min	Predictora
21	- 10 minutos	ventDAcu_10_min	Predictora
22	- 15 minutos	ventDAcu_15_min	Predictora

Tabla 5.1. Variables que se utilizan en el análisis: la variable objetivo (hr) y las 22 variables predictoras

Como puede verse en la *Tabla 5.1*, existen dos tipos en los que se divide el conjunto de variables: objetivo y predictora. El primer tipo hace referencia a la variable que se quiere predecir y el segundo a las variables que servirán como entrada para que el modelo pueda predecir la variable objetivo.

Después de definir el conjunto de variables, se ha realizado un preanálisis en el que se ha observado la correlación entre los distintos predictores y la variable objetivo, que se puede observar en la columna resaltada de la *Figura 5.3*, para analizar la previsible importancia de las variables predictoras en el modelo. A continuación, se muestra el resultado obtenido:

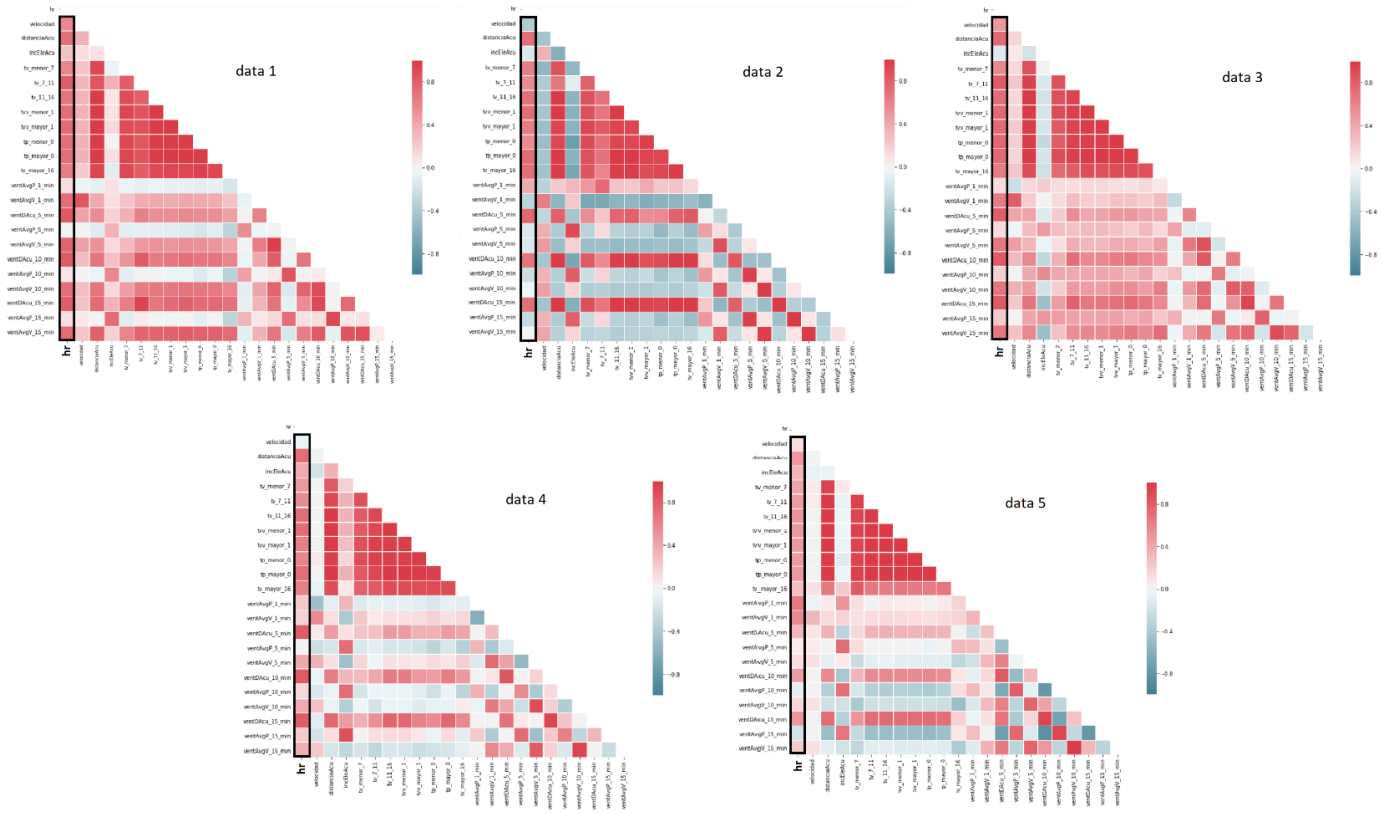

Figura 5.3. Heatmaps de cinco carreras distintas que muestran la correlación entre las distintas variables

Como se puede observar en la *Figura 5.3*, se muestran cinco *heatmaps* de cinco carreras distintas donde podemos ver la correlación entre las distintas variables y que, como es previsible es alta en muchos de los casos. Hay variables que en algunos casos su correlación respecto a la frecuencia cardiaca es débil, pero se ha decidido mantenerlas en el conjunto que se utilizará como entrada en la red neuronal, ya que en otros casos tienen más fuerza y también presenta relación con otras variables predictoras. Por ejemplo, el caso que se muestra en la *Figura 5.4*, muchas las carreras que se presentan varían su elevación durante el recorrido, sin embargo, parece que tiene muy poca importancia según los *heatmaps*, pero es obvio que sí la tiene.

Figura 5.4. Gráfica en la que se representa la variación de la elevación durante el recorrido de cinco carreras

A continuación, se van a comentar más específicamente las variables derivadas y su relación respecto a la frecuencia cardiaca:

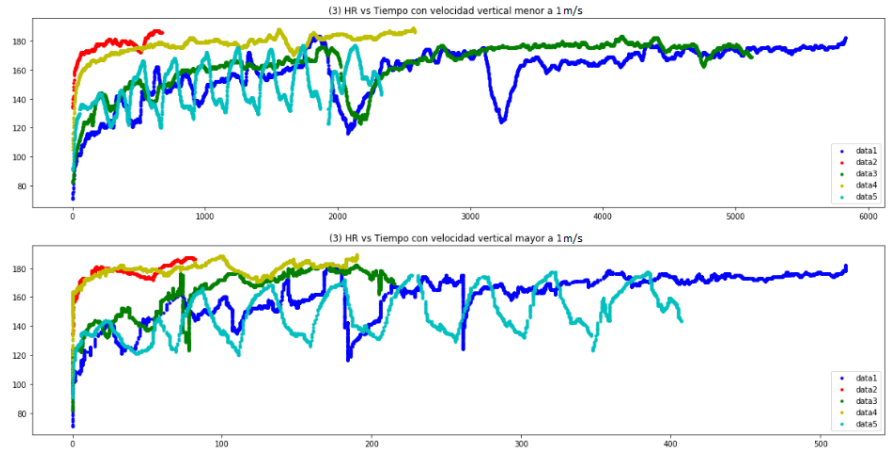

Figura 5.5. Relación entre la frecuencia cardiaca y el tiempo a una determinada velocidad vertical

En las variables que acumulan el tiempo cuanto se cumple un cierto valor, en sus gráficas se muestra una clara relación, ya que a medida que aumenta la cantidad de tiempo la frecuencia cardiaca también aumenta. En este caso, se puede observar una más clara dependencia cuándo se cuenta el tiempo a velocidad vertical mayor que 1 m/s en la *Figura 5.5*.

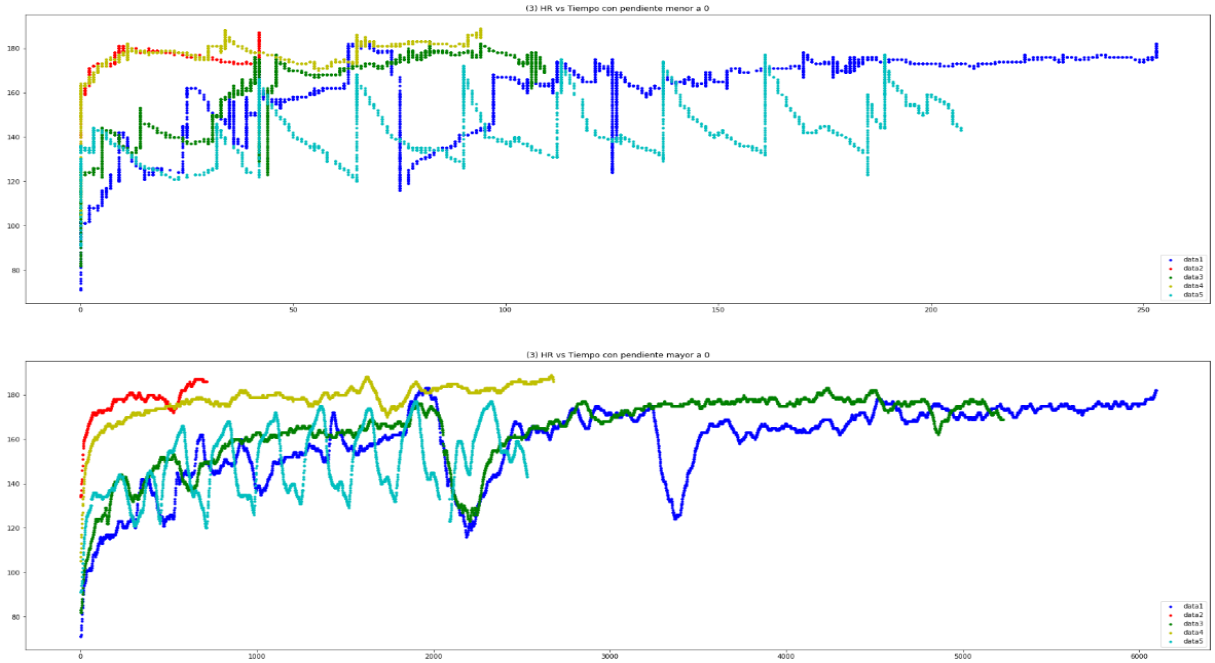

Figura 5.6. Relación entre la frecuencia cardiaca y el tiempo a una determinada pendiente

Por otro lado, se observa en la *Figura 5.6* que el tiempo acumulado con una pendiente mayor a 0 también marca mucho más la relación que tiene esta variable respecto a la frecuencia cardiaca.

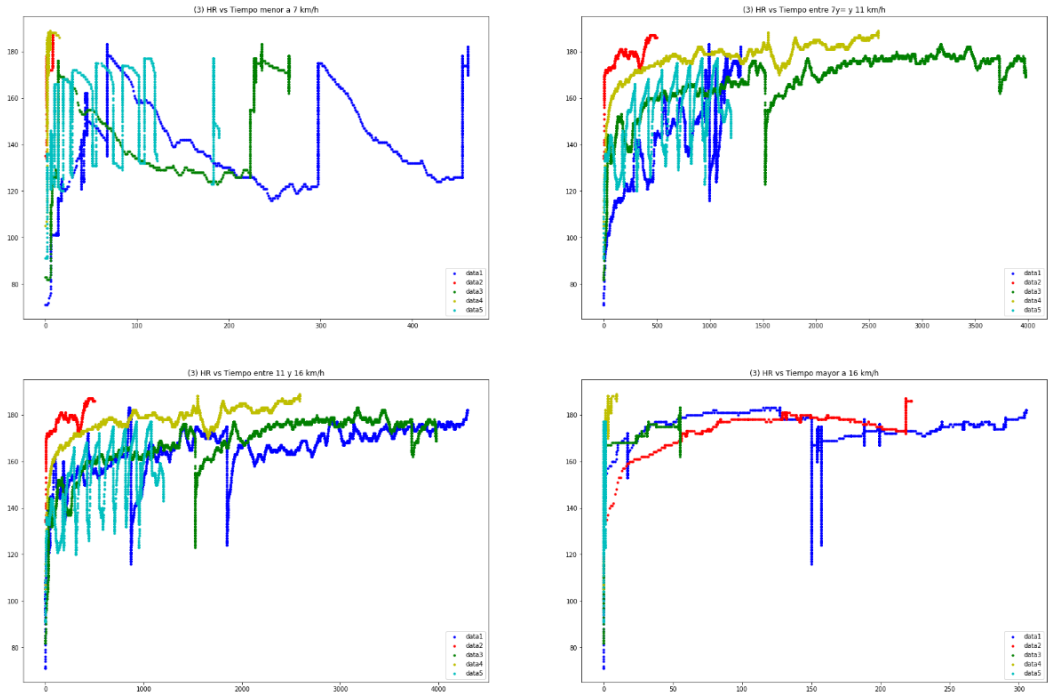

Figura 5.7. Relación entre la frecuencia cardiaca y el tiempo a una determinada velocidad

En este otro caso (*Figura 5.7)*, contemplando las variables que cuentan el tiempo durante el que se corre a una cierta velocidad, cuando la velocidad es menor que 7 km/h y mayor que 16 km/h es cierto que no se ve una clara relación entre las variables y la frecuencia cardiaca, a diferencia de los otros dos intervalos. Sin embargo, en el *heatmap* de la *Figura 5.3* se puede ver que en muchos de los casos los índices de correlación entre las dos primeras variables y la frecuencia cardiaca no son malos, por esta razón, se han mantenido dentro del conjunto de variables predictoras porque, aparentemente, con esta división de intervalos se captura más información.

Por otro lado, en el caso de las variables promediadas, concretamente la de la velocidad (*Figura 5.8*) y la de la distancia acumulada (*Figura 5.9*) sí que se puede ver la relación entre el predictor y la variable objetivo en todos los casos, a excepción de la variable promediada de la velocidad en el intervalo de un minuto, en la que la relación es algo menos clara.

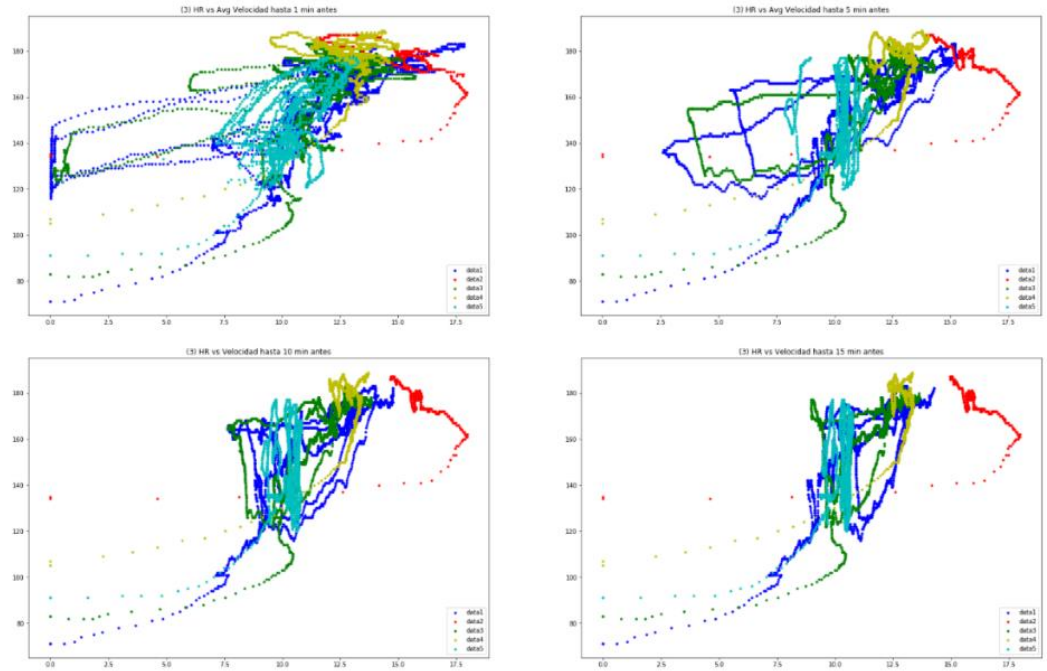

Figura 5.8. Relación entre la frecuencia cardiaca y el promedio de velocidad en cuatro intervalos de tiempo

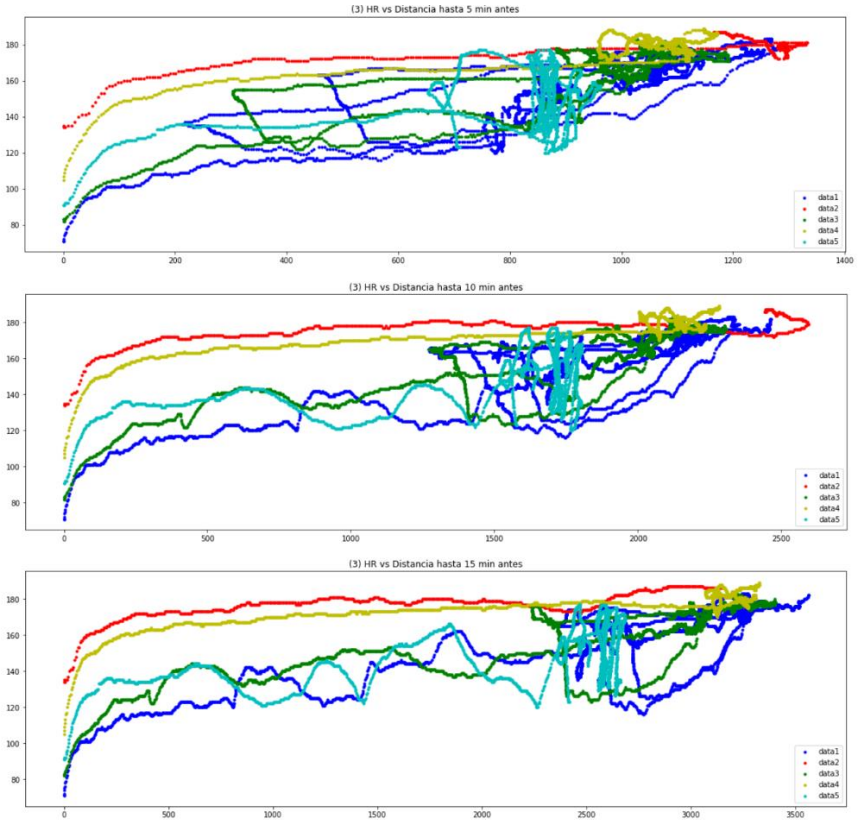

Figura 5.9. Relación entre la frecuencia cardiaca y la distancia acumulada en tres intervalos de tiempo

Sin embargo, en el caso del promedio de la pendiente (*Figura 5.10)*, aunque en algún caso el *heatmap* de correlación sí muestra un índice aceptable, en las gráficas que se muestran a continuación no es evidente que haya algún tipo de relación. Aún así, se han conservado para el proceso de entrenamiento por los resultados del *heatmap* y, si es necesario, será en ese proceso cuando se procederá a eliminar estas variables predictoras.

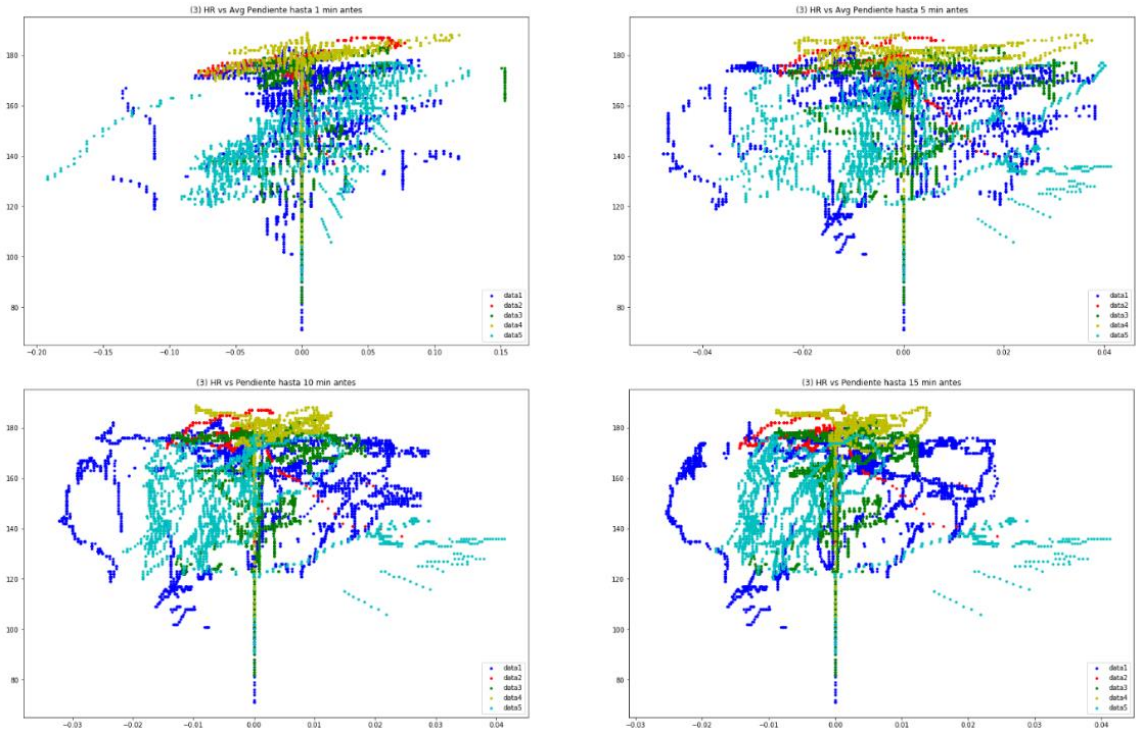

Figura 5.10. Relación entre la frecuencia cardiaca y el promedio de la pendiente en cuatro intervalos de tiempo

5.3. Entrenamiento

Ahora que se han definido las variables que se utilizarán tanto para entrenar el modelo, como para realizar las predicciones, se da comienzo al proceso de entrenamiento con el objetivo de entrenar un modelo que prediga la frecuencia cardiaca lo mejor posible.

Para empezar, se han probado seis modelos distintos:

1. Modelo 1: Este modelo se ha entrenado con una carrera de una persona cualquiera y se han utilizado todas las variables predictoras definidas en un principio.
2. Modelo 2: El segundo modelo se ha entrenado también con la misma sesión de entrenamiento que el primer modelo, pero, en este caso, no se han utilizado todos los predictores.
3. Modelo 3: En el entrenamiento de este modelo se utilizan todas las variables y se realiza sobre varias carreras de la misma persona que en los anteriores modelos.
4. Modelo 4: En el cuarto modelo se ha entrenado igual que el tercer modelo a excepción del número de variables predictoras, que en este caso se ha escogido un subconjunto del total.
5. Modelo 5: Utiliza una arquitectura de red igual que el tercer modelo y es entrenado con cinco carreras de cuatro individuos diferentes.
6. Modelo 6: Este modelo es prácticamente igual que el quinto, a excepción de que el número de variables predictoras utilizadas en este caso es menor.

Después de varias pruebas en las que se ha variado el número de capas ocultas, el número de sus nodos y las funciones de activación utilizadas, para todos los modelos se ha seleccionado una arquitectura de red neuronal sencilla, ya que el *loss* conseguido es menor en este último caso, el tiempo de cómputo también es menor y converge antes. Todos los modelos utilizan una red formada por dos capas intermedias, con 20 y 10 nodos, respectivamente, y una capa de salida de un solo nodo, esto se puede ver en la *Figura 5.11*.

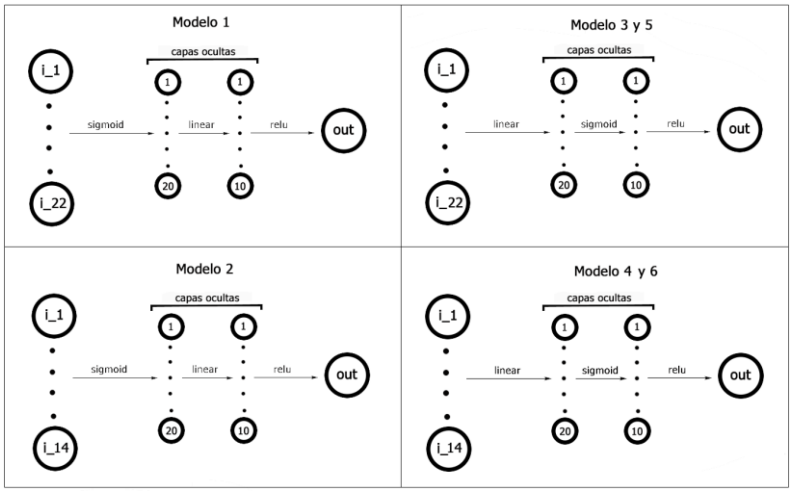

Figura 5.11. Arquitectura de las redes neuronales de los modelos probados

El *loss* de cada modelo se ha medido Mean Absolute Error (MAE), ya que como se ha dicho anteriormente no se esperan datos anómalos en la variable objetivo y como, además, también se utilizan carreras por intervalos en las que hay bastante fluctuación en la frecuencia cardiaca, puede haber valores más alejados de la media que sean totalmente nomrales, por estas razones, se ha considerado que la mejor función de *loss* a utilizar en este caso es MAE. [40]

Por otro lado, en la capa de salida se ha utilizado la función de activación ReLU (Rectified Linear Unit) [41], que se puede ver su representación gráfica en la *Figura 5.12* y se define como: $f(x) = max(0, x)$, donde x es la entrada de la neurona. De esta forma, como se está prediciendo la frecuencia cardiaca, no se obtendrán valores negativos como salida.

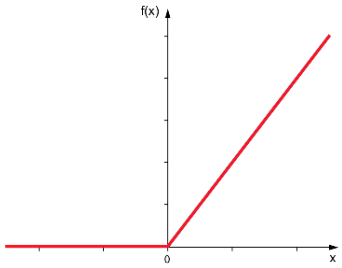

Figura 5.12. Función de activación ReLU

Asimismo, hay que comentar que al tratarse de redes tan sencillas se ha utilizado el optimizador SGD (Stochastic Gradient Descent) y un *learning rate* de 0.01. Al utilizar valores menores este último parámetro no se conseguían buenos resultados y el error permanecía muy alto, por el lado contrario, con valores mayores se descontrolaba y el error oscilaba demasiado, no mostraba un descenso gradual, que es el efecto que sí se consigue con el *learning rate* escogido.

Como se comenta en el próximo capítulo, las variables seleccionadas para los modelos 2 y 4 se han escogido a partir del primer modelo en el que se ha obtenido una buena predicción y se ha realizado un test de correlación entre la salida y variable real medida. Esta decisión de disminuir el número de variables se ha tomado no tanto por simplificar, sino con el objetivo de eliminar variables superfluas y obtener una mejor generalización.

Capítulo 6
Presentación de los resultados

A lo largo de este capítulo se van a resumir los principales resultados de aplicar los modelos descritos anteriormente, a distintas carreras disponibles.

Las pruebas que se van a realizar son las siguientes:

- **Prueba 1**: Los modelos se prueban en el mismo conjunto de datos del entrenamiento.
- **Prueba 2**: Los modelos se prueban en una carrera distinta al conjunto de datos del entrenamiento, pero que pertenece a la misma persona.
- **Prueba 3**: Los modelos se prueban en una carrera distinta al conjunto de datos del entrenamiento y que no pertenece a la misma persona.

Además, para poder conocer correctamente en todo momento las características de los modelos que se están utilizando durante las pruebas, se les nombrará siguiendo la siguiente lógica:

nc_mp_kv, donde:

- n es el número de carreras (c) con las que se ha entrenado el modelo.
- m es el número de personas (p) de las que se han escogido las carreras para el entrenamiento.
- k es el número de variables predictoras (v) que se utilizan en el modelo.

Siguiendo esto, los modelos se identificarán de la siguiente forma:

Modelo	Identificador
Modelo 1	*1c_1p_22v*
Modelo 2	*1c_1p_14v*
Modelo 3	*5c_1p_22v*
Modelo 4	*5c_1p_14v*
Modelo 5	*5c_4p_22v*
Modelo 6	*5c_4p_14v*

Tabla 6.1. Identificación de los modelos

- **Prueba 1**

Predicción vs Test	1c_1p_22v	1c_1p_14v	5c_1p_22v	5c_1p_14v	5c_4p_22v	5c_4p_14v
MAE	4.546	4.135	2.184	2.876	3.105	4.548
Correlación Spearman	0.961	0.932	0.975	0.944	0.960	0.884
Ratio varianzas	0.781	0.962	0.849	0.848	1.001	0.811

Tabla 6.2. Resultados de la primera prueba de los seis modelos donde se prueban en parte del conjunto que ha sido utilizado para entrenarlos

Como ya se ha comentado, en la primera prueba se ha utilizado el 15% de cada conjunto de datos seleccionado para evaluar cada modelo. Los resultados obtenidos en esta primera prueba, presentados en la *Tabla 6.2*, son muy buenos: la correlación entre la predicción y la frecuencia cardiaca del conjunto de test, en todos los casos, es positiva y casi perfecta, al igual que el ratio de varianzas y, por lo tanto, se puede decir que los modelos capturan correctamente el rango de valores observado.

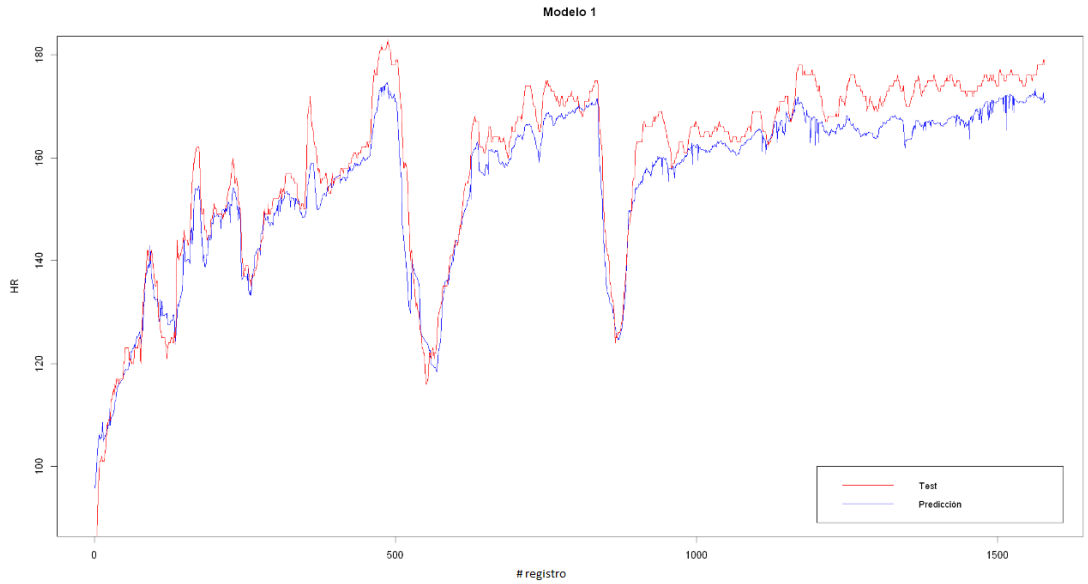

Figura 6.1. Resultado de la prueba 1 en el modelo 1c_1p_22v realizada sobre una parte del conjunto con el que se ha entrenado, donde se representa la frecuencia cardiaca predicha (línea azul) frente a la muestra de test (línea roja)

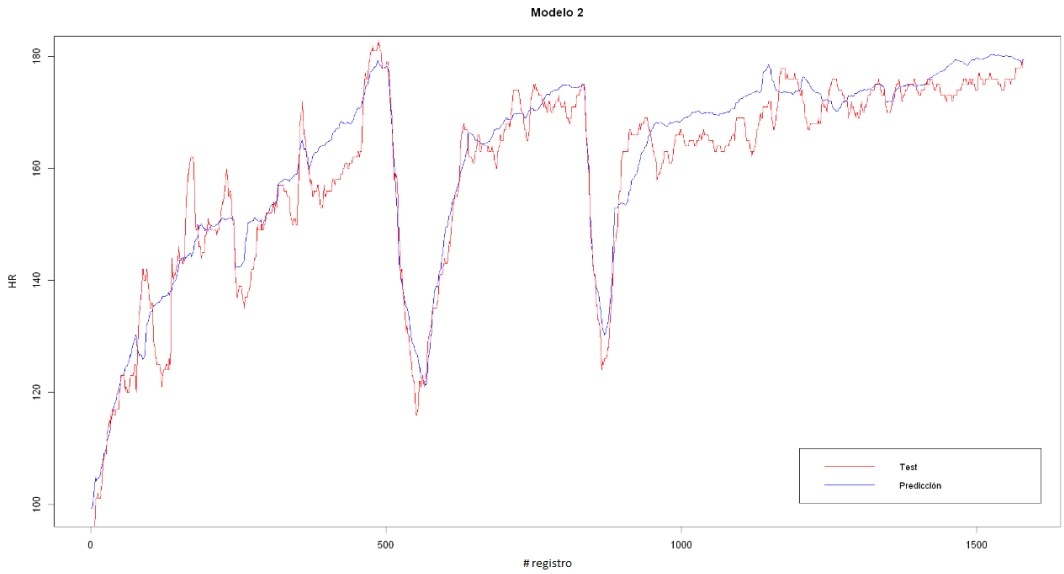

Figura 6.2. Resultado de la prueba 1 en el modelo 1c_1p_14v realizada sobre una parte del conjunto con el que se ha entrenado, donde se representa la frecuencia cardiaca predicha (línea azul) frente a la muestra de test (línea roja)

Como ya se ha dicho, en esta prueba, para todos los modelos, los conjuntos de test utilizados son una parte del conjunto que se ha usado para entrenar cada modelo, pero cada conjunto es diferente en cada caso: Los modelos 1c_1p_22v y 1c_1p_14v utilizan un conjunto de datos

constituido por una sola carrera que pertenece a una persona y contienen 22 y 14 variables predictoras, respectivamente; en cambio, las pruebas de los modelos 5c_1p_22v y 5c_1p_14v se realizan en un conjunto de datos constituido por cinco carreras de una misma persona que utilizan 22 y 14 variables predictoras en cada caso, y los modelos 5c_4p_22v y 5c_4p_14v que se entrenan también con un conjunto de datos de 22 y 14 variables predictoras en cada caso, constituido por cinco carreras de cuatro personas diferentes. Además, todas las carreras utilizadas para los entrenamientos de los modelos son continuas y sin oscilaciones notables de la frecuencia cardiaca.

Los resultados en los seis modelos han sido muy buenos, sin embargo, con los modelos 1c_1p_14v, 5c_1p_22v y 5c_4p_22v, cuyos resultados pueden observarse en las *Figuras 6.2, 6.3 y 6.5,* las predicciones que se han conseguido han sido especialmente buenas. Concretamente, el entrenamiento realizado con cinco carreras de cuatro personas diferentes y 22 variables predictoras ha conseguido una predicción casi perfecta.

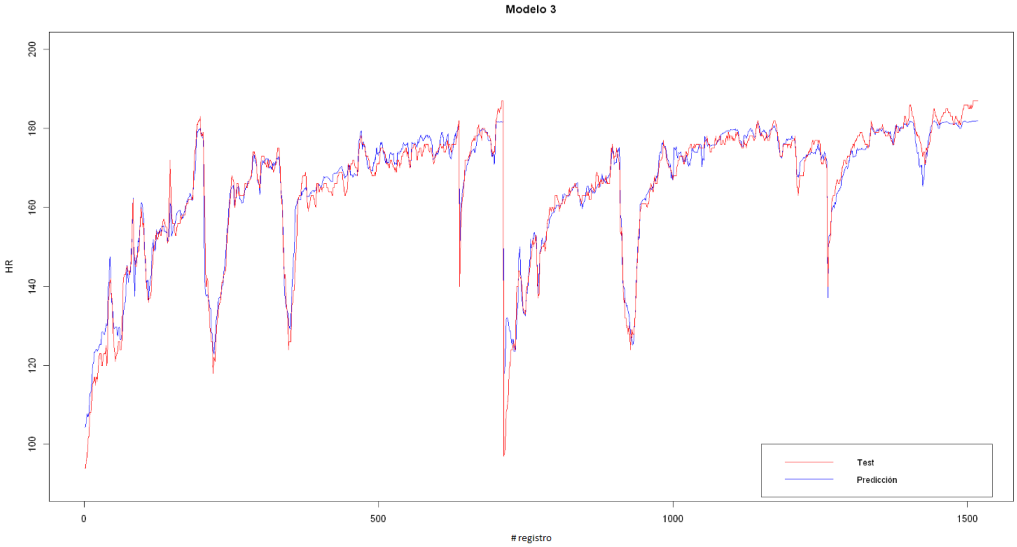

Figura 6.3. Resultado de la prueba 1 sobre el modelo 5c_1p_22v realizada sobre una parte del conjunto con el que se ha entrenado, donde se representa la frecuencia cardiaca predicha (línea azul) frente a la muestra de test (línea roja)

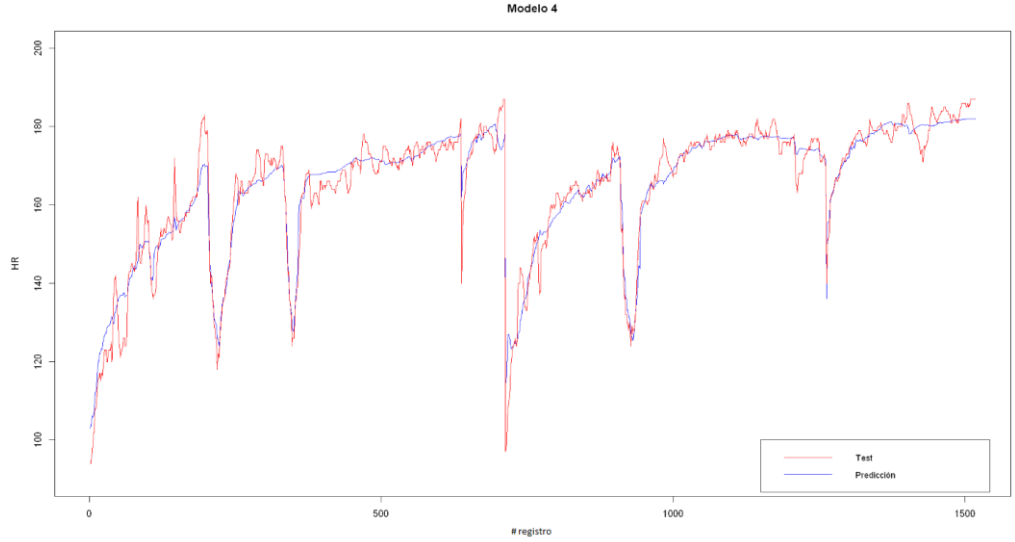

Figura 6.4 Resultado de la prueba 1 sobre el modelo 5c_1p_14v realizada sobre una parte del conjunto con el que se ha entrenado, donde se representa la frecuencia cardiaca predicha (línea azul) frente a la muestra de test (línea roja)

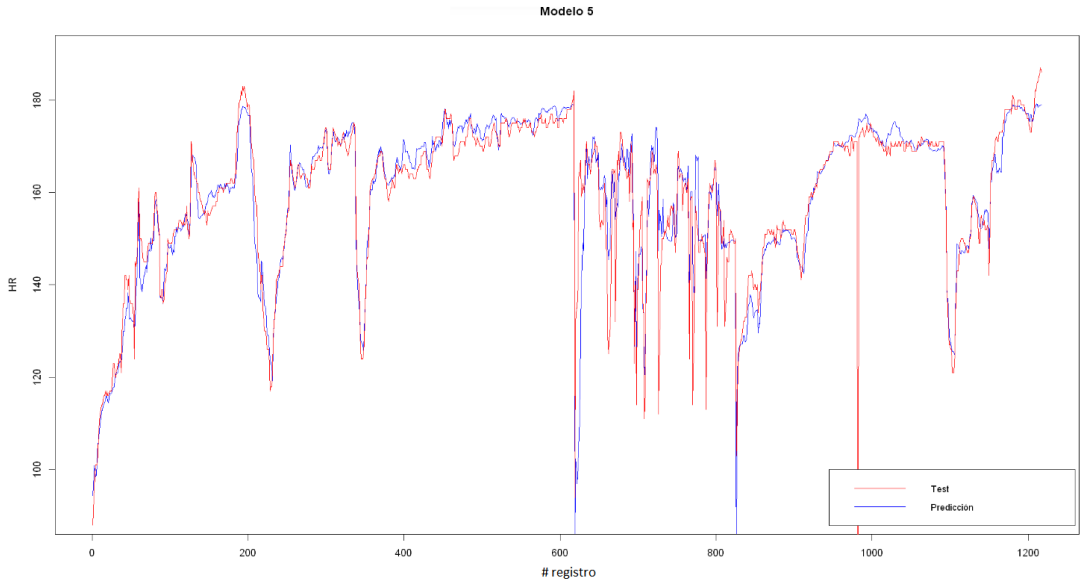

Figura 6.5. Resultado de la prueba 1 del modelo 5c_4p_22v realizada sobre una parte del conjunto con el que se ha entrenado, donde se representa la frecuencia cardiaca predicha (línea azul) frente a la muestra de test (línea roja)

Después de realizar la prueba con el modelo 1c_1p_22v se ha representado en un gráfico que la correlación entre la variable predicha y los predictores para observar qué variables son las más influyentes, seleccionarlas y formar parte del conjunto de variables predictoras para los modelos que solo utilizan 14 variables predictoras. Además, esta información también nos puede ayudar a analizar la dependencia del esfuerzo con los distintos parámetros, lo que se podría utilizar para aumentar el rendimiento deportivo. A continuación, se muestran algunos ejemplos de estas correlaciones, pero se puede encontrar el gráfico completo en el *Anexo 8.1*.

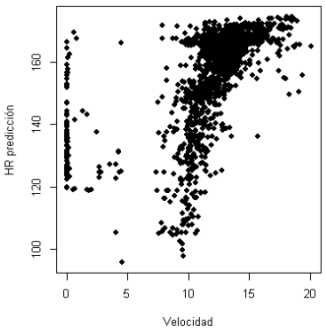

Figura 6.6. Frecuencia cardiaca predicha frente a la velocidad en cada punto del recorrido

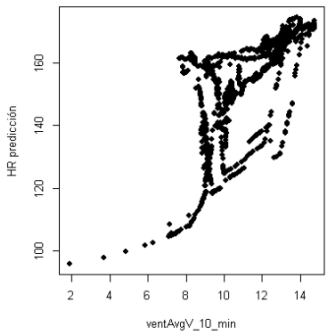

Figura 6.7. Frecuencia cardiaca predicha frente al promedio de la velocidad en un intervalo de 10 minutos

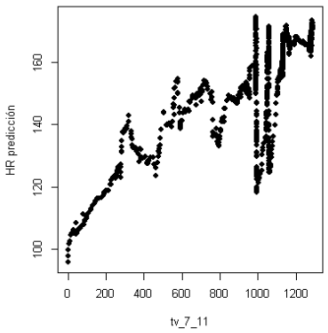

Figura 6.8. Frecuencia cardiaca predicha frente al tiempo que se ha estado a una velocidad de entre 7 y 11 km/h

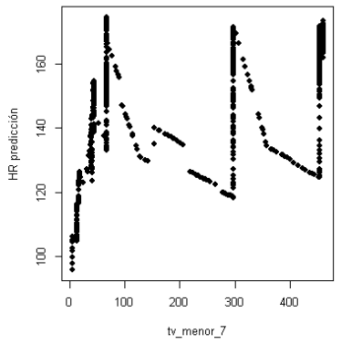

Figura 6.9. Frecuencia cardiaca predicha frente al tiempo que se ha estado a una velocidad menor que 7 km/h

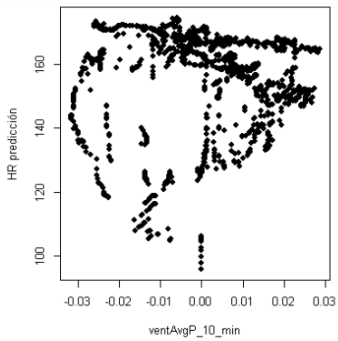

Figura 6.10. Frecuencia cardiaca predicha frente al promedio de la pendiente en un intervalo de 10 minutos

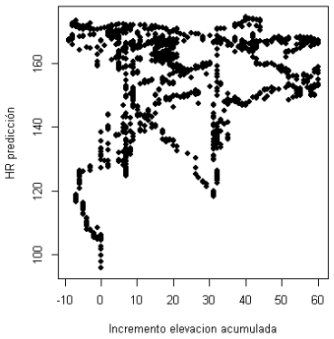

Figura 6.11. Frecuencia cardiaca predicha frente al incremento de elevación acumulado

En el gráfico de correlaciones, se ha podido observar que las variables que no muestran una clara relación son: la velocidad en cada punto concreto (*Figura 6.6*), a diferencia que la velocidad promedio (p.e. *Figura 6.7*) y el recuento de tiempo a cierta velocidad (p.e. *Figura 6.8*) en las que se ha visto en el caso de la velocidad al tener en cuenta un intervalo del entrenamiento se ha podido detectar mejor la dependencia con la variable objetivo; el incremento de elevación acumulado (*Figura 6.11*), los contadores de tiempo a una velocidad menor que 7 (*Figura 6.9*) y mayor a 16 km/h y las variables que presentan el promedio de la pendiente en un intervalo de tiempo (p.e. *Figura 6.10*). De esta forma, se ha pasado de 22 variables predictoras a 14, que son las que se han seleccionado para los modelos 1c_1p_14v, 5c_1p_14v y 5c_4p_14v. Dados los resultados obtenidos, se podría reflexionar sobre el uso de solo 14 variables para reproducir la frecuencia cardiaca correctamente. Sin embargo, hay que examinar el resultado del resto de las pruebas para confirmarlo, porque en el caso de los resultados obtenidos en el modelo 5c_4p_14v (*Figura 6.12*) se puede observar que se reduce la calidad de la predicción. Además, los resultados de correlación entre variables obtenidos no son 100% concluyentes, ya que la información puede venir de la combinación de dos o más variables.

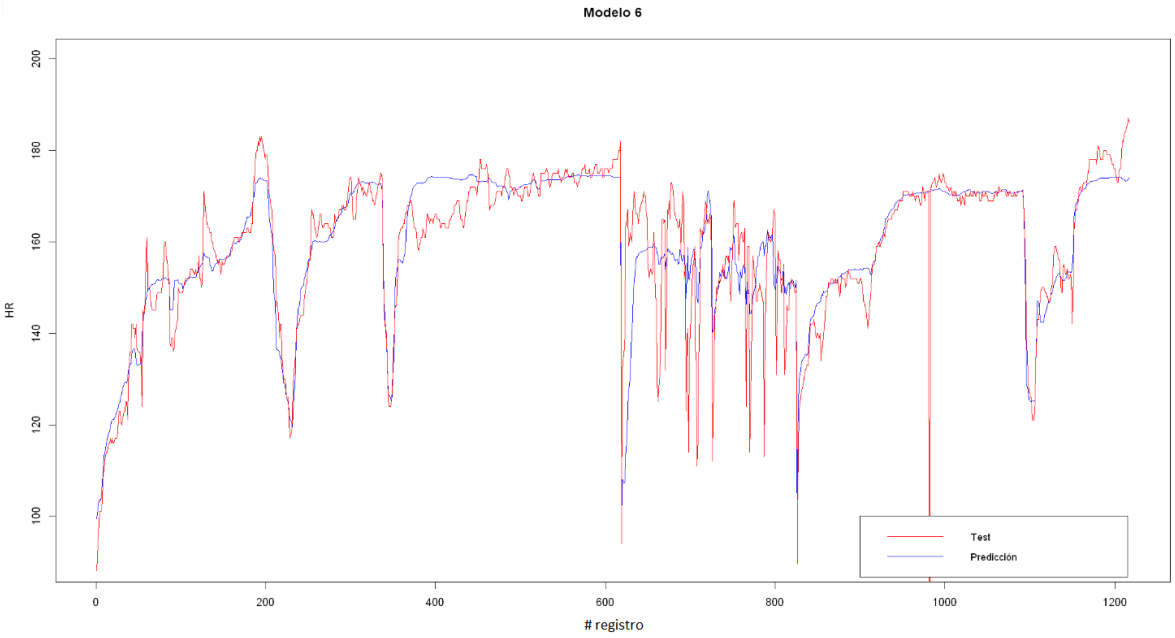

Figura 6.12. Resultado de la prueba 1 en el modelo 5c_4p_14v realizada sobre una parte del conjunto con el que se ha entrenado, donde se representa la frecuencia cardiaca predicha (línea azul) frente a la muestra de test (línea roja)

- **Prueba 2**

La segunda prueba que se ha llevado a cabo ha sido en la que se prueban los modelos en conjuntos de datos diferentes a los utilizados en el entrenamiento, pero que pertenecen a la misma persona.

Los dos primeros modelos se han probado en un total de tres conjuntos de datos, mientras que el resto se han probado solo en uno. Aquí se muestra el resultado de la predicción sobre el conjunto de datos que se ha probado en todos los modelos, los resultados de los dos conjuntos adicionales probados en los modelos 1c_1p_22v y 1c_1p_14v se recogen en el *Anexo 8.*2 y podría decirse que no se han obtenido unas predicciones tan buenas como en la *Prueba 1*, pero prestando atención se puede ver como la curva que forma la frecuencia cardiaca tanto en la predicción, como en el test, es muy similar y que lo que realmente fastidia la predicción es que aparece desplazada, lo que nos hace pensar que falta alguna variable relevante o nos falla algo en la generalización y, por esta razón, se han realizado las pruebas con modelos entrenados con más carreras.

Además, se ha representado, con los resultados del primer conjunto probado en el modelo 1c_1p_22v, un gráfico de correlaciones como en la prueba anterior, que se puede encontrar en el *Anexo 8.3*, y se puede ver que las variables en las que se ve una clara relación respecto a la frecuencia cardiaca coinciden con las seleccionadas en la prueba anterior, teniendo en cuenta que la predicción es peor en este segundo caso.

Predicción vs Test	1c_1p_22v	1c_1p_14v	5c_1p_22v	5c_1p_14v	5c_4p_22v	5c_4p_14v
MAE	9.700	16.041	18.554	19.081	13.333	16.742
Correlación Spearman	0.785	0.682	0.813	0.611	0.778	0.636
Ratio varianzas	0.817	0.903	0.708	0.680	1.481	2.434

Tabla 6.3. Resultados de la segunda prueba de los seis modelos donde se prueban en otro conjunto diferente al que se ha utilizado para entrenar, pero de la misma persona. Corresponden al tercer conjunto que se utiliza en esta prueba para los modelos 1c_1p_22v y 1c_1p_14v y al primer conjunto en el resto de los modelos

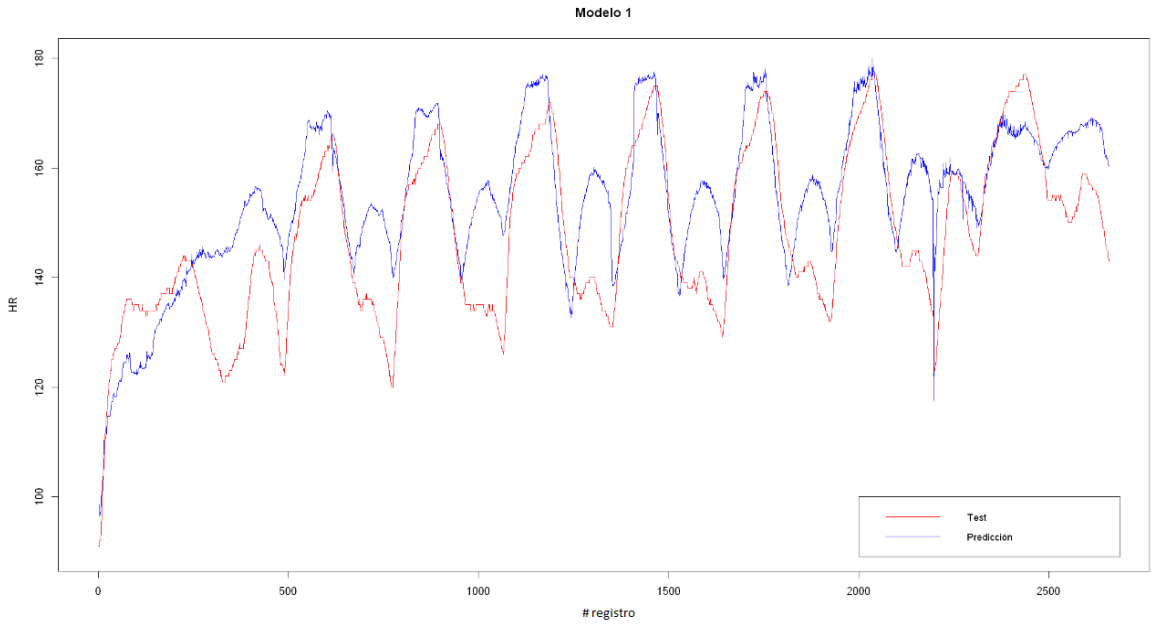

Figura 6.13. Resultado de la prueba 2 del modelo 1c_1p_22v realizada sobre una carrera diferente a la utilizada para entrenar, pero de la misma persona, donde se representa la frecuencia cardiaca predicha (línea azul) frente a la muestra de test (línea roja). Corresponde al tercer conjunto utilizado en esta prueba por el modelo indicado

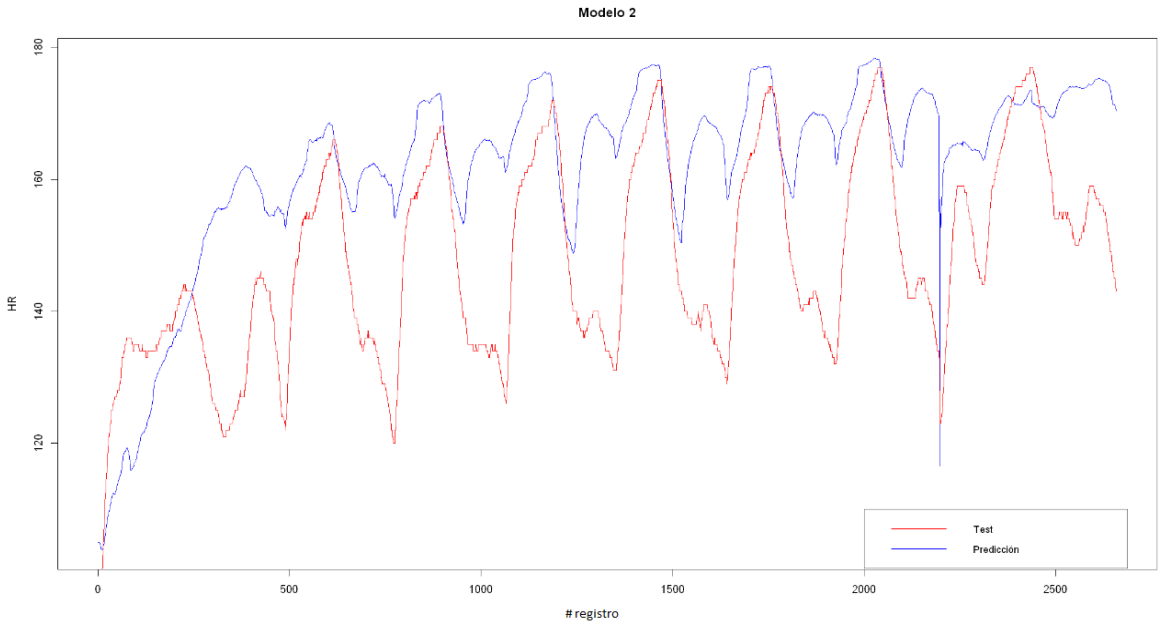

Figura 6.14. Resultado de la prueba 2 del modelo 1c_1p_14v realizada sobre sobre una carrera diferente a la utilizada para entrenar, pero de la misma persona, donde se representa la frecuencia cardiaca predicha (línea azul) frente a la muestra de test (línea roja). Corresponde al tercer conjunto utilizado en esta prueba por el modelo indicado

Por otro lado, se quiere comentar el caso del conjunto de datos que se ha probado en los seis modelos. Los resultados de los modelos que se muestran en las *Figuras 6.13, 6.14, 6.15, 6.16, 6.17 y 6.18*, parecen indicar que se trata de un entrenamiento por series, debido a la periodicidad del patrón de subida y bajada que forma.

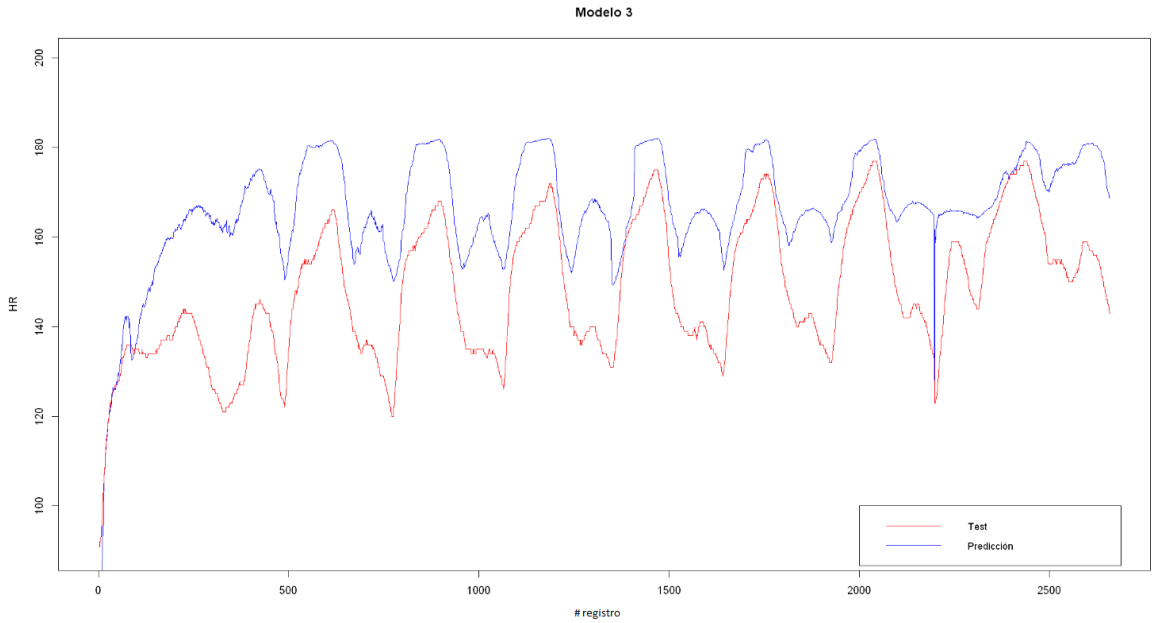

Figura 6.15. Resultado de la prueba 2 del modelo 5c_1p_22v realizada sobre una carrera diferente a la utilizada para entrenar, pero de la misma persona, donde se representa la frecuencia cardiaca predicha (línea azul) frente a la muestra de test (línea roja)

Aunque, por lo general, la predicción no ha capturado correctamente la intensidad de las subidas menos intensas durante el entrenamiento, se han obtenido unos resultados mejores de los esperados sobre todo con el modelo 1c_1p_22v, como puede verse en la *Tabla 6.3*. Cabe destacar que los modelos han sido capaces de interpretar ese patrón de subidas y bajadas en todos los casos, habiendo sido entrenados con carreras continuas y que

probablemente el error que se comete en la predicción se podría corregir utilizando un rango más variado de carreras durante el entrenamiento.

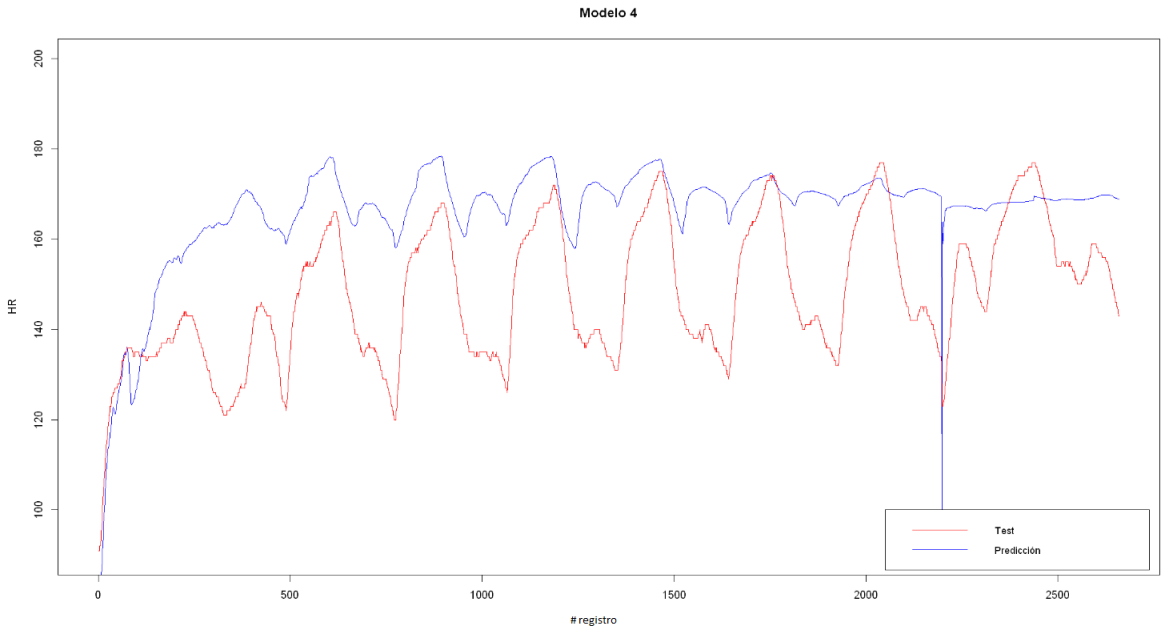

Figura 6.16. Resultado de la prueba 2 del modelo 5c_1p_14v realizada sobre una carrera diferente a la utilizada para entrenar, pero de la misma persona, donde se representa la frecuencia cardiaca predicha (línea azul) frente a la muestra de test (línea roja)

Por otro lado, en el caso del resultado de la prueba de los modelos 1c_1p_14v, 5c_1p_14v y 5c_4p_14v (*Figuras 6.14, 6.16 y 6.18*) se nota claramente que obtienen unas predicciones peores y, como se tratan de los modelos que han sido entrenado con menos variables predictoras, se considera que las variables que no se han utilizado aportan características necesarias.

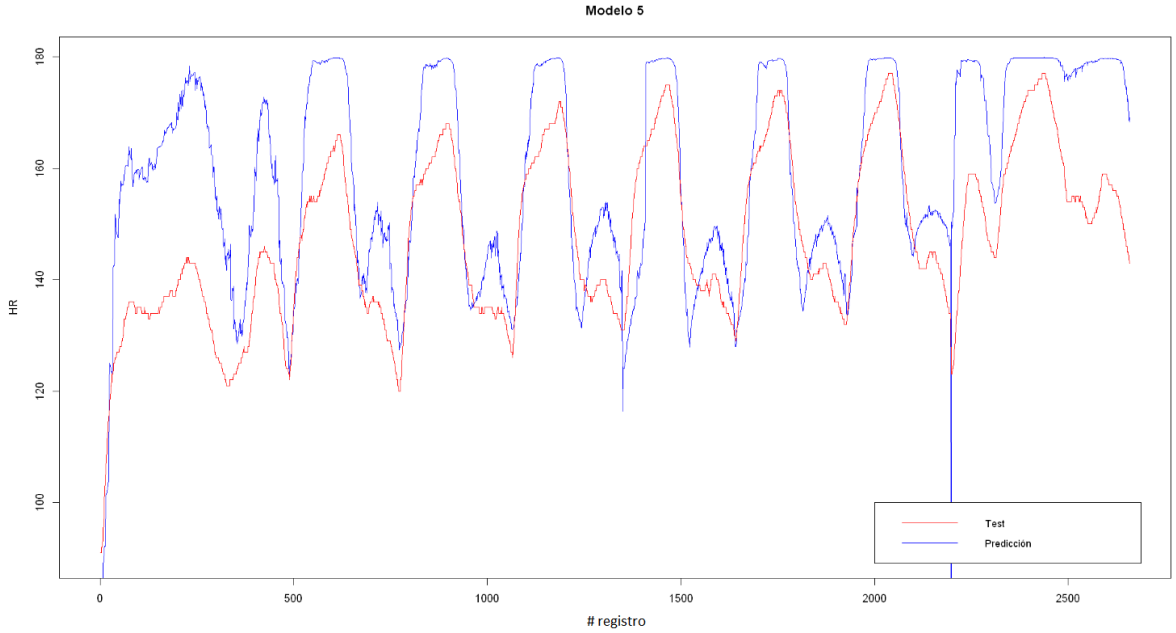

Figura 6.17. Resultado de la prueba 2 del modelo 5c_4p_22v realizada sobre una carrera diferente a la utilizada para entrenar, pero de la misma persona, donde se representa la frecuencia cardiaca predicha (línea azul) frente a la muestra de test (línea roja)

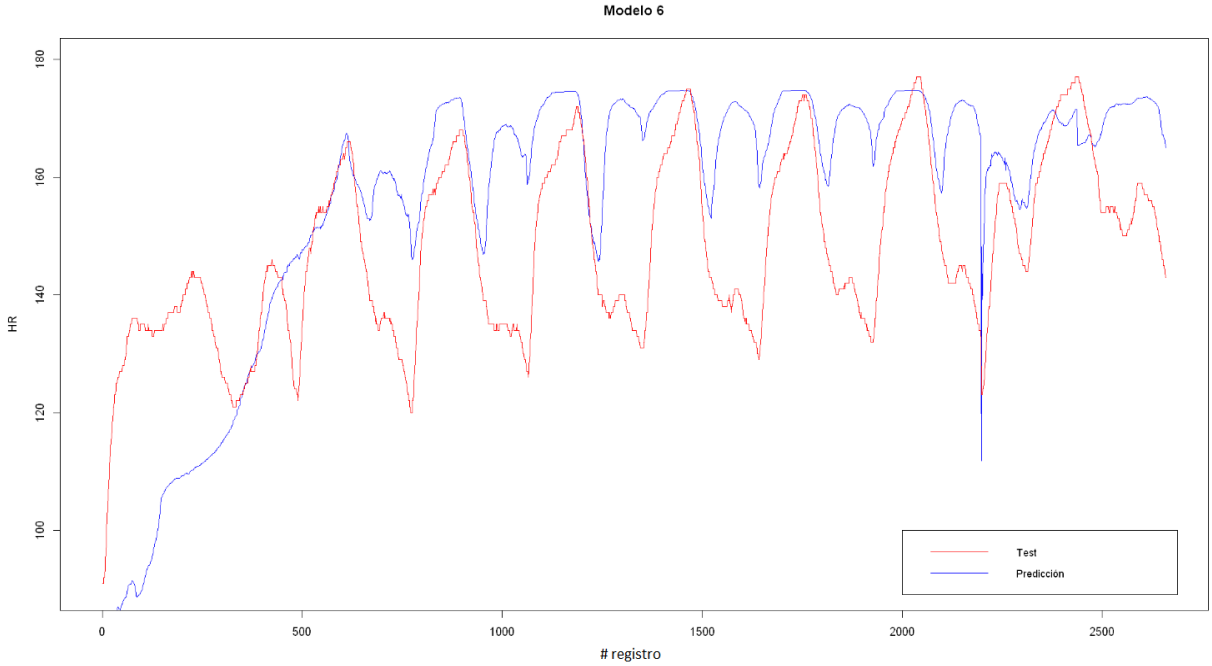

Figura 6.18. Resultado de la prueba 2 del modelo 5c_4p_14v realizada sobre una carrera diferente a la utilizada para entrenar, pero de la misma persona, donde se representa la frecuencia cardiaca predicha (línea azul) frente a la muestra de test (línea roja)

- **Prueba 3**

Por último, se ha realizado el último nivel de prueba en el que los seis modelos se han probado sobre conjuntos de datos de personas diferentes a los utilizados en las pruebas anteriores y en el entrenamiento. Los modelos 1c_1p_22v, 1c_1p_14v, 5c_1p_22v y 4c_1p_14v se han probado en dos conjuntos diferentes. Mientras que los modelos 5c_4p_22v y 4c_4p_14v se han probado en un solo conjunto. Los dos últimos modelos se han probado en conjuntos distintos al resto, ya que los conjuntos utilizados en los otros modelos se han usado para el entreno del quinto y sexto.

Los resultados que se han obtenido de probar los cuatro primeros modelos sobre el primero de sus conjuntos no han sido muy satisfactorios en ninguno de los casos. Aún así, se pueden encontrar las gráficas y los resultados de correlación y varianza en el *Anexo 8.4.*

Predicción vs Test	1c_1p_22v	1c_1p_14v	5c_1p_22v	5c_1p_14v	5c_4p_22v	5c_4p_14v
MAE	3.884	4.899	9.694	9.754	13.232	16.001
Correlación Spearman	0.924	0.914	0.880	0.835	0.777	0.550
Ratio varianzas	0.973	0.896	0.633	1.527	3.695	6.029

Tabla 6.4. Resultados de la tercera prueba de los seis modelos donde se prueban en otro conjunto diferente al que se ha utilizado para entrenar y que pertenece a otra persona distinta. Corresponden al segundo conjunto que se utiliza en esta prueba para los cuatro primeros modelos, y al primer conjunto en los modelos 5c_4p_22v y 5c_4p_14v

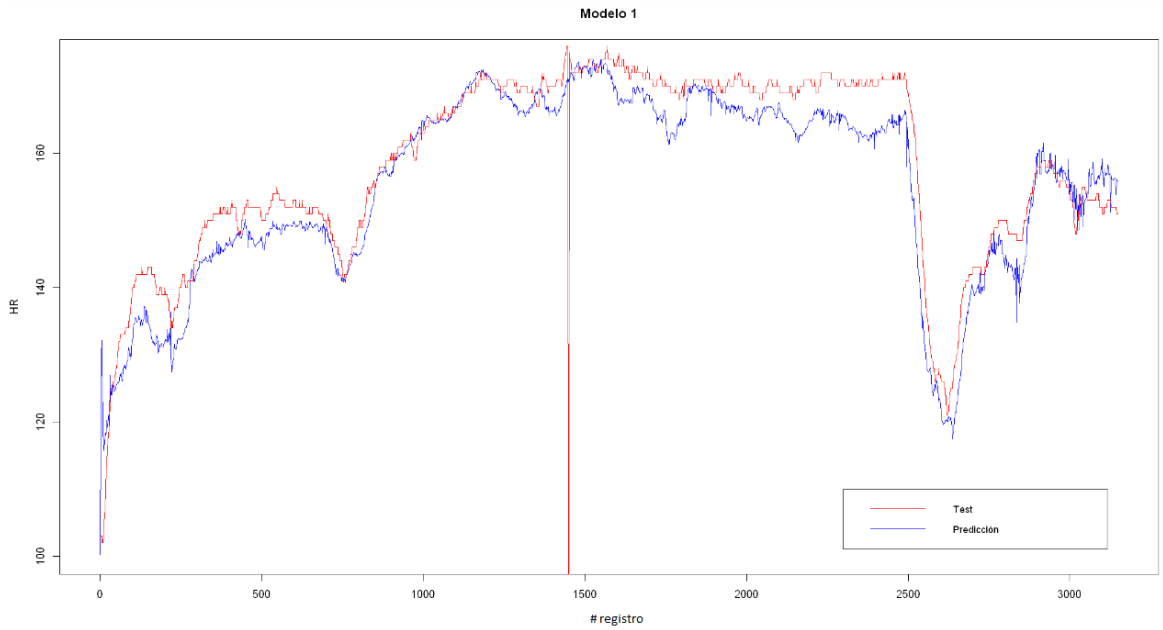

Figura 6.19. Resultado de la prueba 3 del modelo 1c_1p_22v realizada sobre una carrera diferente a la utilizada para entrenar y de una persona diferente, donde se representa la frecuencia cardiaca predicha (línea azul) frente a la muestra de test (línea roja). Corresponde al segundo conjunto utilizado en esta prueba por el modelo indicado

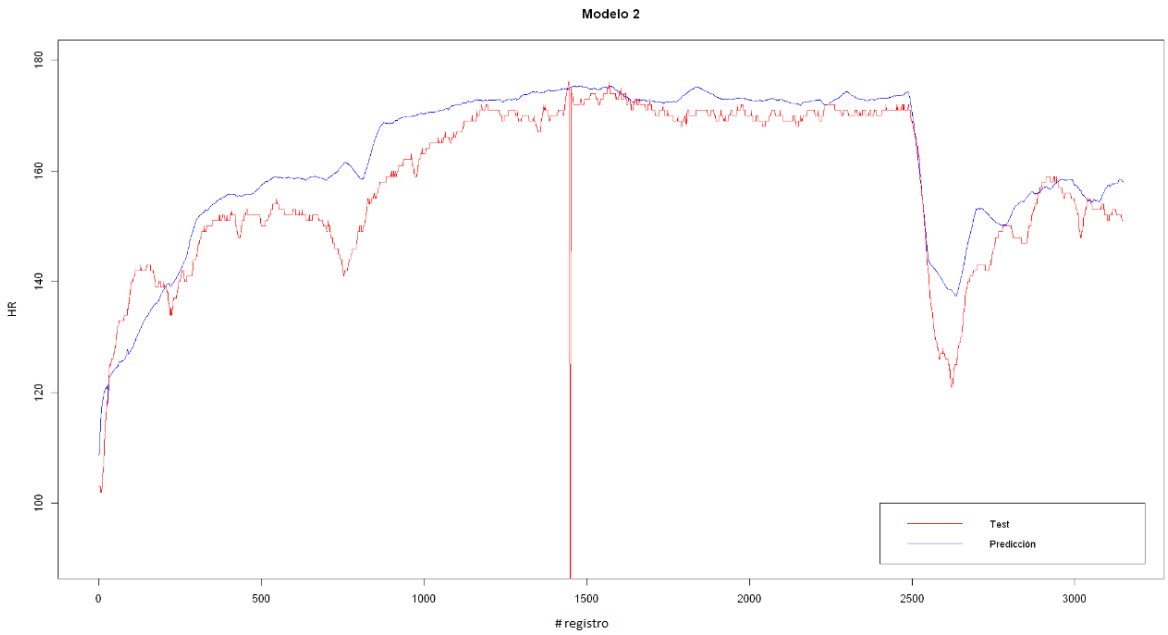

Figura 6.20. Resultado de la prueba 3 del modelo 1c_1p_14v realizada sobre una carrera diferente a la utilizada para entrenar y de una persona diferente, donde se representa la frecuencia cardiaca predicha (línea azul) frente a la muestra de test (línea roja). Corresponde al segundo conjunto utilizado en esta prueba por el modelo indicado

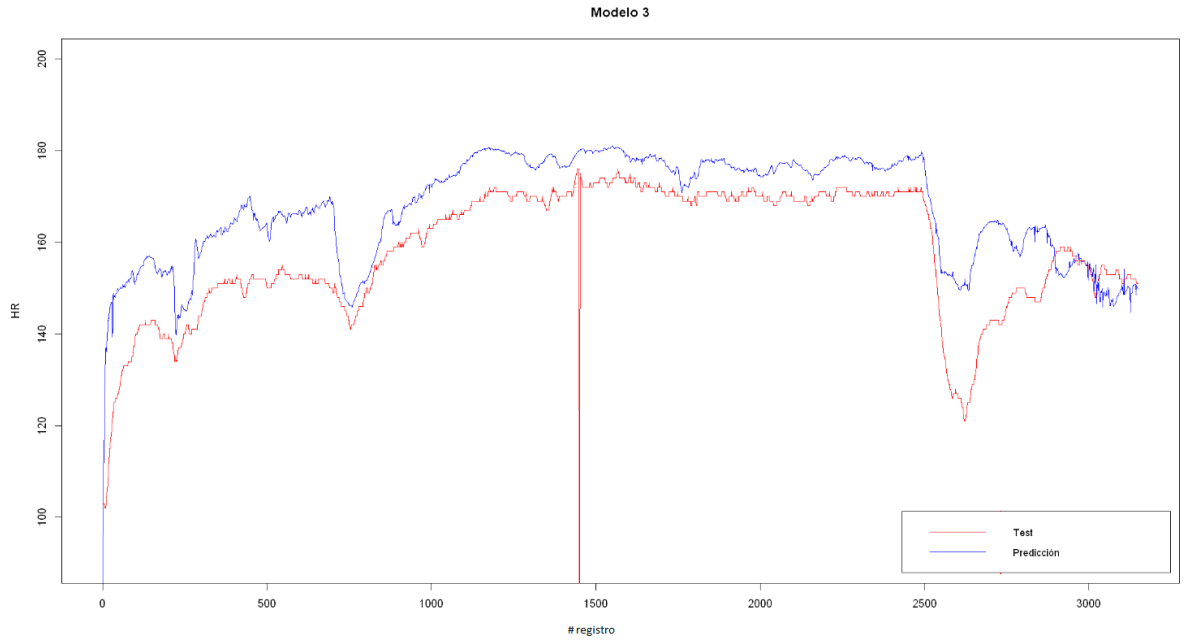

Figura 6.21. Resultado de la prueba 3 del modelo 5c_1p_22v realizada sobre una carrera diferente a la utilizada para entrenar y de una persona diferente, donde se representa la frecuencia cardiaca predicha (línea azul) frente a la muestra de test (línea roja). Corresponde al segundo conjunto utilizado en esta prueba por el modelo indicado

Por otro lado, los resultados conseguidos (*Tabla 6.4*) con los otros conjuntos de datos han sido, en general, sorprendentemente buenos, sobre todo en los casos de los dos primeros modelos, cuyos resultados se presentan en las *Figuras 6.19 y 6.20.* Y, aunque en el modelo 5c_4p_22v se haya obtenido demasiado error y un ratio de varianzas muy elevado, se ha capturado bastante bien la curva de la frecuencia cardiaca, incidiendo en la muy buena predicción de las oscilaciones que se pueden ver en la *Figura 6.23.* Por el contrario, la predicción obtenida con el modelo 5c_4p_14v (*Figura 6.24*) ha sido muy mala, como ya indicaban los datos expuestos en la *Tabla 6.4*.

En esta prueba, se esperaban mejores resultados para los modelos 5c_1p_22v y 5c_1p_14v que los obtenidos (*Figuras 6.21 y 6.22*), ya que al haber sido entrenados con una mayor cantidad de datos se pensaba que tendrían una mejor generalización. Aun así, se puede ver que, aunque no han capturado los valores de la frecuencia cardiaca de forma precisa, han aprendido la tendencia de la frecuencia cardiaca. Esto podría solucionarse entrenando con carreras de más personas, por esta razón se han probado los modelos 5c_4p_22v y 5c_4p_14v (*Figuras 6.23 y 6.24*), pero no se han obtenido una gran mejora en los resultados. Este último resultado, ha llevado a la conclusión de que de una persona a otra hay una diferencia de condición física que es muy relevante, no puedes utilizar la condición física de un campeón olímpico para predecir la de un corredor popular y viceversa.

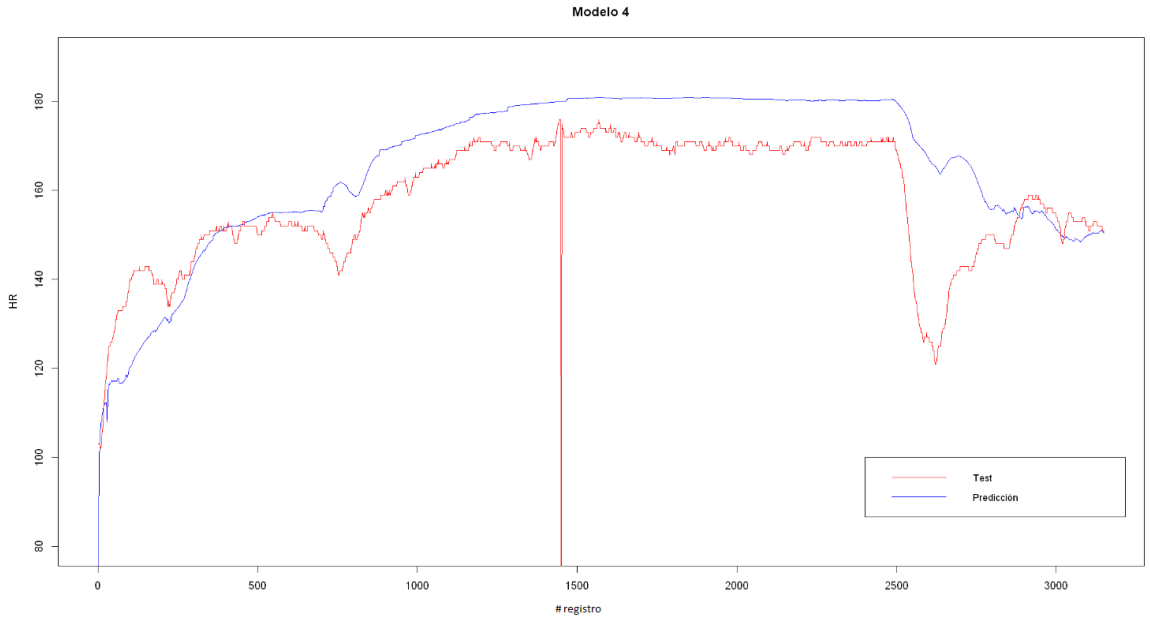

Figura 6.22. Resultado de la prueba 3 del modelo 5c_1p_14v realizada sobre una carrera diferente a la utilizada para entrenar y de una persona diferente, donde se representa la frecuencia cardiaca predicha (línea azul) frente a la muestra de test (línea roja). Corresponde al segundo conjunto utilizado en esta prueba por el modelo indicado

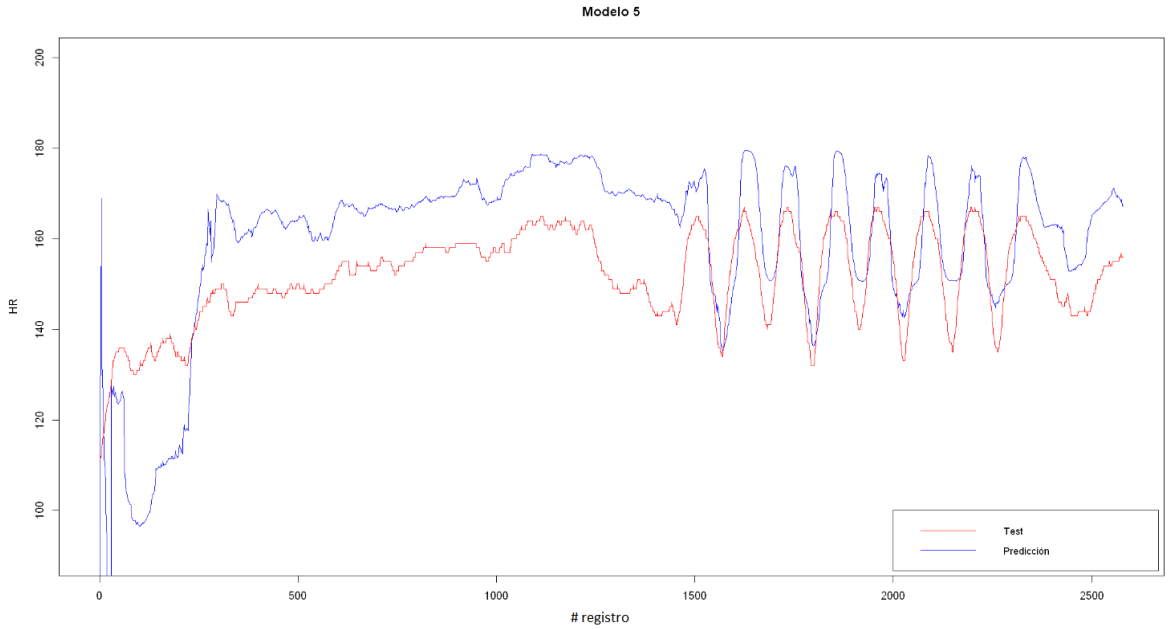

Figura 6.23. Resultado de la prueba 3 del modelo 5c_4p_22v realizada sobre una carrera diferente a la utilizada para entrenar y de una persona diferente, donde se representa la frecuencia cardiaca predicha (línea azul) frente a la muestra de test (línea roja)

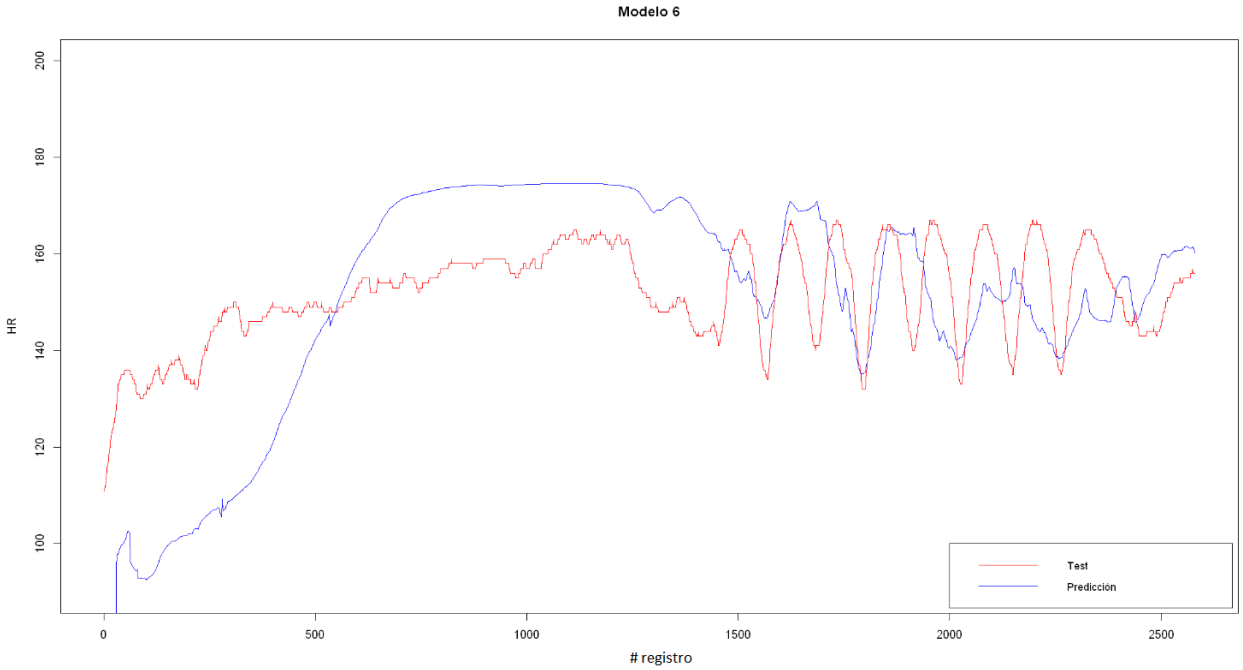

Figura 6.24. Resultado de la prueba 3 del modelo 5c_4p_14v realizada sobre una carrera diferente a la utilizada para entrenar y de una persona diferente, donde se representa la frecuencia cardiaca predicha (línea azul) frente a la muestra de test (línea roja)

En general, se han obtenido unos resultados muy buenos y en los que se puede ver que en ocasiones al hacer una selección más exigente de las variables predictoras se pierde precisión en los valores predichos.

Por otro lado, se ha comprobado que entrenando con varias carreras se consigue captar razonablemente bien la tendencia de la frecuencia cardiaca, incluso en carreras tan variadas como las que se han utilizado en este proyecto. Sin embargo, el hecho de utilizar carreras de varias personas no es conveniente porque cada individuo puede tener una condición física muy dispar. A pesar de ello, este tipo de modelos han conseguido describir correctamente la variación de la frecuencia cardiaca, aunque también fallaban en la precisión de los valores.

Capítulo 7
Conclusiones y trabajos futuros

En este último capítulo se va a valorar de forma general el resultado del proyecto y se proponen las líneas de trabajo que se considera que se deberían abordar para continuar su desarrollo.

7.1. Conclusiones

Siendo el objetivo inicial del presente Trabajo de Fin de Máster confirmar la viabilidad de reproducir la frecuencia cardiaca de un individuo en base a unas variables derivadas de otros datos de su actividad física, se puede decir que ha sido satisfecho. Las predicciones obtenidas han superado las expectativas esperadas y se ha podido confirmar la posibilidad de reproducir la frecuencia cardiaca dadas unas variables. También se considera que se han conseguido unos buenos resultados teniendo en cuenta la simplicidad de los modelos y del proceso de curación de datos y se puede asegurar que añadiendo complejidad a la parte de corrección de datos y selección de variables se podrían mejorar notablemente.

Es importante indicar que en este proyecto me he encargado de recoger los ficheros de datos y transformarlos a un solo formato homogéneo y con una misma estructura. Por otro lado, el proyecto se ha enfocado en los datos de *running* y, por ello, he manejado medidas de GPS y registros de frecuencia cardiaca y, después de una primera inspección de los datos, he comprobado la presencia de medidas imprecisas y he desarrollado procedimientos para que éstas no afectasen excesivamente al análisis. Estos métodos los he probado sobre la medida de la distancia total en varios ejemplos con referencias externas y he conseguido reducir la imprecisión desde varios kilómetros a unas pocas decenas de metros. Tras curar los datos, se han propuesto una serie de variables derivadas de las otras variables que habían sido medidas con los dispositivos y que muestran correlación con la frecuencia cardiaca para hacer el análisis. En esta etapa, he diseñado y configurado una serie de redes neuronales que, utilizando las variables calculadas, han sido capaces de predecir la frecuencia cardiaca. Para terminar, he comprobado que aplicadas las redes sobre la misma carrera con la que se entrenan los modelos, la frecuencia cardiaca se reproduce con precisión y que, cuando se aplican los modelos entrenados con una o varias carreras a otras carreras del mismo individuo o a otros, se reproduce de forma parcial, capturando las tendencias, pero con falta de precisión en los valores.

Por último, quiero comentar que el desarrollo del proyecto desde su fase inicial a su final me ha servido para poner en práctica los conocimientos que he adquirido durante el máster sobre estadística, representación de datos, aprendizaje automático e interpretación de los resultados. Además, también me ha servido para darme cuenta de la importancia que tiene la calidad de los datos utilizados y de la realización de un análisis previo sobre las variables que se van a introducir a la red.

7.2. Trabajos futuros

El desarrollo de este proyecto ha abierto muchas posibilidades que no se han podido cubrir dada la extensión en el tiempo de este Trabajo de Fin de Máster. Sin embargo, se describen a continuación algunas ideas con las que considero que se podría continuar:

- **Implementar un algoritmo de corrección de ruta**
  En este caso, la idea es implementar un algoritmo que, teniendo en cuenta los *trackpoints* anteriores y posteriores, consiga corregir la ruta medida. De esta forma se podrán evitar o disminuir los casos en los que una ruta que es en línea recta sea medida como una línea en zigzag. Así se podría lograr tratar con una aproximación más realista de la ruta que realmente siguió el usuario.

  Además, podría plantearse como una extensión de la aplicación STRAVA.

- **Añadir complejidad al sistema de corrección y detección de anomalías**
  El proceso de curación realizado en este proyecto ha sido muy simple y podrían implementarse técnicas de corrección y detección de anomalías más complejas y para más variables.

- **Análisis de variables predictoras más complejo**
  Este último aspecto podría tener gran influencia en trabajos futuros en los que se aplicase la reproducción de la frecuencia cardiaca de distintos individuos, por ello, como creo que falta buscar métodos que generalicen mejor, creo que se debería realizar un análisis y una selección más complejos y dedicados a las variables predictoras, ya que son las que aportarán la información necesaria para predecir la variable objetivo, o buscando alguna forma de entrenar los modelos en la que no se realice un aprendizaje muy específico de un tipo de carrera.

- **Detección de problemas de salud**
  En el caso de que se consiguiese ajustar el predictor correctamente, podrían detectarse problemas de salud cuando se encontrasen diferencias claras entre la frecuencia cardiaca y su predicción en base a los parámetros.

- **Diseño de ritmo de carrera óptimo para una carrera**
  Otra aplicación de este proyecto sería diseñar el ritmo óptimo para una carrera. Teniendo el predictor y unas condiciones como una pendiente determinada, la distancia, etc. Se podría buscar el ritmo que mantuviese la frecuencia cardiaca constante.

- **Categorización de corredores**
  Si como hemos hipotetizado, la diferencia entre la predicción entrenada en un individuo y su ritmo cardiaco medido se debe a su diferente condición física, se podría intentar revertir el proceso para clasificar su estado de forma o nivel de rendimiento.

Capítulo 8
Anexos

8.1. Anexo I

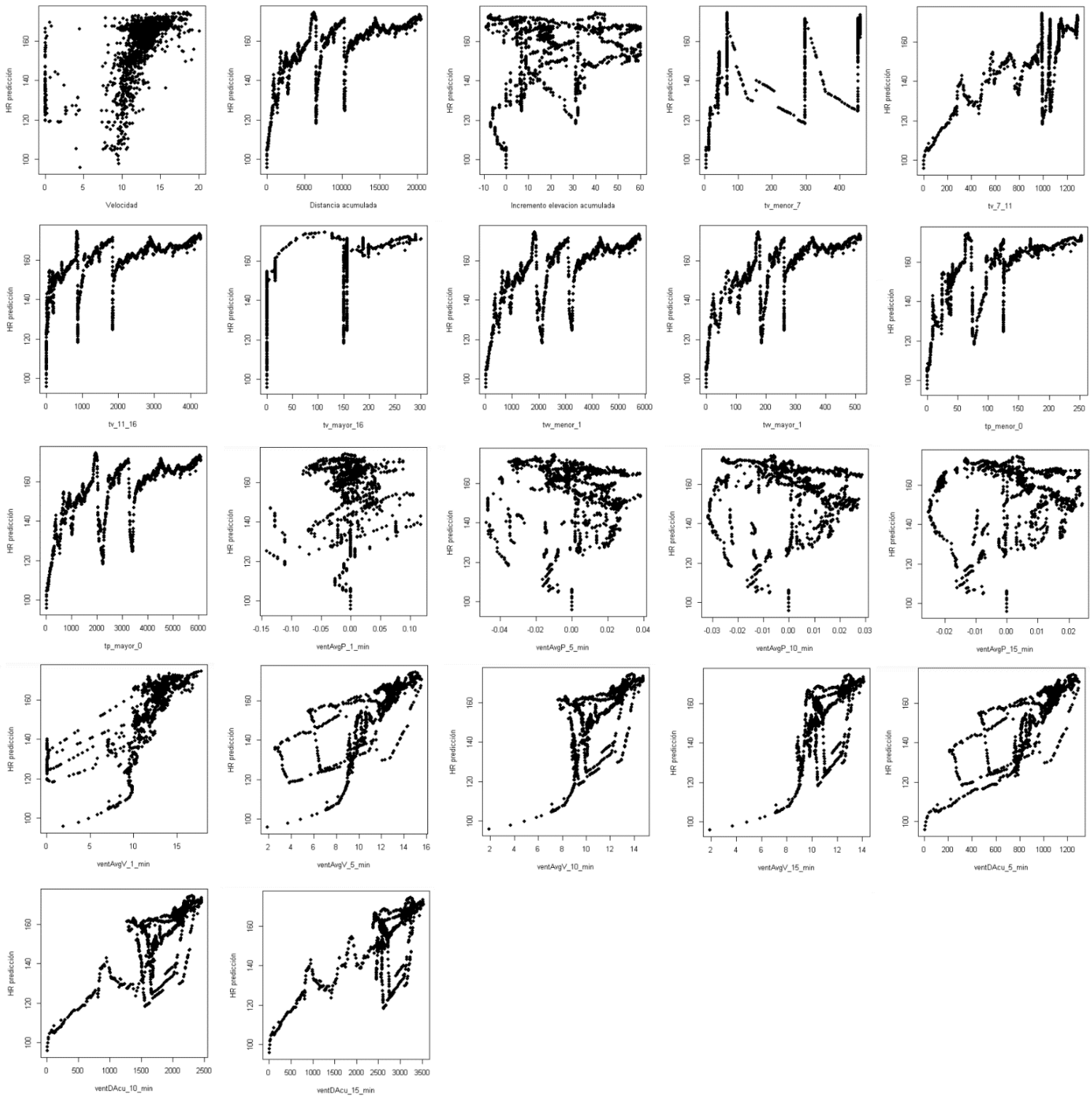

Figura 8.1. Gráfico de correlación entre las variables predictoras y la frecuencia cardiaca predicha en la prueba 1 del modelo 1c_1p_22v

8.2. Anexo II

Predicción vs Test	1c_1p_22v	1c_1p_14v
MAE	20.881	14.348
Correlación Spearman	0.769	0.529
Ratio varianzas	1.021	1.500

Tabla 8.1. Resultados de la prueba 2 de los modelos 1c_1p_22v y 1c_1p_14v, donde se prueban en otro conjunto diferente al que se ha utilizado para entrenar, pero de la misma persona. Corresponden al primer conjunto que se utiliza en esta prueba para los modelos indicados

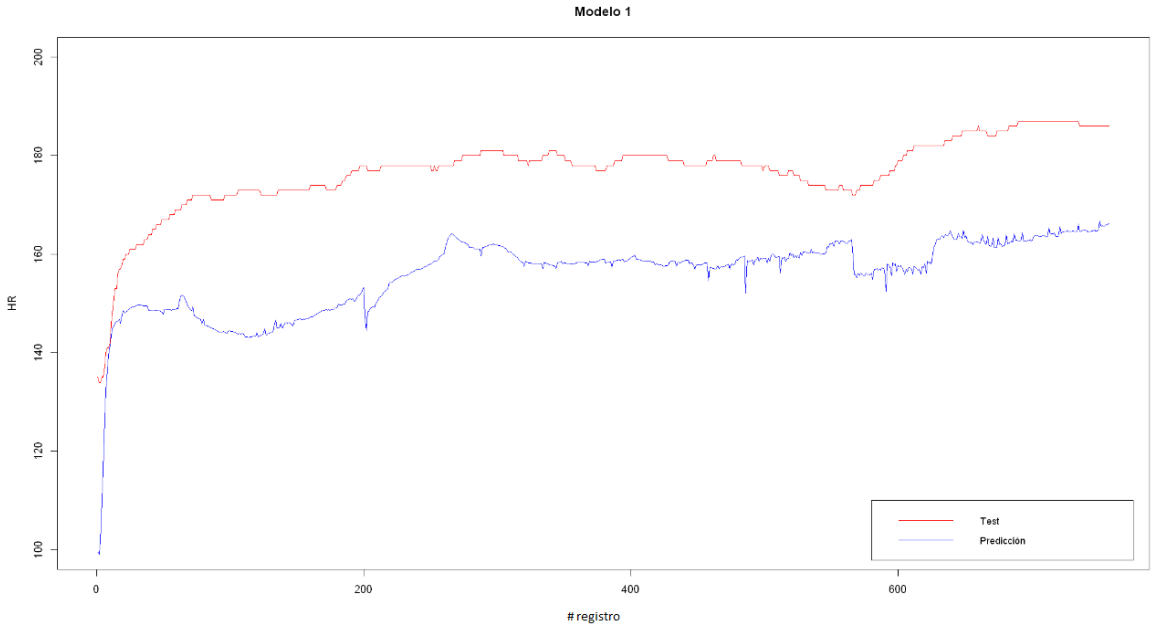

Figura 8.2. Resultado de la prueba 2 del modelo 1c_1p_22v realizada sobre una carrera diferente a la utilizada para entrenar, pero de la misma persona, donde se representa la frecuencia cardiaca predicha (línea azul) frente a la muestra de test (línea roja). Corresponde al primer conjunto de datos utilizado en esta prueba por el modelo indicado

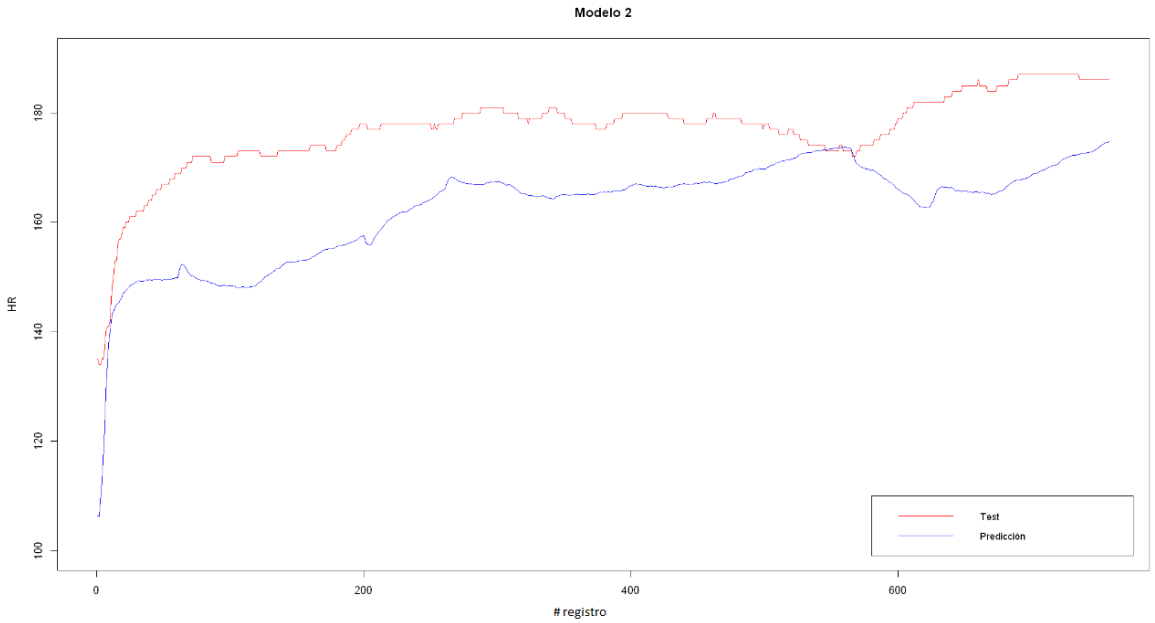

Figura 8.3. Resultado de la prueba 2 del modelo 1c_1p_14v realizada sobre una carrera diferente a la utilizada para entrenar, pero de la misma persona, donde se representa la frecuencia cardiaca predicha (línea azul) frente a la muestra de test (línea roja). Corresponde al primer conjunto de datos utilizado en esta prueba por el modelo indicado

Predicción vs Test	1c_1p_22v	1c_1p_14v
MAE	9.378	4.342
Correlación Spearman	0.920	0.881
Ratio varianzas	0.992	1.203

Tabla 8.2. Resultados de la prueba 2 de los modelos 1c_1p_22v y 1c_1p_14v, donde se prueban en otro conjunto diferente al que se ha utilizado para entrenar, pero de la misma persona. Corresponden al segundo conjunto que se utiliza en esta prueba para los modelos indicados

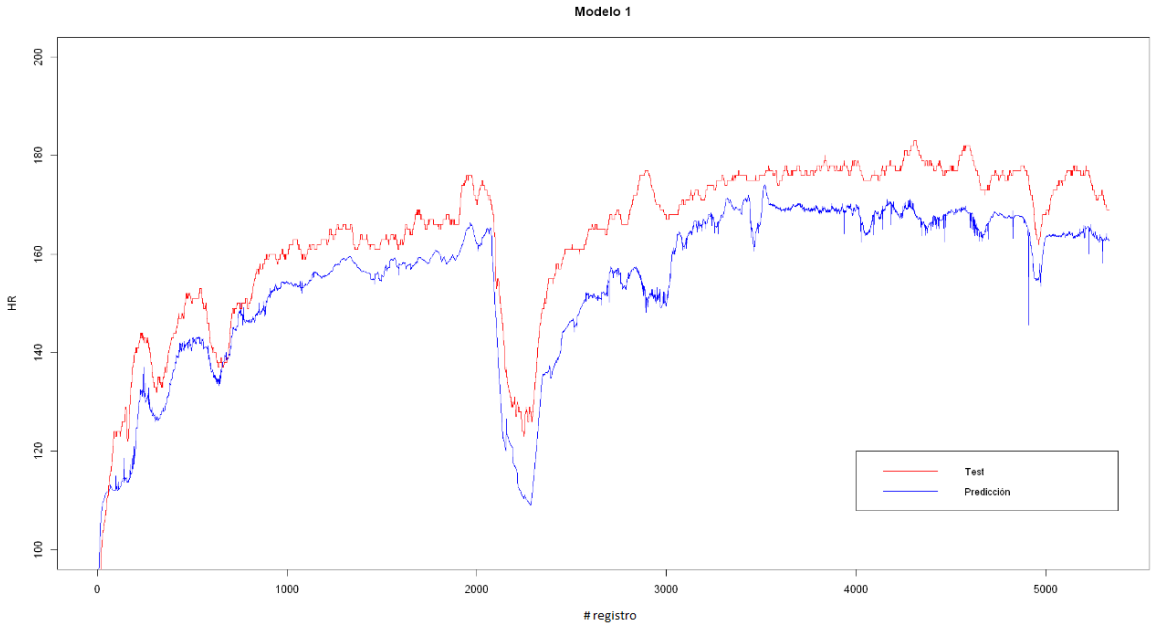

Figura 8.4. Resultado de la prueba 2 del modelo 1c_1p_22v realizada sobre una carrera diferente a la utilizada para entrenar, pero de la misma persona, donde se representa la frecuencia cardiaca predicha (línea azul) frente a la muestra de test (línea roja). Corresponde al segundo conjunto de datos utilizado en esta prueba por el modelo indicado

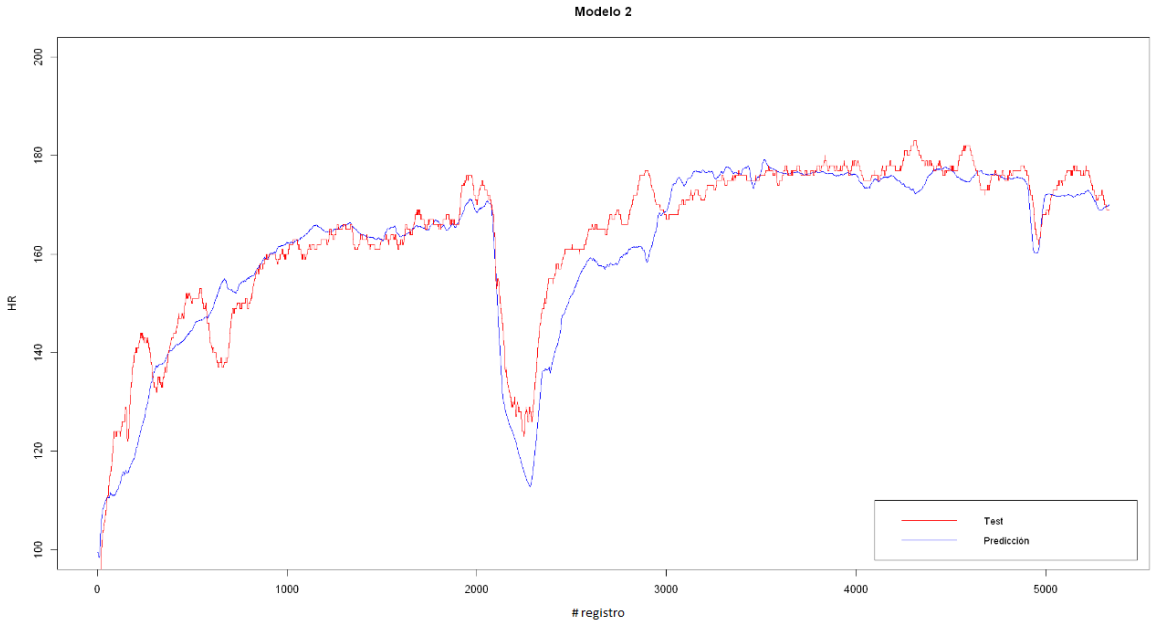

Figura 8.5. Resultado de la prueba 2 del modelo 1c_1p_14c realizada sobre una carrera diferente a la utilizada para entrenar, pero de la misma persona, donde se representa la frecuencia cardiaca predicha (línea azul) frente a la muestra de test (línea roja). Corresponde al segundo conjunto de datos utilizado en esta prueba por el modelo indicado

8.3. Anexo III

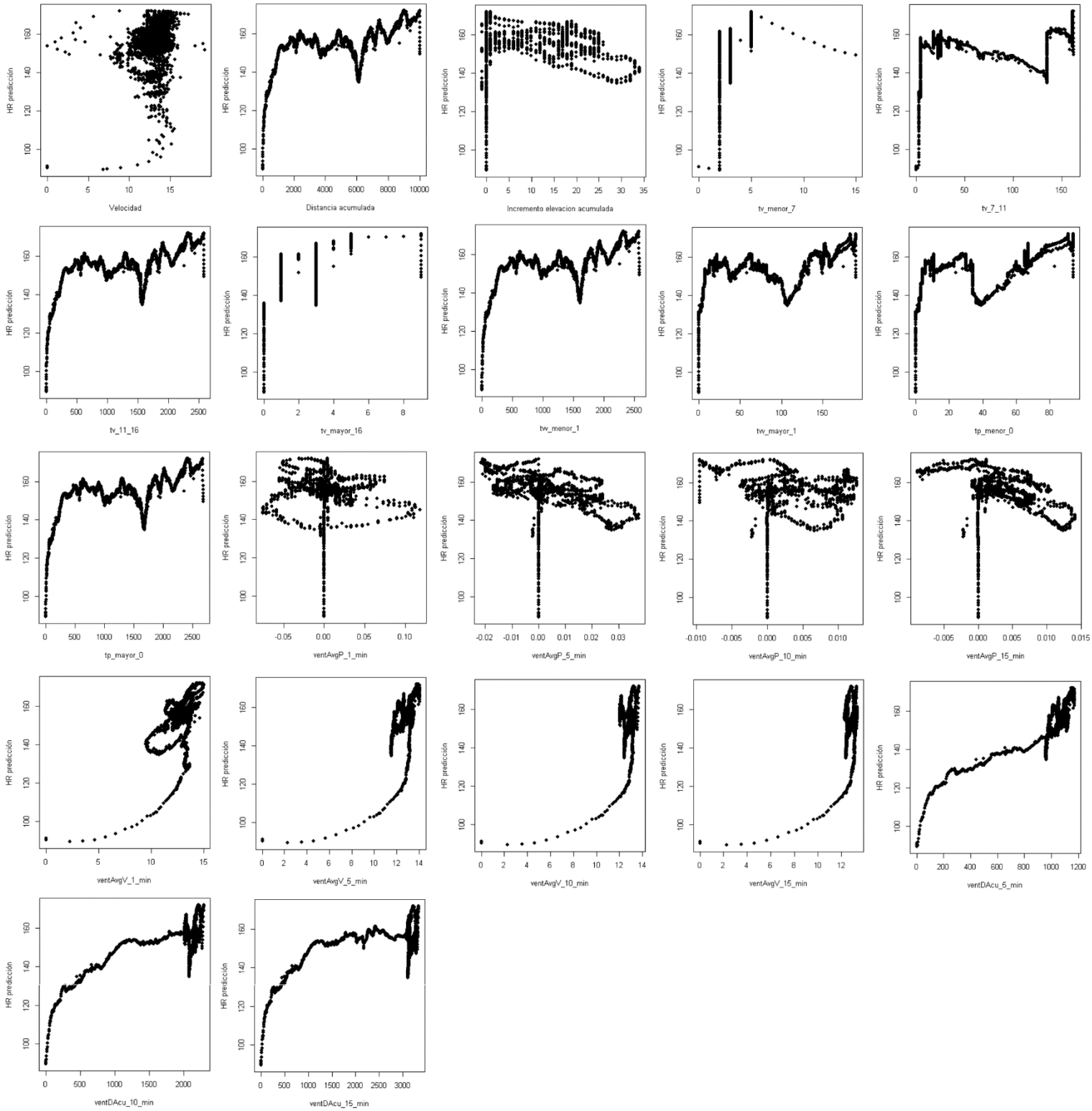

Figura 8.6. Gráfico de correlación entre las variables predictoras y la frecuencia cardiaca predicha en la prueba 2 del modelo 1c_1p_22v sobre el primer conjunto de datos

8.4. Anexo IV

Predicción VS Test	1c_1p_22v	1c_1p_14v	5c_1p_22v	5c_1p_14v
MAE	11.387	14.930	14.057	15.376
Correlación Spearman	0.251	0.009	0.419	-0.088
Ratio varianzas	0.696	0.747	0.310	0.334

Tabla 8.3. Resultados de la prueba 3 de los modelos 1c_1p_22v, 1c_1p_14v, 5c_1p_22v y 5c_1p_14v, donde se prueban en otro conjunto diferente al que se ha utilizado para entrenar y de una persona diferente. Corresponde al primer conjunto de datos que se utiliza en esta prueba para los modelos indicados

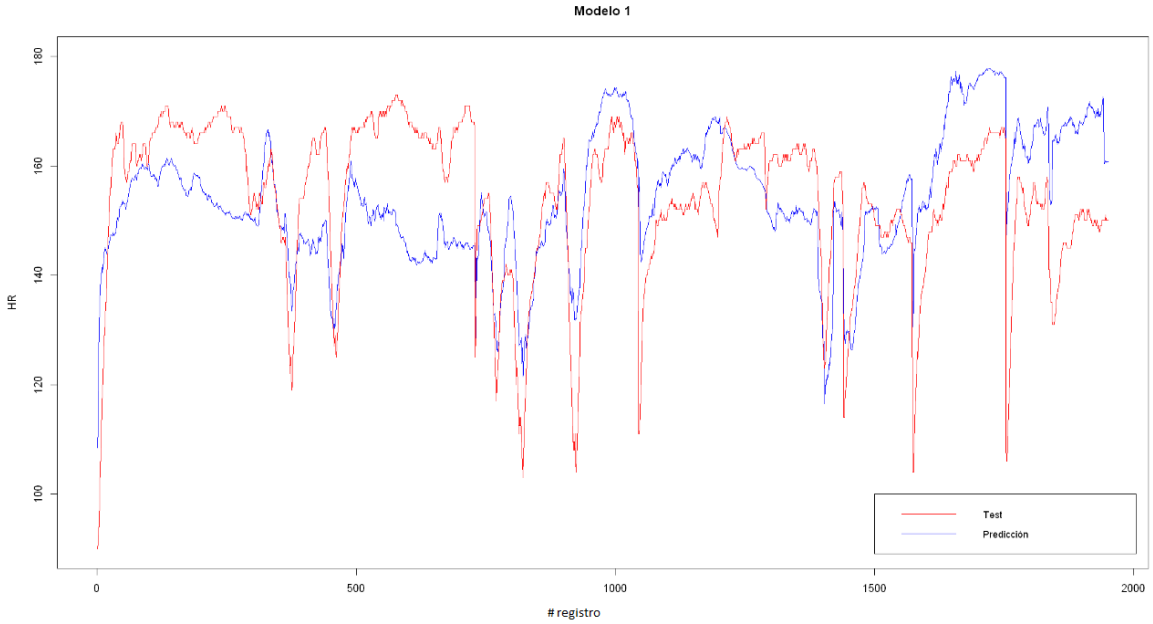

Figura 8.7. Resultado de la prueba 3 del modelo 1c_1p_22v realizada sobre una carrera diferente a la utilizada para entrenar y de una persona diferente, donde se representa la frecuencia cardiaca predicha (línea azul) frente a la muestra de test (línea roja). Corresponde al primer conjunto de datos utilizado en esta prueba para el modelo indicado

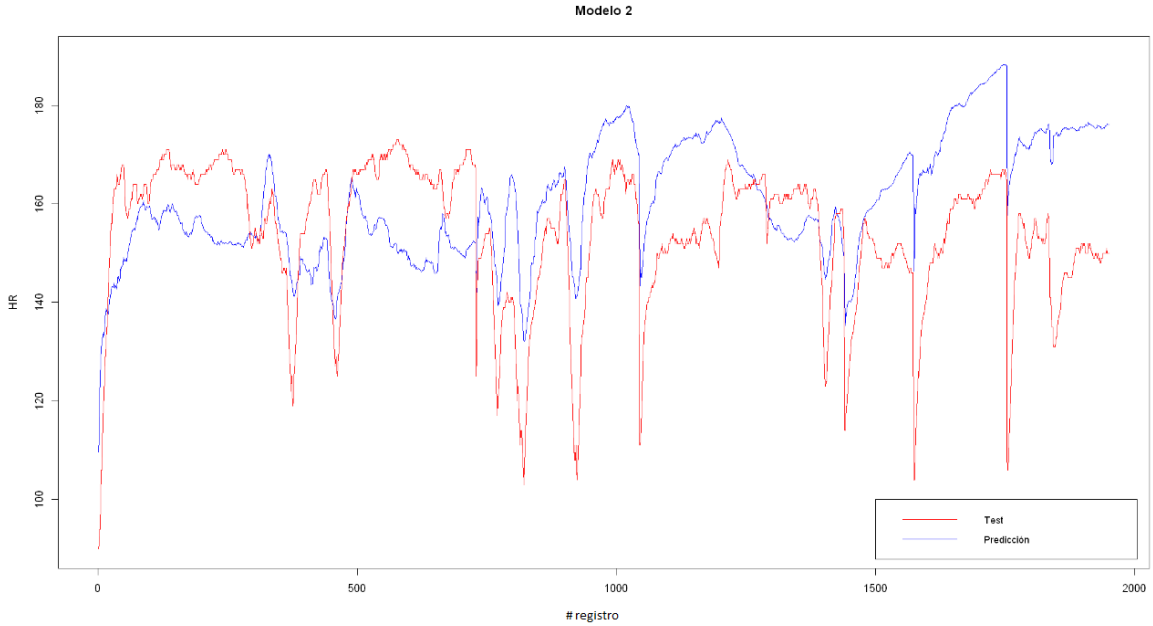

Figura 8.8. Resultado de la prueba 3 del modelo 1c_1p_14v realizada sobre una carrera diferente a la utilizada para entrenar y de una persona diferente, donde se representa la frecuencia cardiaca predicha (línea azul) frente a la muestra de test (línea roja). Corresponde al primer conjunto de datos utilizado en esta prueba para el modelo indicado

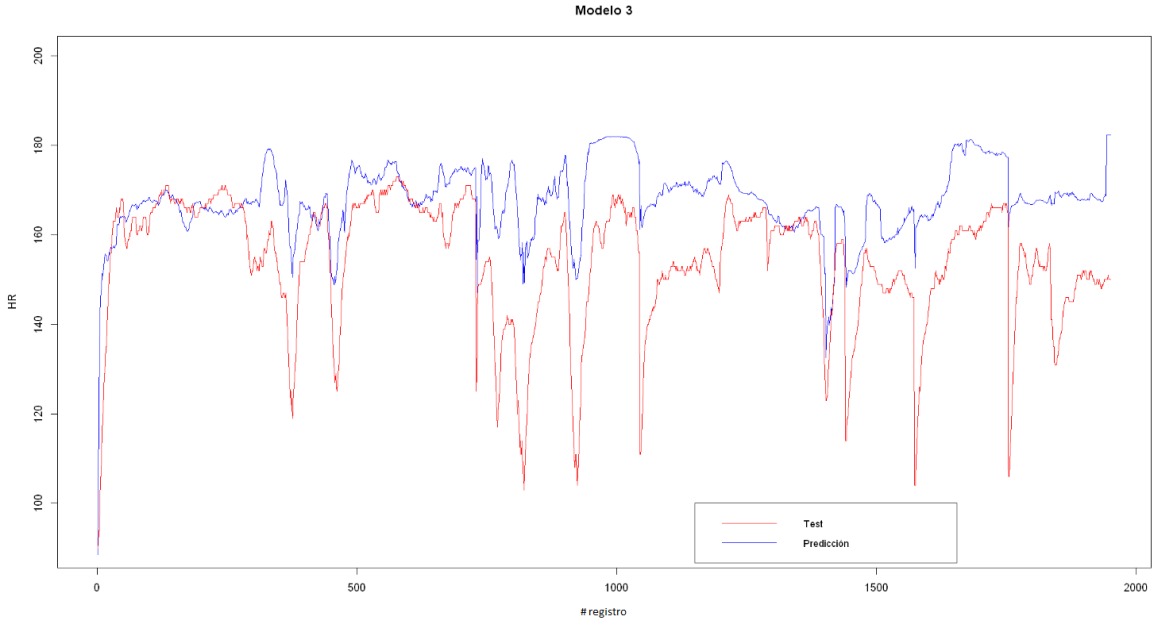

Figura 8.9. Resultado de la prueba 3 del modelo 5c_1p_22v realizada sobre una carrera diferente a la utilizada para entrenar y de una persona diferente, donde se representa la frecuencia cardiaca predicha (línea azul) frente a la muestra de test (línea roja). Corresponde al primer conjunto de datos utilizado en esta prueba para el modelo indicado

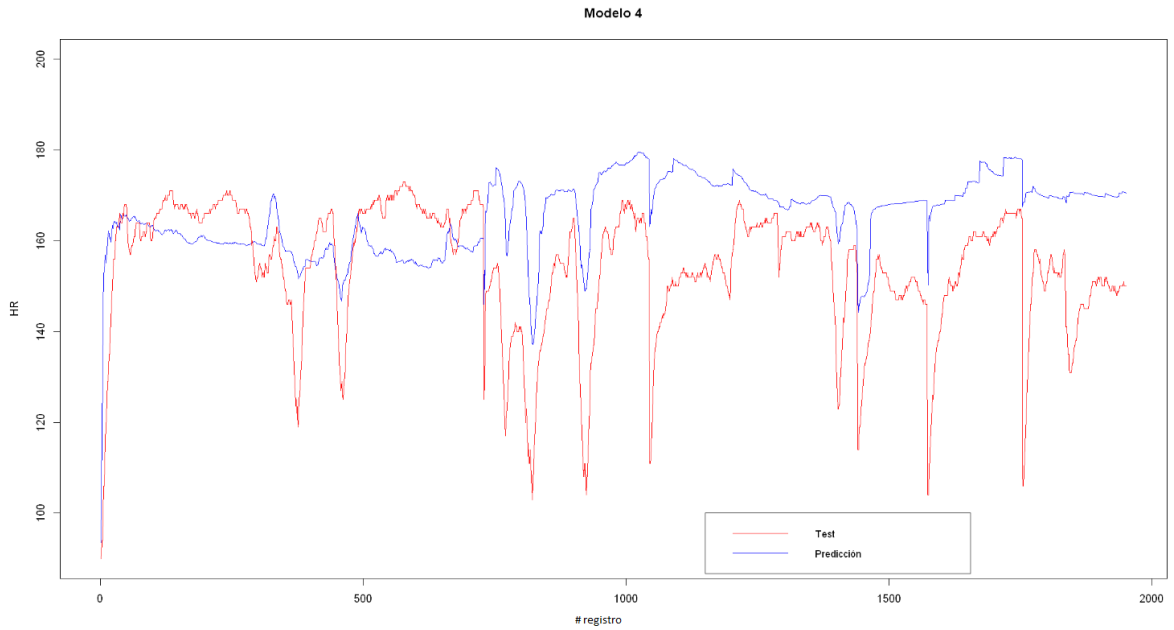

Figura 8.10. Resultado de la prueba 3 del modelo 5c_1p_14v realizada sobre una carrera diferente a la utilizada para entrenar y de una persona diferente, donde se representa la frecuencia cardiaca predicha (línea azul) frente a la muestra de test (línea roja). Corresponde al primer conjunto de datos utilizado en esta prueba para el modelo indicado

Bibliografía

TrackeR y fitdc referencias serias en R con citation("paquete")

1. https://www.redalyc.org/html/2351/235131674015/
2. https://www.larazon.es/historico/7236-las-nuevas-tecnologias-mejoran-el-rendimiento-deportivo-OLLA_RAZON_318590
3. https://www.topografix.com/GPX/1/1/
4. https://www.topografix.com/gpx.asp
5. https://en.wikipedia.org/wiki/Training_Center_XML
6. https://www8.garmin.com/xmlschemas/TrainingCenterDatabasev2.xsd
7. https://fileinfo.com/extension/tcx
8. [pdf fit file types description]
9. https://tools.ietf.org/html/rfc4180
10. https://es.wikipedia.org/wiki/Valores_separados_por_comas
11. https://www.strava.com/about
12. https://www.goldencheetah.org/
13. https://github.com/GoldenCheetah/GoldenCheetah
14. https://www.python.org/
15. https://www.numpy.org/
16. https://pandas.pydata.org/
17. https://docs.python.org/2/library/xml.etree.elementtree.html
18. https://matplotlib.org/
19. https://www.atlassian.com/git
20. https://trello.com/
21. https://jupyter.org/
22. https://www.r-project.org/
23. https://keras.io/getting-started/functional-api-guide/
24. https://www.tensorflow.org/
25. https://www.kdnuggets.com/2018/05/poll-tools-analytics-data-science-machine-learning-results.html
26. https://github.com/tkrajina/gpxpy
27. https://lxml.de/
28. https://docs.python.org/2/library/xml.etree.elementtree.html
29. https://pypi.org/project/python-tcxparser/
30. http://dtcooper.github.io/python-fitparse/
31. https://www.rdocumentation.org/packages/XML/versions/3.98-1.20
32. http://plotkml.r-forge.r-project.org/
33. https://github.com/trackerproject/trackeR
34. https://cran.r-project.org/web/packages/fitdc/fitdc.pdf
35. https://web.archive.org/web/20041108132234/http://www.census.gov/cgi-bin/geo/gisfaq?Q5.1
36. https://fellrnr.com/wiki/GPS_Accuracy

37. https://www.bbc.com/mundo/noticias-45660899
38. https://keras.io/activations/
39. https://www.geosci-model-dev.net/7/1247/2014/gmd-7-1247-2014.pdf