

## Descripción de la práctica

Esta práctica ha sido realizada individualmente por Javier Gómez de Diego.

- GitHub: <https://github.com/javi22gd/Practica2>
- Video: <https://github.com/javi22gd/Practica2/blob/main/Presentacion/Presentacion.mp4>

### Índice

- 1. Descripción del dataset
- 2. Integración y selección de los datos de interés a analizar
- 3. Limpieza de los datos
  - 3.1. Valores vacíos
  - 3.2. Outliers
- 4. Análisis de los datos
  - 4.1. Selección de datos
  - 4.2. Normalidad y homocedasticidad
  - 4.3. Pruebas estadísticas
- 5. Representación visual
- 6. Resolución/Conclusiones
- 7. Código
- 8. Datasets
  - 8.1. Original
  - 8.2. Intermedio
  - 8.3. Final

### 1. Descripción del dataset

Strava es una red social enfocada a deportistas, principalmente ciclistas y corredores, basada en compartir y realizar un seguimiento de las actividades deportivas.



En Strava, cada usuario puede subir y publicar todas las actividades deportivas registradas mediante algún dispositivo como ciclómetros, smartwatches o smartphones, entre otros. Estos dispositivos reciben datos provenientes de un gran número de sensores conectados como input y que se traducen en multitud de parámetros: posición GPS, temperatura, frecuencia cardíaca, cadencia, potencia, etc.

Todos estos datos quedan registrados en archivos de formato .fit, más conocidos como **actividades** en Strava. Tras años de registros y publicaciones, un usuario puede llegar a tener cientos o miles de **actividades**, cada una con una gran variedad de datos. El sitio web de Strava permite acceder a una API para descargar todos estos registros, obteniendo un **dataset** muy interesante para estudiar.

El caso de estudio de esta práctica será el tratamiento de un **dataset** que, originalmente, contiene los datos de 1561 **actividades** de todo tipo desde 2016 hasta 2021 de Strava, con 81 variables cada una. El objetivo de este tratamiento será limpiar este **dataset** crudo para, posteriormente, realizar un análisis sobre la actividad del usuario a nivel deportivo a lo largo de 5 años.

[Sígueme en STRAVA](#)

### 2. Integración y selección de los datos de interés a analizar

Las variables que se mantienen en el **dataset** son las siguientes:

- Fecha.de.la.actividad:** Indica la fecha en la que se inició la actividad.
- Tipo.de.actividad:** Muestra de qué tipo de actividad se trata (p.e. bicicleta, natación, etc.).
- Tiempo.transcurrido:** Segundos transcurridos entre el inicio de la actividad y el final.
- Distancia:** En las actividades que cuentan con datos de ubicación GPS, esta variable indica la distancia total recorrida.
- Ritmo.cardíaco.máximo:** En las actividades que cuentan con datos cardíacos de un sensor de frecuencia cardíaca, este atributo muestra las pulsaciones por minuto máximas registradas durante la actividad.
- Esfuerzo.Relativo:** Este dato es calculado por Strava para cuantificar el esfuerzo que ha requerido la por parte del usuario en función de múltiples parámetros.
- Equipamiento.de.la.actividad:** Para las variables de tipo **Bicicleta** y **Bicicleta virtual**, este atributo indica la bicicleta utilizada para realizar la actividad.
- Tiempo.en.movimiento:** Muestra los segundos transcurridos mientras la actividad se lleva a cabo; es decir, es el tiempo durante el cual se está efectuando el ejercicio sin contabilizar el tiempo en el que la actividad se ha pausado.
- Velocidad.máxima:** En las actividades que cuentan con datos de ubicación GPS, esta variable indica la velocidad máxima que se ha registrado durante la actividad.
- Velocidad.promedio:** En las actividades que cuentan con datos de ubicación GPS, esta variable indica la velocidad media durante la actividad. Se puede calcular como la **Distancia** total recorrida dividida entre el **Tiempo.en.movimiento**.
- Desnivel.positivo:** En las actividades que cuentan con datos de ubicación GPS, esta variable muestra la distancia acumulada total ascendida. No se debe confundir con la diferencia de altitud entre el inicio y el final de la actividad, ya que se trata de los metros acumulados de ascensiones.
- Pendiente.máxima:** En las actividades que cuentan con datos de ubicación GPS, esta variable indica la pendiente máxima recorrida durante la actividad; es decir, la diferencia máxima entre la distancia horizontal y la distancia vertical en un momento dado.
- Pendiente.promedio:** En las actividades que cuentan con datos de ubicación GPS, esta variable indica la pendiente media recorrida durante la actividad, que se calcula teniendo en cuenta la **Distancia** y el **Desnivel.positivo** de la misma.
- Cadencia.máx:** En las actividades que cuentan con datos de sensores de cadencia, esta variable indica las revoluciones por minuto máximas (pedaladas o pasos), que se calculan como las pedaladas o pasos totales divididos entre los minutos del **Tiempo.en.movimiento**.
- Ritmo.cardíaco.promedio:** En las actividades que cuentan con datos cardíacos de un sensor de frecuencia cardíaca, este atributo muestra el número total de pulsaciones dividido entre los minutos del **Tiempo.en.movimiento**.
- Calorías:** Indica una estimación calculada por Strava del número total de kcal quemadas durante la actividad en base a diversos parámetros.
- Tiempo.de.descanso:** Este atributo no está presente en el **dataset** original y se calcula mediante la diferencia entre el **Tiempo.transcurrido** y el **Tiempo.en.movimiento**, indicando el tiempo en el que la actividad ha estado pausada mientras se llevaba a cabo, por lo que se puede entender como el tiempo de descanso total.

### 3. Limpieza de los datos

Lo primero de todo es resolver un pequeño problema con el nombre de algunas variables. Cuando Strava exporta los datos, traduce el nombre de las variables del inglés al idioma predeterminado del usuario: español, en este caso. Al hacer esta traducción, algunos nombres de las variables pueden dar error, lo que resulta en un nombre de la variable que se muestra como una sentencia HTML. Esto se corrige con el código Python utilizado en esta práctica, que exporta el **dataset** **activities2.csv** con todos los nombres correctos de las variables.

Tras eliminar el resto de variables y mantener solo las anteriores, se revisan una a una. El formato y/o tipo de algunas variables no es el correcto, por lo que es necesario modificarlo.

Por ejemplo, el formato de la variable **Fecha.de.la.actividad** tiene el formato de origen "26 nov. 2018 8:03:27" y formato **character** interpretado por R. Para pasarlo a formato **date**, primero hay que transformarlo a **POSIXct** para interpretar la cadena de caracteres anterior. Sin embargo, hay 133 registros que, por alguna razón, no se pueden interpretar y dan como resultado un valor vacío. Analizando estos datos, todos coinciden en ser del mes de septiembre con el siguiente formato: "12 sep. 2015 7:56:13". La función para transformar una cadena de tipo **character** a **POSIXct** interpreta los meses abreviados como una cadena de 3 o 4 caracteres; como "sep.", tiene 5, no podía interpretarlo. Para solucionarlo, se eliminan de estos registros la letra "e" para que el formato del mes sea "sep.", y de esta forma pueda ser interpretado correctamente. Tras comprobar que todos los registros se han convertido correctamente, se transforman a tipo **Date**.

Otro ejemplo es el de la variable **Tipo.de.actividad**. R la interpreta originalmente como **character**, pero el formato idóneo sería **factor** debido a que se tratan de etiquetas más que de cadenas de caracteres individuales y diferentes. De esta forma, se convierte obteniendo las etiquetas **Bicicleta**, **Bicicleta virtual**, **Entrenamiento con pesas**, **Entrenamiento**, **Caminata**, **Carrera**, **Carrera virtual** y **Natación** para todo el conjunto de datos.

Como último ejemplo, la variable **Distancia** también es interpretada por R con **character** siendo todos los datos valores numéricos. Esto debe a que el separador de los valores decimales son originalmente ",". Para solucionarlo, se sustituyen en todos los registros por "," y se convierten a tipo **double**. No obstante, se observa que el valor máximo es irreal (13500 km). Este registro pertenece a la actividad de **Natación**, que originalmente está en metros. Por lo tanto, se pasa a km dividiendo entre 1000, manteniendo así la integridad de los datos.

#### 3.1. Valores vacíos

La mayoría de variables tienen valores vacíos ya que algunas variables solo aplican a un cierto tipo de actividad, dejándose vacías en las demás. Esto ocurre, por ejemplo, con los atributos derivados de datos GPS en actividades que no registran estos datos, como es el caso de **Entrenamiento con pesas**. En estas situaciones, es correcto dejar estos campos vacíos, siempre y cuando en el análisis posterior esto se tenga en cuenta.

No obstante, en otros casos si es necesario realizar una imputación de los valores vacíos. Por ejemplo, la variable **Ritmo.cardíaco.máximo** está vacía en algunos casos, por lo que se calcula la media de esta variable en el resto de registros diferenciando por el **Tipo.de.actividad** y se impone a los valores vacíos de cada tipo.

Otros casos algo más elaborados, como con la variable **Velocidad.promedio** en las actividades que sí deberían tener este dato, se calcula mediante el cociente de la distancia total recorrida y el tiempo en movimiento.

El ejemplo más complejo es el de la variable **Calorías**. Estos datos no dependen directamente del valor aportado por un sensor concreto, sino que lo calcula Strava mediante los valores de diferentes parámetros. Para hacer una estimación lo más aproximada posible a este cálculo, se estudió la correlación de diferentes variables con la variable **Calorías**, resultando estar altamente correlacionado (>.92) con el producto de las variables **Ritmo.cardíaco.promedio** y **Tiempo.en.movimiento**. Posteriormente, se construye un modelo de regresión lineal con estas variables por cada tipo de actividad y se predice el valor de las calorías quemadas a partir de dicho valor.

#### 3.2. Outliers

Un ejemplo del tratamiento de valores extremos es el de la variable **Pendiente.máxima**. Este atributo se obtiene mediante los sensores GPS utilizados durante la actividad, siendo el coeficiente máximo registrado entre la distancia vertical y horizontal recorrida en un momento dado. Cuando el sensor GPS pierde señal durante unos instantes, se pueden falsear los datos de posición espacial obtenidos puntualmente, pudiendo provocar un pico en este coeficiente. No existe una solución óptima para este problema debido a que esta variable es completamente independiente del resto; ya que la distancia, el tiempo o la velocidad (entre otros) no afectan al cálculo del valor de este atributo. Teniendo esto en cuenta, lo único que se puede hacer es detectar los valores irrealistas, diferenciando por tipo de actividad ya que, por ejemplo, no es lo mismo una pendiente del 35% en una actividad de tipo **Bicicleta** (casi imposible) que en una de **Caminata** (subir escaleras); del mismo modo, no es lo mismo una pendiente del 30% en actividades de **Bicicleta** por asfalto que en actividades del mismo tipo por campo, ya que en carretera no es posible alcanzar esa inclinación en una pendiente mientras que en un camino de tierra se podría llegar a dar. Teniendo esto en cuenta, como el tipo de actividad e, incluso, por el material utilizado (la bicicleta utilizada para las actividades de tipo **Bicicleta** indica si es por carretera o por campo) y se establece un criterio para definir a partir de qué valor de la pendiente se considera imposible. Tras esto, se seleccionan los registros cuya **Pendiente.máxima** está por encima de dicho valor y se le impone el valor máximo razonable.

### 4. Análisis de los datos

#### 4.1. Selección de datos

Las variables analizadas son **Fecha.de.la.actividad** y **Tiempo.en.movimiento** para las actividades de tipo **Bicicleta**, **Bicicleta virtual**, **Entrenamiento con pesas**, **Entrenamiento**, **Entrenamiento** y **Caminata**.

#### 4.2. Normalidad y homocedasticidad

El resultado de estudiar la **normalidad** de una de las variables anteriores por cada tipo de actividad es que ninguna de ellas sigue una distribución normal. El procedimiento seguido se explica en el siguiente apartado.

Para la variable **Fecha.de.la.actividad** se analiza la homocedasticidad sobre los siguientes pares de variables, obteniendo los siguientes resultados:

- Bicicleta & Bicicleta virtual:** Varianzas distintas.
- Bicicleta virtual & Entrenamiento con pesas:** Varianzas distintas.
- Entrenamiento con pesas & Entrenamiento:** Varianzas iguales.

Para la variable **Tiempo.en.movimiento** se analiza la **homocedasticidad** sobre los siguientes pares de variables, obteniendo los siguientes resultados:

- Bicicleta & Bicicleta virtual:** Varianzas distintas.
- Bicicleta virtual & Caminata:** Varianzas iguales.
- Entrenamiento con pesas & Caminata:** Varianzas iguales.

#### 4.3. Pruebas estadísticas

Para comprobar la **normalidad** de los datos anteriormente descritos se realiza el test de **Lilliefors**, que consiste en realizar un contraste de hipótesis donde la hipótesis nula es que la distribución de los datos es normal.

H0: Los datos **sí** siguen una distribución normal.

H1: Los datos **no** siguen una distribución normal.

En el caso de la igualdad de varianzas, u **homocedasticidad**, se ha realizado el **F test**, que formula las siguientes hipótesis:

H0:  $\sigma_1^2 = \sigma_2^2$

H1:  $\sigma_1^2 \neq \sigma_2^2$

La tercera prueba estadística realizada es el test de **correlación de Pearson** sobre la variable **Calorías** como dependiente y el producto de las variables **Ritmo.cardíaco.promedio** y **Tiempo.en.movimiento** como independiente. El resultado es una fuerte correlación, por lo que se construye un modelo de **regresión lineal** para estimar los valores vacíos de la variable **Calorías** en función del producto de las otras dos.

### 5. Representación visual

#### 5.1. Regresión lineal

A continuación se muestra la variable **Calorías** en el eje Y y el producto entre las variables **Ritmo.cardíaco.promedio** y **Tiempo.en.movimiento** en el eje X por cada tipo de actividad estudiado. La recta representa el modelo de regresión lineal construido para realizar la imputación de los valores vacíos explicado anteriormente.

En la figura se observa que la recta de regresión lineal es casi vertical, lo que indica que la variable **Calorías** es casi constante.

En la figura se observa que la recta de regresión lineal es casi vertical, lo que indica que la variable **Calorías** es casi constante.

En la figura se observa que la recta de regresión lineal es casi vertical, lo que indica que la variable **Calorías** es casi constante.

En la figura se observa que la recta de regresión lineal es casi vertical, lo que indica que la variable **Calorías** es casi constante.

En la figura se observa que la recta de regresión lineal es casi vertical, lo que indica que la variable **Calorías** es casi constante.

En la figura se observa que la recta de regresión lineal es casi vertical, lo que indica que la variable **Calorías** es casi constante.

En la figura se observa que la recta de regresión lineal es casi vertical, lo que indica que la variable **Calorías** es casi constante.

En la figura se observa que la recta de regresión lineal es casi vertical, lo que indica que la variable **Calorías** es casi constante.

En la figura se observa que la recta de regresión lineal es casi vertical, lo que indica que la variable **Calorías** es casi constante.

En la figura se observa que la recta de regresión lineal es casi vertical, lo que indica que la variable **Calorías** es casi constante.

En la figura se observa que la recta de regresión lineal es casi vertical, lo que indica que la variable **Calorías** es casi constante.

En la figura se observa que la recta de regresión lineal es casi vertical, lo que indica que la variable **Calorías** es casi constante.

En la figura se observa que la recta de regresión lineal es casi vertical, lo que indica que la variable **Calorías** es casi constante.

En la figura se observa que la recta de regresión lineal es casi vertical, lo que indica que la variable **Calorías** es casi constante.

En la figura se observa que la recta de regresión lineal es casi vertical, lo que indica que la variable **Calorías** es casi constante.

En la figura se observa que la recta de regresión lineal es casi vertical, lo que indica que la variable **Calorías** es casi constante.

En la figura se observa que la recta de regresión lineal es casi vertical, lo que indica que la variable **Calorías** es casi constante.

En la figura se observa que la recta de regresión lineal es casi vertical, lo que indica que la variable **Calorías** es casi constante.

En la figura se observa que la recta de regresión lineal es casi vertical, lo que indica que la variable **Calorías** es casi constante.

En la figura se observa que la recta de regresión lineal es casi vertical, lo que indica que la variable **Calorías** es casi constante.

En la figura se observa que la recta de regresión lineal es casi vertical, lo que indica que la variable **Calorías** es casi constante.

En la figura se observa que la recta de regresión lineal es casi vertical, lo que indica que la variable **Calorías** es casi constante.

En la figura se observa que la recta de regresión lineal es casi vertical, lo que indica que la variable **Calorías** es casi constante.</p