***Actividad 2***

05MBIG – Métodos estadísticos para el análisis e interpretación de datos masivos

Contenido

[1. Introducción 3](#_Toc58102339)

[2. Objetivo 3](#_Toc58102340)

[3. Dataset 4](#_Toc58102341)

[4. Proceso KDD 5](#_Toc58102342)

[4.1. Selección 5](#_Toc58102343)

[4.2. Preprocesado 5](#_Toc58102344)

[4.3. Transformaciones y nuevas variables 5](#_Toc58102345)

[4.4. Data Mining 7](#_Toc58102346)

### [Preprocesamiento y calidad de datos](https://campus.viu.es/webapps/assignment/uploadAssignment?content_id=_2881262_1&course_id=_81894_1&group_id=&mode=view)

Actividad 2: Preprocesamiento y calidad de datos

El **preprocesamiento** de datos es un paso preliminar durante el proceso de minería de datos. Se trata de cualquier tipo de procesamiento que se realiza con los datos brutos para transformarlos en datos que tengan formatos que sean más fáciles de utilizar.

Para ello en esta activadad, se buscará un conjunto de datos, se importaran  y se realizarán los pasos necesarios que el alumno considere considerando la información que aparece en el artículo:

Salvador García, Sergio Ramírez-Gallego, Julián Luengo,Francisco Herrera.  Big Data: Preprocesamiento y calidad de datos. Novática, nº 237. Julio-Octubre 2016 17-23

Se debe de entregar un **zip** con el notebook y los dataset utilizados. (En el caso de qu e el repositorio se pueda descargar de manera remota, no es necesario incluirlo)  El valor será el 30% de la nota total.

Recomiendo el uso de Google Colab:   https://colab.research.google.com/notebooks/intro.ipynb#, pero se puede utilizar cualquier herramienta que se pueda utilizar multiplataforma.

-

La entrega retrasada se permitirá hasta el 10 de Enero, pudiendo obtener hasta 2 puntos de la nota total. (20% )

1. **Introducción**

En los últimos años los datos se han vuelto muy importantes en el mundo del deporte profesional tanto a la hora de analizar tácticas y resultados como a la hora de analizar el rendimiento de un deportista.

El ciclismo es uno de esos deportes que ha evolucionado mucho en este aspecto en los últimos 20 años, sobre todo en la generación de datos en cuanto al rendimiento del ciclista se refiere. Actualmente se recogen multitud de datos en periodo de competición y de entrenamiento como pueden ser: velocidades, ritmo cardiaco, potencia, estrés del ejercicio, fatiga acumulada…los cuales ayudan a planificar entrenamientos específicos y personalizados para un ciclista según los objetivos de la temporada.

Entre todos estos datos recogidos, la potencia es uno de los parámetros más importantes a la hora de medir el estado de forma de un ciclista. Una métrica habitual a medir es el FTP (Functional Threshold Power) o lo que es lo mismo, la potencia media que es capaz de mantener un ciclista durante 1 hora. Esta variable dividida por el peso del ciclista daría el W/kg y es una medida comparable con otro atleta donde se puede medir el nivel de cada uno.

1. **Objetivo**

Esta actividad se va a centrar en la variable de potencia. Se pretende diseñar un modelo donde dado un dataset compuesto por cerca de 500 registros de entrenamiento y 300 variables, sea capaz de predecir la potencia media que va a generar el deportista durante una salida en bicicleta de X kilómetros. Esta predicción es muy útil para saber si se afronta una carrera en unas condiciones óptimas de entrenamiento: ni poco entrenado, ni en estado de fatiga.

Para ellos no solo habrá que tener en cuenta la tendencia de potencia de los últimos entrenamientos, si no también variables como el estrés acumulado en los últimos días, los días de recuperación entre entrenamientos o frecuencia cardiaca entre otras.

1. **Dataset**

Como dataset se van a utilizar los datos exportados de la aplicación *Golden Cheetah*. Esta es una aplicación donde se registran diferentes actividades deportivas y las métricas de un atleta (ciclismo, correr, nadar).

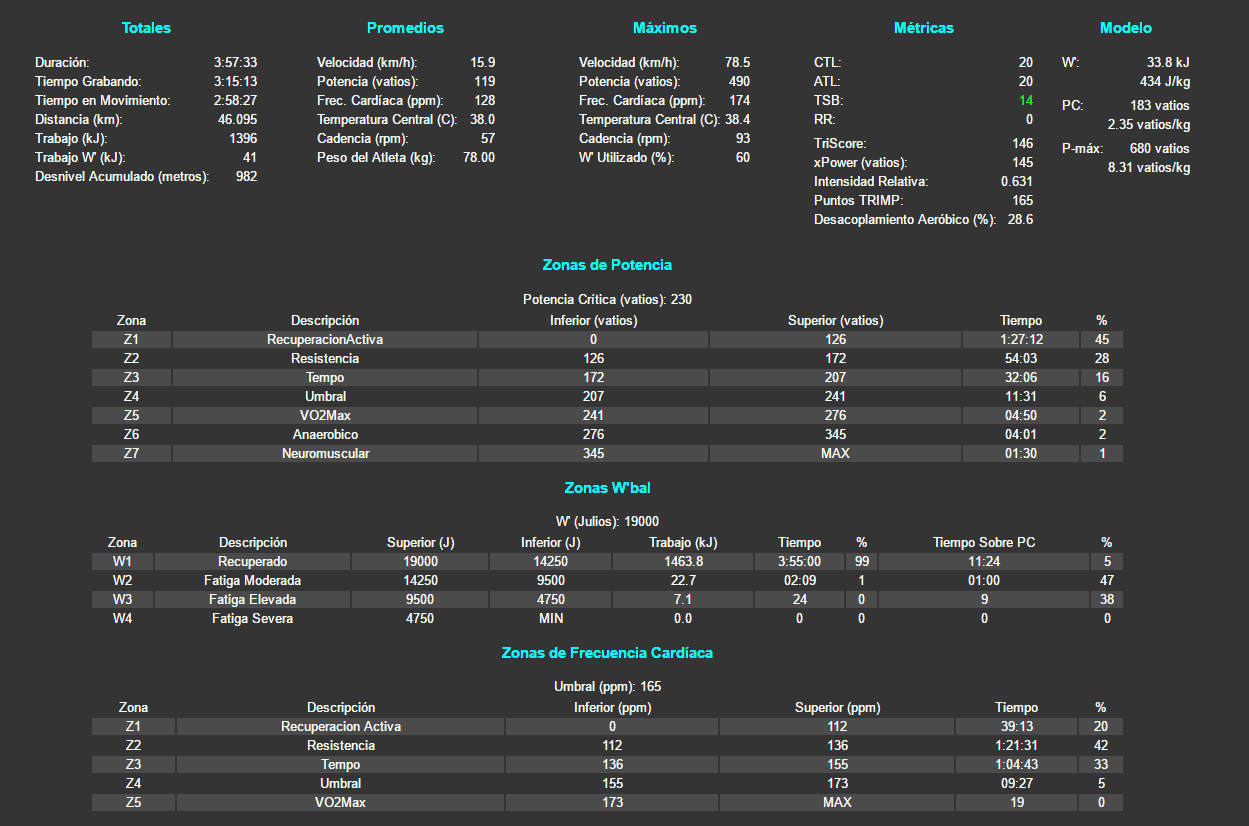


Ilustración 1: Aplicación Golden Cheetah.

En la imagen anterior se puede observar alguno de los datos registrados durante una sola actividad deportiva en *Golden Cheetah*. Los datos de todas las actividades serán exportados a un fichero csv separado por comas (“,”) y es lo que se usará como dataset.

Estos datos necesitan ser pre-procesados pues, por ejemplo, solo interesan las actividades de ciclismo y, además, no todas las actividades de ciclismo aquí guardadas tienen datos de potencia y frecuencia cardiaca por lo que habrá que filtrarlas.

1. Proceso KDD

En este apartado se detalla cada paso a seguir en el proceso KDD para lograr el objetivo propuesto y dado el dataset explicado en el punto anterior.

* 1. Selección

La aplicación Golden Cheetah de donde extraemos el dataset guarda actividades de ciclismo, natación y carrera. Para el objetivo que nos ocupa solo son interesantes en un principio las actividades de ciclismo, por los que descartaremos las demás. En una futura ampliación podría tenerse en cuenta también las otras actividades pues también generan una fatiga que pueden afectar al resultado final.

* 1. **Preprocesado**

Valores nulos

Analizado el dataset, dado que las variables más importantes son “potencia media” y “frecuencia cardiaca”, se eliminan todos los registros donde alguna de estas variables es nula o cero.

Outliers

Del mismo modo que para los valores nulos, se eliminan los registros donde estas mismas variables están por encima de cierto valor. Para decidir si el valor esta fuera de rango se sigue el siguiente criterio:

“Potencia Media”: Se grafican los valores y se establece un límite de forma manual.

“Frecuencia cardiaca media”: Se eliminan los valores por encima de 200 pulsaciones.

* 1. **Transformaciones y nuevas variables**

Antes de comenzar con el proceso de minería de datos, es necesario convertir las unidades de algunas variables, así como crear otras nuevas variables a partir de las existentes.

Fechas

Analizando el dataset se encuentra que en la variable fecha, el año, aparece en algunos registros con 2 dígitos y en otros con 4. Se procede a unificar todas las fechas al formato de año con 4 dígitos (mm/dd/YYYY).

Distancia

Al extraer los datos se observa que la distancia ha sido exportada multiplicada por un factor de x10. Se corrige esta variable dividiendo este valor en todo el dataset entre 10 para tener la medida en km.

Días de Descanso

Para cada entrenamiento, se necesita calcular el número de días que han pasado desde el anterior entrenamiento. Esta nueva variable se llama “DiasDescanso”.

Esta nueva variable es necesario para calcular los dos siguientes parámetros: CTL y ATL.

CTL

El CTL hace referencia al nivel de carga del entrenamiento que ha acumulado un ciclista durante los últimos 42 días. Esta nueva variable se calcula con la siguiente fórmula:

Donde TSS es la intensidad del entrenamiento. Esta variable depende de la potencia media generada en el entrenamiento.

ATL

El parámetro ATL es una estimación del nivel de fatiga producida por los entrenamientos realizados durante los últimos siete días. Esta nueva variable se calcula con la fórmula:

Donde TSS es la intensidad del entrenamiento. Esta variable depende de la potencia media generada en el entrenamiento.

Mejor potencia 20’ últimos 15 días

A partir de la variable de potencia en 20’, calcular el mejor registro de potencia en 20’ de los últimos 15 días. Con esta nueva variable se pretende aproximar de una forma más precisa el estado de forma.

* 1. Data Mining

En esta última parte de proceso, con los datos previamente pre-procesados y con las nuevas variables calculadas, se pretende aplicar un modelo de **Regresión Lineal Múltiple** para estimar la potencia media que podrá dar un ciclista en un entrenamiento de X horas teniendo en cuenta su estrés acumulado y la carga de entrenamiento previa.

Para este modelo la variable dependiente *Y* es la “Potencia Media” mientras que la variable independiente *X* está compuesta de la siguiente forma:

Como paso previo a la generación del modelo, se realiza un estudio de la correlación entre las variables independientes por si hubiese que eliminar alguna variable que aportase información duplicada.