Predicción de tiempos de carrera a pie mediante la aplicación de algoritmos de Machine Learning

José R. Vinueza[1]

[1] *IT Academy, Data Science, Barcelona, España*

Resumen— Salir a correr se ha convertido en una de las actividades físicas más comunes en España, y su práctica ha tenido un incremento importante en los últimos años. Cada vez es más frecuente el uso de dispositivos que nos permiten registrar los datos que se generan alrededor de esta actividad. El presente trabajo, busca crear un modelo de regresión mediante la aplicación de técnicas y algoritmos de machine learning que ayude a predecir el tiempo que una persona puede tardar en correr una distancia determinada. La evaluación de los resultados se realizará mediante el error absoluto medio (MAE) y el análisis del error cuadrático (R2).

Palabras clave— Regresión - Machine Learning - Aprendizaje Supervisado - Algoritmos - Predicción - Dataset - Error Absoluto Medio (MAE)

I. Introducción

En los últimos años, salir a correr se ha convertido en una de las actividades físicas más comunes en España. El porcentaje de la población que practica este deporte (ver fig.1), ha tenido un incremento considerable desde el 2018 [1]. Por otro lado, el desarrollo de aplicaciones Big Data también ha tenido un incremento importante, hoy en día se generan grandes volúmenes de datos en diferentes sectores o industrias, y el deporte es uno de esos [2].

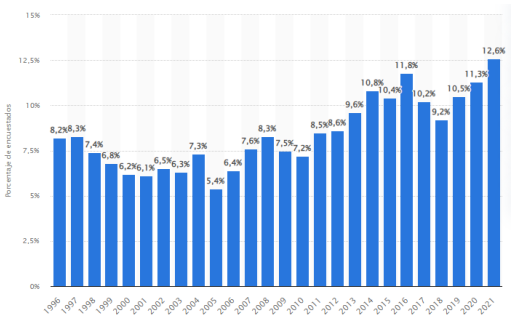

Fig. 1: Porcentaje de individuos que practicó running en España de 1996 a 2021.

Los rastreadores (sensores) de ejercicio portátiles proporcionan datos que codifican la información sobre el rendimiento individual de los corredores. Estos datos tienen un gran potencial para mejorar nuestra comprensión de la compleja interacción entre el entrenamiento y el rendimiento [3].
El objetivo del presente trabajo es crear un modelo de machine learning, que en función de las características de un corredor y los datos recopilados mediante el uso de un reloj de pulsera, nos permita lograr predecir el tiempo que tardará en correr una distancia indicada.

II. Estado del Arte

Existen trabajos realizados, donde aplicando un modelo matemático con datos de 14.000 individuos con 1,6 millones de sesiones de ejercicio que contienen la duración y la distancia, con una distancia total aproximada de 20 millones de km. Adicionalmente, el modelo depende de dos parámetros: un índice de potencia aeróbica y un índice de resistencia. La inclusión de la resistencia, que describe la disminución de la potencia sostenible a lo largo de la duración, ofrece nuevas perspectivas sobre el rendimiento: una predicción muy precisa del tiempo de carrera y la identificación de parámetros clave como el umbral de lactato, comúnmente utilizado en la fisiología del ejercicio[3].
Por otro lado, ya existen dispositivos de gama alta (relojes), que incorporan un visualizador de tiempos de carrera estimados, mismos que están en función de la frecuencia cardiaca máxima y utilizan el valor del VO2 máximo (esto es el máximo volumen de oxígeno en mililitros que puedes consumir por minuto y por kilogramo de peso corporal en el punto de máximo rendimiento) combinado con el historial de varias semanas de entrenamiento del corredor, para ofrecer estimaciones de tiempo de carrera más precisas[4].
Mediante el presente trabajo, lo que buscamos es brindar una herramienta u opción gratiuta, que permita al usuario predecir su tiempo de carrera mediante la aportación de datos/ registros de sus entrenamientos.

III. Metodología

a. Dataset

El conjunto de datos (Dataset) utilizado es de dominio propio, y son actividades deportivas registradas con un dispositivo Garmin (en este caso un reloj de pulsera). Dichas actividades están almacenadas, por medio de un usuario y contraseña, en la aplicación Garmin Connect [5], que es una her-

ramienta para realizar seguimiento, analizar y compartir actividades deportivas.
El Dataset cuenta con un total de 1569 actividades a partir del año 2017 hasta 2022, para un único usuario.

b. Preparación de datos

Se identificaron 49 características (columnas) diferentes dentro del dataset, sin embargo, lo que nos interesa de este Dataset, es principalmente toda la información correspondiente a las actividades de carrera a pie, para las demás actividades, como por ejemplo ciclismo, natación, fuerza, etc., solo se conservarán los parámetros que coincidan con la carrera (ej.: distancia, frecuencia cardíaca), pero no los específicos de cada uno (ej: repeticiones, brazadas), por lo que se decide eliminar esas columnas que no aportan ningún tipo de valor para la predicción del tiempo de carrera.
Se encontraron columnas, que contenían caracteres especiales (NaN) para definir que no existe información en ese campo, y al ser pocos registros se optó por eliminar esas filas. En columnas donde el número era elevado, se optó por reemplazar con valor 0, ya que pueden tratarse de fallas del reloj, o a su vez de actividades indoor, donde no se consideran/ registran esos parámetros puntuales.
Posteriormente, se procedió a corregir el tipo de dato (Dtype) del dataset, ya que todos los valores se encontraban como objeto o string. Se transformaron los formatos de fechas, se asignó Dtype de entero o decimal según corresponde, los tiempos correspondientes a ritmos se unificaron todos como velocidad (km/h), ya que la aplicación registra velocidades para actividades como ciclismo y ritmos (min/km) para actividades restantes, y al tiempo registrado que estaba en formato HH:MM:SS, se lo transformó a segundos.
Finalmente, la columna categórica de actividad y la fecha se transformó a números, para poder utilizar estos valores dentro del algoritmo.

c. Creación de modelos

Para predecir el tiempo de carrera a pie, se emplearon modelos de aprendizaje supervisado, en específico algoritmos de regresión. Al existir varios algoritmos de machine learning dentro de la librería Scikit-learn [6], se escogieron tres en función de su uso, aplicación y requisitos de capacidad de cálculo computacional, siendo estos los siguientes:

- Nearest Neighbors (KNN)
- Support Vector Machines (SVM)
- Random Forest (RF)

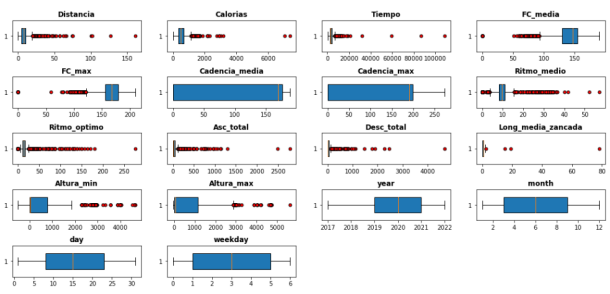

Fig. 2: Anális de las características de cada columna de datos.

Además, posteriormente al análisis de cada columna (ver fig.2), se decidió normalizar todo el conjunto de datos en función de cada columna, por lo que se procedió a transformar de la siguiente manera:

- Columnas categóricas: OneHotEncoder
- Columnas con outliers: RobustScaler
- Columnas sin outliers: MinMaxScaler

Para cada modelo seleccionado, se determinaron diferentes hiperparámetros, con los cuales se efectuó un gridsearch para lograr ajustar los mismos, evaluando su puntaje para obtener un MAE (Mean Absolute Error) lo más bajo posible y un R2 lo más cercano a 1.
Posteriormente, se realizó una validación cruzada (cross validation) con el total de los datos, y finalmente se ajustó el modelo para obtener las métricas de análisis planteadas (MAE y R2).

d. Resultados

En la tabla 1, se muestran los resultados obtenidos de cada modelo, con la transformación de datos correspondiente y los mejores hiperparámetros de cada uno.

TABLA 1: MÉTRICAS FINALES DE LOS DIFERENTES MODELOS.

Classifier Name	R2 score	MAE score
Nearest Neighbors Regressor	0.45	595.60
Support Vector Machines	0.70	480.99
Random Forest Regressor	0.83	300.35

Como se puede observar en la tabla 1 ,las mejores métricas se obtuvieron con el algoritmo Random Forest, y mediante los hiperparámetros de 'max features': 'auto' y 'n estimators': 147.
Este modelo, además de ofrecer el menor MAE, 300.35 segundos, lo que equivale a 5 minutos (00:05:00), también presenta el mayor R2 con un valor de 83%.

Finalmente, en la fig.3, se muestran las variables que tienen mayor importancia dentro del modelo seleccionado, siendo las de mayor peso: el ritmo medio y el ritmo óptimo, la altura máxima, el ascenso total, el año y la distancia.
De estas 5 variables con mayor importancias, es razonable ver que la distancia, el desnivel y altura sobre el nivel del mar afecten al tiempo, ya que un corredor varía sus tiempos y esfuerzos dependiendo de la distancia que corra y el desnivel en subida afectará la velocidad.
Encuanto el año, es evidente que existe una evolución con el pasar del tiempo. Finalmente el modelo también está basado en un ritmo (velocidad) óptimo y medio, que el corredor logra correr durante una distancia y condiciones determinadas, sin embargo, al conocer este dato y la distancia que se correrá es muy simple lograr predecir el tiempo que se requiere, pero al mismo tiempo estas variables son un condicionante para el modelo predictivo ya que son datos que no se pueden saber con antelación pero lo que se puede hacer es un análisis de los últimos entrenamientos o

registros que se tengan para poder estimar un valor medio e introducirlo en el modelo predictivo.

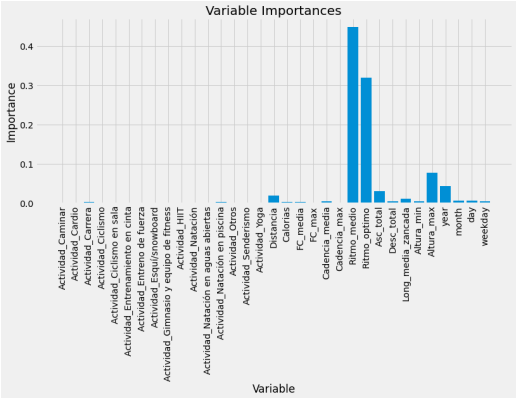

Fig. 3: Importancia de las variables

IV. Conclusiones

R2 es una medida estadística que representa la bondad de ajuste de un modelo de regresión. El valor ideal de R2 es 1, y cuanto más se acerque a 1, mejor se ajusta el modelo; sin embargo, el resultado obtenido de 0.83, se considera un resultado bastante favorable para el modelo utilizado y dentro del campo que se está haciendo uso.

Sin lugar a duda, siempre es posible utilizar más datos o parámetros propios del corredor como es el VO2 máximo para lograr ajustar los resultados.

A penas 17 de las 33 variables utilizadas tienen una importancia considerable dentro del modelo utilizado, lo que nos sugiere eliminar las que no aportan valor y por otro lado implementar nuevas variables que ayuden a incrementar la precisión del modelo.

Referencias

[1] S. R. Department. (2011) Porcentaje de individuos que practicó running en españa de 1996 a 2021. Tomado de https : / / es . statista . com / estadisticas / 569559 / evolucion-del-porcentaje-de-individuos-que-practicaron-running/ (14/10/2022).

[2] A. L. A. Oussous, F.Benjelloun and S. Belfkiha, "Big Data technologies: A survey," *Journal of King Saud University - Computer and Information Sciences*, vol. 30, no. 1, pp. 431–448, 2018.

[3] T. Emig and J. Peltonen, "Human running performance from real-world big data," *Nature Communications*, no. 1, pp. 1–9, 2020.

[4] G. Connect. Manual de usuario. Tomado de https : / / www8 . garmin . com / manuals / webhelp / fenix6-6ssport / ES-XM / GUID-31B2458A-859A-4A34-AB83-224E4A29387A . html# : ~ : text = El % 20dispositivo % 20utiliza % 20la % 20estimaci % C3 % B3n , tiempo%20de%20carrera%20m%C3%A1s%20precisas. (23/11/2022).

[5] ——. (2022) La aplicación garmin connect. Tomado de https://www.garmin.com/es (23/11/2022).

[6] F. P. et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, no. 1, pp. 2825–2830, 2011.