Predicción de la quema de calorias durante el ejercicio

19 de febrero de 2023

Uso de técnicas de Machine Learning: Método Jerarquico + XGB Regressor

AUTOR: Robledo, Nallely.[a,*]

[a] *Universidad Autonoma de Nuevo León, Facultad de Fisicomatematicas*
Pedro de Alba S/N, Niños Héroes, Ciudad Universitaria, San Nicolás de los Garza, N.L.

To cite this article: Robledo, Nallely. 2023. Calorie burn prediction during exercise. Universidad Autonoma de Nuevo León, Facultad de Ciencias Fisicomatematicas 00, 1-5.

Resumen

Si la gente respondiera honestamente a la pregunta '¿Cuáles son las razones por las que haces ejercicio?', una respuesta frecuente sería quemar calorías. De hecho, según el Departamento de Salud y Servicios Humanos de EE. UU. (1992), el 26 porciento de los adultos estadounidenses entre 20 y 74 años tienen sobrepeso, lo que demuestra claramente el impacto de esta preocupación nacional.

Se sabe que la reducción de la grasa corporal puede revertir varios procesos de enfermedades (p. ej., diabetes tipo II, enfermedades cardíacas, etc.), el ejercicio aumenta el gasto calórico total y también maximiza la pérdida de grasa corporal y el mantenimiento o aumento de masa muscular, la participación en el ejercicio es una estrategia muy consecuente y gratificante para perder grasa corporal y mejorar su salud.

El ejercicio como medio para quemar calorías ha sido reconocido por la industria del fitness. Hay muchos tipos de modalidades de ejercicio que se comercializan con el reclamo de "quemar más calorías", y el consumidor se pregunta qué es lo que determina la cantidad de calorías quemadas durante el ejercicio. Esta situación es la razón fundamental para escribir este artículo.

Calorie burn prediction during exercise -

Abstract

If people were to honestly answer the question 'What are the reasons you exercise?' a frequent answer would be to burn calories. In fact, according to the US Department of Health and Human Services (1992), 26 percent of American adults ages 20-74 are overweight, clearly demonstrating the impact of this national concern.

It is known that reducing body fat can reverse various disease processes (eg, type II diabetes, heart disease, etc.), exercise increases total caloric expenditure and also maximizes body fat loss and maintenance or gaining muscle mass, engaging in exercise is a very consistent and rewarding strategy for losing body fat and improving your health.

Exercise as a means of burning calories has been recognized by the fitness industry. There are many types of exercise modalities that are marketed as "burning more calories.[a]nd the consumer wonders what determines the number of calories burned during exercise. This situation is the fundamental reason for writing this article.

1. Introducción

La principal directriz para la creación de este articulo es representar la solución a la interrogante: *¿Cual es el factor de mayor influencia en la quema de calorias durante el ejercicio?*

Se evaluan diferentes variables que involucran las caracteristicas fisicas de la persona que realiza el ejercicio, contra otras variables, que describen la manera de hacer el ejercicio. Las variables de estudio serán: Edad, Sexo, Peso, Estatura, Pulso cardiaco, Temperatura y Duración.

En reposo, el cuerpo gasta energía para mantener las funciones de las células que son esenciales para la vida. El bombeo continuo de sangre por parte del corazón exige energía, al igual que la ventilación continua (movimiento de aire hacia adentro y hacia afuera) de los pulmones. Además, mantener un entorno de soporte vital dentro y alrededor de las células

*Autor para correspondencia: autor1@ceautomatica.es

requiere una descomposición constante de ciertas moléculas liberadoras de energía. Esta energía también se utiliza para formar las moléculas necesarias para reparar las células, almacenar energía (glucógeno y triglicéridos), combatir infecciones y procesar los nutrientes obtenidos de la digestión. Estas funciones exigentes de energía se combinan para formar la tasa metabólica basal del cuerpo, que puede variar de aproximadamente 800 a 1500 Kcal dependiendo del tamaño del cuerpo y la ingesta calórica total (cantidad ingerida de alimentos)..

2. Marco teórico

El trifosfato de adenosina (ATP) es la molécula principal que el cuerpo utiliza como medio para utilizar la energía química para realizar el trabajo celular. El ejercicio aumenta el gasto calórico del cuerpo, ya que la contracción muscular implica la necesidad de formar y descomponer ATP repetidamente. La energía liberada por la descomposición del ATP alimenta la contracción del músculo esquelético, lo que aumenta las demandas de energía del cuerpo y aumenta el gasto calórico. Las investigaciones han demostrado que durante el ejercicio el aumento del gasto calórico se debe casi en su totalidad a la contracción del músculo esquelético; el equilibrio se debe a un aumento en las demandas de energía del corazón y los músculos utilizados durante la ventilación.

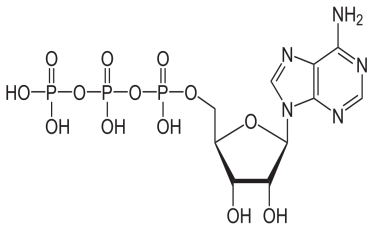

Figura 1: Estructura química del trifosfato, Adenosintriphosphat protoniert.svg, Dominio Publico

La investigación evaluará las variables en consideración para al quema de calorias y se definirá a travéz de modelos mátematicos los de mayor influencia en la quema de calorias. Al final de esta investigación se concluirá de manera lógica si estas variables estan influidas por el efecto de el ATP.

Los módelos mátematicos de Machine Learning seleccionados para el análisis: Método Jerarquico y XGBBoost, estan sustentados bajo la misma teoria de clásificación; que es la de los árboles de decisión. Esto, apesar de que siguen procedimientos totalmente adversos.

En general, los árboles de decisión clasifican datos a partir de su separación en regiones y obtienen una clasificación a partir de las cotas que limitan las regiones. Una vez obtenidas dichas regiones, la función de predicción.

Se busca concluir el resultado de esta investigación en base al modelo que generó un mejor desempeño. Para ello se utilizará la Media de Error Aboluto o MAE; esta es una métrica que mide el promedio de la diferencia absoluta entre las predicciones y los valores reales. Se puede expresar matemáticamente como:

$$\text{MAE}(y,) = \sum_{i=0}^{N-1} |y_i - \hat{y}_i| \overline{N}$$

3. Machine Learning

Esta documento hace una recopilación de conjuntos apropiados para enseñar a nuestros modelos de aprendizaje automático para que logre saber cuál es la cantidad de calorías que el individuo gasta para quemar.

Usaremos el Método jerarquico y Linear Regression como modelos de aprendizaje automático para comparar y luego evaluar estos modelos. La herramienta es Google Colab, el cual es un servicio basado en la nube.

3.1. Selección de caracteristicas

Matriz de Correlación

Primero se muestra la relación entre las variables independientes y su influencia sobre la variable dependiente que es Burned Calories. En color obscuro se marcaron aquellos más significantes para su fácil interpretación.

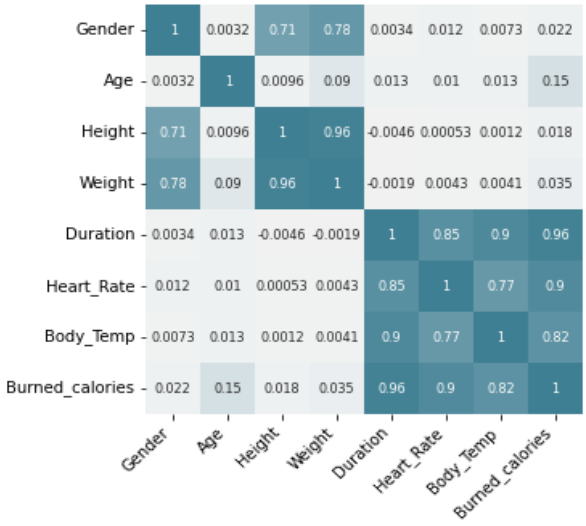

Figura 2: Gráfico Dendrograma

ANOVA

Esta herramienta es una fórmula estadística que sirve para comparar las varianzas entre las medias de diferentes grupos. Tambien se utiliza para determinar si existe alguna diferencia entre las medias de los diferentes grupos.

A continuación se presentan los resultados del valor de F, ordenando de las variables que tienen un valor más alto, al menos representativo.

Tabla 1: Resultados de ANOVA, F-Value por variable

Variable	F Value
Duration	157053.43
Heart Rate	62387.94
Body Temp	31855.44
Weight	366.25
Age	18.904356
Gender	7.5
Height	4.61

En base a esta evaluación se interpreta que un valor F alto, indica alta relación lineal; valores menores, lo contrario. Por lo

tanto asumimos que existen 3 variables con mayor relación con la variable de respuesta *Burned Calories* y estas son Duration, Heart Rate y Body Temp.

Información mutua

Esta mide la cantidad de información transferida cuando x^i= (variable de interes) es transmitido y y^i =(variable de respuesta) es recibido.

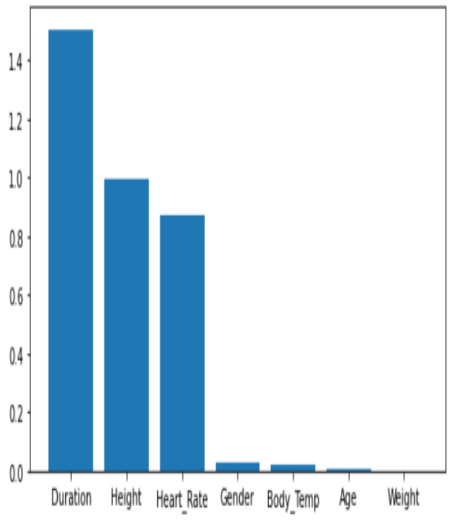

Figura 3: Gráfico de Información mutua entre variables

Con el grafico anterior podemos observar que las variables Duration, Height y Heart Rate comparten información mutua reelevante, en comparacion con el resto de las variables

En base a los análisis previos, se cree que la eliminación de las variables Age, Weight y Gender pudiera ser conveniente para el agrupamiento de caracteristicas, ya que estas variables tienen muy poca significancia y además comparten poca información mutua.

3.2. Agrupamiento de caracterisiticas y análisis de grupos

En esta sección se realizará el análisis de grupos o tambien conocido como clustering, es la tarea de agrupar objetos por similitud, en conjuntos de manera que los miembros del mismo grupo tengan características similares.

Es la tarea principal de la minería de datos exploratoria y es una técnica común en el análisis de datos estadísticos. Se puede realizar a travez del aprendizaje automatico No Supervisado y el Supervisado.

En primer instancia es importante mencionar que para esta sección del análisis, las variables independientes ya fueron filtradas por medio de la selección de caracteristicas precedente, por lo que solo se tomarán en cuenta aquellas que aportan mayor información al modelo: Heart Rate, Duration, Body Temp y Height

A. Método jerarquico

Los datos son escalados, debido a la diferencia de unidades entre las variables. Para esto se hace uso de la libreria sklearn.preprocessing importando StandardScaler.

- Definir cantidad de clústers

  Se realiza el gráfico de dendrograma, el cual de acuerdo a su estructura muestra los datos en subcategorías que se van dividiendo en otros hasta llegar al nivel de detalle deseado.

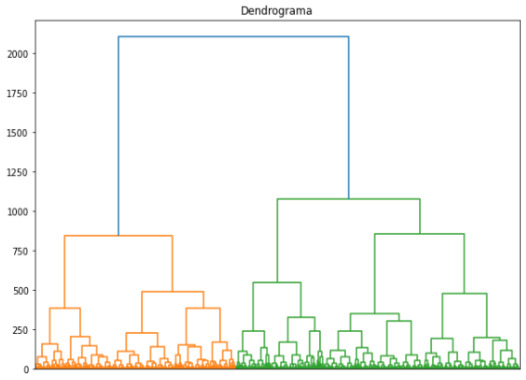

Figura 4: Gráfico Dendrograma

El gráfico muestra la creación de 3 subgrupos de los que se desprenden otros nuevos. Por lo que esta es la cantidad elegida de clusters.

- Visualización de clústers

  Para esta sección se hace uso de la libreria sklearn.cluster importando la AgglomerativeClustering.

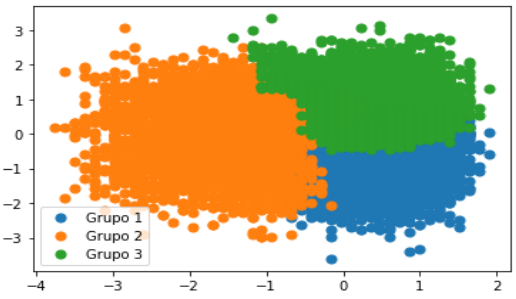

Figura 5: Visualización de clústers

- Evaluación de características de los clústers

  En el punto anterior, se visualizo la unión de los clusters en el plano 2D. Posterior a esto, se evaluan las caracterisiticas reelevantes de los grupos. Esto nos servirá para entender las comunalidades de estos modelos. El primer paso es asignar a cada observación su respectivo cluster. Posteriormente se crea un nuevo conjunto de datos que contenga solamente la variable de interés y la etiqueta de cluster. Y por último se selecciono un diagrama de boxplot, es una herramienta muy utilizada para la evalación de la estadistica descriptiva.

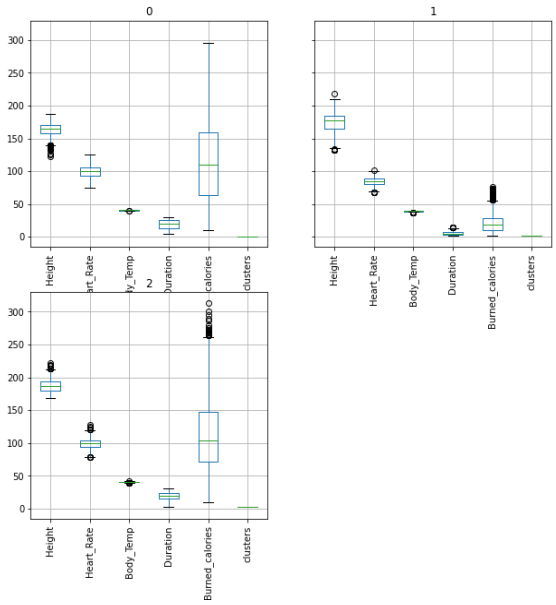

Figura 6: Boxplot de grupos (Método Jerárquico)

Tabla 2: Estadistica descriptiva de los clústes

Cluster	Media	Desv Std
1	93.5123	0.265618
2	36.8728	1.050691
3	145.9825	0.2963

Al observar los grupos, se prueba lo visto en información mutua, la disminución de Duration hace que Burned Calories dismunuya a la par.

- Desempeño de los modelos o clústers

Previamente se han generado 3 clústers, para conocer el desempeño del modelo, se pone en evaluación la siguiente pregunta: ¿Cuál es el clúster que tiene un mejor desempeño?. Para resolver esto se realiza un diseño de experimentos.

Prueba de Hipótesis
Se prueba la siguiente inferencia, importando f_.oneway de la libreria scipy.stats.:
H0= No existe diferencia signficativa entre los grupos
HA= Existe diferencia signficativa entre los grupos

Tabla 3: Resultado prueba de hipotesis

Estadistico	Valor
Valor F:	6300.20
Valor p:	0.001

Por lo tanto con 95 %. de confianza se rechaza la hipotesis nula. Por lo que se asume una diferencia signficativa entre los grupos.

Media de Error Absoluto
Se hace uso de la libreria sklearn en la sección de metrics.

Tabla 4: Resultado Media de Error

Cluster	MAE
1	47.7409
2	20.9387
3	39.8526

B. Método RGBBoost Regression
Este es un algoritmo de aprendizaje automático supervisado que utiliza una técnica de ensamblado de árboles de decisión para mejorar la precisión de la predicción.

Se importa la libreria xgboost y train test split de la libreria sklear modelselection.

Los datos son escalados nuevamente por su diferencia de unidades, utilizando las mismas herramientas que en el método jerarquico.

Se crean las variables para X y Y de prueba y entrenamiento. En este método se utilizan los errores residuales del modelo inicial para entrenar un segundo modelo. El segundo modelo se enfoca en corregir los errores del primer modelo. Se repite el proceso para cada modelo subsiguiente, utilizando los errores residuales del modelo anterior para entrenar el siguiente modelo.

- Desempeño del modelo.
Para conocer el desempeño o conocer la exactitud de predicción brindado por el modelo:

Media de Error Absoluto
El resultado de MAE obtenido da 0.026

KFold
Despues se evaluo la eficacia de un modelo, atravéz de otra métrica. En este caso utilizamos Kfold, para uso se importo de la libreria sklearn.modelselection. En esta técnica, se divide el conjunto de datos en k subconjuntos o ”folds”de aproximadamente el mismo tamaño. Posteriormente, el modelo se entrena k veces, cada vez utilizando k-1 subconjuntos para entrenamiento y el subconjunto restante para validación. El resultado de KFold obtenido da un score de 0.96.
Por ultimo se grafico el comportamiento de la variable dependiente de prueba y la de predicción. La visualización se hace en un plot, la función se importo de la libreria matplotlib.pyplot.

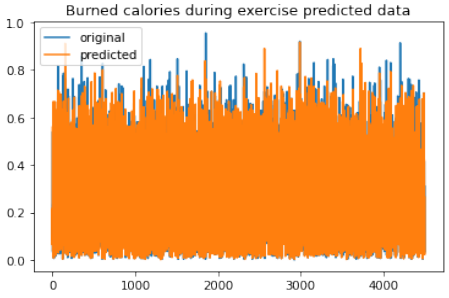

Figura 7: Visualización gráfica del desempeño de la predicción)

4. Conclusiones

Se sabe en base a la selección de caracteristicas realizada en base al análisis descriptivo que existen 3 variables que van a influir significativamente, en comparación con las demás, en la creación de un modelo de mejor desempeño estas son: Body Temp, Duration y Heart Rate.

Duration y Heart Rate demostraron tener una interacción con Burned Calories, durante la evaluación de grupos realizada en el análisis descriptivo en el método jerarquico, se observaba que los grupos con una media superior, tenian un incremento considerable de estas variables. Por lo que se concluye que al combinar estas variables el ejercicio realizado resulta más efectivo. Las variable Body Temp se mostro bastante estable, pero demostro en el análisis de correlación estar muy relacionada con Heart Rate, por lo que se asume van de la mano

Para un modelo con mejor desempeño se recomienda XGBOOST sobre el método jerárquico, ya que probo mejor resultados en MAE y KFold.

Agradecimientos

Este trabajo ha sido realizado con el apoyo de los maestros de la Facultad de Ciencias Fisicomateticas de la UANL.

5. Apendice

[1] Tareas en Google Colab.md GitHub. https://github.com/nalrob/Aprendizaje-Automatico/blob /main/Tareas+en+Google+Colab.md

6. Referencias

[1] Información extraída de Kaggle, creada con fines educativos por Eduardo M. De Mories https://www.kaggle.com/datasets/aadhavvignesh/calories-burned-during-exercise-and-activities

[2] Vinoy Binumon Joseph, S. (2022). Calorie Burn Prediction Analysis Using XGBoost Regressor and Linear Regression Algorithms. Proceedings of the National Conference on Emerging Computer Applications, 4, 5.

[3] Poellabauer, S. V. A. (15/Julio/2019). Multi-modal Biometric-based Implicit Authentication of Wearable Device Users. 1, 3.

[4] Learning, M. (1994). Neural and Statistical Classification. Editors D. Mitchie et. al, 350.

[5] Mitchell, T. M. (1999). Machine learning and data mining. Communications of the ACM,42(11), 30-36.

[6] Downey, A. B. (2011). Think stats. .ºReilly Media, Inc."