

Uso de algoritmos de Aprendizaje Supervisado para la Clasificación del Nivel de Actividad Física en Escolares de la Provincia de Arequipa

Juan J. Gallegos Valdivia, Bach¹, Gonzalo R. Luna Luza, Bach¹, José Sullá-Torres, Dr¹, Rossana Gomez-Campos, Dr², and Marco Cossio-Bolaños, Dr²

¹Universidad Nacional de San Agustín de Arequipa, Perú, jgallegosv@unsa.edu.pe, glunaluza@gmail.com, jsulla@unsa.edu.pe,

²Universidad Católica del Maule, Chile, rossauicamp@gmail.com, mcossio30@hotmail.com

Abstract– *Sedentary lifestyle is one of the main causes of numerous health problems. Today, at school level, the physical activity necessary for their age is not practiced. The objective of this study was to evaluate by applying a validated questionnaire to 1,336 schoolchildren from 11 to 18 years old to classify the level of physical activity. To obtain the results, the KDD methodology used in data mining was used for data processing and transformation, along with the algorithms: CART decision trees and Random Forest. The precision of the algorithms was compared. At the end of the comparison it was determined that the Random Forest algorithm was the most accurate in the classification of Physical Activity of schoolchildren.*

Keywords- *Supervised Learning, Decision Tree, Random Forest, Classification, Physical Activity.*

Resumen– *El sedentarismo es una de las principales causas de numerosos problemas de salud. Hoy en día a nivel escolar no se practica la actividad física necesaria para su edad. El objetivo de este estudio fue evaluar mediante la aplicación de un cuestionario validado a 1336 escolares de 11 a 18 años para clasificar el nivel de actividad física. Para obtener los resultados, se usó la metodología KDD utilizada en la minería de datos para el procesamiento y transformación de los datos, junto con los algoritmos: árboles de decisión CART y Random Forest. Se comparó la precisión de los algoritmos. Al terminar la comparación se determinó que el algoritmo Random Forest fue el más preciso en la clasificación de Actividad Física de los escolares.*

Palabras clave- *Aprendizaje Supervisado, árbol de decisión, Random Forest, Clasificación, Actividad Física.*

I. INTRODUCCIÓN

Según la Organización Mundial de la Salud (2018), la inactividad física genera alrededor del 6% de las muertes registradas en todo el mundo, representando el cuarto factor de riesgo en lo que respecta a la mortalidad mundial. Asimismo, se estima que es la causa principal de entre un 21% y 25% de los cánceres de mama y de colon, el 27% de los casos de diabetes y aproximadamente del 30% de la carga de cardiopatía isquémica [1].

La Actividad Física (AF), definida como cualquier movimiento corporal producido por los músculos esqueléticos que exija gasto de energía, representa una herramienta de vital importancia en la lucha contra las enfermedades no transmisibles. A pesar de tener beneficios la práctica de la Actividad Física, los niveles de AF están disminuyendo significativamente en diversas partes del mundo [2].

En general, el problema principal es que en los colegios del Perú la actividad física no está muy tomada en cuenta ya que lo consideran como pérdida de tiempo y así como hay estudiantes que son buenos para una materia también lo son para algún deporte en específico, Este artículo tiene la función de dar a conocer el nivel de actividad física en los estudiantes de Arequipa; gracias al ejercicio, los escolares mejoran el estado de ánimo, disminuyen la depresión y la ansiedad, elevan la autoestima y la imagen corporal, ofrece oportunidades de distracción e interacción social, y ayuda a mejorar el afrontamiento al estrés de la vida cotidiana [3].

Investigaciones recientes han demostrado que las técnicas de aprendizaje automático pueden predecir con precisión las clases de actividad a partir de los datos de estudio [4].

En ese sentido, algunos estudios sugieren el uso de técnicas de aprendizaje automático como Árboles de Decisión y Random Forest para realizarla clasificación de la Actividad Física de una mejor manera [5].

Considerando todo esto, existe una necesidad apremiante de medir los niveles de actividad física en individuos con riesgo de desarrollar obesidad lo antes posible.

Es en este escenario que el presente trabajo tiene como objetivo clasificar el nivel de actividad física en los escolares, de Centros Educativos Estatales, tomando como caso de estudio la ciudad de Arequipa, Perú. Para ello, se va a comparar los resultados de los algoritmos en evaluación y seleccionar el más preciso para la clasificación de la AF en escolares.

Este artículo está organizado de la siguiente manera: en la sección II describe los trabajos relacionados, la sección III presenta los materiales y métodos, donde se explica el método a utilizar. En la sección IV se reportan los resultados obtenidos luego de la experimentación, finalmente en la sección V se presenta la conclusión con el grado de cumplimiento del objetivo de este trabajo.

II. TRABAJOS RELACIONADOS

En esta sección se revisan los trabajos relacionados a la actividad física y su relación con los algoritmos supervisados, ya que es un tema de sumo interés [6].

En [1] se muestra la influencia de diferentes intensidades de actividad física sobre indicadores que miden el estado de salud en el Perú, donde trata de establecer los determinantes de la participación en actividades físicas a partir de la estructura

económica presentada en Cawley y un modelo de elección de participación propuesto por Humprheys y Ruseski.

Otro artículo efectuado en Perú [7], verificó la validez y confiabilidad a través de un cuestionario que mide el nivel de actividad física que fue validada por medio de análisis factorial confirmatorio y la fiabilidad por medio de consistencia interna, esto se desarrolla a través de percentiles para clasificar estos niveles de AF de escolares según sexo y edad. Los puntos de corte propuestos fueron: Bajo (<p15), moderado (p15-p85) y alto nivel de AF (>p85). Se ha verificado que el cuestionario de AF utilizado es válido y confiable, además los percentiles desarrollados son una herramienta útil para diagnosticar, clasificar y monitorear los niveles de AF.

Los algoritmos del Aprendizaje Automático (AA) sirven para poder clasificar una gran cantidad de datos y predecir cierto comportamiento que han adaptado, estos modelos, a las diferentes áreas de aplicación [8].

En ese sentido, el AA se ha utilizado para la clasificación de la Actividad Física en Escolares. En [9] se compara diversos algoritmos de aprendizaje automático para la predicción de la actividad física, concluyendo con la alta precisión obtenida del algoritmo *Random Forest*, por lo que el uso del aprendizaje automático podría convertirse en un activo invaluable en el proceso de entrenamiento personalizado automatizado. Los algoritmos individualizados permiten predecir la actividad física durante el día y brindan la posibilidad de intervenir a tiempo.

En [10], diseñaron modelos de predicción de frecuencia cardíaca basados en actividad física, para la predicción, para lo cual utilizaron la actividad física en el paso de tiempo inicial y señales de frecuencia cardíaca en una larga duración de predicción. Para afianzar la predicción, hicieron uso de una red neuronal evolutiva, los resultados que obtuvieron demuestran que la frecuencia cardíaca obtenida se asemeja a la frecuencia cardíaca real. Para esto utilizó técnica de Variables Instrumentales (VI) y el modelo de probabilidad Probit. Los resultados que obtuvo evidenciaron que existe una influencia positiva de la práctica de actividad física sobre el estado de salud, medido a través de la glucosa y el Índice de Masa Corporal (IMC).

III. MATERIALES Y MÉTODOS

Para el desarrollo de esta investigación se utilizó la metodología KDD (*Knowledge Discovery in Databases*), la cual es utilizada en el desarrollo de proyectos de Minería de Datos [11] y el algoritmo de clasificación *Random Forest* [9] para clasificar el nivel de Actividad Física en escolares.

A. KDD

El análisis de los datos es una tarea difícil que requiere el uso de enfoques específicos. La metodología KDD (*Knowledge Discovery in Databases*). En la Fig. 1 se muestra los pasos para poder descubrir conocimiento dentro de un área estudiada. Primero se identifica los patrones válidos novedosos y

potencialmente útiles. Esta primera fase del KDD es la fase de integración y recopilación en donde se unifican todos los datos dándoles un formato común para una mejor visualización de los datos, en la segunda fase se verifica la calidad de los datos donde se seleccionan y preparan datos para su posterior clasificación ya que no todos los datos son necesarios, en la tercera fase se utiliza el algoritmo *Random Forest*, en la cuarta etapa se aplica la técnica elegida en los datos preprocesados, por último en la quinta etapa se evalúa los resultados obtenidos.

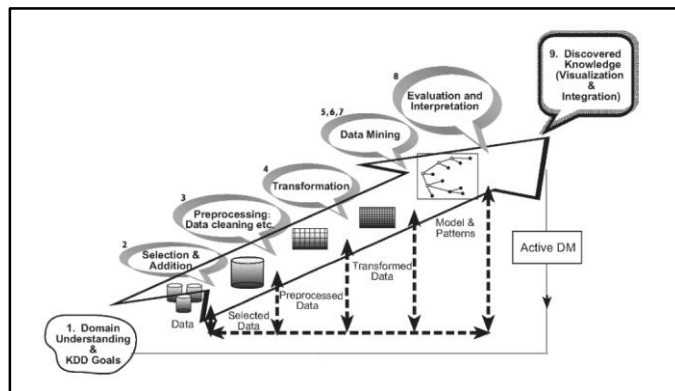


Fig. 1 Fases del KDD [12].

B. Árboles de Decisión

Los árboles de decisión son algoritmos que permiten la representación de un modelo de estudio. Un árbol de decisión consta de dos tipos de nodos; nodos no terminales y nodos terminales. Un nodo no terminal denota una condición basada en una característica. Esta condición divide los registros de datos en función de sus características. El nodo terminal, también llamado nodo hoja, no tiene una condición porque se usa para tomar una decisión. Cada rama del árbol de decisión representa el resultado de las condiciones, mientras que el nodo hoja contiene una etiqueta de clasificación. Según el número de ramas, el árbol de decisión puede ser un árbol binario o de varias ramas [13].

Este modelo de representación ofrece una mejor comprensión de los perfiles de actividad física de las unidades de estudio.

Entre sus usos se encuentran: [14]

- Selección de variables.
- Evaluar la importancia relativa de las variables.
- Manejo de valores faltantes.
- Predicción.
- Manipulación de datos.

Existen diferentes tipos de árboles de decisión, de acuerdo a la métrica de entrada:

- CART: Índice de Gini, Criterios de dosificación
- C4.5: Entropía
- CHAID: Chi-square
- QUEST: Chi-square para variables categóricas; ANOVA J-way para variables continuas

C. Random Forest

El algoritmo *Random Forest* es una colección de árboles de decisión aleatorios. Es utilizado para la clasificación o regresión de datos, combinan las predicciones de un conjunto de árboles para llegar a una respuesta. El error está dado por la precisión y correlación entre los árboles individuales. La mejora en un árbol provoca la disminución del error general [15].

Entre sus características se encuentran:

- Estimación de variables con mayor importancia.
- No sobreentrena al incrementar la cantidad de árboles.
- Robustez ante ruido y anomalías.

En la Fig. 2 se aprecia la comparación con otros algoritmos donde se ve que los arboles de decisión y Random Forest muestran mejores resultados en cuanto a la exactitud (accuracy) y F1-score frente al porcentaje que alcanza en el eje X.

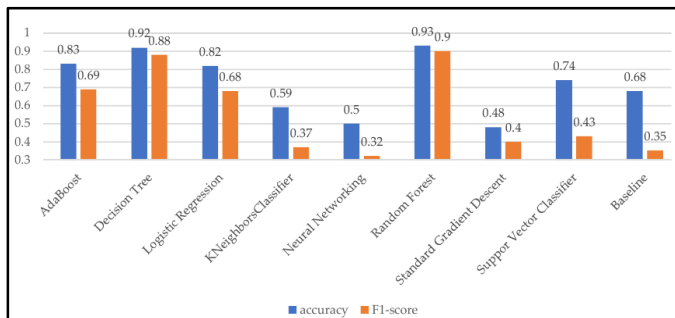


Fig. 2 Resultados comparativos de los Algoritmos utilizados [9].

D. Base de datos

Para esta investigación se cuenta con una base de datos 1341 escolares de 6 Centros Educativos Públicos (condición socioeconómica media) de la Provincia de Arequipa, Perú. Esta ciudad está ubicada a 2320 msnm., el rango de edad es de 11 a 18 años. Los escolares fueron seleccionados de forma probabilística (estratificada).

Para medir el nivel de actividad física se utilizó la técnica de la Encuesta. El instrumento utilizado fue el cuestionario propuesto por Gómez-Campos, Vilcazán, Arruda, Hespagnol & Cossio-Bolaños [16]. La encuesta consta de 11 preguntas que fue validada en el estudio. En la Tabla I, se muestra la estructura del cuestionario utilizado.

TABLA I
ESTRUCTURA DEL CUESTIONARIO A ESCOLARES QUE SERVIRÁ PARA DETERMINAR EL NIVEL DE ACTIVIDAD FÍSICA

Cod.	Nombre	Descripción	Unidad de Medida
AF1	Pregunta 1	¿Realizas algún tipo de actividad física durante la semana?	numérico
AF2	Pregunta 2	¿Cómo acostumbras a ocupar tu tiempo libre? (fuera del horario escolar)	numérico
AF3	Pregunta 3	¿Realizas alguna actividad física durante el receso (recreo escolar)?	numérico
AF4	Pregunta 4	¿En los meses de vacaciones ¿realizas actividad física?	numérico
AF5	Pregunta 5	¿Cuántos días a la semana realizas actividad física?	numérico

AF6	Pregunta 6	En los días que realices actividad física ¿cuánto tiempo (minutos) total ocupas por día?	numérico
AF7	Pregunta 7	¿Cuántos días/semana realizas actividad física vigorosa?	numérico
AF8	Pregunta 8	¿Cuántos minutos/día realizas actividad física vigorosa?	numérico
AF9	Pregunta 9	¿Cuántos días/semana realizas actividad física moderada?	numérico
AF10	Pregunta 10	¿Cuántos minutos/día realizas actividad física moderada?	numérico
AF11	Pregunta 11	¿Acostumbas ir de la casa al colegio en?	numérico
ZG	Zona Geográfica	Lugar donde reside	Numérico
GI	Grado de Instrucción	Etapas escolares del escolar	Numérico

Para la carga, extracción y transformación (ETL) se utilizó la herramienta KNIME, que es uno de los software más utilizados para realizar las operaciones de ETL y encontrar la mejor precisión en los resultados.

Una vez procesada los datos, se utilizó el lenguaje de programación Python, el cual permite trabajar con técnicas de aprendizaje supervisado para realizar la clasificación.

IV. EXPERIMENTACIÓN Y RESULTADOS

Las características de la muestra estudiada se presentan en la Tabla II

TABLA II
CARACTERÍSTICAS DE LA MUESTRA ESTUDIADA [7]

Edades	Tipo AF		Frecuencia AF		Duración AF		Intensidad AF		AF (total)		
	n	X	DE	X	DE	X	DE	X	DE	X	DE
Hombres											
11,0-11,9	28	10,8*	2,9	5,7	1,3	9,1*	8,0	6,6	1,4	32,3*	9,1
12,0-12,9	86	10,3	2,9	5,3	1,4	6,9	2,6	6,2	1,7	28,7*	6,3
13,0-13,9	40	9,7	3,1	5,4	1,7	7,6	2,7	7,0	1,8	29,7*	6,8
14,0-14,9	97	9,7	2,6	5,2	1,4	7,4	4,7	6,6	1,8	29,0*	7,2
15,0-15,9	135	9,8*	2,9	5,4	1,5	7,7*	2,6	7,0	1,6	29,9*	6,7
16,0-16,9	129	9,9*	3,0	5,2	1,6	7,8*	2,8	6,9	1,8	29,9*	7,2
17,0-17,9	82	9,6*	3,1	5,2	1,6	7,5*	3,0	6,8	1,9	29,1*	7,8
18,0-18,9	64	10,0*	2,9	5,1	1,6	7,5*	2,7	6,8	1,5	29,4*	6,8
Total	661	9,9*	2,9	5,3	1,5	7,6*	3,5	6,8	1,7	29,5*	7,1
Mujeres											
11,0-11,9	22	8,8	2,4	6,9	1,6	6,8	2,8	6,1	1,4	28,6	9,6
12,0-12,9	69	9,3	2,8	5,1	1,3	6,6	2,6	6,1	1,8	27,2	6,6
13,0-13,9	26	9,8	3,0	5,4	1,6	7,2	2,7	6,2	1,6	28,7	6,3
14,0-14,9	88	8,3	2,7	5,0	1,5	6,3	2,7	6,4	1,6	26,1	6,1
15,0-15,9	173	8,2	4,2	4,8	1,4	6,4	2,8	6,2	1,6	25,7	7,0
16,0-16,9	166	8,1	2,6	4,6	1,4	6,1	2,8	6,1	1,5	24,9	6,4
17,0-17,9	82	8,1	2,2	4,5	1,3	6,0	2,3	5,9	1,3	24,6	4,9
18,0-18,9	52	7,9	2,6	4,2	1,2	5,9	2,5	6,2	1,7	24,2	6,2
Total	678	8,3	3,1	4,8	2,1	6,3	2,7	6,2	1,5	25,7	6,6

Leyenda: X: Promedio, DE: Desviación estándar, AF: Actividad Física, *: Diferencia significativa ($p < 0.05$)

A. Pre-procesamiento y Limpieza de Datos

Para tener unos datos bien estructurados y con la que se pueda trabajar de forma correcta es necesario verificarlos, para tomar las decisiones que regulen ese problema.

En algunas aplicaciones es común llenar los datos faltantes utilizando un criterio matemático, en esta investigación se optó por eliminar aquellos registros donde algunos datos de interés (como el llenado de las preguntas) se encuentran ausentes, debido a que calcular estos valores empleando criterios

matemáticos no sería correcto ya que estamos hablando de cuestiones de salud donde las aproximaciones podrían afectar los resultados.

B. Clasificación de la Actividad Física

Luego de Procesar los datos se aplicaron los algoritmos seleccionado: árbol de decisión CART y *Random Forest*, se utilizó la herramienta *Spyder* con el lenguaje de programación *Python*. Para facilitar los niveles de clasificación de actividad física, se utilizó la siguiente codificación: 0: Baja, 1: Media y 2: Alta.

En la Fig. 3 se muestra los 3 niveles de actividad física encontrados en los alumnos de la región de Arequipa., donde se aprecia que la mayoría de escolares se encuentran en el nivel de clasificación Media (eje X) en relación a percentil de Actividad Física.

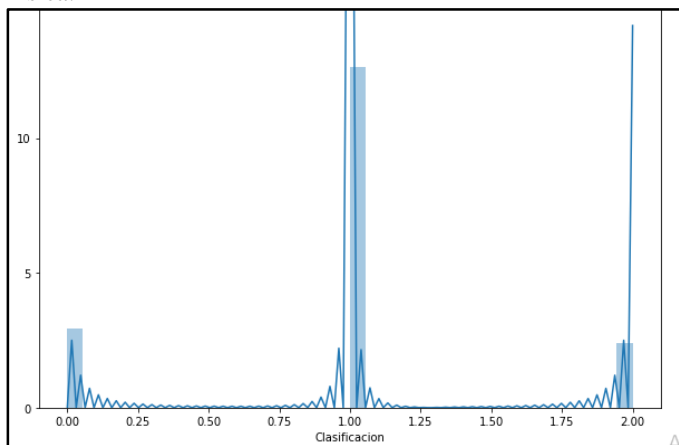


Fig. 3 Niveles de Actividad Física. Resultado en la herramienta Spyder

1) *Random Forest*.

La precisión alcanzada por el algoritmo *Random Forest* fue del 93%, también muestra la matriz de confusión como se aprecia en la Tabla III, donde las filas representan los valores predictivos y las columnas los valores reales.

TABLA III
MATRIZ DE CONFUSIÓN DE RANDOM FOREST

	0	1	2
0	16	3	0
1	0	98	3
2	0	3	11

Al algoritmo *Random Forest* se cargó como entrada un vector de valores para predecir el nivel de actividad física en que se encuentra un nuevo estudiante, dando como resultado que se encuentra en un nivel de actividad física media como se muestra en la Fig. 4.

Nombre	Tipo	Tamaño	Valor
X	DataFrame	(1336, 10)	Column names: AF_2, AF_3, AF_4, AF_5, AF_6, AF_7, AF_8, AF_9, AF_10, A ...
acc	float64	1	0.9328358208955224
ceros	int32	1	219
cm	int64	(3, 3)	[[16 3 0] [0 98 3]]
comp	DataFrame	(134, 2)	Column names: real, preds
df	DataFrame	(1336, 12)	Column names: AF_1, AF_2, AF_3, AF_4, AF_5, AF_6, AF_7, AF_8, AF_9, AF ...
dos	int32	1	179
new_prediction	int64	(1,)	[1]
unos	int32	1	938
x_test	DataFrame	(134, 10)	Column names: AF_2, AF_3, AF_4, AF_5, AF_6, AF_7, AF_8, AF_9, AF_10, A ...
x_train	DataFrame	(1202, 10)	Column names: AF_2, AF_3, AF_4, AF_5, AF_6, AF_7, AF_8, AF_9, AF_10, A ...
y	Series	(1336,)	Series object of pandas.core.series module
y_preds	int64	(134,)	[0 1 1 ... 1 1 1]

pExactitud	float64	1	0.9328358208955224
precision	float64	1	0.9093406593406593
puntaje	float64	1	0.8853658536585365
sensibilidad	float64	1	0.8660388595250502

Fig. 4. Resultado de la predicción realizado con Random Forest.

2) *Árbol de decisión CART*.

Para la evaluación del algoritmo de árbol de decisión CART se utilizó los mismos datos dividido en un 20% para entrenamiento y el 80% para la evaluación, obteniendo la matriz de confusión en la Tabla IV, donde las filas representan los valores predictivos y las columnas los valores reales.

TABLA IV
MATRIZ DE CONFUSIÓN DE CART

	0	1	2
0	42	11	0
1	11	272	10
2	0	27	28

Para validar las predicciones se utilizaron las siguientes métricas de evaluación:

1. *Precisión*: Es la proporción de las predicciones correctas, se representa como:

$$Precisión = \frac{VP}{VP + FP} \quad (1)$$

2. *Sensibilidad*: Es la fracción de verdaderos positivos, se representa como:

$$Sensibilidad = \frac{VP}{VP + FN} \quad (2)$$

3. *F1 score*: Combina las medidas de precisión y sensibilidad en un solo valor, se representa como:

$$F1 = 2 * \frac{Precisión * Sensibilidad}{Precisión + Sensibilidad} \quad (3)$$

Donde,

- VP: Verdaderos Positivos
- FP: Falsos Positivos
- FN: Falsos Negativos

La Tabla V muestra los resultados de la comparación de los algoritmos Random Forest y CART.

TABLA V
COMPARACIÓN DE LAS MÉTRICAS DE EVALUACIÓN

	RandomForest	CART
Precisión	0.90934065934	0.8469081964
Sensibilidad	0.86603885952	0.8528678304
F1 score	0.88536585366	0.8465086148

Entonces los pasos para encontrar el nivel de Actividad Física de los escolares eran los siguientes:

1. Primero se tenía que llenar el cuestionario por parte de los escolares.
2. Con los resultados de los cuestionarios se tenía que escoger las preguntas correspondientes a las categorías de actividad física.
3. Se tenía que realizar una suma de cada pregunta para su categoría correspondiente.
4. Se tenía que evaluar las categorías para encontrar el Nivel físico de los Alumnos.

Ahora los pasos se reducen a:

1. Primero llenar el cuestionario por parte de los alumnos.
2. Aplicar el modelo hallado para predecir su nivel de Actividad Física

V. CONCLUSIONES

Es posible obtener un modelo predictivo que ayude a agilizar el proceso de evaluación de los niveles de actividad física en los alumnos de la región de Arequipa, El modelo más idóneo para este tipo de problemas es el algoritmo de aprendizaje supervisado *Random Forest* por su porcentaje alto de precisión en los datos, según los resultados obtenidos.

Para la validación se usaron métricas de evaluación en donde los resultados obtenidos fueron altos, la relación entre las predicciones correctas y el número total de predicciones es la métrica de la exactitud en donde se obtuvo un 93%, la siguiente métrica mide la precisión del clasificador cuyo resultado fue de un 90% y en la métrica de sensibilidad obteniendo un 87%.

Por lo que el uso de técnicas de aprendizaje supervisado son una buena alternativa para los interesados en determinar la clasificación de niveles de Actividad Física en escolares.

Se concluye que el uso de algoritmos de aprendizaje supervisado ayuda a la clasificación del nivel de actividad física en escolares de la provincia de Arequipa.

AGRADECIMIENTOS

A la Universidad Nacional de San Agustín de Arequipa, quien ha financiado el proyecto «Propuesta normativa para valorar los niveles de Actividad Física de los escolares de la

provincia de Arequipa», con numero de contrato 15-2016-UNSA.

REFERENCIAS

- [1] K. Cansino and H. Gálvez, "Determinantes de la participación en actividades físicas en el Perú," *Rev. Peru. Med. Exp. Salud Publica*, 2014.
- [2] D. Husárová, Z. D. Veselská, D. Sigmundová, and A. Madarasová Gecková, "Age and gender differences in prevalence of screen based behaviour, physical activity and health complaints among Slovak school-aged children," *Cent. Eur. J. Public Health*, 2015.
- [3] O. Cintra and Y. Balboa, "La actividad física: un aporte para la salud," *Lect. Educ. y Deport. Rev. Digit.*, 2011.
- [4] M. Hagenbuchner, D. P. Cliff, S. G. Trost, N. Van Tuc, and G. E. Peoples, "Prediction of activity type in preschool children using machine learning techniques," *J. Sci. Med. Sport*, 2015.
- [5] A. H. Al-Fatlawi, H. K. Fatlawi, and S. H. Ling, "Recognition physical activities with optimal number of wearable sensors using data mining algorithms and deep belief network," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 2017.
- [6] M. Jones, E. Defever, A. Letsinger, J. Steele, and K. A. Mackintosh, "A mixed studies systematic review and meta-analysis of school-based interventions to promote physical activity and/or reduce sedentary time in children," *J. Sport Heal. Sci.*, 2019.
- [7] M. A. Cossio-Bolanos, R. Vidal-Espinoza, J. Sulla-Torres, C. Luarte-Rocha, J. Pacheco-Carrillo, and R. Gomez-Campos, "Validity, reliability, and percentiles of a questionnaire measuring activity levels in adolescents living at moderate altitude in Peru," *Retos*, 2019.
- [8] M. Allamanis, E. T. Barr, P. Devanbu, and C. Sutton, "A survey of machine learning for big code and naturalness," *ACM Computing Surveys*. 2018.
- [9] T. B. Dijkhuis, F. J. Blaauw, M. W. van Ittersum, H. Velthuisen, and M. Aiello, "Personalized physical activity coaching: A machine learning approach," *Sensors (Switzerland)*, 2018.
- [10] F. Xiao, Y. M. Chen, M. Yuchi, M. Y. Ding, and J. Jo, "Heart rate prediction model based on physical activities using evolutionary neural network," in *Proceedings - 4th International Conference on Genetic and Evolutionary Computing, ICGEC 2010*, 2010.
- [11] S. Aljawarneh, A. Anguera, J. W. Atwood, J. A. Lara, and D. Lizcano, "Particularities of data mining in medicine: lessons learned from patient medical time series data analysis," *Eurasip J. Wirel. Commun. Netw.*, 2019.
- [12] O. Maimon and L. Rokach, "Introduction to Knowledge Discovery and Data Mining," in *Data Mining and Knowledge Discovery Handbook*, 2009.
- [13] M. K. Jiawei Han and J. Pei, "Data Mining: Concepts and Techniques: Concepts and Techniques." Elsevier, Amsterdam, 2011.
- [14] Y. Y. Song and Y. Lu, "Decision tree methods: applications for classification and prediction," *Shanghai Arch. Psychiatry*, 2015.
- [15] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
- [16] R. Gómez Campos, É. Vilcazán, M. De Arruda, J. E. Hespagnol, and M. A. Cossio-Bolaños, "Validación de un cuestionario para la valoración de la actividad física en escolares adolescentes Validation of a questionnaire assessing school physical activity in adolescents," *An. la Fac. Med.*, 2013.