Revista Innovación y Software
Vol. 3, No. 2, Mes Septiembre-Febrero, 2022
ISSN: 2708-0935
Pág. 99-108
https://revistas.ulasalle.edu.pe/innosoft

Tipo de artículo: Artículos originales
Temática: Inteligencia artificial
Recibido: 20/06/2022 | Aceptado: 31/07/2022 | Publicado: 30/09/2022

Identificadores persistentes:
ARK: ark:/42411/s9/a71
PURL: 42411/s9/a71

Predicción del nivel de obesidad en personas usando el modelo de árbol de decisión

Prediction of the level of obesity in people using the decision tree model

Renato Eduardo Delgado Huacallo [1][0000-0002-6222-5294], **Christian Ilachoque Hanccoccallo** [2][0000-0003-4468-6713], **Felman Luque Sanabria** [3][0000-0001-6322-0228] , **Jose Maykol Paniura Huamani** [4][0000-0002-3031-9317]

[1] Universidad Nacional de San Agustín de Arequipa. Perú. rdelgadoh@unsa.edu.pe
[2] Universidad Nacional de San Agustín de Arequipa. Perú. cilachoque@unsa.edu.pe
[3] Universidad Nacional de San Agustín de Arequipa. Perú. fluques@unsa.edu.pe
[4] Universidad Nacional de San Agustín de Arequipa. Perú. jpaniura@unsa.edu.pe

[*] Autor para correspondencia: rdelgadoh@unsa.edu.pe

Resumen

La obesidad es un problema para la salud pública que afecta a la población mundial es por eso que el presente trabajo está orientado a presentar una solución informática para la estimación y predicción de niveles de obesidad haciendo posible que una persona pueda conocer su estado físico actual, para esto se usó un dataset de personas con obesidad de distintos países como Perú, México y Colombia basándose en sus hábitos alimenticios y su condición física, creando con todos estos datos un árbol de decisión.

Palabras clave: Obesidad, árbol de decisión, nivel de obesidad.

Abstract

Obesity is a public health problem that affects the world population, that is why the present work is oriented to present a computer solution for the estimation and prediction of obesity levels, making it possible for a person to know their current physical condition for this we used a dataset of people with obesity from different countries like Peru, Mexico and Colombia based on their eating habits and their physical condition, creating a decision tree with all these data.

Keywords: *Obesity, decision tree, level of obesity.*

Introducción

Es un hecho en el mundo que el problema de obesidad va en aumento, muchos años atrás está tan solo era un número pequeño dentro de la población mundial, pero este problema en la actualidad va en crecimiento, lo cual puede presentar un gran problema para la salud pública de infantes y adultos, cuyas consecuencias propias de un actuar negligente, representa pérdidas sustanciales, en razón a que la obesidad implica un estado vulnerable hacia otras enfermedades como puede ser el cáncer o problemas cardiovasculares que afectan al corazón [1,2].

Según la Organización Mundial de la Salud (OMS) [3] desde 1975 los problemas de obesidad presentaron un incremento cercano al triple de su media habitual de aquel entonces, trasladándonos al año 2016 que servirá como punto de referencia y del cual destacan los siguientes datos: El 39% de adultos, considera aquellos de 18 años a más tenían sobrepeso y el 13% eran obesos; más de 340 millones de niños y adolescentes, considera aquellos de 5 a 18 años contaban con sobrepeso u obesidad y alrededor de 41 millones de niños menores a los 5 años de edad contaban con sobrepeso u obesidad. Manifestando entonces, la poca o nula preocupación frente al problema, que en determinados casos deviene en consecuencias irreversibles y más aún teniendo en cuenta que el tratamiento para restablecer la salud y estado físico no es de periodicidad inmediata a lo cual este lapso de tiempo será de vital importancia para los individuos pertinentes al caso.

Con el fin de adecuar cuidadosamente la información sobre este tema se ha tomando en cuenta algunos trabajos relacionados, de los que se pueden destacar los siguientes: El primero, hace un estudio en la población española de adultos, donde se evalúa la prevalencia de obesidad considerando mediciones antropométricas individuales, factores sociodemográficos, consumo alimentario, actividad física, estilos de vida y problemas de salud. Con el que se llegó a la conclusión de que existía una prevalencia de obesidad alta en varones con mayor edad, y este factor presenta relación inversa con el nivel socioeconómico. Además de que la probabilidad mínima de obesidad estaba relacionada a las personas con estilos de vida que incluían la actividad física y un sedentarismo moderado.[4]

Como segundo trabajo, también realizado en España, en esta ocasión el estudio fue realizado a una población infantil. En este caso se evalúa la estimación de la prevalencia de sobrepeso en niños de entre 2 y 14 años, tomando en cuenta el Índice de Masa Corporal (IMC) y propiamente dicho la edad y sexo de la población de niños y niñas. De este trabajo se obtuvo que la prevalencia de sobrepeso y obesidad alta era mayor en los varones [5].

Un tercer trabajo, nos describe un proyecto de software realizado en Arequipa en el cual se analiza, diseña e implementa un modelo de minería de datos con datos recolectados de diferentes colegios del Perú para estimar en nivel de obesidad de una persona mediante el IMC, en el cual se hace uso del algoritmo de árboles de decisión para poder predecir un resultado de acuerdo a datos ingresados por el usuario en este caso el estudiante. Se usan varios algoritmos para esta tarea, tales como el J48, Multilayer Perceptrion, ForestPA, NaiveBayes, BayestNet asi obteniendo como resultado el mejor algoritmo que es J48 con una precisión del 94.38% [6].

Como último trabajo relacionado tenemos una investigación en la cual se identifican las variables más influyentes en determinar el grado de obesidad por medio de técnicas de minería de datos, en esta se tiene 16 variables independientes y una variable dependiente la cual se identifica como grado de obesidad, se hace uso del algoritmo J48 y otras técnicas de inteligencia artificial para poder cumplir el objetivo, los resultados de la investigación muestran que las variables independientes más influyentes son el género, estatura, peso y el IMC, así también se obtiene que por el algoritmo J48 se obtiene un éxito superior al 97% mediante validación cruzada [7].

En este trabajo se pretende utilizar un dataset recuperado del repositorio de Machine Learning UCI [8], el cual presenta datos para la estimación de distintos niveles de obesidad en personas de los países de México, Perú y Colombia, basados en sus hábitos alimenticios y condición física. Por todo lo descrito anteriormente, el presente trabajo tiene como objetivo principal presentar una solución informática para la estimación y/o predicción de niveles de obesidad de las personas que deseen conocer su estado físico; haciendo uso de la data set mencionado anteriormente y utilizando como técnica de análisis predictivo un árbol de decisión, el cual está implementado en el lenguaje de programación de Python.

Trabajos relacionados

En el trabajo de Moral et.al. [9], desarrollan una estrategia para evaluar variables que puedan influir en la aparición o evolución de la Diabetes tipo 2. La información almacenada de los pacientes incluye historias clínicas y datos de laboratorio; dicha información se encuentra almacenada en distintas bases de datos. La técnica utilizada es del Random Forest, el cual crea árboles de decisión para la clasificación; concluyendo que su mejor modelo obtenido es un Random Forest con 40 árboles y una profundidad de 5 para cada uno, listando además las variables más importantes para la predicción.

Otro trabajo que se puede rescatar es el de Fierro et.al. [10], en donde realizan un análisis de tres modelos predictivos que están basados en Machine Learning, para obtener la predicción de la tendencia de los jóvenes al alcoholismo. Dentro del dataset se tiene información del estado familiar, lugar de vivienda de un joven, entre otros como variables de entrada. Como resultado de su análisis obtuvieron que el modelo con mayor precisión fue el modelo de Regresión Lineal, quedando por debajo el modelo KNN y el Árbol de decisión.

Materiales y métodos o Metodología computacional

Como se mencionó anteriormente, se recuperaron los datos acerca de la estimación de niveles de obesidad del repositorio de Machine Learning UCI [8], presentando información recolectada de personas de los países de Colombia, México y Perú; este dataset se compone de distintas variables de entrada (como el género, edad, altura, peso, historia familiar con sobrepeso, entre otras) y una variable de salida que representa el nivel de obesidad. A continuación, se hará una breve descripción de las variables de entrada restantes:

- FAVC: Consumo frecuente de alimentos ricos en calorías
- FCVC: Frecuencia de consumo de verduras
- NCP: Número de comidas principales
- CAEC: Consumo de alimentos entre comidas
- SMOKE: Fuma (Si o No)
- CH2O: Consumo de agua diario
- SCC: Seguimiento del consumo de calorías
- FAF: Frecuencia de actividad física
- TUE: Tiempo usando dispositivos tecnológicos
- CALC: Consumo de alcohol
- MTRANS: Transporte usado

Entre las herramientas que se utilizaron, se tiene en primer lugar Google Collab, el cual nos permite ejecutar código de Python en el navegador para distintos usos, tales como IA, análisis de datos entre otros. Este servicio no requiere instalación y nos brinda la posibilidad de acceso a recursos computacionales sin costo.

Revista Innovación y Software
Vol. 3, No. 2, Mes Septiembre-Febrero, 2022
ISSN: 2708-0935
Pág. 99-108
https://revistas.ulasalle.edu.pe/innosoft

Nuestra otra herramienta usada es Scikit-learn es cual es una biblioteca, considerada como la más útil y sólida para lo que se refiere aprendizaje automático en Python, se basa principalmente en Numpy, SciPy y Matplolib. También se ha hecho uso de Pandas, usado para tareas destinadas a ciencias de datos y aprendizaje automático, está construido sobre el paquete Numpy.

Como se indicó en la introducción, el modelo que se planea usar es el árbol de decisión; para ello se necesita tener el dataset en formato CSV y guardarlo en Google Drive para acceder desde el código Python. Antes de realizar el modelo, se necesita hacer una limpieza de datos para detectar datos anómalos, para ello se lee el archivo CSV y se grafican histogramas de los datos que pertenecen a las variables que se consideran necesarias a analizar, en este caso se verifican las variables de edad, altura y peso.

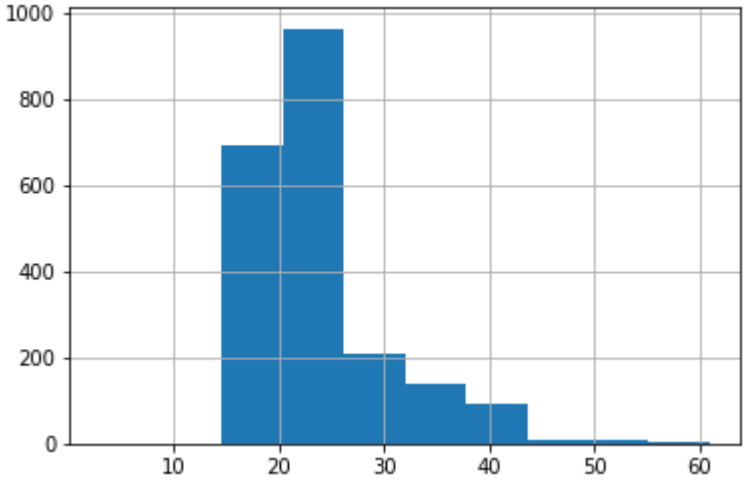

Figura 1. Histograma de Edad

Revista Innovación y Software
Vol. 3, No. 2, Mes Septiembre-Febrero, 2022
ISSN: 2708-0935
Pág. 99-108
https://revistas.ulasalle.edu.pe/innosoft

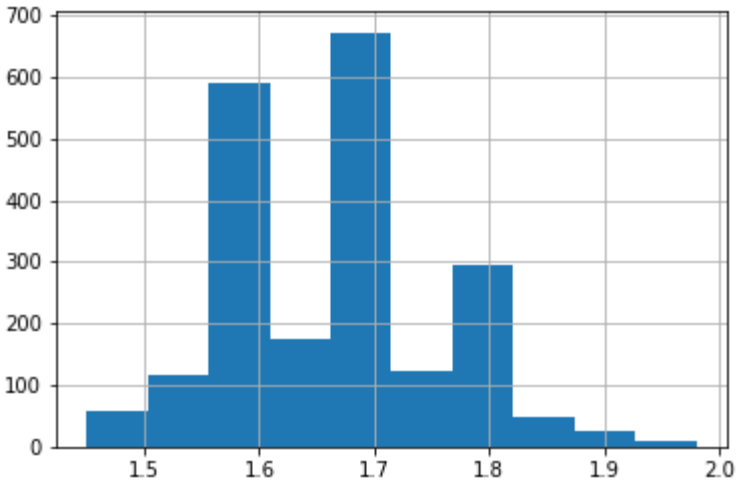

Figura 2. Histograma de Altura

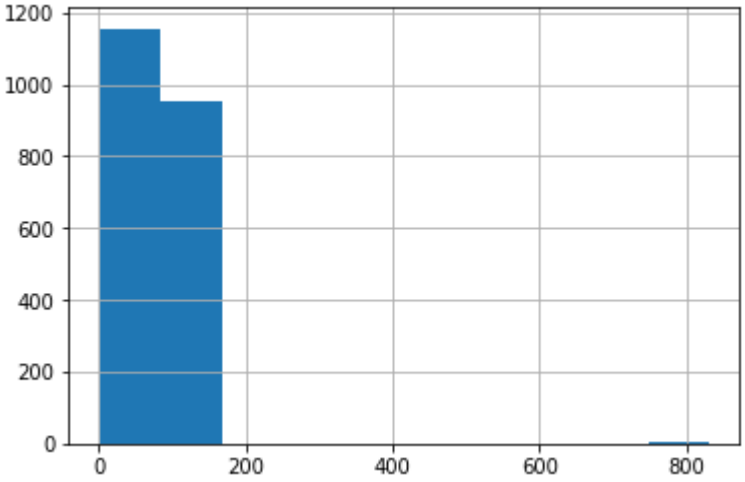

Figura 3. Histograma de Peso

En dichos histogramas se pueden observar algunos datos que están muy alejados al conjunto con mayor cantidad de datos, por ende, se procede a filtrar el dataset, usando el siguiente código:

```
filtered_dataset = dataset[(dataset['Age'] > 15) & (dataset['Age'] < 45) & (dataset['Height']<1.88) & (dataset['Weight']<155)]
```

Figura 4. Código para filtrar el dataset

Revista Innovación y Software
Vol. 3, No. 2, Mes Septiembre-Febrero, 2022
ISSN: 2708-0935
Pág. 99-108
https://revistas.ulasalle.edu.pe/innosoft

Una vez realizada la filtración del dataset, se procede a graficar de nuevo los histogramas, en este caso con todas las variables de entrada del dataset.

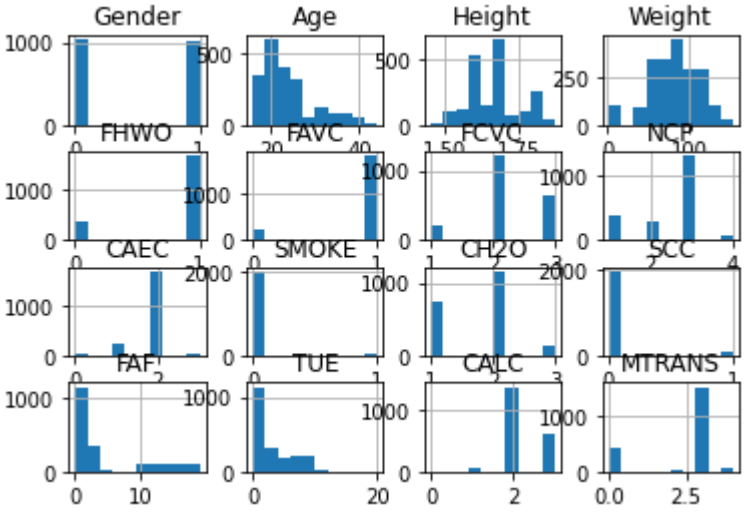

Figura 5. Histogramas después de realizar el filtro

Después se utiliza un 20% de los datos para la realización de pruebas, y el 80% restante de los datos para realizar el entrenamiento. Ahora se procede con la creación del modelo de árbol de decisión, en esta oportunidad se plantea desarrollar un árbol con profundidad de 4. En adición se realiza la validación del entrenamiento usando la matriz de confusión y la métrica de exactitud del modelo; finalmente se procede a crear el árbol de decisión.

Resultados y discusión

En la validación del entrenamiento, primero se realizó la matriz de confusión dando como resultado lo siguiente:

Revista Innovación y Software
Vol. 3, No. 2, Mes Septiembre-Febrero, 2022
ISSN: 2708-0935
Pág. 99-108
https://revistas.ulasalle.edu.pe/innosoft

```
matriz = confusion_matrix(Y_test, Y_pred)
print("Matriz de confusion")
print(matriz)
```

```
Matriz de confusion
[[33 12  0  3  1  0  0]
 [ 3 45  0  0  0  1  2]
 [ 0  0 46  8  0  1 15]
 [ 0  0  1 60  0  0  0]
 [ 0  0  0  0 57  0  0]
 [ 0 15  0  2  1 26 22]
 [ 0  2 14  4  0  6 30]]
```

Figura 6. Matriz de confusión

En cuanto a la exactitud del modelo, se obtuvo el valor de 0,724390243902439, lo cual indica que el modelo no es tan exacto, pero se consideraría aceptable.

```
accuracy_score = accuracy_score(Y_test, Y_pred)
print('Exactitud del modelo:')
print(accuracy_score)
```

```
Exactitud del modelo:
0.724390243902439
```

Figura 7. Resultado exactitud del modelo

Para finalizar, se presenta el árbol de decisión que se generó en Python.

Revista Innovación y Software
Vol. 3, No. 2, Mes Septiembre-Febrero, 2022
ISSN: 2708-0935
Pág. 99-108
https://revistas.ulasalle.edu.pe/innosoft

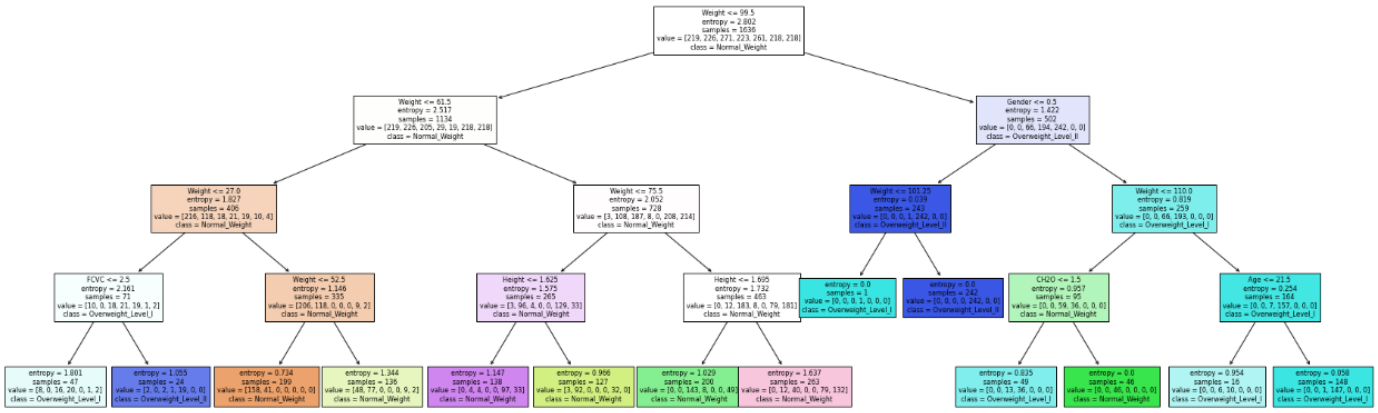

Figura 8. Árbol de decisión generado

Conclusiones

Los resultados obtenidos a lo largo del proyecto en el que se ha utilizado un árbol de decisiones para aplicación práctica a un Dataset sobre nivel de obesidad, resultan en una exactitud de 0.7243 que depende de la cantidad de datos de entrenamiento que se proporciona; aunque no es un valor bastante exacto, es un resultado aceptable.

Para trabajos futuros se tiene pensado implementar y estudiar casos similares para realizar una comparación entre diferentes modelos de Machine Learning para poder evidenciar diferentes perspectivas de solución, así como analizar ventajas y desventajas que se encuentren para cada modelo.

Referencias

[1] M. Malo Serrano, N. Castillo M. y D. Pajita D., "La obesidad en el mundo", Anales de la Facultad de Medicina, vol. 78, n.º 2, p. 67, julio de 2017. Accedido el 21 de junio de 2022. [En línea]. Disponible: https://doi.org/10.15381/anales.v78i2.13213

[2] L. M. T. Garcia, R. F. Hunter, K. Haye, C. D. Economos y A. C. King, "Un marco conceptual orientado a la acción para soluciones sistémicas de prevención de la obesidad infantil en Latinoamérica y en las poblaciones latinas de

Estados Unidos", Obesity Reviews, vol. 22, S5, octubre de 2021. Accedido el 21 de junio de 2022. [En línea]. Disponible: https://doi.org/10.1111/obr.13354

[3] Organización Mundial de la Salud (OMS), Nota descriptiva N°311 junio de 2016. Disponible en: http://www.who.int/mediacentre/factsheets/ fs311/es/

[4] Pérez-Rodrigo, C., Hervás Bárbara, G., Gianzo Citores, M. y Aranceta-Bartrina, J. (2021). Prevalencia de obesidad y factores de riesgo cardiovascular asociados en la población general española: estudio ENPE. Revista Española de Cardiología. https://doi.org/10.1016/j.recesp.2020.12.013

[5] Lasarte-Velillas, J. J., Hernández-Aguilar, M. T., Martínez-Boyero, T., Soria-Cabeza, G., Soria-Ruiz, D., Bastarós-García, J. C., Gil-Hernández, I., Pastor-Arilla, C. y Lasarte-Sanz, I. (2015). Estimación de la prevalencia de sobrepeso y obesidad infantil en un sector sanitario de Zaragoza utilizando diferentes estándares de crecimiento. Anales de Pediatría, 82(3), 152–158. https://doi.org/10.1016/j.anpedi.2014.03.005

[6] M. Ticona, "Sistema Para la Predicción de Obesidad en la Adolescencia Utilizando Técnicas de Minería de Datos", Universidad Católica de Santa María, Arequipa, 2018. Accedido el 21 de junio de 2022. [En línea]. Disponible en: http://tesis.ucsm.edu.pe/repositorio/handle/UCSM/8305

[7] O. D. Castrillón, "Las variables más influyentes en la obesidad: un análisis desde la minería de datos", Información tecnológica, vol. 32, n.º 6, pp. 123–132, diciembre de 2021. Accedido el 23 de junio de 2022. [En línea]. Disponible en: https://doi.org/10.4067/s0718-07642021000600123.

[8] F. Mendoza Palechor y A. de la Hoz Manotas. (2019). UCI Machine Learning Repository: Estimation of obesity levels based on eating habits and physical condition Data Set. [En línea] Available: https://archive.ics.uci.edu/ml/datasets/Estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition+. [Accedido: Jun 21, 2022]

[9] MORAL, D. R., & GARCIA, L. C. ANALISIS PREDICTIVO EN DIABETES TIPO 2 USANDO ESTRUCTURAS BIG DATA A. RODRIGUEZ 1, 2, V. SUAREZ-ULLOA1, C. TILVE ALVAREZ1, P. PUIG GALLEGO1, A. SOTO GONZALEZ1.

[10] Fierro, F. S., Castañeda, J., & Revelo-Aldás, M. (2022). Modelos predictivos para la estimación de adolescentes con tendencia al alcoholismo. AXIOMA, 1(26), 74-79.