Las variables más influyentes en la obesidad: un análisis desde la minería de datos

Omar Danilo Castrillón

Universidad Nacional de Colombia, Facultad de Ingeniería y Arquitectura, Departamento de Ingeniería Industrial, Campus La Nubia Bloque Q piso 2, Manizales - Colombia. (correo-e: odcastrillong@unal.edu.co)

Recibido Abr. 14, 2021; Aceptado Jun. 15, 2021; Versión final Jun. 30, 2021, Publicado Dic. 2021

Resumen

El principal objetivo de esta investigación es identificar las variables más influyentes en el grado de obesidad por medio de técnicas de minería de datos y la plataforma WEKA. Se reestructura una base de datos existente con 2.111 registros, 16 variables independientes y una variable dependiente (grado de obesidad). Se agrega la variable índice de masa corporal y se emplea el algoritmo J48 para realizar un proceso de selección estadística. Los resultados muestran que las variables independientes más influyentes son: género, estatura, peso e índice de masa corporal. El porcentaje de éxito supera el 97% por medio del algoritmo J48 y otras técnicas inteligentes ejecutadas con la misma plataforma mediante validación cruzada. Al suprimir las variables más influyentes, la edad e historia familiar tienen una influencia moderada. Se concluye que es responsabilidad de cada persona ser consiente y controlar cada una de estas variables si desea un control efectivo de su peso corporal.

Palabras clave: minería; datos; Weka; obesidad; selección estadística

The most influential variables associated with obesity: a data mining analysis

Abstract

The main objective of this research study is to identify the most influential variables on obesity degree by applying data mining techniques and by using the platform WEKA. An existing database is restructured. It contains 2,111 records, 16 independent variables, and 1 dependent variable (degree of obesity). The body mass index variable is added and the J48 algorithm is used to perform a statistical selection process. The results show that the most influential independent variables are: gender, height, weight, and body mass index. Success rate exceeds 97% when using the J48 algorithm and other intelligent techniques on the same platform. By suppressing the most influential variables, age and family history show a moderate influence. It is concluded that every person is responsible for controlling each of these variables to effectively control body weight.

Keywords: data mining; Weka; obesity; statistical selection

INTRODUCCIÓN

La obesidad puede ser considerada como la enfermedad de salud más grave del mundo (Hossain et al., 2018). La misma es asociada con un sin número de enfermedades y trastornos en las personas, con todos los problemas que esto último puede acarrear. Pudiéndose definir la obesidad como una alteración en el consumo de energía del cuerpo, lo cual generalmente se ve reflejado en el volumen de grasa corporal, aumentando desproporcionalmente. La misma puede ser diagnosticada por medio del Índice de Masa Corporal (IMC). Este problema puede estar asociado con algunas dietas inadecuadas y/o múltiples factores, causantes de cierto grado de obesidad o sobrepeso en algunas poblaciones o personas (Dalili et al., 2020). Así mismo, la obesidad también puede ser definida como un exceso de grasa corporal, donde usualmente (como se expresó antes) se emplean métodos indirectos para poder ser diagnosticada (IMC), pues los métodos directos no resultan prácticos. En este mismo sentido otros autores (Suárez-Carmona et al., 2017) definen la obesidad como una enfermedad crónica, siendo también uno de los principales problemas sociales del siglo XXI.

El objetivo fundamental de esta investigación es establecer por medio de una técnica de inteligencia artificial las variables más influyentes en el grado de obesidad de una persona. Se reestructura una base de datos existente (Palechor y De la Hoz, 2019; Dua, y Graff, 2019). Esta base de datos consta de una variable dependiente denominada grado de obesidad (GOB) con 7 estados (falta de peso, peso normal, sobre peso nivel I, sobre peso nivel II, obesidad tipo I, obesidad tipo II, obesidad tipo III) y 16 variables independientes: Género (G), Edad (E), Estatura (A), Peso (P), Historia Familiar (HF), Consumo frecuente de alimentos ricos en calorías (CFAC), Frecuencia de consumo de verduras (FCV), Número de comidas principales (NCP), Consumo de alimentos entre comidas (CAEC), Fuma (F), Consumo diario de agua (CDA), Monitoreo del consumo de calorías (MCC), Frecuencia de actividad física (FAF), Tiempo usando dispositivos tecnológicos (TUDT.), Consumo de alcohol (CA), Transporte usado (TU). Adicionalmente, la variable IMC es adicionada a la anterior Base de datos; posteriormente una selección estadística de variables es realizada.

La anterior selección de las variables, se realiza mediante la plataforma Weka (Witten, 2017), empleando el algoritmo J48, como se expresó en el párrafo anterior. Con la ayuda de este algoritmo, se seleccionan las variables independientes más influyentes. Este proceso de selección se realiza suprimiendo cada variable y mirando su influencia en el porcentaje de Éxito final, entre menor sea el porcentaje de Éxito, mayor es la influencia de la variable suprimida. Este último aspecto, permite determinar el porcentaje de éxito por medio de un proceso de validación cruzada de las variables más influyentes. Así mismo, este porcentaje de éxito es comparado con los resultados obtenidos por otras técnicas inteligentes. Adicionalmente, en la parte final de la metodología, son suprimidas las variables más influyentes, con el fin de evitar que estas oculten la influencia moderada de otras variables y poder descubrir nuevas variables incidentes.

Por otra parte, hoy en día se puede demostrar que la obesidad desarrolla muchos problemas en todo el organismo. La obesidad puede estar asociada con problemas en mecanismos de aprendizaje neurológicos (Mathar et al., 2017), considerándose la obesidad como una amenaza de rápido crecimiento para la salud mundial, la cual puede ser heredable (Khera et al., 2019). Recientemente, también se ha encontrado que la obesidad es un factor de riesgo de hospitalización asociado a las personas que padecen Covid - 19 (Bellini et al., 2021; Rossi et al., 2021). Siendo la obesidad y el sobre peso muy frecuentes en pacientes con trastornos depresivos (Si et al., 2021), igualmente se ha encontrado que la obesidad puede estar asociada con la diabetes mellitus tipo 2 (Zafra-Tanaka et al., 2020). Así mismo otros autores (Mulugeta y Hoque, 2021) determinaron el aumento de la obesidad en niños durante el encierro por el COVID-19. Adicionalmente, muchos otros problemas asociados a la obesidad y sobre peso pueden ser encontrados en la literatura (De-La-Hoz y Mendoza, 2019; Majd et al., 2021; Van et al., 2020).

Desde otro punto de vista, la inteligencia artificial y en especial las técnicas como la minería de datos también han sido usadas en el tratamiento de problemas de salud pública como la obesidad. Algunos autores (Triantafyllidis et al., 2020) han desarrollado estudios sobre el empleo de máquinas de aprendizaje para el análisis de la obesidad en niños, como medida de prevención y tratamiento. En este mismo sentido, por medio de técnicas inteligentes basadas en lógica difusa (Giabbanelli et al., 2012) también se han desarrollado mapas cognitivos para identificar las características psicosociales de la obesidad. Igualmente, otros autores (Hossain et al., 2018) han logrado predecir el riesgo de obesidad en personas de mediana edad por medio de técnicas de minería de datos. Por medio de estas mismas técnicas (Akben, 2019) también se ha logrado determinar los valores de los parámetros incluidos la obesidad, en la detección del cáncer de mama. En diversas bases de datos, pueden ser encontradas otras referencias relacionadas con la aplicación de estas técnicas (Zusi et al., 2019; Cañas y Martínez, 2020; Jaafar et al., 2021).

Es importante resaltar que, si bien existen diversos trabajos, desde varios puntos de vista como: Clínicos, sociales, computacionales, matemáticos, lógicos, basados en inteligencia artificial entre muchos otros. Muy pocos de estos estudios se han enfocado en determinar computacionalmente (técnicas de minerías de datos)

las principales variables que pueden influir en el grado de obesidad de una persona. Se expresa, que muchas de las investigaciones realizadas no son concluyentes o presentan resultados contradictorios respecto a otros estudios previamente realizados. Se vuelve así, de gran importancia establecer las variables más importantes que puedan influir en este problema, máxime cuando el mismo no solo es considerado el mayor problema de salud a nivel mundial y uno de los más importantes del siglo XXI, sino también un gran problema social, como se referencia en las diferentes citaciones relacionadas en esta introducción. La inclusión de técnicas inteligentes en la solución de este problema, permitirá establecer objetivamente y con precisión la incidencia de cada una de las variables objeto de análisis y determinar si las mismas tienen o no alguna relevancia sobre el comportamiento de la variable dependiente, en este caso denominada grado de obesidad.

Finalmente, con el fin garantizar una adecuada comprensión de este artículo, el mismo es organizado de la siguiente forma: a) *Primero*. Se presenta la metodología del documento, donde se explica en detalle cada uno de los pasos desarrollados en esta investigación y la forma como es estructurada la base de datos con sus respectivas variables independientes y dependientes, así como los procesos de selección de variables empleados. b) *Segundo*. Se muestran los resultados obtenidos, los cuales son una consecuencia lógica de aplicar la metodología propuesta en el paso uno. Los resultados también son desarrollados paso a paso, en concordancia con cada una de las partes estructuradas en la metodología. c) *Tercero*. Se realiza una comparación, entre los resultados obtenidos en esta investigación con otros resultados similares mostrados por diversos autores, estableciéndose así una sección de discusiones sobre los mismos. d) *Cuarto*. Se desarrollan las respectivas conclusiones del documento, según los resultados obtenidos después de aplicar la metodología propuesta. e) *Quinto*. Se incluyen dos secciones más denominadas agradecimientos y referencias bibliográficas.

MATERIALES Y METODOS

El desarrollo de la metodología propuesta, se inicia con la reestructuración de la base de datos tomada como referencia (Dua y Graff, 2019; Palechor y De la Hoz, 2019). La información estructurada en las anteriores referencias, permite construir las bases de datos que será analizada por medio de un algoritmo de comportamiento similar, a los algoritmos Bayesianos, denominado J48. Este programa es ejecutado por medio de la plataforma de aprendizaje automático y minería de datos denominada Weka (Witten, 2017). La anterior estructuración, permite definir la metodología que será aplicada. Esta metodología se hace de forma similar a otras metodologías previamente desarrolladas por los autores de este trabajo de investigación (Galeano et al., 2021). Esta metodología consta de los siguientes pasos: a) Paso 1. Reestructuración de la base de datos y adición de la variable IMC. b) Paso 2. Selección de variables por medio de la plataforma Weka. c) Paso 3. Porcentajes de éxito con las variables seleccionadas. d) Paso 4. Tablas de resumen, precisión y matriz de confusión. Generación del árbol de clasificación y comparación con otras técnicas de inteligencia artificial. e) Paso 5. Supresión de las variables más influyentes y determinación de otras variables.

Paso 1. Reestructuración de la base de datos y adición de la variable IMC.

La base de datos enunciada en el párrafo anterior, es reestructurada. Inicialmente la variable IMC definida como $\text{Peso (Kgs)} / \text{Estatura}^2 \text{ (Mts)}$, es adicionada. Posteriormente los valores de las variables independientes cualitativas son transformados en valores cuantitativos. Respecto a la variable dependiente solo se reduce el nombre de los valores que la misma puede tomar. Estos cambios permiten obtener una mayor claridad en el árbol de decisión final, obtenido por medio de la plataforma Weka. El renombramiento de los valores que toman estas variables es realizado conforme se referencia en la Tabla 1. En esta tabla, también se ilustra el estado de variables cualitativas que no cambiaron de valor, al igual que la nueva variable adicionada, esto con el fin de mostrar la estructura completa de la nueva base de datos. Adicionalmente, se calcula la efectividad de la base de datos inicial y se compara con la efectividad de la nueva base de datos reestructurada; aspecto que permite establecer la efectividad de los cambios realizados.

Paso 2. Selección de variables por medio de la plataforma Weka.

Tomando como referencia las 16 primeras variables, descritas en la Tabla 1, se seleccionan todos los posibles grupos de 15 variables (sin repeticiones y sin importar el orden de las mismas). En total son seleccionados y analizados 16 grupos por medio del algoritmo J48 de la plataforma de aprendizaje automático y minería de datos denominada Weka (Witten, 2017). Para cada uno de los grupos previamente establecidos, se determina la efectividad de las variables independientes del grupo seleccionado, en la predicción de la variable dependiente, en este caso Obesidad. Un menor porcentaje de éxito o acierto sobre la variable dependiente, indica una mayor influencia de la variable faltante en cada grupo seleccionado. Una vez realizado el anterior proceso, se seleccionan todas las variables independientes que presenten una influencia (disminución) superior al 2% respecto al éxito en la predicción de la variable dependiente. Siendo así, importante establecer previamente el porcentaje de éxito, respecto a la variable dependiente. Esta forma de selección permite establecer las variables independientes predominantes.

Tabla 1: Transformación variable independientes cualitativas y variable dependiente.

#	Variable	Sigla	Valor anterior	Valor nuevo	Observación
1	Genero	G	Femenino	0	
			Masculino	1	
2	Edad	E	Numérico	Numérico	No cambió
3	Estatura	A	Número	Numérico	No cambió
4	Peso	P	Numérico	Numérico	No cambió
5	Historia Familiar con sobrepeso	HF	Si	1	
			No	0	
6	Consumo de alimentos frecuentes ricos en calorías	CFAC	Si	1	
			No	0	
7	Frecuencia de consumo de verduras	FCV	1,2,3 (veces)	1,2,3	No cambió
8	Numero de comidas principales	NCP	1,2,3 (veces)	1,2,3	No cambió
9	Consumo de alimentos entre comidas	CAEC	No	0	
			Algunas veces	1	
			Frecuentemente	2	
			Siempre	3	
10	Fuma	F	Si	1	
			No	0	
11	Consumo diario de agua	CDA	1,2,3 (veces)	1,2,3	No cambió
12	Monitoreo del consumo de calorías	MCC	Si	1	
			No	0	
13	Frecuencia actividad física	FAF	0,1,2,3 (veces)	0,1,2,3	No cambió
14	Tiempo usando dispositivos tecnológicos	TUdT	0,1,2,3 (veces)	0,1,2,3	No cambió
15	Consumo de alcohol	CA	No	0	
			Algunas veces	1	
			Frecuentemente	2	
			Siempre	3	
16	Transporte Usado	TU	Caminar	0	
			Transporte publico	1	
			Bicicleta	2	
			Moto	3	
			Automóvil	4	
17	Obesidad	GOB	Falta de peso	INSW	
			Peso normal	NW	
			Sobre peso nivel I	OWI	
			Sobre pesos nivel II	OWII	
			Obesidad tipo I	OTI	
			Obesidad tipo II	OTII	
			Obesidad Tipo III	OTIII	
18	Índice de masa corporal ($V4/V3^2$)	IMC		Numérico	Adicionada

Paso 3. Porcentajes de éxito con las variables seleccionadas.

La ejecución del paso anterior, permite la reducción del conjunto variables independientes, por medio del proceso de selección previamente descrito. Con las variables seleccionadas en este proceso (Paso 2 de la metodología), la base de datos es reestructurada (se realiza una reducción dimensional). Además, sobre el conjunto de variables seleccionadas, se adiciona una nueva variable calculada denominada índice de masa corporal (IMC), la cual es descrita en la Tabla 1 (Variable Número 18). Con este conjunto reducido de variables seleccionadas y la nueva variable adicionada, se establece el nuevo porcentaje de éxito en la predicción de la variable dependiente denominada Obesidad. Nuevamente, este porcentaje de éxito es calculado por medio del algoritmo J48 de comportamiento similar a los algoritmos bayesianos, el cual es ejecutado por medio de la plataforma de aprendizaje automático y minería de datos denominada Weka. El porcentaje de éxito obtenido, en este paso, será comparado con otras técnicas inteligentes. Se resalta que los algoritmos de compartimento bayesiano pueden obtener muy buenos resultados con pocos datos. (Valencia et al., 2015)

Paso 4. Tablas de resumen, precisión y matriz de confusión.

Con el fin de establecer la consistencia de los resultados anteriores, en este punto se genera: a) Una tabla con el resumen de la clasificación, la cual además de los porcentajes de clasificación contiene las estadísticas Kappa, error absoluto medio, error cuadrático medio, error absoluto relativo, error cuadrático relativo b) Tabla de precisión detallada, la cual indica para cada clase el porcentaje de: positivos verdaderos, falsos positivos, precisión, recalcado, la medida F y el área ROC. c) Matriz de confusión d) árbol de clasificación. e) por medio de una técnica de validación cruzada (90%-10%), el porcentaje de éxito obtenido en el paso 3 de esta metodología, es comparado con los resultados de éxito que pueden ser obtenidos por medio de otra serie de técnicas inteligentes bayesianas. Con los resultados de las anteriores técnicas se obtiene el respectivo promedio de clasificación, para ser comparado.

Paso 5. Supresión de las variables más influyentes y determinación de otras variables.

Para determinar la influencia de otras variables independientes, sobre la variable dependiente, las variables más influyentes encontradas en el Paso 3 de la metodología son suprimidas. Una vez suprimidas estas variables, los Pasos 2, 3 y 4 de la metodología son adaptados al nuevo conjunto de variables y repetidos. Este aspecto permite encontrar otra serie de variables independientes, las cuales pueden estar enmascaradas por las variables más predominantes en este proceso, pero que de igual forma pueden tener una influencia moderada sobre la variable dependiente. Finalmente, en la sección siguiente se ilustran los resultados obtenidos, una vez realizado el desarrollo de cada uno de los pasos de la metodología presentada en esta sección del artículo.

RESULTADOS

Como resultado de ejecutar cada uno de los pasos de la metodología descritos en la sección anterior, se obtiene los siguientes resultados, los cuales son descritos a continuación:

Paso 1. Restructuración de la base de datos y adición de la variable IMC. La ejecución de este paso, permite obtener una nueva base de datos como la ilustrada en la Tabla 2, donde por razones de espacio solo es mostrada parte de esta tabla. Con la base de datos inicial, es decir con las 16 primeras variables independientes, por medio del algoritmo J48, de la plataforma WEKA, se logra una efectividad 93.47% en la predicción de la variable dependiente GOB. No obstante, cuando la variable independiente calculada IMC es adicionada, la efectividad en la predicción de la variable dependiente GOB, sube al 97.48%, empleando el mismo algoritmo J48 de la plataforma de WEKA.

Tabla 2: Parte de la nueva base de datos obtenida (Dua y Graff, 2019; Palechor y De la Hoz Manotas, 2019)

G	E	A	P	HF	CFAC	FCV	NCP	CAEC	F	CDA	MCC	FAF	TUDT	CA	TU	GOB	IMC
0	21	1,62	64	1	0	2	3	1	0	2	0	0	1	0	1	1	24,39
0	21	1,52	56	1	0	3	3	1	1	3	1	3	0	1	1	1	24,24
1	26	1,85	105	1	1	3	3	2	0	3	0	2	2	1	1	4	30,68

Paso 2. Selección de variables por medio de la plataforma Weka. Siguiendo la metodología ilustrada en el paso de la sección denominada *Materiales y Métodos*, se seleccionaron todos los subgrupos posibles de 15 variables (tomado de las 16 primeras variables, sin repeticiones y sin importar el orden). Con cada uno de esto subgrupos se calcula el porcentaje de éxito en la predicción de la variable dependiente. Este aspecto es ilustrado en la Tabla 3. En la tabla 3, las líneas resaltadas en mayúsculas (líneas 1, 3 y 4), representan aquellos conjuntos de variables, en los cuales el porcentaje de éxito fue menor. Este último aspecto permite identificar tres variables ausentes (una en cada conjunto - renglón) las cuales tienen una influencia significativa sobre la variable dependiente. Estas variables independientes son: Género (G), Estatura (A) y Peso (P).

Paso 3. Porcentajes de éxito con las variables seleccionadas. Con base en las variables seleccionadas en el paso 2 (G, A, P) se realiza un nuevo proceso de clasificación por medio del algoritmo J48 de la plataforma WEKA, mediante un proceso de validación cruzada (90 -10), lográndose una efectividad del 95.26%. Cuando la variable IMC es adicionada y se realiza el mismo proceso anterior, el porcentaje de efectividad sube al 97.15%. Con diferentes porcentajes de partición, en el proceso de validación cruzada, el porcentaje de éxito varía muy levemente, lo que demuestra la estabilidad del sistema propuesto, dado que se puede demostrar que no existe diferencia significativa en los datos obtenidos. Resultados similares son obtenidos por medio del método de particiones. Los anteriores aspectos son ilustrados en las Tablas 4 y 5:

Tabla 3: Predicción variable dependiente en subgrupos de 15 variables sobre 16 variables iniciales

G	E	A	P	HF	CFAC	FCV	NCP	CAEC	F	CDA	MCC	FAF	TUDT	CA	TU	% Éxito
	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	91.22
x		x	x	x	x	x	x	x	x	x	x	x	x	x	x	93.33
X	X		X	X	X	X	X	X	X	X	X	X	X	X	X	87.30
X	X	X		X	X	X	X	X	X	X	X	X	X	X	X	78.35
x	x	x	x		x	x	x	x	x	x	x	x	x	x	x	93.84
x	x	x	x	x		x	x	x	x	x	x	x	x	x	x	94.12
x	x	x	x	x	x		x	x	x	x	x	x	x	x	x	93.93
x	x	x	x	x	x	x		x	x	x	x	x	x	x	x	93.93
x	x	x	x	x	x	x	x		x	x	x	x	x	x	x	94.31
x	x	x	x	x	x	x	x	x		x	x	x	x	x	x	93.74
x	x	x	x	x	x	x	x	x	x		x	x	x	x	x	93.79
x	x	x	x	x	x	x	x	x	x	x		x	x	x	x	93.55
x	x	x	x	x	x	x	x	x	x	x	x		x	x	x	93.79
x	x	x	x	x	x	x	x	x	x	x	x	x		x	x	93.88
x	x	x	x	x	x	x	x	x	x	x	x	x	x		x	93.88
x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		93.65

Tabla 4: %Éxito por medio de una validación cruzada

Validación aplicada	% Éxito
90 -10	97.15
80 -20	97.29
70 -30	97.15

Tabla 5: %Éxito por medio de particiones

Validación aplicada	% Éxito
66 -33	97.49
50-50	97.34

Paso 4. Tablas de resumen, precisión y matriz de confusión. Generación del árbol de clasificación y comparación con otras técnicas de inteligencia artificial. Con el fin de establecer la consistencia y visualizar mejor los datos, se generan las Tablas 6,7,8,9 y la figura 1, las cuales ilustran los resultados obtenidos:

Tabla 6: Resumen clasificación

Estadística	Número	Porcentaje
Instancias clasificadas correctamente	2051	97.1577 %
Instancias clasificadas incorrectamente	60	2.8423 %
Estadísticas Kappa	0.9668	
Error absoluto medio	0.0129	
Error cuadrático medio	0.089	
Error absoluto relativo	5.27%	
Error cuadrático relativo	25.4563%	
Número total de instancias	2111	

Tabla 7: Tabla de precisión detallada

Positivos Verdaderos	Falsos Positivos	Precisión	Recalculo	Medida F	Área ROC	Clase
0.978	0.001	0.996	0.978	0.987	0.997	INSW
0.986	0.008	0.95	0.986	0.968	0.988	NW
0.931	0.005	0.964	0.931	0.947	0.981	OWI
0.959	0.007	0.955	0.959	0.957	0.988	OWII
0.969	0.005	0.974	0.969	0.971	0.991	OTI
0.98	0.004	0.973	0.98	0.977	0.99	OTII
0.997	0.002	0.988	0.997	0.992	0.997	OTIII
0.972	0.005	0.972	0.972	0.972	0.99	←Promedio

Tabla 8: Matriz de Confusión

a	b	c	d	e	f	g	← Clasificado
266	6	0	0	0	0	0	a = INSW
1	283	3	0	0	0	0	b = NW
0	9	270	11	0	0	0	c = OWI
0	0	7	278	5	0	0	d = OWII
0	0	0	2	340	7	2	e = OTI
0	0	0	0	4	291	2	f = OTII
0	0	0	0	0	1	323	g = OTIII

Fig. 1: Árbol de clasificación.

Tabla 9: Comparación con otras técnicas inteligentes

Técnica	% Éxito	Descripción
Bayesnet	96.54	https://weka.sourceforge.io/doc.dev/weka/classifiers/bayes/BayesNet.html
NaiveBayes	94.36	https://weka.sourceforge.io/doc.dev/weka/classifiers/bayes/NaiveBayes.html
NaiveBayessimle	94.36	https://weka.sourceforge.io/doc.stable/weka/classifiers/bayes/NaiveBayesSimple.html
Promedio	95.08	2.06% inferior al porcentaje logrado con la metodología propuesta (Tabla 6)

Paso 5. Supresión de las variables más influyentes y determinación de otras. Al suprimir las cuatro variables más predominantes (género, peso, altura e IMC) y repetir los pasos 1 - 4 de la metodología, se encuentran otras dos variables independientes con una influencia moderada sobre la variable dependiente: edad (E) e historia familiar con sobre peso (HF). No obstante, se resalta que sin las 4 variables suprimidas el porcentaje de clasificación solo llega al 75.13%.

DISCUSIÓN

En el paso 4 de la metodología, se muestra como género, peso, altura y la variable calculada IMC, son suficientes para el control de la obesidad en una persona, sin embargo, esta investigación pretende ir más allá y encontrar la influencia de otra serie de variables, además de las tradicionalmente usadas. Este último aspecto va en sentido similar a lo planteado por algunos autores (Dalili et al., 2020), quienes consideran que el IMC puede ser una medida limitada y abierta a interpretaciones u omisión de datos. Es así, como en esta investigación se muestra que, al incluir el cálculo de esta variable (IMC) en la base de datos, los resultados de éxito mejoran notoriamente. No obstante, se observa que la variable (IMC) encapsula la influencia de otra serie de variables independientes (Edad - E e Historia familiar con sobre peso HF) como se muestra en el paso 5 de la metodología. Por ello, al suprimir las variables predominantes, se encuentran estas dos variables (E) y (HF), las cuales presentan una influencia moderada sobre la variable dependiente.

En el mismo sentido otros autores (Dent et al., 2020) encuentran que la obesidad, puede estar ligada a factores como adherencia a la dieta, la actividad física, el sexo, la edad y medicamentos específicos, lo cual es concordante con los resultados encontrados en esta investigación dado que las variables edad y sexo fueron halladas directamente en esta investigación y las demás variables como actividad física, dietas, y medicamentos están ligadas a una historia familiar con sobre peso (HF), la cual arrastra todos los hábitos de vida y forma de alimentación de padres a hijos. Si bien el control de la obesidad siempre va estar ligado a una serie de variables predominantes como género, peso, estatura, IMC, en este trabajo se encuentran otra serie de variables como edad e historia familiar con sobre peso, las cuales aunque presenten una influencia moderada sobre la variable dependiente objeto de estudio, son de gran importancia en el control de la obesidad y las personas de formas individual deben ser consientes sobre lo que cada una de ellas implica en especial la variable historia familiar con sobre peso.

Los estudios realizados son válidos en el caso de las muestras analizadas, sería aconsejable extender este trabajo a otra serie de poblaciones y grupos étnicos con el fin de lograr obtener una mayor validez del trabajo desarrollado y establecer de esta forma una serie de conclusiones más generales, las cuales pueden redundar en el beneficio de toda la humanidad y ayudar al tratamiento de este grave problema de salud pública. Este aspecto haría posible caracterizar una serie de variables influyentes en el grado de obesidad de una persona las cuales pueden diferir según el contexto poblacional.

CONCLUSIONES

El desarrollo de la metodología propuesta, permite encontrar 6 variables influyentes en el grado de obesidad de una persona: Género (G), peso (P), altura (A), índice de masa corporal (IMC), edad (E) e historia familiar con sobre peso (HF). Sin embargo, y en concordancia con diversos autores es responsabilidad de cada persona ser consiente y controlar el estado de estas variables, si desea un control efectivo sobre su peso corporal. Finalmente, se resalta que la metodología propuesta bajo estas técnicas de inteligencia artificial es igual o superior a otras técnicas inteligentes (Tabla 9) además, puede ser fácilmente aplicada a diversas bases de datos, para lo cual solo es necesario repetir cada uno de los pasos propuestos en la misma, aspecto que permitirá contribuir de una mejor forma en la solución de este problema a nivel mundial.

AGRADECIMIENTO

Se agradece la colaboración a la Universidad Nacional de Colombia Sede Manizales y al Departamento de Ingeniería Industrial de la misma Sede. Igualmente se agradece a: Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [<http://archive.ics.uci.edu/ml>]. Irvine, CA: University of California, School of Information and Computer Science.

REFERENCIAS

- Akben, S.B., Determination of the blood, hormone and obesity value ranges that indicate the breast cancer, using data mining based expert system, <https://doi.org/10.1016/j.irbm.2019.05.007>, IRBM, 40, 355–360 (2019)
- Bellini, B., Cresci, B., y otros ocho autores, Obesity as a risk factor for hospitalization in coronavirus disease-19 (COVID-19) patients: analysis of the Tuscany regional database, <https://doi.org/10.1016/j.numecd.2020.11.030>, Nutrition, Metabolism & Cardiovascular Diseases, 31(3), 769-773 (2021)
- Cañas, R., y Martínez, U., Estimation of obesity levels based on computational intelligence, <https://doi.org/10.1016/j.imu.2020.100472>, Informatics in Medicine Unlocked, 21, 00472 (2020)
- Dalili, D., Bazzocchi, A., y otros tres autores, The role of body composition assessment in obesity and eating disorders, <https://doi.org/10.1016/j.ejrad.2020.109227>, European Journal of Radiology, 131, 109227 (2020)

- De-La-Hoz-Correa, E., Mendoza-Palechor, F., y otros tres autores, Obesity level estimation software based on decision trees, <https://doi.org/10.3844/jcssp.2019.67.77>, Journal of Computer, 15 (1), 67-77 (2019)
- Dent, R., McPherson, R., y Harper, M., Factors affecting weight Loss variability in obesity, <https://doi.org/10.1016/j.metabol.2020.154388>, Metabolism Clinical and Experimental, 113, 154388 (2020)
- Dua, D., y Graff, C., UCI Machine Learning Repository <http://archive.ics.uci.edu/ml> Irvine, CA: University of California, School of Information and Computer Science, (2019)
- Galeano, L., Aguirre, S., y Castrillón, O., Análisis de calidad del vino por medio de técnicas de inteligencia artificial, <http://dx.doi.org/10.4067/S0718-07642021000100017>, Información tecnológica, 32(1), 17-26 (2021)
- Giabbanelli, P., Torsney-Weir, T., y Mago, V-K., A fuzzy cognitive map of the psychosocial determinants of obesity, <https://doi.org/10.1016/j.asoc.2012.02.006>, Applied Soft Computing, 12, 3711–3724 (2012)
- Hossain, R., Mahmud, S.M., y otros tres autores, PRMT: predicting risk factor of obesity among middle-aged people using data mining techniques, <https://doi.org/10.1016/j.procs.2018.05.022>, Procedia Computer Science, 132, 1068–1076 (2018)
- Jaafar, Z., Kreidieh, D., y otros tres autores, Cross-validation of prediction equations for estimating the body fat percentage in adults with obesity, <https://doi.org/10.1016/j.clnesp.2020.11.003>, Clinical Nutrition ESPEN, 41, 346e350 (2021)
- Khera, A., Chaffin, M., y otros diecisiete autores, Polygenic prediction of weight and obesity trajectories from birth to adulthood, <https://doi.org/10.1016/j.cell.2019.03.028>, Cell, 177, 587–596 (2019)
- Majd M., Smyth, J., y otros ocho autores, The factor structure of depressive symptoms in patients with obesity enrolled in the rainbow clinical trial, <https://doi.org/10.1016/j.jad.2020.11.105>, Journal of Affective Disorders, 281, 367–375 (2021)
- Mathar, D., Neumann, J., y otros dos autores, Failing to learn from negative prediction errors: obesity is associated with alterations in a fundamental neural learning mechanism, <http://dx.doi.org/10.1016/j.cortex.2017.08.022>, Cortex, 95, 222e2237 (2017)
- Mulugeta, W., y Hoque, L., Impact of the COVID-19 lockdown on weight status and associated factors for obesity among children in Massachusetts, <https://doi.org/10.1016/j.obmed.2021.100325>, Obesity Medicine, 22, 100325 (2021)
- Palechor, F. M., y De la Hoz Manotas, A., Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico, Data in Brief, 104344 (2019)
- Rossi, A., Gottin, L., y otros seis autores, Obesity as a risk factor for unfavourable outcomes in critically ill patients affected by COVID 19, <https://doi.org/10.1016/j.numecd.2020.11.012>, Nutrition, Metabolism & Cardiovascular Diseases, 31(3), 762-768 (2021)
- Si, T., Yang, K., y otros cinco autores, Prevalence and risk factors of overweight and obesity in Chinese patients with first episode drug-naïve major depressive disorder, <https://doi.org/10.1016/j.jad.2021.01.037>, Journal of Affective Disorders, 286 (1), 351 - 359 (2021)
- Suárez-Carmona, W., Sánchez-Oliver, A., y González-Jurado, J., Fisiopatología de la obesidad: perspectiva actual, <http://dx.doi.org/10.4067/s0717-75182017000300226>, Rev. Chil. Nutr., 44(3), 226 - 233 (2017)
- Triantafyllidis, A., Polychronidou, E., y otros trece autores, Computerized decision support and machine learning applications for the prevention and treatment of childhood obesity: a systematic review of the literatura, <https://doi.org/10.1016/j.artmed.2020.101844>, Artificial Intelligence In Medicine 104, 101844 (2020)
- Valencia, M., Correa, J., y Díaz, F., Métodos estadísticos clásicos y Bayesianos para el pronóstico de demanda. Un Análisis Comparativo, <https://doi.org/10.15446/rev.fac.cienc.v4n1.49775>, Revista Facultad de Ciencias Universidad Nacional de Colombia, 4(1), 52 -67 (2015)
- Van, T., Uyen, T., y otros ocho autores, The impact of sexual harassment on obesity in female adolescents: an update and perspective to control, <https://doi.org/10.1016/j.dsx.2020.09.036>, Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 14, 1931e1939 (2020)
- Witten, I., Frank, E., y otros dos autores, Data mining practical machine learning tools and techniques, Morgan and Kaufman publication (Elsevier), Cambridge, USA (2017)
- Zafra-Tanaka, J., Miranda, J., y otros cuatro autores, Obesity markers for the prediction of incident type 2 diabetes mellitus in resource-Poor settings: The Cronicas Cohort Study, <https://doi.org/10.1016/j.diabres.2020.108494>, Diabetes Research and Clinical Practice, 170, 108494 (2020)
- Zusi, C., Mantovani, A., y otros nueve autores, Contribution of a genetic risk score to clinical prediction of hepatic steatosis in obese children and adolescents, <https://doi.org/10.1016/j.dld.2019.05.029>, Digestive and Liver Disease, 51, 1586–1592 (2019)

Página en blanco