Algoritmos de clasificación para la detección de obesidad en adolescentes: Un estudio comparativo entre KNN y árboles de decisión

Classification algorithms for detecting obesity in teenagers: A comparative study between KNN and decision trees

Samuel Erasto López Díaz
Universidad del Istmo, campus Tehuantepec
lopezdiaz200028@gmail.com

José Alejandro Pérez Sibaja
Universidad del Istmo, campus Tehuantepec
japs_misterio@hotmail.com

Adonay Flores Martínez
Universidad Alfred Novel, campus Tehuantepec
trabajos.afm@gmail.com

Sergio Juárez Vázquez
Universidad del Istmo, campus Tehuantepec
sjuarez@sandunga.unistmo.edu.mx
ORCID: 0000-0002-2080-4861

doi: https://doi.org/10.36825/RITI.11.23.007

Recibido: Febrero 20, 2023
Aceptado: Mayo 17, 2023

Resumen: El aprendizaje automático es una rama de la inteligencia artificial que se centra en desarrollar algoritmos y modelos estadísticos para que los sistemas informáticos puedan aprender y mejorar su desempeño a partir de datos. Los algoritmos de clasificación son un tipo de modelo de aprendizaje automático que se utilizan para predecir la clase o categoría de un objeto en función de las características o atributos observados. En el caso de la clasificación de la obesidad, estos algoritmos se han utilizado para desarrollar modelos que permitan predecir si un individuo tiene obesidad a partir de datos como el índice de masa corporal, la edad, el género y otros factores de riesgo. Esto puede ayudar a identificar la obesidad tempranamente y a implementar intervenciones preventivas y de tratamiento más eficaces. En este artículo se compara la eficacia de dos algoritmos para predecir la obesidad en adolescentes utilizando un conjunto de datos de 200 participantes de entre 15 y 19 años y cuatro variables (peso, edad, talla y género). Se comparan los algoritmos de árboles de decisión y *k* vecinos más cercanos, y se concluye que ambos son efectivos en la clasificación de la obesidad en adolescentes, aunque los árboles de decisión son una opción más precisa.

Palabras clave: *KNN, Árboles de Decisión, Obesidad en Adolescentes.*

Abstract: Machine learning is a branch of artificial intelligence that focuses on developing statistical algorithms and models for computer systems to learn and improve their performance from data. Classification algorithms are a type of machine learning model used to predict the class or category of an object based on observed features or attributes. In obesity classification, these algorithms have been used to develop models that predict whether an individual has obesity based on data such as body mass index, age, gender, and other risk factors. This can help identify obesity early and implement more effective preventive and treatment interventions. This article compares the effectiveness of two algorithms in predicting obesity in adolescents using a dataset of 200 participants aged 15 to 19 and four variables (weight, age, height, and gender). The decision tree and k-nearest neighbor algorithms are compared, and it is concluded that both are effective in the classification of obesity in adolescents, although decision trees are a more accurate option.

Keywords*: KNN, Decision Trees, Obesity in Adolescents.*

1. Introducción

La obesidad es un problema de salud grave que afecta a millones de personas en todo el mundo. Según [1] "La obesidad es uno de los principales factores de riesgo para numerosas enfermedades crónicas, entre las que se incluyen la diabetes, las enfermedades cardiovasculares, la hipertensión y los accidentes cerebrovasculares, así como varios tipos de cáncer". Este es un problema que afecta a todas las personas de todas las edades y grupos sociales en México y en todo el mundo. México es el más afectada, con una prevalencia del 70% de adultos con sobrepeso u obesidad, la más alta entre todas las regiones de la Organización Mundial de la Salud [2]. En el grupo de niños y adolescentes de 5 a 19 años, el 33,6% están afectados por sobrepeso u obesidad, según la UNICEF, la OMS y el Banco Mundial.

La obesidad en adolescentes se ha convertido en un problema de salud pública creciente en todo el mundo. La mayoría de los jóvenes con obesidad mantienen este problema hasta la edad adulta [3]. Esto puede causar complicaciones médicas como las que anteriormente fueron mencionadas. Además de las complicaciones médicas que puede causar la obesidad adolescente, el problema puede tener un impacto negativo en el bienestar social y emocional de los jóvenes. Por ejemplo, los jóvenes obesos pueden experimentar baja autoestima, ansiedad y depresión, lo que a su vez puede afectar su capacidad para relacionarse con sus compañeros y participar en actividades sociales.

Además, la obesidad adolescente puede afectar el desarrollo y la capacidad de los jóvenes para lograr ciertas metas a largo plazo, como terminar la escuela o ingresar al mercado laboral. Es importante recordar que durante la adolescencia se producen cambios fisiológicos y de desarrollo importantes, como el aumento de la masa muscular y la maduración sexual, que pueden verse afectados negativamente por la obesidad. Por lo tanto, se deben considerar cuidadosamente los aspectos médicos, psicológicos y sociales para abordar la obesidad adolescente de manera integrada y multidisciplinaria [4].

La Encuesta Nacional de Salud y Nutrición (ENSANUT) reveló que la obesidad ha aumentado en México desde 2006 hasta 2021. En el estudio participaron 2.230 adolescentes de un total de 17.107.800 mujeres y hombres entre 12 y 19 años. Los resultados mostraron que los jóvenes masculinos experimentaron un mayor incremento en la prevalencia de obesidad, llegando al 21,5%, mientras que las mujeres tuvieron una prevalencia del 15%. La obesidad alcanzó su pico más alto en adolescentes de 14, 15 y 16 años, con un 20%, 20% y 20,3%, respectivamente. Además, se analizó la prevalencia de sobrepeso y obesidad según el tipo de localidad de residencia de los adolescentes. Las localidades rurales presentaron una menor prevalencia de obesidad (14,6%) en comparación con las zonas urbanas, que registraron la prevalencia más alta de obesidad en el total de adolescentes, con un 19,3% [5].

Por lo tanto, la obesidad en adolescentes en México es un problema de salud pública que debe ser abordado debido a sus efectos negativos en la salud física y mental de los jóvenes. Es importante combatir la obesidad en adolescentes para mejorar su salud a corto y largo plazo, y reducir el costo económico y social del tratamiento de las enfermedades relacionadas con la obesidad. Pero para poder combatirla, primero debemos detectar el estado en el que se encuentra cada adolescente, ya sea que tenga un peso normal, sobrepeso u obesidad. Para ello, se pueden utilizar varios métodos.

El índice de masa corporal (IMC), es uno de los indicadores más utilizados, el cual considera la altura como el peso de una persona. Además, el índice permite clasificar a los jóvenes en diferentes categorías según su estado

nutricional: bajo peso, normopeso, sobrepeso u obesidad. Sin embargo, es conocido que el IMC no es una medida absoluta y puede tener algunas limitaciones, especialmente en personas jóvenes con alta masa muscular. Por esta razón, se recomiendan otras medidas, como la circunferencia de la cintura, para evaluar la distribución de la grasa corporal y el riesgo de enfermedades metabólicas relacionadas con la obesidad, como la diabetes tipo 2 y las enfermedades cardiovasculares. En cualquier caso, la detección precoz del estado nutricional de los adolescentes puede intervenir eficazmente en la prevención y tratamiento de la obesidad.

Una forma de clasificar el estado de un objeto es mediante el uso de algoritmos de clasificación, que emplean el aprendizaje automático para asignar objetos a diferentes categorías o clases. Estos algoritmos utilizan las características o atributos de los objetos como entrada y, basándose en un modelo previamente entrenado, los clasifican en una de varias categorías predefinidas [6]. Los algoritmos de clasificación son ampliamente utilizados en una gran variedad de áreas, como el reconocimiento de imágenes, la detección de spam y el fraude con tarjetas de crédito, etc. Además, se puede mencionar que estos algoritmos son muy útiles para automatizar tareas que antes requerían mucho tiempo y esfuerzo humano.

Los algoritmos de clasificación basados en inteligencia artificial pueden ofrecer una forma precisa y automatizada de clasificar el peso en adolescentes, ayudando así a prevenir la obesidad. En este estudio, se desarrolló y validó un algoritmo de aprendizaje automático para clasificar a los adolescentes con obesidad utilizando medidas antropométricas. Se utilizó un conjunto de datos prospectivos de adolescentes y se aplicó el algoritmo a través de un programa desarrollado en MATLAB. Se evaluó el rendimiento de cada uno de los algoritmos de clasificación y se comparó su precisión. En conclusión, el programa de aprendizaje automático ofrece una forma precisa y automatizada de clasificar el peso en adolescentes, lo que puede resultar una herramienta efectiva en la prevención de la obesidad en los adolescentes.

2. Estado del arte

En [7], los autores presentan una solución para la estimación y predicción de niveles de obesidad. Esta solución permite a una persona conocer su estado físico actual. Para ello, se utilizó un *dataset* de personas con obesidad, basándose en sus hábitos alimenticios y su condición física. El *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. Con todos estos datos, se creó un árbol de decisión, obteniendo así una exactitud del 72.43%, que depende de la cantidad de datos de entrenamiento que se proporciona. En este caso, se empleó el 20% de los datos para la realización de pruebas y el 80% restante para el entrenamiento.

En [8], se utilizó la técnica *Naive Bayes* para predecir la obesidad en niños menores de 5 años. Se utilizó un conjunto de datos de 770 registros y 27 variables extraídos del aplicativo *e-Qhali* para implementar el modelo. La técnica *Naive Bayes* se aplicó al 41% de los datos recolectados, y se destacaron los siguientes resultados:

- 214 casos Verdaderos Negativos
- 89 casos Falsos Positivos
- 1 caso de Falso Negativo
- 13 casos de Verdaderos Positivos

Las pruebas fueron efectuadas sobre 317 registros obteniendo un modelo con 72% de precisión y 93% de sensibilidad. Esta técnica fue comparada con otras como Regresión Logística, *Random Forest* y SVM, alcanzando el mayor porcentaje de sensibilidad.

En [9] se basa en un modelo de inferencia difusa para detectar tempranamente el sobrepeso y la obesidad en niños y adolescentes en el ámbito escolar. Se utilizó un conjunto de 81 reglas para clasificar la composición corporal de 92 niñas y 88 niños de entre 10 y 14 años. El modelo se basó en cuatro atributos: peso, índice de masa corporal, porcentaje de masa grasa y porcentaje de peso habitual. El resultado obtenido por el modelo difuso fue una precisión promedio de 97,6% con un 4,2% de error para la clasificación corporal de las niñas, y de 93,4% con un 4,6% de error para los niños.

El siguiente trabajo [10] tiene como objetivo comparar tres tipos de algoritmos evolutivos diferentes: Real *Encoding Particle Swarm Optimization* (REPSO–C), Incremental *Learning with Genetic Algorithms* (ILGA) y *Decision Tree with Genetic Algorithm* (DT-GA), para determinar el porcentaje de mejora durante cada iteración utilizando la herramienta *Keel* sobre una base de datos relacionada con la obesidad escolar de niños y adolescentes

entre cinco y diecisiete años. Se tomaron en cuenta las variables de estatura (m), peso (kg) e índice de masa corporal (IMC=peso/estatura^2). La base de datos utilizada se muestra en la Tabla 1.

Tabla 1. Distribución original de los datos.

Categoría	Niños	Niñas
Normal	2331	2429
Sobrepeso	436	459
Obesidad	171	136
Total	2938	3024

Fuente: Elaboración propia.

El resultado obtenido destacó el método DT-GA, el cual mostró un mejor porcentaje de precisión (93.76%). Esto se debe a que se utiliza un modelo híbrido que combina un árbol de decisión mejorado con algoritmos genéticos. En este modelo, el componente de evaluación es un árbol de decisión, mientras que el componente de búsqueda es un GA.

El objetivo del artículo [11] es clasificar la obesidad en niños y adolescentes varones de entre 6 y 17 años utilizando un modelo neuro-difuso ANFIS (*Artificial Neural Network Fuzzy Inference System*) que se encuentra en el *toolbox* de Matlab. El estudio se realizó en escolares agrupados por edad, realizando doce estudios en un total de 2,938 escolares. Las pruebas realizadas mostraron una exactitud del 96.96% en la clasificación y un error del 3.04%. Esta técnica se comparó con otras, como la red neuronal MLP con un 96.80%, SVM con un 96.28% y *Naive Bayes* con un 73.42%, y se encontró que el modelo propuesto obtuvo una precisión superior.

El estudio realizado en [12] examinó la relación entre el estado de peso y las actividades físicas en los seres humanos, además de comparar algunos modelos de aprendizaje automático y estadísticos clásicos utilizados en la predicción del nivel de obesidad. Utilizaron el conjunto de datos de la Encuesta Nacional de Salud y Nutrición para su modelo, y se emplearon once algoritmos diferentes, como subespacio aleatorio, regresión logística, tabla de decisión, *Naive Bayes*, función de base radial, vecino más cercano (*k-nearest neighbor*), clasificación mediante regresión, J48 y percepción multicapa. Se utilizó la métrica de evaluación ROC y AUC, y el algoritmo que obtuvo la mayor precisión general fue el algoritmo de clasificación de subespacio aleatorio.

3. Materiales y métodos

La metodología utilizada para clasificar y predecir la obesidad en adolescentes está dividida en 7 etapas, según se muestra en la Figura 1. Cada una de las cuales son explicadas en las siguientes subsecciones.

Figura 1. Metodología propuesta para la clasificación de obesidad.

3.1. Conjunto de datos

Para crear su conjunto de datos, los autores buscaron en la literatura los factores o hábitos clave asociados con la obesidad. El conjunto de datos consta de 4 variables que pueden determinar si una persona tiene obesidad. Para recolectar la información, se utilizaron encuestas para recopilar la edad, el sexo, el peso en kg y la altura en metros.

En la encuesta participaron 200 estudiantes de preparatoria, 99 hombres y 101 mujeres, con edades entre 15 y 19 años.

A continuación, se presenta una tabla en donde se describe cada uno de las características utilizadas en el estudio de predicción de obesidad en adolescentes, de igual manera se proporciona una breve explicación de cada uno (véase la Tabla 2).

Tabla 2. Características utilizadas para la Predicción de Obesidad en Adolescentes.

Característica	Descripción
Edad	La edad del individuo en años.
Sexo	El sexo del individuo (masculino o femenino).
Peso	El peso del individuo, medido en kilogramos.
Altura	La altura del adolescente, medido en metros.

Fuente: Elaboración propia.

3.2. Definición de algoritmo k-NN

El algoritmo *k-Nearest Neighbors* (k-NN), también conocido como KNN, es un clasificador de aprendizaje supervisado no paramétrico que utiliza la cercanía para hacer predicciones sobre la agrupación de un punto de datos en particular. Aunque puede ser utilizado para problemas de regresión o clasificación, generalmente se aplica como un algoritmo de clasificación. El algoritmo parte de la idea de que los puntos similares suelen estar cerca entre sí [11].

En problemas de clasificación, la etiqueta de clase se asigna en base a la mayoría de los votos, es decir, la etiqueta que aparece con más frecuencia en los puntos de datos alrededor de un punto determinado. Aunque se utiliza comúnmente el término "votación mayoritaria" para describir este proceso, técnicamente no es correcto. La "votación mayoritaria" implica una mayoría superior al 50%, lo que sólo es válido en casos con dos categorías. Con varias clases, como cuatro categorías, no es necesario obtener el 50% de los votos para llegar a una conclusión sobre una clase; se puede asignar una etiqueta de clase con un voto mayor al 25% [11].

3.3. Función knnclassify

La función *knnclassify* en Matlab es una herramienta de clasificación de aprendizaje no supervisado que utiliza el algoritmo *k-Nearest Neighbors* (KNN). Esta función toma como entrada un conjunto de datos de prueba y un conjunto de datos de entrenamiento, junto con sus respectivas etiquetas de clase y un valor de *k*. A continuación, devuelve las etiquetas de clase predichas para cada punto de prueba.

La función *knnclassify* se encuentra en el paquete de aprendizaje automático y minería de datos de Matlab, y es útil para realizar tareas de clasificación en datos no etiquetados o para evaluar el desempeño de un modelo de clasificación. Esta función se utiliza comúnmente en aplicaciones de análisis de datos, visión por computadora, procesamiento de imágenes y otras áreas relacionadas con el aprendizaje automático. Para este proyecto, se utilizó la función *knnclassify* de Matlab [13].

3.4. Métricas de distancia

3.4.1. Distancia euclidiana

La medida de distancia más comúnmente utilizada se limita a vectores con valores reales. Esta medida se usa para calcular la distancia entre dos puntos a través de una línea recta, utilizando la fórmula correspondiente [11].

$$Distancia\ Euclidiana = d(x,y) = \sqrt{\sum_{i=1}^{n}(yi - xi)^2} \quad (1)$$

3.4.2. Distancia de Manhattan

Esta es una métrica de distancia comúnmente utilizada que mide la diferencia absoluta entre dos puntos. También es conocida como *taxicab distance* o *city block distance*, ya que se puede visualizar como un viaje en una cuadrícula que representa las calles de una ciudad [11].

$$Distancia\ de\ Manhattan\ = d(x,y) = (\sum_{i=1}^{m} |xi - yi|) \tag{2}$$

3.4.3. Distancia de Minkowski

Esta es una medida de distancia ampliada que comprende tanto la distancia euclidiana como la distancia de Manhattan. La fórmula incluye un parámetro, *p*, que permite personalizar la métrica de distancia según las necesidades del problema. Cuando se establece en *p*=2, se refiere a la distancia euclidiana y cuando se establece en *p*=1, se refiere a la distancia de Manhattan [11].

$$Distancia\ de\ Minkowski = (\sum_{i=1}^{n} |xi - yi|)^{\frac{1}{p}} \tag{3}$$

3.4.4. Distancia de Hamming

Esta técnica es comúnmente empleada con vectores de valores booleanos o de caracteres, donde se identifican los lugares en los que los vectores no son iguales. Por esta razón, también se la conoce como *overlap metric* [11].

$$Distancia\ de\ Hamming = \ D_H = \left(\sum_{i=1}^{k} |xi - yi|\right) \tag{4}$$

3.5. Valor k

La variable *k* en el algoritmo KNN es un factor importante para clasificar un punto de datos específico. Esta variable determina la cantidad de vecinos que se evaluarán para determinar su categoría. Por ejemplo, cuando *k*=1, la clasificación se basará en la categoría del vecino más cercano. Seleccionar el valor adecuado de *k* puede ser un desafío, ya que valores bajos pueden resultar en un ajuste inadecuado y una alta variabilidad, mientras que valores altos pueden resultar en un sesgo alto y una variabilidad más baja. La elección óptima de *k* depende en gran medida de los datos, y aquellos con valores atípicos o ruido generalmente requieren valores de *k* más altos [11].

3.6. Descripción de código utilizando el algoritmo k-NN

La primera línea de código muestra la asignación de una matriz previamente creada en un archivo Excel, en donde se selecciona cuatro datos (edad, sexo, peso, talla) para obtener una matriz que servirá de entrenamiento al algoritmo.

```
trainingData  = xlsread('ObesityP.xlsx', 'Hoja1!A2:D161');
```

Posteriormente, se etiquetan cada uno de los registros en la matriz de datos de entrenamiento con una clase correspondiente, donde se utiliza el número 1 para indicar bajo peso, el 2 para peso normal, el 3 para sobrepeso y el 4 para obesidad.

```
classLabels = xlsread('ObesityP.xlsx', 'Hoja1!E2:E161');
```

Se generó una segunda matriz de datos de prueba que contenía los mismos cuatro atributos utilizados en el entrenamiento (peso, edad, talla y género) para evaluar la capacidad del modelo para predecir la obesidad en adolescentes.

La generación de esta matriz de datos de prueba es una práctica común en el desarrollo de modelos de aprendizaje automático, permite evaluar la capacidad del modelo para generalizar a nuevos datos de entrada que no fueron empleado en la etapa de entrenamiento. Utilizando los mismos atributos de entrada para ambas matrices de datos, permite medir la precisión de las predicciones realizadas por el modelo en datos no vistos anteriormente.

```
testData = xlsread('ObesityP.xlsx', 'Hoja1!J2:M41');
```

Después de realizar varias evaluaciones de los datos y la aplicación de la técnica de validación cruzada, se seleccionó un valor de 3 para la variable *k* en el algoritmo k-vecinos más cercanos (KNN). Se encontró que, para este conjunto de datos en particular, un valor de *k*=3 logró un buen equilibrio entre la precisión del modelo y la capacidad de generalización. Se realizar varias pruebas con varios valores de *k* y se evaluaron las métricas de rendimiento del modelo para cada uno. Se observó que valores de k demasiado bajos (*k*=1) tendían a sobreajustar los datos de entrenamiento, mientras que valores de *k* demasiado altos (*k*=5 o más) tendían a subajustar los datos y reducían la precisión del modelo. Por lo tanto, después de ser considerado cuidadosamente estos factores, se decidió que un valor de *k*=3 era el mejor ajuste para este conjunto de datos y para este algoritmo de clasificación en particular. Los datos de prueba se clasificaron utilizando el algoritmo KNN con la función previamente creada (*knnclassify*) y, finalmente, se imprimieron los resultados de la clasificación obtenidos.

```
k = 3;
predictedClasses = knnclassify(testData, trainingData, classLabels, k);
disp(predictedClasses);
```

3.7. Definición árbol de decisión

Un árbol de decisión es un modelo de aprendizaje automático supervisado que se utiliza para clasificar o realizar regresiones. Es una herramienta de aprendizaje no paramétrica que se construye recursivamente a partir de un conjunto de datos de entrenamiento. Se puede visualizar como una estructura de ramificación, en la que cada nodo representa una prueba sobre una característica, y cada hoja representa una predicción [14].

En una variedad de estudios previos de distintas áreas: en el sector salud, financiero, manufactura, social entre otros han demostrado la efectividad de los árboles de decisiones en la clasificación de problemas complejos y en la identificación de características relevantes. Por ejemplo, en [15] realizan un análisis comparativo entre los árboles de decisión y las redes neuronales como clasificadores llegando a la conclusión que los árboles de decisión proporcionan mejores resultados. Por otro lado, en [16] propuso una metodología para delimitar los tipos de cobertura vegetal en la subcuenca Quillcay. Los resultados que se obtuvieron demostraron la eficacia de los árboles de decisión en la clasificación de la cobertura vegetal, lo que respalda la relevancia de esta técnica en el estudio y gestión de los recursos naturales.

Los árboles de decisión utilizan una técnica de segmentación de datos basada en una serie de reglas de prueba que se aplican a cada característica de los datos de entrada. La idea detrás de esta técnica es dividir el espacio de entrada en una serie de regiones cada vez más pequeñas. El objetivo es que cada región contenga un número limitado de puntos de datos con la misma etiqueta de clase.

En cada nodo del árbol de decisión, se determina la característica a seleccionar utilizando una fórmula que se basa en una medida de impureza, como la entropía o la ganancia de información. La medida de impureza se utiliza para evaluar la calidad de las divisiones que se pueden realizar en los datos y para seleccionar la característica y el punto de corte que resultan en la división más "pura". El objetivo es dividir el espacio de entrada en regiones cada vez más pequeñas, donde cada región contiene un número limitado de puntos de datos con la misma etiqueta de clase.

3.8. Función fitctree

En MATLAB, la función *fitctree* es utilizada para entrenar árboles de decisión de clasificación mediante aprendizaje automático. Para ello, se debe ingresar como parámetros un conjunto de datos de entrenamiento y una

matriz de etiquetas de clase, obteniendo como resultado un modelo de árbol de decisión entrenado capaz de realizar predicciones en nuevos datos.

En la función *fitctree* de MATLAB, el objetivo principal es construir un modelo de árbol de decisión que pueda asignar las etiquetas de clase correctas a nuevos puntos de datos con la mayor precisión posible. Para lograr esto, la función utiliza técnicas de segmentación recursivas para dividir el espacio de entrada en regiones cada vez más pequeñas. De esta manera, se construye una estructura de ramificación que representa las reglas de prueba que se aplican a cada característica del conjunto de datos [17].

3.9. Descripción de código utilizando árbol de decisión

Lo primero que hacemos al igual que con el algoritmo de KNN definimos los datos de entrada y las etiquetas de clase correspondientes.

```
datos = xlsread('Obesity.xlsx', 'Hoja1!A2:D161');
clases = xlsread('Obesity.xlsx', 'Hoja1!E2:E161');
```

A continuación, se llama a la función *fitctree* con los datos de entrenamiento y las etiquetas de clase como entrada. La función ajusta un árbol de decisión a los datos de entrenamiento utilizando técnicas de segmentación recursivas basadas en medidas de impureza como la entropía o la ganancia de información. El árbol de decisión resultante se puede visualizar y analizar para comprender cómo el modelo toma decisiones y cómo se relacionan las diferentes características con las etiquetas de clase.

```
Varbol = fitctree(datos, clases, 'SplitCriterion', 'gdi');
```

Mediante la función *view* de Matlab se grafica el árbol resultante para tener una visualización del modelo (véase Figura 2). Esta visualización del árbol de decisión también puede ser útil para identificar áreas donde el modelo puede estar sobreajustando los datos de entrenamiento, lo que puede ser corregido mediante técnicas de poda del árbol.

```
view(Varbol, 'mode', 'graph');
```

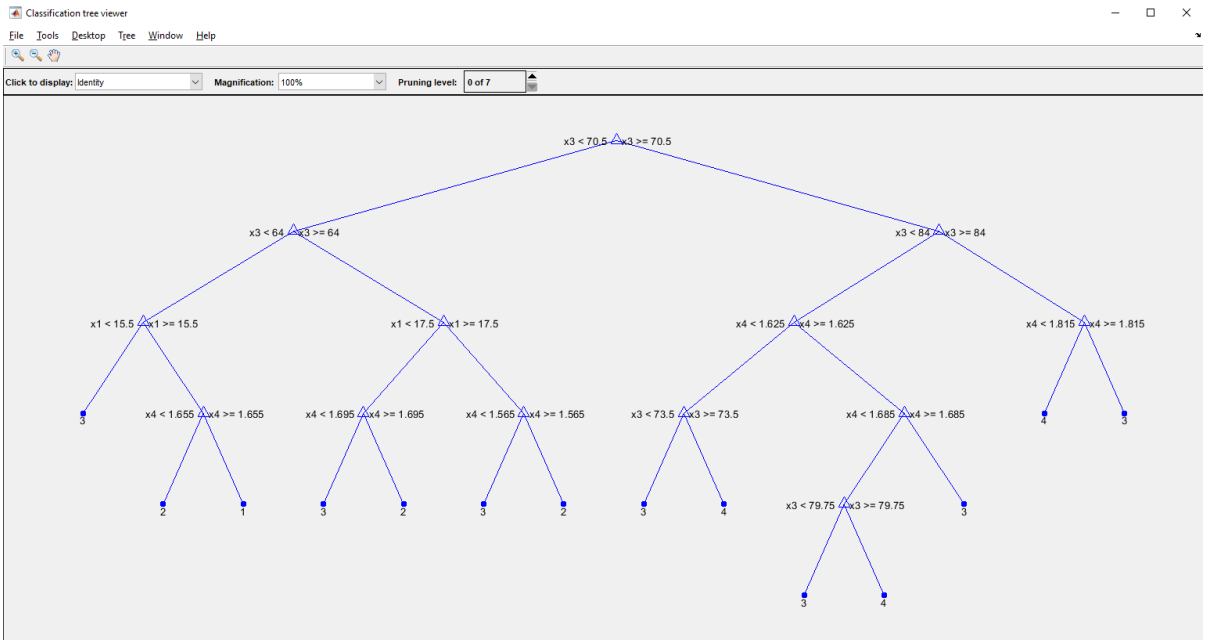

Figura 2. Gráfica del árbol de decisión con los datos de entrada proporcionados.

Una vez entrenado el modelo, se pueden utilizar los datos de prueba para evaluar la precisión del modelo y hacer predicciones sobre nuevas instancias de datos. La función *predict* se utiliza para predecir las etiquetas de clase de los datos de prueba utilizando el modelo entrenado, y luego se pueden calcular las métricas de rendimiento, como la precisión y el *recall*, para evaluar la calidad del modelo.

```
nuevo_caso = xlsread('Obesity.xlsx', 'Hoja1!J2:M41');
prediccion = predict(Varbol, nuevo_caso);
```

4. Resultados

Los resultados obtenidos se presentan en la Tabla 3, la cual compara los datos de la sección "Tipo peso" con los resultados obtenidos mediante el uso de los algoritmos KNN y Árbol de decisión. Se observan las siguientes precisiones: el algoritmo KNN obtuvo una precisión del 80%, ya que tuvo 8 errores de clasificación de un total de 40 datos. Por otro lado, el árbol de decisión mostró una precisión del 87.7%, con 5 errores de clasificación de los 40 datos totales. Estos resultados indican que el árbol de decisión logró clasificar con mayor precisión los datos de entrada en comparación con el algoritmo KNN. Además, el análisis del árbol de decisión permitió comprender cómo se tomaron las decisiones en cada nodo y cómo se relacionaron las diferentes características con las etiquetas de clase.

Tabla 3. Resultados del algoritmo KNN y árbol de decisión.

Edad	Sexo	Peso (kg)	Talla (cm)	Tipo peso	KNN	Árbol de decisión
17	2	67	1.57	3	3	3
15	1	52	1.5	2	2	3
16	1	50	1.49	2	2	2
17	2	70	1.66	3	3	3
15	2	65	1.56	3	3	3
15	1	56	1.5	3	2	3
15	2	71	1.69	3	3	3
15	1	82	1.68	4	4	4
16	2	60	1.6	2	2	2
16	1	69	1.65	3	3	3
16	2	79	1.7	3	4	3
16	2	72	1.62	3	3	3
16	1	68	1.66	3	3	3
16	1	65	1.6	3	3	3
16	1	55	1.64	2	2	2
16	2	90	1.63	4	4	4
16	2	88	1.65	4	4	4
16	1	58	1.6	2	2	2
16	2	78	1.67	4	3	3
16	2	80	1.69	4	3	3
16	2	85	1.71	4	4	4
17	2	82	1.69	4	3	3
17	2	79.5	1.7	3	4	3
17	1	60	1.62	2	2	2
17	2	80	1.55	4	4	4
17	2	75	1.7	3	3	3

17	2	79	1.6	4	4	4
17	2	88	1.6	4	4	4
17	2	90	1.66	4	4	4
17	1	47	1.65	2	2	2
15	2	90	1.65	4	4	4
17	2	82	1.69	4	3	3
16	1	70	1.66	3	3	3
15	1	74	1.74	3	3	3
16	2	92	1.76	4	4	4
15	2	75	1.62	4	3	4
16	2	90	1.63	4	4	4
17	2	50	1.64	2	2	2

Fuente: Elaboración propia.

5. Conclusiones

Después de realizar una comparativa en la clasificación de la obesidad en adolescentes, se ha observado que los árboles de decisión son más efectivos en comparación con el algoritmo de *k* vecinos más cercanos. Los árboles de decisión son una técnica de aprendizaje automático que permite realizar predicciones basadas en una serie de reglas y decisiones. Esta técnica es especialmente útil en la clasificación de la obesidad en adolescentes, ya que permite identificar las características más relevantes en la determinación de la obesidad.

Por otro lado, el algoritmo de *k* vecinos más cercanos es una técnica de clasificación que se basa en la identificación de los *k* vecinos más cercanos a una muestra dada. Sin embargo, este enfoque puede ser menos preciso en comparación con los árboles de decisión, ya que no tiene en cuenta todas las características relevantes en la determinación de la obesidad. En la comparativa realizada, se encontró que los árboles de decisión obtuvieron una precisión del 87.7%, mientras que el algoritmo k vecinos más cercanos presentó una precisión del 80%.

Los resultados obtenidos pueden deberse a varias razones. Una de ellas es que los árboles de decisión tienen la capacidad de identificar las características más importantes para la detección de la obesidad, lo que les permite tomar decisiones más precisas al clasificar nuevos datos. Además, tienen la capacidad de manejar relaciones no lineales entre las características y las clases. Los árboles de decisión pueden construir divisiones jerárquicas basadas en múltiples características, lo que les permite procesar relaciones complejas y no lineales en los datos. Debido a su estructura en forma de árbol, los árboles de decisión pueden manejar patrones más complejos y adaptarse mejor a la variabilidad de los datos, lo que podría haber contribuido a su mayor precisión en la clasificación de la obesidad en adolescentes. Por otro lado, el algoritmo del *k*-vecino más cercano se basa en la identificación de los *k*-vecinos más cercanos para una muestra dada sin considerar la importancia de cada característica en la detección de la obesidad. Además, los árboles de decisión pueden manejar datos faltantes o incompletos de manera más eficiente que el algoritmo del vecino más cercano, lo que puede contribuir a una mayor precisión en la clasificación de la obesidad adolescente.

El análisis de los resultados reveló dos características importantes de la clasificación de la obesidad adolescente. El primero es el índice de masa corporal (IMC), que está estrechamente relacionado con la clasificación y es crucial para la toma de decisiones en el modelo de árbol de decisiones. La otra característica es la medida de la cintura, que está estrechamente relacionada con la obesidad y juega un papel importante en la clasificación. Estos hallazgos resaltan la importancia de evaluar tanto el IMC como las medidas de la cintura cuando se trata la obesidad adolescente, ya que brindan una imagen más completa y precisa de su salud y riesgo de obesidad.

Estudios previos han demostrado la efectividad del IMC en la clasificación de problemas complejos y en la identificación de características relevantes. Por ejemplo, en un estudio realizado en [18], se compararon tres tipos de algoritmos evolutivos para determinar el porcentaje de la obesidad escolar en niños y adolescentes entre cinco y diecisiete en Brasil, El índice de masa corporal (IMC) fue utilizado como una de las variables en las reglas del algoritmo evaluado que obtuvo el mejor resultado en la clasificación de la obesidad en escolares. Además, en [19] se realizó un estudio para evaluar la fiabilidad IMC como medida de la composición corporal, especialmente en

personas mayores y jóvenes. Los resultados concluyeron que el IMC por sí solo no es suficiente. Por lo tanto, en este trabajo se optó por combinar el IMC con la medida de la circunferencia de la cintura, lo cual permitió obtener resultados más precisos y completos.

En resumen, los resultados de la comparativa sugieren que los árboles de decisión son más adecuados en la clasificación de la obesidad en adolescentes en comparación con el algoritmo de k vecinos más cercanos. Este hallazgo es importante, ya que puede ayudar a los profesionales de la salud a desarrollar estrategias más efectivas para prevenir y tratar la obesidad en esta población.

6. Agradecimientos

Agradecemos sinceramente a los alumnos del plantel 18 del CECYTE de Tehuantepec por su valiosa colaboración y por proporcionar la información necesaria para la obtención del conjunto de datos utilizado en este artículo.

7. Referencias

[1] Organización Panamericana de la Salud. (2023). *Prevención de la Obesidad.* Paho. https://www.paho.org/es/temas/prevencion-obesidad#:~:text=Si%20se%20examina%20%C3%BAnicamente%20la,31%25%20de%20las%20mujeres).

[2] Morales Márquez, L. E., Carrillo Ruiz, M., García Juárez, P., Colmenares Guillén, L. E. (2022). Determinación del riesgo de diabetes en México mediante un sistema difuso optimizado por recocido simulado. *Revista de Investigación en Tecnologías de la Información*, *10* (20), 130–144. https://doi.org/10.36825/RITI.10.20.011

[3] Pérez-de-Celis Herrero, C., Lara Muñoz, C., Somodevilla García, M. J., Pineda Torres, I. H., Colmenares Guillen, E. (2016). Estilo de Vida de los Estudiantes de Informática. *Revista de Investigación en Tecnologías de la Información*, *4* (8), 7–13. https://doi.org/10.36825/RITI.04.08.002

[4] Cardel, M. I., Atkinson, M. A., Taveras, E. M., Holm, J. C., Kelly, A. S. (2020). Tratamiento de la obesidad entre adolescentes: una revisión de la evidencia actual y las direcciones futuras. *JAMA Pediatrics*, *174* (6), 609–617. https://doi.org/10.1001/jamapediatrics.2020.0085

[5] Instituto Nacional de Salud Pública. (2022). *Encuesta Nacional de Salud y Nutrición 2021 sobre COVID-19*. https://www.insp.mx/resources/images/stories/2022/docs/220804_Ensa21_digital_4ago.pdf

[6] Alpaydin, E. (2019). *Introduction to machine learning* (3era Ed.). MIT Press.

[7] Delgado Huacallo, R. E., Ilachoque Hanccoccallo, C., Luque Sanabria, F., Paniura Huamani, J. M. (2022). Predicción del nivel de obesidad en personas usando el modelo de árbol de decisión. *Revista Innovación y Software*. https://n2t.net/ark:/42411/s9/a71

[8] Becerra Romero, N. E., Huayna Dueñas, A. M. (2022). Aplicación Web Basado en Minería de Datos usando la Técnica de Naive Bayes para la Predicción de la obesidad en edad infantil en los Hospitales Públicos de Lima. *Revista de investigación de Sistemas e Informática*, *14* (2), 89–98. https://doi.org/10.15381/risi.v14i2.23150

[9] Alva, L., Laria, J., Ibarra, S., Castán, J., Terán, J. (2020). Propuesta de un modelo difuso para determinar sobrepeso y obesidad en niños y adolescentes. *Revista chilena de nutrición*, *47* (4), 545-551. https://dx.doi.org/10.4067/S0717-75182020000400545

[10] Arenas Rodríguez, A., Torres Naira, C., Vizcarra Huyhua, F., Sulla Torres, J., Méndez Cornejo, J. (2016). Comparación de algoritmos evolutivos para la optimización en la clasificación de la obesidad en escolares. *UCMaule*, (51), 25-42. https://revistaucmaule.ucm.cl/article/view/15

[11] IBM. (2023). *What is the k-nearest neighbors algorithm?* https://www.ibm.com/topics/knn

[12] Cheng, X., Lin, S. Y., Liu, J., Liu, S., Zhang, J., Nie, P., Fuemmeler, B. F., Wang, Y., Xue, H. (2021). Does physical activity predict obesity—a machine learning and statistical method-based analysis. *International Journal of environmental research and public Health*, *18* (8), 1-11. https://doi.org/10.3390/ijerph18083966

[13] MathWorks. (2014). *knnclassification.m*. Archivo de software. https://la.mathworks.com/matlabcentral/fileexchange/47033-knnclassification-m

[14] Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C. J. (1984). *Classification and Regression Trees*. Routledge.

[15]Origel-Rivas, C. G., Lara, E. R., Barrera, I. A., Alejo-Eleuterio, R. (2020). Redes neuronales artificiales y árboles de decisión para la clasificación con datos categóricos. *Research in Computing Science*, *149* (8), 541-554. https://rcs.cic.ipn.mx/rcs/2020_149_8/Redes%20neuronales%20artificiales%20y%20arboles%20de%20decision%20para%20la%20clasificacion%20con%20datos%20categoricos.pdf

[16]Santiago Bazan, F., Mallqui Meza, H., Rios Recra, R. (2021). Mapeo de la cobertura vegetal en la subcuenca Quillcay (Ancash–Perú) con el clasificador de Árbol de decisiones. *Aporte Santiaguino*, *14* (1), 78-91. https://doi.org/10.32911/as.2021.v14.n1.761

[17]MathWorks. (2023.). *fitctree (MATLAB)*. https://la.mathworks.com/help/stats/fitctree.html?s_tid=srchtitle_fitctree_1

[18]Arenas Rodríguez, A. C., Torres Naira, C. A., Vizcarra Huyhua, F. M., Sulla-Torres, J., Méndez-Cornejo, J. (2016). Comparación de algoritmos evolutivos para la optimización en la clasificación de la obesidad en escolares. *Revista Académica UCMaule*, (51), 25-42. https://revistaucmaule.ucm.cl/article/view/15

[19]Grecco Dos Santos, R. R., Carra Forte, G., Mundstock, E., Azambuja Amaral, M., Gomes da Silveira, C., Chaves Amantéa, F., Frota Variani, J., Booij, L., Mattiello, R. (2020). Body composition parameters can better predict body size dissatisfaction than body mass index in children and adolescents. *Eat Weight Disord, 25*, 1197–1203. https://doi.org/10.1007/s40519-019-00750-4

