Ensayo de investigación

Análisis con técnicas de clustering de los hábitos alimenticios de estudiantes de la Universidad Tecnológica de la Mixteca, Oaxaca, México

Analysis using clustering techniques of the eating habits of students at the Technological University of the Mixteca, Oaxaca, Mexico

Antonio Orantes-Molina[1]*, Rosebet Miranda-Luna[1]

[1]Universidad Tecnológica de la Mixteca

Autor de correspondencia:
*tonito@mixteco.utm.mx

Recibido: 20-01-2025 Aceptado: 29-04-2026 (Artículo Arbitrado)

Resumen

El algoritmo de clustering no supervisado denominado K-Prototype fue aplicado a una base de datos mixta acerca de los hábitos alimenticios de los estudiantes de la Universidad Tecnológica de la Mixteca (UTM), con el fin de determinar si la percepción de los alumnos acerca de su alimentación es acertada o si por el contrario su percepción es errónea. Se aplicó una encuesta a 121 estudiantes que incluye variables cuantitativas y categóricas. Las variables que contribuyeron al análisis son: cantidad de comidas al día, género, tipos de comida chatarra consumidas, consumo de alcohol u otras drogas, percepción de salud, edad, cantidad económica percibida mensual. Se aplicó la metodología CRISP-DM, el cual es un método probado para orientar los trabajos en análisis de datos. Las fases de la metodología CRISP-DM son: comprensión del proceso, comprensión de los datos, preparación de los datos, modelado, evaluación e implementación. Para la fase de modelado, se aplicó el algoritmo de clustering K-Prototype porque permite emplear datos categóricos y numéricos simultáneamente. Los resultados determinaron que las mujeres con hábitos saludables, se consideran poco saludables, mientras que la población masculina se considera saludable, no obstante, presentan hábitos poco saludables.

Palabras clave: Disimilaridad, K-Means, K-Modes, K-Prototype, Metodología CRISP-DM.

Abstract

The K-Prototype unsupervised clustering algorithm was applied to a mixed dataset examining students' eating habits at the Technological University of the Mixteca (UTM) to assess the accuracy of students' self-perceptions regarding their diets. A survey was conducted with 121 students, incorporating both quantitative and qualitative variables. Variables analyzed included number of meals per day, gender, junk food consumption, alcohol or drug use, health perception, age, and perceived monthly income. The CRISP-DM methodology, a widely recognized framework for data analysis, guided the process through the phases of process. The phases of the CRISP-DM methodology are: understanding, data understanding, data preparation, modeling, evaluation, and deployment. The K-Prototype algorithm was selected for the modeling phase due to its capacity to handle both categorical and numerical data simultaneously. Findings indicate that women with healthy habits are consider themselves as unhealthy, whereas men with unhealthy habits are classified as healthy.

Keywords: Dissimilarity, K-Means, K-Modes, K-Prototype, CRISP-DM methodology.

Introducción

El presente trabajo tiene como objetivo determinar el número de alumnos que son realmente conscientes de lo saludable que son sus hábitos alimenticios, para ello el análisis de los datos se realiza con base a las técnicas de clasificación no supervisada denominada clustering, ya que agrupan distintos objetos según sus similitudes en un solo grupo o "clúster" que describe la estructura del conjunto de datos (Ordoñez, Ordoñez, Méndez & Ordoñez, 2020). Estas técnicas son muy utilizadas en áreas donde se requiera hacer una clasificación, como lo son el área médica, biológica, clasificación de productos, la detección de anomalías, aplicaciones empresariales, entre otros (Guojun, Chaoqu & Jianhong, 2007).

Las técnicas de clustering se pueden realizar por medio de distintos algoritmos, entre ellos el K-Prototypes, el cual es una combinación de K-Means y K-Modes, siendo el algoritmo K-Means (Morissette y Chartier, 2013) uno de los más reconocidos por sus facilidades de empleo. El algoritmo K-Means puede emplear diversos tipos de métricas de distancia como son la distancia Euclidiana, Chebyshev, Manhattan, Minkowsky, Mahalanobis, Similitud de Coseno, Kullback Leibler, los cuales se utilizan para determinar el grupo o clúster al que pertenece cada dato. Cabe destacar que este algoritmo solo acepta datos numéricos para su análisis, sin embargo, los datos recopilados en el mundo real no presentan únicamente valores numéricos sino también valores categóricos por lo que los métodos de clustering numéricos no serán útiles para el análisis de este proyecto. Para analizar los datos del presente trabajo es necesario utilizar un algoritmo distinto, uno que sea capaz de trabajar con ambos tipos de datos de manera simultánea, tanto numéricos como categóricos. El método a ocupar es la unión de dos algoritmos clásicos: el K-Means (Morissette, 2013) y el K-modes (Huang, 1998) denominado K-Prototype.

La técnica de clustering K-Prototype es un algoritmo de agrupamiento que permite el análisis de datos numéricos y categóricos mediante la disimilaridad que presentan los datos y al igual que otros algoritmos, el algoritmo también se encuentra incluido en la biblioteca de funciones en los lenguajes de programación de alto nivel como Octave, Python o R-Studio. En este trabajo se hará uso de la herramienta matemática Octave que es un software de uso libre para analizar los datos obtenidos por medio de una encuesta aplicada a los estudiantes de la Universidad Tecnológica de la Mixteca (UTM) ubicada en la Ciudad de Huajuapan de León, Oaxaca, México y de esta forma obtener el número de alumnos que realmente son conscientes de sus hábitos alimenticios y el número de los que no lo son, tomando en cuenta factores como cantidad de comidas al día, tipo de comida chatarra preferida, consumo de alcohol u otras sustancias nocivas para la salud y sus ingresos económicos mensuales.

Metodología

Los datos a analizar fueron obtenidos de una encuesta realizada por la M.A.N. Martha Angélica Ruiz González (perteneciente al Instituto de Ciencias Sociales y Humanidades de la UTM) a 121 alumnos de la comunidad universitaria de la Universidad Tecnológica de la Mixteca como parte de un proyecto interno de investigación. En este proyecto se aplicó una encuesta de 18 preguntas de las cuales 7 variables fueron seleccionadas para el análisis: 3 cuantitativas y 4 categóricas. Estas son: los ingresos económicos, la edad, el género, comida chatarra más consumida, uso de drogas y alcohol, la percepción acerca de sus hábitos alimenticios y el número de alimentos consumidos al día.

Para la gestión de proyectos que involucran el análisis de datos se pueden utilizar varios tipos de metodologías para la minería de datos. Huber, Wiemer, Schneider e Ihlenfeldt (2019) señalan que la minería de datos es introducida para tener una visión más allá de la aplicación de algoritmos estadísticos o aprendizaje automático, con el objetivo de determinar las preguntas más relevantes mediante el tratamiento de los datos generando nueva información.

La metodología CRISP-DM (*Cross-Industry Standard Process for Data Mining*) contiene las fases de proyecto que se muestran en la Figura 1, notando que algunas fases de la metodología son bidireccionales, lo que significa que estas fases pueden recurrir a las etapas anteriores de ser necesario. En este trabajo, las dos últimas fases: evaluación e implementación, se presentan en el rubro de los resultados. A continuación, se detallan las primeras 4 fases de esta metodología (Huber et al., 2019).

Fase 1: Comprensión del proceso

Se realizó un análisis de los ámbitos en los que fue enfocada la encuesta sobre hábitos alimenticios de la comunidad universitaria. Se entrevistó a la profesora-investigadora M.A.N Martha Angélica Ruiz González sobre la finalidad de la encuesta, y a partir de ello se determinaron los parámetros del proyecto.

Fase 2: Comprensión de los datos

Se realizó un estudio detallado con los datos obtenidos, evaluando la lógica de las preguntas y respuestas, así como la relación existente entre las preguntas realizadas, seleccionando las variables que se utilizan a lo largo de la investigación. Se identificaron las variables categóricas y las variables cuantitativas, los cuales se muestran en la Tabla 1.

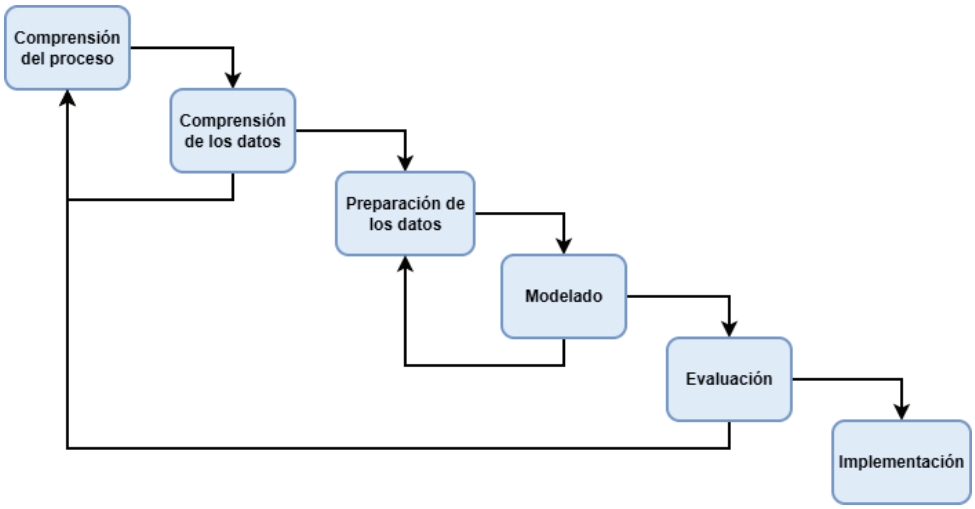

Figura 1. Metodología CRISP-DM.
Fuente: Elaboración propia.

Fase 3: Preparación de los datos

En esta fase se realizó la selección y eliminación de los datos atípicos o outliers que se presentan en la base de datos. Los datos atípicos son una parte importante para el estudio de los datos, Da Vila (2020). Estos Indican que son datos anómalos en comparación a los demás. En el contexto del aprendizaje automático, pueden representar un problema significativo por las siguientes razones:

Distorsión de modelos: Los modelos de aprendizaje automático, especialmente los que se basan en distancia como el k-NN o los modelos de regresión, pueden ser muy sensibles a los outliers. Estos valores atípicos pueden sesgar el modelo, afectando la precisión y capacidad de generalización.

Ruido y errores: Los datos atípicos a menudo pueden ser ruido o errores, lo que puede llevar a que el modelo aprenda patrones irrelevantes o incorrectos. Esto puede disminuir la calidad del modelo y llevar a un sobreajuste.

Tabla 1. Variables cuantitativas y categóricas de la base de datos.

Variable	Tipo de variable
Ingresos	Cuantitativa
Edad	Cuantitativa
Cantidad de comidas al día	Cuantitativa
Género	Categórica
¿Cuál es la comida chatarra que más consumes?	Categórica
¿Estás acostumbrado a las drogas y al alcohol? ¿si es así, cuáles?	Categórica
¿Qué tan saludable te consideras?	Categórica

Fuente: Elaboración propia.

Impacto en la evaluación de modelos: En tareas de clasificación, pueden afectar las métricas de evaluación, ya que pueden ser clasificados incorrectamente con mayor frecuencia.

De 121 alumnos entrevistados, se eliminó uno debido a la ausencia de información. En la Tabla 2 se presenta el total de alumnos para cada una de las variables categóricas de la base de datos. Se muestra en la primera columna las variables categóricas relevantes para nuestro estudio obtenidos de la encuesta realizada. En la segunda columna se muestran las diferentes modalidades correspondientes a cada variable categórica y en la última columna, el total de alumnos contabilizados en cada rubro.

Para mitigar la problemática de los datos atípicos, se pueden utilizar técnicas como la normalización de los datos.

La normalización de los datos es un paso importante para su procesamiento. Morante (2018) señala que para optimizar el aprendizaje de las máquinas es necesario comprimir o extender los valores de las variables de entrada con la finalidad de obtener un rango definido. También auxilia para eliminar la influencia de la dimensión y magnitud en cada una de las características además de que permite mejorar la velocidad de convergencia en este tipo de algoritmos, ya que sólo trabaja con valores comprendidos entre 0 y 1. Existen diversos métodos de normalización, cada uno con sus ventajas y desventajas, para este proyecto se utilizó la mostrada en la ecuación (1).

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}(max - min) + min \qquad (1)$$

Donde X_{norm} es el dato normalizado, X es el dato de entrada, X_{max} es el dato con mayor valor, X_{min} es el dato con menor valor, *max* y *min* son los valores máximo y mínimo del rango al cual deseamos normalizar (en este caso 1 y 0, respectivamente).

Fase 4: Modelado

Las técnicas de aprendizaje no supervisado como las técnicas de clustering tienen como objetivo encontrar patrones o agrupamientos naturales en una base de datos. Esto lo hacen a través de una métrica de distancia que les permite determinar que un dato está cerca del otro.

En esta fase se presentan 4 métricas de distancia de las más utilizadas dentro de las técnicas de clustering que son: Euclidiana, Chebyshev, Manhattan y Minkowsky.

La distancia Euclidiana es la distancia más corta entre dos puntos en cualquier dimensión *p*, su fórmula es la raíz cuadrada de la suma de cuadrados de la diferencia entre los puntos P_1 y P_2 (Fórmula de Pitágoras).

Para los puntos $P_1(x_1, y_1)$ y $P2(x_2, y_2)$ de dos dimensiones, la distancia Euclidiana se presenta en la Figura 2.

En dos dimensiones, la distancia Euclidiana se calcula con la ecuación (2).

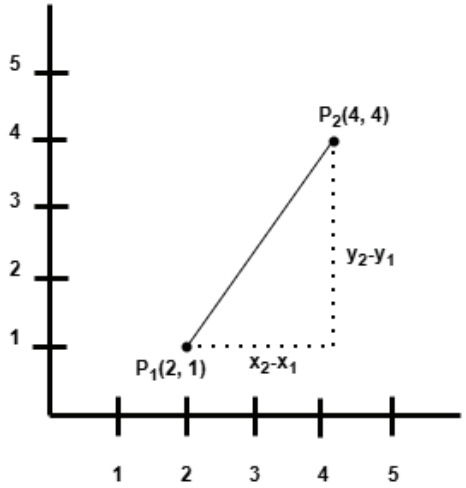

Figura 2. Distancia Euclidiana entre dos puntos.
Fuente: Elaboración propia.

En el caso de *p* dimensiones, la ecuación (3) permite el cálculo de la distancia.

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \qquad (2)$$

$$D(\bar{x}, \bar{y}) = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2} \qquad (3)$$

La distancia Chebyshev se calcula con la ecuación (4). Se le conoce también como la distancia del tablero de ajedrez porque es la distancia que recorrería la pieza del rey entre dos puntos del tablero y se define como la mayor de las diferencias entre sus dimensiones.

Tabla 2. Contabilización de las modalidades de los datos categóricos.

Variable Categórica	Modalidad	Cantidad
Género	1. Mujer 2. Hombre 3. Prefiero no decirlo	59 61 0
¿Cuál es la comida chatarra que más consumes?	1. Sabritas/Frituras 2. Golosinas (paletas, chicles, chocolate, etc.) 3.Galletas. 4.Comida rápida (hamburguesas, hot-dog, pizza, etc.) 5.Pasteles empaquetados (mantecadas, nito, bigote, panqué) 6.Refrescos y bebidas embotelladas.	24 26 20 36 4 10
¿Estás acostumbrado al alcohol y otras drogas?	1. Sólo drogas. 2. Sólo alcohol. 3. Ambas. 4. Ninguna.	2 26 2 90
¿Qué tan saludable te consideras?	1. Poco saludable. 2. Saludable. 3. Muy saludable.	32 77 11

Fuente: Elaboración propia.

$$D_{Chebyshev}(\bar{x},\bar{y}) = max_i\left(\left|x_i - y_i\right|\right) \quad (4)$$

Para los datos bidimensionales de la Figura 2, $d = max(|x_1\text{-}x_2|, |y_1\text{-}y_2|) = max(2, 3) = 3$.

La distancia Manhattan, conocida como la distancia con ángulos rectos, por definición, contempla el camino más corto de un punto a otro considerando una curva en ángulo recto. La distancia Manhattan se calcula con la ecuación (5), la cual nos dice que la distancia entre dos puntos es la suma de las diferencias absolutas de sus coordenadas.

$$D(\bar{x},\bar{y}) = \sum_{i=1}^{p}\left|x_i - y_i\right| \quad (5)$$

Para los puntos en dimensión 2 de la Figura 2, $d = |x_1\text{-}x_2| + |y_1\text{-}y_2| = |2\text{-}4| + |1\text{-}4| = 5$.

La distancia Minkowsky puede considerarse una generalización de las distancias Euclideana y Manhattan y está definida por la ecuación (6).

$$D(\bar{x},\bar{y}) = \left(\sum_{i=1}^{p}\left|x_i - y_i\right|^r\right)^{\frac{1}{r}} \quad (6)$$

Obsérvese que para $r=1$, la expresión coincide con la distancia Manhattan y para $r=2$, coincide con la distancia Euclidiana.

En términos generales, determinar cuál es la métrica de distancia más adecuada está en función tanto del conjunto de datos como la técnica de clusterizado que se utiliza. Normalmente se realiza experimentación con diversas métricas para obtener la más adecuada.

Para obtener el modelo del proceso se utilizó el algoritmo K-Prototypes, el cual es una combinación de los algoritmos K-Means y K-Modes. De acuerdo con Morissette y Chartier (2013) el algoritmo K-Means es una agrupación inicial para algoritmos más complejos. K-Means es un algoritmo de clasificación no supervisada que agrupa objetos en "*k*" grupos basándose en sus características. Se realiza minimizando la suma de distancias entre cada objeto y el prototipo de su grupo o centroide. Es necesario definir el número de clúster (agrupaciones) desde un inicio. Para definir el número óptimo de *k*, se emplean diversas técnicas como el codo de Jambú, coeficiente de Silouhette, índice Davies-Bouldin o el criterio de la inercia. El algoritmo puede tomar diferentes criterios de paro como son: los individuos ya no son reasignados a otro clúster, el valor de la inercia Intra-Clase es menor a un valor de tolerancia dado o el número máximo de iteraciones se cumple. Para llevar con éxito este algoritmo se tienen los siguientes pasos:

1. Definir el número de clústeres.
2. Establecer los K-Clúster en el espacio de datos.
3. Asignar datos al clúster más cercano.
4. Se actualizan los prototipos o centroides tomando el promedio de las coordenadas de los datos que pertenecen a cada clúster.
5. Se repiten los pasos 3 y 4 hasta que se cumpla algún criterio de paro.

Para Bonthu (2024) la técnica de K-Modes es un algoritmo de aprendizaje automático no supervisado utilizado para agrupar variables categóricas. El algoritmo se puede reducir en los siguientes pasos (Cerón, 2018):

1. Deleccionar la cantidad K de clústeres, asignando un dato al azar como clúster.
2. Calcular las diferencias y asignar cada elemento a su grupo más cercano, comparando iterativamente los puntos de datos del conglomerado con cada una de los prototipos. Los puntos de datos similares dan el valor 0, los puntos diferentes dan un valor 1.
   Dados dos elementos $x_i = (c_{i1}, c_{i2}, \ldots, c_{ip})$ y $x_j = (c_{j1}, c_{j2}, \ldots, c_{jp})$ diremos que entre ellos hay una discrepancia en la *k*-ésima variable si $i \neq j$ la cual se representa mediante la métrica discreta de la ecuación (7).

$$\delta(c_{ik}, c_{jk}) = \begin{cases} 0 & \text{si } i = j \\ 1 & \text{si } i \neq j \end{cases} \quad (7)$$

Para todo par de elementos x_i, x_j se define la disimilaridad como lo indica la ecuación (8).

$$d_{ij} = d(\bar{x}_i, \bar{y}_j) = \sum_{r=1}^{p}\delta(c_{ir}, c_{jr}) \quad (8)$$

3. Se actualizan los clústeres de acuerdo a la moda en cada categoría.
4. Se repetirán los pasos 2 y 3 hasta que se cumpla algún criterio de paro.

El algoritmo K-Prototype propuesto por Huang (1998) es la integración de los algoritmos K-Means y K-Modes. K-Prototype es capaz de manejar datos de tipo mixto (numéricos y categóricos) de manera simultánea, los cuales predominan en las bases de datos del mundo real. La medida de disimilaridad que utiliza se obtiene tomando en cuenta las medidas de distancia del algoritmo de K-Means y la medida de disimilaridad del K-Modes unidos mediante la variable γ la cual se encarga de evitar alguna influencia por parte de algún tipo de dato (Zhezue, 1997).

$$d(X,Y)=\sum_{j=1}^{p}\left(x_j-y_j\right)^2+\gamma\sum_{j=1}^{p}\delta(x_j,y_j) \qquad (9)$$

El primer término de la ecuación (9) representa la distancia Euclidiana al cuadrado mientras que el segundo término representa la medida de coincidencia simple de los datos categóricos. El peso γ normalmente toma el valor de 1 cuando el rango de las variables categóricas y numéricas son muy cercanas entre sí. La elección de γ depende de la distribución de atributos numéricos. En términos generales, γ se relaciona con la desviación estándar promedio σ de los atributos numéricos. En la práctica, σ puede usarse como una guía para determinar γ. Según Huang (1998) un valor γ adecuado se encuentra entre $1/3\sigma$ y $2/3\sigma$ para los conjuntos de datos.

Los pasos para desarrollar el algoritmo son los siguientes:

Requiere: k (el número de clústeres)

1. Seleccionar k prototipos iniciales desde la base de datos, uno para cada clúster.
2. Asignar cada objeto de la base de datos a un clúster cuyo prototipo es el más cercano a él de acuerdo a la medida de disimilaridad.
3. Actualizar el prototipo del clúster después de cada asignación.

Repetir:

4. Volver a probar la similaridad entre cada objeto y el prototipo: si un objeto es encontrado que es más perteneciente a otro prototipo en lugar del actual, reasignar el objeto al clúster más cercano.
5. Actualizar el prototipo de ambos clústeres.
6. El algoritmo termina hasta que no haya más cambios en los clústeres.

Resultados

A continuación, se muestran los resultados obtenidos de la aplicación del algoritmo de clusterización K-Prototype a la base de datos mixta de la Tabla 1. Se realizó una comparación de los resultados de datos sin normalizar y con los datos normalizados para observar las diferencias e influencias entre ellas.

Aplicando el algoritmo K-Prototype a los datos sin normalizar, se obtuvieron los prototipos descritos en la Tabla 3. La distribución de la población en cada clúster se muestra en la Figura 3.

De la distribución de la población total se tienen 3 clústeres de interés, de los cuales se puede deducir lo siguiente:

1. En el clúster 2, se concentra el 35.5% de la población la cual posee un ingreso medio de $1861 y cuya edad promedio se sitúa en los 21 años. En este clúster se concentra la población femenina cuya percepción de salud considera saludable sin consumo de drogas o alcohol cumpliendo con 3 comidas al día, sin embargo, el consumo de la comida chatarra se centra más en la comida rápida.
2. El clúster 6 concentra el 18.7% de la población de 21 años de media, femenina, sin consumo de drogas o alcohol y con una percepción de salud alimenticia considerada como saludable. El único problema que presenta los miembros de este clúster es de índole económico, pues presenta la menor percepción económica lo que es un causante de la baja cantidad de comidas al día que presenta.
3. El clúster 4 presenta la mayor cantidad percibida mensual con una media de $5500 y una edad media de 23 años. La percepción de salud alimenticia de este clúster también es considerada como saludable, sin embargo, presenta consumo de alcohol en la población masculina y una preferencia a la comida chatarra.

Los prototipos obtenidos en la aplicación del algoritmo para los datos normalizados se presentan en la Tabla 4, dando una distribución de la población como se muestra en la Figura 4.

En comparación al análisis con los datos sin normalizar, existe una distribución un tanto más homogéneo de los datos en todos los clústeres. Normalizando los datos, podemos determinar lo siguiente:

Tabla 3. Prototipos finales de la base de datos sin normalizar (Ver Tabla 2 para identificar la modalidad).

Número de clúster	Cantidad de comidas al día	Género	Consumo de comida chatarra	Consumo de alcohol o drogas	Percepción de salud	Edad promedio	Cantidad percibida mensual (MXN)
1	3	2	2	4	2	23.64	4000
2	3	1	4	4	2	21.34	1861.6
3	2	1	1	4	2	21.31	3062.5
4	3	2	4	2	2	23.5	5500
5	3	1	4	4	2	22.4	2446.7
6	2	1	2	4	2	21.1	860

Fuente: Elaboración propia.

Tabla 4. Prototipos finales de la base de datos normalizados.

Número de clúster	Cantidad de comidas al día	Género	Consumo de comida chatarra	Consumo de alcohol o drogas	Percepción de salud	Edad promedio	Cantidad percibida mensual (MXN)
1	2	2	4	4	2	0.509	0.27054
2	3	2	4	2	2	0.611	0.41228
3	3	1	3	4	1	0.363	0.2386
4	3	2	4	4	2	0.5	0.66374
5	2	1	2	4	2	0.378	0.33474
6	3	1	1	4	2	0.167	0.25185

Fuente: Elaboración propia.

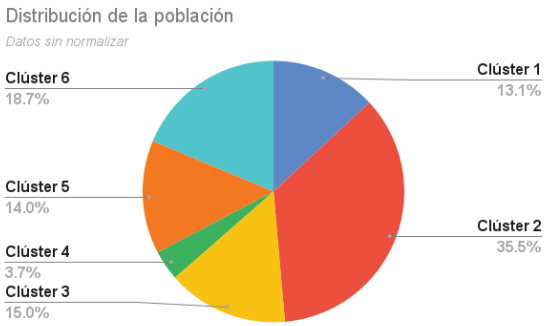

Figura 3. Distribución de la población por clúster de los datos sin normalizar. **Fuente:** Elaboración propia.

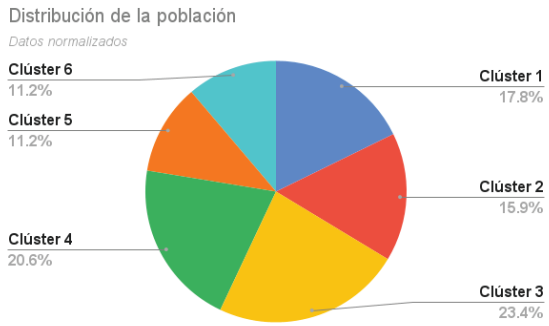

Figura 4. Distribución de la población por clúster de los datos normalizados. **Fuente:** Elaboración propia.

1. El clúster 3 presenta el porcentaje de población más grande con una media de edad de 21 años y género femenino. Presentan ingresos medios de $1660 y una cantidad de alimentos al día de 3 veces y cuyo consumo de comida chatarra se centra en galletas, aunque su percepción de salud alimenticia se encuentra como poco saludable.
2. El clúster 4 posee la percepción económica más alta con una media de $4083 con una población masculina de 22 años en promedio. Este clúster considera su salud alimenticia como saludable presentando un consumo de comida chatarra centrada en la comida rápida y con 3 alimentos al día.
3. En el clúster 2 se centra la población masculina de 23 años con una percepción mensual media de $2650. Esta población considera su salud alimenticia como saludable, pero presenta un consumo de alcohol, así como una preferencia a la comida chatarra.

Como puede observarse, existe una diferencia marcada entre los clústeres formados de los datos normalizados con respecto a los datos no normalizados, pues existe una alta influencia en los datos sin normalizar del rango de valores de las variables.

Conclusiones

El análisis de datos por medio de algoritmos de clusterización permite observar resultados agrupados por características similares, en este caso de aplicación, permite comparar las respuestas de los alumnos con respecto a su definición de vida saludable y lo que reflejan sus hábitos alimenticios.

En los resultados se observó que la gran mayoría de clústeres presentan una negación en el consumo de alcohol, de igual forma existe una discrepancia entre la percepción de su salud y sus hábitos alimenticios, debido a que la mayoría de los encuestados consideran que su alimentación es saludable, sin embargo, tienen hábitos como comer 2 veces al día, el consumo de alcohol y comida rápida, demostrando con ello, que existe una falsa percepción en cuanto a la alimentación saludable.

Al generar agrupaciones los factores como la edad, el género e ingresos económicos, se pueden analizar como un conjunto, sin necesidad de realizar gráficas independientes, los datos categóricos no necesitan ser evaluados independientemente, por el contrario, se consideran características de los grupos permitiendo caracterizarlos ya que, por ejemplo, un grupo de hombres de 22 años puede dividirse de acuerdo a la cantidad de veces que comen al día y su consumo de alcohol, por lo que la aplicación del algoritmo K-Prototype permite unificar de manera exitosa el análisis de datos categóricos y numéricos.

Respecto a la comparación realizada entre datos normalizados y no normalizados, se observaron cambios en cuanto a los prototipos, pues existe una desviación debido a la magnitud de los datos y la forma de tratar dichos datos. Al realizar una comparación de los prototipos uno a uno, se observó mayor uniformidad en la repartición de los clústeres utilizando los datos normalizados, contrario a los datos sin normalizar que presentan una distribución menos homogénea y menos significativa de las características en cada grupo dando como resultado una mala interpretación de la clasificación.

Referencias

Bonthu, H. (2024). *KModes Clustering Algorithm for Categorical data*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/06/kmodes-clustering-algorithm-for-categorical-data/.

Cerón, F. (2018). *Análisis de algoritmos de clustering para datos categóricos* [Tesis de grado, Universidad de los Andes]. Repositorio institucional de la Universidad de los Andes. https://repositorio.uniandes.edu.co/handle/1992/39125?show=full.

Da Vila, S. (2020). *Detección de outliers en grandes bases de datos*. [Tesis de maestría, Universidad de Santiago de Compostela]. Recuperado de: http://eio.usc.es/pub/mte/descargas/ProyectosFinMaster/Proyecto_1780.pdf.

Guojun, G., Chaoqun, M. & Jianhong, W. (2007). Data clustering Theory, Algorithms, and Applications. Philadelphia, Pennsylvania. Siam.

Huang, Z. (1998). Extensions to the K-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Mining and Knowledge Discovery*, vol. 2, 283–304.

Huang, Z. (1997). *Clustering large data sets with mixed numeric and Categorical values*. CSIRO Mathematical and Information Sciences GPO Box 664 Canberra ACT 2601, AUSTRALIA.

Huber, S., Wiemer, H., Schneider, D., & Ihlenfeldt, S. (2019). DMME: Data mining methodology for engineering applications a holistic extension to the CRISP-DM model. ELSEVIER. Vol. 79, 403-408. Recuperado de: https://www.sciencedirect.com/science/article/pii/S2212827119302239

Morante, S. (2018). *Precauciones a la hora de normalizar datos en Data Science - Think Big Empresas*. Think Big. https://empresas.blogthinkbig.com/precauciones-la-hora-de-normalizar/

Morissette, L., y Chartier, S. (2013). The k-means clustering technique: General considerations and implementation in Mathematica. *Tutorials in Quantitative Methods for Psychology*, vol. 9(1), 15-24.

Ordoñez, C., Ordoñez, J., Méndez C. & Ordoñez, H. (2020). Evaluación e implementación de técnicas de clustering para un sistema de recuperación de documentos judiciales. *Revista Ibérica de Sistemas e Tecnologias de Informação*. E33, 141–151.