Aplicación del algoritmo K-Prototypes a un conjunto de datos relacionados a la estatura y peso de los estudiantes de la Universidad Tecnológica de la Mixteca

Application of the K-Prototypes algorithm to a set of data related to the height and weight of the students of the Technological University of the Mixteca

García-Rivera B.

*Instituto de electrónica y mecatrónica  
Universidad tecnológica de la mixteca  
E-mail: garb981202@gs.utm.mx*

Lucas-Sánchez L.

*Instituto de electrónica y mecatrónica  
Universidad tecnológica de la mixteca  
E-mail: lusl991026@gs.utm.mx*

Quiroz-Bautista U.

*Instituto de electrónica y mecatrónica  
Universidad tecnológica de la mixteca  
E-mail: qubj980830@gs.utm.mx*

Reyes-López D.

*Instituto de electrónica y mecatrónica  
Universidad tecnológica de la mixteca  
E-mail: reld981213@gs.utm.mx*

Vásquez-Cruz O.

*Instituto de electrónica y mecatrónica  
Universidad tecnológica de la mixteca  
E-mail: vaco990811@gs.utm.mx*

Resumen

En este trabajo se analizan los resultados obtenidos de una encuesta que relaciona la edad, peso, altura, alimentación y ejercicio dirigida a los alumnos de la Universidad Tecnológica de la Mixteca (UTM) con el objetivo de interpretar el impacto que genera hacia la salud y bienestar de los estudiantes el cursar sus estudios durante todo el trayecto de la universidad.

La metodología utilizada para este trabajo es la siguiente: inicia al realizar una encuesta a diferentes semestres que actualmente se encuentran activos en la UTM, esta encuesta se centra en los aspectos mencionados con anterioridad; una vez con los datos recolectados se procede al tratamiento de la información a través del algoritmo denominado K-Prototype, este algoritmo fue implementado en lenguaje de programación Python y se encarga del agrupamiento de los datos de acuerdo a la similitud de los mismos; finalmente se brinda una interpretación de los datos de acuerdo con los grupos formados por el algoritmo terminando con esto el proceso.

Los resultados que se han obtenido con base en la muestra realizada arrojan un desbalance en el desarrollo de los estudiantes entre su vida académica y salud física que se muestra escalonada conforme se pasa el tiempo en la universidad.

Este trabajo está delimitado a los aspectos del cuidado de la salud básicos como es la alimentación y actividad física que llevan a cabo los estudiantes de acuerdo a la edad y carrera que se cursa, sin tomar en cuenta otros factores como el económico, diferentes tipos de trastornos o genética. La finalidad de este estudio es poder brindar un panorama general de la situación actual de los estudiantes además la información recabada junto con el algoritmo utilizado pueden servir de guía para la prevención de enfermedades dentro del alumnado.

***Palabras clave – Algoritmo K-Prototype, UTM, Estatura, Peso, Estudiantes.***

Abstract

In this work, the results obtained from a survey that relates age, weight, height, diet and exercise addressed to the students of the Technological University of the Mixteca (UTM) are analyzed with the objective of interpreting the impact that it generates towards health and welfare of students to pursue their studies throughout the course of the university.

The methodology used for this work is as follows: it begins by conducting a survey of different semesters that are currently active in the UTM, this survey focuses on the aspects mentioned above; Once the data is collected, the information is processed through the algorithm called K-Prototype, this algorithm was implemented in the Python programming language and is responsible for grouping the data according to their similarity; Finally, an interpretation of the data is provided according to the groups formed by the algorithm, thus ending the process.

The results that have been obtained based on the sample carried out show an imbalance in the development of the students between their academic life and physical health that is shown to be staggered as time passes in the university.

This work is limited to basic health care aspects such as nutrition and physical activity carried out by students according to their age and career, without taking into account other factors such as economics, different types of disorders or genetics. The purpose of this study is to be able to provide a general overview of the current situation of the students, in addition to the information collected together with the algorithm used, which can serve as a guide for the prevention of diseases within the student body.

***Keywords – Algorithm K-Prototype, UTM, Height, Weight, Students.***

Introducción

El análisis estadístico a realizar se centra en los estudiantes inscritos dentro de la Universidad Tecnológica de la Mixteca, universidad ubicada en Huajuapan de León, Oaxaca. La modalidad de trabajo de esta institución pública se centra en un esquema de tiempo completo brindando una vida puramente universitaria en todos sus aspectos.

Este esquema de vida posiciona a los estudiantes ante una nueva realidad ya que en su mayoría la institución está conformada por estudiantes foráneos que en múltiples ocasiones suelen poner los trabajos escolares y el cumplimiento de las asignaciones por encima de un desarrollo saludable completo.

Debido a esto el análisis que se lleva a cabo ayuda a conocer el estado actual de los estudiantes con este estilo de vida por lo que se pueden brindar alternativas para el desarrollo y mejoramiento hacia una forma de vida más saludable y balanceada en los estudiantes.

La información es recabada por medio de encuestas a los estudiantes dentro de las cuales se despliegan opciones variadas para obtener la mayor cantidad de información acerca del estilo de vida personal, siendo obtenida de manera confidencial para evitar cualquier inseguridad por parte de los encuestados. Los datos recabados son tanto numéricos como categóricos, por ello para el procesamiento de los mismos se utilizará el algoritmo K-Prototypes pues el algoritmo permite agrupar de manera sencilla ambos tipos de datos tomando como base la distancia euclidiana para los indicadores de similaridad entre los datos numéricos y la medida de disimilaridad utilizada para los datos categóricos que se tienen, el codo de Jambú es usado para determinar la cantidad óptima de clústeres a formar.

Desarrollo

Distancia Euclidiana

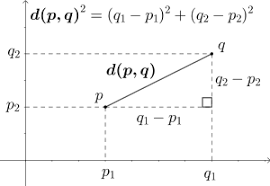
La mayoría de los algoritmos de clustering requieren de la medición de distancias entre datos. Existen distintos tipos de distancia, sin embargo para este caso únicamente se realizará una distancia de tipo euclidiana.

La distancia euclidiana es un número mayor o igual a cero la cual indica la separación de dos puntos en un espacio a través de la geometría propuesta por Euclides [1].

La fórmula de la distancia Euclidiana se muestra en la ecuación 1 donde el conjunto pertenecen a las coordenadas de puntos en el espacio y son los puntos de los centros que se encuentran en otro conjunto de puntos.

(1)

Para un arreglo bidimensional este se muestra en la figura 1.



*Figura 1. Distancia euclidiana bidimensional [1]*

Indicadores de disimilaridad y similaridad

En la actualidad existen diversas maneras de realizar mediciones de distancia donde normalmente se toman en cuenta valores numéricos, sin embargo en muchas ocasiones es necesario recurrir a la relación de distancias entre valores categóricos como lo son algunos exámenes de opción múltiple.

Los valores categóricos son cualidades o características que comparten un grupo de objetos, individuos, procedimientos, condición, entre otros. Estos valores son altamente utilizados en estadística principalmente [2]. Así entonces las mediciones de distancia pueden ser tanto numéricas como categóricas. Con este concepto anteriormente dado se puede decir que la similitud es una manera de poder realizar la medición de los datos adquiridos en base a sus relaciones entre las mismas o cerradas. Por otra parte las medidas de disimilaridad permite realizar la medición de qué tanto un objeto es distinto de otro, los cuales son ampliamente utilizados en el agrupamiento [3].

Según la Universitat de València*:* ***“Se da, en general, el nombre de distancia o disimilaridad entre dos individuos i y j a una medida, indicada por d(i,j) , que mide el grado de semejanza, o a mejor decir de desemejanza, entre ambos objetos o individuos, en relación a un cierto número de características cuantitativa y / o cualitativas. El valor de d(i,j) es siempre un valor no negativo, y cuanto mayor sea este valor mayor será la diferencia entre los individuos i y j.”*** [4].

Matemáticamente la disimilaridad entre dos individuos se puede expresar a través de la ecuación 2.

El algoritmo de K-Means

El agrupamiento de datos (también llamado Clustering) existe desde hace mucho tiempo, existen diversos métodos para dicha agrupación ya que cada uno se enfoca a distintas variables de datos de los cuales los más comunes son los datos categóricos y los datos numéricos, además de tener clasificación como lo es un algoritmo de tipo supervisado y algoritmos de tipo no supervisados.

El algoritmo de K-means es un tipo de algoritmo de clustering de tipo no supervisado el cuál es usualmente aplicado en minería de datos y reconocimiento de patrones [5]. Este tipo de algoritmo trabaja únicamente con variables numéricas, las variables categóricas no pueden ser consideradas para trabajar con este algoritmo. Las bases de este algoritmo se establecen en el criterio del error, el error cuadrático y el rendimiento de la agrupación.

Todos los procedimientos tienen un grado de optimización que para este caso se intentan encontrar K divisiones para satisfacer un criterio de convergencia, esto es el mínimo número de agrupaciones requeridas para que no existan grandes dispersiones de datos. Normalmente se requiere de un filtro antes de aplicar el algoritmo para esta optimización.

Para poder aplicar este algoritmo primeramente se debe elegir un punto arbitrario inicial, el cual normalmente está dado por un punto aleatorio en el espacio del conjunto de datos establecidos inicialmente. Cabe destacar que sin este primer paso el algoritmo no es capaz de poder realizar el proceso; como segundo paso se debe reunir los puntos mostrados restantes a través de la distancia euclidiana, para posteriormente obtener una clasificación inicial y en dado caso de que esta clasificación no sea razonable se modificará. Este proceso se seguirá ejecutando hasta que no exista una diferencia notable con la anterior evaluación.

Este algoritmo es ampliamente utilizado en bases de datos muy grandes que deban reorganizar y agrupar los datos tal como bases de datos de organismos gubernamentales o buscar la distancia más cercana de un local a través del GPS.

En general el algoritmo depende mucho de los puntos inicialmente elegidos y la diferencia en elegir las primeras muestras.

Procedimiento del algoritmo

Como se explicó con anterioridad la determinación del punto focal inicial del agrupamiento es normalmente aleatorio o estocástico. En este caso se usa la distancia mínima próxima para determinar los primeros K clústers [6]. Se deberán seguir una serie de pasos para poder realizar el algoritmo.

Dado un conjunto de N muestras con modelo en espera de ser clasificados.

1. Se elige cualquier elemento de las muestras del modelo para actuar como punto pivote inicial, este llevará de nombre .
2. Se elige cualquier punto que se encuentre lo más lejos posible de que será el punto focal del segundo cluster y además se calcula para posteriormente calcular la distancia entre y cada muestra del conjunto de datos:

Si:

(3)

Que se elige para ser punto focal del segundo cluster, y =

1. Calcular la distancia entre cada muestra a través de y muestra por muestra

(4)

(5)

Se elige el mínimo de los resultados:

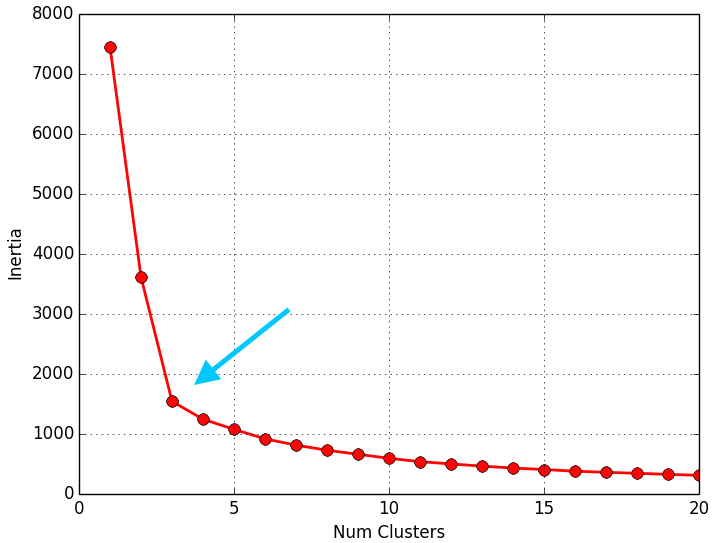
1. Posterior a la elección del mínimo se establece que cada uno de los datos estarán localizados con el centroide más cercano. Esto se establece con la distancia euclidiana.
2. Se actualiza la posición del centroide a la media aritmética de las posiciones de los datos asignados al grupo
3. Repetir los pasos 4 y 5 iterativamente hasta que no haya más cambios, o los cambios sean mínimos.

El codo de Jambú

El codo de Jambú es un algoritmo utilizado para poder realizar la elección óptima de clusters. Este método hace uso de los valores de inercia el cual se obtiene con la reproducción del algoritmo de k-means (que va desde el primer valor hasta N), donde la inercia es la sumatoria de las distancias euclidianas elevadas al cuadrado de cada uno de los objetos hacia el centroide [7]. La fórmula de la inercia está dada por:

(6)

Una vez aplicado el algoritmo este tiene una representación de una gráfica lineal el cual se visualiza la forma de un codo humano, donde esta parte relaciona la inercia y el número óptimo de clusters para seleccionar, es decir esta caída de la gráfica muestra el punto para la elección de clusters óptimos para que no exista la dispersión. En la figura 2 se muestra la aplicación de un ejemplo del codo de Jambú donde el algoritmo se empleó a un número de clusters igual a 20. Se puede mostrar que de acuerdo a los datos obtenidos de la inercia el máximo punto está en aproximadamente en 7500 en el eje de las ‘Y’ mientras que en el eje de las ‘X’ se representa el número de clusters. De acuerdo a la gráfica para este caso la selección óptima del número de clusters es aproximadamente de K=3.



*Figura 2. Representación del codo de Jambú ante un conjunto de datos por el método de la inercia [7].*

El algoritmo de K-Modes

El algoritmo de k-modes al igual que el k-means son algoritmos de tipo no supervisado, es decir que no existe una respuesta correcta con la elección de clusters o agrupamiento ideal, por lo que es una cuestión subjetiva. La principal diferencia que existe entre ambos es el tratamiento de los datos: k-means trata con objetos de tipo numérico mientras k-modes con datos de tipo categóricos.

En este algoritmo se asume que los datos que serán agrupados tienen un conjunto de atributos en común denotado como donde cada atributo describe un dominio de valores denotado por y asociado con un significado de la palabra definida y un tipo de datos [8].

Se denota como un conjunto de n objetos donde el objeto se representa como . Se dice que si para 1≤ j ≤ m.

Las siguientes modificaciones del algoritmo de k-means con respecto al k-modas son: 1) Se usa un emparejamiento de mediciones de disimilaridad para objetos de tipo categórico, 2) El uso de las medias es usado en el algoritmo de k-means donde en el de k-modas se reemplaza esta medición, y 3) Se usa principalmente método basado en frecuencias para encontrar las modas. Con esto dicho anteriormente se siguen las siguientes reglas básicas para realizar el algoritmo: Se denota como y como dos objetos de tipo categórico los cuales son representados como y respectivamente. La medición de disimilaridad entre y se define como la ecuación 7.

(7)

donde

Entonces el k-modas usa directamente el paradigma del k-means, es decir que el objetivo de los algoritmos de agrupamiento es reunir un conjunto de n datos categóricos dentro de k clusters para encontrar una W y Z que minimice

(8)

Donde K es el número de clusters conocidos.

Procedimiento del algoritmo

Los siguientes pasos son necesarios para poder aplicar el algoritmo k-modas son los siguientes:

1. Se eligen k clusters de manera aleatoria o aplicando el algoritmo del codo de Jambú para determinar el número de agrupaciones. De igual manera elegir un punto inicial del conjunto de datos igualmente de manera aleatoria o realizar.
2. Calcular las disimilaridades y asignar cada observación al clúster más cercano.
3. Define las nuevas modas para los clusters.
4. Repetir los pasos 2 y 3 siempre y cuando no haya nuevas reasignaciones requeridas posteriormente.

El algoritmo de K-Prototype

De acuerdo a lo establecido por Huang [9] el algoritmo de k-prototype es un tipo de algoritmo para agrupamiento de datos de tipo mixto, es decir de variables categóricas y numéricas, además de esto es una alternativa para la combinación de ideas de los algoritmos k-means y k-modes. Con lo anterior dicho el algoritmo de k-prototypes combina las ‘medias’ de la parte numérica y las ‘modas’ de las partes categóricas para construir un nuevo centro de cluster híbrido llamado ‘Prototype’ [10]. La medición de distancia de un cluster a otro se representa con la siguiente fórmula:

Donde

(9)

La combinación de esta fórmula mide la disimilaridad y la suma con la distancia euclidiana de los datos numéricos. Existe un parámetro el cual ajusta el grado de disimilaridad para que no exista una variación grande con respecto a la distancia euclidiana.

Procedimiento del algoritmo.

Los siguientes pasos describe el procedimiento a seguir para poder realizar el algoritmo:

1. Se seleccionan k objetos aleatoriamente seleccionados del conjunto de datos establecido inicialmente de acuerdo a los centros de los clusters iniciales.
2. Se calcula la distancia con la ecuación (9). De acuerdo con este cálculo se asigna el valor al cluster más cercano.
3. Acorde a los centros de clusters actuales la disimilaridad del conjunto de datos se recalcula. Se reasigna los objetos al subcluster más cercano; en la parte categórica se utilizan los valores de mayor frecuencia, y en la parte numérica se utiliza el método de valor medio para determinar la cercanía. Actualizar los Centros de Clúster.
4. Se repiten los pasos 2 y 3 mientras no exista variación con respecto a los anteriores cálculos o de igual manera cuando no exista una variación establecida. En dado caso de que no se cumpla esta condición se termina el algoritmo.

Procedimiento

Una vez que se contestaron las encuestas los datos se almacenaron en un archivo separado por comas o .CSV para posteriormente realizar el análisis de los datos con el algoritmo K-Prototipos que se programó en el lenguaje python, la estructura del programa consiste en lo siguiente:

1. Remover outliers : En este paso lo que se busca es eliminar los valores poco creíbles que sean ingresados ya sea intencional o accidentalmente, en este caso se tienen mayores outliers en los valores que son tecleados manualmente por el usuario como edad, peso y estatura, para el filtro de edad se escogió se colocaron límites de no mayor a 28 y no menor a 16 años, para el filtro de estatura se escogieron límites de no mayor a 2 metros ni menor a 1.4 metros y finalmente se escogieron límites de peso de no mayor a 110 Kg ni menor a 40 Kg.

1. Eliminar valores Nan : Para este caso a los valores Nan se les tomó como cero o en su caso categórico como nada, estos valores deben ser eliminados ya que de otro modo el algoritmo no se comportaría de la manera deseada o arrogaría errores al momento del cálculo de los clusters.

1. Calcular los cluster adecuados : En este punto se calculan los clusters adecuados para la implementación de algoritmo, esto se hace usando el codo de Jambú, este valor se ingresará en la función que ofrece python y se iniciara con el claculo de los clusters.

1. Cálculo de los clusters : Para este paso se usarán las funciones “KPrototypes” y “kproto.fit\_predict” para formar los clusters, estas funciones son una herramienta que ofrece la librería “kmodes” de python la cual es de gran ayuda para la implementación del algoritmo K.prototipos

Una vez se tienen los resultados se procede a hacer un análisis de los datos con base en los clusters formados.

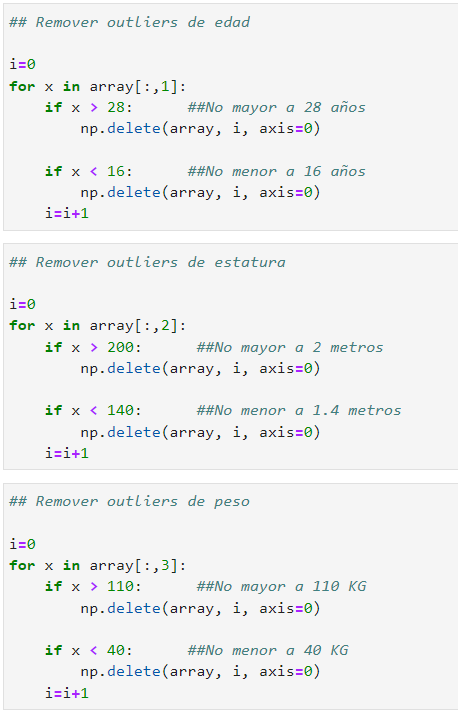
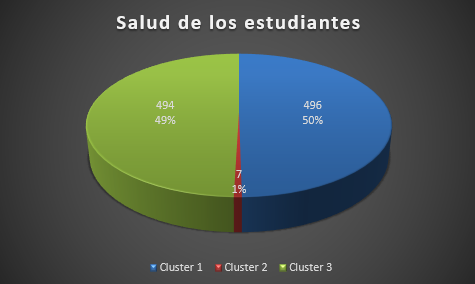


Figura 3. Código preliminar para remover los outliers en Python.

Interpretación y resultados



*Figura 4. Resultados del agrupamiento de los datos de acuerdo a su clusters correspondientes.*

Siguiendo el método de Elbow, usando tres agrupaciones o clusters por la dispersión de los datos y la información recaudada, se realizo una grafica que muestra el porcentaje de personas que tiene cada cluster contra el total de personas encuestadas.

**

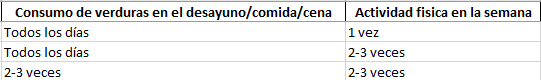
**

1. *Cluster 1*

**

**

1. *Cluster 2*

**

1. *Cluster 3*

*Figura 5. Resultados obtenidos de acuerdo a los datos almacenados de las encuestas donde se muestran:(a) las personas que pueden sufrir algún percance en la salud,(b) los integrantes que tienen un desbalance en su peso pero con buena alimentación y (c ) los que tienen un buen estado de salud.*

Como se puede observar en la ***Figura 4***, los datos se agruparon en 3 diferentes clusters, de los cuales se llegaron a la siguiente interpretación:

* Cluster 1: Se refiere a aquellos estudiantes que tienen una muy mala alimentación, además de tener sobrepeso y carecer de alguna actividad física que aporte a su salud.
* Cluster 2: Es el cluster que muestra la información de las personas que llevan una buena alimentación o realizan ejercicio frecuentemente pero que no se encuentran dentro de su rango de peso ideal, sufriendo de sobrepeso u obesidad.
* Cluster 3: Se caracteriza por tener aquellas personas que se alimentan de una manera balanceada, comiendo frutas y verduras, realizando la actividad física mínima en la semana para estar dentro del rango ideal de su peso.

Conclusiones

El estilo de vida de al menos el 50% de los estudiantes, se encuentran en un estado crítico en la salud ya que, aunque se encuentren dentro de su peso ideal su alimentación no es del todo buena y la falta de ejercicio abunda por los malos hábitos, falta de tiempo y muchos otros factores que se pudieron observar al momento de revisar los datos.

Esto es algo preocupante, ya que actualmente estamos viviendo una pandemia global que afecta a la salud de muchas personas, y sumando el pésimo estado de salud de muchos estudiantes, puede llegar afectar en los estudios de la comunidad estudiantil. Para contrarrestar estos problemas, algunas medidas que se pueden tomar son:

* Inversión en equipamiento para realizar actividades físicas obligatorias dentro de la escuela como la natación, el atletismo, artes marciales, clases de baile, etc.
* Reducción en los costos de los platillos en la cafetería de la escuela, sirviendo más frutas y verduras con su respectiva porción de carne.
* Mejorar el horario de los estudiantes para evitar el descontrol en la administración del tiempo.

Bibliografía

*[1] Lehmann C. (1972) Geometría analítica. UTEHA.*

*[2] Zapata, F. (2019, 11 de julio). Variable categórica: características y ejemplos. Lifeder. https://www.lifeder.com/variable-categorica/*

*[3]* Harmouch, M. (2021, 13 de marzo). *17 types of similarity and dissimilarity measures used in data science.* Towards data science. <https://towardsdatascience.com/17-types-of-similarity-and-dissimilarity-measures-used-in-data-science-3eb914d2681#:~:text=A%20given%20distance(e.g.%20dissimilarity,all%20p,%20q,%20r.>

*[4]* Marquez, A. (2013, 2 de agosto). *Análisis Cluster*. Universitat de València. [https://www.uv.es/ceaces/multivari/cluster/criterios\_de\_similitud.htm#:~:text=Se%20da,%20en%20general,%20el,características%20cuantitativa%20y%20/%20o%20cualitativas.](https://www.uv.es/ceaces/multivari/cluster/criterios_de_similitud.htm#:~:text=Se%20da,%20en%20general,%20el,caracter%C3%ADsticas%20cuantitativa%20y%20/%20o%20cualitativas.)

[5] Li, Y. y Wu, H. (2012). A Clustering Method Based on K-Means Algorithm. *Physics Procedia*, *25*, 1104–1109. <https://doi.org/10.1016/j.phpro.2012.03.206>

[6] Martínez, J. (2020, 20 de septiembre). *Clustering (Agrupamiento), K-Means con ejemplos en python - IArtificial.net*. IArtificial.net. <https://www.iartificial.net/clustering-agrupamiento-kmeans-ejemplos-en-python/#K-Means>

[7] Moya, R. (2016, 16 de septiembre). Selección del número óptimo de Clusters. Jarroba. https://jarroba.com/seleccion-del-numero-optimo-clusters/

[8] Ng, M., Li, M., Huang, J. y He, Z. (2007). On the Impact of Dissimilarity Measure in k-Modes Clustering Algorithm. IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(3), 503–507. https://doi.org/10.1109/tpami.2007.53

[9] Z. Huang, “Clustering large data sets with mixed numeric and categorical values,” in Proceedings of the 1st Pacific-Asia Conference on Knowledge Discovery and Data Mining, Singapore, 1997.

[10] Ziqi Jia, Ling Song, "Weighted k-Prototypes Clustering Algorithm Based on the Hybrid Dissimilarity Coefficient", Mathematical Problems in Engineering, vol. 2020, Article ID 5143797, 13 pages, 2020. https://doi.org/10.1155/2020/5143797