Máster universitario en Rehabilitación Neuropsicológica y Estimulación Cognitiva

8ª Edición 2017-2018

Clasificación de hábitos de vida saludable en población adulta sana mediante la aplicación de técnicas de inteligencia artificial

Trabajo de fin de Máster

Alumna: Alba Roca Ventura

Tutor: Javier Solana Sánchez

31 de mayo de 2018

El presente Trabajo de Fin de Máster, titulado “Clasificación de hábitos de vida saludable en población adulta sana mediante la aplicación de técnicas de inteligencia artificial" realizado por Alba Roca Ventura, ha sido desarrollado con la colaboración del proyecto Barcelona Brain Health Initiative (BBHI) en el Institut Guttmann, que ha proporcionado las bases de datos sobre los casos.

Abstract

Introduction and objectives: This work is part of the Barcelona Brain Health Initiative (BBHI), a research project of *Institut Guttmann*. It consists of a longitudinal study that intends to deepen the knowledge of the indicators and mechanisms related to maintain the health of our brain throughout life. Within this framework, the aim of this study is to find the clustering solution of lifestyles in healthy adult population to establish different profiles that helps to guide the promotion and prevention of health. In order to achieve this goal, the unsupervised learning algorithms within the Data Mining techniques were used.

Methodology: The data was obtained in the first phase of the project, involving 3488 volunteers between 40 and 65 years, free of neurological or neuropsychiatric diagnoses at the time they joined the study. Participants answered on-line questionnaires related to brain health, general health, socialization, nutrition, physical activity, sleep and life plan. In addition, demographic, socioeconomic, anthropometric and other important aspects related to brain health were also collected.

Results and conclusion: Weka® software was used for the analysis, which allowed the application, comparison and interpretation if different algorithms. The best results were achieved using the Expectation-Maximization (EM) and Self-Organizing Maps (SOM), depending on the algorithm used, between six and seven clusters were obtained. The attributes that best discriminated the different clusters were those related to work and the perception of health, along with the indices of cognitive performance.

Key words: clustering, Data mining, brain health, lifestyle, healthy lifestyle habits

Resumen

Introducción y objetivos: Este trabajo se encuadra dentro del Barcelona Brain Health Initiative, un proyecto de investigación promovido por el Institut Guttmann y la Obra Social La Caixa. Consiste en un estudio longitudinal que pretende profundizar en el conocimiento de los indicadores y mecanismos que ayudan a mantener la salud de nuestro cerebro a lo largo de la vida. Dentro de este marco, el objetivo de este trabajo es el agrupamiento de los hábitos de vida de la población adulta para establecer diferentes perfiles que ayuden a orientar la promoción y prevención de la salud. Para conseguir esta meta se analizan los algoritmos de aprendizaje no supervisados, dentro de las técnicas de Data Mining.

Metodología: Los datos se obtuvieron en la primera fase del proyecto, donde participaron 3488 voluntarios entre 40 y 65 años, libres de diagnósticos neurológicos o neuropsiquiátricos en el momento que se unieron al estudio. Los participantes respondieron una serie de cuestionarios on-line relacionados con la salud cerebral, salud general, socialización, nutrición, actividad física, sueño y plan vital. A parte, también se recogió información demográfica, socioeconómica, antropométrica.

Resultados y conclusión: Para el análisis se utilizó el programa Weka®, que permitió aplicar los diferentes algoritmos, compararlos e interpretarlos. Los mejores resultados se lograron utilizando los algoritmos Expectation-Maximization (EM) y Self-Organizing Maps (SOM), dependiendo del algoritmo utilizado, se obtuvieron entre 6 y 7 grupos. Los atributos que mejor discriminaban los diferentes clústeres eran los relacionados con el trabajo y la percepción de salud, junto con los índices de rendimiento cognitivo.

Palabras clave: Data Mining, Salud Cerebral, Hábitos de vida , Clustering

Índice

1. Introducción y Objetivos

El impacto de las enfermedades neurológicas es un problema creciente y su incidencia (nuevos casos por año) y su prevalencia (individuos afectados) es, actualmente, uno de los principales problemas de salud pública de los países desarrollados. De acuerdo con la Organización Mundial de la Salud (OMS, (Rodríguez & Minoletti, 2013), una de cada cuatro personas desarrollará un trastorno psiquiátrico o neurológico a lo largo de sus vidas y, en Cataluña, en el año 2030 se prevé que se den unos 30.000 nuevos casos de discapacidad neurológica. En una sociedad donde se esta incrementado la esperanza de vida, el tratamiento de las enfermedades neurodegenerativas es necesario pero no suficiente, ya que los síntomas normalmente se manifiestan décadas después de que los daños cerebrales causados por la enfermedad hayan empezado (Kaeberlein et al., 2015). Por esto, es esencial mantener la salud cerebral a lo largo de la vida, haciendo el cerebro más resistente a los cambios o a las enfermedades que puedan aparecer. Esta aproximación, para prevenir enfermedades promoviendo la salud cerebral, representa un cambio de paradigma y requiere un replanteamiento de la salud pública, las políticas de salud, y por encima de todo de los hábitos y estilos de vida de cada persona.

Mantener la capacidad funcional del sistema nervioso a lo largo de la vida del ser humano es un objetivo prioritario de la investigación biomédica en el siglo 21. Un cerebro sano y activo tiene mas recursos para afrontar los años y los posibles daños o enfermedades que pueden ocurrir. Por este motivo, el Institut Guttmann, centro de neurorehabilitación especializado en el tratamiento de la discapacidad, ha abierto una nueva línea de investigación más centrada en la promoción de hábitos saludables y la capacidad de recuperación del cerebro. En Marzo de 2017 lanzó el Barcelona Brain Health Initiative (BBHI) con el apoyo de Obra Social “la Caixa”, un proyecto de investigación cuyo objetivo es descubrir qué podemos hacer para mantener el cerebro saludable durante toda nuestra vida. El proyecto invita a ciudadanos de Cataluña a conocer más y mejor el estado de salud del cerebro pero, sobre todo, a incorporar los hábitos de vida saludables que tienen un efecto positivo en la salud de su cerebro. Esto

requiere voluntarios sanos (sin diagnóstico médico de enfermedad neurológica o condición psiquiátrica) entre 40 y 65 años de edad, un rango de edad adecuado para la prevención de trastornos que pueden aparecer a estas edades y con el que se ha llevado a cabo poca intervención de investigación.

Los principales objetivos del Barcelona Brain Health Initiative son por un lado, identificar aquellos factores del estilo de vida y los mecanismos biológicos que subyacen una buena salud cerebral en adultos de mediana edad y ancianos "jóvenes" y que predicen el mantenimiento de salud cognitiva y conductual. Por otro lado, se quiere evaluar los efectos de una intervención multidimensional de un programa personalizado, focalizado en estilos de vida para promover la salud cerebral individual, preservar la cognición y reducir la carga de trastornos crónicos en personas mayores al potencial la salud cerebral mediante mecanismos que ayuden a mantener la reserva cognitiva y la plasticidad cerebral.

Con este propósito, el proyecto propone un estudio de cohortes longitudinal prospectivo a lo largo de tres años desarrollado en tres fases, incluyendo un ensayo clínico de intervención.

En la fase I del BBHI, se administran periódicamente cuestionarios a los participantes sobre sus estilos de vida, en diferentes áreas, con el objetivo de descubrir cuales son los determinantes asociados al mantenimiento de la salud cerebral. Los cuestionarios se administran cada año con el objetivo de conocer, monitorizar y evaluar el impacto de los posibles cambios. Los cuestionarios que se administran en esta fase son los que se utilizan en este estudio y se explicaran más adelante.

En la Fase II, que empieza en 2018, un grupo seleccionado de voluntarios participará en evaluaciones presenciales. Se recogerán datos personales, biográficos, ambientales y sociales relacionados con la salud cerebral, que permitirán estudiar los correlatos neurobiológicos asociados con una mayor o menor resiliencia a desarrollar enfermedades psiquiátricas o neurológicas. El protocolo incluye , además de los datos ya comentados, una evaluación cognitiva, una evaluación médica, una resonancia magnética, un electroencefalograma, y la recogida de muestras de sangre y saliva. Además de una prueba para evaluar la capacidad de respuesta cerebral a la interferencia en multitarea y una prueba de perturbación bioeléctrica mediante

estimulación cerebral no invasiva. Las evaluaciones se repetirán periódicamente a lo largo del tiempo (previsiblemente cada 2 años) para determinar los indicadores biológicos que pueden formar un "índice de salud cerebral" para cada persona.

En la Fase III, se invitará a un subgrupo de la Fase II de voluntarios a participar en un programa de intervención multidimensional para promocionar hábitos saludables, diseñados específicamente para mejorar los factores anteriormente comentados: los hábitos dietéticos, la actividad física, la actividad cognitiva, la socialización, el sueño, orientación del plan vital y salud médica general. Se evaluará el impacto de los cambios en los hábitos y en los parámetros biológicos que constituirían el "Brain Health Index", incrementando la resistencia de cada persona a desarrollar enfermedades neurológicas o psiquiátricas. Y también se compararán diferentes estrategias de intervención que incluyen programas educacionales generales, frente a programas de "coaching" personalizado, mediante tecnologías móviles para la interacción con usuarios (apps) y supervisión de diferentes hábitos de salud.

En este proyecto, la tecnología desempeña un papel clave, en el cual la aplicación de técnicas de inteligencia artificial y de minería de datos brinda una gran oportunidad para investigar y adquirir conocimientos sobre las relaciones entre los hábitos saludables y el estado de salud autopercibido. Gracias a estas técnicas, podemos mejorar el "Brain Health Index" del que hemos hablado. Además, las técnicas de *clustering* pueden ayudar a determinar grupos de participantes con características similares, por lo que podremos comparar y explorar abordajes basados en tecnología para aumentar la personalización y aumentar la motivación y la adhesión a hábitos saludables. De esta forma, y en línea a los objetivos del BBHI, el presente trabajo pretende: A) Clasificar los hábitos de salud de una amplia muestra de población adulta sana mediante técnicas de inteligencia artificial. B) Seleccionar el mejor algoritmo de agrupación. C) Hacer un análisis descriptivo de cada clúster para poder hacer una interpretación del perfil de cada grupo.

La Salud cerebral

La Salud cerebral se define como el desarrollo y preservación de las funciones cerebrales y redes neuronales, acorde con la edad, necesarias para disfrutar completamente de la vida, y también la buena capacidad de recuperarse de una lesión o una enfermedad. En este sentido, es conocido que hay importantes diferencias entre diferentes personas en la habilidad para responder a un mismo trastorno neurológico (Redolat & Mesa-Gresa, 2015). La capacidad de **resiliencia** es el proceso de adaptarse bien a la adversidad, a un trauma, tragedia, amenaza, o fuentes de tensión significativas, como problemas familiares o de relaciones personales, problemas serios de salud o situaciones estresantes del trabajo o financieras. Esta difiere entre individuos, reflejando una combinación entre aspectos genéticos y ambientales, pero no es una característica que la gente tiene o no tiene. Incluye conductas, pensamientos y acciones que pueden ser aprendidas y desarrolladas por cualquier persona (Pagina web APA, Asociación Americana de Psicología). Uno de los conceptos más ligados al termino de resiliencia es la reserva cognitiva, un término neuropsicológico que se refiere a la capacidad de tolerar los cambios en las estructuras cerebrales. Cambios que pueden ser causados tanto por la edad como por alguna patología sobrevenida y que se manifiestan en la propia estructura del cerebro o en las formas de procesamiento de la información. Durante el proceso de envejecimiento, el cerebro experimenta una reducción del volumen cerebral, pérdida de tejido neuronal y modificaciones en la actividad de los neurotransmisores, que al final se traduce en un cierto declive natural o neuropatológico de funciones cognitivas como la memoria, las funciones ejecutivas, la atención, las habilidades visoespaciales o la velocidad de procesamiento de la información. Sin embargo, pese que es un proceso universal, estos cambios no son iguales para todos los sujetos. Existe un incremento de variabilidad interindividual que puede derivarse tanto de diferencias genéticas innatas como ambientales (nivel educativo, sexo, condición socioeconómica, estilos de vida, tipo de actividades de ocio o tipo de trabajo entre otros) antes y durante la etapa del envejecimiento. Estas diferencias pueden determinar un mejor o peor curso del envejecimiento cognitivo (Bastin et al., 2012, Lee, 2003; Sánchez et al., 2010; Stern, 2009;

Stern et al., 2003). Así surge la hipótesis de la “reserva cognitiva”(RC), que sostiene que la mayor capacidad intelectual innata y la mayor estimulación cognitiva a través de un ambiente enriquecido, brindaría ciertos beneficios al cerebro maduro para afrontar los cambios ocasionados por el paso del tiempo o los efectos de un proceso neurodegenerativo que pueden ser causa de demencia, a través de la implementación de procesos cognitivos (Fratiglioni, Paillard-Borg y WinBlad, 2004; La Rue, 2010; Whalley, Deary, Appleton y Starr, 2004). Esto puede estar debido a la utilización más eficiente de las redes neuronales cerebrales o una mayor capacidad para reclutar redes alternativas según fuera necesario. La reserva cognitiva tiene un efecto neuroprotector que permitiría responder más adecuadamente a los cambios del cerebro (Stern, 2002). El término de reserva cognitiva (también denominado reserva cerebral), se acuñó a partir de varias observaciones repetidas donde no existía una relación directa entre la severidad del daño cerebral y el grado de afectación a la hora de la realización de una actividad o tarea, ni con las características clínicas de los pacientes (Bosch Capdevilla, B., 2010, Snowdon, 1986). Esta capacidad cognitiva e intelectual, empieza a formarse ya desde los primeros momentos, durante el período de desarrollo el cerebro va formando conexiones gracias a las influencias del entorno y se va acumulando a lo largo de su vida. Entre los factores más importantes que influyen a la reserva cognitiva encontramos el cociente intelectual, la educación y nivel cultural, el ocio (lectura, juegos, idiomas..) y relaciones sociales, el ejercicio físico y el ejercicio mental. También es importante llevar una correcta alimentación, donde ingerimos alimentos que contienen vitaminas para el cerebro, ejercicio físico moderado, no fumar o no beber alcohol en exceso ya que son factores que también están relacionados con mayores niveles de reserva cognitiva. Además, influyen también las capacidades innatas y los factores genéticos y los biomarcadores, lo que explicaría la variabilidad individual (Montañés Ibáñez, A., 2012; Redolat, R. y Carrasco, M.C. , 1998).

En relación a esto, se sabe que la participación en actividades cognitivamente estimulantes contribuye a la reserva cognitiva. De hecho, el abordaje frecuente de tareas que impliquen un reto intelectual, estaría asociado con un nivel de

reserva cognitiva más alto (Rodríguez Álvarez, M. y Sánchez Rodríguez, J.L., 2004).

Esta capacidad intrínseca de modificar su función y estructura para desarrollar nuevas capacidades y adaptarse a nuevos retos o cambios en el ambiente es gracias a la llamada plasticidad neuronal, postulada por primera vez por Santiago Ramón y Cajal a finales del sigo XIX. Postula a través de su hipótesis de “gimnasia cerebral” que gracias a la plasticidad de las neuronas corticales, un estimulo continuado produciría un incremento de las conexiones entre neuronas, aumentando así la capacidad del cerebro. Esta capacidad le permite al cerebro cambiar, modificar y reorganizarse dinámicamente en respuesta a estimulación sensorial, cognitiva o al aprendizaje, así como también, le permite el recuperar o compensase tras una lesión cerebral (Bergado y Almaguer, 2000; Boakye, 2009; Muresanu, 2007). Gracias a estas propiedades del cerebro, el objetivo, es saber de estos cambios para potenciarlos y dirigirlos. Aprovechar la plasticidad neuronal y promover la reserva cognitiva para mejorar la salud cerebral de cada individuo. Además, el sistema nervioso ejerce una función reguladora al resto de sistemas que al optimizar la salud cerebral nos permitirá mejorar la salud general y reducir el impacto de diferentes enfermedades crónicas.

2. Metodología

Los datos se obtuvieron de los cuestionarios que se administraron vía on-line en la fase 1 del proyecto BBHI, a un total de 3488 participantes entre 40 y 65 años de edad. Los participantes eran de ambos sexos y libres de enfermedades neurológicas o neuropsiquiátricas autoreportadas cuando se inició el estudio.

Los cuestionarios recogían información sobre percepción subjectiva de salud y sobre estilos de vida relacionadas con 7 pilares básicos de la salud cerebral definidos en el proyecto: salud cognitiva, actividad física, socialización, salud general, nutrición, sueño y plan vital. Además, también se recogió información sociodemográfica, la presencia de diferentes factores de riesgo A continuación se explican los cuestionarios administrados que se han utilizado en la formación de los clústeres.

Cuestionarios auto-informados online

Los participantes se apuntaron en la fase 1 del estudio de forma voluntaria mediante la página web o la aplicación móvil del proyecto. Primeramente, se les administró una instrumento de *screening* multidominio como base-line.

Este cuestionario, creado por el equipo de BBHI utilizando partes de instrumentos previamente validados en otros estudios, se construyó para recoger información sobre el participante y su estilo de vida relacionada con el mantenimiento de la salud cerebral (Skoog et al., 2017; Livingston et al., 2017).

La primera parte del cuestionario recoge información demográfica, socio-económica y antropométrica, así como la presencia de diagnósticos médicos y otros factores de riesgo (por ejemplo, fumar, consumo de alcohol y drogas).

Para evaluar la auto-percepción de la salud general, dolor y quejas cognitivas y mentales se utilizó el *Patient-Reported Outcomes Measurement Information System of global health and pain interference short forms (PROMIS)*, el Neuro-

Qol y el Patients Health Questionnaire for Depression and Anxiety (DASS) (Ader, 2007; Cella et al., 2012; Kroenke, Spitzer, Williams, & Löwe, 2009).

Los hábitos de vida en nutrición, actividad física y sueño se exploraron utilizando ítems del *Mediterranean Diet Adherence Scale* (Schröder et al., 2010) , el *Pyshical Activity Scale* y el *Jenkins Sleep Evaluation Questionnarie* (JSEQ, Jenkins, Stanton, Niemcryk, & Rose, 1988) respectivamente.

Finalmente, el *Ryff Scale* (Ryff, 1995) que evalúa el bienestar psicológico y se utiliza para medir las dimensiones de "propósito en la vida" y "crecimiento personal" dentro del área de plan vital.

Periódicamente, se les pidió que rellenaran otros cuestionarios validados para evaluar más especifica y concretamente aspectos destacados como importantes para la salud cerebral.

Aprendizaje no supervisado.

En las ultimas décadas ha habido un aumento constante en el uso de técnicas de minería de datos en un amplio número de disciplinas. La Minería de Datos es una forma de descubrir conocimiento y es un proceso muy útil para descubrir patrones en los datos mediante la exploración y modelaje de grandes cantidades de datos (Witten, Frank, & Hall, 2011). La distinción entre estadística y minería de datos se ha relacionado con la naturaleza del análisis; la estadística maneja un análisis primario, mientras que la minería de datos se encarga de un análisis secundario (Yoo et al., 2012) que aprende de los datos (Friedman, 1997). La minería de datos incorpora algoritmos de aprendizaje automático que pueden aprender, extraer e identificar información útil y conocimiento de amplias bases de datos (Joanna F. Dipnall et al., 2016; Witten et al., 2011).

Dentro de la aplicación de la Minería de datos podemos describir cuatro estilos de aprendizajes diferentes. En el aprendizaje de clasificación, el esquema de aprendizaje se presenta con un conjunto de ejemplos ya clasificados, datos de entrenamiento, y se espera que aprenda a clasificar para poder pasar a los datos en que no conoce el resultado, datos test. En el aprendizaje de asociación, se busca cualquier asociación entre características, no solo aquellas que predicen un valor de clase particular. En la agrupación o

clustering, se busca hacer grupos o clústeres en que los ejemplos o datos están relacionados. Y en la predicción numérica, el resultado a predecir no es una clase discreta sino una cantidad numérica (Witten et al., 2011).

Mientras que en otros campos, como por ejemplo en el marqueting, los algoritmos de aprendizaje automático se han utilizado desde hace tiempo, en el campo de la salud se están empezando a utilizar recientemente. La caracterización de pacientes permite mejorar la calidad de atención hospitalaria. Si tratamos de entender la distribución de enfermedades en multimorbilidad nos enfrentamos a un laberinto lleno de posibilidades. Todas las enfermedades están más o menos asociadas estadísticamente entre sí. Comenzar en cualquier punto de este laberinto y no tener ninguna guía para encontrar a el camino hace que nos perdamos fácilmente. Por esa razón, es importante descubrir la estructura subyacente en la distribución de enfermedades, es decir, qué vías pueden conducir a través de este laberinto de multimorbilidad (Schäfer et al., 2010). En estos casos, un análisis de clúster es una forma muy útil de identificar diferentes perfiles presentes en la heterogeneidad de la población, al identificar grupos de personas que comparten características similares ayudará a entender mejor estas características (Ortoleva Bucher, Dubuc, von Gunten, Trottier, & Morin, 2016). El uso de algoritmos de *clustering* cada vez se está utilizando más en la salud para distinguir entre diferentes patologías, pero no hay estudios dirigidos a la población sana y así poder utilizar estos datos para la prevención y promoción de la salud.

El *Clustering* o algoritmo de agrupamiento implica separar un conjunto de datos en unos subconjuntos denominados clústeres. Un clúster es una colección de datos donde los elementos que lo componen son similares entre ellos y diferentes de los elementos de otro clúster. El objetivo de la agrupación es encontrar una estructura inherente en los datos y mostrar esta estructura como un conjunto de clústeres. Numerosos algoritmos de *clustering* se han introducido y usado durante las ultimas décadas (Yoo et al., 2012).

Los algoritmos de *clustering* se pueden dividir en varios grupos según sus características. La principal división es entre técnicas o tipos de *clustering* Jerárquicos y no Jerárquicos. En los Jerárquicos, los datos se van agrupando en conjuntos cada vez más numerosos hasta que sólo queda uno de ellos que

reúne a todos los elementos. Se clasifican como métodos jerárquicos aglomerativos, en que se comienza con los objetos o individuos de modo individual; de este modo, se tienen tantos clúster iniciales como objetos. Luego se van agrupando de modo que los primeros en hacerlo son los más similares (Mínima distancia o vecino más próximo) y al final, todos los subgrupos se unen en un único clúster. Y , por otro lado, los métodos jerárquicos divididos que se actúa al contrario. Se parte de un grupo único con todas las observaciones y se van dividiendo según lo lejanos que estén.

Los no jerárquicos se denominan normalmente particionales, en que los elementos se dividen en un número determinado de grupos (prefijado de antemano). Dentro la literatura podemos encontrar otro algoritmo no jerárquico, el probabilista, en que un elemento puede pertenecer a varios grupos simultáneamente con distintas probabilidades (mixtura de Gaussianas) aunque normalmente lo incluyen dentro del tipo particional. En la figura 1 vemos un ejemplo de esta clasificación de los distintos algoritmos.

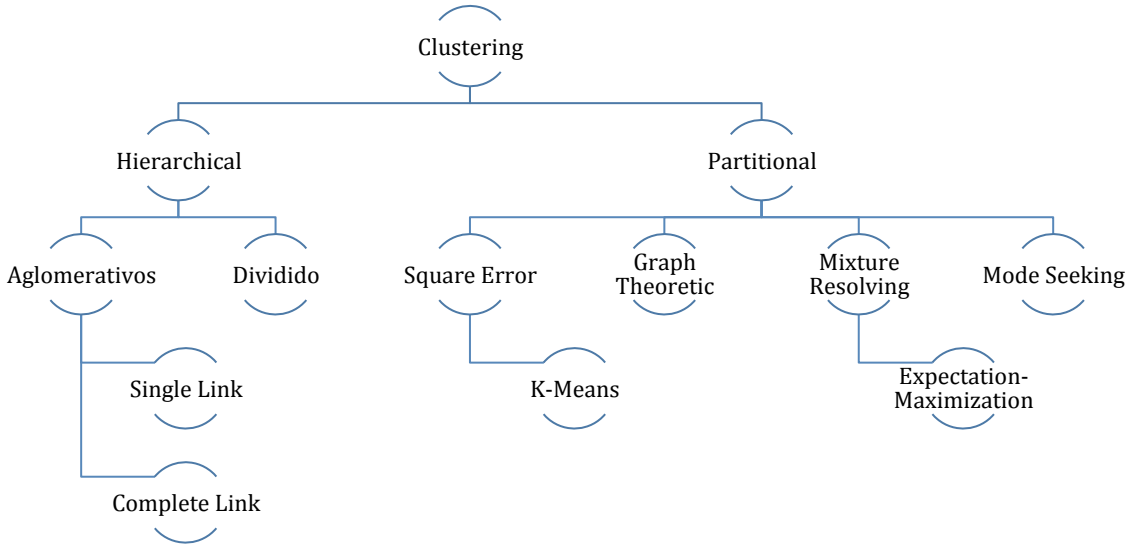


Figura 1. Clasificación de las técnicas de clustering dentro del aprendizaje no supervisado

A continuación, se explicarán algunos de los algoritmos que se encuentran en la literatura relacionada.

Métodos de *Clustering* Jerárquico

Este método se basa en una descomposición jerárquica del conjunto de los datos. La formación de las capas, se puede llevar a cabo mediante operaciones de división (arriba-abajo, *top-down*) en que cada observación empieza en su propio clúster, y se van emparejando a medida que se sube en la jerarquía; o por aglomeración, también llamado aproximación (de abajo hacia arriba, *bottom-up*) dónde todas las observaciones comienzan perteneciendo a un único clúster, que se va dividiendo repetidamente a medida que se desciende en la jerarquía.

Este método genera los clústeres anidados y la precisión es alta. Este método comporta una serie de problemas ya que cada vez que se aglomera, las similitudes entre los clústeres deben compararse globalmente antes de elegir los mejores clústeres, como resultado, es un proceso muy lento y pesado computacionalmente y el método no es adecuado para aplicaciones con alto volumen de datos. Además, las fuentes de error y variación no son considerados con los métodos jerárquicos y esto implica una gran sensibilidad a observaciones anómalas o *outliers*. Si un objeto se ha colocado erróneamente en un grupo al principio del proceso, ya no se puede arreglar en una etapa posterior. Una forma de solucionar este problema, es usar varias distancias o similitudes con los mismos objetos y observar si se mantienen los mismos clústeres o grupos. Así, se comprueba la existencia de grupos naturales.

Cobweb

Es un algoritmo incremental de *clustering* jerárquico de tipo conceptual que construye los clústeres sin un numero predefinido previamente. Fue inventado por el profesor Douglas H. Fisher (1987).

Este algoritmo organiza incrementalmente las observaciones en un árbol de clasificación. Cada nodo del árbol de clasificación representa una clase (concepto) y es designado por un concepto probabilístico que resume las distribuciones de atributos y valores de los objetos clasificados bajo el nodo.

Este árbol de clasificación puede ser utilizado para predecir atributos ausentes o clases de un nuevo objeto (William Iba and Pat Langley).
Se encuentran cuatro operaciones básicas que COBWEB utiliza para construir el árbol de clasificación. La operación se selecciona dependiendo de la utilidad de categoría de la clasificación que se logra al aplicarla, se empieza con una red de nodos vacía y las instancia se van añadiendo una por una. Se pueden colocar dentro de un nodo ya creado, creando uno nuevo, combinando dos clases en una sola clase y situando la nueva instancia dentro de la jerarquía, partiendo una clase en dos.

Métodos de *Clustering* Particionales

k-Means

El método denominado k-means (o k-medias) es uno de los algoritmos más sencillo y más ampliamente utilizado. Este algoritmo divide los N objetos en K particiones (K siendo un valor arbitrario) en donde un objeto irá al clúster con la media más cercana. El algoritmo asigna K centros aleatoriamente, luego asigna los objetos al centro más cercano. El centro se recalcula como la media de los puntos que tiene asignado, una vez actualizado se vuelven a reasignar los objetos al más cercano y así hasta tener convergencia. Depende mucho de la asignación inicial de los centros, nos puede dar un resultado u otro por lo que es mejor hacer varias pruebas con diferentes valores. Si dos centroides iniciales caen por casualidad en un único clúster natural, entonces los clústeres que resultan están poco diferenciados entre sí.
Es un algoritmo simple, eficiente y muy general pero la principal desventaja de este algoritmo es que hay que especificar el número de clústeres que queremos que nos forme, y dependiendo de este numero los resultados varían y puede dar lugar a grupos artificiales o bien a juntar grupos distintos. También es un algoritmo muy sensible a ruido y a *outliers,* ya que si aparecen, se obtiene por lo menos un clúster con sus objetos muy dispersos. Una posible solución es considerar varias elecciones del número k de clústeres comparando luego sus coeficientes de la F de Snedecor.
La investigación en el método k-means ha sido muy extensa y todavía sigue activa. Existen otras variantes, las más conocidas son k-medoid (Kaufman &

Rousseuw, 1991) y Fuzzy C-means (Dunn, 1973). La primera utiliza la mediana en lugar de la media para calcular los centroides de los clusters. De esta forma, el centroide es un punto de los pertenecientes al conjunto original, en lugar de un punto "ficticio" generado artificialmente. La segunda variante, Fuzzy C-means, se diferencia en que calcula funciones de pertenencia difusas para cada clúster, en lugar de un valor de pertenencia único y concreto.

Esperanza-Maximización

Otro algoritmo bastante difundido es el Esperanza-Maximización (Dempster, Laird, & Rubin, 1977) o algoritmo EM (del inglés Expectation-Maximization) es un método iterativo para encontrar estimadores de parámetros de modelos estadísticos con la mayor probabilidad (maximum likelyhood), donde el modelo depende de variables latentes no observables. El algoritmo EM alterna pasos de esperanza (paso E), donde se computa la esperanza de la verosimilitud mediante la inclusión de variables latentes como si fueran observables, y un paso de maximización (paso M), donde se computan estimadores de máxima verosimilitud de los parámetros mediante la maximización de la verosimilitud esperada del paso E. Los parámetros que se encuentran en el paso M se usan para comenzar el paso E siguiente, y así el proceso se repite. Es decir, se asignan unos valores arbitrarios a uno de los dos conjuntos desconocidos, y se usan para estimar los valores del otro. A continuación se usan esos nuevos valores para encontrar una mejor estimación del primer conjunto, y así se van alternando entre ambos, hasta que los valores resultantes converjan hacia puntos fijos. El valor obtenido es un máximo de la función de similitud. En general, puede haber múltiples máximos y no hay garantía de que se encuentre el máximo global.

El algoritmo se utiliza frecuentemente para algoritmos de agrupamiento en aprendizaje automático y visión artificial, para aprender Modelos ocultos de Márkov y Mixturas Gaussianas, utilizadas en procesos de clasificación o reconocimiento. Además, en psicometría, es casi indispensable para estimación de parámetros de ítems y habilidades latentes de teoría de respuesta al ítem. En los casos en los que las ecuaciones de los modelos estadísticos no pueden ser resueltas directamente, se usa el algoritmo EM para encontrar los parámetros de máxima verosimilitud. Generalmente, estos

modelos incluyen variables latentes, además de parámetros desconocidos y datos conocidos obtenidos de las observaciones. Es decir, hay valores desconocidos entre los datos o el modelo no puede ser formulado de manera más sencilla asumiendo la existencia de datos no observados adicionales.

SOM

Las redes neuronales artificiales (RNA) han emergido como una potente herramienta para el modelado estadístico, orientadas principalmente al reconocimiento de patrones, tanto en la vertiente de predicción como de clasificación (Martínez, s.f). Entre los paradigmas de RNA que son usados en diversos campos de aplicación, se destacan los SOM de Kohonen (2001). Este algoritmo se introdujo en 1981 y sus primeras aplicaciones se realizaron principalmente en ingeniería. Posteriormente, el algoritmo se convirtió progresivamente en un método estándar para el análisis de datos, en una gran variedad de campos donde se utiliza el aprendizaje no supervisado: agrupamiento, visualización, organización de datos, caracterización y exploración, aplicado a la solución de problemas en múltiples áreas del conocimiento.

El mapa auto-organizado o SOM (de las siglas en ingles Self-Organizing Map), también llamado red Kohonen por su creador, realiza clústeres de aprendizaje no supervisado. Intenta descubrir patrones en el conjunto de datos de entrada y los agrupa en grupos distintos sin un objetivo previo. Los registros dentro de un grupo o clúster tienden a ser similares entre sí, y los registros de los otros grupos son diferentes. Un SOM usa aprendizaje competitivo (es decir, el algoritmo ganador se lleva todo). Cuando se impone un patrón de entrada en la red neuronal, el algoritmo selecciona el nodo de salida con la menor distancia euclidiana entre el vector de patrón de entrada presentado (X) y su vector de ponderación (j W). Solo la neurona ganadora genera una señal de salida desde la capa de salida; todas las otras neuronas en la capa tienen una señal de salida cero. Debido a que el aprendizaje implica el ajuste del vector de peso, solo las neuronas en el vecindario de la neurona ganadora pueden aprender con este patrón de entrada particular. Lo hacen ajustando sus pesos más cerca del vector de entrada, de acuerdo con la ecuación(Li, X; Zaïane, O.R.; Li, Z., 2006). La principal ventaja es que el SOM es un método de visualización

multivariante muy útil que permite que los datos multidimensionales se muestren como un mapa bidimensional(Behbahani & Nasrabadi, 2009). A parte, puede manejar bases de datos grandes, sin depender de suposiciones distribucionales, lo que permite su uso en una amplia variedad de disciplinas.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

DBSCAN es un algoritmo de agrupamiento de datos basado en la densidad. Fue propuesto por Martin Ester, Hans-Peter Kriegel, Jörg Sander y Xiaowei Xu (1996) (Instituto de Ciencias de la Computación - Universidad de Munich. Modela los clústeres como grupos de alta densidad de puntos. Por lo cual, si un punto pertenece o no a un clúster, debe estar cerca de un montón de otros puntos de dicho clúster.

Se definen dos parámetros, un número ***épsilon*** positivo y un número natural ***minPoints***, y se empieza eligiendo un punto arbitrario en el conjunto de datos. Si hay una cantidad de puntos mayor o igual a *minPoints* a una distancia *épsilon* del punto arbitrario, a partir de ese momento se consideran todos los puntos como parte de un clúster. A continuación, se expande ese grupo mediante la comprobación de todos los nuevos puntos y ver si ellos también tienen más puntos *minPoints* a una distancia *épsilon*, creciendo el clúster de forma recursiva en caso afirmativo. A medida que se generan los clúster, quedan puntos sin añadir al clúster. A continuación, se elige un nuevo punto arbitrario y se repite el proceso. Es posible que el punto arbitrario escogido tenga menos de *minPoints* puntos en su circulo de radio *épsilon*, y tampoco sea parte de cualquier otra agrupación. Si ese es el caso, se considera un "punto de ruido" que no pertenecen a ningún grupo. El proceso se repite hasta que todos los puntos del clúster están determinados dentro de un clúster o como ruido.

DBSCAN posee varias ventajas frente o otros algoritmos de *clustering*. Primero, no requiere un numero de clústeres fijado previamente. También identifica *outliers* como ruido contrariamente a otros clústeres que los engloba a otros clústeres aunque los datos sean muy distintos. Además, puede encontrar clústeres no separados linealmente ni depende de las condiciones de inicio. Las desventajas es que asume densidades similares en todos los

clústeres y puede tener dificultades para separar los clústeres cuando las densidades entre ellos varían.

Optimización por nube de partículas (PSO)

La Optimización por nube de partículas o PSO (por sus siglas en ingles "*particle swarm optimization")* es una técnica de búsqueda aleatoria basada en la población que evoca el comportamiento de partículas en la naturaleza. Originalmente propuesta por J. Kennedy y R. Eberhart (1995) se ha convertido en una fascinante rama de la computación evolutiva. La motivación para el desarrollo de este algoritmo fue el comportamiento social de los animales, como el agrupamiento de aves, bancos de peces y inteligencia de enjambre. Posteriormente el algoritmo se simplificó y se comprobó que era adecuado para problemas de optimización.

La PSO permite optimizar un problema a partir de una población de soluciones candidatas, denotadas como "partículas", moviendo éstas por todo el espacio de búsqueda según reglas matemáticas que tienen en cuenta la posición y la velocidad de las partículas. El movimiento de cada partícula se ve influido por su mejor posición local hallada hasta el momento, así como por las mejores posiciones globales encontradas por otras partículas a medida que recorren el espacio de búsqueda. El fundamento teórico de esto es hacer que la nube de partículas converja rápidamente hacia las mejores soluciones.
PSO es una metaheurística, ya que asume pocas o ninguna hipótesis sobre el problema a optimizar y puede aplicarse en grandes espacios de soluciones candidatas. Sin embargo, como toda metaheurística, PSO no garantiza la obtención de una solución óptima en todos los casos.

Análisis de los datos

Para el análisis y formación de clústeres se ha utilizado el programa WEKA (Waikato Environment for Knowledge Analysis) desarrollado en la Universidad de Waikato. Es una plataforma de software libre utilizado para el aprendizaje automático y la minería de datos escrito en Java. Weka contiene una colección

de herramientas de visualización y algoritmos para el análisis de datos y una interfaz gráfica de usuario (GUI) para acceder fácilmente a sus funcionalidades. Para poder escoger mejor aquellos atributos que más valor aportaban al agrupamiento se utilizó un selector de atributos como evaluador "*InfoGainAttributeEval"* y de método de búsqueda *Ranker*.

Validación de datos

La evaluación de los resultados de los algoritmos de *clustering* es importante. Aún así, es difícil definir cuándo el resultado de un agrupamiento es aceptable. Por esta razón existen técnicas e índices para la validación de un agrupamiento realizado. Existen dos categorías para la validación de *clustering*, la validación externa y la validación interna. La validación externa usa información externa para la validación, es decir, información que no es producto de la técnica de agrupación utilizada. Principalmente es usada para escoger un algoritmo de *clustering* óptimo sobre un data set especifico. A diferencia de las técnicas de validación externas, las de validación interna miden el *clustering* únicamente basadas en información de los datos. Evalúan que tan buena es la estructura del *clustering* sin necesidad de información ajena al propio algoritmo y su resultado. Pueden usarse para escoger el mejor algoritmo de *clustering*, así como el número de clúster óptimo sin ningún tipo de información adicional.

En nuestro caso no disponemos de información externa, por lo que hemos aplicado técnicas de validación interna. Se ha utilizado el Índice de Silhouette por su fácil interpretación y la opción de aplicarlo mediante el programa Weka con el paquete KValid que analiza el número óptimo de k mediante este índice. El coeficiente de Silhouette utiliza tanto criterios de cohesión como de separación. La cohesión (a(x)) es la distancia promedio de x a todos los demás puntos en el mismo clúster. Y la separación (b(x)) es la distancia promedio de x a todos los demás puntos del clúster más cercano. El resultado puede ir de -1 a 1, donde -1 se considera un mal agrupamiento, 0 indiferente y 1 es un buen agrupamiento (León Guzmán, 2016).

3. Resultados y discusión

Filtrado y normalización de datos

En un primer momento, se introdujeron todas las variables, incluyendo cada ítem de cada test. Dado que los resultados con todas las variables no daba ningún agrupamiento se seleccionaron solo los resultados finales de los índices de cada escala. La presencia de enfermedades se contabilizo y se creo una nueva variable llamada "Comorbilidad" donde quedaba representado si la persona tenia más de una enfermedad comórbida.

Por otro lado se estandarizaron las puntuaciones y se normalizaron los datos mediante el filtro *Normalize* de Weka de 0 a 1.

Validación de los datos

Utilizando todas las variables y con el método de inicialización de k-means, el índice de Silhouette era de 0,1082 con la mejor k de 6 (Figura 2 y Figura 3), no encontrándose una estructura substancial. Y de 0,2356 utilizando el método de inicialización *Farthest first* con una k de 4 tampoco encontrándose una estructura substancial. Por este motivo se procedió a seleccionar los mejores atributos y hacer un análisis a partir de esta selección.

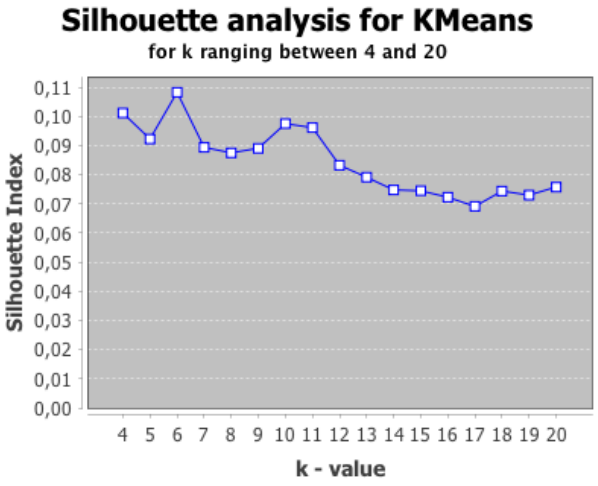


Figura 2. Gráfico a partir de los índices de Silhoutte utilizando el método de inicialización k-means++

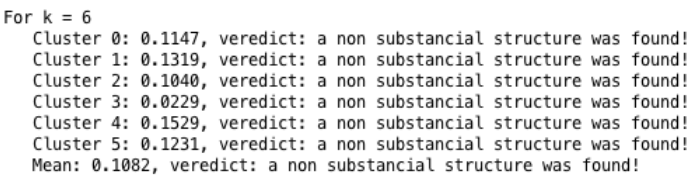


Figura 3. Índices de Silhoutte de k = 6 utilizando el método de inicialización k-means++

En el Apéndice (1. Evaluación de los atributos.)se puede encontrar una tabla con los resultados de la evaluación de los atributos ordenados. Se basa en el cálculo de la ganancia de información (también llamada entropía) de cada atributo para el output de la variable. Los valores de entrada varían de 0 (sin información) a 1 (información máxima). Aquellos atributos que aportan más información tendrán un mayor valor de ganancia de información y se pueden seleccionar, mientras que aquellos que no agregan mucha información tendrán una puntuación más baja y se pueden eliminar.

Los atributos que resultaron no tener ninguna correlación para el agrupamiento fueron: Pueblo/ciudad de la infancia, Índice del Consumo de alcohol, Estado civil, Pueblo/ciudad (actual), Prematuridad, Peso al nacer, Complicaciones al nacer.

Se analizó cuál era el mejor k eliminando estos atributos de los datos, utilizando el método de inicialización *k-means++, Canopy y Farthest First*. Los resultados detallados se pueden encontrar en el apartado de Apéndices 2. Resultados validación de clústeres

Cuando se utilizó *k-means++,* ni la k más óptima ni ninguna más obtuvieron estructura que se pudiera considerar, los mismos resultados se obtuvieron mediante *Canopy*. En cambio, cuando se utilizó el método de *Farthest First* se encontraron estructuras en k 3, k 4 y k 5 aunque esta estructura era débil. En la figura número 4 se muestran los diferentes Índices de Silhouette para un número de clúster de 3 a 20.

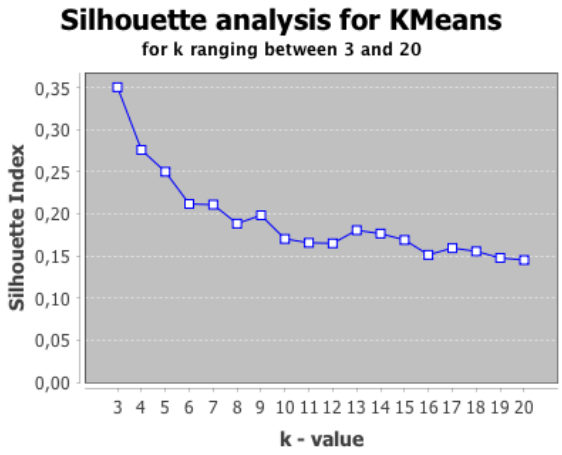


Figura 4. Índices de Silhouette de las diferentes k utilizando farthest first

Para obtener índices de Silhoutte más altos y mejorar la validez de los clústeres, se estableció un punto de corte de 0,1 y se seleccionaron aquellos atributos con una correlación más alta, no se observaban diferencias significativas cuando el punto de corte era más alto. Con este cribado se eliminaron los atributos: índice del *RYFF , PHYSICAL , PROMIS 5 Socialization, MENTAL 3 Nervous , Nivel educacional , Ingresos , Con quien vive, Educación del padre, Educación de la Madre* y *Género.* Y se volvió a analizar el mejor k mediante KValid. Los índices de silhoutte para cada clúster y las gráficas se pueden encontrar en los Apéndices 3. Resultados al reducir atributos.

Los resultados con k-means++ volvieron a ser negativos, y no se obtuvo una estructura substancial. A diferencia, utilizando *Canopy* si se observo estructura esta vez con k de 7,8 y 9, aunque también resulto ser débil. Por último, utilizando *Farthest First* el k más óptimo fue de 4 (iS: Mean: 0.3467) con el que se encontró estructura. También se encontró estructura con k de 5 a k de 8.

Resultados por Algoritmos

Para poder explicar mejor los resultados que se han obtenido con los diferentes algoritmos, se ha resumido en una tabla todos los algoritmos analizados y sus respectivos resultados. En algunos casos, se ha analizado el modelo utilizando

diferentes k, basadas en la validación anterior del mejor k o a partir de la recomendación del propio algoritmo utilizado.

Algoritmo	Nº of k	Distribución	SSE	Otros
K means	4	0 485 (14%) 1 787 (23%) 2 400 (11%) 3 1816 (52%)	2809.4 4	
	6	0 850 (24%) 1 185 (5%) 2 414 (12%) 3 504 (14%) 4 915 (26%) 5 620 (18%	2471.9	
	7	0 271 (8%) 1 237 (7%) 2 373 (11%) 3 218 (6%) 4 1042 (30%) 5 880 (25%) 6 467 (13%)	2280.5 0	
Hierarchal clustering (birch)				
EM	4	0 507 (15%) 1 1840 (53%) 2 627 (18%) 3 514 (15%)		
	6	0 83 (2%) 1 387 (11%) 2 284 (8%) 3 784 (22%) 4 487 (14%) 5 1463 (42%)		
	7	0 78 (2%) 1 617 (18%) 2 630 (18%) 3 282 (8%) 4 267 (8%) 5 1122 (32%) 6 492 (14%)		

Algoritmo	Nº of k	Distribución	SSE	Otros
MTree	4	0 272 (8%) 1 451 (13%) 2 1592 (46%) 3 1173 (34%)	3908.4 0	
	5	0 189 (5%) 1 450 (13%) 2 457 (13%) 3 1575 (45%) 4 817 (23%)	3244.8 0	
	7	0 188 (5%) 1 449 (13%) 2 344 (10%) 3 982 (28%) 4 174 (5%) 5 695 (20%) 6 656 (19%)	3051	
Farthest first	4	0 598 (17%) 1 8 (0%) 2 291 (8%) 3 2591 (74%)		
	7	0 401 (11%) 1 8 (0%) 2 186 (5%) 3 2248 (64%) 4 241 (7%) 5 34 (1%) 6 370 (11%)		
Canopy	13	0 2253 (65%) 1 411 (12%) 2 320 (9%) 3 23 (1%) 4 195 (6%) 5 195 (6%) 6 15 (0%) 7 35 (1%) 8 3 (0%) 9 8 (0%) 10 6 (0%) 11 8 (0%) 12 16 (0%)		
	7	0 2274 (65%) 1 414 (12%) 2 342 (10%) 3 201 (6%) 4 221 (6%)		

Algoritmo	Nº of k	Distribución	SSE	Otros
		5 28 (1%) 6 8 (0%)		
LVQ	7	0 141 (4%) 1 8 (0%) 2 64 (2%) 3 2447 (70%) 4 372 (11%) 5 409 (12%) 6 47 (1%)		
	4	0 423 (12%) 1 82 (2%) 2 456 (13%) 3 2527 (72%)		
DBSCAN	2	0 2465 (77%) 1 755 (23%)		Instancias no agrupadas : 268
	3	0 13 (36%) 1 19 (53%) 2 4 (11%)		Instancias no agrupadas: 3452
	9	0 2481 (75%) 1 816 (25%) 2 6 (0%) 3 2 (0%) 4 5 (0%) 5 2 (0%) 6 2 (0%) 7 2 (0%) 8 2 (0%)		Instancias no agrupadas: 170
SOM	8	0 880 (25%) 1 579 (17%) 2 367 (11%) 3 474 (14%) 4 633 (18%) 5 298 (9%) 6 90 (3%) 7 167 (5%)		

Algoritmo	Nº of k	Distribución	SSE	Otros
	16	0 277 (8%) 1 249 (7%) 2 267 (8%) 3 447 (13%) 4 121 (3%) 5 56 (2%) 6 360 (10%) 7 286 (8%) 8 132 (4%) 9 89 (3%) 10 191 (5%) 11 252 (7%) 12 8 (0%) 13 221 (6%) 14 250 (7%) 15 282 (8%)		
	4	0 1062 (30%) 1 297 (9%) 2 1547 (44%) 3 582 (17%)		
	6	0 730 (21%) 1 829 (24%) 2 1076 (31%) 3 204 (6%) 4 171 (5%) 5 478 (14%)		
MakeDensity BasedClusterer	7	0 254 (7%) 1 117 (3%) 2 1491 (43%) 3 305 (9%) 4 174 (5%) 5 413 (12%) 6 734 (21%)	2312.3 0	
CascadeSimpleK Means	4	0 433 (12%) 1 377 (11%) 2 585 (17%) 3 2093 (60%)		CH: 1663228139

Algoritmo	Nº of k	Distribución	SSE	Otros
	5	0 644 (18%) 1 1519 (44%) 2 386 (11%) 3 646 (19%) 4 293 (8%)		CH: -33050424
	6	0 659 (19%) 1 1498 (43%) 2 499 (14%) 3 234 (7%) 4 222 (6%) 5 376 (11%)		CH: -55339627
	7	0 474 (14%) 1 228 (7%) 2 936 (27%) 3 373 (11%) 4 222 (6%) 5 496 (14%) 6 759 (22%)		CH: -1827216221

K- Means

La suma de errores (SoE, sum of errors) es elevada lo que implica un variedad alta dentro del clúster. Se observa un SoE más bajo cuando la k es de 7, ya que el SoE esta influenciado por el numero de observaciones, cuando menor es el numero de observaciones dentro de cada clúster más bajo tiende a ser el SoE y viceversa.

Aún así, cuando se compara con otros algoritmos utilizados vemos que sí hay un agrupamiento similar y se pueden obtener conclusiones significativas y comparables a otros algoritmos tanto con k de 6 como de 7.

Jerárquico

Como podemos ver en la tabla, no se obtuvieron resultados cuando se utilizó el método clásico de *clustering* jerárquico aglomerativo. Probablemente, la base de datos es muy grande y el algoritmo no consigue completar el proceso.

EM

Este algoritmo por sus características obtiene muy buenos resultados cuando las bases de datos están formadas por modos gaussianos, donde obtiene los clúster correctos incluso cuando hay solapamiento.

En nuestros datos podemos ver que se identifican clústeres comparables con los otros algoritmos y, por lo tanto, es uno de los algoritmos que más nos sirve para poder hacer un perfil de agrupamiento de los datos.

MTree

Es un algoritmo que también utiliza EM para determinar cuándo y cómo se dividirá cada nodo. Pero lo mejora la posibilidad de usar métodos específicos de MTree , como RangeQuery y KNN estos métodos permiten una forma más rápida de obtener vecinos cercanos ofreciendo la posibilidad de excluir algunos clústeres (nodos) de la búsqueda.

Al estar basado en EM, también se han obtenido buenos resultados al utilizar este algoritmo y los agrupamientos que se observan son comparables con los encontrados en los otros algoritmos. El único problema es que no acepta variables nominales y nos interesaba poder utilizarlas, por ejemplo la variable comorbilidad.

Farthest First , LVQ y Canopy

Los resultados de estos algoritmos muestran un clúster que agrupa la gran mayoría de instancias dejando los otros clústeres muy débiles. Por lo cual, no permite hacer un análisis interpretativo preciso.

DBSCAN

Debido a las características de este algoritmo que va enlazando las vecindades más pobladas de cada punto, cuando las bases de datos tienen grupos muy solapados, es incapaz de detectarlos. Este es el caso de nuestros datos, sobre la que no puede obtener los modos. El resultado de disminuir el radio de la vecindad, trae como consecuencia un aumento de los puntos ruido, debido a que a medida que el radio disminuye, la mayoría de puntos se convierten en

puntos ruido. Así se obtienen tres grupos separados con una cantidad muy alta de puntos ruido (o instancias no agrupadas) en dependencia del valor del radio.

SOM

EL algoritmo no permitía la elección de una k de 7 pero se obtuvieron muy buenos resultados utilizando una k de 6 y los agrupamientos se podían equiparar a los conseguidos con los otros algoritmos. Sus resultados se utilizaron junto a los otros algoritmos para realizar el perfil de los agrupamientos.

MakeDensity BasedClusterer

Devuelve la distribución y la densidad utilizando otro algoritmo. Los datos que se extraen son equiparables con los otros algoritmos.

CascadeSimpleKMeans

Es otro algoritmo de validación de número de clúster. Sirvió para ver que con k de 6 y 7 las instancias quedan más equilibradas entre los diferentes grupos.

Descripción de los clústeres

En un primer momento, cuando se quisieron añadir todos los datos y cuestionarios utilizados no se observó ninguna agrupación. Se tuvieron que utilizar diferentes métodos de inicialización (*k-means++, canopy, Farthest First*) y un cribado de los atributos a utilizar para poder hacer la validación de los clústeres y la elección de la k óptima. Después del filtrado, los resultados se mostraron con estructuras débiles y poco robustas, esto se debe a la gran complejidad de la muestra y la cantidad de variables que hace que los resultados se muestren homogéneos. Aún así, se han podido interpretar los grupos y sacar conclusiones. Se seleccionaron k de 4 a 7 como valores de agrupamiento, ya que después del análisis tenían coeficientes de silhoutte más altos. Se preestableció un número mayor de 4 ya que en principio un número menor haría muy difícil la interpretación de los resultados. Por otro lado, dada la falta de estabilidad, los clústeres más altos de 7 tampoco se tuvieron en

cuenta. Una vez se analizaron, los clústeres de 6 y 7 eran los que tenían más valor interpretativo y por lo tanto los que se seleccionaron para la descripción de los diferentes grupos. En los gráficos de visualización que se encuentran en los apéndices (5. Visualizaciones del agrupamiento) podemos ver que tanto los algoritmos que dividen la muestra en 6 grupos como los que la dividen en 7 obtienen perfiles parecidos aunque con algunas diferencias. Los algoritmos que se escogieron por la posterior interpretación fueron el SOM y EM ya que fueron los que mejor agruparon los datos y con más valor interpretativo. Además, en la literatura se encuentran otros grupos que obtuvieron buenos resultados utilizando estos algoritmos. El EM fue utilizado en el proyecto PRESSILA (Sánchez, Barreiro, Pérez, Opisso, & Aguilera, 2016) obteniendo buenos resultados. El grupo de Dipnall et al. (2017) utilizo el SOM , que se ha empezado a utilizar en el ámbito de la salud recientemente, para detectar clústeres de depresión. Cabe resaltar que no se han encontrado diferencias entre grupos referentes al género. Tampoco ha habido diferencias en la variable edad excepto en algún grupo que se observa que están ligeramente por encima de la media. En los apéndices (4. Resultados de la aplicación de los algoritmos) se pueden observar los resultados de medias y desviaciones típicas que se han usado para describir los distintos perfiles pertenecientes a cada *clúster* que se describirán a continuación.

El SOM divide la muestra en 6 clústeres. En la Figura 5 se pueden observar como se agrupan los clústeres a partir del PROMIS I y la variable tipo de Trabajo.

1. Verde. Podemos identificar el primer grupo como personas trabajadoras que acostumbran a tener estudios cualificados y puestos de trabajo altos. Su rendimiento cognitivo y mental es muy alto y con uno resultados en el PROMIS I también elevados. Es decir, son personas con muy buen rendimiento y que perciben su estado de salud como excelente. Tienden a dormir bien y a no fumar. Este grupo es estable utilizando todos los algoritmos y con las diferentes k.

2. Rosa. El segundo grupo tiene la media de edad más alta. Actualmente se encuentra desempleado o jubilado, aún así su nivel de estudios es alto. Tienden a estar casados y vivir con alguien. Su percepción de salud es excelente y sus resultados tanto en el índice mental como cognitivo también son altos. Tienden a dormir bien y no fumar.
3. Rojo. Este grupo consta de gente empleada con estudios superiores. Tienen un BMI ligeramente superior, es decir, pueden tener un peso superior a la media. Consideran que su estado subjetivo de salud es bueno o normal, en ocasiones incluso bajo, contrariamente a los otros grupos que hemos visto que lo consideraban excelente. Aún así, tienen puntuaciones elevadas tanto en las partes del cuestionario relativas a salud mental y aspectos cognitivos.
4. Azul Claro. El siguiente grupo consta de personas trabajadoras. Es el grupo que percibe su estado de salud más bajo. Y tiene un rendimiento entre medio y bajo tanto en índice mental como cognitivo. Son no fumadores, y se observan algunas dificultades para dormir. No están tan determinados a expandir horizontes o crecer personalmente. Este grupo es inestable y es uno de los clúster que más cambios se observan entre diferentes algoritmos y números de k.
5. Azul oscuro: Se caracteriza por un grupo con trabajo remunerado y edad de rango medio. Los resultados del cuestionario mental es alto y el rendimiento cognitivo es intermedio, por debajo de la media de los otros grupos. Su estado subjetivo de la salud es medio, no consideran que sea excelente pero tampoco pobre, empiezan a presentar quejas cognitivas.
6. Amarillo: El último grupo se caracteriza por ser personas no trabajadoras con edad por encima de la media. Presentan algunas dificultades para dormir. Y tienen un rendimiento bajo-medio en el test cognitivo pero es bueno en el test mental. Es el grupo con más enfermedades comórbidas. Este grupo, probablemente, lo constituyen mayores que empiezan su declive cognitivo.

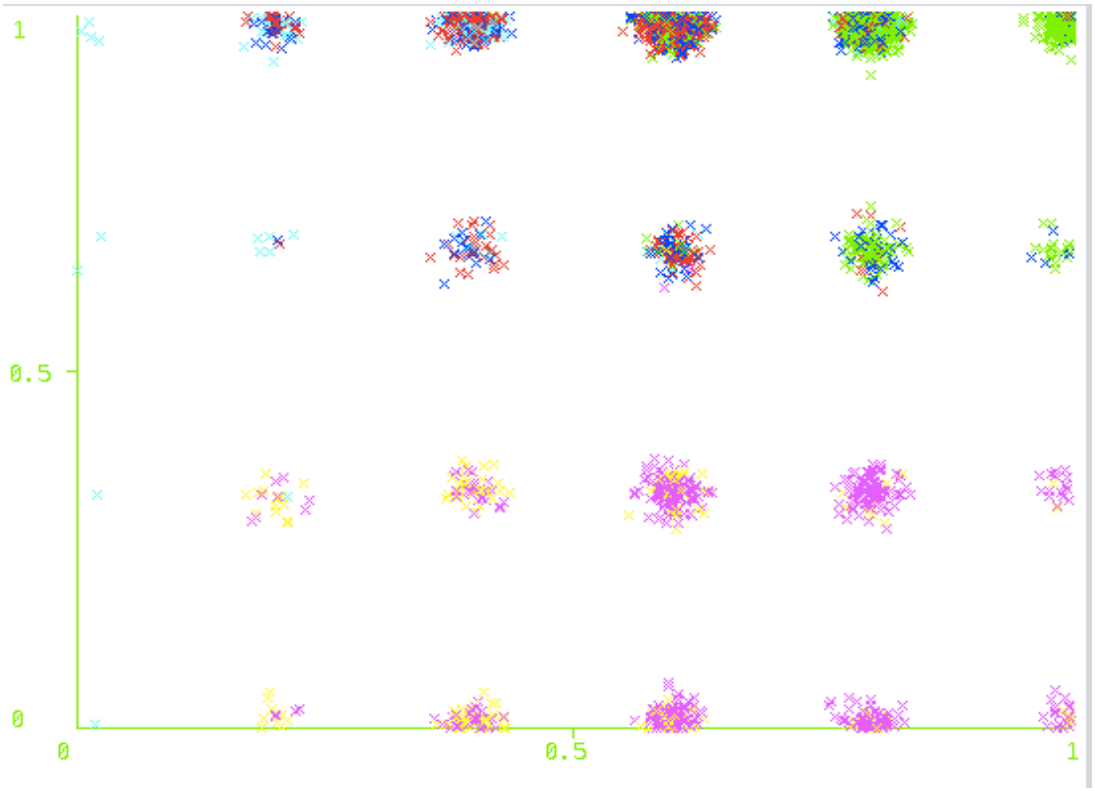


Figura 5. Visualización utilizando SOM de los índices PROMIS I. Salud General (eje x) y Trabajo (Eje y)

Podemos observar que en el SOM una de las variables que más diferencia entre grupos es el atributo “Trabajo”, es decir cual es la situación laboral de cada sujeto. Respecto al EM se describe la agrupación de 7 clústeres (Figura 6). En este caso, la comorbilidad ha tenido más peso a la hora de distribuir los clústeres que en el SOM como se puede observar en la Figura 7):

1. Rojo: Grupo caracterizado por no padecer comorbilidades. Con una edad dentro de la media y con empleo. Su percepción del estado de salud es medio pero tendría un buen rendimiento mental y cognitivo.
2. Amarillo: Contrariamente al anterior tendría una tendencia muy elevada a las comorbilidades, también se observaría un BMI ligeramente superior y su estado de salud subjetivo es medio. Las puntuaciones en los índices cognitivo y mental son elevados. Con una edad dentro de la media y su situación laboral es activa. También muestran ligeros problemas para dormir.

3. Azul oscuro: Este grupo muestra una percepción de la salud medio. Las puntuaciones en los índices mentales son altas pero el rendimiento cognitivo es medio-bajo, es el segundo grupo con peores resultados. Se aprecian ligeros problemas para dormir, pero no tienden a padecer comorbilidades.
4. Azul claro: El siguiente grupo se caracteriza por ser el grupo con las puntuaciones más bajas en percepción de la salud y en los índices cognitivos, también se observa un BMI ligeramente superior. El rendimiento mental, aunque se sitúa en rango medio, también es inferior que en los otros grupos. Se observan problemas para dormir y tendencia a padecer comorbilidades.
5. Negro: Son el grupo de mayor edad y se encuentran en situación de desempleo. Su percepción de la salud es medio-bajo, tienen un rendimiento mental alto y los índices cognitivos en rango medio-alto, con tendencia a padecer comorbilidades.
6. Rosa: Este grupo se caracteriza por personas no trabajadoras y de edad superior. Su rendimiento mental es alto y el cognitivo también. Tiene tendencia a no padecer comorbilidades.
7. Verde: Muy buen estado subjetivo de salud percibida, y muy buen rendimiento tanto cognitivo como mental. Observamos una proporción menor de fumadores. Y una alta tendencia a no padecer comorbilidades.

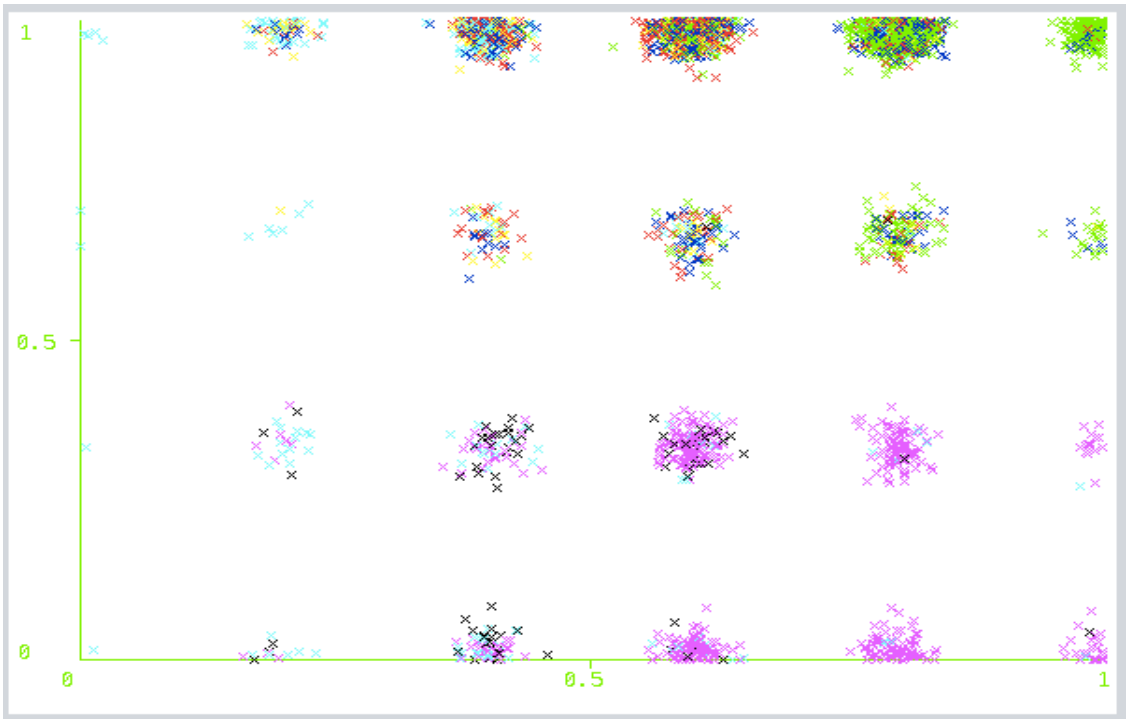


Figura 6. Visualización clasificación mediante EM. PROMIS I. Salud General (eje x) y Trabajo (eje y).

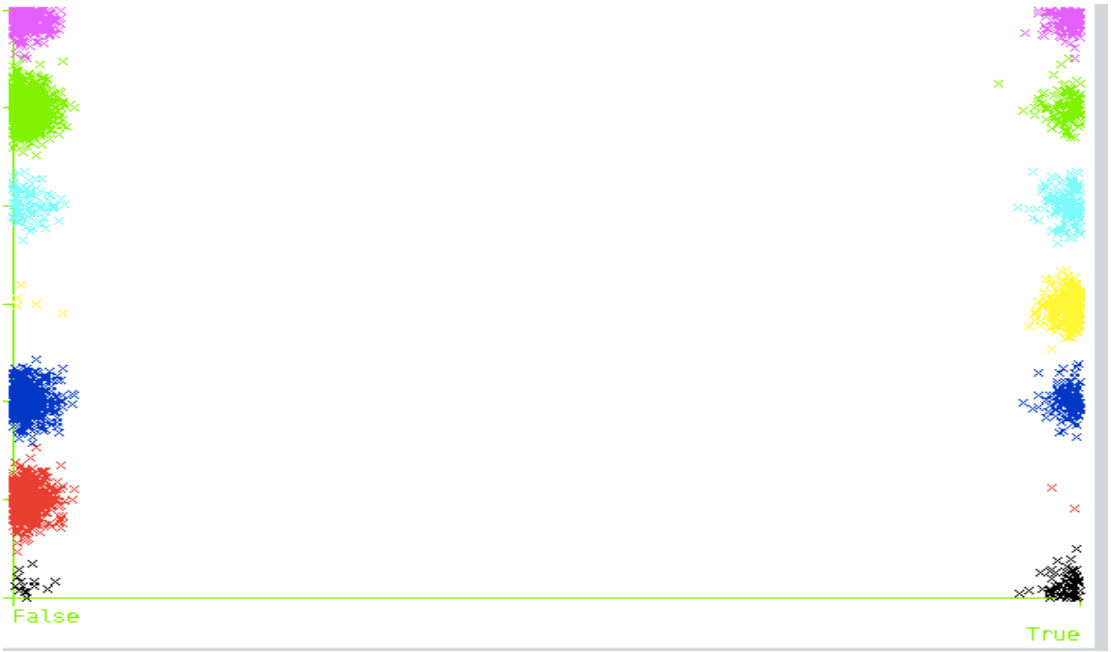


Figura 7. Visualización de clústeres utilizando EM según la comorbilidad.

4. Discusión y conclusiones

En la literatura no se encuentran otros estudios que intenten clasificar los hábitos de vida en población sana utilizando técnicas de inteligencia artificial. Se puede encontrar algunos trabajos que clasifican los estilos de vida en población con diversos tipos de patología como en la depresión (J. F. Dipnall et al., 2017; Joanna F. Dipnall et al., 2016) , Enfermedad de Alzheimer (Nieto-Reyes, Duque, Montaña, & Lage, 2017; Zhang, Wang, Zhou, Yuan, & Shen, 2011), síndrome metabólico (Ha, Choi, & Lee, 2017) o enfermedades cardiovasculares (Sandbakk et al., 2016). Por otro lado, encontramos trabajos de clasificación de los hábitos de vida sin utilizar técnicas de inteligencia artificial (Dodd, Al-Nakeeb, Nevill, & Forshaw, 2010; Morris, D'Este, Sargent-Cox, & Anstey, 2016; Schuit, Van Loon, Tijhuis, & Ocké, 2002), lo que no permite utilizar tantas variables, y su enfoque es más bien confirmatorio y no exploratorio como es nuestro caso ya que no partimos de un supuesto de partida y pretendemos buscar conocimiento nuevo y susceptible de proporcionarnos nueva información.

A mayor dimensionalidad del problema el data mining ofrece mejores soluciones y nos permite encontrar relaciones inéditas para luego concretar la investigación sobre las variables más interesantes. Una vez hecho el análisis, uno de los atributos que más valor nos aporta a la hora de discriminar entre grupos es el PROMIS I, sobretodo el índice de salud general. La percepción del adulto acerca de su estado de salud y calidad de vida son influidos por su salud mental y capacidad funcional (Azpiazu et al., 2002; Love, Goldman & Rodríguez, 2008). Por tanto, la percepción de salud es un constructo asociado a otras variables psicológicas como autoestima, satisfacción con la vida y depresión (Mella et al., 2004; Winocur, Palmer, Dawson, Binns, Bridges & Stuss, 2007) y ha probado tener asociaciones significativas con otros indicadores más objetivos, como son el número de enfermedades crónicas, el periodo de tiempo que han vivido con una enfermedad, la agudización de problemas crónicos, etc. (Beaman, Reyes, García-Peña & Cortés, 2004). Estas asociaciones se reflejan, en los datos, positivamente en el grupo número 7, ya que muestran un alto rendimiento cognitivo y tendencia a no tener

comorbilidades y un muy buen estado subjetivo y , de forma negativa en el grupo 4, que tiene puntuaciones bajas en rendimiento y un bajo estado subjetivo de salud. El rendimiento se ha evaluado con el test cognitivo, que también es un buen atributo para segregar los grupos, contrariamente al test mental que todas las puntuaciones tienden a encontrarse en valores altos y por lo tanto no separa entre clústeres. El PROMIS II, aunque sí ha afectado la clasificación de los clústeres no nos permite hacer una diferenciación ya que todos los grupos se agrupan de forma homogénea en este test.

Schuit et al., (2002) encontraron altas asociaciones entre el consumo de alcohol y de tabaco como factores de riesgo en deterioro de salud. En los datos del presente trabajo, aunque se pueden sacar pocas conclusiones del índice "smoking" ya que los datos no se agrupan, sí se puede ver cómo los grupos con puntuaciones altas en estado subjetivo de salud tienen menos instancias con puntuaciones altas, es decir tienen menos sujetos fumadores. Otras variables de gran peso que se encuentran en la literatura son la actividad física y el consumo de alcohol, pero nuestros resultados los señalan como los peores índices ya que no han mostrado agrupaciones entre los diferentes clústeres y no han tenido el peso que esperábamos. Esto puede ser porque al formar parte de una amplia batería de cuestionarios su peso haya quedado atenuado otra posibilidad es que la forma en que se puntúan estos tests haya afectado a la clasificación negativamente, o bien porque los cuestionarios empleados no sean fácilmente traducibles en una cantidad específica de actividad física o consumo de alcohol (sirven para medir cualitativamente, pero no tanto cuantitavemente). Por tanto, habrá que valorar si la incorporación de otros cuestionarios que nos den una información más cuantitativa aportan valor a la clasificación.

Nuestros resultados también apuntan a la correlación del sueño con la salud cerebral. Son bien conocidas las graves consecuencias físicas y psicosociales de trastornos del sueño, como el insomnio o la apnea (Roth & Ancoli-Israel, 1999). Asimismo, es muy frecuente la presencia de problemas del sueño en diversas afecciones médicas y trastornos psicopatológicos (Benca, Obermeyer, Thisted & Gillin, 1992). Mientras, los que duermen aproximadamente 7-8 horas son los que gozan de mayores ventajas tanto a nivel físico como psicológico (Miró, E., & Iáñez, M., & Cano-Lozano, M. , 2002). Finalmente no se han

podido tener en cuenta índices como la nutrición o la reserva cognitiva que podrían haber aportado más información a al clasificación por clústeres y son variables que también son clave en la salud cerebral (Yannakoulia, Kontogianni, & Scarmeas, 2015).

El trabajo consta con numerosas fortalezas como son el elevado número de muestra. Aún así se tienen que destacar algunas de las limitaciones con las que nos hemos encontrado. Una de ellas se pudo detectar con K-means, método finalmente no utilizado porque no proporcionaba un buen agrupamiento al ser muy sensible a los *outliers,* aunque como ya se ha comentado los clústeres resultantes se podían comparar con los encontrados en el EM y en el SOM. Este algoritmo segregó un clúster con aquellos sujetos que de forma errónea no habían respondido a los test mental y cognición pero la respuesta se había contabilizado como 0. Este aspecto, aunque sucedía con pocas instancias, puede haber afectado a nuestros resultados contando algunos sujetos como cognitivamente en mal estado cuando en realidad no habían llegado a responder el cuestionario. Por otro lado, esperábamos que las variables relacionadas con fumar y consumir alcohol tuvieran más peso, para futuros trabajos se tendría que utilizar otros test que evaluaran estas variables o estandarizarlas de forma distinta. Otra variable que tiene mucha importancia es la comorbilidad, cuando se normalizaron los datos se conto como comorbilidad todo tipo de enfermedades sin hacer distinción de causa, gravedad, o órgano afectado. Probablemente si se dividiera esta variable se obtendrían clústeres más robustos.

En conclusión, los hábitos de vida en la población adulta sana se agrupan. La tendencia de los hábitos de vida a agregarse tiene implicaciones importantes tanto para la promoción como para la prevención de la salud ya que señala aquellos factores de riesgo y aquellos estilos de vida que actúan como protectores. Las variables que más han agrupado han sido la situación laboral del sujeto, su autopercepción de salud, si padecen comorbilidades y su rendimiento en tests cognitivos.

La información sobre los subgrupos ayudará a planificar futuras estrategias y ha identificar aquellos factores que puedan aumentar la motivación y la adhesión a hábitos saludables.

5. Bibliografía

Ader, D. (2007). Developing the Patient-Reported Outcomes Measurement Information System (PROMIS). *Medical Care*, *45*(5 Supplement 1), S3–S11. http://doi.org/10.1097/01.mlr.0000258615.42478.55.The

Azpiazu, M., Jentoft, A., Villagrasa, J., Abanades, J., García, N., & Alvear, F. (2002). Factores asociados a mal estado de salud percibido o mala calidad de vida en personas mayores de 65 años. *Revista Española de Salud Publica, 76,* 683-699.

Beaman, P., Reyes, S., García, C., & Cortés, A. (2004). Percepción de la salud entre los adultos mayores derechohabientes del Instituto Mexicano del Seguro Social. En O. Muñoz, C. García, & L. Durán (Eds.), *La salud del adulto mayor* (pp. 117-138). México: Instituto Mexicano del Seguro Social. Recuperado el 10 de septiembre de 2009 desde: http://www.bibliotecas.salud.gob.mx/gsdl/collect/publin1/index/assoc/HASHd0ef.dir/doc.pdf

Behbahani, S., & Nasrabadi, A. M. (2009). Application of SOM neural network in clustering. *J. Biomedical Science and Engineering*, *2*(December), 637–643. http://doi.org/10.4236/jbise.2009.28093

Benca, R. M., Obermeyer, W. H., Thisted, R. A. & Gillin, J. C. (1992). Sleep and psychiatric disorders: A meta-analysis. *Archives of General Psychiatry*, 49, 651-668.

Bosch Capdevilla, B. (2010). Influencia de la reserva cognitiva en la estructura y funcionalidad cerebral en el envejecimiento sano y patológico. Barcelona: Universidad de Barcelona.

Calabria, M., Cattaneo, G., & Costa, A. (2017). It is time to project into the future: 'Bilingualism in healthy and pathological aging.' *Journal of Neurolinguistics*, *43*(August), 1–3. http://doi.org/10.1016/j.jneuroling.2017.03.003

Cella, D., Lai, J. S., Nowinski, C. J., Victorson, D., Peterman, A., Miller, D., … Moy, C. (2012). Neuro-QOL: Brief measures of health-related quality of life for clinical research in neurology. *Neurology*, *78*(23), 1860–1867. http://doi.org/10.1212/WNL.0b013e318258f744

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). *Maximum likelihood from*

incomplete data via the EM algorithm. *J. R. Statist. Soc. A* (Vol. 39). http://doi.org/10.2307/2984875

Dipnall, J. F., Pasco, J. A., Berk, M., Williams, L. J., Dodd, S., Jacka, F. N., & Meyer, D. (2016). Fusing data mining, machine learning and traditional statistics to detect biomarkers associated with depression. *PLoS ONE*, *11*(2), 1–23. http://doi.org/10.1371/journal.pone.0148195

Dipnall, J. F., Pasco, J. A., Berk, M., Williams, L. J., Dodd, S., Jacka, F. N., & Meyer, D. (2017). Why so GLUMM? Detecting depression clusters through graphing lifestyle-environs using machine-learning methods (GLUMM). *European Psychiatry*, *39*, 40–50. http://doi.org/10.1016/j.eurpsy.2016.06.003

Dodd, L. J., Al-Nakeeb, Y., Nevill, A., & Forshaw, M. J. (2010). Lifestyle risk factors of students: A cluster analytical approach. *Preventive Medicine*, *51*(1), 73–77. http://doi.org/10.1016/j.ypmed.2010.04.005

Dunn, J. C. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, *3*(3), 32–57. http://doi.org/10.1080/01969727308546046

Ester, M., Kriegel, H. P., Sander , J. and Xu, X.(1996) A density-based algorithm for discovering clusters in large special databases with noise. *International Conference on Knowledge Discovery and Data Mining* (KDD´96). 226 – 231

Fisher, Douglas (1987). *"Knowledge acquisition via incremental conceptual clustering" (PDF). Machine Learning. 2* (2): 139–172.*doi*:*10.1007/BF00114265.*

Fisher, Douglas H. (1987). "Improving inference through conceptual clustering". *Proceedings of the 1987 AAAI Conferences. AAAI Conference. Seattle Washington.* pp. 461–465.

Friedman, J. H. (1997). Data Mining and Statistics: what's the connection? *Computing Science and Statistics*, (May), 3–9.

Ha, S., Choi, H. R., & Lee, Y. H. (2017). Clustering of four major lifestyle risk factors among Korean adults with metabolic syndrome. *PLoS ONE*, *12*(3), 1–9. http://doi.org/10.1371/journal.pone.0174567

Jenkins, C. D., Stanton, B.-A., Niemcryk, S. J., & Rose, R. M. (1988). a Scale for the Estimation of Sleep Problems in Clinical Research. *J Cltn Epbmiil*, *41*(4), 313–321. http://doi.org/10.1016/0895-4356(88)90138-2

Kaeberlein, M., Rabinovitch, P. S., & Martin, G. M. (2015). Healthy aging: The ultimate preventative medicine. *Science* (New York, N.Y.), 350 (6265), 1191-1193

Kaufman, L., & Rousseuw, P. J. (1991). Finding Groups in Data: An Introduction to Cluster Analysis. *Biometrics*, *47*(2), 788. http://doi.org/10.2307/2532178

Kennedy ', J., & Eberhart2, R. (1995). Particle Swarm Optimization. *IEEE International Conference on Neural Networks*, *Vol. IV*(Piscataway, NJ), 1942–1948. Retrieved from https://www.cs.tufts.edu/comp/150GA/homeworks/hw3/_reading6 1995 particle swarming.pdf

Kroenke, K., Spitzer, R. L., Williams, J. B. W., & Löwe, B. (2009). An ultra-brief screening scale for anxiety and depression: the PHQ-4. *Psychosomatics*, *50*(6), 613–621. http://doi.org/10.1016/S0033-3182(09)70864-3

León Guzmán, E. (2016). Métricas para la validación de Clustering Contenido. Retrieved from http://www.disi.unal.edu.co/profesores/eleonguz/cursos/md/presentaciones/Sesion13_validacion_Clustering.pdf

Li, X; Zaïane, O.R.; Li, Z., E. (2006). *Advanced Data Mining and Applications: Second International Conference, ADMA 2006, Xi'An, China, August 2006, Proceedings*. (J. Carbonell, J. G. and Siekmann, Ed.). Berlin: Springer.

Livingston, G., Sommerlad, A., Orgeta, V., Costafreda, S. G., Huntley, J., Ames, D., … Mukadam, N. (2017). Dementia prevention, intervention, and care. *The Lancet*, *6736*(17). http://doi.org/10.1016/S0140-6736(17)31363-6

Love, A., Goldman, N., & Rodríguez, G. (2008). Is positive well-being protective of mobility limitations among older adults?. *Journal of Gerontology, 63,* 321-327.

Mella, R., González, L., D' Appolonio, J., Maldonado, I., Fuenzalida, A., & Díaz, A. (2004). Factores asociados al bienestar subjetivo en el adultomayor. *Psykhe, 13,* 79-89

Miró, E., & Ibáñex, M. & Cano-Lozano, M. (2002). Patrones de sueño y salud. *International Journal of Clinical and Health Psychology, 2* (2), 301-326

Montañés Ibáñez, A. (2012). Enfermedad de Alzheimer y Reserva cognitiva. Publicaciones didácticas.

Morris, L. J., D'Este, C., Sargent-Cox, K., & Anstey, K. J. (2016). Concurrent lifestyle risk factors: Clusters and determinants in an Australian sample.

Preventive Medicine, *84*, 1–5. http://doi.org/10.1016/j.ypmed.2015.12.009

Nieto-Reyes, A., Duque, R., Montaña, J. L., & Lage, C. (2017). Classification of alzheimer's patients through ubiquitous computing. *Sensors (Switzerland)*, *17*(7), 1–18. http://doi.org/10.3390/s17071679

Ortoleva Bucher, C., Dubuc, N., von Gunten, A., Trottier, L., & Morin, D. (2016). Development and validation of clinical profiles of patients hospitalized due to behavioral and psychological symptoms of dementia. *BMC Psychiatry*, *16*(1), 1–12. http://doi.org/10.1186/s12888-016-0966-7

Redolat, R. y Carrasco, M.C. (1998) ¿Es la plasticidad cerebral un factor crítico en el tratamiento de las alteraciones cognitivas asociadas al envejecimiento? *Anales de psicología*, 14, 45-53

Redolat, R., & Mesa-Gresa, P. (2015). Brain health as a key concept in the development of stretegies for dalaying age-related cognitive decline and Alzeimer's disease, *Journal of Parkinson's Disease & Alzheimer's Disease*, 2-4

Rodríguez Álvarez, M. y Sánchez Rodríguez, J.L. (2004). Reserva cognitiva y demencia. *Anales de psicología,* 20 (2), 175-186.

Rodríguez, J., & Minoletti, A. (2013). Manual de salud mental para trabajadores de atención primaria. *Paltex Para Técnicos Médios y Auxiliares*, (25), 235. Retrieved from https://xa.yimg.com/kq/groups/70925778/1769151768/name/SaludMental_paratrabajadores_APS.pdf

Roth, T. & Ancoli-Israel, S. (1999). Daytime consequences and correlates of insomnia in the United States: Results of the 1991 National Sleep Foundation Survey II. *Sleep*, 22 , 354-358

Sánchez, J. S., Barreiro, F. J. G., Pérez, E. H., Opisso, E., & Aguilera, E. J. G. (2016). Plataforma PERSSILAA : Algoritmos y herramientas de ayuda a la decisión para la prevención de la fragilidad en personas mayores, 143–146.

Sandbakk, S. B., Nauman, J., Zisko, N., Sandbakk, Ø., Aspvik, N. P., Stensvold, D., & Wisløff, U. (2016). Sedentary Time, Cardiorespiratory Fitness, and Cardiovascular Risk Factor Clustering in Older Adults–the Generation 100 Study. *Mayo Clinic Proceedings*, *91*(11), 1525–1534. http://doi.org/10.1016/j.mayocp.2016.07.020

Schäfer, I., von Leitner, E. C., Schön, G., Koller, D., Hansen, H., Kolonko, T., …

van den Bussche, H. (2010). Multimorbidity patterns in the elderly: A new approach of disease clustering identifies complex interrelations between chronic conditions. *PLoS ONE*, *5*(12). http://doi.org/10.1371/journal.pone.0015941

Schröder, H., Fitó, M., Estruch, R., Martínez-González, M A Corella, D., Salas-Salvadó, J., Lamuela-Ravento, R., … Covas, M. I. (2010). A Short Screener Is Valid for Assessing Mediterranean Diet Adherence among Older. *The Journal of Nutrition*, *141*(6), 1140–1145. http://doi.org/10.3945/jn.110.135566.dietary

Schuit, A. J., Van Loon, A. J. M., Tijhuis, M., & Ocké, M. C. (2002). Clustering of lifestyle risk factors in a general adult population. *Preventive Medicine*, *35*(3), 219–224. http://doi.org/10.1006/pmed.2002.1064

Skoog, I., Börjesson-Hanson, A., Kern, S., Johansson, L., Falk, H., Sigström, R., & Östling, S. (2017). Decreasing prevalence of dementia in 85-year olds examined 22 years apart: the influence of education and stroke. *Scientific Reports*, *7*. http://doi.org/10.1038/s41598-017-05022-8

William Iba and Pat Langley. "Cobweb models of categorization and probabilistic concept formation". In Emmanuel M. Pothos and Andy J. Wills,. *Formal approaches in categorization*. Cambridge: Cambridge University Press. pp. 253–273. *ISBN 9780521190480*.

Winocur, G., Palmer, H., Dawson, D., Binns, M., Bridges, K., & Stuss, D. (2007). Cognitive rehabilitation in the elderly: An evaluation of psychosocial factors. *Journal of the International Neuropsychological Society, 13,* 153-165.

Witten, I. H., Frank, E., & Hall, M. a. (2011). *Data Mining: Practical Machine Learning Tools and Techniques (Google eBook)*. *Complementary literature None*. http://doi.org/0120884070, 9780120884070

Yannakoulia, M., Kontogianni, M., & Scarmeas, N. (2015). Cognitive health and Mediterranean Diet: Just diet or lifestyle pattern? *Ageing Research Reviews*, *20*, 74–78. http://doi.org/10.1016/j.arr.2014.10.003

Yoo, I., Alafaireet, P., Marinov, M., Pena-Hernandez, K., Gopidi, R., Chang, J. F., & Hua, L. (2012). Data mining in healthcare and biomedicine: A survey of the literature. *Journal of Medical Systems*, *36*(4), 2431–2448. http://doi.org/10.1007/s10916-011-9710-5

Zhang, D., Wang, Y., Zhou, L., Yuan, H., & Shen, D. (2011). Multimodal classification of Alzheimer's disease and mild cognitive impairment. *NeuroImage*, *55*(3), 856–867. http://doi.org/10.1016/j.neuroimage.2011.01.008

6. Apéndices

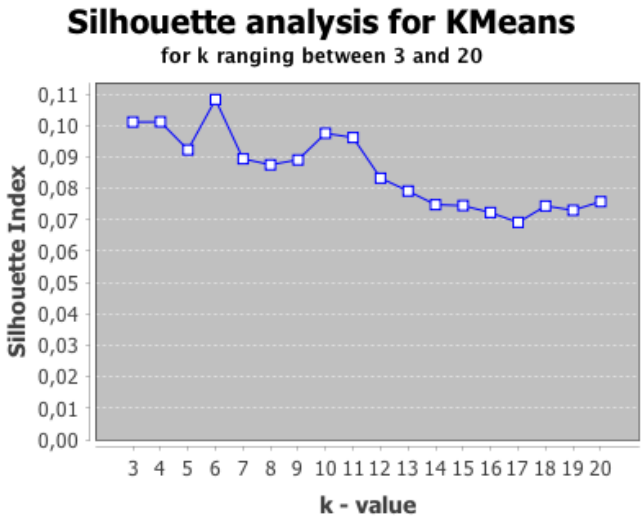
1. Evaluación de los atributos

Ranked attributes:	Attribute number	Attribute Name
0.082015	18	PROMIS 1 health
0.059083	20	PROMIS 3 Physical
0.033922	1	Age
0.029437	5	BMI N
0.0264	32	COGNITIVE 1 Memory
0.026292	47	SLEEP N
0.026013	21	PROMIS 4 Mental
0.025196	19	PROMIS 2 QoL
0.02184	40	COGNITIVE 9 Concentrating
0.021248	41	COGNITIVE 10 Remembering
0.020891	38	COGNITIVE 7 Slower
0.020248	43	COGNITIVE 12 Planning
0.019744	34	COGNITIVE 3 Interrupted
0.019524	7	Work
0.019443	37	COGNITIVE 6 Thinking
0.019284	25	PROMIS II 2 Pain
0.019084	8	Kind job
0.018273	36	COGNITIVE 5 Recalling
0.017899	39	COGNITIVE 8 Pay attention
0.016861	33	COGNITIVE 2 Read
0.015017	28	MENTAL 1 Interest
0.014221	35	COGNITIVE 4 Two things
0.013049	46	SMOKING N
0.012756	23	PROMIS 6 Daily life
0.012494	24	PROMIS II 1 Daily living
0.012423	29	MENTAL 2 Sad
0.0117	42	COGNITIVE 11 Decisions
0.010914	31	MENTAL 4 Uncontrol
0.010624	27	PROMIS II 4 Emotional
0.01042	26	PROMIS II 3 Fatigue
0.008888	48	RYFF N
0.008001	44	PHYSICAL N
0.007902	22	PROMIS 5 Socialization

0.007594	30	MENTAL 3 Nervous
0.006068	6	Educational level
0.005463	10	Incomes
0.004952	4	Live with
0.003559	14	Father´s education
0.003354	15	Mother´s education
0.000196	2	Gender
0	17	Town/village child
0	45	DRINKING N
0	3	Marital status
0	9	Town/village
0	16	House childhood
0	11	Premature
0	13	Birth weight
0	12	Complications at birth

2. Resultados validación de clústeres

2.1. Resultados utilizando k-means++



For k = 6

clúster 0: 0.1147, veredict: a non substancial structure was found!

clúster 1: 0.1319, veredict: a non substancial structure was found!

clúster 2: 0.1040, veredict: a non substancial structure was found!

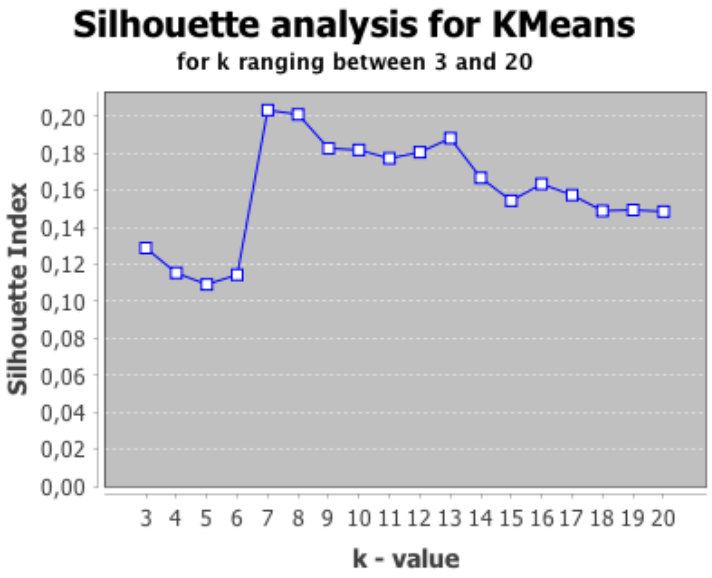
clúster 3: 0.0229, veredict: a non substancial structure was found!

clúster 4: 0.1529, veredict: a non substancial structure was found!

clúster 5: 0.1231, veredict: a non substancial structure was found!

Mean: 0.1082, veredict**: a non substancial structure was found**!

2.2. Resultados utilizando Canopy



For k = 7

clúster 0: 0.2444, veredict: a non substancial structure was found!

clúster 1: 0.1485, veredict: a non substancial structure was found!

clúster 2: 0.0725, veredict: a non substancial structure was found!

clúster 3: 0.0552, veredict: a non substancial structure was found!

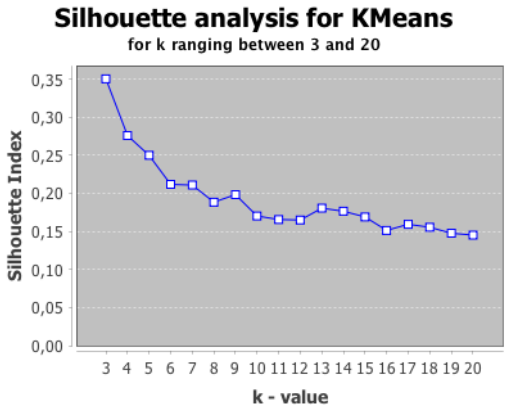
clúster 4: 0.0579, veredict: a non substancial structure was found!

clúster 5: 0.1495, veredict: a non substancial structure was found!

clúster 6: 0.6960, veredict: reasonably structure!

Mean: 0.2034, veredict: **a non substancial structure was found**!

2.3. Resultados utilizando Farther First



For k = 3

clúster 0: 0.2285, veredict: a non substancial structure was found!

clúster 1: 0.7560, veredict: strong structure!

clúster 2: 0.0675, veredict: a non substancial structure was found!

Mean: 0.3506, veredict: **weak structure**!

For k = 4

clúster 0: 0.2060, veredict: a non substancial structure was found!

clúster 1: 0.7164, veredict: strong structure!

clúster 2: 0.0739, veredict: a non substancial structure was found!

clúster 3: 0.1081, veredict: a non substancial structure was found!

Mean: 0.2761, veredict: **weak structure**!

For k = 5

clúster 0: 0.1773, veredict: a non substancial structure was found!

clúster 1: 0.7263, veredict: strong structure!

clúster 2: 0.0433, veredict: a non substancial structure was found!

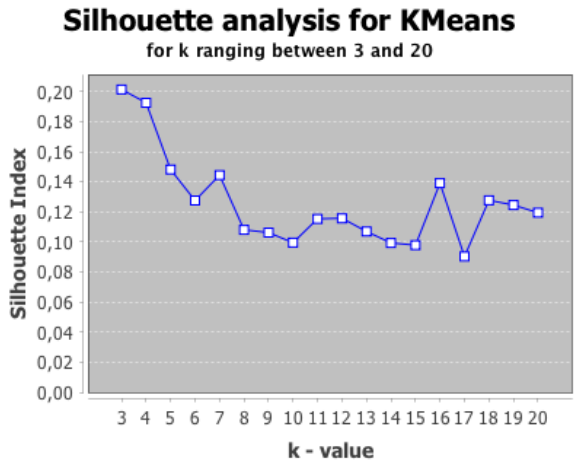
clúster 3: 0.1179, veredict: a non substancial structure was found!

clúster 4: 0.1855, veredict: a non substancial structure was found!

Mean: 0.2501, veredict: **weak structure**!

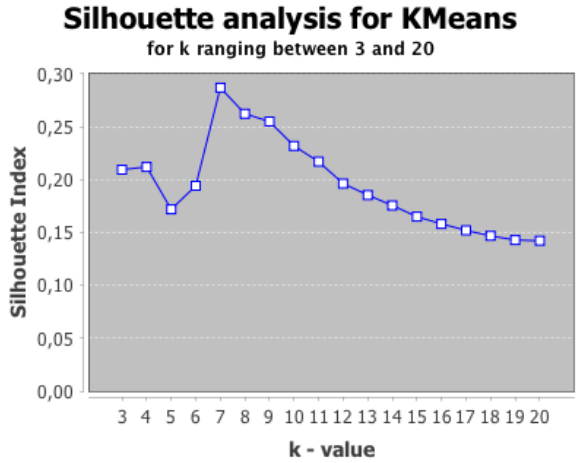
3. Resultados al reducir atributos

3.1. Resultados utilizando k-means++



Ni la k más óptima ni ninguna más obtuvieron estructura utilizando k-means++ para inicializar el proceso.

3.2. Resultados utilizando Canopy (menos atributos)



For k = 7

clúster 0: 0.2803, veredict: weak structure!

clúster 1: 0.2787, veredict: weak structure!

clúster 2: 0.2295, veredict: a non substancial structure was found!

clúster 3: 0.0846, veredict: a non substancial structure was found!

clúster 4: 0.0927, veredict: a non substancial structure was found!

clúster 5: 0.2218, veredict: a non substancial structure was found!

clúster 6: 0.8205, veredict: strong structure!

Mean: 0.2869, veredict**: weak structure**!

For k = 8

clúster 0: 0.2780, veredict: weak structure!

clúster 1: 0.2847, veredict: weak structure!

clúster 2: 0.2349, veredict: a non substancial structure was found!

clúster 3: 0.0872, veredict: a non substancial structure was found!

clúster 4: 0.0935, veredict: a non substancial structure was found!

clúster 5: 0.1011, veredict: a non substancial structure was found!

clúster 6: 0.8234, veredict: strong structure!

clúster 7: 0.1951, veredict: a non substancial structure was found!

Mean: 0.2622, veredict: **weak structure!**

For k = 9

clúster 0: 0.2784, veredict: weak structure!

clúster 1: 0.2732, veredict: weak structure!

clúster 2: 0.2296, veredict: a non substancial structure was found!

clúster 3: 0.1230, veredict: a non substancial structure was found!

clúster 4: 0.0889, veredict: a non substancial structure was found!

clúster 5: 0.1039, veredict: a non substancial structure was found!

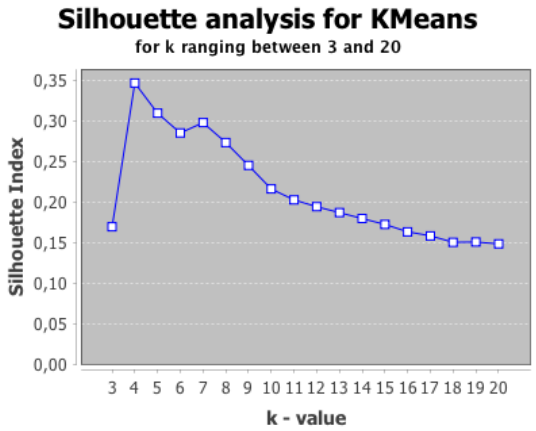
clúster 6: 0.8272, veredict: strong structure!

clúster 7: 0.1121, veredict: a non substancial structure was found!

clúster 8: 0.2575, veredict: weak structure!

Mean: 0.2549, veredict: **weak structure**!

3.3. Resultados utilizando Farther First



For k = 3

clúster 0: 0.3111, veredict: weak structure!

clúster 1: 0.0578, veredict: a non substancial structure was found!

clúster 2: 0.1390, veredict: a non substancial structure was found!

Mean: 0.1693, **veredict: a non substancial structure was found!**

For k = 4

clúster 0: 0.2514, veredict: weak structure!

clúster 1: 0.8305, veredict: strong structure!

clúster 2: 0.0942, veredict: a non substancial structure was found!

clúster 3: 0.2107, veredict: a non substancial structure was found!

Mean: 0.3467, **veredict: weak structure!**

For k = 5

clúster 0: 0.3138, veredict: weak structure!

clúster 1: 0.8551, veredict: strong structure!

clúster 2: 0.0428, veredict: a non substancial structure was found!

clúster 3: 0.2382, veredict: a non substancial structure was found!

clúster 4: 0.0974, veredict: a non substancial structure was found!

Mean: 0.3095, **veredict: weak structure!**

For k = 6

clúster 0: 0.2918, veredict: weak structure!

clúster 1: 0.8361, veredict: strong structure!

clúster 2: 0.0799, veredict: a non substancial structure was found!

clúster 3: 0.2387, veredict: a non substancial structure was found!

clúster 4: 0.0984, veredict: a non substancial structure was found!

clúster 5: 0.1651, veredict: a non substancial structure was found!

Mean: 0.2850, **veredict: weak structure!**

For k = 7

clúster 0: 0.2641, veredict: weak structure!

clúster 1: 0.8230, veredict: strong structure!

clúster 2: 0.0991, veredict: a non substancial structure was found!

clúster 3: 0.2681, veredict: weak structure!

clúster 4: 0.1035, veredict: a non substancial structure was found!

clúster 5: 0.2543, veredict: weak structure!

clúster 6: 0.2722, veredict: weak structure!

Mean: 0.2977, **veredict: weak structure!**

For k = 8

clúster 0: 0.2798, veredict: weak structure!

clúster 1: 0.8255, veredict: strong structure!

clúster 2: 0.0932, veredict: a non substancial structure was found!

clúster 3: 0.2672, veredict: weak structure!

clúster 4: 0.1278, veredict: a non substancial structure was found!

clúster 5: 0.2297, veredict: a non substancial structure was found!

clúster 6: 0.2721, veredict: weak structure!

clúster 7: 0.0909, veredict: a non substancial structure was found!

Mean: 0.2733, **veredict: weak structure!**

4. Resultados de la aplicación de los algoritmos

4.1. Resultados aplicando algoritmo EM con 7 clústers

Number of clústeres : 7

Number of iterations performed: 1

Attribute	0	1	2	3	4	5	6
	(0.05)	(0.21)	(0.15)	(0.11)	(0.08)	(0.28)	(0.11)
Age							
mean	0.6761	0.4506	0.4507	0.527	0.5197	0.4507	0.6385
std. dev.	0.1634	0.1565	0.1489	0.1575	0.1819	0.164	0.1866
BMI N							
mean	0.4246	0.3756	0.3604	0.4288	0.417	0.3494	0.3679
std. dev.	0.1357	0.1291	0.1162	0.1508	0.1549	0.1078	0.1212
Work							
mean	0.2254	0.9655	0.96	0.9602	0.7152	0.9703	0.1739
std. dev.	0.1695	0.1016	0.1083	0.1081	0.3809	0.0949	0.1665
Kind job							
mean	0.0091	0.8421	0.8421	0.8432	0.5681	0.8754	0
std. dev.	0.0739	0.1387	0.151	0.1469	0.3979	0.1295	0.0001
PROMIS 1 health							
mean	0.5475	0.6135	0.6506	0.5795	0.4401	0.8058	0.6978
std. dev.	0.1605	0.1359	0.1649	0.1643	0.1981	0.1224	0.1676
PROMIS 2 QoL							
mean	0.6344	0.627	0.6631	0.653	0.5111	0.8131	0.7276
std. dev.	0.1613	0.1311	0.1573	0.1498	0.2014	0.1158	0.1507
PROMIS 3 Physical							
mean	0.5061	0.5494	0.5981	0.5439	0.4073	0.7684	0.6572
std. dev.	0.1711	0.1502	0.1818	0.1684	0.2017	0.1347	0.1724

PROMIS 4 Mental							
mean	0.6515	0.6543	0.6083	0.6687	0.441	0.8348	0.7211
std. dev.	0.1747	0.1446	0.169	0.1677	0.1797	0.1163	0.1645
PROMIS 6 Daily life							
mean	0.6651	0.6618	0.6564	0.7163	0.4872	0.8382	0.7597
std. dev.	0.1766	0.1684	0.1866	0.1743	0.2121	0.1245	0.1641
PROMIS II 1 Daily living							
mean	0.937	0.9678	0.9805	0.9744	0.8474	0.9996	0.9872
std. dev.	0.1535	0.1068	0.0717	0.0853	0.2521	0.009	0.058
PROMIS II 2 Pain							
mean	0.844	0.8712	0.8787	0.8537	0.7384	0.9364	0.9078
std. dev.	0.1839	0.1536	0.155	0.1712	0.2554	0.1097	0.1379
PROMIS II 3 Fatigue							
mean	0.8671	0.8531	0.8421	0.8618	0.7154	0.9311	0.9228
std. dev.	0.167	0.1545	0.1756	0.1615	0.2524	0.1078	0.1242
PROMIS II 4 Emotional							
mean	0.8262	0.7995	0.7781	0.813	0.6368	0.9119	0.8698
std. dev.	0.1936	0.1913	0.2181	0.2027	0.2848	0.1372	0.1727
MENTAL 1 Interest							
mean	0.8575	0.8738	0.826	0.8737	0.6967	0.9388	0.8936
std. dev.	0.1271	0.1174	0.1218	0.1185	0.2045	0.0966	0.1187
MENTAL 2 Sad							
mean	0.8975	0.8898	0.8458	0.8846	0.7282	0.9511	0.914
std. dev.	0.1097	0.1099	0.1194	0.1194	0.2073	0.0864	0.1082
MENTAL 4 Uncontrol							
mean	0.9061	0.9071	0.8476	0.9031	0.7274	0.9586	0.9372

std. dev.	0.1272	0.117	0.1311	0.1195	0.2167	0.0823	0.1041
COGNITIVE 1 Memory							
mean	0.7516	0.8419	0.624	0.8054	0.4666	0.8994	0.8673
std. dev.	0.1883	0.1597	0.1656	0.1689	0.2047	0.1413	0.1551
COGNITIVE 2 Read							
mean	0.8059	0.8532	0.649	0.8376	0.4899	0.9037	0.8598
std. dev.	0.1588	0.1332	0.1651	0.1551	0.2029	0.1185	0.1446
COGNITIVE 3 Interrupted							
mean	0.7831	0.8516	0.6318	0.8257	0.479	0.8861	0.8526
std. dev.	0.1811	0.1266	0.1569	0.1518	0.1986	0.1262	0.1531
COGNITIVE 4 Two things							
mean	0.8543	0.892	0.701	0.8643	0.5329	0.9251	0.8901
std. dev.	0.1471	0.1288	0.163	0.1456	0.2216	0.1137	0.1457
COGNITIVE 5 Recalling							
mean	0.7727	0.8573	0.6286	0.8253	0.4637	0.8912	0.8318
std. dev.	0.2116	0.1463	0.1758	0.169	0.2097	0.1405	0.1683
COGNITIVE 6 Thinking							
mean	0.9003	0.9336	0.7308	0.9182	0.5403	0.969	0.9393
std. dev.	0.1319	0.0967	0.1381	0.1078	0.2018	0.0734	0.1051
COGNITIVE 7 Slower							
mean	0.846	0.9154	0.708	0.9008	0.5077	0.9583	0.9242
std. dev.	0.1519	0.1062	0.1427	0.1224	0.2023	0.0843	0.122
COGNITIVE 8 Pay							

attention							
mean	0.8532	0.9026	0.6709	0.8809	0.4766	0.9411	0.897
std. dev.	0.1616	0.1121	0.1419	0.1231	0.1998	0.0969	0.1236
COGNITIVE 9 Concentrating							
mean	0.8009	0.8541	0.6297	0.8267	0.4292	0.9045	0.8582
std. dev.	0.1693	0.1197	0.1441	0.1355	0.177	0.1156	0.1376
COGNITIVE 10 Remembering							
mean	0.8317	0.8867	0.6909	0.8582	0.5205	0.9231	0.879
std. dev.	0.1499	0.122	0.1582	0.141	0.2164	0.1083	0.134
COGNITIVE 11 Decisions							
mean	0.8689	0.8902	0.7224	0.8727	0.529	0.9411	0.8894
std. dev.	0.1513	0.1315	0.1703	0.141	0.2128	0.1	0.1417
COGNITIVE 12 Planning							
mean	0.9186	0.9443	0.7882	0.9294	0.6004	0.9766	0.9406
std. dev.	0.1202	0.0954	0.1393	0.1008	0.2345	0.0643	0.1085
SMOKING N							
mean	0.1621	0.0989	0.0878	0.1465	0.1425	0.0651	0.1067
std. dev.	0.2094	0.1389	0.1213	0.179	0.1854	0.1022	0.1471
SLEEP N							
mean	0.3349	0.3307	0.3779	0.36	0.5165	0.2511	0.2952
std. dev.	0.1642	0.1578	0.1701	0.1719	0.228	0.1156	0.142
Comorbility							
False	86.753	7.180.548	471.803	52.667	805.302	9.364.241	353.246
True	1.653.472	42.479	694.789	3.938.678	1.899.955	594.028	456.599
[total]	1.740.225	7.223.027	5.412.819	3.991.344	2.705.257	9.958.269	3.989.059

4.2. Resultados aplicando algoritmo SOM con 6 clústers

Attribute	0	1	2	3	4	5
Instancias	730	829	1076	204	171	478
Age						
value	0.479	0.4601	0.4631	0.4953	0.6366	0.6613
min	0.1667	0	0.1667	0	0.1667	0.1667
max	0.8333	0.8333	0.8333	0.8333	1	1
mean	0.4653	0.4692	0.4545	0.4886	0.6189	0.6506
std. dev.	0.1567	0.1598	0.1633	0.1608	0.2049	0.1776
BMI N						
value	0.3681	0.4065	0.363	0.3984	0.3892	0.3977
min	0.125	0	0.125	0.125	0.125	0.125
max	1	1	0.875	0.875	1	1
mean	0.3683	0.3976	0.3547	0.3958	0.3838	0.3912
std. dev.	0.1194	0.144	0.113	0.1426	0.1513	0.1269
Work						
value	0.9575	0.974	0.9607	0.9558	0.1935	0.2151
min	0.6667	0.6667	0.6667	0	0	0
max	1	1	1	1	0.6667	0.6667
mean	0.958	0.9662	0.9681	0.9526	0.1832	0.1904
std. dev.	0.1107	0.1006	0.0981	0.1382	0.1702	0.1693
Kind job						
value	0.7484	0.7578	0.7843	0.809	0.0027	0.007
min	0.125	0.125	0.25	0	0	0
max	1	1	1	1	0.25	0.375
mean	0.8389	0.8382	0.877	0.8217	0.0015	0.0018
std. dev.	0.1496	0.1443	0.1282	0.1664	0.0191	0.0235
PROMIS 1 health						
value	0.652	0.5736	0.8087	0.4667	0.5191	0.6524
min	0.2	0.2	0.4	0	0.2	0.2
max	1	1	1	1	1	1
mean	0.6455	0.5698	0.8058	0.4686	0.5064	0.6699
std. dev.	0.1637	0.1359	0.1192	0.2034	0.1919	0.1753
PROMIS 2 QoL						
value	0.6634	0.5962	0.8234	0.5202	0.6175	0.6928
min	0.2	0.2	0.4	0	0.2	0.2

Attribute	0	1	2	3	4	5
max	1	1	1	1	1	1
mean	0.6622	0.6012	0.8184	0.5206	0.6035	0.7113
std. dev.	0.1519	0.1272	0.1101	0.1939	0.1909	0.1559
PROMIS 3 Physical						
value	0.6065	0.5121	0.7634	0.4371	0.4941	0.6032
min	0.2	0.2	0.2	0	0.2	0.2
max	1	0.8	1	1	1	1
mean	0.5973	0.5107	0.7684	0.4275	0.483	0.6259
std. dev.	0.1765	0.1419	0.1318	0.2045	0.1997	0.1831
PROMIS 4 Mental						
value	0.6278	0.6345	0.8428	0.4287	0.5318	0.7244
min	0.2	0.2	0.4	0	0.2	0.4
max	1	1	1	0.8	1	1
mean	0.623	0.6417	0.8362	0.4363	0.5228	0.7243
std. dev.	0.1586	0.1479	0.1182	0.1709	0.1805	0.1644
PROMIS 6 Daily life						
value	0.6617	0.6545	0.8476	0.4958	0.5922	0.7318
min	0.2	0.2	0.4	0	0.2	0.2
max	1	1	1	1	1	1
mean	0.6605	0.6581	0.8435	0.502	0.5895	0.7456
std. dev.	0.1815	0.1698	0.1228	0.2087	0.1979	0.1734
PROMIS II 1 Daily living						
value	0.9888	0.9784	0.9935	0.8656	0.9403	0.9765
min	0.2	0.2	0.6	0	0.2	0.2
max	1	1	1	1	1	1
mean	0.9781	0.9679	0.9948	0.8637	0.9287	0.9728
std. dev.	0.0893	0.1072	0.0392	0.2416	0.1647	0.102
PROMIS II 2 Pain						
value	0.9148	0.8839	0.9491	0.7317	0.8516	0.909
min	0.4	0.4	0.4	0	0.2	0.4
max	1	1	1	1	1	1
mean	0.8858	0.8608	0.9283	0.7206	0.8316	0.8925
std. dev.	0.1501	0.1653	0.1185	0.2553	0.2062	0.1501
PROMIS II 3 Fatigue						
value	0.8831	0.8932	0.9493	0.6836	0.862	0.9249
min	0.4	0.4	0.6	0	0.4	0.4

Attribute	0	1	2	3	4	5
max	1	1	1	1	1	1
mean	0.8556	0.8533	0.9243	0.6696	0.8421	0.9134
std. dev.	0.1674	0.16	0.1149	0.2457	0.1856	0.1369
PROMIS II 4 Emotional						
value	0.8316	0.82	0.9163	0.5819	0.7991	0.8889
min	0.2	0.2	0.4	0	0.4	0.4
max	1	1	1	1	1	1
mean	0.8019	0.7957	0.9045	0.5716	0.786	0.8657
std. dev.	0.2117	0.198	0.1431	0.2667	0.2218	0.1751
MENTAL 1 Interest						
value	0.8352	0.8705	0.9465	0.7031	0.7833	0.8857
min	0.4	0.4	0.4	0	0.4	0.4
max	1	1	1	1	1	1
mean	0.8373	0.8731	0.9361	0.698	0.7836	0.89
std. dev.	0.1232	0.1184	0.0999	0.2049	0.16	0.123
MENTAL 2 Sad						
value	0.8461	0.8814	0.9485	0.7426	0.819	0.9064
min	0.4	0.4	0.6	0	0.4	0.6
max	1	1	1	1	1	1
mean	0.8575	0.8842	0.9483	0.7294	0.8199	0.9159
std. dev.	0.1189	0.118	0.088	0.2159	0.1498	0.107
MENTAL 4 Uncontrol						
value	0.8588	0.9211	0.9515	0.7128	0.8368	0.9387
min	0.4	0.4	0.4	0	0.4	0.4
max	1	1	1	1	1	1
mean	0.8584	0.9141	0.9513	0.7137	0.8421	0.936
std. dev.	0.1289	0.1182	0.0905	0.2229	0.1506	0.1099
COGNITIVE 1 Memory						
value	0.6387	0.882	0.8927	0.4295	0.5609	0.8725
min	0.2	0.4	0.2	0	0.2	0.4
max	1	1	1	1	1	1
mean	0.6479	0.8642	0.8946	0.4422	0.5626	0.8724
std. dev.	0.1645	0.1489	0.1423	0.1872	0.1736	0.1506
COGNITIVE 2 Read						
value	0.6753	0.8783	0.9029	0.4592	0.6032	0.8815
min	0.2	0.4	0.2	0	0.2	0.4

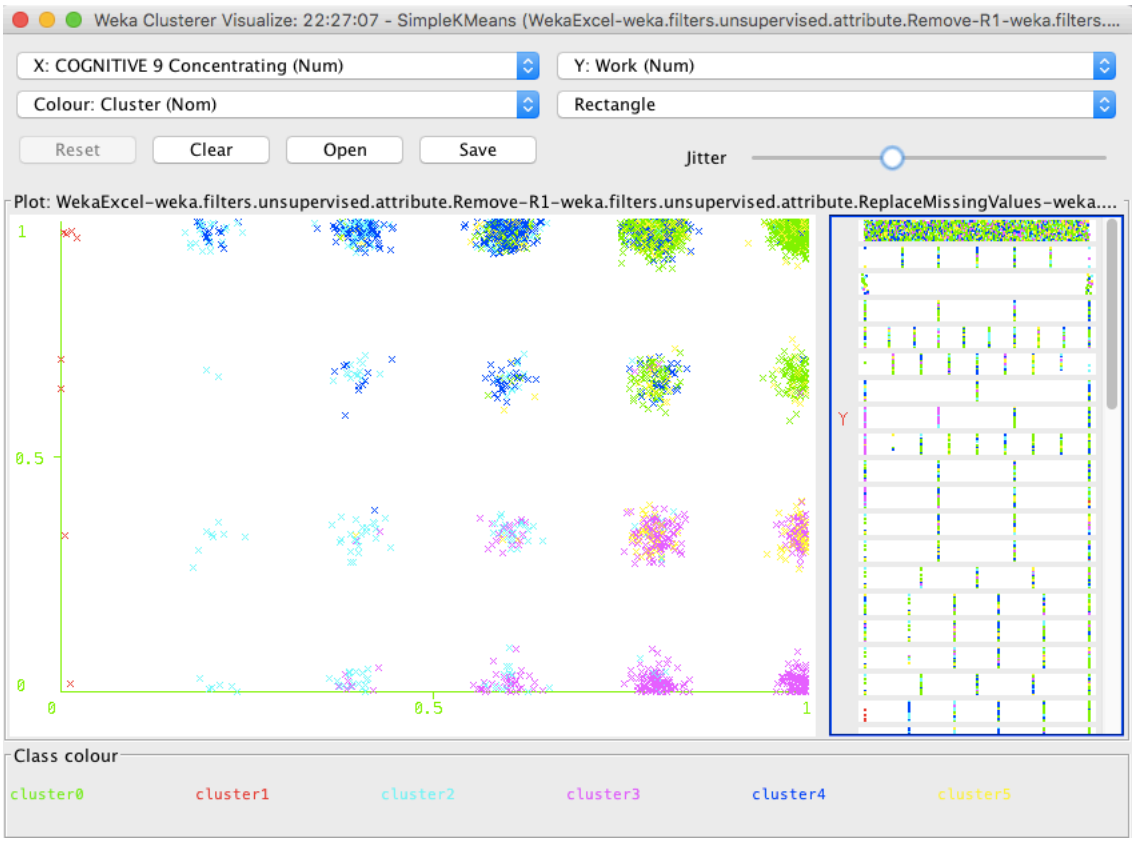
Attribute	0	1	2	3	4	5
max	1	1	1	0.8	1	1
mean	0.6795	0.8767	0.9009	0.4578	0.6058	0.8736
std. dev.	0.158	0.1266	0.1197	0.1883	0.1794	0.1343
COGNITIVE 3 Interrupted						
value	0.6719	0.8585	0.8814	0.4401	0.5636	0.8594
min	0.2	0.4	0.4	0	0.2	0.4
max	1	1	1	1	1	1
mean	0.6674	0.8704	0.8835	0.4529	0.5684	0.869
std. dev.	0.1524	0.1246	0.1264	0.1903	0.1699	0.1395
COGNITIVE 4 Two things						
value	0.7403	0.9054	0.9215	0.4938	0.6417	0.921
min	0.2	0.2	0.4	0	0.2	0.2
max	1	1	1	1	1	1
mean	0.7318	0.9078	0.9229	0.5088	0.6316	0.91
std. dev.	0.1515	0.1259	0.1159	0.2085	0.1945	0.1263
COGNITIVE 5 Recalling						
value	0.6589	0.8825	0.884	0.4255	0.5263	0.8625
min	0.2	0.4	0.2	0	0.2	0.4
max	1	1	1	1	1	1
mean	0.6616	0.8811	0.8868	0.4382	0.5345	0.8561
std. dev.	0.1733	0.139	0.1405	0.1849	0.1926	0.1539
COGNITIVE 6 Thinking						
value	0.7871	0.9468	0.9618	0.5117	0.674	0.9582
min	0.4	0.6	0.4	0	0.2	0.6
max	1	1	1	1	1	1
mean	0.7762	0.952	0.9623	0.5137	0.6643	0.9569
std. dev.	0.1314	0.0888	0.082	0.1787	0.1936	0.0853
COGNITIVE 7 Slower						
value	0.7509	0.9353	0.9465	0.4819	0.6331	0.9324
min	0.2	0.6	0.6	0	0.2	0.6
max	1	1	1	0.8	1	1
mean	0.7482	0.9353	0.9533	0.4853	0.6246	0.9347
std. dev.	0.1366	0.0996	0.0897	0.1914	0.1856	0.104
COGNITIVE 8 Pay attention						
value	0.7112	0.9136	0.9353	0.4562	0.5968	0.9179
min	0.2	0.4	0.6	0	0.2	0.6

Attribute	0	1	2	3	4	5
max	1	1	1	1	1	1
mean	0.7156	0.9197	0.9375	0.4569	0.5918	0.9201
std. dev.	0.1386	0.108	0.0989	0.1849	0.182	0.1055
COGNITIVE 9 Concentrating						
value	0.6603	0.8659	0.8993	0.4127	0.5458	0.8762
min	0.2	0.4	0.4	0	0.2	0.4
max	1	1	1	0.8	0.8	1
mean	0.6685	0.8685	0.9006	0.4176	0.5439	0.8795
std. dev.	0.1418	0.1193	0.1162	0.1699	0.167	0.1217
COGNITIVE 10 Remembering						
value	0.7387	0.9078	0.9253	0.4904	0.6343	0.898
min	0.2	0.4	0.4	0	0.2	0.4
max	1	1	1	1	1	1
mean	0.7227	0.9006	0.9212	0.4912	0.6316	0.8946
std. dev.	0.1509	0.1166	0.1103	0.2018	0.1858	0.1212
COGNITIVE 11 Decisions						
value	0.7529	0.8998	0.934	0.5248	0.6409	0.9095
min	0.2	0.2	0.4	0	0.2	0.4
max	1	1	1	1	1	1
mean	0.751	0.9025	0.9372	0.5304	0.6351	0.9113
std. dev.	0.1671	0.1295	0.1038	0.2121	0.1957	0.1222
COGNITIVE 12 Planning						
value	0.8102	0.9523	0.9736	0.575	0.7323	0.9592
min	0.2	0.6	0.6	0	0.2	0.4
max	1	1	1	1	1	1
mean	0.8178	0.9561	0.9734	0.5824	0.7193	0.9594
std. dev.	0.1356	0.0885	0.0695	0.2203	0.1923	0.0894
SMOKING N						
value	0.1124	0.0957	0.0719	0.1132	0.1449	0.1162
min	0	0	0	0	0	0
max	0.875	1	0.75	1	1	1
mean	0.1026	0.1074	0.0737	0.1275	0.1382	0.1198
std. dev.	0.1416	0.1493	0.116	0.1728	0.1819	0.165
SLEEP N						
value	0.3687	0.3257	0.271	0.5118	0.4484	0.3048
min	0	0	0	0	0.1667	0.1667

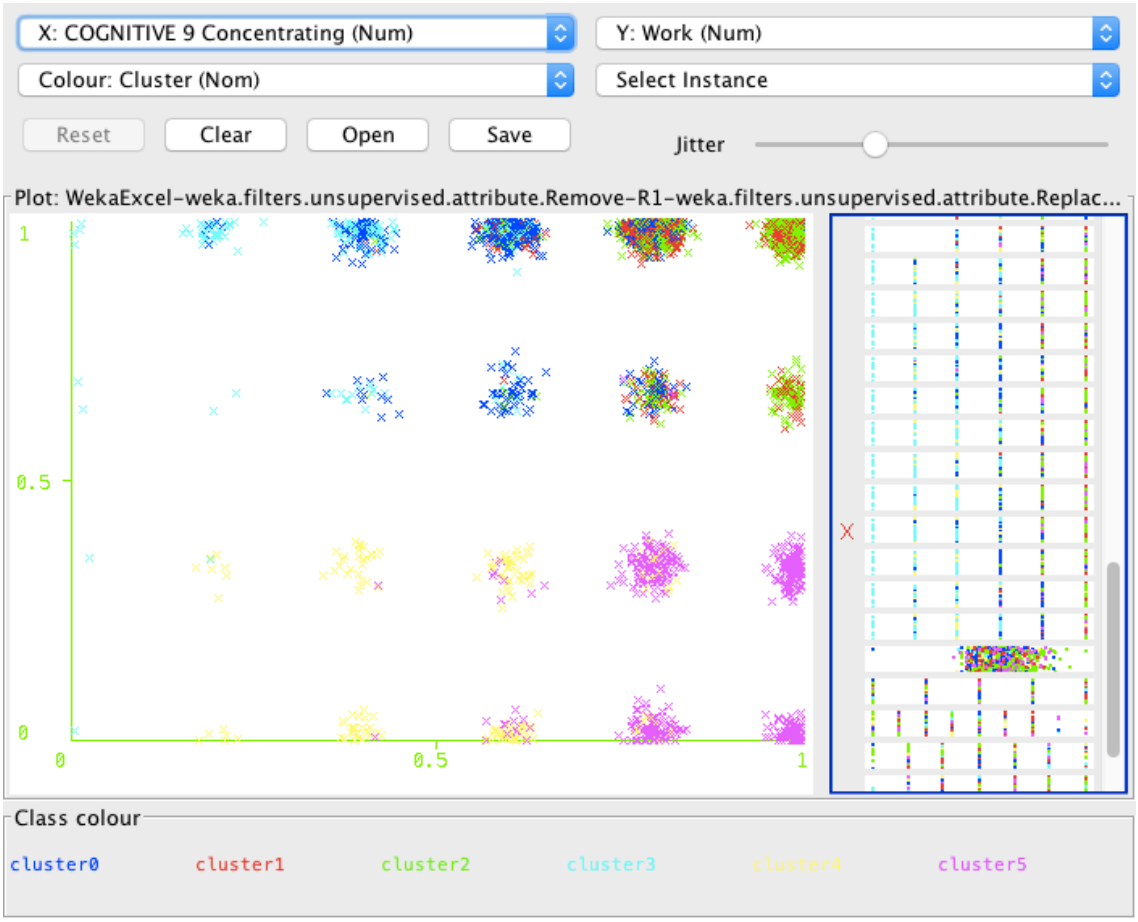
Attribute	0	1	2	3	4	5
max	1	1	1	1	1	1
mean	0.3724	0.3323	0.2619	0.5	0.4357	0.2978
std. dev.	0.1673	0.1621	0.1271	0.2176	0.2289	0.1456
Comorbility						
value	0.2494	0.3386	0.149	0.5627	0.6125	0.3787
min	0	0	0	0	0	0
max	1	1	1	1	1	1
mean	0.263	0.269	0.1208	0.5441	0.6199	0.3326
std. dev.	0.4406	0.4437	0.3261	0.4993	0.4868	0.4717
Time taken to build model (full training data) : 157.03 seconds						
Clustered Instances						
0 730 (21%)						
1 829 (24%)						
2 1076 (31%)						
3 204 (6%)						
4 171 (5%)						
5 478 (14%)						

5. Visualizaciones del agrupamiento

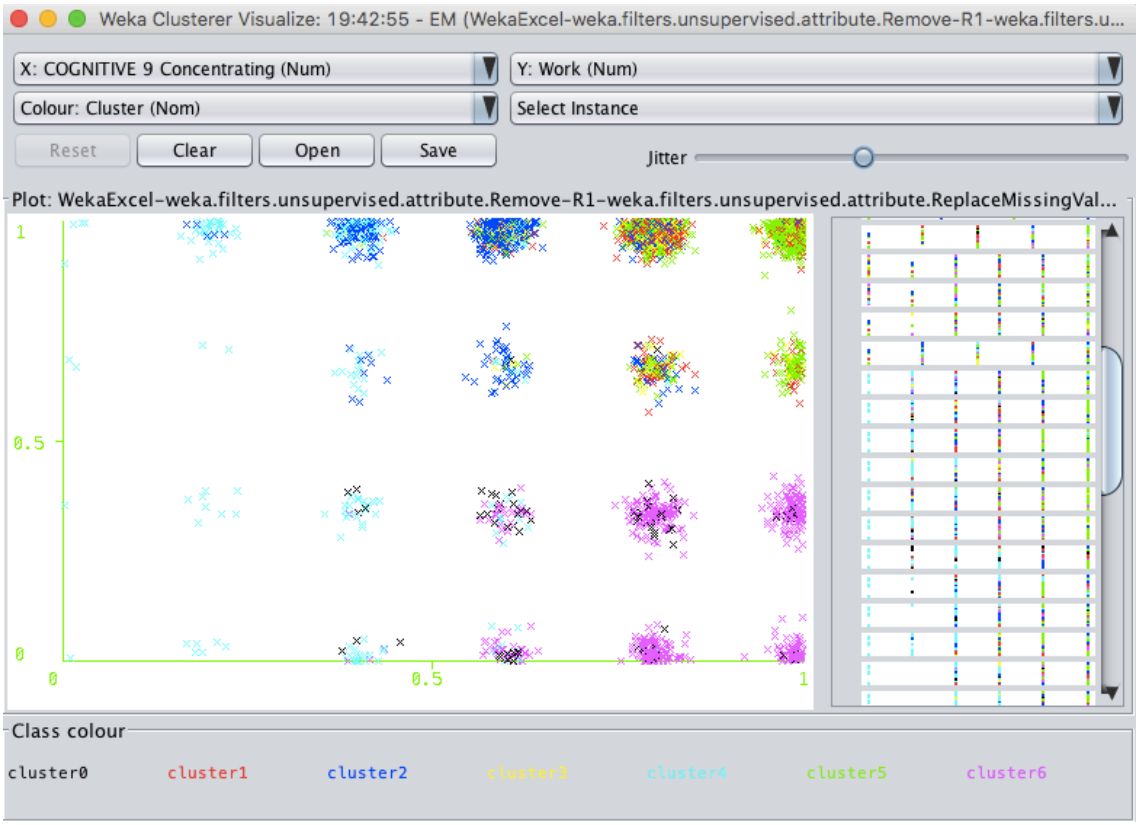
Visualización con SimpleKMeans con 7 clústeres



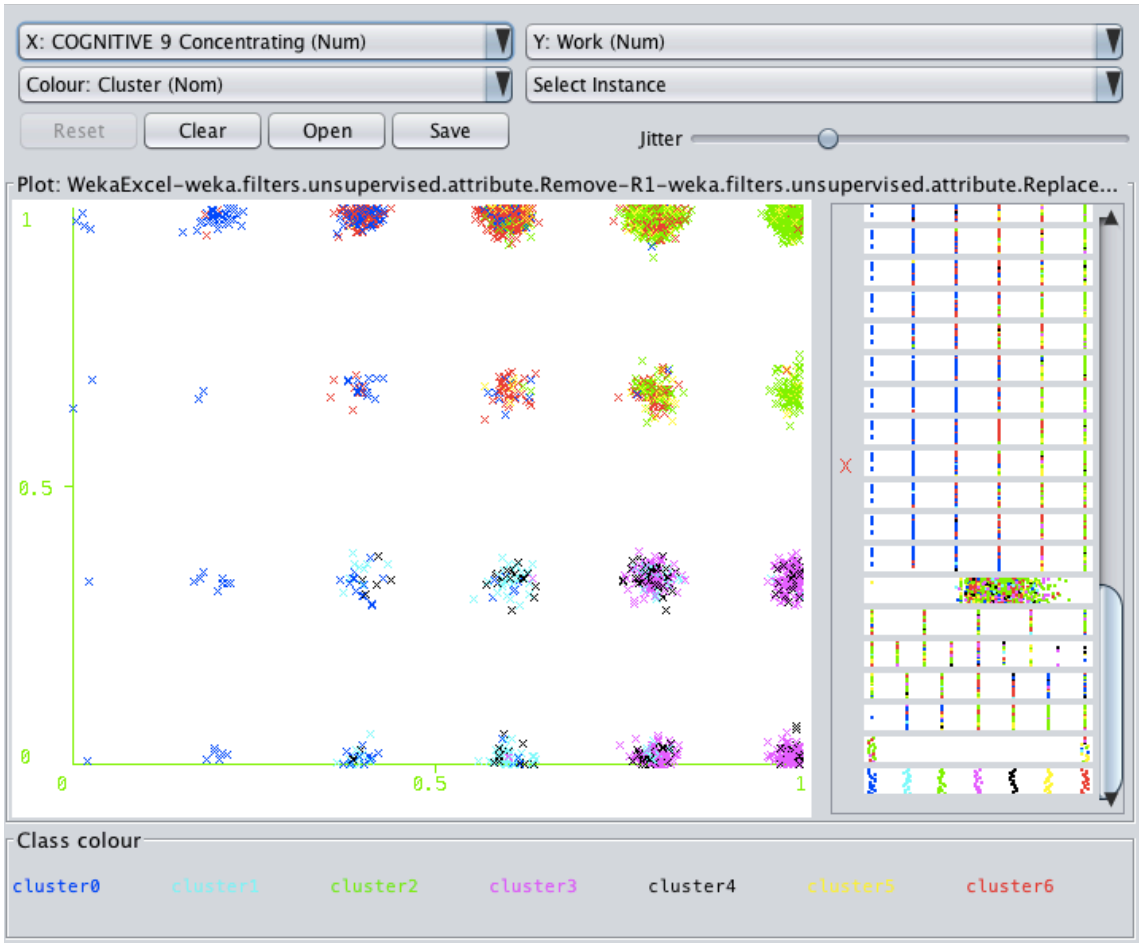
Visualización con SOM con 6 clústeres



Visualización con EM con 7 clústeres



Visualización MakeDensityBasedClusterer 7 clústeres



Visualización con MTree de 5 clústeres