Exploración de perfiles de datos generales, de personalidad y de niveles de estrés, depresión y ansiedad

Juan Pablo Maldonado Castro
Tecnologías para la Información en Ciencias
UNAM - ENES Morelia
maldonadocastrojp@gmail.com

1. INTRODUCCIÓN

Algunos de los fenómenos más estudiados en áreas como la psicología son aquellos relacionados con las enfermedades mentales. Entre ellas se encuentran el estrés, la depresión y ansiedad. Con el objetivo de poder identificar sujetos que padezcan alguna de éstas y determinar qué tan grave es su estado, se hace del uso de herramientas como lo son los cuestionarios y las encuestas. A partir de los resultados de éstas es posible obtener una idea general acerca del estado de salud mental de una persona.

También ocurre algo similar en el estudio de la personalidad de una persona. Dependiendo del tipo de clasificación de éstas que se esté haciendo, se hace un determinado cuestionario y a partir de los resultados se extraen las características de la personalidad de alguien.

Cuando relacionamos a estos dos tipos de análisis o estudio, no sería extraño hacer preguntas como, ¿se puede determinar en cierta medida el nivel de depresión, estrés o ansiedad a partir de los resultados de un cuestionario de personalidad o viceversa? O, ¿qué tanta relación hay con la personalidad de una persona y la severidad de las enfermedades mentales que pueda tener? O incluso, ¿puede que también otros factores fuera de los perfiles de personalidad y de estados de salud mental, como lo son la edad, género, religión, raza, etc., tengan influencia sobre estos resultados?

El objetivo general de este proyecto será hacer una exploración de un conjunto de datos que cuenta con cuestionarios del perfil general, personalidad, niveles de estrés, depresión y ansiedad de varias personas. A partir del uso de técnicas y herramientas de minería de datos como patrones frecuentes, reglas de asociación y clustering trataremos de buscar respuestas a las preguntas planteadas hace un momento.

2. DESCRIPCIÓN DE LOS DATOS

El conjunto de datos a utilizar, *Predicting Depression, Anxiety and Stress*[1], consta de los resultados recolectados en una encuesta en línea entre 2017 y 2019. Contiene los datos de casi 35,000 encuestados y estos se podrían dividir en n categorías:

- Respuestas para determinar la Escala para Depresión Estrés y Ansiedad (Depression Anxiety Stress Scales - DASS)[2]: El objetivo principal de éstas es la identificación de personas que padezcan problemas de salud mental; en particular depresión, estrés y ansiedad. Son un total de 42 y el nombre de sus columnas tienen un formato del tipo QnA, donde n es el número de pregunta.

  La forma en que funcionan es la siguiente: entre las 42 preguntas, hay tres categorías (depresión, estrés y ansiedad); donde cada una de las preguntas puede obtener un puntaje entre 0 y 3 (a mayor puntaje, mayor peso se le da a que la enfermedad mental esté presente). El puntaje final corresponde a la suma de los puntajes obtenidos en cada pregunta de la categoría asociada. Este puntaje final, se relaciona con un rango que al final ayuda determinar uno de cinco niveles de la enfermedad:
  - Normal (0)
  - Ligero (1)
  - Moderado (2)
  - Severo (3)
  - Extremadamente severo (4)
- Datos técnicos de las preguntas y respuestas: son datos que toman rastro del tiempo en milisegundos que tardó el usuario en responder y el orden en que fue contestada la pregunta. Solo va enfocado a preguntas de la sección mencionada anteriormente.
- Respuestas hacia un test de personalidad. Éstas son diez preguntas basadas en El Inventario de Personalidad de Diez Elementos (*The Ten Item Personality Inventory*)[3]. Esto con el fin de determinar una personalidad general de quien está tomando la encuesta. La forma en que funciona es que se le da una serie de diez preguntas que le piden a la persona indicar en una escala qué tan identificado se siente con cierta característica. Hay un total de 7 niveles que se definen por:
  - Totalmente en desacuerdo (1)
  - Moderadamente en desacuerdo (2)
  - Ligeramente en desacuerdo (3)
  - Ni acuerdo ni en desacuerdo (4)
  - Un poco de acuerdo (5)
  - Moderadamente de acuerdo (6)
  - Totalmente de acuerdo (7)

  Al final de todo esto, se devuelve un puntaje basado en los mismos niveles, para cada característica de personalidad del individuo. Las características son:
  - Extraversión
  - Simpatía

[1] Conjunto de datos en Kaggle: https://www.kaggle.com/yamqwe/depression-anxiety-stress-scales

[2] Los valores de esta escala se pueden calcular con base en las reglas planteadas en la siguiente página https://www.psytoolkit.org/survey-library/depression-anxiety-stress-dass.html

[3] una propuesta de cuestionario para clasificar características de personalidades. Explicada a detalle en el siguiente artículo http://gosling.psy.utexas.edu/wp-content/uploads/2014/09/JRP-03-tipi.pdf

  - Escrupulosidad
  - Estabilidad Emocional
  - Franqueza
- Respuestas a una sección de identificación de palabras. Estos datos solo consisten de una respuesta de verdadero o falso donde el encuestado indicaba qué palabras conocía o no; algunas de éstas eran inventadas.
- Datos generales de la persona entrevistada: esto contiene información del perfil general de quien respondió la encuesta. Esto incluye datos como la edad, el género, tipo de educación, raza, etc.

Para ver a más detalle cuáles fueron las preguntas, el tipo de respuestas que puede haber y sus significados, se puede consultar el *codebook* que viene incluido con el conjunto de datos.

3. OBJETIVOS Y JUSTIFICACIÓN

A lo largo del desarrollo de este proyecto se plantean varios objetivos. El primero de ellos es el análisis de los perfiles generales de los encuestados a partir de los datos proporcionados. La razón de esto es que es importante entrar en contexto de quiénes están tomando la encuesta. ¿Qué tipo de perfiles hay? ¿Predomina la gente de cierta edad o sí llega a variar en distintos rangos de manera significativa? ¿Las categorías que hay como raza, género y educación varían? Una vez que hayamos esto, podremos responder éstas y más preguntas de similares. Además, esto nos puede servir en algún futuro para hacer comparaciones entre características generales de una persona y otros tipos de perfiles que discutiremos más adelante.

Lo siguiente a realizar sería definir de manera explícita los niveles de estrés, ansiedad y depresión de los encuestados con base en las respuestas de la encuesta. Un inconveniente que tiene el conjunto de datos, es que no muestra directamente estos resultados que se obtienen una vez se finaliza la encuesta. Esto se hace simplemente siguiendo las reglas de *DASS* sobre las respuestas proporcionadas. Con este tipo de información de manera explícita se puede resumir en general el propósito de 42 de las preguntas de la encuesta en solo 6 columnas de datos (tres de puntaje específico y tres de la categoría/nivel).

Una vez definidos estos niveles y puntajes definidos para la ansiedad, estrés y depresión se pueden empezar a hacer distintos análisis. Uno de ellos sería observar si hay patrones frecuentes entre estos tres estados, por ejemplo nos podemos preguntar ¿si se tiene estrés, entonces también ansiedad o viceversa? Podríamos ver si estos tres problemas están correlacionados entre sí o si se pueden tratar como fenómenos ajenos.

Otro aspecto que se podría manejar a partir de esta información son los distintos grupos de personas que se pueden llegar a formar tomando en consideración sus niveles de estrés, ansiedad y depresión. A partir de esto podemos plantearnos y responder algunas preguntas como las siguientes: ¿cuántos grupos puede haber? ¿Los resultados se dividirán simplemente entre gente que no tiene ninguno de estos problemas y la que los padece mucho, o habrá grupos de gente con distintos niveles de cada una de las enfermedades? ¿Acaso los grupos generados a partir del nivel de cada una de las enfermedades mentales tienen cierto parecido con los generados a partir de los perfiles generales?

Todo lo anterior planteado para los perfiles definidos a partir de la ansiedad, depresión y estrés se puede repetir de manera similar para la formación de los perfiles de personalidad de los entrevistados. Estas características de personalidad se definen a partir de lo establecido por el *Ten Item Personality Test*. Con los grupos encontrados, también podríamos realizar alguna comparación de ellos con los distintos niveles que hay para cada enfermedad mental. ¿Hay alguna personalidad que tienda a cierto tipo de comportamiento relacionado a alguna de estas enfermedades?

Los últimos objetivos que se tienen para este proyecto es el análisis directo de las preguntas relacionadas a la *DASS*. En lugar de depender directamente de los puntajes que calculemos, también puede que encontremos algo de información a partir de una inspección de las preguntas de manera individual. Podríamos buscar patrones frecuentes entre ellas mismas, reglas de asociación, correlaciones que puede haber entre ellas o alguna otra propiedad como perfiles acorde a la personalidad o el nivel de enfermedad mental, etc.

El propósito general de este proyecto es simplemente una exploración de los datos. Ver sus características, cuáles patrones puede haber entre ellos, cómo se pueden llegar a correlacionar, los distintos grupos en que se pueden clasificar las personalidades, los perfiles generales y los niveles de enfermedad mental. La justificación de este proyecto no va más allá de eso, no hay ningún motivo oculto para descubrir algo completamente nuevo o inventar una nueva forma de plantear cuestionarios de este estilo. Simplemente se busca un mayor entendimiento de estos datos y ver cómo se relacionan entre sí.

4. HERRAMIENTAS Y TÉCNICAS A UTILIZAR

Para el desarrollo de este proyecto se estará utilizando la técnica de clustering por medio de *K-means*, en menor medida también se usarán dendrogramas. Como técnicas para apoyar esto, se estará usando el método del codo, el índice de Silhouette y el valor de Información Normalizada Mutua entre agrupamientos. Otra técnica que se va a manejar serán patrones frecuentes y reglas de asociación que haya entre los datos que se estén utilizando.

Lo que se va a utilizar para obtener las gráficas, tablas y resultados en general del proyecto será Python con las librerías de *sklearn, mlxtend, yellowbrick, pandas, matplotlib, scipy* y *numpy*. Adicionalmente, hay un tablero realizado con la librería de *Dash* usando el mismo lenguaje de programación.

5. EXPERIMENTOS Y ANÁLISIS

5.1. Perfiles generales

5.1.1. Primeras observaciones.

Como habíamos comentado en los objetivos, lo primero que haremos será hacer un análisis sobre los perfiles generales de los entrevistados que están en el conjunto de datos. Antes de empezar a manipular los datos generales, comencemos por visualizar algunos de estos. Uno de los datos que más pueden destacar en un perfil

general es la edad, así que comencemos por graficar un histograma de eso. Dentro de los datos contamos con edades que van de los 13 hasta los 117 años, graficando esto obtenemos lo que se muestra en la Figura 1.

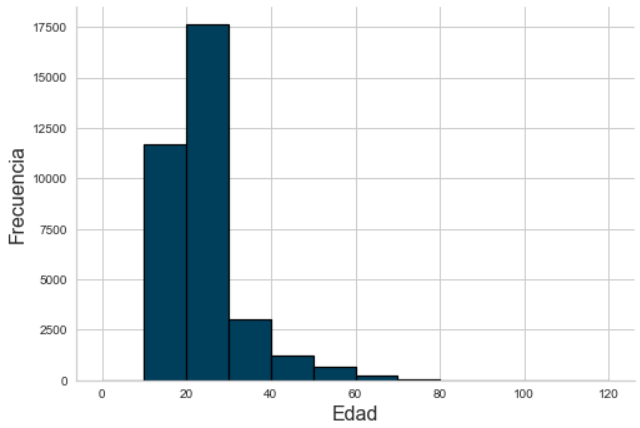

Figura 1: Histograma de las edades de los encuestados

Ya que tenemos la edad, veamos solo algunos otros de los datos generales. Algunos que podríamos acomodar en gráficas de barras son: género, nivel de educación y tipo de zona en donde vivieron su infancia. Esto se puede apreciar en la Figura 2.

Ya tenemos una noción general de nuestro conjunto de datos. Observando las figuras mostradas y las otras que están disponibles junto con el tablero de este proyecto, podemos darnos una idea general de cómo es nuestro conjunto de datos. En primera instancia, vemos cómo la mayoría de las personas encuestadas son jóvenes entre los 10 y 30 años; sin embargo, también podemos ver que alrededor de otras 5000 son adultos con más de 30 años, siendo todavía una cantidad considerable.

Pasando a otros aspectos como el género, tenemos hombres, mujeres y gente que se identifica con otros. En esta encuesta hay una porción significativamente mayor de mujeres que de hombres y los otros géneros son una minoría. También el nivel general de educación que se presenta son personas que ya terminaron la licenciatura o la preparatoria (lo cual tiene sentido considerando la edad de la mayoría de los participantes); y hay todavía una porción menor, pero no insignificante, de gente que no completó la preparatoria y otros que tienen estudios de posgrado.

Podemos seguir haciendo este tipo de observaciones para todos los atributos generales de cada participante. El siguiente paso para estos datos generales es determinar no solo las cantidades individuales en cada categoría, sino dividir los perfiles generales en distintos grupos (clusters) para poder clasificar a toda esta gente y ver qué tanto varían sus características entre ellos.. A primera vista, parece que la mayoría de los encuestados son muy similares entre sí, y aquellos que difieren de estos en sus atributos son una minoría

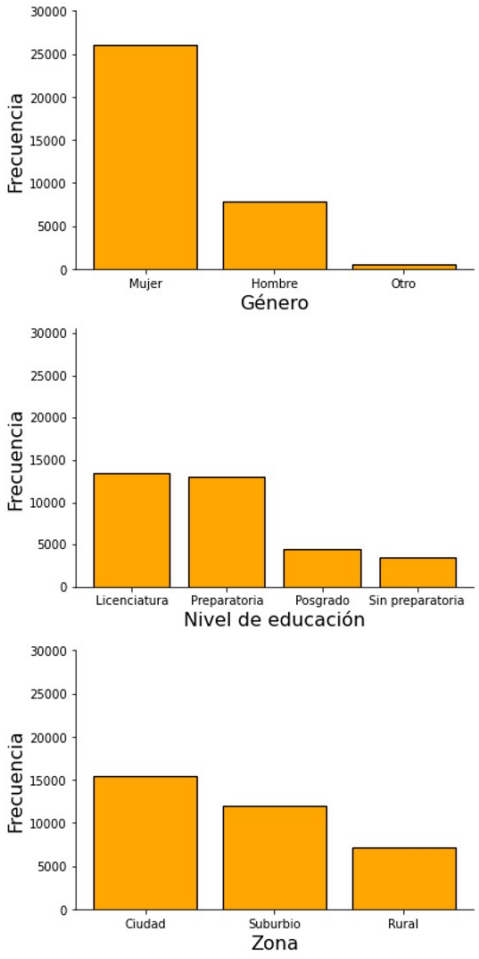

Figura 2: Gráficas de barras del género, máximo nivel de educación y tipo de zona en que crecieron los encuestados

(ya que a veces ni son la mitad de los encuestados).

5.1.2. Formación de clusters.
Para esta sección, revisaremos cuántos clusters convendría formar a partir de los datos proporcionados. Una vez hecho esto, se formarán dichos clusters y se hará una descripción del perfil general que corresponda. De esta manera nos damos una mejor idea de los distintos grupos de gente que toma la encuesta. Como nota, vamos a manejar nuestros datos categóricos aplicándoles One Hot Encoding; los resultados numéricos y explícitos ya se muestran en

el código anexo a este proyecto[4].

Lo primero es determinar el número k de clusters, para ello experimentemos con dos distintas opciones: visualizar un dendrograma a partir de los datos y usar el método del codo. Los resultados son los que vemos en las Figuras 3 y 4.

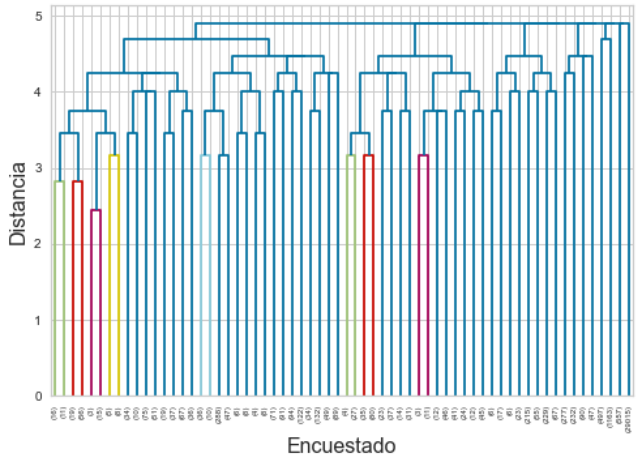

Figura 3: Dendrograma a partir de las características generales de los encuestados

Notemos cómo resultó muy inconveniente utilizar un dendrograma para este problema. Todo inicia con una partición en 7 grupos alrededor de la distancia de 7. A partir de ahí se van formando aún más clusters a pesar de que la distancia no disminuya tanto. Esto hace que sea muy difícil determinar un k conveniente solo observando esa figura.

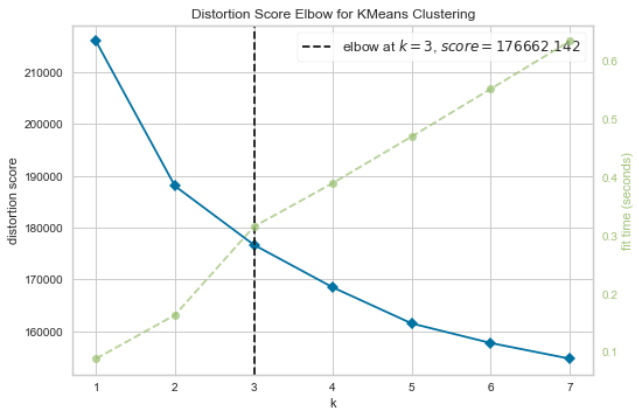

Figura 4: Método del codo para encontrar un k conveniente para clusters de los perfiles generales

[4]Los resultados que se obtengan de este punto, tablas, clusters y gráficas, se pueden recrear en el *Notebook* que viene junto a este proyecto. Está bajo la sección de *Perfile generales*

Como podemos observar el método del codo parece ser una mejor opción porque al menos ilustra en la gráfica el k óptimo acorde a nuestros datos. En este caso se nos indica que $k = 3$, por lo tanto vamos a formar 3 clusters.

Una vez que terminamos de hacer este procedimiento, obtenemos una tabla con varios valores. Las columnas representan un cluster y las filas un valor entre 0 y 1 asignado a cada categoría y su valor. Si el valor numérico es muy cercano a 0 podemos decir que el valor de esa categoría no se asocia con ese grupo; en cambio, si es cercana a 1 decimos que sí está asociada. A partir de esa tabla generada, podemos describir los 3 cluster de la siguiente manera:

- Cluster 0.
  - Llegaron a completar la educación preparatoria.
  - Está conformado principalmente por mujeres.
  - El inglés no es su lengua natal.
  - Su edad está por debajo de los 20 años.
  - La religión que predomina es la musulmana.
  - La raza es asiática.
  - No han votado.
- Cluster 1.
  - Han completado una licenciatura.
  - Son principalmente mujeres.
  - El inglés no es su lengua natal.
  - Tienen entre 20 y 30 años.
  - Su religión es musulmana.
  - Su raza es asiática.
  - Algunas de las personas han votado y otras no.
- Cluster 2.
  - Su nivel de educación es muy variado entre los miembros.
  - El género tampoco está bien definido, no tiende a hombres, mujeres, ni otro.
  - El inglés sí es su lengua natal.
  - Su rango de edad es mucho más diverso (entre 10 y 40 años la mayoría).
  - Un porcentaje significativo de ellos es agnóstico o ateo, pero también hay algunas otras religiones en el grupo.
  - Su raza es de gente blanca.
  - Algunos han votado y otros no.

Notemos que dos de estos clusters son extremadamente parecidos: el 0 y el 1. Ambos son conformados principalmente por mujeres, cuya lengua natal no es el inglés, su religión es la musulmana y su raza es asiática. Lo único que las diferencía es el rango de edad que va a implicar probablemente las diferencias en características de si han sido votantes y nivel de educación. Desde mi punto de vista, estaría mejor tener estos dos grupos como uno porque lo único que los diferencía son dos propiedades que además están muy cercanas entre sí.

Para lograr esto, hagamos una división en $k = 2$ clusters; pero antes, como apoyo visual, podemos observar los clusters formados sobre los componentes principales 1 y 2. Esto se muestra en la Figura 5.

Notemos cómo los datos están muy unidos entre sí sobre esta nube de puntos, esto probablemente se debe a que no hay mucha varianza entre varias de las propiedades. A pesar de ello, todavía es

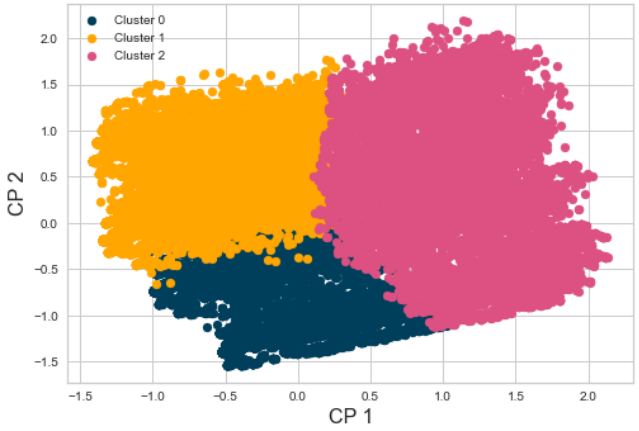

Figura 5: Clusters con $k = 3$ para los perfiles generales visualizados en los componentes principales 1 y 2

posible extraer algunas características que nos permiten distinguir los grupos; que en este caso fueron los tres clusters que formamos.

Algo que es curioso observar es cómo los puntos de los clusters 0 y 1 comparten un rango muy similar en el eje del componente principal 1, mientras que varían en el componente principal 2. Por lo que vimos en los cluster, los que más se parecen entre sí son el 0 y 1 en cuanto a religión, raza, lengua natal y género. Puede que ese hecho influencie la posición en CP1, lo cual también atribuye un poco a cómo el que se separa notablemente en este mismo eje es el cluster 2, ya que difiere en estos parámetros del cluster 0 y 1. Lo que también hace el cluster 2 es abarcar mucho más espacio en el eje de CP2 y recordemos que este cluster abarca un rango grande de edades (compartiendo aquellas del cluster 0 y 1).

Esto es simplemente una intuición detrás de la figura que se generó, sin embargo, no podemos saber con exactitud qué representa cada uno de estos componentes principales. Por lo tanto no llegaremos a conclusiones a partir de ésta, solo está aquí con fines de visualización.

Siguiendo con la formación de clusters con $k = 2$, los centroides que obtenemos, que se encuentran en el código correspondiente a este proyecto, nos indican lo siguiente:

- Cluster 0.
  - Su rango de edades es muy variado, anda entre 10 y 30 años. La mayoría siendo entre 20 y 30.
  - Su lengua natal no es el inglés.
  - El género principal es de mujeres.
  - La religión principal es musulmana.
  - La raza es asiática.
- Cluster 1.
  - Su rango de edades varía mucho entre los 10 y 40 años. La mayoría siendo menores a 20 años.
  - Su lengua nata es el inglés.
  - Hay más variación entre los géneros, pero sigue habiendo más mujeres que hombres,
  - En cuanto a las religiones, hay más agnósticos y ateos, pero también hay algunas otras religiones presentes.
  - La raza es de gente blanca.

Parece ser que los clusters formados si juntaron principalmente a los que estaban en los cluster 0 y 1 del agrupamiento anterior. Ahora sí tenemos una división, en mi opinión, mucho más distinguible entre clusters. Notemos que los atributos que marcan la mayor diferencia son la lengua natal, la religión y raza. Probablemente se trate de que esta encuesta se hizo en dos distintas regiones que tenían estas características.

Visualicemos el nuevo agrupamiento sobre los componentes principales 1 y 2 en la Figura 6.

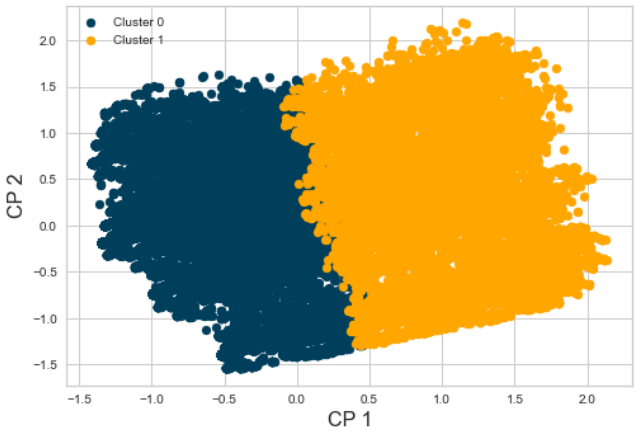

Figura 6: Clusters con $k = 2$ para los perfiles generales visualizados en los componentes principales 1 y 2

Podemos ver claramente que gran parte de los puntos del cluster 0 y 1 se unieron en uno solo.

Para uso futuro, estaremos usando el agrupamiento con $k = 2$.

5.2. Análisis sobre niveles de depresión, ansiedad y estrés

Como siguiente punto para este proyecto, lo que sigue es calcular los puntajes de depresión[5], estrés y ansiedad de cada uno de los encuestados. Para ello, nos basamos en la guía del cuestionario *DASS* que usa el conjunto de datos.

Una vez que obtenemos los niveles de depresión, estrés y ansiedad para cada persona, podemos buscar lo siguiente: patrones frecuentes, reglas de asociación y posibles grupos/clusters.

5.2.1. Patrones frecuentes.

Lo que se realizará sobre este conjunto de datos para analizar los patrones frecuentes será, por medio de la librería de *mlxtend*, identificar a estos si aparecen en al menos el 10 % de las observaciones.

[5]El cálculo de estos puntajes se puede replicar en el *Notebook* del proyecto. Están bajo la sección de *Calculando escala de estrés, ansiedad y depresión (DASS)*

Si fuera un porcentaje muy grande, probablemente sería muy difícil encontrar patrones ya que estamos contando con miles de datos, pero de momento está bien identificar como mínimo patrones con este porcentaje de apariciones. Una vez hecho esto, se obtiene un resultado como el que se muestra en la Tabla 1.

Cuadro 1: Patrones frecuentes en niveles de depresión, ansiedad y estrés

Umbral de frecuencia	Conjunto
0.224110	(DepressionCat_0)
0.176476	(DepressionCat_2)
0.162443	(DepressionCat_3)
0.342311	(DepressionCat_4)
0.249565	(AnxietyCat_0)
0.178853	(AnxietyCat_2)
0.151745	(AnxietyCat_3)
0.350719	(AnxietyCat_4)
0.298040	(StressCat_0)
0.124348	(StressCat_1)
0.219152	(StressCat_2)
0.214253	(StressCat_3)
0.144207	(StressCat_4)
0.145483	(AnxietyCat_0, DepressionCat_0)
0.173055	(StressCat_0, DepressionCat_0)
0.224719	(DepressionCat_4, AnxietyCat_4)
0.119709	(DepressionCat_4, StressCat_3)
0.120869	(DepressionCat_4, StressCat_4)
0.193146	(StressCat_0, AnxietyCat_0)
0.136495	(StressCat_3, AnxietyCat_4)
0.128639	(StressCat_4, AnxietyCat_4)
0.132350	(StressCat_0, AnxietyCat_0, DepressionCat_0)
0.111765	(DepressionCat_4, StressCat_4, AnxietyCat_4)

A partir de lo que nos muestra la tabla podemos hacer estas primeras observaciones.

- Todas las enfermedades están presentes a lo largo del conjunto de datos en casi todos sus niveles, desde el normal hasta el extreamadamente severo.
- Entre los encuestados, los niveles de depresión y ansiedad se cargan principalmente entre los dos extremos: nivel 0 y nivel 4. Los rangos medios de manera individual aparecen de manera menos frecuente. Esto nos indica que es más común que los encuestados establezcan que no se sienten mal con alguna de estas enfermedades o que se sienten de la peor forma posible, a que estén en un punto medio entre estos niveles.
- Los niveles de los extremos también aparecen juntos, por ejemplo tanto ansiedad, como estrés y depresión en el nivel máximo aparecen de forma seguida en una misma persona. De la misma manera, es frecuente el hecho de que las personas no padezcan ninguna de las 3. Esto nos indica que lo más común, es que las personas sufran al máximo estas tres enfermedades a la vez o ninguna. Por ejemplo, no es tan común ver gente que tenga niveles muy altos de algo como estrés, pero bajos de ansiedad.

5.2.2. *Reglas de asociación.*

Ya obtuvimos los patrones frecuentes, algo que podría ayudar a complementar lo comentado sobre éstos, serían las reglas de asociación. Usando la misma librería del punto anterior y estableciendo una confianza mínima del 80 % nos dan los resultados de la Tabla 2.

Cuadro 2: Reglas de asociación en niveles de depresión, ansiedad y estrés

Antecedente	Consecuencia	Confianza
(StressCat_4)	(DepressionCat_4)	0.838158
(StressCat_4) (	AnxietyCat_4)	0.892039
(AnxietyCat_0, DepressionCat_0)	(StressCat_0)	0.909725
(DepressionCat_4, StressCat_4) (	AnxietyCat_4)	0.924682
(StressCat_4, AnxietyCat_4)	(DepressionCat_4)	0.868830

Con estas reglas, es posible decir con altos niveles de confianza lo que ya habíamos mencionado en el punto anterior, pero de manera más explícita: gente que padece de una o dos de las enfermedades en niveles altos, muy probablemente también padecerá del resto. Por el contrario, gente que no padezca al menos dos de las enfermedades (en este caso ansiedad y depresión), no padecerá entonces de la otra (estrés).

5.2.3. *Formación de clusters.*

Por último, tratemos de clasificar a las personas en distintos grupos acorde a sus niveles de estrés, ansiedad y depresión. Así identificamos cuántos y cuáles grupos existen acorde a este criterio. A diferencia del clustering de perfiles generales, no parece necesario tratar a las respuestas de la encuesta como datos categóricos. La razón de esto, es que estamos trabajando con una escala incremental. A mayor valor de la categoría, decimos que es más severa la enfermedad.

Primero obtenemos la k correspondiente usando también un dendrograma y el método del codo como se hizo en la sección de los perfiles generales. Esto se muestra en las Figuras 7 y 8.

Parece ser que hay varias k posibles. La que nos recomienda el método del codo es de $k = 2$. Sin embargo, el dendrograma que se generó, también muestra que puede ser razonable llegar inclusive hasta 4 o 5 grupos. En este caso vamos a probar con dos posibilidades $k = 2$ y $k = 4$. A partir de los resultados determinaremos cuáles conjuntos aparentan ser mejores acorde a sus características.

Usando *K-means* con $k = 2$, obtendremos los resultados de la Tabla 3.

A partir de los centroides, podemos hacer las siguientes conclusiones de cada cluster:

- El cluster 0 es de gente que sí padece de todas las enfermedades en una medida principalmente severa. Los que están en este grupo probablemente oscilan entre el nivel medio al más alto de todas las enfermedades.

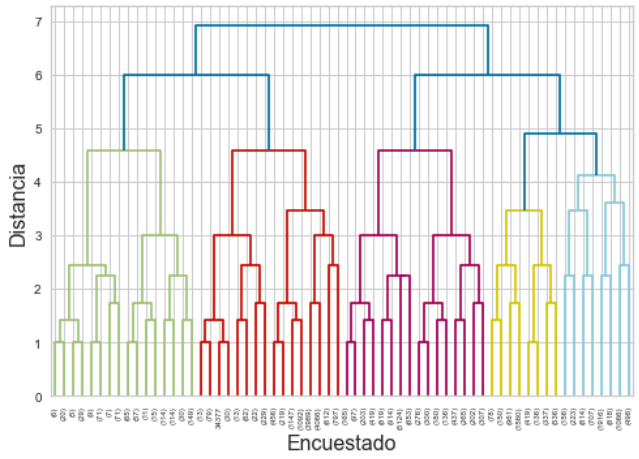

Figura 7: Dendrograma de los niveles de estrés, ansiedad y depresión

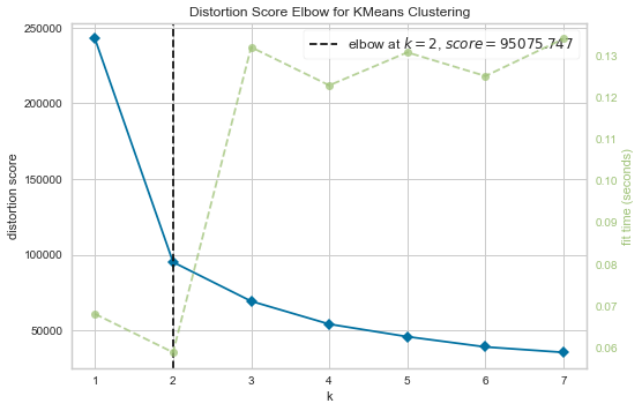

Figura 8: Método del codo para encontrar k en los datos de niveles de depresión, ansiedad y estrés

Cuadro 3: Centroides de los clusters con $k = 2$ para niveles de depresión, ansiedad y estrés

Enfermedad	Cluster 0	Cluster 1
Depresión	3.291261	0.901888
Ansiedad	3.350245	0.771493
Estrés	2.730821	0.434627

- El cluster 1 es de gente que padece en menor medida las enfermedades. Sus valores probablemente se encuentren entre los dos niveles más bajos de severidad de las enfermedades.

Repitiendo lo mismo, pero con $k = 4$, obtendremos los resultados que se muestran en la Tabla 4.

Con estos centroides, ya tenemos una idea muy clara de qué representa cada uno:

Cuadro 4: Centroides de los clusters con $k = 3$ para niveles de depresión, ansiedad y estrés

Enfermedad	Cluster 0	Cluster 1	Cluster 2	Cluster 3
DepressionCat	3.674405	0.387134	1.660455	3.253985
AnxietyCat	3.738092	0.410767	2.867429	1.068863
StressCat	3.184579	0.197541	1.570239	1.324973

- El cluster 0 representa a la gente que tiene niveles severos de depresión, ansiedad y estrés.
- El cluster 1 es de aquellas personas que tienen niveles muy bajos de las tres enferemdades.
- El cluster 2 representa a la gente que tiene en un nivel moderado de depresión y estrés, pero una ansiedad que casi es severa.
- El cluster 3 representa a la gente que tiene un nivel severo de depresión y uno moderado de estrés y ansiedad.

Lo que nos permite concluir esta agrupación, es que no existen solo dos extremos al clasificar gente con distintos niveles de una enfermedad mental. Éstas pueden estar presentes en un punto medio. Lo que es curioso es observar como en aquellos grupos donde no todo se va a un extremo, los niveles varían más entre las enfermedades. Por ejemplo, en el cluster 3, la depresión es severa a diferencia de la ansiedad y el estrés que están dos niveles más abajo.

Otro característica que podemos concluir de manera inicial con estos clusters es que tener al menos una de estas enfermedades en un nivel cercano al severo, va a provocar que las otras dos estén presentes por lo menos de manera moderada. En cambio, no parece que vaya a ser muy común tener tan solo uno de estas en un nivel casi nulo, cuando haya alguna otra enfermedad presente en niveles mayores.

Ya que analizamos los dos clusterings con los k propuestos, por opinión personal, usaremos de ahora en adelante el de $k = 4$. La razón de esto es que parece más razonable tener grupos en los puntos medios de los espectros de depresión, estrés y ansiedad, a tener solo dos grupos que representan un extremo cada uno.

De manera visual, usando los primeros dos componentes principales, podemos graficar en dos dimensiones los cuatro grupos formados. Resultando en lo que se muestra en la Figura 9.

Con esta visualización de los clusters, es posible apreciar cómo cada uno tiende a cierta región del plano formado por los dos primeros componentes principales. Aún así, la división entre ellos no es muy clara debido a que la separación entre clusters es casi inexistente. De todas formas, es posible formar los grupos como ya vimos anteriormente para enmarcar de manera general las distintas clasificaciones posibles cuando se están analizando depresión, estrés y ansiedad de manera simultánea.

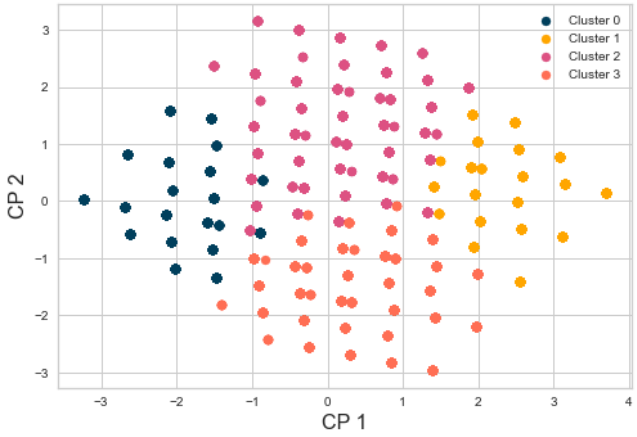

Figura 9: Clasificación de los niveles de depresión, estrés y ansiedad sobre componentes principales 1 y 2

5.3. Análisis sobre las personalidades

La última clasificación por realizar es la de las personalidades de quienes tomaron la encuesta. Para ello, primero calculamos los puntajes de cada característica de personalidad acorde a lo indicado en la referencia del *codebook*[6].

Una vez hecho esto, tendremos varios puntajes numéricos asignados sobre la escala del 1 al 7 respecto a cada característica. En este caso, los puntajes en el rango del 1 a menos de 4 representan lo negativo a lo planteado (en este caso 1 siendo el peor); el puntaje igual a 4 representa un estado neutro de la característica; y finalmente los puntajes mayores a 4 representan de manera incremental la característica planteada de manera positiva.

Como estamos trabajando con valores de tipo flotante, si queremos repetir lo mismo que con los perfiles acorde a niveles de estrés, ansiedad y depresión; debemos discretizar nuestros valores. Lo que vamos a hacer es discretizar acorde al rango del puntaje:

- 1: Si se encuentra en el rango [1,3), representa personalidad negativa.
- 2: Si se encuentra en el rango [3,5), representa personalidad neutra.
- 3: Si se encuentra en el rango [5,7], representa personalidad positiva.

Dados estos valores para todas las personas, ya se puede comenzar a analizar patrones frecuentes y reglas de asociación. En cuanto a los clusters se va a trabajar con los valores numéricos para perder menos variabilidad de los datos.

5.3.1. *Patrones frecuentes.*

Buscando los patrones frecuentes en los datos, si establecemos umbrales de frecuencia de algo como 10 %, obtendremos una cantidad enorme de conjuntos. Para filtrar varios de los resultados, mejor lo subiremos al 30 %. Con ayuda de la librería *mlxtend* de Python se obtienen los resultados mostrados en la Tabla 5.

[6]El cálculo de estos valores se encuentra en la sección del *Notebook* con el título de *Cálculo del tipo de personalidad*

Cuadro 5: Patrones frecuentes de las características de personalidad

Soporte	Conjunto
0.460600	(ExtraversionCat_1)
0.394497	(ExtraversionCat_2)
0.596776	(AgreeablenessCat_2)
0.482605	(ConscientiousnessCat_2)
0.545054	(EmotionalStabilityCat_1)
0.337122	(EmotionalStabilityCat_2)
0.514003	(OpennessCat_2)
0.336716	(OpennessCat_3)
0.302302	(ConscientiousnessCat_2, AgreeablenessCat_2)
0.335237	(EmotionalStabilityCat_1, AgreeablenessCat_2)
0.323611	(AgreeablenessCat_2, OpennessCat_2)

A primera vista, lo que podemos notar es que parece haber muy pocos patrones frecuentes para las características de personalidad positivas. Casi todos los conjuntos que aparecen son de características neutras o negativas. Una de las características de personalidad más frecuentes son la simpatía neutra, incluso podemos ver que se relaciona frecuentemente con estabilidad emocional baja, franqueza neutra y escrupulosidad neutra.

5.3.2. *Reglas de asociación.*

Ya que tenemos una noción de lo que sucede con los patrones frecuentes, tratemos de observar las reglas de asociación que pueden llegar a surgir a partir de lo que obtuvimos. Los resultados se muestran en la Tabla 6.

Cuadro 6: Reglas de asociación para las características de personalidad

Antecedente	Consecuencia	Confianza
(ConscientiousnessCat_2)	(AgreeablenessCat_2)	0.626397
(EmotionalStabilityCat_1)	(AgreeablenessCat_2)	0.615053
(OpennessCat_2)	(AgreeablenessCat_2)	0.629590

Estas reglas de asociación no son de confianza muy alta, por lo que no es tan confiable guiarse por lo que nos digan en futuros análisis. Dejando eso de lado, nos están indicando tres reglas de asociación con un 60 % de confianza:

- La gente con escrupulosidad, estabilidad emocional baja o franqueza neutral neutral tienden a tener una simpatía neutra.

Viendo estas reglas de asociación, no parece haber ningún tipo de regla destacable que nos ayude a determinar qué características implican a otras o cuáles tendencias hay entre las distintas características de personalidad. Esto probablemente nos está diciendo que cuando se evalúan las propiedades de personalidad, no se pueden establecer claramente patrones fijos; ya que las características van a estar variando bastante de manera independiente a las demás.

5.3.3. Formación de clusters.
Ya vimos en las reglas de asociación, cómo es muy posible que todos los puntos de nuestro conjunto de datos de personalidad no tengan patrones bien determinados. Esto se podría ver reflejado también en los clusters que formemos, la consecuencia siendo que no haya una partición bien definida entre cada uno de ellos. A pesar de ello, sigue siendo conveniente ver los resultados que se pueden generar a partir de esto para así observar si al menos se pueden hacer algunas distinciones un tanto significativas entre los grupos generados.

Para la formación de clusters, como ya mencionamos, usaremos el puntaje numérico de cada personalidad. Primero encontremos un k usando el método del codo y un dendrograma. Esto se muestra en las Figuras 10 y 11.

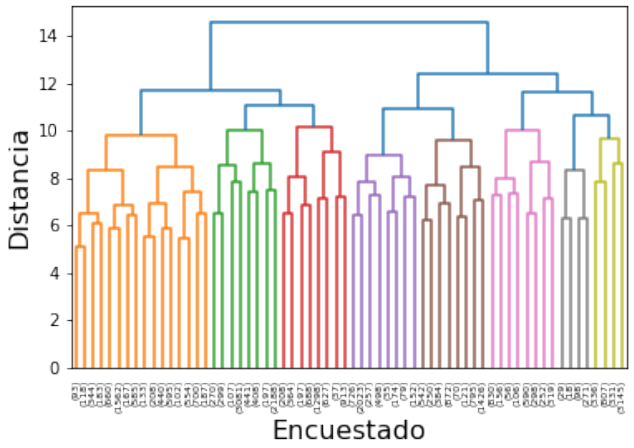

Figura 10: Dendrograma generado a partir de las características de personalidad

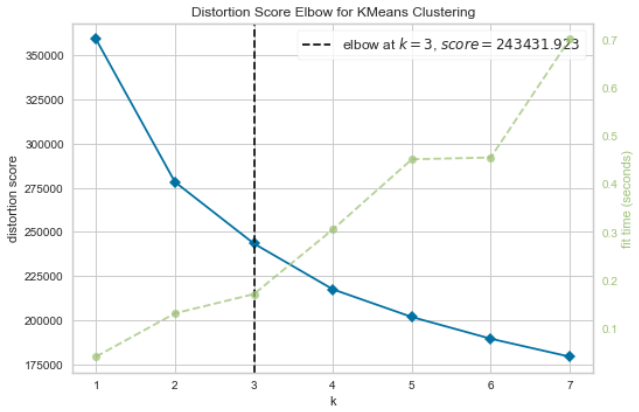

Figura 11: Método del codo para encontrar k en los datos de características de personalidad

Con el dendrograma, podemos observar cómo empieza a aumentar muy rápido el número de clusters conforme bajamos a partir de $k = 3$. Luego en el método del codo, el valor recomendado es $k = 3$. Como el uso de los clusters, dados los resultados del punto anterior, siguen siendo un tanto cuestionables, también visualicemos gráficas del índice de Silhouette para ver qué tan bien resulta hacer este procedimiento. Esto se muestra en la Figura 12.

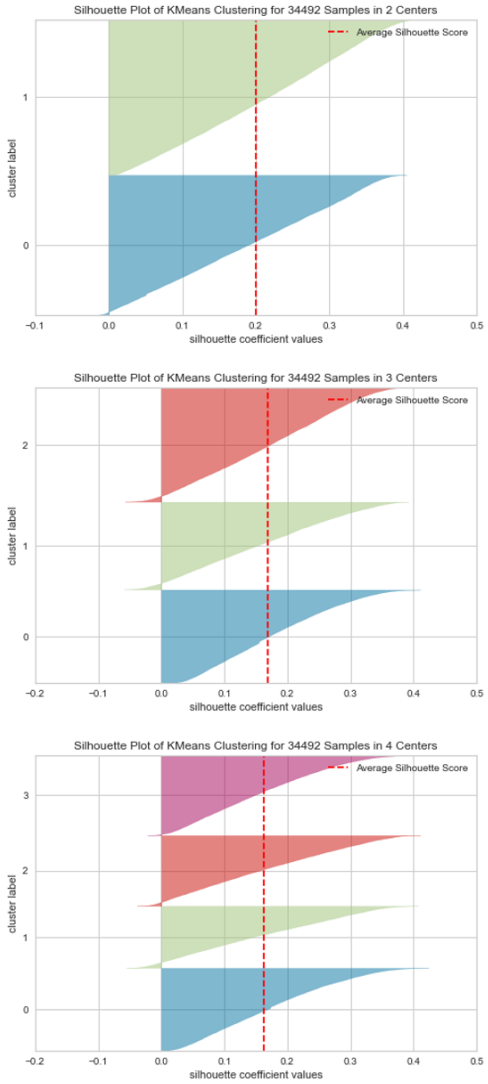

Figura 12: Índices de Silhouette con $k = 2, 3, 4$, para clusters de características de personalidad

Por lo que nos indican las figuras del índice de Silhouette, no parece muy conveniente hacer la división de clusters. Vamos a tener muchos errores al momento de clasificar ya que no podemos hacer grupos bien distinguidos los unos de los otros. Esto lo podemos ver reflejado en los descensos muy rápidos que se forman al evaluar los índices de Silhouette.

A pesar de que no parece que vayamos a obtener información relevante a partir de esta división en clusters; no es mala idea osbervar qué ocurriría cuando aplicamos *K-means* con $k = 3$. Además, los categorías de grupos formados pueden ayudarnos en algunas secciones de más adelante.

Primero obtenemos los centroides de cada cluster generado como se muestra en la Tabla 7.

Cuadro 7: Centroides de los clusters para las características de personalidad

Enfermedad	Cluster 0	Cluster 1	Cluster 2
Extraversión	4.268702	3.267441	1.650431
Simpatía	3.694169	4.571080	3.910011
Escrupulosidad	3.513889	4.796907	3.100075
Estabilidad Emocional	2.190136	4.395453	1.907649
Franqueza	4.371333	4.743487	3.451444

A partir de esa tabla podemos hacer las siguientes anotaciones acerca de los grupos con base en las características de personalidad:

- El cluster 2 es el que tiene los valores más bajos en general. Vemos cómo tienen un nivel de extraversión y estabilidad emocional especialmente bajo. Indicándonos que son el grupo con las personalidades menos estables del resto.
- El cluster 1 contiene valores ligeramente arriba del medio de la escala, esto nos dice la mayoría de su gente tiene aspectos neutrales o ligeramente positivos en cuanto a su personalidad.
- El cluster 0 tiene valores que oscilan en el medio, con una ligera tendencia hacia características negativas. Por esto decimos que son gente que se encuentra con ligeras propiedades negativas de personalidad; nótese el puntaje general bajo en la estabilidad emocional.
- Se cumplió en su mayor parte lo que habíamos mencionado previamente, la distinción entre los clusters no es muy clara. Donde hay mayor variación es en la estabilidad emocional y extraversión; sin embargo la mayoría de las características de personalidad están ocurriendo en un rango muy cerrado que está en el medio de la escala.

Ya que contamos con estos cluster, no está de más visualizarlos sobre los primeros dos componentes principales de los datos correspondientes. Esto se muestra en la Figura 13.

A partir de la Figura 13, podemos observar claramente una idea general de cómo se llevó a cabo la divisón. Lo curioso es observar cómo la figura está partida en tres partes de tamaño similar. Cada una de ellas se inclina hacia cierta parte del plano. Algo que cabe destacar, es que de todas formas, toda la nube de puntos parece parte de un solo grupo, debido a que todos los puntos están muy unidos entre sí. Indicándonos que no hay una separación muy clara entre los distintos tipos de personalidad que podemos encontrarnos. Probablemente lo que influye en la división son los atributos que más diferían en los clusters que definimos (extraversión y estabilidad emocional).

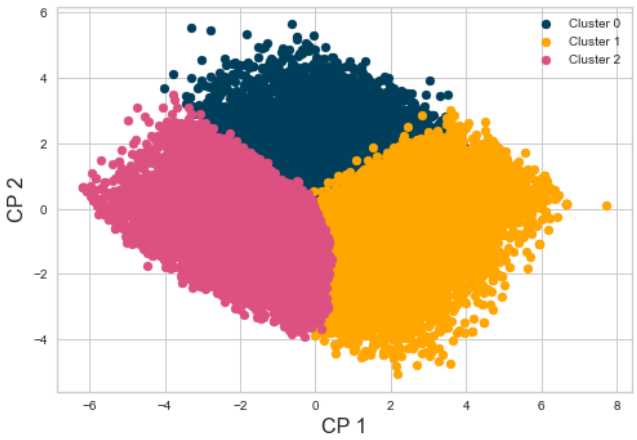

Figura 13: Clusters a partir de las características de personalidad sobre los componentes principales 1 y 2

5.4. Comparación entre los tres distintos agrupamientos de perfiles

Ya analizamos de manera general los distintos tipos de perfiles que presenta el conjunto de datos: el general, el basado en características de personalidad y el que se basa en niveles depresión, estrés y ansiedad. En cada uno de ellos formamos cluster, y como vimos sí había similitudes entre los mismos cluster, pero eran lo suficientemente distinguibles en algunos de sus atributos como para justificar los agrupamientos. Ahora tratemos de observar si entre los cluster formados de cada grupo existe alguna relación con la personalidad, el perfil general y los niveles de estrés, depresión y ansiedad.

5.4.1. Similitud entre clusters.

Lo primero que mediremos será la similitud de los clusters. Para ello, fijémonos en el valor de la información normalizada mutua (*Normalized Mutual Information - NMI*).

Cuadro 8: *NMI* entre cada una de las combinaciones posibles de los clusters de perfiles: general, DASS y de características de personalidad

Tipo de perfil	General	DASS	Personalidad
General	1		
DASS	0.00848	1	
Personalidad	0.00494	0.099767	1

A partir de aquí, ya se nos está mostrando que cada uno de los clusters generados son completamente distintos entre sí. Todas las combinaciones posibles de conjuntos distintos tienen valores numéricos extremadamente pequeños porque los clusters son muy distintos entre sí.

5.4.2. Relacionando los clusters de perfiles con patrones frecuentes.

Ya observamos que los clusters no se parecen entre sí. A pesar de ello, no está de más observar si hay algún patrón frecuente de

perfiles que se presente en el conjunto de datos. Los resultados de los patrones frecuentes entre los tres tipos de perfiles, usando un soporte mínimo de 15 % se muestran en la Tabla 9.

Cuadro 9: Patrones frecuentes entre los tres tipos de perfiles

Soporte	Conjunto
0.621854	(General_0)
0.378146	(General_1)
0.384669	(DASS_0)
0.273571	(DASS_1)
0.205352	(DASS_2)
0.315435	(Personality_0)
0.297402	(Personality_1)
0.387162	(Personality_2)
0.238577	(General_0, DASS_0)
0.180042	(General_0, DASS_1)
0.212716	(General_0, Personality_0)
0.189116	(General_0, Personality_1)
0.220022	(Personality_2, General_0)
0.167140	(Personality_2, General_1)
0.217123	(Personality_2, DASS_0)
0.169894	(Personality_1, DASS_1)

Con estos resultados, podemos observar lo siguiente:

- Todos los perfiles generales, aparecen frecuentemente. El único que está significativamente por debajo de ellos sería el perfil general 0. Aunque la diferencia de su frecuencia no es tanta.
- Entre los perfiles de niveles de estrés, ansiedad y depresión; lo más común es encontrarse con gente que tiene las tres enferemdades de manera severa (cluster 0). En una menor cantidad está la gente que tiene niveles bajos de estas escalas (cluster 1), seguido de personas que tienen niveles de ansiedad severos, pero de depresión y estrés moderados (cluster 3).
- Aparecen los tres grupos principales de perfiles de personalidad. El más común de ellos es el de personas con el nivel más negativo en general, con estabilidad emocional y extraversión bajas (cluster 2). Por debajo, están las personas con personalidades neutras que tienden ya sea a un lado ligeramente negativo por tener baja estabilidad emocional (cluster 0) y las que tienen tendencias ligeramente positivas (cluster 1).
- Algunos patrones de perfiles que aparecen juntos de forma un tanto frecuente son las personalidades negativas (cluster 2) ya sea con nivel de depresión estrés y ansiedad severos o con gente del perfil general 2. También sucede que gente con personalidades neutras que tienden a lo positivo (cluster 1) aparecen frecuentemente con niveles de estrés, ansiedad y depresión bajos.
- Los perfiles generales del cluster 0 se juntan con todo el tipo de personalidades. No podemos llegar a buenas conclusiones con solo esto ya que no se ve un patrón que nos indique que el perfil general 0 tienda o no a cierta personalidad en específico.
- Los perfiles generales que se asocian con un nivel de depresión, estrés y ansiedad; contiene tanto a los de depresión severa como los que la tienen en un nivel muy bajo. Una vez más, esto no nos permite hacer una conclusión apropiada para decir que el perfil general no influye en el perfil de nivel de las enfermedades mentales.

5.4.3. Relacionando los clusters de perfiles con reglas de asociación.

Para complementar los patrones frecuentes, ahora consideremos las reglas de asociación que pueden llegar a ocurrir. Esto se muestra en la Tabla 10.

Cuadro 10: Reglas de asociación entre los tres tipos de perfiles

Antecedente	Consecuencia	Confianza
(DASS_0)	(General_0)	0.620214
(DASS_1)	(General_0)	0.658118
(Personality_0)	(General_0)	0.674357
(Personality_1)	(General_0)	0.635894
(Personality_2)	(General_0)	0.568294
(Personality_2)	(DASS_0)	0.560806
(DASS_0)	(Personality_2)	0.564441
(Personality_1)	(DASS_1)	0.571261
(DASS_1)	(Personality_1)	0.621026

Al obtener esto, podemos confirmar tres de los puntos que establecimos al observar los patrones frecuentes:

- Gente con personalidad neutra que tiende al lado positivo (cluster 1), usualmente tiene bajos o casi nulos niveles de estrés, ansiedad y depresión (cluster 1) y viceversa.
- Gente con personalidad negativa, específicamente con baja estabilidad emocional y extraversión (cluster 2), tienden a tener depresión, estrés y ansiedad en niveles severos.
- Los perfiles generales del cluster 0 se están relacionando con todas las personalidades, por lo que no nos indica un patrón claro entre estos dos tipos de perfiles. Lo mismo ocurre entre perfiles generales de cluster 0 y perfiles de niveles de estrés, ansiedad y depresión de los cluster 0 y 1 (los dos extremos del nivel).

Sin embargo, debemos considerar que esta confianza no es muy alta. Tenemos valores entre 0.56 y 0.66, lo cual no sería algo completamente fiable como para decir que esto casi siempre ocurre. Solo podemos decir que sucede una cantidad considerable de veces como para no pasar desapercibido. De esta forma confirmamos que existe cierta probabilidad ligeramente significativa para que ocurra alguna de estas reglas.

5.4.4. Formando clusters.

Pudimos notar que los conjuntos no se parecen y algunos patrones frecuentes y reglas de asociación. Ahora tratemos de ver si es posible formar clusters que tengan sentido al momento de interpretarlos. Si tratamos a los clusters como datos categóricos y aplicamos

K-means, obtendremos los resultados.

Empezando por encontrar una k, nos basaremos en el método del codo que se muestra en la Figura 14.

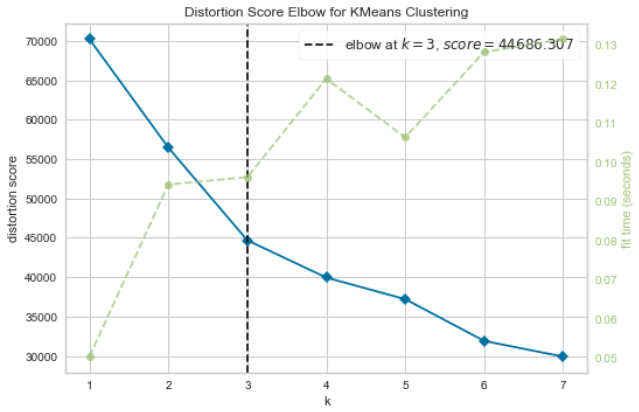

Figura 14: Método del codo para encontrar k en los datos de perfiles generados

Vemos que un k óptimo se forma en $k = 3$, por lo tanto formaremos 3 grupos. Los centroides de los clusters generados se muestran en la Tabla 11.

Cuadro 11: Centroides de los clusters entre los tres tipos de perfiles

Perfil	Cluster 0	Cluster 1	Cluster 2
General_0	0.57	0.64	0.67
General_1	0.43	0.36	0.33
DASS_0	0.56	0.13	0.41
DASS_1	0.10	0.57	0.21
DASS_2	0.18	0.18	0.26
DASS_3	0.16	0.12	0.13
Personality_0	0.00	0.00	1.00
Personality_1	0.00	1.00	0.00
Personality_2	1.00	0.00	0.00

A partir de estos resultados, podemos definir a cada cluster de la siguiente manera:

- Cluster 0: contiene gente con personalidades de niveles bajos en la estabilidad emocional y la extraversión. No hay un perfil general que caracterize a este grupo, y en cuanto a los niveles de depresión, estrés y ansiedad, tenemos que gran parte de ellos tiende hacia los valores severos.
- Cluster 1: contiene gente con personalidades neutrales que tienden a características positivas. De la misma manera que en el cluster 0, no parece haber un perfil general que predomine. En cuanto a los niveles de estrés, ansiedad y depresión, en su mayoría parecen ser bajos.
- Cluster 2: contiene gente con personalidades que están en un valor principalmente neutro, pero tienen tendencias hacia una baja estabilidad emocional. Lo más común es tener personas con niveles severos de depresión, estrés y ansiedad; pero también parece ser normal que haya gente con niveles de depresión y estrés moderados y a la vez ansiedad severa o sino gente con nivel bajo o casi nulo de estas enfermedades mentales. Tampoco parece ser que haya un grupo general bien definido para este cluster.

Una de las características más destacables de este agrupamiento es cómo ningún perfil general parece enfocarse en uno solo de los clusters; todos ellos tienen centroides muy parecidos para este tipo de perfil. Esto probablemente nos está diciendo que el perfil general es algo que no tiene mucho sentido relacionar con la personalidad ni el nivel de depresión. Entonces las propiedades de religión, edad, raza, género, etc., para este conjunto de datos en específico tienen una correlación extremadamente baja o nula con las características de personalidad y niveles de depresión, ansiedad y estrés de una persona.

Por otro lado, volvemos a ver reflejado en los cluster lo mismo que descubrimos en las reglas de asociación. Existe cierta tendencia por algunas personalidades de sufrir mayormente algún nivel general de estrés, depresión y ansiedad. Por ejemplo, está el cluster 0 que representa a las personas de niveles severos de estas enfermedades y personalidades mayormente negativas en la estabilidad emocional y extraversión. El cluster 1 se asocia a la otra regla que habíamos encontrado, que relaciona personalidades positivas neutras con niveles bajos de las enfermedades mentales. Finalmente, el tercero refleja el punto medio entre los extremos tanto en personalidad como en niveles de severidad de las enfermedades.

Otro aspecto interesante es observar cómo los clusters se dividen de manera muy clara por los perfiles de personalidad. Cada uno de ellos tiene asignado un grupo completamente relacionado a un solo tipo de personalidad. Esto tal vez nos está diciendo que aquel tipo de perfiles que *K-means* pudo distinguir con más claridad fue el de personalidad. Lo que genera como resultado es una idea general del nivel de estrés, depresión y ansiedad que tiene cada perfil de personalidad.

Graficando esto sobre los dos componentes principales nos da lo que se muestra en la Figura 15.

En esta figura, la división entre clusters es mucho menos clara que en las anteriores. Esto puede ocurrir por distintas razones como que falte algún porcentaje significativo de la explicación de la varianza o que los datos son tan similares al punto de que no se pueden formar clusters apropiados. A pesar de que no podamos visualizar claramente los clusters sobre estos componentes, de todas formas fue posible observar e interpretar los resultados generados a partir de los centroides obtenidos. Por ello, no descartaría totalmente esta clasificación y lo que implica.

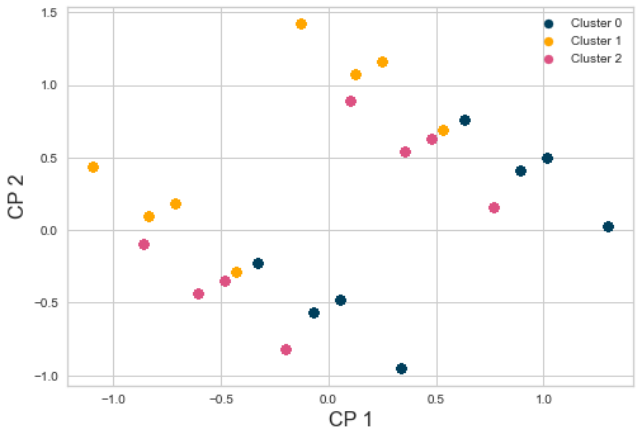

Figura 15: Clusters para perfiles generales, de personalidad y niveles de depresión, estrés y ansiedad

5.5. Análisis sobre las preguntas principales

Hasta ahora, el papel que jugaron las preguntas principales de la encuesta (las de la escala DASS) era calcular directamente el nivel de severidad de cada enfermedad mental. Por esto mismo no sería muy relevante hacer comparaciones entre algo como el puntaje de DASS y estas preguntas. El objetivo para esta sección será el de analizar de manera un poco más detallada las preguntas de la encuesta.

Trataremos de responder las preguntas de, ¿qué patrones y reglas de asociación se pueden encontrar en las respuestas de los encuestados? También, ¿si tuviéramos que separar en categorías los resultados de las preguntas seguiremos obteniendo algo parecido a los clusters formados en la sección 4.2?

5.5.1. Patrones frecuentes.

Al estar analizando patrones frecuentes, tratamos de observar aquellas respuestas y combinaciones de ellas que aparecen de manera más frecuente. Si ponemos un soporte mínimo de 25 %, tenemos una gran cantidad de patrones. Algunos de ellos se muestran en la Tabla 12.

Debido a que hay una gran cantidad de patrones, solo vamos a comentar algunas observaciones respecto a los más frecuentes (los que se muestran en la Tabla 12). Lo que más destaca de estos patrones frecuentes es que lo más común entre todo son las respuestas de categoría 1. Esto implica que las situaciones planteadas por las preguntas no estaban relacionadas en lo absoluto al estado del paciente. Entre estas preguntas tenemos:

- 23: Tuve dificultar al tragar.
- 15: Tuve una sensación de desmayo.
- 19: Transpiraba notablemente (por ejemplo, manos sudorosas) en ausencia de altas temperaturas o esfuerzo físico.
- 7: Tuve una sensación de temblores (por ejemplo, las piernas iban a ceder).

Cuadro 12: Patrones frecuentes entre las respuestas del cuestionario

Soporte	Conjunto
0.636263	(Q23A_1)
0.495100	(Q15A_1)
0.469848	(Q19A_1)
0.457584	(Q7A_1)
0.449061	(Q4A_1)
0.439000	(Q41A_1)
0.385916	(Q35A_2)
0.385510	(Q7A_1, Q23A_1)
0.382871	(Q4A_1, Q23A_1)
0.377595	(Q23A_1, Q19A_1)
0.406500	(Q15A_1, Q23A_1)

- 4: Experimenté dificultad para respirar (por ejemplo, respiración excesivamente rápida, disnea en ausencia de esfuerzo físico).
- 41: Experimenté temblores (por ejemplo, en las manos).

Notemos que todas estas preguntas están relacionadas a síntomas que se manifiestan de manera física. Dentro del cuestionario también hay preguntas más relacionadas con los sentimientos y pensamientos de la persona. Éstas tratan de otro tipo de efectos que podrían derivar de tener algunas de estas enfermedades mentales; pero notemos como las respuestas más comunes son de negación respecto a aspectos que no tienen mucho que ver con esto. Esto nos puede estar diciendo que independientemente de que alguien sufra alguna de estas enfermedades mentales o no, lo menos común es presentar síntomas físicos de estilo.

Dentro del conjunto de patrones frecuentes completo también podemos ver respuestas que van al otro extremo, es decir aquellas con valor 4. Esto implica que la persona estaba altamente relacionada con lo que establecía la pregunta. Algunas de ellas eran:

- 13: Me sentía triste y deprimido.
- 17: Sentía que no valía mucho como persona.
- 11: Me encontré molesto fácilmente.
- 34: Me sentí bastante inútil.
- 40: Me preocupaban las situaciones en las que podría entrar en pánico y hacer el ridículo.

Al contrario de las preguntas que identificamos al inicio, estas son mucho más directas al tratar de determinar el estado de depresión, estrés o ansiedad de una persona. Ya no se van por síntomas físicos, sino por los pensamientos y sentimientos internos de la persona. Esto es algo que ahora sí ayuda a identificar de manera mucho más explícita a alguien con depresión, estrés y/o ansiedad.

Recordemos que al momento de encontrar patrones frecuentes acorde a niveles de estas enfermedades, encontramos que algunos de los grupos más frecuentes eran los de personas con depresión y ansiedad en niveles severos. Esto concuerda con la frecuencia de las respuestas que estamos viendo a estas preguntas.

5.5.2. Reglas de asociación.

Ya tenemos una idea general de lo que está ocurriendo con los

patrones frecuentes de las preguntas. Lo que sigue es formar las reglas de asociación que podrían ayudarnos a complementar lo que encontramos. Al hacer esto con una confianza mínima del 70 % se obtiene una gran cantidad de reglas de decisión. Algunas de las encontradas se muestran en la Tabla 13.

Cuadro 13: Reglas de asociación para las respuestas del cuestionario

Antecedente	Consecuencia	Confianza
(Q4A_1, Q7A_1)	(Q23A_1)	0.897410
(Q4A_1, Q41A_1)	(Q23A_1)	0.895012
(Q15A_1, Q4A_1)	(Q23A_1)	0.893066
(Q34A_4)	(Q17A_4)	0.804365
(Q17A_4)	(Q34A_4)	0.779134

Al estar observando las reglas de decisión de todas las observaciones obtenidas lo más frecuente era encontrar reglas que relacionaban respuestas de puntaje 1 entre sí. Estas respuestas en su mayoría eran combinaciones de los patrones frecuentes que mencionamos en el punto anterior. El mismo caso se repite para las respuestas con puntaje 4.

Esto solo confirma nuevamente lo que ya mencionamos, pero no añade mucha información relevante.

5.5.3. *Formación de clusters.*

Lo último que haremos con este conjunto de preguntas será formar clusters. Lo que haremos será comparar los clusters que se formen apartir de analizar directamente las preguntas con los formados a partir de las sección 4.2. De esta manera podremos determinar qué tanto se parecen estos dos agrupamientos, uno calculado a partir del puntaje general y otro a partir de las preguntas.

Para encontrar la k inicial usemos un dendrograma como referencia visual y el método del codo. Estos se muestran en las Figuras 16 y 17.

Acorde al método del codo, tenemos que $k = 2$ y en el dendrograma se puede hacer el corte alrededor de la distancia 17.5 ya sea para formar 2 o 3 clusters principales. Usaremos en este caso $k = 2$.

Al formar los dos clusters se puede crear una tabla de 42 filas y 2 columnas que representan los centroides de cada cluster en relación a cada respuesta[7]. Lo más importante de esta tabla, es que hay dos patrones muy distintivos entre cada una de ellas:

- En el cluster 0, gran parte de las respuestas tienen centroides de puntaje 3 o más, lo cual nos indica que la gente se sentía identificada con lo que decía la pregunta. Todas las preguntas de esta sección de la encuesta están diseñadas para que a mayor puntaje, mayor nivel se asigne a una de las enfermedades mentales. Por lo tanto, este grupo tiene que ser el que está en el extremo severo de padecer depresión, ansiedad y estrés.
- En el cluster 1, casi todas las respuestas tienen centroides de puntaje menor a 2, indicándonos que la gente no se sentía casi identificada con lo que decía la pregunta. Esto nos lleva a la conclusión de que este grupo tiene que ser el del extremo de gente que sufría un nivel muy bajo o casi nulo de estas enfermedades.

[7]Los resultados completos se muestran en el *Notebook* bajo l a sección de *Relacionando preguntas*

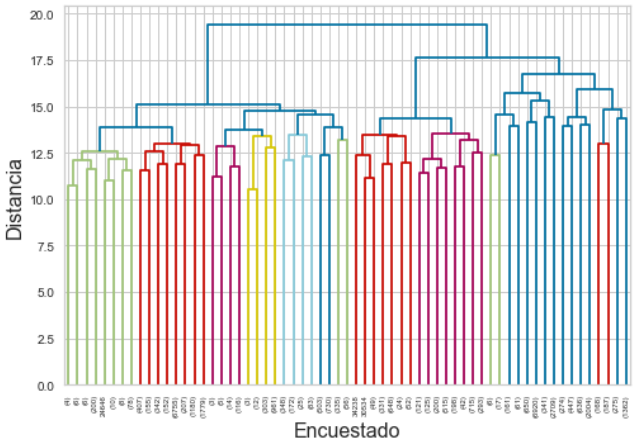

Figura 16: Dendrograma de las respuestas del cuestionario

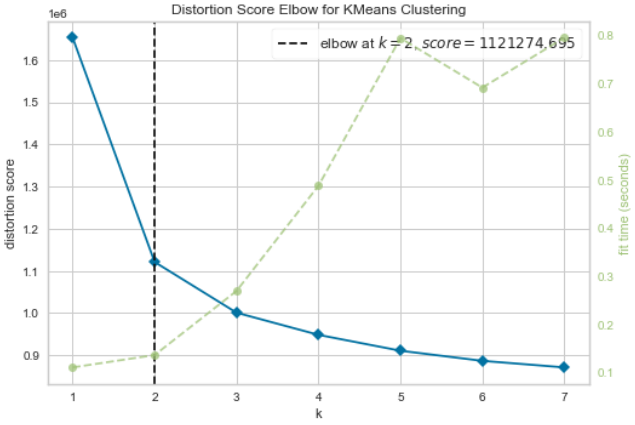

Figura 17: Método del codo para encontrar k con las respuestas del cuestionario

Podemos visualizar los clusters sobre sus dos primeros componentes principales en la Figura 18.

Aquí podemos ver una clara división casi a la mitad del eje del componente principal 1. De igual manera que en casos anteriores, la distancia que hay entre los clusters parece ser muy baja al visualizarla de esta forma; sin embargo sigue siendo posible analizar bien las características de cada conjunto. Otra observación que podemos hacer es que la forma de la nube de puntos se asemeja bastante

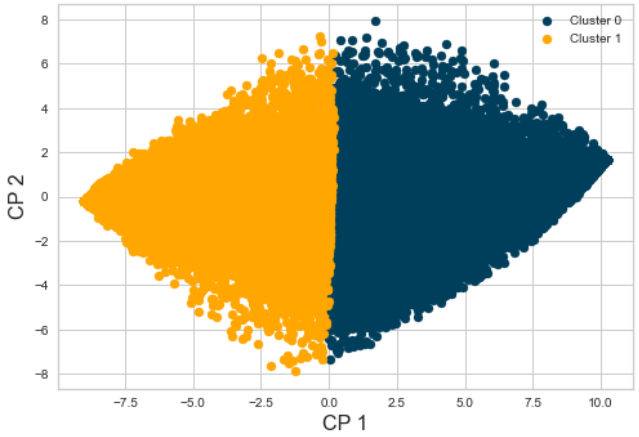

Figura 18: Clusters para perfiles generales con base en las respuestas

Ya que están los clusters formados, se puede calcular su valor de Información Mutua Normalizada (NMI). Como estuvimos trabajando con $k = 2$, parece mejor idea compararlo con el agrupamiento de $k = 2$ en valores DASS. Esto resulta en

$$NMI(\text{PerfilesDASS2, PerfilesRespuestas}) = 0{,}59$$

Notemos que esto ya es un resultado bastante alto si lo comparamos a las previas evaluaciones que se hicieron con este índice. Ahora la pregunta es ¿por qué no llegó a un valor más alto si los valores del puntaje DASS son calculados directamente de estas preguntas?

Una posibilidad detrás de esta ocurrencia es que perdemos variabilidad de los datos al momento de pasar los puntajes obtenidos a partir de las preguntas a una categoría dividida en cuatro niveles. Cuando manejamos las preguntas, estamos trabajando con puntajes con mucha mayor variación en los resultados, pero todo esto se reduce significativamente en diversos rangos de cada pregunta. Además, acorde a la DASS, para poner pesos a cada categoría en teoría equivalentes, los rangos de definición de cada categoría son únicos por enfermedad; provocando aún más causas de que los clusters generados difieran.

Para probar esta teoría, podemos hacer un experimento muy simple. Formemos 2 clusters a partir de los puntajes que se ocupan para la DASS, pero dejémoslos expresados como los puntajes directos. Así no perderemos varianza en los puntajes obtenidos a diferencia de los niveles.

Los clusters generados se muestran en la Tabla 14.
A primera vista nos están volviendo a dividir en los dos extremos que ya hemos venido mencionando varias veces: aquellos que no sufren de manera muy baja de las enfermedades mentales y los que las tienen presentes en un nivel severo.

Ahora saquemos el valor de Información Mutua Normalizada.

$$NMI(\text{PerfilesDASSScore, PerfilesRespuestas}) = 0{,}93$$

Cuadro 14: Centroides de los clusters formados a partir de puntajes usados para calcular niveles depresión, estrés y ansiedad

Enfermedad	Cluster 0	Cluster 1
Depresión	30.999582	11.668021
Ansiedad	23.570892	8.717606
Estrés	29.349299	13.315733

Notemos, que ahora es mucho más elevado que lo obtenido con el puntaje de los niveles de depresión, estrés y ansiedad. Por lo tanto concluimos que la razón por la que los clusters formados a partir de las preguntas y los formados a partir de los niveles DASS no tenían un puntaje tan alto fue por la perdida de varianza que hay al resumir en 4 niveles los puntajes obtenidos.

Para concluir con esta sección, repasemos qué información ganamos a partir de los resultados.

En primer lugar, identificamos el tipo de preguntas que parecen ser menos relevantes en la encuesta con base en la frecuencia de sus respuestas. En este caso las menos importantes eran las que estaban relacionadas con síntomas físicos de la persona.

Después, observamos qué tanto iban a diferir los clusters que se generan a partir de las preguntas de los que se hicieron con los niveles de estrés, ansiedad y depresión calculados. Lo que vimos, fue que ambos agrupamientos tenían un valor de *NMI* alto. Diciéndonos de esta forma que usar las preguntas de manera directa o el nivel general eran ambas opciones que generarían resultados extremadamente parecidos. Con la diferencia de que al momento de hacer el clustering por niveles ocupamos menos columnas en el sentido de que estamos trabajando con 4 niveles de 3 enfermedades; mientras que con las preguntas trabajamos con 4 niveles de 42 preguntas.

A pesar de que los resultados fueran parecidos, seguía habiendo ciertas diferencias. Pero al final se concluyó que esto se debía a una pérdida de variabilidad en los datos al clasificar en categorías los niveles de depresión, estrés y ansiedad más que nada. Por lo tanto la conclusión que podemos obtener es que al pasar estos resultados en una escala, los resultados son más fáciles de interpretar ya que tenemos menos columnas que analizar; y además no perdemos una porción significativa de la interpretación que le podemos dar a los resultados. Por esto mismo, en este conjunto de datos es más útil trabajar directamente con los valores de la escala que se puede calcular a usar directamente las preguntas del cuestionanrio.

6. INTERPRETACIÓN GENERAL SOBRE LOS RESULTADOS

Durante la sección anterior del proyecto logramos hacer clusters para obtener distintos grupos de perfiles. Lo más importante fue que generamos a partir de *K-means* agrupamientos para tres tipos perfiles: general, de personalidad y de niveles de depresión, estrés y ansiedad. De manera adicional, se hizo una comparación de qué

tanto cambiarían los resultados en la comparación de perfiles si trabajáramos con las preguntas del cuestionario directamente en lugar de la escala de estrés, depresión y ansiedad.

Ahora, resumiremos los resultados obtenidos de tal manera que podamos agrupar e interpretar en manera conjunta los resultados. Lo primero a responder es si encontramos patrones o agrupamientos que nos permitan distinguir cada tipo de perfiles de manera clara.

6.0.1. Comentarios sobre los perfiles generales.

Comenzando con los perfiles generales, trabajamos con $k = 2$ y $k = 3$. Los agrupamientos generados al usar $k = 3$ no fueron muy convincentes tras interpretarlos debido a que había dos clusters que eran muy similares en varias de sus propiedades. Por otro lado, usar $k = 2$, nos permitió identificar dos grupos bien distinguidos:

- Cluster 0: principalmente mujeres de reiligión musulmana, raza asiática y cuya lengua nativa no es el inglés.
- Cluster 1: mujeres y hombres (principalmente mujeres), agnósticos o ateos, raza blanca y cuya lengua nativa sí es el inglés.

A partir de esto podemos notar ciertas características sobre la gente que tomó la encuesta. En primer lugar, parece ser que quienes realizaron la encuesta se enfocaron principalmente en dos regiones distintas. Esto lo concluimos a partir de las características de religión, raza y lengua nativa. Por otra parte, las propiedades que se mantuvieron casi iguales a lo largo entre los agrupamientos fueron los rangos de edades, niveles de educación, tipo de zona donde cada persona creció, orientación sexual, participación en votaciones y algunos otros.

En el caso de la edad, la población no era muy variada ya que en ambos cluster la gran mayoría de las edades andaban entre los 10 y 30 años; a partir de los 30 ya empezaba a haber una gran disminución de la población. Lo mismo ocurría con otras propiedades como el estado civil, la orientación y el género. En cuanto a otros datos, como el tipo de zona y la educación sí variaban en cantidad considerable por sí solos; pero parece ser que en cada región que se tomó la encuesta, las proporciones de estos valores eran tan similares que al momento de hacer clustering no se diferenciaron.

En otros casos, como el de $k = 3$ vimos que sí llegaba a variar la participación de votación, la edad y el nivel de educación. Aunque esto solo ocurría cuando partíamos los datos de una sola región (como en este caso la región donde se encontraba la gente asiática), haciendo que no pudiéramos diferenciar todos los cluster por las mismas características.

Concluyendo con esta primer clasificación, diremos que lo más relevante de aquí fue la identificación de dos regiones principales, se llegó a esto debido que las propiedades que mayor cambio tenían entre los clusters formados eran la religión, raza y lengua natal. Por otro lado, si hubiéramos intentado hacer clustering con base en otras propiedades, habríamos tenido que hacer una selección apropiada de una cantidad de datos que cumplieran con ciertas características.

6.0.2. Comentarios sobre los perfiles de niveles de estrés, ansiedad y depresión.

Después de eso pasamos a calcular los puntajes de la escala de estrés, depresión y ansiedad (DASS). A partir de esto se encontraron algunos patrones en nuestra población que fueron remarcados tanto por los patrones frecuentes, como las reglas de asociación y el clustering.

Algunas cosas que se alcanzaron a distinguir antes de formar asociaciones entre categorías y formar clusters fue observar cómo había unos niveles mucho más comunes que otros. Por ejemplo, lo más común era observar que la mayoría de las respuestas confirmaban que entre la ansiedad y la depresión, la mayoría de las personas presentaban los niveles más severos de esas enfermedades. La excepción de esto fue el estrés, deonde la mayoría parecía tenerlo en un nivel bajo.

Formamos dos agrupamientos: uno con $k = 2$ y otro con $k = 4$. Lo que destacó de ambos es que había un cluster con una población que se caracterizaba por sufrir depresión, ansiedad y estrés a niveles severos; mientras que el otro cluster era de gente que no tenía niveles extremadamente bajos de esto. Esto permitió observar cómo era muy común que gente que tuviera dos o una de las tres enfermedades mentales en un nivel severo muy probablemente iba a padecer de por lo menos una más a niveles también severos.

Algo similar ocurría para los niveles muy bajos. Si alguien tenía un nivel extremadamente bajo con dos de las enfermedades, muy probablemente tampoco sufriría de la otra.

Lo último que se concluyó con estos agrupamientos fue que con $k = 4$ se mostró que todavía es posible hacer distinciones en las escalas de los niveles de cada grupo. La gente no se va a únicamente a uno de los dos extremos, sino que alguien puede padecer en niveles moderados estrés, ansiedad o depresíon. Lo interesante fue observar que esta vez, las tres enfermedades no estaban al mismo nivel como en los casos de los dos extremos. Había algunas enfermedades que se acercaban a un nivel severo, otras estaban en rangos moderados o casi normales. Por lo tanto, la depresión, ansiedad y estrés no siempre se manifiestan al mismo nivel, sino que va variando; esto ya empieza a cambiar cuando una o dos de estas enfermedades se encuentran en uno de los niveles extremos.

6.0.3. Comentarios sobre los perfiles de personalidades.

Luego de eso tenemos el análisis de las personalidades. Se obtenían los puntajes de las propiedades de personalidad generadas a partir de algunas de las preguntas de la encuesta y a partir de ahí se evaluaban los datos. Aquí aplicamos patrones frecuentes, reglas de asociación y clustering; sin embargo, los primeros métodos no nos dieron mucha información interesante o que pareciera relevante. Cuando se hizo el clustering en $k = 3$ grupos, sí se empezaron a notar ciertas características en el conjunto.

Los resultados fueron un tanto parecidos a los que se obtuvieron con los niveles de estrés, depresión y ansiedad. En este caso había un cluster que tenía características de personalidad generalmente en niveles bajos, especialmente en la extraversión y estabilidad

emocional. Otro cluster tenía valores en el nivel medio, pero que por lo general estaban un poco arriba de la mitad de la escala. El último consistía de valores que también estaban a la mitad de la escala, pero el puntaje general de la estabilidad emocional estaba en un rango bajo.

Todos estos agrupamientos eran un tanto similares por el hecho de que había valores de centroides entre clusters que se llegaban a parecer mucho. A pesar de ello, había suficientes diferencias como para establecer las diferencias entre cada uno de ellos. Lo que sucedía es que el espectro de personalidades que se mostraba llegaba a variar mucho entre las personas, pero todavía era posible trazar una sección de personalidades negativa, una neutra con algunas características negativas y una que tendía a puntajes de personalidad positiva.

6.0.4. Comentarios sobre las relaciones entre los perfiles.

El último análisis relevante fue el de comparación entre los tres distintos agrupamientos de perfiles. Se siguió el mismo procedimiento que en los análisis anteriores. Solo que esta vez el objetivo era relacionar todos los resultados generados para tratar de ver si había patrones entre la personalidad, el perfil general y/o el nivel de estrés, depresión y ansiedad.

A partir de lo mostrado en esa sección llegamos a algunas conclusiones.

La primera de ellas es que en este conjunto de datos, el perfil de la persona no se relaciona con la personalidad ni con el nivel de estrés, ansiedad y depresión. Esto nos quiere decir que en este caso las propiedades de una persona como lo son la lengua natal, la religión y la raza no van a afectar de manera significativa los niveles de las tres enfermedades mentales que hemos venido trabajando. La justificación de ello, es que cuando buscamos patrones frecuentes, reglas de asociación y clusters; en todos los casos el perfil general no marcaba información relevante. No se encontraba en algún cluster en específico y sus patrones de asociación no decían nada especial.

En segundo lugar, sí pudimos observar una ligera relación entre los grupos de personalidad y los de niveles de estrés, ansiedad y depresión. Estos patrones aparecían en los extremos de cada uno de los perfiles. Por un lado, pudimos notar que aquellas personas que tenían niveles bajos de aspectos como la estabilidad emocional y la extraversión eran comunmente asociadas con niveles severos de salud mental. Esto tiene mucho sentido, es normal pensar que un problema de las personas que tienen baja estabilidad emocional, por alguna u otra razón, están en un mayor riesgo de sufrir de estas enfermedades como consecuencia.

Por otro lado, observamos que las personas con características de personalidad ligeramente positivas tenían una tendencia a no tener estrés, depresión y/o ansiedad. En este caso, las características de personalidad que más sobresalían a lo positivo eran la escrupulosidad, la simpatía y la franqueza; la estabilidad emocional estaba estaba en un valor neutro y la extraversión estaba un poco debajo de éste. Probablemente, lo que más contribuía en este caso para establecer la relación entre los grupos fue la estabildad emocional establecida por la personalidad. Esto es porque es el valor que más lo diferenciaba del cluster de personalidad negativa. En el caso del valor de extraversión, no parece que haya habido tanta influencia por el hecho de que que este valor andaba por debajo del puntaje neutro.

En conclusión, podríamos decir que al evaluar la personalidad de alguien y querer obtener una primera aproximación hacia sus niveles de estrés depresión y ansiedad; el valor más importante es la estabilidad emocional. Si está en un nivel muy bajo, la posibilidad de que la persona tenga depresión, estrés y ansiedad severos puede ser un tanto probable. En cambio, si está en un nivel al menos neutro, la posibilidad de que la persona sufra alguna de esas enfermedades en niveles bajos o casi nulos también puede ser algo común.

Lo que también hay que tener en consideración es que esto no es una tendencia absoluta o casi definitiva. Por lo que vimos al formar los cluster, cada una de las personalidades va a tener gente de todos los cuatro grupos de severidad de depresión, estrés y ansiedad que marcamos. Lo que ocurre es que ciertos grupos de severidad de las enfermedades son más propensos a pertenecer en determinadas personalidades.

6.0.5. Comentarios sobre los datos y la información obtenida.

Como últimas observaciones de esta sección quisiera plantear dos preguntas: ¿qué tanto contribuye esta exploración de los datos al entendimiento de los mismos? Y, ¿acaso lo encontrado es algo representativo de una población más allá de la que presenta este conjunto?

Empecemos por tratar de responder la segunda.

Mientras estuvimos evaluando los datos, hubo una cosa que me llamó la atención. Esta fue la gran cantidad de similaridades que había entre las personas que tomaron la encuesta. Más allá de las diferencias que encontramos en la religión, raza y lengua nativa, no había muchas más comparaciones significativas que se pudieran hacer. Esto lo podemos notar en cómo casi toda las respuestas que tenemos son por parte de jóvenes (gente entre 10 y 30 años). Luego está la cuestión del género, tuvimos respuestas de una gran cantidad de mujeres, pero eran tantas a comparación de los hombres que terminó siendo una característica ignorada al momento de evaluar algo como perfiles generales.

Claro que había gente con características distintas a las de la mayoría, pero el problema está en que las muestras que teníamos de éstas eran significativamente menores. Al final de todo, pienso que es algo que no afectó mucho a la exploración de los datos en general. Pero si el objetivo final de algún trabajo consistiera de hacer un análisis de estos datos para ver algo como la influencia que tiene la región, edad, raza, género, etc., en las personalidades y/o niveles de estrés, depresión y ansiedad; pienso que este conjunto carece de suficientes muestras de varios sectores de población como para llegar a conclusiones buenas. Una solución podría ser cortar los datos de tal forma que se tengan proporciones no tan desniveladas entre los sectores de la población con la que contamos, o sino solo

usar las características de la población que buscamos relacionar con los demás perfiles.

Esto mismo pudo haber tenido influencia en los resultados del momento en que tratamos de relacionar los distintos perfiles y obtuvimos que el perfil general no tiene casi nada que ver. De todas formas, esto no es una afirmación definitiva, ya que puede ser que realmente los perfiles generales no tienen relación alguna con los demás a pesar de que contemos con suficientes datos que sí tengan porciones significativas de distintas poblaciones.

Dejando de lado el aspecto de los perfiles generales, donde sí tuvimos una buena cantidad de variación en los datos fue para los perfiles de personalidad y niveles de estrés, depresión y ansiedad.

Pudimos observar relaciones entre características de la personalidad y cómo tienen cierta tendencia a influir sobre estas enfermedades mentales. También dentro de cada perfil se alcanzó a distinguir cómo las propiedades de cada uno se llegan a comportar en conjunto (por ejemplo cuando tenemos dos enfermedades mentales en un extremo, es muy probable que la otra también esté en ese mismo).

Puede que la encuesta se haya hecho probablemente sobre dos sectores principales de personas. Sin embargo, los resultados de los distintos niveles de depresión, estrés y ansiedad es algo que no se puede saber de manera adecuada hasta que se obtienen los datos. Por esta razón, pienso que no sería raro encontrar lo que observó aquí para otros conjuntos que se basen en estas escalas y categorías de personalidad y niveles de las enfermedades mentales correspondientes.

Entonces concluimos que la respuesta para la segunda pregunta que planteamos es que lo que encontramos sí puede ser algo que se repita en otros conjuntos de datos que traten los mismos tipos de datos de estrés, depresión, ansiedad y personalidad. Pero hay que tomar en consideración que el caso de los perfiles generales no es algo que aplique muy bien para otro tipo de poblaciones, debido a las limitaciones que había en los grupos de gente que tomó la encuesta.

Pasando a la primera pregunta, pienso que al final de toda esta experimentación expusimos algunos aspectos de las personalidades, los niveles de estrés, depresión y ansiedad, de las preguntas del cuestionario principal e incluso de los perfiles generales que no eran tan fáciles de identificar sin haber hecho esta exploración de los datos.

En cuanto a los perfiles generados y sus relaciones, ya mencionamos todo lo que encontramos al inicio de esta sección y también en la de experimentos y análisis. En cuanto a estos, pudimos ver que sí hubo algunos aspectos que estaban muy relacionados entre sí, como lo que ocurría en los extremos de los niveles de las enfermedades o las personalidades. A pesar de ello, parece ser que en nuestros datos hay tanta variación entre las características de las personalidades y los niveles de estrés, depresión y ansiedad que resulta no ser tan fácil encontrar un agrupamiento o conjunto de reglas tan evidentes que nos permitan establecer las características específicas de una persona con estas enfermedades.

En conclusión, esto sirve para darnos ideas muy generales de cómo se relacionan las características de los perfiles generales de este conjunto de datos en particular, la personalidad y de los niveles de estrés, depresión y ansiedad. Sin embargo, al ver los niveles de similaridad entre los patrones frecuentes y clusters encontrados, podemos decir que esto no es algo que sirva de mucho para decir con alta confianza las relaciones formadas.

7. CONCLUSIÓN

La exploración de estos datos por medio de procesos de clustering con *K-means*, patrones frecuentes y reglas de asociación nos permitió tener un mejor entendimiento de los datos. A pesar de que se hayan llegado a ciertas conclusiones generales respecto a su comportamiento, también pudimos apreciar que, al menos con estos métodos aplicados de esta forma, es difícil identificar o establecer alguna regla, patrón o agrupamiento definitivo que nos permita caracterizar de manera adecuada y específica los niveles de depresión, estrés y ansiedad de una persona con base en otras características como lo es la personalidad.

8. BIBLIOGRAFÍA

Gosling, S. D., Rentfrow, P. J., Swann, W. B., Jr. 2003. *A Very Brief Measure of the Big Five Personality Domains*. Journal of Research in Personality, 37, 504-528. http://gosling.psy.utexas.edu/wp-content/uploads/2014/09/JRP-03-tipi.pdf

Peleg, Yam. 2021. *Predicting Depression, Anxiety and Stress*. Kaggle. https://www.kaggle.com/yamqwe/depression-anxiety-stress-scales

Psytoolkit. 2021. *Depression Anxiety Stress Scales (DASS)*. https://www.psytoolkit.org/survey-library/depression-anxiety-stress-dass.html

9. RECREACIÓN DE RESULTADOS

Todos los resultados, incluyendo, tablas y gráficas generadas se pueden recrear en el *Notebook* que viene junto a este proyecto. También se cuenta con un tablero hecho en Dash con fines de visualización de algunos datos y resultados obtenidos.