Universidad de Buenos Aires

Facultad de Ciencias Exactas y Naturales

Maestría en Exploración de Datos y Descubrimiento del Conocimiento

Análisis sobre depresión ansiedad y estrés basado en respuestas online al test DASS-42

Autor: Juan Jose Iguaran Fernandez

Índice

Índice de figuras

Índice de cuadros

1. Introducción

Las salud mental es un componente fundamental dentro de la salud humana, y esta relevancia esta evidenciando tener aumento creciente tanto a nivel social, como de programas de política publica relacionadas con este tema [**world2018mental**]. Dentro de este contexto, cabe resaltar que las condiciones mas prevalentes en la población mundial son la ansiedad y la depresión respectivamente [**james2018global**].

La adecuada identificación y diagnostico de las condiciones mentales en general y de la ansiedad y la depresión en particular, suponen un desafió a los profesionales de la salud debido en gran medida por la interpretación subjetiva de los patrones de pensamiento y conducta evidenciados en los pacientes [**beck1961inventory**].

Dentro de este marco surgen las técnicas de diagnostico conocidas como los cuestionarios de auto relato cuyo objetivo es proveer una detección objetiva sobre la condición del paciente mediante basado en la enunciación de declaraciones propias de síntomas y actitudes características de la condición en cuestión, en donde el paciente proporciona un valor numérico dentro de una escala en la medida en que se sienta mas o menos identificado con dicha declaración.

El Depression Anxiety Stress Scale, DASS 42, es uno de dichos cuestionarios en donde se presentan 42 enunciados, 14 para cada condición y el evaluado califica la enunciación de 1 a 4 dependiendo su nivel de afinidad en la ultima semana. Luego, para cada condición se suman los valores de todos los enunciados que la componen y este resultado se ubica dentro de un grado de severidad que va desde normal hasta extremadamente severo dependiendo de su valor [**parkitny2010depression**].

El siguiente trabajo tiene por objetivo aplicar técnicas de aprendizaje supervisado y no supervisado sobre una base de datos de respuestas al DASS 42 ofrecidos por usuarios de manera anónima a través de la pagina web https://openpsychometrics.org, cuyo objetivo es proveer data abierta y anónima que permita la investigación en psicología.

Inicialmente se entrena un algoritmo de aprendizaje supervisado, específicamente un random forest en donde una vez seleccionado los mejores hiperparametros, se estimara la importancia de las variables obtenido para intentar entender el poder predictivo de las variables a la hora de diagnosticar. Luego se aplican técnicas de clustering sobre los datos que incluyen además de las respuestas al cuestionario, información demográfica y sobre la personalidad de las personas y se aprecia si existe una semejanza entre los clusters formados y el diagnostico obtenido a través del test.

2. Datos

EL dataset esta constituido por los 42 enunciados que constituyen el DAAS 42, los 10 enunciados que constituyen el Ten Item Personality Inventory(TIPI) test[**gosling2003very**], cuyo propósito es que el paciente de una valoración de 1 a 7 para cada enunciado que corresponde a características de la personalidad de la personalidad, además de las siguientes variables demográficas: educación, tipo de zona urbana en la que se creció, genero, edad, religión, orientación sexual, raza, estado civil y tamaño de la familia en donde se creció, para dar un total de 61 campos. Información detallada sobre los diferentes campos esta disponible en los cuadros 19 y 18 en la sección de anexos. Para estos campos, el dataset cuenta con un total de 39775 registros que corresponden a las respuestas al test dadas por cada voluntario de diferentes nacionalidades y registrada de manera anónima. De este total fueron removidos aquellos registros que registraban un tamaño familiar superior a 20 y una edad superior a 90 pues son consideradas registros corruptos, quedando así con un total de 39759 registros.

En el cuadro 1 se presenta un perfil medio de las variables demográficas de la población total, en donde las variables numéricas fueron promediadas y las variables categóricas se obtuvo su media:

Cuadro 1: Perfil medio dentro de cada diagnostico para la población general

education	urban	gender	religion	orientation	race	married	TIPI1	TIPI2	TIPI3	TIPI4	TIPI5	TIPI6	TIPI7	TIPI8	TIPI9	TIPI10	age	familysize
3	3	2	10	1	10	1	3.79	4.19	4.74	5.17	4.93	4.85	5.27	4.28	3.65	3.73	23.4	3.5

Se evidencia como los valores mas frecuentes para los atributos demográficos son educación universitaria, haber crecido en un ambiente urbano, genero femenino, religión musulmana, de orientación heterosexual, raza asiática, nunca han estado casados, de un poco mas de 23 años de edad y con un numero de hermanos de un poco mas de3 incluyéndolos. Así mismo se aprecia que los enunciados de TIPI que mas puntaje tienen son el 4 y el 7 y los que menos tienen son el 1, el 9 y el 10.

El primer paso para realizar un diagnostico dentro del DAAS 42 es sumar la valoración dada en todos los enunciados que corresponden a cada condición, que se puede apreciar en el cuadro 20 de los anexos obteniendo así un puntaje. En el cuadro 2 se presenta una matriz con el valor obtenido al calcular la correlación de Pearson en el puntaje para cada par de condiciones:

Cuadro 2: Correlación entre los puntajes para las distintas condiciones

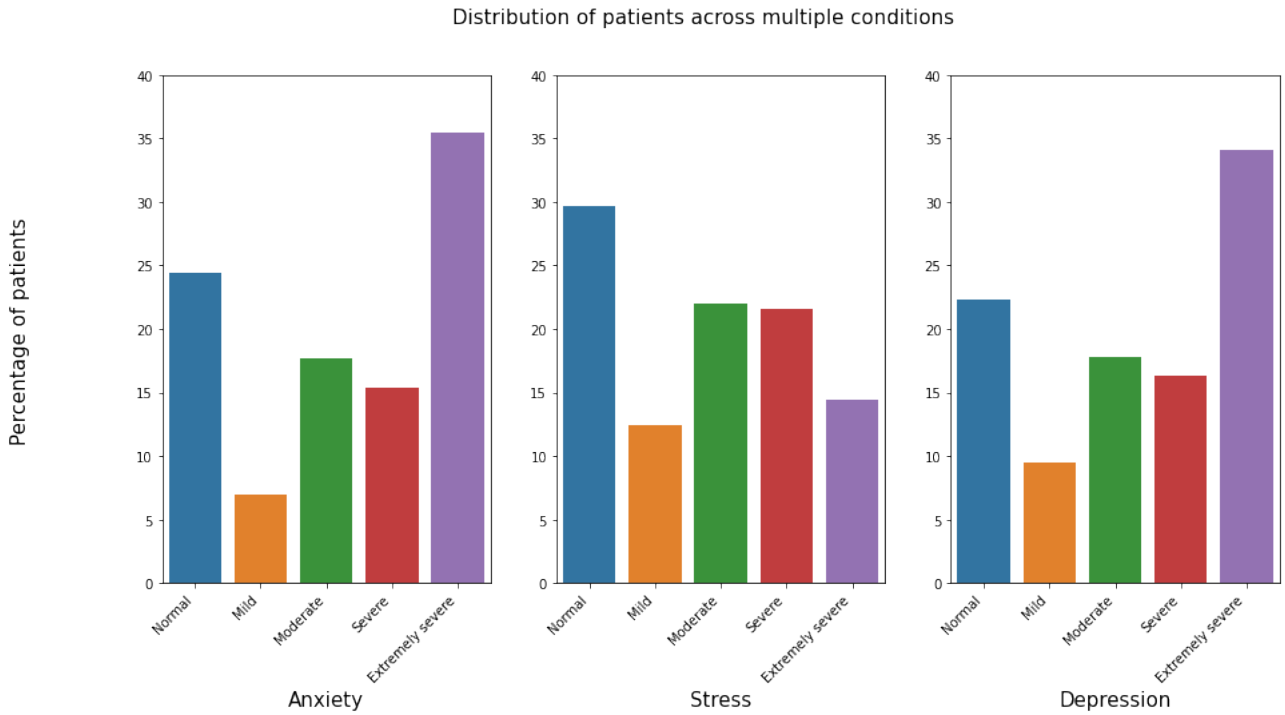
	Stress	Anxiety	Depression
Stress	1.00	0.80	0.74
Anxiety	0.80	1.00	0.67
Depression	0.74	0.67	1.00

Se aprecia que existe una correlación significativa en el puntaje obtenido entre las tres condiciones, siendo la condición de estrés la que se relaciona mas fuertemente con las otras dos, en particular la ansiedad y el estrés. Así mismo la correlación mas débil se da entre la depresión y la ansiedad.

Una vez obtenido el puntaje de cada condición, la siguiente etapa es asignar un diagnostico dependiendo del valor del puntaje, pues al comparar dicho puntaje, con el umbral mínimo de cada diagnostico, se le asigna al paciente determinada condición si su puntaje es mayor a su umbral mínimo. Los valores de dichos umbrales se pueden encontrar en el cuadro 21 de la sección de anexos.

Luego de diagnosticados los pacientes, se procede a cuantificar que porcentaje de la población se encuentra dentro de cada diagnostico para cada condición y el resultado de dicha cuantificación se aprecia en la figura 1.

Figura 1: Porcentaje de personas en cada diagnostico para cada condición



Se evidencia que para las condiciones de ansiedad y depresión, el mayor numero de personas se encuentra dentro del diagnostico extremadamente severo, seguido del diagnostico normal siendo el de menor preponderancia los de síntomas leves. Por otro lado la condición de estrés presenta el mayor numero de personas en el diagnostico normal, seguido por moderado y severo en forma semejante y los de menor preponderancia son leve y extremadamente severo respectivamente. Es interesante que aun cuando la condición de estrés tiene una alta correlación en su puntaje con las otras dos, las distribuciones difieren en cuanto a su diagnostico, esto se explicaría por el hecho de que los umbrales mínimos para cada condición son diferentes y esto afecta directamente la distribución de los diagnósticos.

Así mismo a partir del diagnostico de los pacientes, se puede calcular el perfil medio de cada diagnostico para cada condición como se evidencia en los cuadros 3, 4 y 5:

Cuadro 3: Perfil medio dentro de cada diagnostico para ansiedad

diagnosis	education	urban	gender	religion	orientation	race	married	TIPI1	TIPI2	TIPI3	TIPI4	TIPI5	TIPI6	TIPI7	TIPI8	TIPI9	TIPI10	age	familysize
Normal anxiety	3	3	2	10	1	10	1	4.17	3.80	5.15	3.81	5.41	4.42	5.27	3.67	4.74	3.42	26.94	3.48
Mild anxiety	3	3	2	10	1	10	1	4.01	4.01	4.89	4.71	5.17	4.65	5.28	4.07	4.10	3.63	24.09	3.52
Moderate anxiety	3	3	2	10	1	10	1	3.80	4.17	4.78	5.11	5.00	4.84	5.25	4.24	3.77	3.76	23.45	3.54
Severe anxiety	2	3	2	10	1	10	1	3.68	4.24	4.63	5.53	4.86	4.92	5.29	4.40	3.41	3.81	22.36	3.48
Extremely severe anxiety	2	3	2	10	1	10	1	3.52	4.49	4.47	6.08	4.56	5.16	5.28	4.71	2.86	3.92	21.25	3.50

Cuadro 4: Perfil medio dentro de cada diagnostico para depresión

diagnosis	education	urban	gender	religion	orientation	race	married	TIPI1	TIPI2	TIPI3	TIPI4	TIPI5	TIPI6	TIPI7	TIPI8	TIPI9	TIPI10	age	familysize
Normal depression	3	3	2	10	1	10	1	4.45	3.75	5.34	3.92	5.49	4.24	5.41	3.54	4.91	3.38	25.38	3.63
Mild depression	3	3	2	10	1	10	1	4.13	4.07	4.99	4.84	5.20	4.60	5.35	4.04	4.16	3.63	24.16	3.61
Moderate depression	3	3	2	10	1	10	1	3.91	4.16	4.79	5.24	5.02	4.80	5.31	4.23	3.75	3.73	23.37	3.55
Severe depression	2	3	2	10	1	10	1	3.69	4.32	4.62	5.50	4.85	4.92	5.28	4.48	3.39	3.79	22.65	3.45
Extremely severe depression	2	3	2	10	1	10	1	3.24	4.47	4.32	5.90	4.50	5.31	5.14	4.76	2.76	3.95	22.27	3.38

Se aprecia que algunas de estas variables cambian de valor proporcionalmente de manera a análoga en todos las condiciones: el nivel educativo tiende a ser menor para diagnósticos severos, así mismo la edad tiende a disminuir con la severidad. En cuanto al TIPI, los enunciados 1, 3, 5, y 9 tienden a

Cuadro 5: Perfil medio dentro de cada diagnostico para estrés

diagnosis	education	urban	gender	religion	orientation	race	married	TIPI1	TIPI2	TIPI3	TIPI4	TIPI5	TIPI6	TIPI7	TIPI8	TIPI9	TIPI10	age	familysize
Normal stress	3	3	2	10	1	10	1	4.15	3.65	5.10	3.79	5.36	4.51	5.32	3.76	4.84	3.51	25.47	3.61
Mild stress	3	3	2	10	1	10	1	3.86	4.02	4.78	4.97	5.06	4.81	5.30	4.21	3.91	3.74	23.66	3.48
Moderate stress	2	3	2	10	1	10	1	3.72	4.26	4.67	5.50	4.90	4.92	5.28	4.38	3.45	3.75	22.80	3.46
Severe stress	2	3	2	10	1	10	1	3.58	4.54	4.55	6.01	4.69	5.02	5.24	4.57	2.91	3.84	22.16	3.44
Extremely severe stress	2	3	2	10	1	10	1	3.39	4.84	4.37	6.42	4.37	5.23	5.20	4.82	2.38	3.96	21.69	3.43

disminuir, lo cual tiene sentido pues están asociados a características positivas de la personalidad como entusiasmo, disciplina, apertura a nuevas experiencias y calma. Por otro lado los enunciados 2, 4, 6 y 8 tienden a aumentar con el diagnostico; estos enunciados están a su vez asociados con características negativas de la personalidad como conflictividad, irritabilidad, reserva y desorden.

3. Metodología

El presente análisis tiene por objetivo descubrir y exponer conocimiento a través de dos tipos de análisis: Aprendizaje supervisado y aprendizaje no supervisado.

3.1. Aprendizaje supervisado

Para el aprendizaje supervisado, se utiliza como variable objetivo el diagnostico del paciente. Puesto que son tres condiciones diferentes, cada uno con distintos valores para el diagnostico, el análisis efectuado fue la construcción de tres modelos por separado en donde se usan todas las variables y registros de los datos y la variable objetivo fue el diagnostico de cada condición en particular.

Puesto que los modelos son entrenados con la data proveniente de los distintos test suministrados y estos tienen a su vez una manera establecida para establecer el diagnostico, resulta poco interesante que el objetivo del análisis sea entrenar un modelo cuyo objetivo sea predecir la clase dado el set de datos. Por esta razón, el foco del presente análisis esta en la explicabilidad de los modelos, obteniendo de esta manera información en cuanto a que tan importantes son las distintas variables que constituyen el set de datos a la hora de efectuar los distintos diagnósticos de la condición; esto permite identificar que enunciados del test o atributos demográficos resultan determinantes a la hora de diagnosticar una condición.

El modelo escogido para el análisis fue un random forest. Para poder entrenar y evaluar dicho modelo, se separa la totalidad de los datos en el set de entrenamiento y set de test con una proporcion del 75 % y 25 % respectivamente.

Para determinar los hiperparametros del modelo, se realizo una grid search utilizando una validación cruzada de 3 particiones, buscando maximizar la medida de desempeño, en este caso el accuracy, teniendo en cuenta el numero de arboles, la profundidad de los arboles y el criterio de partición usando los valores de búsqueda presentes en la tabla 6:

Cuadro 6: Hiperparametros considerados en el Grid Search

	Hyperparameter Values
Number of trees	50, 100, 200, 500, 1000
Max depth	2, 4, 6, 8, 10
Criterion	'gini', 'entropy'

Una vez establecido los hiperparametros que otorgan una mayor accuracy, se procede a evaluar su desempeño en el set de test mediante el calculo del accuracy, así como la estimación de la importancia de las distintas variables para la estimación del modelo. Esta importancia es calculada para cada atributo, a partir de que tan bien es capaz dicho atributo de separar las clases objetivos, en este caso los diagnósticos para cada condición, según el valor que tome el atributo. Esta importancia es calculada para todos las variables del modelo y se ordenan de mayor a menor para apreciar cuales son las variables mas importantes a la hora de realizar la clasificación.

3.2. Aprendizaje no supervisado

Para el aprendizaje no supervisado, se procede a realizar un análisis de clusters con el objetivo de determinar posibles agrupaciones que puedan surgir a partir de la información contenida en los tests además de los datos demográficos, y comparar dichos clusters con las diagnósticos realizados para cada condición.

Puesto que dentro del set de datos contamos con datos mixtos, es decir, variables numéricas y categóricas, los métodos tradicionales de clustering no son suficientes para elaborar el análisis pues se centran en algún tipo especifico de variable. Por consiguiente, para elaborar el modelo de clustering, se recurre a un algoritmo conocido como K-prototype.

El concepto detrás de este algoritmo es homologo al de K-means [**huang1998extensions**] : se establecen la pertenencia de cada registro a alguno de los n clusters a formar basado en la distancia que existe de entre dicho punto y el centroide del cluster. Esta distancia debe ser establecida para cada uno de las columnas que constituyen el dataset, teniendo pues una distancia de tantas dimensiones como variables se tengan. EL objetivo pues del algoritmo es minimizar una función de costo que computa las distancias entre los puntos y el centroide al que pertenece, reubicando los centroides en cada iteración.

Esta función de costo esta constituida por dos partes una para las distancias numéricas y otra para las distancias categóricas como lo enuncia la ecuación 1:

$$P(W,Q) = \sum_{l=1}^{k} \left(\sum_{i=1}^{n} w_{il} \sum_{j=1}^{p} (x_{ij} - q_{lj})^2 + \gamma \sum_{i=1}^{n} w_{il} \sum_{i=p+1}^{m} \delta(x_{ij}, q_{lj}) \right) \tag{1}$$

Donde :

k = Numero de clusters establecidos

n = Cantidad de registros en el set de datos

w_{il} = variable binaria que indica si el registro i pertenece al cluster l

W = Es una matriz de dimensiones n x k que contiene las variables w_{il}

x_{ij} = valor del registro i y la variable j

q_{lj} = valor del centroide para el cluster l y la variable j

Q = Es una matriz de dimensiones k x p que contiene los centroides q_{lj}

p = Cantidad de atributos numéricos

m = Cantidad de atributos categóricos

γ = Peso que se utiliza para evitar favorecer alguno de los tipos de variables

$\delta(x_{ij}, q_{lj})$ = Función de discrepancia entre el punto x_{ij} y la moda q_{lj} que es 0 si el valor coincide y 1 si no

A partir de esta función de costo, se concluye que para cada registro que constituye el set de datos X, es necesario hacer dos tipos de cálculos diferentes para estimar la distancia del registro respecto al centroide: Uno para las variables numéricas que sera las diferencias cuadradas y otra para las variables categóricas que sera la función de discrepancia. Así pues, el objetivo del algoritmo sera encontrar los valores de las variables w_{il} y q_{lj} que minimicen esta función de costo.

El numero de clusters k se escoge mediante la implementación del método del codo, el cual permite establecer el numero de óptimo a partir de la identificación de un quiebre significativo en la pendiente de la curva que tiene en el eje x el numero de clusters y en el eje y el costo.

Una vez entrenado que el modelo haya asignado cada paciente a un cluster, se procede para cada condición, a generar una matriz de confusión en donde se aprecie la coincidencia entre el diagnostico de la condición especifica y los clusters constituidos. Esta matriz a su vez servirá para calcular medidas de validación externa. En particular se calculara el criterio de Van Dongen, que calcula la pureza de la agrupación a partir del calculo de cuan cercana esta la matriz de confusión de tener un solo elemento por fila y columna, siendo 0 en este caso 1 en el extremo opuesto, es decir, clases por completo repartidas en todos los clusters [**van2000performance**]; y además se calculara el índice de rand ajustado que mide, para dos conjuntos de clusters, que tan bien convergen las asignaciones

de estos ajustado por el azar, es decir asignaciones aleatorias siendo 0 completamente aleatorio y 1 asignaciones idénticas [**hubert1985comparing**].

Para la visualización de los clusters, puesto que se cuenta con tantas dimensiones como variables, se procede a realizar una reducción de dimensionalidad de las variables originales, buscando encontrar una proyección en dos dimensiones que expliquen un porcentaje importante de la variación de los datos. Esto se consigue a través de la implementación del algoritmo FAMD que en términos generales combina las técnicas de PCA para las variables numéricas y MCA para las variables categóricas [**pages2002analyse**]. Esta reducción de dimensionalidad se aplicara para visualizar los datos y a que cluster corresponden, así como una comparación de esta asignación, con el diagnostico para cada condición

4. Resultados

4.1. Aprendizaje supervisado

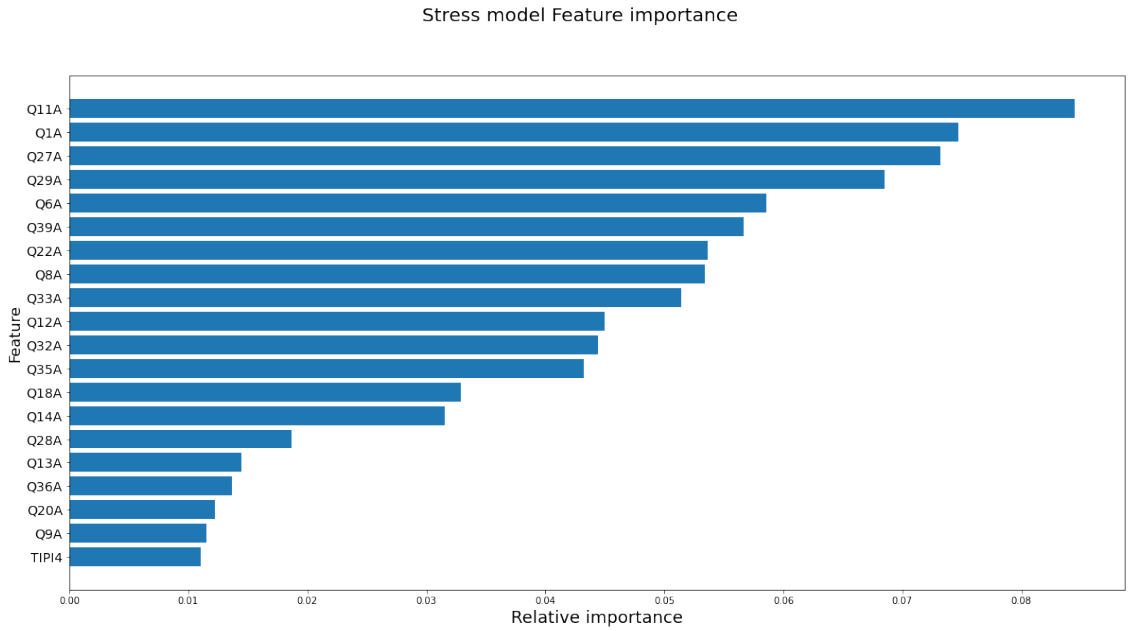
Se puede apreciar que el accuracy para todos los modelos fue superior al 80 %. Por un lado es esperable altos valores del desempeño puesto que la etiqueta del diagnostico, que es la variable objetivo, esta construida a partir de un calculo con algunas de las variables independientes, es decir, los enunciados que corresponden al diagnostico. Por otra parte, lo realmente interesante es observar la distribución de la importancia de las variables que se utilizan para predecir la clase, así como que enunciados que no pertenecen al diagnostico y otro tipo de variables, terminan siendo importantes para el modelo.

Los resultados obtenidos por el grid search para modelo entrenado con los datos de la condición de estrés, se pueden apreciar en el cuadro 7 y la importancia de las 20 variables mas importantes de dicho modelo en la figura 2.

Cuadro 7: Métricas para el modelo de estrés

n_estimators	max_depth	criterion	accuracy
500	10	entropy	0.827264

Figura 2: Importancia relativa de los atributos para el modelo de estrés



Las variables que principalmente salen a las vista son aquellas que son utilizadas para el calculo del diagnostico, sin embargo, se puede apreciar que esta importancia no es constante para todas las preguntas que constituyen el test y que por el contrario, hay una diferencia significativa en la importancia que el modelo otorga a determinadas preguntas

Hay un enunciado cuya importancia resalta que es la numero 11: "Me percibo molestándome con facilidad". Luego los enunciados 1, 27 y 29 tienen también un a importancia significativa y su valor es semejante; los enunciados son "Me percibo molestándome por cosas triviales", "Me he encontrado siendo muy irritablez "Encuentro difícil calmarme después de que algo me ha molestado". Todos estos enunciados corresponden a la categoría de estrés y se aprecia como tienen en común la facilidad y frecuencia con la que el paciente suele molestarse.

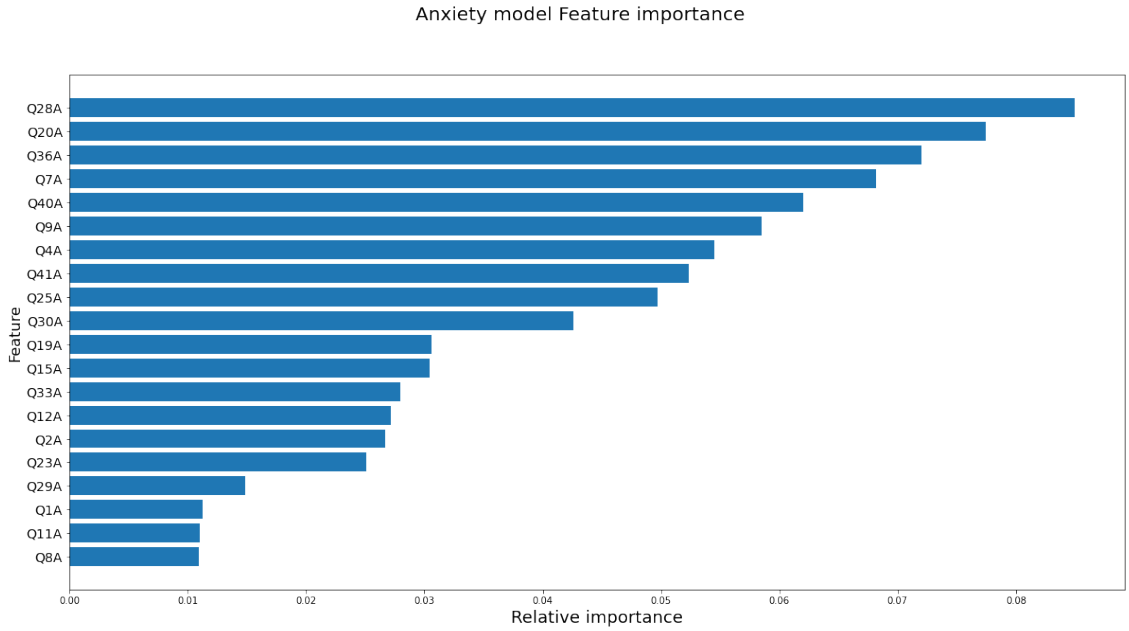
Por otro lado, hay algunos enunciados que no corresponden a la categoría de estrés y que también son considerados importantes por el modelo, estos son 28, 36, 20 y 9 que pertenecen a la categoría de ansiedad y cuyos enunciados son "Me he sentido cerca al pánico", "Me he sentido aterrorizado", "Me he sentido asustado sin ninguna razónz "Me he encontrado en situaciones que me hacen sentir tan ansioso que me he sentido casi aliviado cuando terminaron". Estos enunciados tienen en común el tema de la sensación de miedo sentido por el paciente. Además, también esta el enunciado 13 que corresponde a la categoría de depresión: "Me he sentido triste y deprimido". Finalmente, se aprecia que el enunciado 4 del TIPI también se considera importante para el modelo, cuyo enunciado es "Ansioso, molesto fácilmente", lo cual va en concordancia con los enunciados encontrados mas importantes para el modelo.

Para la condición de ansiedad, los resultados del grid search se encuentran en el cuadro 8 y la importancia delas 20 variables mas importantes esta en la figura 3

Cuadro 8: Métricas para el modelo de Ansiedad

n_estimators	max_depth	criterion	accuracy
1000	10	gini	0.817404

Figura 3: Importancia relativa de los atributos para el modelo de Ansiedad



Para este modelo, resalta que los principales enunciados son el 28, el 20, el 36 y el 7, que enuncian "Me he sentido cerca al pánico", "Me he sentido asustado sin ninguna razón", "Me he sentido aterrorizadoz "He tenido sensación de temblor". Resulta interesante que 3 de estos cuatro enunciados, es decir el 28, el 20 y el 36 también resultaron importantes para el modelo de estrés.

Se puede apreciar que hay un quiebre en la importancia de las variables a partir del enunciado 19. Aquí encontramos mezclados enunciados que corresponden a la ansiedad y algunos que corresponden al estrés. Los de el estrés son 33, 12, 29, 1, 8 que enuncian "Me he sentido en un estado de tensión nerviosa", "He sentido que he usado mucha energía", "Encuentro difícil calmarme después de que algo me ha molestado", "Me percibo molestándome por cosas triviales" y "Encuentro difícil relajarme".Es de resaltar que las variable 29 y la 1 fueron de las mas importantes para el modelo de estrés, lo

cual junto con las variables previamente mencionadas que también aparecen en el modelo de estrés, demuestra la alta correlación entre ambos observada en los datos.

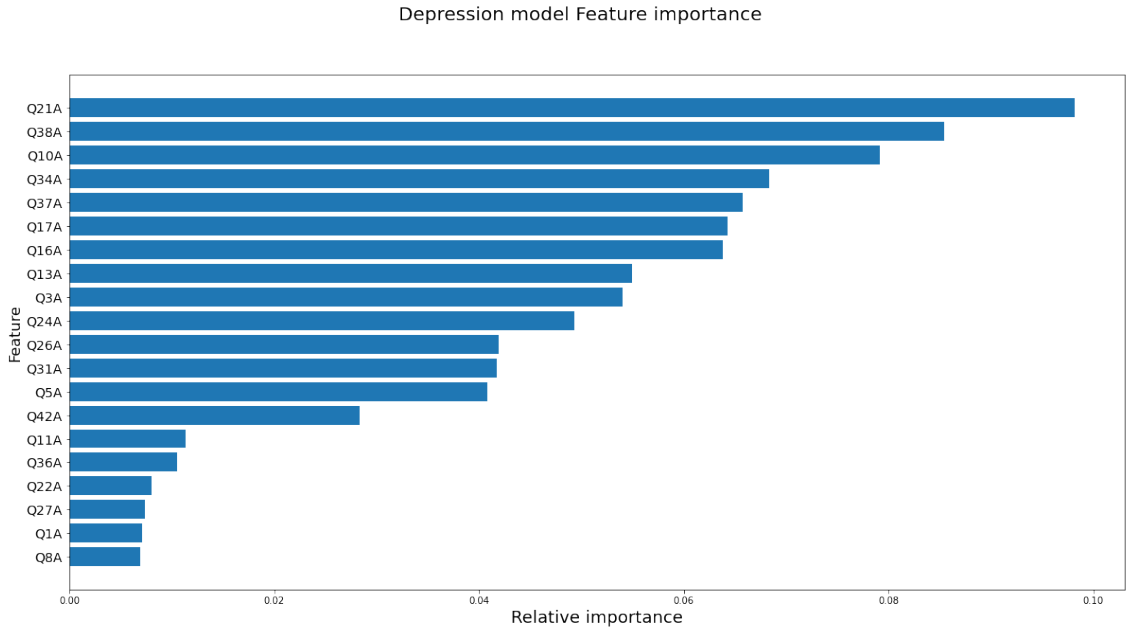
Por otro lado, los enunciados que corresponden a ansiedad y que se encuentran luego de este quiebre de importancia son 19, 15, 2, 23 que enuncian "Transpiro considerablemente en la ausencia de altas temperaturas o ejercicio fisico", "He tenido sensación de desvanecimiento", "He sido consciente de la sequedad en mi boca" y "He tenido dificultad tragando" respectivamente. Esto daría a entender que a la hora de diagnosticar esta condición, estos enunciados son igualmente importantes que los enunciados de estrés previamente mencionados.

Finalmente, para la condición de depresión, los resultados del grid search se encuentran en el cuadro 9 y la importancia delas 20 variables mas importantes esta en la figura 4

Cuadro 9: Métricas para el modelo de Depresión

n_estimators	max_depth	criterion	accuracy
1000	10	gini	0.878471

Figura 4: Importancia relativa de los atributos para el modelo de Depresión



Para el caso de la condición de depresión, resulta evidente que el enunciado mas importantes es el numero 21, que enuncia "He sentido que la vida no vale la pena". Luego los siguientes dos enunciados mas importantes son los numero 38 y 10 que enuncian "He sentido que la vida no tiene sentido", "He sentido que no tengo nada a lo que aspirar". Es interesante ver como estos tres enunciados denotan un desanimo general respecto a la vida.

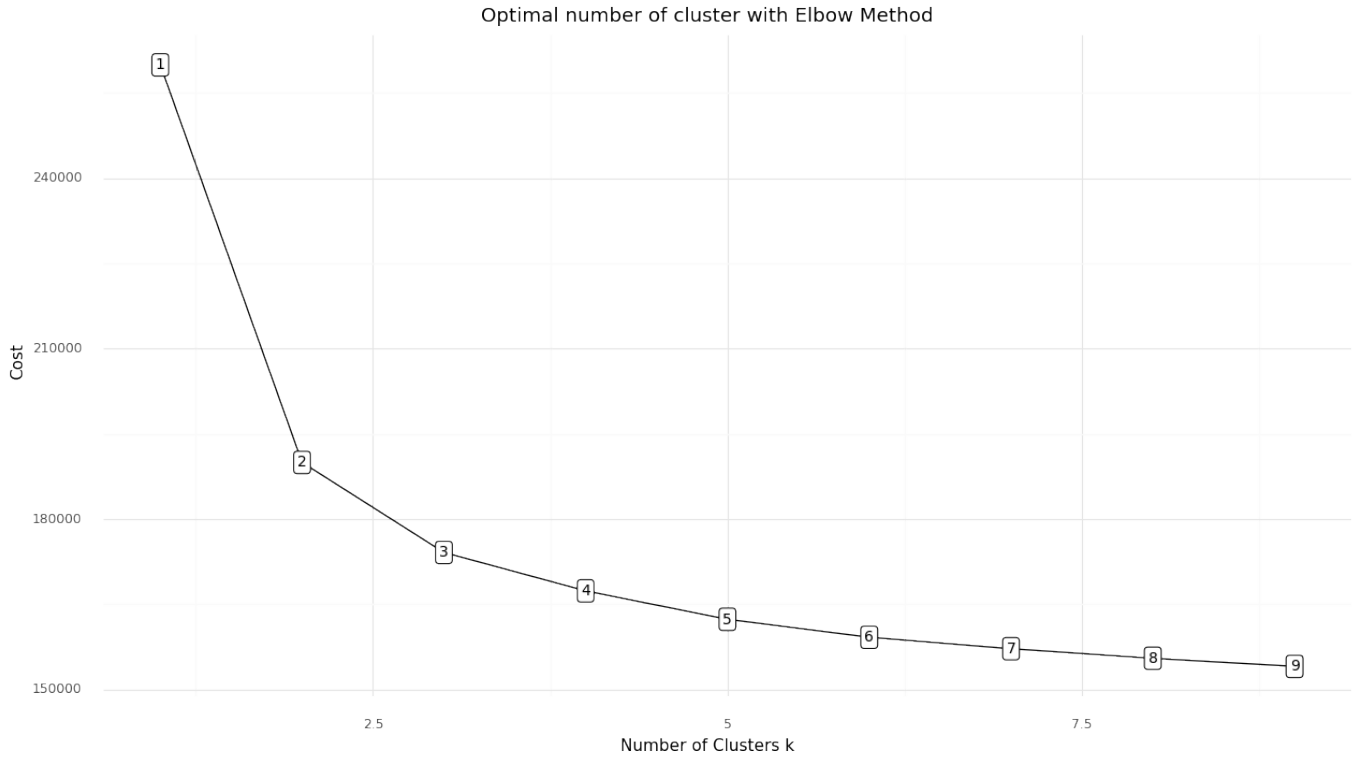
Por otro lado, dentro de estas 20 enunciados mas importantes se encuentran al final varios que tienen que ver con la condición de ansiedad; estos son los enunciados 11, 22, 27, 1 y 8 que enuncian "Me percibo molestándome con facilidad", "Encuentro difícil calmarme", , "Me he encontrado siendo muy irritable", "Me percibo molestándome por cosas triviales" y "Encuentro difícil relajarmerespectivamente. De estos 4 enunciados, tres aparecen como los mas importantes del modelo de estrés, y estos son los enunciados 11, 1 y 27. A su vez el enunciado 1 apareció también como importante para el modelo de ansiedad. Esto muestra la alta correlación que tienen la depresión y el estrés, así como el hecho de que molestarse por cosas pequeñas se encuentra como importante en las tres condiciones.

Finalmente, el enunciado numero 36 que corresponde a los enunciados de ansiedad también apareció dentro de los mas importantes y su enunciado es "Me he sentido aterrorizado", lo cual habla de la relación entre el miedo y la depresión.

4.2. Aprendizaje no supervisado

En la figura 5 se encuentran los resultados de la implementación del método del codo en donde se puede apreciar el valor de la función de costo para los distintos números de clusters

Figura 5: Costo total para diferente numero de Clusters



Se puede evidenciar un quiebre significativo de la pendiente a partir del numero 4 como cantidad de clusters, por lo que se elige esta cantidad.

Una vez entrenado el modelo y asignado un numero de cluster para cada caso, se procede calcular un perfil medio de los atributos demográficos para cada cluster cuyos resultados se encuentran en la figura **??**

Cuadro 10: Perfil medio dentro de cada cluster

Cluster	education	urban	gender	religion	orientation	race	married	TIPI1	TIPI2	TIPI3	TIPI4	TIPI5	TIPI6	TIPI7	TIPI8	TIPI9	TIPI10	age	familysize
1	2.0	2.0	2.0	10.0	1.0	10.0	1.0	3.39	4.66	4.35	6.30	4.40	5.27	5.22	4.87	2.55	4.00	21.20	3.48
2	2.0	2.0	2.0	10.0	1.0	10.0	1.0	3.25	4.23	4.38	5.39	4.71	5.18	5.12	4.53	3.18	3.84	23.22	3.29
3	3.0	2.0	2.0	10.0	1.0	10.0	1.0	4.33	3.68	5.23	3.75	5.45	4.36	5.36	3.65	4.95	3.45	25.97	3.64
4	3.0	2.0	2.0	10.0	1.0	10.0	1.0	3.99	4.28	4.86	5.47	5.04	4.73	5.36	4.23	3.65	3.71	22.83	3.53

Haciendo un paralelo con los perfiles encontrados previamente, se puede observar que el nivel educativo es mayor para los clusters 3 y 4. Así mismo la edad es mayor para el cluster 3 y menor para el cluster 1. En cuanto al TIPI, los enunciados 1, 3, 5, y 9 tienden a ser mayores para el cluster 3 y menores para el cluster 1 y 4 siendo menores para los clusters 1 y 2; y los enunciados 2, 4, 6 y 8 son mayores para los clusters 1 y 2 y menores en los clusters 3 y 4. Esto invitaría a pensar que se esperan diagnósticos de severidad mayores principalmente en el cluster 1 y luego en el 2, encontrando las menores severidades en el cluster 3 y luego en el cluster 4

En la figura 11 se puede apreciar para cada cluster, cual fue la condición mas común para cada una de las condiciones. Se evidencia que el cluster 3 presenta principalmente diagnósticos normales, el 4 condiciones moderadas, el 2 una mezcla de distintos desde moderado a extremadamente severo y el cluster 1 son principalmente diagnósticos extremadamente severos. Esto es congruente con la información contenida en los perfiles demográficos.

Cuadro 11: Frecuencia máxima de diagnósticos por condición y cluster

Cluster	stress diagnosis	anxiety diagnosis	depression diagnosis
1	Extremely severe	Extremely severe	Extremely severe
2	Moderate	Severe	Extremely severe
3	Normal	Normal	Normal
4	Moderate	Moderate	Moderate

En el cuadro 12 se aprecia que la mayor cantidad de diagnósticos para el cluster 1 son severo y extremadamente severo, para el cluster 2 son moderado y severo y en menor medida leve y normal, para el cluster 3 son normal y leve y para el cluster 4 mayormente moderado, seguido de leve y severo y en menor medida normal.

Cuadro 12: Matriz de confusión para estrés

Clusters Stress_Diagnosis	1	2	3	4
Normal	3	907	9743	1143
Mild	9	1114	1090	2708
Moderate	402	3442	212	4671
Severe	3745	2717	6	2102
Extremely severe	5214	249	0	282

En el cuadro 13 se aprecian los valores de los índices de Van Dogen y y Rand ajustado, cuyos valores son 0.54 y 0.37 respectivamente, lo cual indica que aun cuando hay cierta afinidad entre los clusters creados y los diagnósticos para esta condición, estos no terminan de ser asignados de una manera completamente coincidente, en parte por que el numero de clusters es diferente al numero de condiciones y además hay cierta aleatoriedad en las asignaciones.

Cuadro 13: Métricas de los clusters de estrés

Van Dongen Criterion	Adjusted Rand Score
0.54	0.37

En las figuras 6, 7 y 8 se muestran la comparación entre los clusters constituidos a la izquierda y el diagnostico de cada condición a la derecha, en una representación de dos dimensiones a través de la implementación de FAMD, en donde se puede apreciar la inercia, es decir, cuanta varianza es capaz de capturar, en cada eje.

La figura 6 muestra como el eje x esta capturando la severidad de la condición de estrés, pues mientras los datos se ubican mas a la derecha, la severidad aumenta en la imagen de la derecha. En ese sentido, se aprecia que las condición normal esta claramente a la izquierda y la condición extremadamente severo bastante a la derecha, quedando así un solapamiento entre las condiciones severa, moderada y media, estando la condición moderada en el centro. Por otro lado, la imagen de la izquierda nos

muestra que los miembros del cluster 3 se ubican a la izquierda mientras que los miembros del cluster 1 se ubican a la derecha, quedando así los miembros del cluster 24 en la mitad con un poco de solapamiento entre ellos, estando el 4 a la izquierda y el 2 a la derecha. Esto es permite apreciar de una manera gráfica, cuan semejante es la distribución de las asignaciones previamente discutidas en el cuadro 12

Figura 6: Análisis de clusters para el diagnostico de estrés

En el cuadro 14 vemos como al cluster 1 son asignados principalmente casos con diagnostico extremadamente severo y en menor medida severo; el cluster dos tiene en proporcione similares diagnósticos moderados, severo y extremadamente severo, para luego tener mas presencia de de diagnostico normal que leve. El cluster 3 presenta principalmente diagnósticos normales y en una menor medida, diagnósticos leve, moderado y finalmente para el cluster 4 se presenten principalmente diagnósticos moderado y severo. Cabe resaltar la semejanza a las asignaciones que se dan para la condición de ansiedad con respecto a la condición de estrés, principalmente para los clusters 1 y 3 que presentan en su mayoría condiciones extremadamente severa y normales respectivamente, existiendo variaciones mas pronunciadas en los clusters 4 y 4 que tienen las condiciones intermedias. Esta semejanza puede verse así mismo en el cuadro 15 ya que se tienen valores similares a los de la condición de estrés.

Cuadro 14: Matriz de confusión para ansiedad

Clusters Anxiety_Diagnosis	1	2	3	4
Normal	0	1031	7928	766
Mild	0	549	1380	834
Moderate	10	2216	1519	3302
Severe	228	2561	213	3109
Extremely severe	9135	2072	11	2895

En la figura 7 puede verse que al igual que para la condición de estrés, para ansiedad el eje x esta capturando la severidad de la condición estando mas a la derecha condiciones mas severas. En ese sentido, si se compara con los clusters encontrados, se observa también que el cluster 3 al estar mas a la izquierda coincide con casos cuyos diagnósticos son menos severos, teniendo en el otro extremo

Cuadro 15: Métricas de los clusters de ansiedad

Van Dongen Criterion	Adjusted Rand Score
0.58	0.34

al cluster 1, tal con mayor presencia de severidad en los diagnósticos. Se aprecia así mismo que los casos de ansiedad severa están mas dispersos hacia el centro superponiendo se con mas frecuencia a casos de diagnósticos severa y moderada, lo cual tendría que ver con el hecho de que para esta condición hay muchos mas casos con este diagnostico. Esto también tendría que ver con observar casos con diagnósticos de severidad extrema presentes hacia el centro de la gráfica, que corresponden a los clusters 2 y 4. Por otro lado, es también interesante el apreciar como los diagnósticos moderado y severa presentan una gran dispersión a lo lardo de la figura, lo cual coincide con su presencia en distintos clusters como lo muestra la matriz de confusión.

Figura 7: Análisis de clusters para el diagnostico de ansiedad

En el cuadro 16 se evidencia que así como para las condiciones anteriores, los clusters 1 y 3 tienen la mayor cantidad de casos con diagnósticos extremadamente severo y normal respectivamente. Lo particular para esta condición es que el cluster 2 presenta únicamente casos con diagnósticos extremadamente severo y severo con excepción de unos pocos casos con diagnostico moderado. Finalmente, el cluster 4 presenta en su mayoría casos con diagnostico moderado y un numero considerable de los demás diagnósticos con excepción de extremadamente severo. El cuadro 17 permite apreciar que las métricas para este modelo tienen valores similares a las condiciones anteriores pero son un poco menores.

La figura 8 muestra que al igual que para las dos condiciones previas, el eje x coincide con la severidad de la condición. Al igual que para las previas condiciones, los diagnósticos leve, moderado y severo se superponen en el centro de la gráfica, estando particularmente presentes en el cluster 4 y en menor medida en el cluster 2. Así mismo, al igual que para la condición de ansiedad, se observa que la condición de extremadamente severo se encuentra dispersa hacia en centro del gráfico, lo cual se explica, al igual que para la condición de ansiedad, por una mayor presencia de casos con este diagnostico, estando de esta manera presente en los clusters 1 y 2.

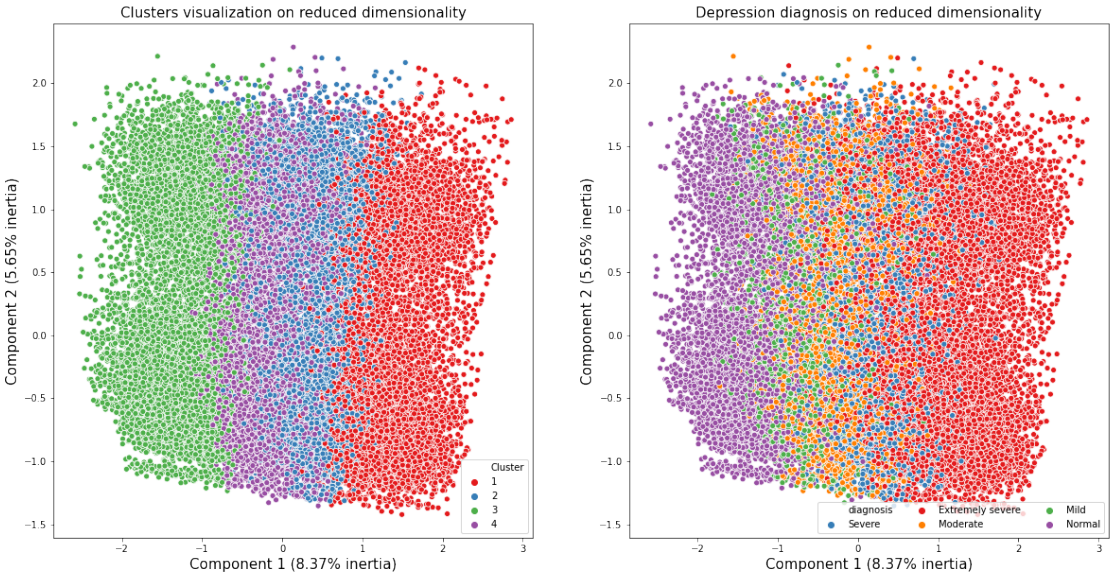
Cuadro 16: Matriz de confusión para depresión

Clusters Depression_Diagnosis	1	2	3	4
Normal	0	0	7655	1198
Mild	13	0	1836	1937
Moderate	209	7	1278	5583
Severe	1218	2803	267	2188
Extremely severe	7933	5619	15	0

Cuadro 17: Métricas de los clusters de depresión

Van Dongen Criterion	Adjusted Rand Score
0.49	0.4

Figura 8: Análisis de clusters para el diagnostico de depresión



5. Conclusiones

Se evidencio que existe una correlación significativa entre las tres condiciones, estando principalmente el estrés ligado a la ansiedad y a la depresión y en una menor medida la ansiedad y la depresión siendo sin embargo una correlación significativa. De igual manera se evidencio que para la presente muestra, existen diagnósticos de una severidad mayor para la ansiedad y la depresión que para el estrés. Se evidencio así mismo que las personas con un mayor nivel educativo y mayor edad tienden a tener diagnósticos menos severos. De igual manera los enunciados 1, 3, 5, y 9 del TIPI que están asociados a características positivas de la personalidad como entusiasmo, disciplina, apertura a nuevas experiencias y calma tienden a ser mayores en personas con diagnósticos leves, así como los enunciados 2, 4, 6 y 8 que están asociados con características negativas de la personalidad como conflictividad, irritabilidad, reserva y desorden tienden a aumentar con el diagnostico.

Los random forest entrenados para cada condición, encontraron que existe diferente poder predictivo en los enunciados que constituyen el test para cada condición, siendo particularmente importantes

unos cuatro enunciados para cada condición a la hora de realizar el diagnostico. Para el caso del estrés estos enunciados son 11,1, 27 y 29 que enuncian "Me percibo molestándome con facilidad", "Me percibo molestándome por cosas triviales", "Me he encontrado siendo muy irritable ", .Encuentro difícil calmarme después de que algo me ha molestado" respectivamente; para la ansiedad estos enunciados son 28, el 20, el 36 y el 7, que enuncian "Me he sentido cerca al pánico", "Me he sentido asustado sin ninguna razón", "Me he sentido aterrorizadoz "He tenido sensación de temblorrespectivamente y para el caso de la depresión son 21, 38 y 10 que enuncian "He sentido que la vida no vale la pena", "He sentido que la vida no tiene sentido" y "He sentido que no tengo nada a lo que aspirar". Así mismo, dentro de las variables importantes del modelo de cada una de las condiciones aparecieron enunciados correspondientes a otras condiciones, particularmente enunciados de estrés en las otras dos y de ansiedad y depresión en el de estrés lo que corresponde con la correlación encontrada en los diagnósticos.

Se encontró que el numero óptimo de clusters para los datos disponibles es 4. Al encontrar el diagnostico mas frecuente para cada cluster se aprecia que en el cluster 3 son mas frecuentes los diagnósticos normales, el cluster 1 los diagnósticos extremadamente severo, quedando en los cluster 2 y 4 una mezcla de los diagnósticos intermedios. Esto coincide así mismo con los datos demográficos medios dentro de cada uno de estos clusters con aquellos, pues poseen atributos similares a los observados en los perfiles medio de los diagnósticos y condiciones correspondientes.

Al representar los datos en una reducción de dos dimensiones y comparar las etiquetas de los diagnósticos de cada condición y los clusters encontrados, se aprecia que el principal componente, es decir el eje de las x, coincide con la severidad de los diagnósticos en las tres condiciones y es clara el solapamiento entre el cluster 3 y el diagnostico normal y el cluster 1 y el diagnostico extremadamente severo. Por otro lado los cluster 2 y 4 presentan una cantidad considerable de múltiples diagnósticos y esta distribución cambia según la condición. Esto se refleja en las métricas de Rand y Van Dongen que evidencian que la distribución de los clusters, si bien coincide de alguna manera con los diagnósticos, no refleja completamente los mismos.

A. Anexos

Cuadro 18: Atributos demográficos incluidos en la encuesta

Field	Description
TIPI1	Extraverted, enthusiastic.
TIPI2	Critical, quarrelsome.
TIPI3	Dependable, self-disciplined.
TIPI4	Anxious, easily upset.
TIPI5	Open to new experiences, complex.
TIPI6	Reserved, quiet.
TIPI7	Sympathetic, warm.
TIPI8	Disorganized, careless.
TIPI9	Calm, emotionally stable.
TIPI10	Conventional, uncreative.
education	How much education have you completed?, 1=Less than high school, 2=High school, 3=University degree, 4=Graduate degree
urban	What type of area did you live when you were a child?, 1=Rural (country side), 2=Suburban, 3=Urban (town, city)
gender	What is your gender?, 1=Male, 2=Female, 3=Other
age	How many years old are you?
religion	What is your religion?, 1=Agnostic, 2=Atheist, 3=Buddhist, 4=Christian (Catholic), 5=Christian (Mormon), 6=Christian (Protestant), 7=Christian (Other), 8=Hindu, 9=Jewish, 10=Muslim, 11=Sikh, 12=Other
orientation	What is your sexual orientation?, 1=Heterosexual, 2=Bisexual, 3=Homosexual, 4=Asexual, 5=Other
race	What is your race?, 10=Asian, 20=Arab, 30=Black, 40=Indigenous Australian, 50=Native American, 60=White, 70=Other
married	What is your marital status?, 1=Never married, 2=Currently married, 3=Previously married
familysize	Including you, how many children did your mother have?

Cuadro 19: Enunciados del test

Field	Description
Q1	I found myself getting upset by quite trivial things.
Q2	I was aware of dryness of my mouth.
Q3	I couldnt seem to experience any positive feeling at all.
Q4	I experienced breathing difficulty (eg, excessively rapid breathing, breathlessness in the absence of physical exertion).
Q5	I just couldnt seem to get going.
Q6	I tended to over-react to situations.
Q7	I had a feeling of shakiness (eg, legs going to give way).
Q8	I found it difficult to relax.
Q9	I found myself in situations that made me so anxious I was most relieved when they ended.
Q10	I felt that I had nothing to look forward to.
Q11	I found myself getting upset rather easily.
Q12	I felt that I was using a lot of nervous energy.
Q13	I felt sad and depressed.
Q14	I found myself getting impatient when I was delayed in any way (eg, elevators, traffic lights, being kept waiting).
Q15	I had a feeling of faintness.
Q16	I felt that I had lost interest in just about everything.
Q17	I felt I wasnt worth much as a person.
Q18	I felt that I was rather touchy.
Q19	I perspired noticeably (eg, hands sweaty) in the absence of high temperatures or physical exertion.
Q20	I felt scared without any good reason.
Q21	I felt that life wasnt worthwhile.
Q22	I found it hard to wind down.
Q23	I had difficulty in swallowing.
Q24	I couldnt seem to get any enjoyment out of the things I did.
Q25	I was aware of the action of my heart in the absence of physical exertion (eg, sense of heart rate increase, heart missing a beat).
Q26	I felt down-hearted and blue.
Q27	I found that I was very irritable.
Q28	I felt I was close to panic.
Q29	I found it hard to calm down after something upset me.
Q30	I feared that I would be thrown by some trivial but unfamiliar task.
Q31	I was unable to become enthusiastic about anything.
Q32	I found it difficult to tolerate interruptions to what I was doing.
Q33	I was in a state of nervous tension.
Q34	I felt I was pretty worthless.
Q35	I was intolerant of anything that kept me from getting on with what I was doing.
Q36	I felt terrified.
Q37	I could see nothing in the future to be hopeful about.
Q38	I felt that life was meaningless.
Q39	I found myself getting agitated.
Q40	I was worried about situations in which I might panic and make a fool of myself.
Q41	I experienced trembling (eg, in the hands).
Q42	I found it difficult to work up the initiative to do things.

Cuadro 20: Set de preguntas para cada condición

	Questions
Depression	3, 5, 10, 13, 16, 17, 21, 24, 26, 31, 34, 37, 38, 42
Anxiety	2, 4, 7, 9, 15, 19, 20, 23, 25, 28, 30, 36, 40, 41
Stress	1, 6, 8, 11, 12, 14, 18, 22, 27, 29, 32, 33, 35, 39

Cuadro 21: Puntajes mínimos para cada diagnostico en cada condición

	Depression	Anxiety	Stress
Normal	0	0	0
Mild	10	8	15
Moderate	14	10	19
Severe	21	15	26
Extremely severe	28	20	34