Evaluación del efecto del tratamiento psicológico y farmacológico en el contexto de pandemia

3 de octubre de 2022

Universidad Internacional de Valencia

Titulación:

Master en Big Data & Data Analysis

Curso académico

2021-2022

Alumno/a: Catalán Torralbo, Sergio

D.N.I: 79094381A

Directora de TFM: Dra. Vanessa Moscardó

Convocatoria: Tercera

De:
Planeta Formación y Universidades

Universidad Internacional de Valencia

Índice

Índice de ilustraciones

Índice de tablas

Resumen

En este proyecto se ha estudiado el impacto de la pandemia sobre la población española y su salud mental, a través de una caracterización de los perfiles de los sujetos de estudio en función de los tipos de tratamiento a los que estuvieron sometidos, psicológico y/o farmacológico, y a cuando estuvieron sometidos a ellos, antes o después de la pandemia Covid-19. Para ello se hace uso de una encuesta realizada por el CIS en febrero de 2021, en la que se entrevistaron a 3083 personas, donde se obtuvieron para cada una de ellas 354 variables, que representaban preguntas realizadas por los entrevistadores o datos sobre las entrevistas. La mayoría de las preguntas de esta encuesta que resultan útiles para nuestro objetivo son aquellas de carácter sociodemográfico, como el sexo o el nivel de estudios, aunque otras como el consumo de fármacos por parte de familiares o la muerte de familiares a causa del Covid entre otras también nos resultan relevantes.

Para lograr esta caracterización de perfiles de tratamiento, se realizó un proceso KDD completo, donde comenzamos con un preprocesado de datos, consistente en tratar los valores nulos o vacíos, eliminar las variables no relevantes para el caso de estudio y buscar respuestas a preguntas muy desbalanceadas entre otras técnicas. Después definimos las variables objetivo, que representan los tipos de tratamientos que caracterizamos, para luego comenzar con el estudio estadístico propiamente dicho.

En este estudio se aplicaron técnicas estadísticas, como contrastes de hipótesis, típicos de la estadística inferencial, con el objeto de probar dependencias entre los tipos de tratamientos y las preguntas realizadas por los entrevistadores. También se realizó una búsqueda de similitudes entre distintos perfiles de tratamientos y una selecciones de variables para cada tratamiento. Todo este proceso se llevó a cabo a través de métodos de programación y técnicas de inteligencia artificial, sobre el lenguaje de programación de código abierto Python, que facilitaron la obtención de resultados, que más adelante analizamos.

En este trabajo encontramos importantes resultados, como un cambio relevante en la distribución de las edades para las personas que consumían fármacos por problemas de salud mental antes de la pandemia y para aquellas que comenzaron a consumirlos después de ella. También se hallaron diferencias significativas en el consumo de fármacos por problemas de salud mental entre mujeres y hombres, donde las primeras son más propensas a su consumo. Además se observó una mayor tendencia entre los sanitarios a la recepción de tratamiento psicológico, tanto antes como después de la pandemia, señalando la vulnerabilidad de este colectivo ante los problemas de salud mental.

Palabras clave: Covid-19, Tratamiento farmacológico, Tratamiento Psicológico, Pandemia, Estadística Inferencial

Abstract

In this project, the impact of the pandemic on the Spanish population has been studied, through a characterization of the profiles of the study subjects based on the types of treatment to which they were subjected, psychological and/or pharmacological, and when were subjected to them, before or after the Covid-19 pandemic. To do this, a survey carried out by the CIS in February 2021 is used, in which 3083 people were interviewed, where 354 variables were obtained for each of them, which represented questions asked by the interviewers or data of the interviews. Most of the questions in this survey that are useful for our objective are those of a sociodemographic nature, such as gender or level of education, although others such as the consumption of drugs by family members, deaths of family members due to Covid, among others are also relevant to us.

To achieve this characterization of treatment profiles, a complete KDD process was carried out, where we began with a data pre-processing, consisting of treating null or empty values, eliminating variables not relevant to the case study, and looking for answers to very unbalanced questions among other techniques. Then we define the objective variables, which represent the types of treatments that we are going to characterize, and then begin with the statistical study itself.

In this study, statistical techniques were applied, such as hypothesis tests, typical of inferential statistics, in order to test dependencies between the types of treatments and the questions asked by the interviewers. A search for similarities between different treatment profiles and a selection of variables for each treatment was also carried out. All these processes were done through programming methods and artificial intelligence techniques, on the open-source programming language, Python, which facilitated the obtaining of results, which we will analyse later.

In this work we find important results, such as a relevant change in the distribution of ages for people who used drugs for mental health problems before the pandemic and for those who began to use them after it. Differences were also found in the use of drugs for mental health problems between women and men, where the women are more likely to use them. Spanish population to be able to extrapolate these conclusions to a national panorama. A greater tendency was also observed among health workers to receive psychological treatment, both before and after the pandemic, indicating the vulnerability of this group to mental health problems.

Keywords: Covid-19, Pharmacological treatment, Psychological treatment, Pandemic, Inferential Statistics

1. Introducción

La pandemia ocasionada por la enfermedad Coronavirus 2019 (Covid–19) alteró significativamente la vida de una gran parte de la población mundial, afectando enormemente a la salud, tanto mental como física de las personas. Las cifras aportadas por la OMS hablan por sí solas, más de 6,5 millones de fallecidos a causa del Covid, y más de 600 millones de casos en todo el mundo (Organización Mundial de la Salud [OMS], 2022). Periodos de confinamiento prolongado para evitar el esparcimiento de la enfermedad, la incertidumbre antes una situación desconocida para la mayoría, contracción de la enfermedad por parte de familiares amigos o conocidos, y en algunos casos la muerte de estos, son algunos de los muchos motivos que influyeron a la hora de generar este enorme impacto en salud mental, y que aún se mantiene en buen parte de la población. Algunas de las causas más inmediatas de la pandemia son aumento del aislamiento social, y la soledad en los individuos, Holmes et al. (2020), que van fuertemente asociados con la ansiedad, la depresión, las autolesiones y el suicidio, Elovainio et al. (2017), pero van mucho más allá, también ha favorecido a la aparición, entre otros, de trastornos alimenticios, Rodgers et al. (2020) o "burn-out" entre los trabajadores de la salud mental, Hongjin et al. (2022).

Existen muchas consecuencias derivadas de la pandemia, por lo que resulta interesante realizar un estudio a nivel de España con el objeto de caracterizar cuales son aquellos factores más influyentes sobre la salud mental de la población. De manera más concreta, en nuestro trabajo analizamos cuáles son los factores que más influyen a la hora de someterse a tratamientos de tipo psicológico y farmacológico a causa de problemas de salud mental, y como los perfiles de las personas que se someten a estos han variado antes y después de la pandemia. Esto nos aportará una visión diferente sobre cuáles han sido los colectivos más vulnerables a esta situación epidemiológica, y como han variado según los tipos de tratamientos recibidos.

Para poder realizar el estudio hemos hecho uso de una encuesta realizada por el CIS en febrero de 2021, nombrada *"Salud mental de los españoles durante la pandemia de la Covid-19"*. El CIS, Centro de Investigaciones Sociológicas, es un organismo autónomo del Ministerio de la Presidencia, cuya función principal es contribuir al conocimiento científico de la población española, y que se encarga de recoger datos de diversas fuentes y ámbitos, a través de encuestas en general, para ponerla después a disposición de los ciudadanos (son de dominio público) y apoyar así a la investigación en ciencias sociales. Esta encuesta en concreto se realizó a 3083 personas, y nosotros la usaremos para hacer un estudio profundo de los perfiles asociados a los distintos tipos de tratamientos a los que se han sometido los entrevistados.

En el proyecto hemos comenzado primero definiendo los objetivos específicos que nos hemos planteado, para luego realizar un resumen de los proyectos de temáticas similares al nuestro y una introducción teórica de todos aquellos conceptos que son necesarios conocer, aunque sea de manera superficial, para poder comprender el desarrollo del trabajo. Después seguimos avanzando con el desarrollo del proyecto *per-*

se, que consistió en un proceso KDD completo, y que describimos con todo detalle, desde la extracción de la información y su procesado, hasta el análisis de resultados mediante técnicas estadísticas y de inteligencia artificial, y por último las conclusiones obtenidas de este proyecto. La parte más gruesa del trabajo está realizada en el proceso de limpieza de datos y su estudio estadístico mediante técnicas de programación, debido al gran número de variables que poseíamos, y las dificultades que conlleva por tanto analizarlos correctamente.

2. Objetivos

El principal objetivo de este trabajo es caracterizar los perfiles afines a los distintos grados de tratamiento al que estaban sometidos los encuestados. Estos distintos grados de tratamiento y las variables objetivo asociadas a cada tipo en nuestros datos se han definido teniendo en cuenta si el tratamiento recibido ha sido farmacológico, psicológico o ambos, y también si se ha sometido a este antes, después o antes y después de la pandemia.

Estos perfiles, con los distintos niveles de tratamientos farmacológicos y/o psicológicos, se caracterizan en función de variables de interés obtenidas del dataset de la encuesta, correspondiéndose cada variable con una pregunta realizada por el entrevistador. Estas variables (preguntas de encuesta), pueden ser de tipo socio-demográfico u otros, y en su mayoría, son tratadas para hacer más eficaz este análisis, tal y como se describirá más adelante.

Los objetivos específicos de este trabajo son:

- **Revisar la bibliográfica sobre trabajos de temáticas similares**
  Comprobar los resultados obtenidos por estudios en el mismo campo o similares, ya sea caracterización de perfiles en el contexto de la salud mental o el análisis estadístico de los efectos de la pandemia sobre la población y su salud mental y tratamientos recibidos en consecuencias.

- **Comprender el dataset y realizar un tratamiento de datos para poder llevar a cabo un análisis correcto.**
  Este proceso abarcó desde la extracción de los datos aportados por el CIS hasta la limpieza y transformación de estos, mediante el tratamiento de missing values, el agrupamiento de diferentes respuestas a las preguntas en menos categorías, creación de nuevas variables de interés a partir de las presentes y comprobación de las correlaciones entre variables.

- **Seleccionar las variables de interés.**
  Como se explicará con más detalle más adelante, esta selección se hizo en base a diferentes criterios, como la selección de aquellas relevantes para el marco en el que se engloba el proyecto o la calidad de la información aportada por estas, ya que, por ejemplo, si la cantidad de valores nulos era demasiado elevada, o había una distribución de valores muy pobre (muchos registros con un mismo valor, y muy pocos o ninguno con valores diferentes), esta característica es descartada por su poca influencia. También los valores de la correlación entre características fue un factor para eliminar algunas, ya que dos variables muy correlacionadas entre si no nos aportan información relevante, y pueden suponer un peor rendimiento en modelos que queramos aplicar a nuestros datos.

- **Crear las variables objetivo que definan los distintos niveles de tratamiento recibidos.**
  Para ello usamos algunas de las características ya presentes en el dataset, y a partir de ellas construimos las nuevas, generando una variable por cada nivel de tratamiento que queremos caracterizar.

- **Caracterizar de los perfiles afines a los distintos grados de tratamiento manifestado.**
  Se hará uso de dos tests de independencia (contraste de hipótesis) para encontrar las relevancias de las variables presentes en la encuesta respecto a los distintos grados de tratamiento definidos en las variables objetivo.

- **Evaluar las diferencias significativas existentes entre los distintos grados de tratamiento teniendo en cuenta las variables de interés.**
  Comprobaremos como difieren entre sí los perfiles de los encuestados que hemos obtenido para cada tipo de tratamiento mediante la búsqueda de similitudes entre ellos.

- **Obtener la relevancia de las características seleccionadas del dataset para cada una de las variables objetivo definidas.**
  Para ello usaremos una selección de características realizada por un algoritmo de Machine Learning, que devuelve las relevancias de cada una respecto a la variable objetivo que se está estudiando.

- **Extraer conclusiones a partir de los resultados obtenidos.**

3. Estado del Arte y Marco teórico

3.1. Estado del Arte

En el presente apartado se lleva a cabo una revisión sobre proyectos con temáticas similares al que estamos abordando. Se buscaran trabajos enfocados en la aplicación de técnicas estadísticas o de inteligencia artificial para buscar relaciones entre las características sociodemográficas de la población y los problemas de salud mental asociados a la pandemia y al consiguiente confinamiento, y en caso de que los hubiera, las relaciones también con los tratamientos psicológicos o farmacológicos a causa de la pandemia.

Investigaciones consultadas

El primer estudio encontrado relacionado con los objetivos de nuestro proyecto fue desarrollado por Wang et al. (2020) . En él se realizó un estudio estadístico para calcular asociaciones entre variables sociodemográficas y de otros tipos en relación con los niveles de afecciones a la salud mental que había sufrido la población china a causa de la pandemia. Sin embargo en este estudio se aplicaron únicamente regresiones lineales para calcular estas relaciones, mientras que en nuestro proyecto hemos aplicado otras técnicas como la estadística inferencial o Machine Learning.

Otro artículo enfocado en el mismo concepto, pero sobre un brote de SARS que condujo a un confinamiento en Toronto, Canadá, es el desarrollado por Hawryluck et al. (2004), donde a través de una encuesta realizada sobre la población que se vio obligada a someterse a una cuarentena a causa del SARS, se obtuvieron datos sobre la salud mental de estas y sobre aspectos sociodemográficos relevantes de cada una. Se aplicaron análisis de variancia (ANOVA) y tests de independencia como el Chi-Square para determinar estas relaciones entre las características demográficas y los estados de salud mental. Es en la aplicación de estas técnicas estadísticas sobre la relación de la salud mental con las variables demográficas donde encontramos gran similitud con nuestros procedimientos en el proyecto.

Estos dos estudios fueron los más cercanos a abordar los objetivos de este trabajo, sin embargo comprobamos como ninguno abarca la influencia de las características sociodemográficas u otros tipos para relacionarlas con la tipología de tratamientos (psicológico o farmacológico) a los que se sometió la población a causa de la epidemia Covid-19, por lo que nosotros abarcaremos una temática que parece aún sin explorar.

3.2. Marco teórico

En esta sección vamos a sentar las bases teóricas que nos ayudaran a comprender el contexto el trabajo. Definiremos algunos de los conceptos claves para poder entender los métodos aplicados para la extracción de información relevante.

En primer lugar, comenzaremos definiendo lo que es el proceso KDD, ya que es el enfoque usado para la realización de todo el trabajo, y luego entraremos en detalle sobre el análisis estadístico, los métodos de búsquedas de similitudes, y por último, la selección de características.

3.2.1. Proceso KDD

El proceso de KDD tiene diferentes definiciones dependiendo de la fuente consultada, sin embargo, puede resumirse en general en el proceso no trivial de encontrar patrones en los datos que sean válidos, novedosos, útiles y finalmente entendibles (Fayyad, Paitetsky-Shapiro y Smyth, 1996a). El principal problema abordado por los procesos KDD es transformar la información en sus formatos de origen, que suelen ser complejos de interpretar, a formar más entendibles y analizables con el objeto de sacar conclusiones. En núcleo central de este proceso es la aplicación de técnicas de minería de datos (data mining) para el descubrimiento y extracción de los patrones de datos, aplicando herramientas de estadística, inteligencia artificial, Machine Learning y otras ramas, aunque en sí el proceso KDD abarca muchas más fases.

A continuación, en la Ilustración 1, se observa la estructura típica que conforma a estos procesos, aunque en ocasiones encontraremos bibliografía que añada algún paso más o elimine otros.

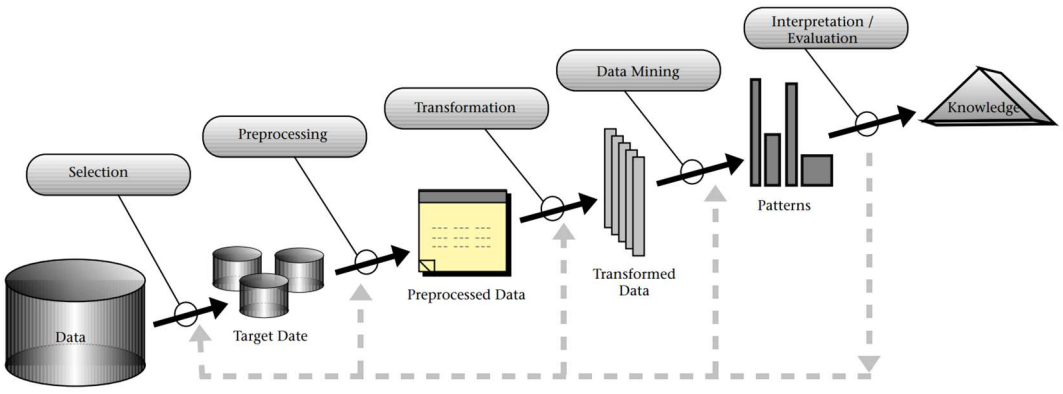


Ilustración 1. Visión genera pasos que componen el proceso KDD. Fuente: Fayyad, Paitetsky-Shapiro y Smyth, 1996b, p.5

El origen del data mining y el proceso KDD puede situarse en la década de 1980, cuando una gran cantidad de compañías comenzaron a almacenar información transaccional, generando cantidades de datos que no podían ser analizadas por los métodos tradicionales. Ante este problema y el avance de la inteligencia artificial, se comenzó a aplicar esta, junto con la estadística a problemas de extracción de conocimiento, generando el campo del KDD. Una de las primeras aplicaciones realmente exitosas del KDD y el data mining fue la detección de fraude de tarjetas de crédito, a través del estudio del comportamiento de los consumidores, demostrando así el potencial de esta tecnología, Clifton (2022).

Si dividimos el proceso KDD tal y como aparece en la Ilustración 1, tendremos las siguientes fases:

Selección/Extracción:

Se basa en extraer de las fuentes de datos aquella información que necesitamos para afrontar nuestro problema y extraer conocimientos.

Preprocesamiento:

Fundamentado en la "limpieza" de los datos, es decir, en tratarlos de manera que los modelos y técnicas aplicadas en la sección de data mining resulten efectivas. La eliminación o tratamiento de missing values, el estudio de los outliers y la selección de las características relevantes podrían ser ejemplos de preprocesamientos que podemos aplicar a nuestros datos, aunque existen muchos otros.

Transformaciones:

Consiste en la transformación de los datos que tenemos en otros más útiles, ya sea combinando diferentes variables que tengamos ya, escalando los datos para que los algoritmos no den importancia solo a los más altos (normalización), agrupando variables para reducir la dimensionalidad, etc. Mediante este proceso presentamos nuestros datos en una forma óptima para su uso.

Data Mining:

Aunque algunas fuentes definen el Data Mining como un sinónimo de proceso KDD, nosotros lo definiremos como el proceso de descubrimiento de patrones en los datos ya procesados, a través de herramientas de estadística e inteligencia artificial. Un ejemplo de data mining podría ser el uso de un árbol de decisión para obtener cuales son las características más relevantes cuando estamos abordando un problema de clasificación, o la aplicación de estadística inferencial para determinar propiedades de los datos analizados.

Evaluación de resultados/Interpretación:

En este paso se interpretan los resultados obtenidos por los modelos, para poder generar información relevante para el problema que estamos abordando, ya que, sin

esta interpretación final, los resultados de los modelos no dejan de ser simples cifras sin un significado real, sin un contexto los datos no pueden transformarse en conocimiento.

A parte de estos pasos que hemos mencionado, algunas referencias añaden otros como la reducción de la dimensionalidad, la validación, la integración, etc, aunque en general estos se pueden englobar en algunos de los que ya están mencionados.

3.2.2. Análisis estadístico de datos

Nos apoyamos en un análisis profundo de los datos para poder obtener información relevante de nuestro set de datos. En el proyecto, este proceso de análisis consiste en la realización de un contraste de hipótesis sobre las variables categóricas y continuas, mediante dos tests de independencia, cada uno enfocado a un tipo de variables, además de una búsqueda de similitudes entre los distintos tipos de perfiles de tratamiento, y por último una selección de características relevantes.

Contraste de hipótesis

El contraste de hipótesis se utilizará en el contexto de nuestro trabajo para determinar cuándo dos conjuntos de datos son independientes, pero antes de definir lo que es un contraste de hipótesis de manera más extensa, se explicará en que consiste formalmente una hipótesis.

Una hipótesis es una proposición acerca de una característica o propiedad de una población de estudio. Un ejemplo de hipótesis podría ser, “la variable x toma valores entre los intervalos (a, b)”. En el campo de los contrastes de hipótesis, dos conceptos importantes son los de hipótesis nula e hipótesis alternativa.

La hipótesis nula, que se representa como H_0, es aquella que se desea contrastar, la planteada originalmente, mientras que una hipótesis alternativa, representada por H_1,es la negación de la hipótesis nula, y estas dos son la base sobre la que se sustentan los contrastes de hipótesis (también conocido como test de hipótesis o prueba de significancia). Estos se definen como un procedimiento para determinar si una propiedad o característica que se supone a una población estadística es compatible con lo observado en una muestra de dicha población. Relacionándolo con los conceptos anteriores, en el contraste de hipótesis se considera la hipótesis H_0, y mediante diferentes métricas se trata de encontrar si se puede negar la hipótesis nula H_0, y por tanto dar como cierta la hipótesis alternativa H_1.

Para determinar cuál de las hipótesis será la seleccionada, se pueden aplicar diferentes tests sobre las muestras de las poblaciones estadísticas, que devolverán un cierto valor estadístico conocido como P-value, y que definirá si es posible descartar o no la hipótesis nula y aceptar o no la hipótesis alternativa por consiguiente.

P-value

A lo largo de los estudios estadísticos que realizaremos en este trabajo nos encontraremos varias veces con el concepto P-value, por lo que merece la pena pararse a definir este importante término.

Los P-values, también conocidos como significancias estadísticas, y representados muchas veces por α, se definen como la probabilidad de que un valor estadístico calculado sea posible dada una hipótesis nula cierta, y como habían indicado anteriormente, se pueden considerar como el resultado a la aplicación del contraste de hipótesis, Schervish (1996).

Una definición alternativa podría ser, el P-value es la probabilidad de observar los resultados del estudio, u otros más alejados de la hipótesis nula (o hipótesis de partida), si la hipótesis nula fuera cierta.

Los valores que puede tomar esta medida oscilan entre 0 y 1, y arbitrariamente se define una condición a partir de la cual se considera que el resultado es estadísticamente significativo. En general este valor "umbral" se define en el 0.05 (en ocasiones el 0.01). Un contraste de hipótesis permite rechazar la hipótesis nula H_0, pero nunca probarla, por lo que para valores de $\alpha > 0.05$, no podremos rechazar la hipótesis nula, pero tampoco probarla, mientras que si $\alpha < 0.05$, estamos rechazando la hipótesis nula, y por consiguiente implícitamente estamos aceptando la hipótesis alternativa H_1. Dependiendo de la hipótesis nula planteada, y los métodos utilizados para tratar de probarla o desmentirla, tendremos diferentes tipos de test. En este trabajo usamos concretamente dos, el test de independencia Chi Square, y el test de independencia T-Test.

Test de independencia Chi Square

Este test de independencia se aplica sobre variables de tipo categóricas, y sirve para probar la hipótesis nula de que las distribución de frecuencia de dos variables (poblaciones) son independientes entre sí, por lo que la hipótesis alternativa será la dependencia entre las distribuciones de frecuencia de las variables categóricas. Fue desarrollado en 1900 por Karl Pearson, y es muy utilizado en el campo de la estadística inferencial.

Test de independencia T-Test

Se usa sobre variables continuas. Este test asume como hipótesis nula que las medias de dos poblaciones distribuidas en forma normal son iguales, por lo que, si la hipótesis nula es rechazada, la hipótesis alternativa, que probaría que las poblaciones tienen una distribución diferente, se puede asumir como verdadera. Fue introducido en 1908 por William Sealy Gosset.

3.2.3. Selección de características

Podemos definirla como el proceso de seleccionar aquellas características más relevantes dentro de un conjunto de datos. Esta relevancia se puede obtener por distintos métodos, como los de filtro, de envoltura o integrados.

En nuestro proyecto vamos a emplear el primero de estos, el método de filtro. Este se basa en una clasificación de las variables según los puntuajes estadísticos que tienden a determinar la correlación entre las características con la variable objetivo. Existen diferentes funciones para evaluar estas correlaciones, como la Correlación de Pearson, método ANOVA, Chi-cuadrado o, la información mutua. El método aplicado es el ANOVA, y está basado en la aplicación de tests F-Test para probar estadísticamente la igualdad de la media entre los dos conjuntos que estamos estudiando.

3.2.4. Métricas de similitud

En estadística, la similitud se define como una función que cuantifica la semejanza entre dos objetos, aunque no podemos considerar esta una definición única . Existen muchas medidas para definir la similitud entre dos objetos, sin embargo, nosotros nos vamos a centrar en nuestro trabajo en dos en concreto, la similitud coseno, y la distancia euclídea, que definiremos a continuación.

Similitud coseno

Es una medida de la similitud existente entre dos vectores distintos de cero dentro de un espacio interno del producto, que nos devuelve el coseno del ángulo entre los dos vectores. Dos vectores con la misma dirección tendrán un ángulo de 0º entre ellos, por lo que el valor de la similitud coseno será de 1, mientras que si los dos vectores están orientados con 90º entre sí, el valor de la similitud será de 0. Por tanto, cuanto mayor sea el valor de la similitud coseno en el rango (0,1), más similares serán nuestros dos vectores. Se calcula a través de la siguiente expresión:

$$s_c(P,Q) = \frac{\sum_{i,j}^{n} p_i * q_j}{\sqrt{\sum_{i,j}^{n} p_i * p_j} * \sqrt{\sum_{i,j}^{n} q_i * q_j}}$$

Donde S y Q son los vectores para los que estamos calculando la similitud, y los índices i, j recorren cada elemento de los vectores.

Distancia euclídea

La distancia euclídea mide la distancia existente entre dos puntos de un espacio vectorial euclídeo, o lo que es lo mismo, entre dos vectores que partan del origen hasta dichos puntos, y se deduce esta métrica de distancia a través del teorema de Pitágoras. Para un espacio euclídeo n-dimensional, la distancia euclídea se calcula a través de la siguiente expresión:

$$d_E(P,Q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

Donde P y Q representan cada punto del espacio vectorial y p_i, q_i representan el elemento i-ésimo de cada vector que parte del origen de coordenadas hasta los puntos P y Q para los que estamos calculando la distancia euclídea.

Cuanto menor sea el valor de la distancia euclídea, más similares serán los dos vectores.

4. Desarrollo del proyecto y resultados

4.1. Metodología

En este apartado vamos a explicar los métodos utilizados para el desarrollo del proyecto.

En primer lugar, comenzaremos hablando del entorno utilizado. Este consistió en la plataforma Google Colab, un producto de Google Research que permite al usuario, nosotros en este caso, escribir y ejecutar código en Python (y otros lenguajes de programación) en el propio navegador, sin que exista una carga computacional en nuestro entorno local. Esta herramienta es de uso gratuito y muy utilizada en las tareas de análisis de datos, ya que la posibilidad de ir añadiendo cuadros de texto y de ejecutar el código por bloques nos aporta una gran versatilidad. La versión utilizada de Python es la 3.7.14.

Para poder realizar el proceso de importación de los datos, su limpieza, y su transformación, se hizo uso principalmente de las librerías Pandas y NumPy, siendo la primera la más utilizada, y para el análisis estadístico y la selección de variables se usaron, a parte de las anteriores, Sklearn y SciPy. Sklearn (Scikit-learn) es una de las herramientas de aprendizaje automático y modelado estadístico más populares en Python, y está basada en las librerías NumPy, SciPy y matplotlib. SciPy por otro lado es una librería formada por una colección de algoritmos matemáticos y otras funciones creadas sobre NumPy, y el módulo que nosotros hemos empleado es stats, que contiene un gran número de distribuciones de probabilidad, funciones de correlación, test estadísticos, etc. Por último, la visualización de los resultados se obtuvo mediante Seaborn, una librería muy famosa para la representación gráfica en el mundo del análisis, y Matplotlib. Como apunte extra cabe destacar que cada vez que se mencionan en el trabajo las tablas, los análogos que se usa para generarlas en Python serán los objetos conocidos como DataFrames, característicos de la librería Pandas.

El código completo a través del que se desarrollará todo el trabajo de programación, así como los datos en crudo, y un archivo pdf con las preguntas de la encuesta y sus codificaciones se puede encontrar en el siguiente repositorio de GitHub:

Enlace repositorio GitHub:

https://github.com/sergi1307/Trabajo-Final-Master-Sergio-Catalan-Torralbo-

4.2. Desarrollo del proyecto

A continuación, aportaremos una descripción detallada del proceso KDD que seguimos en el proyecto.

4.2.1. Origen de los datos

Los datos utilizados en este proyecto provienen de una encuesta realizada por el CIS entre el 19 y el 25 de febrero, llamada “Salud mental de los españoles durante la pandemia de la Covid-19”. Esta se realizó en el ámbito nacional (España) a población de ambos sexos de 18 años o más. La población encuestada abarcó un total de 1080 municipios y 50 provincias, y el contacto se llevó a cabo usando aleatoriamente una selección de teléfonos móviles y fijos, aplicando cuotas de sexo y edad para para poder obtener muestras representativas de los diferentes rangos de edad y del sexo.

Se realizaron un total de 3083 entrevistas de manera concluyente de las 3200 que se tenían diseñadas originalmente, por lo que el tamaño del dataset original consistía en 3083 registros. El número de variables que encontramos en nuestros datos es de 354, sin embargo, no todas se corresponden a preguntas formuladas a los encuestados, buena parte de ellas son variables que rellenadas a *posteriori* por el encargado de realizar las preguntas, y contienen información irrelevante, ya que no nos aportan información sobre los individuos que son objeto del estudio. Nuestro dataset tendrá, por tanto, un tamaño original de 3083 filas x 354 columnas, siendo todas las características de tipo numérico, ya que cada número representa una de las posibles respuestas a las preguntas. La codificación de cada pregunta la mostraremos una vez hayamos explicado cuales son las variables relevantes que hemos seleccionado para nuestro estudio estadístico, ya que mostrar las codificaciones de las 354 columnas es innecesario.

En el siguiente enlace podremos encontrar la ficha del estudio publicada por el CIS, en la que podremos descargar los datos en crudo de las entrevistas, así como a un archivo con los resultados de cada pregunta (porcentajes de cada posible respuesta) o la ficha técnica.

Enlace ficha del estudio:

https://www.cis.es/cis/opencm/ES/1_encuestas/estudios/ver.jsp?estudio=14551

4.2.2. Preprocesamiento

En esta sección damos una explicación detallada de todo el preprocesado aplicado a nuestros datos con el objetivo de mejorar la calidad de estos, para permitirnos realizar un análisis posterior de mayor calidad. Estará compuesto por una preselección de las características que resulten de interés y tratamientos de los missing values, de aquellos registros con respuesta N.C (No Contesta), y por último de las características desbalanceadas.

Preselección:

Hemos realizado una selección manual de aquellas variables de interés de acuerdo con el objetivo del presente proyecto, y centrándonos sobre todo en las que tenían relación con la salud mental o los factores que puedan afectar a esta. Por ejemplo, preguntas como las relacionadas con la intención de voto, la comunidad en la que vive el encuestado, y otras similares no se han seleccionado. El resultado de esta primera preselección, fueron las siguientes características, de ahora en adelante nos referiremos a ellas como "Variables de encuesta":

Variables de encuesta:

['SEXO', 'EDAD', 'P2', 'P3', 'P4', 'P4B', 'P10', 'P16A', 'P16B', 'P17A_1', 'P17A_2', 'P17A_3', 'P17A_4', 'P17A_5', 'P17A_6', 'P17A_7', 'P17A_8', 'P17A_9', 'P17A_10', 'P17A_11', 'P17A_12', 'P17A_13', 'P17A_14', 'P17A_15', 'P17A_16', 'P17A_17', 'P17A_18', 'P17A_19', 'P17A_96', 'P17A_99', 'P18', 'P18A', 'P26', 'P26A', 'P27', 'P27A', 'P28', 'P29', 'P30', 'P30C_1', 'P30C_2', 'P30C_3', 'SITLAB', 'SITCONVIVEN', 'ESCUELA', 'CNO11R', 'ESTUDIOS', 'RELIGIONR', 'CLASESUB']

Código de pregunta	Descripción de la pregunta	Posibles respuestas
SEXO	Sexo	Hombre/Mujer
EDAD	Edad	Edad
P2	Sanitario	Si/No/N.C
P3	Tenencia Covid	Si/No/N.C
P4	Hospitalización Covid	Si/No/N.C
P4B	Fallecimiento Familiar por Covid	Si/No/N.C
P10	Ataques de ansiedad desde la pandemia	Si/No/N.C
P16A	Enfermedad Crónica	Si/No/N.C
P16B	Convivencia con persona con enfermedad crónica	Si/No/N.C
P17A_1	Enfermedad cardiovascular	Menciona/No Menciona/N.C
P17A_2	Diabetes	Menciona/No Menciona/N.C
P17A_3	Cáncer	Menciona/No Menciona/N.C
P17A_4	Enfermedad respiratoria	Menciona/No Menciona/N.C
P17A_5	Enfermedad renal	Menciona/No Menciona/N.C
P17A_6	Esclerosis múltiple	Menciona/No Menciona/N.C
P17A_7	Enfermedad traumatológica	Menciona/No Menciona/N.C
P17A_8	Osteoporosis	Menciona/No Menciona/N.C
P17A_9	Depresión grave	Menciona/No Menciona/N.C
P17A_10	Demencia (alzhéimer, párkinson, etc.)	Menciona/No Menciona/N.C
P17A_11	Enfermedad autoinmune	Menciona/No Menciona/N.C
P17A_12	Otras enfermedades reumáticas	Menciona/No Menciona/N.C
P17A_13	Enfermedades gastrointestinales	Menciona/No Menciona/N.C
P17A_14	Enfermedad de sistema nervioso, neurológica/neuromuscular	Menciona/No Menciona/N.C
P17A_15	Colesterol	Menciona/No Menciona/N.C
P17A_16	Alergias	Menciona/No Menciona/N.C

P17A_17	Otras enfermedades y trastornos mentales	Menciona/No Menciona/N.C
P17A_18	Enfermedades hepáticas	Menciona/No Menciona/N.C
P17A_19	Patologías del trastorno del sueño	Menciona/No Menciona/N.C
P17A_96	Otra enfermedad crónica	Menciona/No Menciona/N.C
P17A_99	Tiene enfermedad crónica pero no contesta el tipo	Menciona/No Menciona/N.C
P18	Tenencia hijos menores convivientes en pandemia	Si/No/N.C
P18A	Número de hijos convivientes en pandemia	1/2/3/4/5 o más/N.C/N.P
P26	Tratamiento psicológico antes de la pandemia	Si/No/N.C
P26A	Trastorno mental sufrido antes de pandemia	*
P27	Tratamiento psicológico desde la pandemia	Si/No/N.C
P27A	Trastorno mental sufrido desde la pandemia	*
P28	Tratamiento farmacológico antes de pandemia (por salud mental)	Si/No/N.C
P29	Tratamiento farmacológico desde la pandemia (por salud mental)	Si/No/N.C
P30	Familiar medicándose por salud mental desde la pandemia	Si/No/N.S/N.C
P30C_1	Consumo actual fármacos familiar 1	Si/No/N.S/N.C/N.P
P30C_2	Consumo actual fármacos familiar 2	Si/No/N.S/N.C/N.P
P30C_3	Consumo actual fármacos familiar 3	Si/No/N.S/N.C/N.P
SITLAB	Situación laboral	*
SITCONVIVEN	Situación de convivencia	*
ESCUELA	Escolarización	No, es analfabeto/No, pero saber leer y escribir/Si, asistió escuela/N.C
CNO11R	Tipo ocupación laboral	*
ESTUDIOS	Rango de estudios	*
RELIGIONR	Religiosidad persona entrevistada	*
CLASESUB	Clase social subjetiva	*

Tabla 1. Variables de encuesta y su significado. Fuente: Elaboración propia.

**Las variables poseen demasiados tipos de respuestas, para encontrarlas dirigirse a la ficha de la encuesta.*

Missing Values:

Dado que estamos trabajando con una encuesta del CIS que ya ha sido previamente tratada, en teoría no debería haber Missing Values o valores erróneos, sin embargo, cuando lo comprobamos se observa que para un porcentaje elevado (> 50%) de columnas existen muchos valores vacíos. Tras realizar un estudio de estos nos dimos cuenta de que estos se correspondían siempre con preguntas en las que se encontraba la opción N.P (No Procede) como respuesta. Un ejemplo de esto es, por ejemplo, la pregunta P4 (Hospitalización por Covid ahora o en el pasado de la persona entrevistada). El motivo de que existan muchos registros vacíos para esta pregunta (N.P como respuesta) es que, anteriormente, en la pregunta P3, se pregunta si la persona entrevistada ha sufrido Covid, por lo que cuando las personas responden a esta P3 con un "No", ya no procede preguntar si la persona ha estado hospitalizada por Covid. Esta es la misma lógica que siguen todas aquellas preguntas que presentan Missing Values, por lo que decidimos que para todos los registros sustituiríamos el valor vacío por un 0, que de ahora en adelante definiremos como la respuesta negativa a preguntas de Si/No. Esto lo hacemos ya que en el caso de una persona que no ha sufrido Covid, posee el mismo significado declarar que no ha estado hospitalizada a señalar la respuesta de NP. Aplicando esto a todas las columnas (se comprobó para cada una que se podía seguir el mismo criterio) eliminamos la totalidad de los valores vacíos de nuestro dataset.

Registros con respuestas N.C (No contesta)

A lo largo de la encuesta se da la posibilidad de no contestar a una pregunta si esta no es cómoda para el encuestado, por lo que en estos casos no tenemos información extraíble de esa característica. En primer lugar, comprobamos que cantidad de registros contenían al menos uno de sus campos con respuesta N.C, obteniendo un total de 285. Al no ser este un número muy elevado sobre el tamaño del set de datos original, y comprobar que las respuestas N.C no están centradas en una sola columna, sino repartidas en muchas, decidimos eliminar los 285 registros, por lo que el tamaño de nuestro set de datos se vio reducido en este momento a 2798 filas x 49 columnas.

Variables desbalanceadas

Algo a tener en cuenta a la hora de estudiar datos categóricos es que las variables estén balanceadas, con esto nos referimos a que, si una pregunta tiene 3 posibles respuestas “a”, ”b” y “c”, pero dentro de nuestros registros la gente solo ha respondido la opción “a” la variable estará desbalanceada, por lo que esta pregunta a la hora de realizar estudios estadísticos o modelos de clasificación, no nos aportará ninguna información útil, y podremos descartarla. En el caso de preguntas como la variable ‘ESTUDIOS’, por ejemplo, al existir muchas posibles respuestas, es normal que algunas de ellas representen porcentajes muy reducidos del total, por lo que no eliminaríamos esta variable, sin embargo, para los casos en el que las posibles respuestas sean Si/No (codificadas como 1/0 respectivamente), si observamos que una de las dos opciones representa menos del 1% de las respuestas , eliminaremos dicha columna, ya que no nos aportará información.

Aplicando este criterio eliminamos 13 columnas:

['P17A_5', 'P17A_6','P17A_8', 'P17A_9', 'P17A_10', 'P17A_13', 'P17A_14', 'P17A_15', 'P17A_16', 'P17A_17', 'P17A_18', 'P17A_19', 'P17A_99']

Vemos que todas estas son variables tipo P17A_, que se corresponden con la enfermedad crónica que tiene la persona entrevistada. Tiene sentido que muchas de estas enfermedades (cada una es una columna y esta codificada con 1:Menciona, 2:No menciona) no contengan una gran representación dentro de nuestros datos, ya que son enfermedades poco comunes. Por ello existirán muy pocos registros que mencionen tener la enfermedad y podremos eliminarlas sin perder información relevante.

4.2.3. Transformación

En primer lugar, definimos una transformación general que aplicamos a todas las columnas que lo requerían en el dataset. En muchos casos, debido al método que seguimos para rellenar los missing values, tendremos preguntas con respuestas de tipo “N.P”/ ”Menciona”/ ”No Menciona” (o lo que es lo mismo, “No Procede”/”Si”/ “No”), y con una codificación tipo 0/1/2. Sin embargo, como ya explicamos, tiene el mismo significado

la respuesta “No Menciona (2)” que la respuesta “No procede (0)”, por lo que para mejorar la efectividad de los modelos aplicados a los datos decidimos agrupar estas dos categorías en una sola, con codificación 0. Esto se aplicó a la gran mayoría de las preguntas, ya que muchas eran de este tipo.

El resto de las transformaciones son algo más específicas, por lo que las mencionaremos individualmente:

P18 y P18A: Tenencia de hijos con la que hayan convivido en pandemia y su número:

La variable P18 tendrá de opciones Si/No, y la P18A 0, 1 , 2, 3, 4, 5, que representan la cantidad de hijos, siendo la última opción 5 hijos o más. Con el objeto de reducir las categorías de la pregunta P18A, se agruparon en las opciones: 0, 1, 2 y 3 hijos, donde 3 significa 3 o más hijos. Sin embargo, tras realizar un estudio de la correlación entre todas las variables de nuestros datos, observamos que el coeficiente correlación de Pearson entre la P18 y la P18A modificada resultó ser superior a 0.9, por lo que finalmente decidimos descartar la P18A, quedándonos solo con la que nos habla sobre si la persona tenía hijos con los que convivio durante la pandemia.

P26A y P27A: Trastornos mentales sufridos antes y después de la pandemia:

Las dos variables que tratamos a continuación son idénticas, únicamente difieren en que, en la primera pregunta por antes de la pandemia, y en la segunda por después de la pandemia. El enunciado de estas es “Trastorno mental sufrido antes/después de la pandemia por la persona entrevistada”, y como posibles valores tiene una larga lista de posibles trastornos. Este tipo de codificación, en la que cada número corresponde con un tipo de trastorno, no nos interesa, ya que a nivel de computación, a nuestros modelos no le aporta información relevante, dado que no es una variable continua sino categórica. Es por ello que decidimos para cada una de las 17 opciones de respuesta presente, crear una variable, con el valor 0 si la persona no poseía la enfermedad, y con un 1 si la persona sí que sufría este trastorno mental. Sin embargo, esto aumentaría en 16 variables nuestros datos, y además podría generarnos algunas columnas en las que no existieran prácticamente valores positivos (cosa que habíamos corregido en la sección de “Variables desbalanceadas”). Por ello se decidió que solo para aquellas enfermedades cuyos valores positivos representaran un porcentaje superior al 2%, crearíamos una nueva columna en los datos. Tras realizar el pertinente código, los trastornos crónicos que se usaron para generar nuevas columnas codificadas con 0/1 (No/Si) y sus respectivos nombres de variables fueron:

- Trastorno depresivo antes de la pandemia: Trastorno_depresivo_antes (211 valores positivos, 6.84 % del total)
- Trastorno ansioso antes de la pandemia: Trastorno_ansioso_antes (179 valores positivos, 5.81 % del total)
- Trastorno depresivo después de la pandemia: Trastorno_depresivo_despues (70 valores positivos, 2.27 % del total)

- Trastorno ansioso después de la pandemia: Trastorno_ansioso_despues (86 valores positivos, 2.79 % del total)

Tras esto procedimos a eliminar las variables P26A y P27A, ya que están incluidas en las que hemos generado.

P26, P27, P28, P29: Generación de las variables objetivo:

Este apartado posee mucha relevancia, ya que en él se aborda el procedimiento para generar las variables objetivo, aquellas que nos servirán para evaluar los niveles de tratamiento psicológico o farmacológico a los que estuvieron sometidos los entrevistados antes y después de la pandemia.

Para ello haremos uso de las preguntas P26, P27, P28 y P29, que se corresponden con:

- P26: Tratamiento psicológico antes de la pandemia
- P27: Tratamiento psicológico después de la pandemia(*)
- P28: Tratamiento farmacológico antes de la pandemia(**)
- P29: Tratamiento farmacológico después de la pandemia(*) (**)

(*) En la encuesta realmente pregunta por el tratamiento desde el inicio de la pandemia, es decir, a cualquier momento desde el inicio de la pandemia hasta el momento de la entrevista.

(**)El tratamiento farmacológico ha de ser asociado a problemas de salud mental.

A partir de ellas generamos las variables objetivo, que como habíamos mencionado previamente, se encargan de definir los distintos niveles/perfiles de tratamiento que han recibido los pacientes, en base al tipo de tratamiento del que se trata (psicológico y/o farmacológico), y también en base a cuando lo recibieron (antes y/o después de la pandemia)

Las categorías definidas son:

- Solo tratamiento farmacológico y solo antes de la pandemia → **"farm_antes"**
- Solo tratamiento farmacológico y solo después de la pandemia → **"farm_despues"**
- Solo tratamiento farmacológico antes y después de la pandemia → **"farm_antes_y_despues"**
- Solo tratamiento psicológico y solo antes de la pandemia → **"psic_antes"**
- Solo tratamiento psicológico y solo después de la pandemia → **"psic_despues"**
- Solo tratamiento psicológico antes y después de la pandemia → **"psic_antes_y_despues"**
- Tratamiento farmacológico y psicológico solo antes de la pandemia → **"psi_y_farm_antes"**

- Tratamiento farmacológico y psicológico solo después de la pandemia → **“psi_y_farm_despues”**
- Tratamiento farmacológico y psicológico antes y después de la pandemia → **“psi_y_farm_antes_y_despues”**
- Ningún tratamiento, ni antes ni después de la pandemia → **“nunca_psic_o_farm”**

4.2.4. Características finales:

Tras realizar todo el preproceso y las transformaciones, las variables que usaremos serán las siguientes (las dividimos entre las que están presentes en la encuesta y las variables objetivo).

Variables objetivo (representan los distintos perfiles de tratamiento):

“farm_antes”, “farm_despues”, “farm_antes_y_despues”, “psic_antes”, “psic_despues”, “psic_antes_y_despues”, “psic_y_farm_antes”, “psic_y_farm_despues”, “psic_y_farm_antes_y_despues” y “nunca_psic_o_farm”

Variables de encuesta:

'SEXO', 'EDAD', 'P2', 'P3', 'P4', 'P4B', 'P10', 'P16A', 'P16B', 'P17A_1', 'P17A_2', 'P17A_3', 'P17A_4', 'P17A_7', 'P17A_11', 'P17A_12','P17A_96', 'P18', 'P30', 'SITLAB', 'SITCONVIVEN', 'ESCUELA', 'ESTUDIOS', 'RELIGIONR', 'CLASESUB', 'Trastorno_Depresivo_Antes’, ‘Trastorno_ansioso_antes', 'Trastorno_depresivo_despues', 'Trastorno_ansioso_despues'

Debido a que el nombre de muchas de estas variables no es reflejo de la información que aportan, decidimos renombrarlas con el objetivo de mejorar la información que transmitirán a la hora de representarlas.

Variables de encuesta renombradas:

'Sexo', 'Edad', 'Sanitario', 'Tenencia_Covid', 'Hospitalizacion_Covid', 'Fallecimiento_Familiar_por_Covid', 'Ataques_Ansiedad', 'Enfermedad_Cronica', 'Enfermedad_Cronica_Familiar', 'Enfermedad_Cardiovascular', 'Diabetes', 'Cancer', 'Enfermedad_Respiratoria', 'Enfermedad_Traumatologica', 'Enfermedad_Autoinmune', 'Otra_Enfermedad_Cronica','Trastorno_Sueño', 'Tenencia_de_Hijos', 'Medicacion_Familiar_por_Salud_Mental', 'Situacion_Laboral', 'Situacion_Convivencia', 'Escolarizacion', 'Estudios', 'Religion', 'Clase_Social', 'Trastorno_Depresivo_Antes', 'Trastorno_Ansioso_Antes', 'Trastorno_Depresivo_Despues', 'Trastorno_Ansioso_Despues'

Como resultado de este proceso, se llegó a un dataset final consistente en 29 “Variables de encuesta” y 10 “Variables objetivo”, con un total de 2798 registros ya preprocesados. Con estos datos ya fuimos capaces de comenzar a realizar un estudio estadístico sobre los encuestados y obtener información relevante.

4.2.5. Data Mining

En el proceso de data mining se realizó el estudio estadístico, cuyos resultados explicaremos en la sección “4.3. Resultados”. Este paso consistió en un análisis estadístico de los datos, mediante contrate de hipótesis, realizando dos test de independencia, Chi Square y T-Test, para variables categóricas y continuas respectivamente, seguido de una selección de características, y a continuación un estudio de similitudes entre las variables objetivo.

A parte, aplicamos un árbol de decisión sobre una de las variables objetivos para comprobar si las características más relevantes seleccionadas por este coincidían con las que mostraban los resultados de nuestro estudio estadístico.

A continuación, hablaremos más en profundidad sobre cada uno de estos procesos.

Contraste de hipótesis

Estudio estadístico de variables categóricas mediante Chi-Square

En esta sección hicimos un estudio estadístico para las variables de encuesta categóricas de nuestro dataset, que son todas exceptuando la edad. Para ello usamos la librería de Python SciPy, y su módulo Stats. Nuestro objetivo era determinar si había relación entre las categorías de cada variable y las categorías de cada variable objetivo. Para ello, primero obtuvimos para cada par “variable de encuesta – variable objetivo” una tabla de contingencia. A modo de ejemplo mostraremos a continuación la *Tabla 2*:

Sexo	farm_antes	
	0	1
0	1031	22
1	1054	9

Tabla 2. Tabla de contingencia entre variables Sexo y farm_antes

Observamos en ella que la variable de encuesta “Sexo” tiene los posibles valores 0 y 1, y que la variable objetivo “farm_antes” tiene los posibles valores 0 y 1 también. Gracias a la tabla de contingencia podemos observar las distintas posibles combinaciones de estas variables y observar cuantos registros hay en cada caso. No hubiera sido complejo realizar esta tabla de manera manual, pero gracias a la función crosstab() de la librería Pandas, simplemente pasándole ambas columnas, la obtuvimos automáticamente. A continuación, utilizando la función scipy.stats.chi2_contingency, fuimos capaces de obtener para tabla de contingencia de cada par “variable de encuesta – variable objetivo” (sexo y farm_antes en este ejemplo), un valor de la independencia entre estas dos variables, mediante el test Chi-Square, y otro valor para la significancia estadística (p-value) de dicho test. Este último es el resultado que nos interesa, ya que es el que nos determina si la hipótesis de independencia entre ambas variables asumida por el test es real o no. Por tanto, cuando el valor de p-value sea más cercano a cero, más lejos estarán de ser independientes ambas variables, lo que implicara que esta más relacionadas. Repetimos este proceso completo para todos los pares “variable de

encuesta – variable objetivo", obteniendo así una tabla final en el que en las columnas colocamos las variables objetivo, y en las filas las variables de encuesta, y como valores los resultados de los p-values del test.

Tras esto decidimos generar otra tabla, con la misma forma que la anterior, pero en la que en lugar de colocar como valores los p-values, colocáramos 1 para aquellos valores en los que el p-value era menor que 0.05, y un 0 en aquellos p-values mayores a 0.05. Esto nos sirvió para crear una tabla en la que los 0 representaban aquellas características que no eran relevantes para la variable objetivo, y en la que el 1 representa que sí que son relevantes.

Estudio estadístico de variables continuas mediante T-Test

La única variable continua presente en nuestro dataset es "Edad", y aunque podríamos considerarla categórica si agrupáramos las edades en distintos rangos, se ha usado de la manera original con el objeto de poder aplicar sobre ella un tipo de test estadístico enfocado a variables continuas, y así enriquecer el estudio que estamos realizando. Antes de comenzar con el análisis propiamente dicho, realizamos una comprobación de la distribución de las edades, para verificar que los resultados no estan sesgados a costa de una distribución irregular.

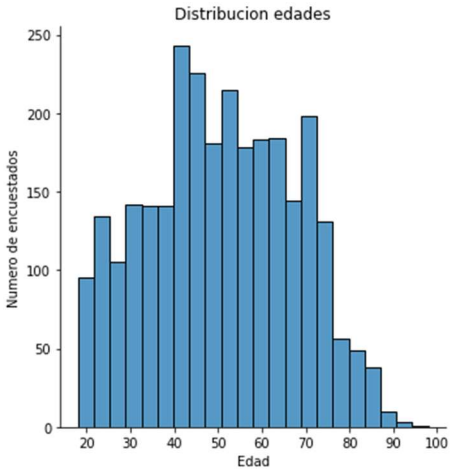


Ilustración 2. Distribución de edades en la población encuestada. Fuente: Elaboración propia.

Podemos comprobar en la Ilustración 3 que efectivamente la edad está distribuida de manera que ningún rango de edad se queda sin representación (exceptuando los menores de 18 años que no entraban dentro de la encuesta), y ninguno de los rangos es significativamente más elevado que los demás.

Nuestro objetivo aquí será determinar cómo afecta la edad a la distribución de las variables objetivos, para comprobar si existe una dependencia clara o no, sin embargo, no podemos hacerlo por el mismo método que con las categóricas, por ello aplicamos otro enfoque.

Primero tendremos que almacenar, para cada una de las variables objetivo, en una lista, aquellos registros de la edad para los que dicha variable objetivo es 0, y en otra lista

aquellos en los que vale 1. A continuación haciendo uso de la función de SciPy stats.ttest_ind() realizaremos un T-Test sobre ambas listas, que nos permitirá determinar si la distribución de ambas listas es independiente, o, por el contrario, existe dependencia entre ambas. Esto nos hará comprobar si está la edad relacionada con cada uno de los tipos de tratamientos que definimos. En el caso de que no estuvieran relacionadas las variables, deberíamos observar que la distribución de edades para el valor 0 de la variable objetivo, no presenta dependencia con distribución de edades para el valor 1 de la variable objetivo.

Las métricas que nos devuelve la función para comprobar si son independientes o no son el estadístico t, y el p-value asociado, que hemos calculado para cada variable objetivo, y que almacenamos en dos tablas diferentes. Dado que en este test la hipótesis nula es la independencia entre variables, cuando encontremos valores de p-value que sean menores a 0.05, podremos negar esta hipótesis de partida, y afirmar como cierta la hipótesis alternativa, que afirma que ambas distribuciones son dependientes.

Evaluación de similitudes

Similitud coseno

Nuestro objetivo en este paso será encontrar la similitud coseno entre cada una de las variables objetivos. Para ello haremos uso de la última tabla generada en el "Estudio estadístico de variables categóricas mediante Chi-Square", que posee una columna por cada variable objetivo, y como registros los 0 y 1 que indicaban si cada una de las variables de encuesta eran significativas o no. Evaluar la similitud coseno entre cada una de las columnas nos permitirá obtener una métrica de cuan similares son las variables objetivo entre si dependiendo de aquellas variables de encuesta que son relevantes para cada una. Como cada variable objetivo corresponde con un tipo de tratamiento, cada una se asocia a un perfil de tratamiento de los encuestados, y obtener esta métrica de similitud nos permite por tanto comprobar como de diferentes son los perfiles entre ellos. Un ejemplo de para que podría servir podría ser comprobar si las características que son relevantes para las personas que solo han recibido tratamiento psicológico después de la pandemia son similares a las de aquellas personas que llevan en tratamiento psicológico desde antes de la pandemia, o por el contrario son perfiles diferentes.

Para lograrlo usamos la función spatial.distance.cosine de la librería SciPy, que cuando le proporcionamos dos listas (arrays), nos devuelve la similitud coseno entre ellas. En nuestro caso, como habíamos mencionado, las listas que aportaremos serán dos columnas de la tabla con las relevancias de las características codificadas con 0 y 1, así que si quisiéramos encontrar la similitud coseno entre la variable objetivo "farm_antes" y "psic_antes", simplemente pasaríamos a la función como primer argumento la lista con los 0 y 1 de las relevancias para farm_antes, y como segundo argumento la lista con los 0 y 1 de las relevancias de psic_antes. Esto nos devolverá directamente un valor para esta similitud coseno, que almacenamos en una matriz, que como filas y columnas tendrás las variables objetivo y como valores en ella los resultados obtenidos.

Distancia euclídea

Como siguiente paso de búsqueda de similitudes entre de los perfiles de tratamiento, obtuvimos una tabla de distancia euclídea entre las distintas variables objetivo, sin embargo, este caso no usaremos la tabla con las relevancias, aplicaremos otro método.

Primero tendremos que generar una nueva matriz, en la que cada columna corresponderá con una variable objetivo, y en las filas colocaremos la tabla de contingencia de cada variable de encuesta, pero fila por fila, y en porcentajes. Para visualizar esto vamos a seguir con el ejemplo que usamos antes, con la columna "farm_antes" y la primera variable que aparece, que era "Sexo". Como vimos en *Tabla 2*, las posibles combinaciones de la tabla de contingencia son:

Sexo = 0 y farm_antes = 0 con 1361 registros

Sexo = 0 y farm_antes = 1 con 37 registros

Sexo = 1 y farm_antes = 0 con 1387 registros

Sexo = 1 y farm_antes = 1 con 13 registros

Estas posibles combinaciones habremos de colocarlas entonces de la manera en la que están ordenadas aquí en la tabla que estamos creando, y habiéndolas transformado al porcentaje que representan sobre el total de registros, de forma que para esta columna ("farm_antes"), añadiríamos los valores 48.64%, 1.32%, 49.57%, 0.46% en filas consecutivas, y repetiríamos lo mismo para el resto de variables, colocando una detrás de otra. Esto nos generará una columna, donde el número de filas será la suma de cantidad de categorías de todas las variables de encuesta, multiplicado por los dos posibles valores de la variable objetivo (0 y 1).

En la *Tabla 3* se mostrará un ejemplo de cómo quedarían representadas las tablas de contingencia de las dos primeras variables de encuesta. La columna "Combinación categorías" de la *Tabla 3* solo está añadidas para facilitar la lectura, no se usará a la hora de calcular las distancias euclídeas:

Variables	Combinacion categorias	psic_antes	psic_despues	psic_antes_y_despues	farm_antes	farm_despues	farm_antes_y_despues
Sexo	(0, 0)	45.14	49.25	49.04	48.64	49.0	49.5
Sexo	(0, 1)	4.82	0.71	0.93	1.32	0.96	0.46
Sexo	(1, 0)	45.85	49.68	49.61	49.57	49.61	49.96
Sexo	(1, 1)	4.18	0.36	0.43	0.46	0.43	0.07
Sanitario	(0, 0)	85.28	92.67	92.42	91.85	92.21	93.07
Sanitario	(0, 1)	8.22	0.82	1.07	1.64	1.29	0.43
Sanitario	(1, 0)	5.72	6.25	6.22	6.36	6.4	6.4
Sanitario	(1, 1)	0.79	0.25	0.29	0.14	0.11	0.11

Tabla 3. Distribución de categorías de variables de encuesta para cada perfil de tratamiento. Fuente: Elaboración propia.

Este mismo proceso se repetirá para todas las variables objetivo y variables de encuesta, hasta completar la tabla completa.

El objetivo de crearla ha sido analizar a continuación la distancia euclídea entre cada columna objetivo, para obtener una métrica diferente de cuan similares son los perfiles de las personas encuestadas, pero esta vez teniendo en cuenta como se distribuyen las categorías de cada variable de encuesta y de las variables objetivo, cosa que no sucedía en la búsqueda de la similitud coseno, donde solo se tenía en cuenta que características eran relevantes.

Para hacer este último estudio volvimos a hacer uso de la librería SciPy, esta vez de la función spatial.distance.euclidean(), a la que le pasamos para cada caso las dos columnas de las que quisiéramos obtener la distancia euclídea, por ejemplo, entre "farmacos_despues" y "psic_despues", y nos devolverá un valor numérico con el valor de la distancia. Cuanto menor sea este valor, menor será la distancia euclídea entre ambas variables objetivo.

Selección de características

Haciendo uso de la función SelectKBest de la librería Sklearn (orientada a Machine Learning), módulo "Feature Selecction", fuimos capaces de obtener para cada variable objetivo, una lista con los k-scores, donde los k-scores son los valores que representan la relevancia de una característica respecto a las demás presentes, y cuanto mayor sean, más relevante es.

El algoritmo usado para la obtención de estos valores nos permite usar diferentes métricas para evaluar la relevancia de las características, entre las que se encuentran "f-classif", "chi2", "mutual_info_classif", "f-regresion" y varias más, siendo "f-classif" la que escogimos (viene por defecto si no seleccionamos otro algoritmo).

Para ejecutar este algoritmo de Machine Learning simplemente tuvimos que pasarle como primer argumento al método .fit() el dataframe en el que estaban contenidas todas las variables de encuesta con sus respectivos valores, y como segundo argumento la variable objetivo para la que vamos a obtener la métrica k-scores, con lo que nos quedaba una función como la siguiente:

*k-scores = SelectKBest(k='all').fit(X, Y) .scores_*

Donde X representa todas las variables de encuesta almacenadas en un DataFrame, e Y representa la variable objetivo para la que estemos calculando las relevancias.

Tras repetir este proceso para cada una de las variables objetivos almacenamos todos los k-scores en una misma tabla.

Por último, realizamos otra tabla para cada métrica, en la que almacenamos los mismos resultados, pero en este caso normalizados, es decir en un rango entre 0 y 1. El proceso para llegar a esta normalización fue simplemente, para cada variable objetivo, dividir todos los resultados del test entre el valor más alto de k-score para esa variable objetivo,

logrando que los resultados se encuentren entre 0 y 1. Esta normalización tiene un problema, y es que si una variable es mucho más relevante que las demás, el resto pasan a tener valores muy bajos, ya que en comparación su k-score es bajo, por lo que estaremos perdiendo información. Como explicaremos en la sección de resultados, este fue un problema al que nos enfrentamos en nuestros datos, por lo que para resolverlo simplemente repetimos todo el proceso, pero esta vez sin estas columnas que son tan relevantes frente a las demás.

4.3. Resultados

En este apartado vamos a analizar y comentar los resultados obtenidos en el proyecto. Primero comenzaremos observando los datos resultantes del contraste de hipótesis aplicado con los dos tests que ya mencionamos. Acto seguido analizaremos las métricas resultantes de los análisis de similitudes aplicados, y por último comprobaremos cuales son aquellas características más relevantes según la selección de variables que hemos aplicado. Dado que originalmente poseemos 10 variables objetivo, y a la hora de representarlas, las tablas quedaban poco legibles, decidimos representar únicamente 7 de ellas, mejorando así la interpretabilidad de los resultados, sin embargo, las versiones completas se pueden encontrar en el apéndice del trabajo. Las características que no representaremos en esta parte serán:

"psic_y_farm_antes", "psic_y_farm_despues" y "psic_y_farm_antes_y_despues"

Nos quedaremos por tanto con aquellas variables correspondientes a los perfiles de tratamiento psicológico y farmacológico por separado, y al perfil que no tiene ningún tratamiento asociado.

4.3.1. Contraste de hipótesis

Test de independencia Chi Square

En primer lugar, observamos la *Tabla 4*, que contiene los p-values correspondiente a aplicar el test de independencia Chi Square para cada par "Variable encuesta- Variable Objetivo".

Variables	psic_antes	psic_despues	psic_antes_y_despues	farm_antes	farm_despues	farm_antes_y_despues	nunca_psic_o_farm
Sexo	0.257	0.098	0.033	0.001	0.024	0.01	0.0
Sanitario	0.171	0.001	0.001	0.886	1.0	0.11	0.004
Tenencia_Covid	0.752	0.256	0.589	0.311	0.982	0.901	0.973
Hospitalizacion_Covid	0.562	0.019	1.0	1.0	0.817	1.0	0.472
Fallecimiento_Familiar_por_Covid	0.552	0.561	0.0	1.0	0.074	0.948	0.029
Ataques_Ansiedad	0.354	0.0	0.0	0.73	0.0	0.002	0.0
Enfermedad_Cronica	0.133	0.431	1.0	0.142	0.547	0.228	0.0
Enfermedad_Cronica_Familiar	0.753	0.733	0.301	0.228	0.096	0.326	0.0
Enfermedad_Cardiovascular	0.604	0.675	0.767	1.0	0.727	0.452	0.783
Diabetes	0.339	0.443	0.333	0.222	0.581	0.818	0.09
Cancer	0.681	0.837	0.465	1.0	0.69	0.768	0.026
Enfermedad_Respiratoria	0.777	1.0	0.02	0.32	0.118	0.686	0.001
Enfermedad_Traumatologica	0.766	0.447	0.86	0.126	0.18	1.0	0.004
Enfermedad_Autoinmune	0.849	1.0	1.0	1.0	1.0	0.898	0.038
Otra_Enfermedad_Cronica	0.445	0.378	0.637	0.003	1.0	0.836	0.007
Trastorno_Sueño	0.15	0.715	0.586	0.064	1.0	1.0	0.109
Tenencia_de_Hijos	0.993	0.184	0.707	0.172	0.786	1.0	0.658
Medicacion_Familiar_por_Salud_Mental	0.871	0.354	0.594	1.0	1.0	0.001	0.0
Situacion_Laboral	0.015	0.085	0.087	0.023	0.479	0.0	0.0
Situacion_Convivencia	0.06	0.844	0.02	0.52	0.571	0.592	0.001
Escolarizacion	0.805	0.853	0.817	0.766	0.813	0.924	0.221
Estudios	0.052	0.294	0.128	0.005	0.258	0.0	0.823
Religion	0.017	0.328	0.455	0.014	0.651	0.634	0.502
Clase_Social	0.271	0.215	0.189	0.353	0.683	0.159	0.033
Trastorno_Depresivo_Antes	0.0	0.264	0.056	0.102	0.17	0.597	0.0
Trastorno_Ansioso_Antes	0.0	0.334	0.0	0.145	0.227	0.686	0.0
Trastorno_Depresivo_Despues	0.02	0.0	0.0	0.539	0.672	1.0	0.0
Trastorno_Ansioso_Despues	0.009	0.0	0.0	0.445	0.572	1.0	0.0

Tabla 4.P-Values Test de independencia estadística Chi Square. Fuente: Elaboración propia.

Recordamos que como habíamos mencionado, aquí estamos representando los p-values, por lo que solo aquellos registros con $\alpha < 0.05$ son considerados relevantes para nuestra variable objetivo. Dado que en este formato de tabla resulta complejo discernir que variables son relevantes bajo este criterio y cuales no, a continuación, colocaremos la tabla *Tabla 5* con las relevancias codificadas como {0: No relevante, 1: Relevante}.

index	psic_antes	psic_despues	psic_antes_y_despues	farm_antes	farm_despues	farm_antes_y_despues	nunca_psic_o_farm
Sexo	0.0	0.0	1.0	1.0	1.0	1.0	1.0
Sanitario	0.0	1.0	1.0	0.0	0.0	0.0	1.0
Tenencia_Covid	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Hospitalizacion_Covid	0.0	1.0	0.0	0.0	0.0	0.0	0.0
Fallecimiento_Familiar_por_Covid	0.0	0.0	1.0	0.0	0.0	0.0	1.0
Ataques_Ansiedad	0.0	1.0	1.0	0.0	1.0	1.0	1.0
Enfermedad_Cronica	0.0	0.0	0.0	0.0	0.0	0.0	1.0
Enfermedad_Cronica_Familiar	0.0	0.0	0.0	0.0	0.0	0.0	1.0
Enfermedad_Cardiovascular	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Diabetes	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Cancer	0.0	0.0	0.0	0.0	0.0	0.0	1.0
Enfermedad_Respiratoria	0.0	0.0	1.0	0.0	0.0	0.0	1.0
Enfermedad_Traumatologica	0.0	0.0	0.0	0.0	0.0	0.0	1.0
Enfermedad_Autoinmune	0.0	0.0	0.0	0.0	0.0	0.0	1.0
Otra_Enfermedad_Cronica	0.0	0.0	0.0	1.0	0.0	0.0	1.0
Trastorno_Sueño	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Tenencia_de_Hijos	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Medicacion_Familiar_por_Salud_Mental	0.0	0.0	0.0	0.0	0.0	1.0	1.0
Situacion_Laboral	1.0	0.0	0.0	1.0	0.0	1.0	1.0
Situacion_Convivencia	0.0	0.0	1.0	0.0	0.0	0.0	1.0
Escolarizacion	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Estudios	0.0	0.0	0.0	1.0	0.0	1.0	0.0
Religion	1.0	0.0	0.0	1.0	0.0	0.0	0.0
Clase_Social	0.0	0.0	0.0	0.0	0.0	0.0	1.0
Trastorno_Depresivo_Antes	1.0	0.0	0.0	0.0	0.0	0.0	1.0
Trastorno_Ansioso_Antes	1.0	0.0	1.0	0.0	0.0	0.0	1.0
Trastorno_Depresivo_Despues	1.0	1.0	1.0	0.0	0.0	0.0	1.0
Trastorno_Ansioso_Despues	1.0	1.0	1.0	0.0	0.0	0.0	1.0

Tabla 5. Variable de encuesta categóricas relevantes para cada perfil de tratamiento. Fuente: Elaboración propia

Con el objeto de facilitar la identificación de las características, hemos resaltado en gris aquellas relevantes para cada columna tras aplicar el criterio mencionado anteriormente.

Si quisiéramos realizar un análisis completo de este dataset, deberíamos realizar análisis de cómo han variado los porcentajes de cada categoría relevante según el perfil de tratamiento, sin embargo, dado la complejidad y amplitud de este proceso, hemos decidido que quedará fuera del marco de este trabajo. Por ello vamos a centrarnos para cada tipo de tratamiento (psicológico o farmacológico), en analizar una de las características relevantes, y estudiarla en profundidad. En el caso de los perfiles de tratamiento psicológico, estudiaremos la variable "Sanitario", mientras que para los perfiles de tratamiento farmacológico nos centraremos en la variable "Sexo"

Perfil de tratamiento psicológico

Antes de comenzar con la variable "Sanitario", nos centramos en las variables objetivos relacionadas únicamente con el tratamiento psicológico. Comprobamos, que tal y como cabría esperar, variables como "Hospitalización_Covid" o "Ataques_Ansiedad" (esta última referente a la tenencia de ataques de ansiedad postpandemia) resultan relevantes únicamente para los perfiles de tratamiento que involucran momentos posteriores a la pandemia, ya que, al no estar basados en características previas a la pandemia (no se puede estar hospitalizado por Covid antes de la pandemia), no deberían afectar a los individuos antes de esta.

A continuación, profundizaremos en el análisis de cómo han variado los porcentajes de la variable de encuesta “Sanitario”, en la cual el valor 0 significa que no es sanitario/a, y el valor 1 que si lo es. Para comenzar este estudio de manera más profunda, analizaremos las tablas de contingencia entre la variable y las tres variables objetivo del tratamiento psicológico.

	psic_antes		
Sanitario	0	1	All
0	85.275	8.22	93.495
1	5.718	0.786	6.505
All	90.994	9.006	100.0

	psic_despues		
Sanitario	0	1	All
0	92.673	0.822	93.495
1	6.254	0.25	6.505
All	98.928	1.072	100.0

	psic_antes_y_despues		
Sanitario	0	1	All
0	92.423	1.072	93.495
1	6.219	0.286	6.505
All	98.642	1.358	100.0

Tabla 6. Tablas de contingencia de perfiles de tratamiento solo psicológicos. Fuente: Elaboración propia.

Antes de comentar como varían los porcentajes de los sanitarios que se medican, cabe destacar otro dato importante. Como vemos, el porcentaje de personas que estaban sometidas a tratamiento psicológico prepandemia era de un 10.37% del total de los encuestados (el 9.006% que solo estuvo prepandemia y el 1.354% que continuo tras ella), mientras que, tras la pandemia, solo continuaron con el tratamiento un 1.36% del total. Esto quiere decir que solo un 13% de aquellas personas que estaban en tratamiento antes de la pandemia lo continuaron tras ella. Esto supone un resultado sorprendente, y que comentaremos con más detalle en las conclusiones, donde especularemos sobre el posible motivo. Si continuamos observando las tablas, veremos que antes de la pandemia, los sanitarios suponen un 6.505% del total de los encuestados, mientras que aquellos que recibían tratamiento antes son un 1.072% (0.786% antes más 0.286% antes y después), esto implica que un 16% de los sanitarios estaban sometidos a tratamiento psicológico antes de la pandemia. Sin embargo, si calculamos el porcentaje de sanitarios que lo reciben tras la pandemia (sumando solo después y antes y después), el porcentaje de sanitarios que lo reciben es de un 8% del número de sanitarios total.

Vemos que al igual que en la población general, donde el porcentaje se había reducido de un 10.31% al 1.36% (solo un 13% continuaron con tratamiento), en los sanitarios este se reduce del 16% al 8%, implicando que un 50% de los sanitarios que estaban sometidos a uno, continuaron con su tratamiento. En primer lugar, se concluye que el porcentaje pre-pandemia de pacientes con tratamiento psicológico es superior en sanitarios que, en la población general, y, en segundo lugar, aunque siguen los sanitarios la tendencia general de disminuir el porcentaje de personas sometidas a este tratamiento, en ellos el ritmo de abandono de esta terapia es mucho menor, concretamente para la población general solo un 13% de los pacientes la continuaron, y en sanitarios lo hicieron un 50%.

Perfil de tratamiento farmacológico

Ahora vamos a analizar, para los perfiles de tratamiento farmacológico, la variable "Sexo". El motivo tras usar esta característica es que aparece como relevante para los tres tipos de tratamiento farmacológico, por lo que será interesante comprobar si el paso de la pandemia afectó a como se distribuyen los tratamientos entre ambos sexos.

Seguiremos el mismo enfoque que en estudio anterior, obtener las tablas de contingencia, y a partir de ellas obtener resultados relevantes. En ellas el sexo "Mujer" esta codificado con un 0, y el sexo "Hombre" con un 1.

	farm_antes		
Sexo	0	1	All
0	48.64	1.32	49.96
1	49.57	0.46	50.04
All	98.21	1.79	100.0

	farm_despues		
Sexo	0	1	All
0	49.0	0.96	49.96
1	49.61	0.43	50.04
All	98.61	1.39	100.0

	farm_antes_y_despues		
Sexo	0	1	All
0	49.5	0.46	49.96
1	49.96	0.07	50.04
All	99.46	0.54	100.0

Tabla 7. Tablas de contingencia de perfiles de tratamiento solo farmacológicos. Fuente: Elaboración propia.

En primer lugar, comprobamos la existencia de la misma tendencia que en el caso del tratamiento psicológico, la población entrevistada general ha interrumpido el tratamiento farmacológico tras la pandemia, del 2.33% (1,79% antes y 0.54% antes y después) que estaba sometido al tratamiento antes del Covid, solo un 0.54% lo ha mantenido, un 23% de los pacientes únicamente. Sin embargo, a diferencia del caso anterior, vemos que el porcentaje de gente que se comenzó a someter a tratamiento farmacológico tras la pandemia es comparable al que había antes de ella, provocando que un 1.93% de los entrevistados estuviese sometido a tratamiento farmacológico, un ligero descenso sobre el 2.33% presente antes de la pandemia. En cuanto a cómo afecto la pandemia a la distribución del tratamiento en cada sexo, observamos que tanto antes, como después del covid, las mujeres estaban sometidas al tratamiento farmacológico en mayor proporción, y el porcentaje se mantiene similar tanto para los perfiles de tratamiento solo antes de la pandemia como solo después (un 1.32% frente a un 0.46%, y un 0.96% frente a un 0.43% respectivamente). Para el perfil de tratamiento farmacológico antes y después del covid, la diferencia de relación de porcentajes entre mujeres y hombres es más acentuada aún, pero al tratarse de porcentajes tan pequeños (y representar tan pocos registros) no lo tomaremos como una medida relevante. Un punto que debimos tener en cuenta antes de realizar este estudio era la distribución de los sexos, ya que, si las categorías "Hombre" y "Mujer" estuviesen desbalanceadas, este estudio por porcentajes que acabamos de realizar no tendría validez. En nuestro caso, representan un 49.96% y un 50.04%, por lo que podemos asumir las categorías como equitativas y dar por válidas las conclusiones que obtengamos. Como primer resultado cabe destacar la elevada tasa de abandonos del tratamiento farmacológico sobre el porcentaje de

personas tratándose que existían (solo un 23% de los pacientes continuaron con el tratamiento), mientras que asociado al sexo podemos destacar que el estudio refleja un porcentaje superior de mujeres bajo tratamiento farmacológico en cualquiera de los 3 tipos de perfil de tratamiento farmacológico.

Test de independencia T-Test

Este test solo se aplica sobre variables continuas, lo cual se corresponde únicamente con la edad dentro de nuestro set de datos. Uno de los requisitos de este test es que la variable continua que estemos analizando tenga una distribución normal, y como comprobamos en la figura X, esto se cumple. Recordamos también que este test de independencia se llevo a cabo para cada par "Edad-Variable Objetivo" de manera independiente, y que los dos conjuntos para los que se estudió la hipótesis nula de independencia fueron la edad cuando la variable objetivo equivalía a 0, y la edad cuando la variable objetivo equivalía a 1.

Los resultados obtenidos de este estudio los representaremos en las dos tablas **X, Y**, en las que incluiremos los estadísticos t de cada variable objetivo, y los p-values de cada test, que es la medida que nos interesa. Como se explicó en el apartado de desarrollo de proyecto, el p-value obtenido es una medida de la "fiabilidad" de la hipótesis nula asumida por el test, por lo que para aquellos valores de $\alpha < 0.05$ (alta significancia estadística), consideramos que esta no se cumple, y que si lo hace la hipótesis alternativa (dependencia entre los conjuntos).

Variables	psic_antes	psic_despues	psic_antes_y_despues	farm_antes	farm_despues	farm_antes_y_despues	nunca_psic_o_farm
Edad	-4.568	-4.14	-4.457	3.993	-0.282	1.798	3.691

Tabla 8. T-test scores de los perfiles de tratamientos frente a la edad.

Variables	psic_antes	psic_despues	psic_antes_y_despues	farm_antes	farm_despues	farm_antes_y_despues	nunca_psic_o_farm
Edad	<0.001	<0.001	<0.001	<0.001	0.778	0.072	<0.001

Tabla 9. P-Value test de independencia T-Test entre la edad y los diferentes perfiles de tratamiento

Observando los distintos niveles de tratamiento, se observa que los p-values de todos aquellos perfiles de tratamientos prepandemia poseen alta significancia estadística. Esto es un indicativo de que existe una relación entre la edad, y aquellos pacientes que se sometían a tratamiento únicamente antes la pandemia. Para el caso del tratamiento psicológico, vemos que este comportamiento se mantiene tras la pandemia, la edad sigue siendo un factor relevante sobre el recibir o no tratamiento psicológico, sin embargo, para los perfiles de tratamiento farmacológico, vemos como la relevancia estadística de la edad disminuye, en especial en el tratamiento únicamente posterior al covid. Esto nos da una clara indicación de que la edad deja de ser tan relevante para recibir tratamiento farmacológico tras el covid.

Para observar los cambios en las distribuciones de la edad de las personas que recibieron tratamientos de una manera más visual, vamos a generar dos diagramas de cajas (boxplots), uno para los tres tipos de tratamientos solo psicológicos, y otra para los otros tres tipos de tratamiento solo farmacológico. De forma que no se visualizaran

en estos gráficos la distribución de edades de aquellos individuos que no estaban sometido a alguno de estos tratamientos.

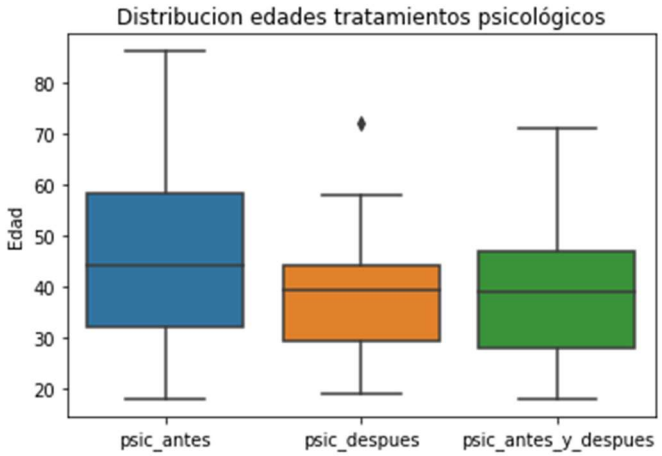


Ilustración 3.Diagrama de cajas de la edad para cada tipo de tratamiento solo psicológico.

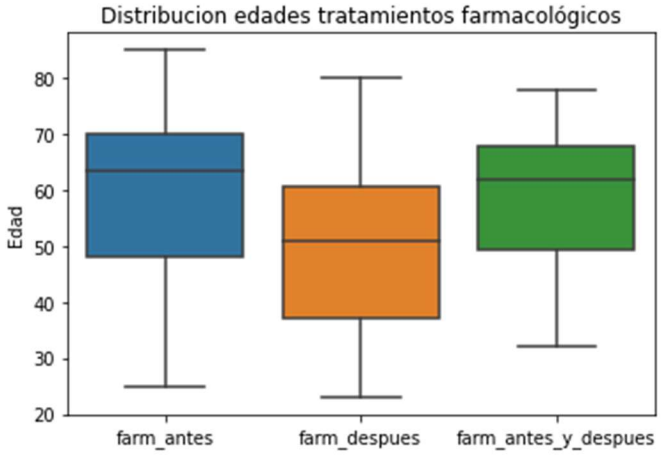


Ilustración 4. Diagrama de cajas de la edad para cada tipo de tratamiento solo farmacológico.

Para los tratamientos psicológicos, comprobamos como la distribución de edades es algo más alargada, sin embargo, vemos que las medias (representadas por la raya situada en el interior de cada cuadrado) se sitúan a una altura similar, aunque se aprecia que para los dos tratamientos que implican momentos posteriores a la pandemia (naranja y verde), esta edad media ha disminuido ligeramente.

Para los tratamientos farmacológicos vemos que si existe una mayor diferencia entre los perfiles. Mientras que los individuos que realizaron tratamiento solo antes, o antes y después, poseen una distribución de edades muy similar, aquellos que recibieron solo tratamiento después de la pandemia poseen una edad media significativamente menor, y en general la distribución de las edades esta más centrada en edades menores, implicando que la población que se sometió a tratamiento farmacológico justo después del covid no se correspondía con personas de edad en su mayoría (> 60 años), a diferencia que en los otros dos perfiles de tratamiento.

Si comparamos ahora las gráficas de tratamientos psicológicos y farmacológicos, comprobamos que, para la primera, las edades están desplazadas más hacia abajo, indicando que os individuos que se someten a tratamiento psicológico, por norma general, son más jóvenes que aquellos que se someten a tratamiento farmacológico.

4.3.2. Similitud entre variables

En este estudio comprobaremos por dos métricas diferentes, la similitud entre los distintos perfiles de tratamientos que poseemos en nuestros datos. Siguiendo el criterio aplicado anteriormente, las variables referentes al tratamiento psicológico y farmacológico al mismo tiempo no están incluidas en esta sección, dado que dificultan la legibilidad de las tablas, debido al elevado número de columnas, pero se pueden consultar en el apéndice del trabajo. Estas dos métricas de similitud son, la similitud coseno y la distancia euclídea.

Similitud coseno

Tras realizar el cálculo pertinente se obtuvo la *Tabla 10* con los valores de la similitud coseno entre las columnas que representaban las relevancias de las variables de encuesta para cada variable objetivo.

Variables	psic_antes	psic_despues	psic_antes_y_despues	farm_antes	farm_despues	farm_antes_y_despues	nunca_psic_o_farm
psic_antes	1	0.365	0.408	0.365	<0.001	0.183	0.468
psic_despues	0.365	1	0.596	<0.001	0.316	0.2	0.41
psic_antes_y_despues	0.408	0.596	1	0.149	0.471	0.298	0.688
farm_antes	0.365	<0.001	0.149	1	0.316	0.6	0.308
farm_despues	<0.001	0.316	0.471	0.316	1	0.632	0.324
farm_antes_y_despues	0.183	0.2	0.298	0.6	0.632	1	0.41
nunca_psic_o_farm	0.468	0.41	0.688	0.308	0.324	0.41	1

Tabla 10. Similitudes coseno entre las características relevantes para los distintos perfiles de tratamiento.

Como vemos en ella, se trata de una matriz simétrica, ya que se obtendrán los mismos valores si calculamos la similitud coseno entre dos variables objetivos, pero en ordenes diferentes (similitud [psic_antes-psic_despues] = similitud [psic_despues-psic_antes]).

Cuanto menor sea el valor entre dos perfiles de tratamiento, más diferentes serán estos en cuanto a cuales son las variables de encuesta relevantes para ellos. Esto no quiere decir que, si dos perfiles tienen un alto valor de similitud, las variables de encuesta que son relevantes para ellos tengan la misma distribución de valores. Podría darse el caso que, por ejemplo, para “farm_antes” y “farm_despues” la variable de encuesta “Sexo” fuera relevante, pero para “farm_antes”, los valores positivos se relacionen con ser mujer, y para “farm_despues” los valores positivos se relacionen con ser hombre, es por ello que esta medida no es una muestra fiable de cuanto se parecen dos perfiles de tratamiento, sino de cuan similares son en cuanto a que variables son relevantes para ambos.

Se comprueba que no existen en la tabla valores de similitud en el que esta supere el 0.8 entre dos variables objetivo diferentes, siendo el valor máximo de similitud entre

"nunca_psic_o_farm" y "psic_antes_y_despues". Nos será más relevante observar por tanto aquellos valores de la similitud que difieren más, con resultados menores a 0.2.

En este caso encontramos las parejas:

"psic_antes" y "farm_despues" < 0.001

"psic_antes" y "farm_antes_y_despues" = 0.183

"psic_despues" y "farm_antes" < 0.001

"psic_antes_y_despues" y "farm_antes" = 0.149

El patrón en estas bajas similitudes está claro, los perfiles de los pacientes asociados a tratamiento psicológico solo antes de la pandemia, o antes y después, difieren de aquellos pacientes que reciben tratamiento farmacológico después del covid en cuanto a que variables son relevantes para ellos.

Para poder ahora estudiar la relación entre perfiles de tratamiento, pero esta vez teniendo en cuenta como se distribuyen las categorías de las variables de encuesta para cada variable objetivo, usaremos la medida de distancia euclídea.

Distancia euclídea

La distancia euclídea mide la distancia entre dos vectores. El método para llevar a los valores que se mostraran en la *Tabla 11* está descrito en el apartado de "Desarrollo del proyecto", así que aquí pasaremos directamente a comentar los resultados. Cabe recordar que en este caso sí que se tienen en cuenta como se distribuyen las categorías de las variables de encuesta, por lo que este método para encontrar similitudes sí que nos sirve para determinar cuales son los perfiles de los encuestados más similares dependiendo del tratamiento al que están sometidos. Los valores más cercanos a cero se corresponden con aquellos perfiles más similares, mientras que cuanto mayor sean, más difieren.

index	psic_antes	psic_despues	psic_antes_y_despues	farm_antes	farm_despues	farm_antes_y_despues	nunca_psic_o_farm
psic_antes	0.0	51.195	49.562	46.436	49.014	54.427	449.855
psic_despues	51.195	0.0	2.448	5.416	2.788	3.657	500.189
psic_antes_y_despues	49.562	2.448	0.0	4.551	2.612	5.366	498.753
farm_antes	46.436	5.416	4.551	0.0	3.164	8.259	495.314
farm_despues	49.014	2.788	2.612	3.164	0.0	5.631	497.976
farm_antes_y_despues	54.427	3.657	5.366	8.259	5.631	0.0	503.466
nunca_psic_o_farm	449.855	500.189	498.753	495.314	497.976	503.466	0.0

Tabla 11. Distancia euclídea entre la distribución de categorías de las variables de encuesta para cada perfil de tratamiento.

En primer lugar, nos centraremos en aquellos valores más elevados. Estos se corresponden con todas las distancias calculadas entre el perfil "nunca_psic_o_farm" y cualquiera de los otros tratamientos, y la diferencia con cualquiera de las otras distancias euclídeas es muy elevada, la distancia más baja encontrada para el tratamiento "nunca_psic_o_farm" es casi 10 veces mayor que la siguiente distancia más alta presente en la tabla. Es por ello que se puede afirmar que aquellos encuestados que

declararon no estar sometidos a ningún tipo de tratamiento psicológico o farmacológico, tienen perfiles muy diferentes (aportaron respuestas diferentes a las preguntas de encuesta) a aquellos que afirmaron haber estado sometidos a alguno antes y/o después de la pandemia. Vemos que estos resultados no concuerdan con los expuestos en la *Tabla 10* (de similitud coseno), demostrando así que efectivamente el echo de tener similares características relevantes, no implicaba similitud entre perfiles de encuestados.

El siguiente tipo de tratamiento que posee valores de la distancia euclídea superior a los demás es el representado por la variable “psic_antes”, que posee distancias entre 46 y 55, mientras que en el resto de la tabla (sin contar “nunca_psic_o_farm”), las distancias oscilan entre 2 y 9, por lo que podremos afirmar que los pacientes que solo estuvieron sometidos a tratamiento psicológico antes de la pandemia poseen perfiles bastante diferentes a aquellos que recibieron cualquier otro tipo de tratamiento.

De resto comprobamos que las distancias euclídeas son bajas, especialmente en aquellos tratamientos que se desarrollan después de la pandemia, o antes y después de la pandemia, independientemente de si son tratamientos psicológicos o farmacológicos.

4.3.3. Selección de características

Como último paso de nuestro análisis aplicamos un algoritmo de selección de variables sobre nuestros datos, con el fin de realizar una ordenación de la relevancia de las características, través del método “f_classif”, que habíamos indicado ya que se basaba en métodos ANOVA.

En la *Tabla 12* se representan las relevancias, siendo mayor la importancia de una característica respecto a una variable objetivo, cuanto mayor sea el valor de la intersección de ambas en la tabla.

Variables	psic_antes	psic_despues	psic_antes_y_despues	farm_antes	farm_despues	farm_antes_y_despues	nunca_psic_o_farm
Sexo	1.441	3.386	5.255	11.806	5.881	8.144	61.399
Edad	20.866	17.144	19.864	15.943	0.08	3.233	13.623
Sanitario	2.256	14.183	13.461	0.187	0.092	4.52	8.788
Tenencia_Covid	0.186	2.116	0.679	1.59	0.092	0.331	0.013
Hospitalizacion_Covid	0.852	10.878	0.361	0.477	1.148	0.141	0.908
Fallecimiento_Familiar_por_Covid	0.593	1.147	16.257	0.018	4.926	0.57	5.306
Ataques_Ansiedad	1.036	40.76	36.392	0.295	52.726	11.541	240.312
Enfermedad_Cronica	2.476	0.961	0.015	2.62	0.595	2.183	49.674
Enfermedad_Cronica_Familiar	0.152	0.304	1.489	1.873	3.421	1.635	16.622
Enfermedad_Cardiovascular	0.392	0.513	0.314	0.025	0.371	1.37	0.121
Diabetes	1.243	1.454	1.847	2.441	0.88	0.723	3.245
Cancer	0.403	0.687	1.651	0.011	0.896	1.378	5.646
Enfermedad_Respiratoria	0.181	0.047	7.167	1.692	3.644	0.92	10.881
Enfermedad_Traumatologica	0.208	1.442	0.323	3.506	2.985	0.157	8.984
Enfermedad_Autoinmune	0.153	0.033	0.003	0.116	0.007	0.837	4.903
Otra_Enfermedad_Cronica	0.949	2.157	0.988	11.776	0.001	1.072	8.204
Trastorno_Sueño	2.693	0.858	1.09	5.245	0.005	0.427	3.03
Tenencia_de_Hijos	0.007	2.353	0.312	2.329	0.205	0.006	0.244
Medicacion_Familiar_por_Salud_Mental	0.1	1.844	0.839	0.064	0.012	16.076	20.745
Situacion_Laboral	0.088	0.211	0.274	2.748	0.218	2.767	11.455
Situacion_Convivencia	0.22	0.191	2.646	1.706	0.083	3.209	1.525
Escolarizacion	0.188	0.285	0.362	0.478	0.372	0.142	0.501
Estudios	10.061	0.1	8.639	11.88	1.425	2.595	0.0
Religion	10.082	3.187	1.477	6.671	1.152	0.375	0.182
Clase_Social	1.204	1.17	1.353	0.215	2.724	4.325	2.058
Trastorno_Depresivo_Antes	315.12	2.197	4.998	3.69	2.866	1.092	898.88
Trastorno_Ansioso_Antes	481.571	1.851	207.293	3.109	2.415	0.92	721.139
Trastorno_Depresivo_Despues	6.493	43.195	31.757	1.191	0.925	0.353	238.556
Trastorno_Ansioso_Despues	7.853	277.9	410.408	1.44	1.119	0.427	293.617

Tabla 12. Scores asociados a las relevancias de cada variable de encuesta.

Si fijamos nuestra atención en las últimas 4 filas de la *Tabla 12*, podremos comprobar que los valores de estas son muy superiores a los presentes en el resto de las variables de encuesta, aunque las 4 no lo sean para todos los perfiles de tratamiento. A continuación, vamos a incorporar una tabla con estos mismos valores, pero normalizados, siendo este proceso de normalización el producto de dividir cada columna, por el valor más alto encontrado en ella. Se realiza la normalización con el objetivo de realizar una comparación más sencilla entre las filas de la tabla, sin embargo, si esto lo hacemos mientras las variables de los trastornos siguen presentes, provocará que no seamos capaces de distinguir para el resto de variables de encuesta cuales son las más relevantes, dado que al dividir entre un valor que será mucho más elevado, las demás devolverán un resultado muy cercano a cero. Es por ello que antes de realizar la normalización, repetiremos el proceso, sin tener en cuenta estas últimas 4 variables, aunque a la hora de analizar los resultados tendremos presente en todo momento que estas variables han sido eliminadas de la *Tabla 13*, en la que se indican los resultados de las relevancias de manera normalizada.

Variables	psic_antes	psic_despues	psic_antes_y_despues	farm_antes	farm_despues	farm_antes_y_despues	nunca_psic_o_farm
Sexo	0.069	0.083	0.144	0.741	0.112	0.507	0.255
Edad	1.0	0.421	0.546	1.0	0.002	0.201	0.057
Sanitario	0.108	0.348	0.37	0.012	0.002	0.281	0.037
Tenencia_Covid	0.009	0.052	0.019	0.1	0.002	0.021	0.0
Hospitalizacion_Covid	0.041	0.267	0.01	0.03	0.022	0.009	0.004
Fallecimiento_Familiar_por_Covid	0.028	0.028	0.447	0.001	0.093	0.035	0.022
Ataques_Ansiedad	0.05	1.0	1.0	0.019	1.0	0.718	1.0
Enfermedad_Cronica	0.119	0.024	0.0	0.164	0.011	0.136	0.207
Enfermedad_Cronica_Familiar	0.007	0.007	0.041	0.117	0.065	0.102	0.069
Enfermedad_Cardiovascular	0.019	0.013	0.009	0.002	0.007	0.085	0.001
Diabetes	0.06	0.036	0.051	0.153	0.017	0.045	0.014
Cancer	0.019	0.017	0.045	0.001	0.017	0.086	0.023
Enfermedad_Respiratoria	0.009	0.001	0.197	0.106	0.069	0.057	0.045
Enfermedad_Traumatologica	0.01	0.035	0.009	0.22	0.057	0.01	0.037
Enfermedad_Autoinmune	0.007	0.001	0.0	0.007	0.0	0.052	0.02
Otra_Enfermedad_Cronica	0.045	0.053	0.027	0.739	0.0	0.067	0.034
Trastorno_Sueño	0.129	0.021	0.03	0.329	0.0	0.027	0.013
Tenencia_de_Hijos	0.0	0.058	0.009	0.146	0.004	0.0	0.001
Medicacion_Familiar_por_Salud_Mental	0.005	0.045	0.023	0.004	0.0	1.0	0.086
Situacion_Laboral	0.004	0.005	0.008	0.172	0.004	0.172	0.048
Situacion_Convivencia	0.011	0.005	0.073	0.107	0.002	0.2	0.006
Escolarizacion	0.009	0.007	0.01	0.03	0.007	0.009	0.002
Estudios	0.482	0.002	0.237	0.745	0.027	0.161	0.0
Religion	0.483	0.078	0.041	0.418	0.022	0.023	0.001
Clase_Social	0.058	0.029	0.037	0.013	0.052	0.269	0.009

Tabla 13. Scores de las relevancias de cada característica normalizados

En primer lugar, vemos como para el tratamiento psicológico y farmacológico antes de la pandemia, la edad resultaba el factor más determinante excluyendo el haber sufrido trastorno ansioso o depresivo antes del covid, mientras que para cualquier tratamiento que implicara un momento posterior a la pandemia, esta ya dejaba de ser la principal característica relevante. Este comportamiento se acentúa especialmente cuando vemos el cambio de relevancia entre el perfil de tratamiento farmacológico solo antes y solo después del covid. Este comportamiento concuerda con el análisis de resultados que habíamos realizado mediante el T-Test sobre la variable de la Edad. Si observamos otras variables en busca de cambios en la relevancia según los distintos campos de tratamiento, cabe destacar también la característica “Sexo”, que para los tratamientos psicológicos no resulta demasiado relevante en comparación con otras, mientras que para los tratamientos farmacológicos si que es un factor a tener en cuenta, siendo en la columna “farm_antes” la tercera característica más relevante (sin contar las 4 descartadas), solo tras “Edad” y “Estudios”.

En la variable objetivo “farm_antes_y_despues”, la característica más relevante es “Medicacion_Familiar_por_Salud_Mental”, y se trata este del único perfil de tratamiento en el que esta característica resulta relevante. Deducimos por tanto que poseer familiares que están consumiendo medicación para tratar un problema de salud mental, se relaciona con consumir fármacos antes y después de la pandemia. Para ver este comportamiento en más detalle, observaremos la tabla de contingencia (*Tabla 14)* de esta característica y el perfil de tratamiento que estamos mencionando.

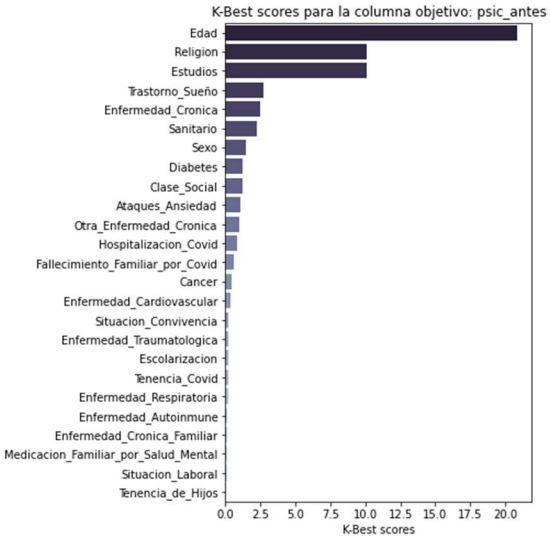
	farm_antes_y_despues		
Medicacion_Familiar	0	1	All
0	94.853	0.393	95.247
1	4.61	0.143	4.753
All	99.464	0.536	100.0

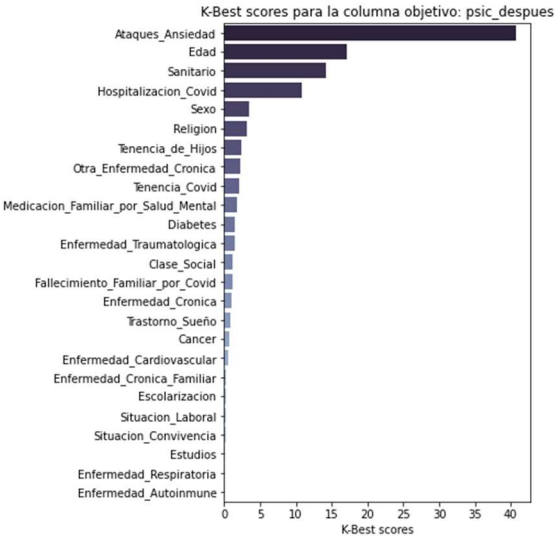
Tabla 14. Tabla de contingencia entre variables Medicacion_Familiar_por_Salud_Mental y farm_antes_y_despues.

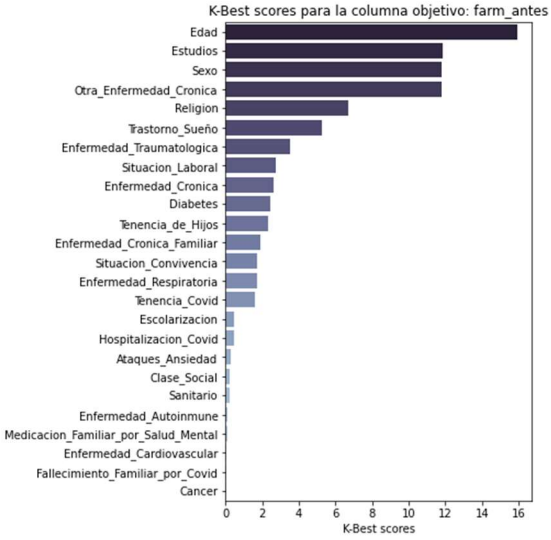
Primero observamos que el porcentaje de encuestados con familiares medicándose es de un 4.753%, y solo un 0.143% tienes familiares medicándose y además reciben medicación ellos mismos. Esto implica que un 3% de los encuestados con familiares bajo medicación también la reciben. Si ahora repetimos el análisis para aquellos sujetos que no tienen familiares medicándose, se comprueba que un 0.393% de los encuestados toman fármacos sin tener familiares haciéndolo. Esto señala que, del 95.247% que representan los que no tienen familiares bajo este tratamiento, solo un 0.4% consumieron fármacos antes y después de la pandemia. Comparando las dos opciones, comprobamos que el porcentaje de encuestados que recibió tratamiento farmacológico antes y después de la pandemia entre los que un familiar suyo se medicaba (3%), es significativamente superior al porcentaje que recibió el mismo tratamiento sin tener ningún familiar medicándose (0.4%), por lo que efectivamente resulta una característica significativa a la hora de recibir o no el tratamiento.

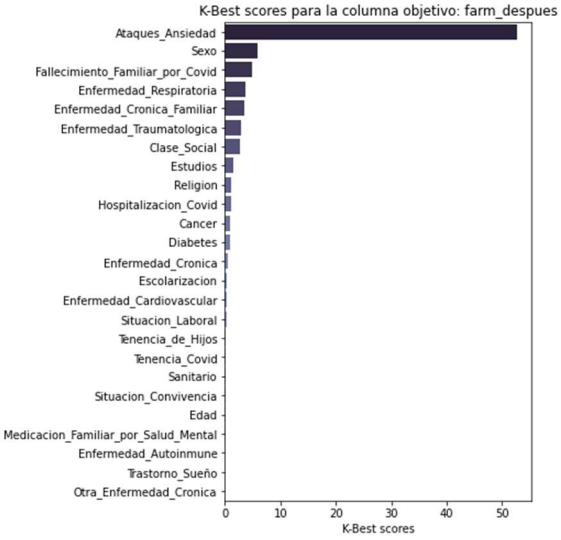
Si analizamos ahora las variables que habíamos descartado, en la *Tabla 12* podemos ver como para los tratamientos psicológicos, haber sufrido o no trastorno ansioso y/o trastorno depresivo antes del covid, resultaba de mucha relevancia respecto a la variable relacionada con haber recibido tratamiento psicológico antes, mientras que haber sufrido estos trastornos después de la pandemia, resulta de mucha importancia respecto a haber recibido tratamiento psicológico después de la pandemia. También comprobamos que el trastorno de ansiedad, sea antes o después del covid, posee mucho peso a la hora de haber recibido tratamiento psicológico antes y después de la pandemia.

Ahora vamos a representar estas relevancias de manera gráfica para que se puedan analizar de una manera más sencilla estos resultados que hemos comentado, generando una gráfica por cada nivel de tratamiento, y ordenando las variables de encuesta en ella de más relevantes a menos. Al igual que en el caso de la *Tabla 13* de relevancias normalizadas, aquí no incluiremos las últimas cuatro características ya que poseen una relevancia demasiado elevada, y nos provocarían una pérdida de información visual.


K-Best scores para la columna objetivo: psic_antes
Edad
Religion
Estudios
Trastorno_Sueño
Enfermedad_Cronica
Sanitario
Sexo
Diabetes
Clase_Social
Ataques_Ansiedad
Otra_Enfermedad_Cronica
Hospitalizacion_Covid
Fallecimiento_Familiar_por_Covid
Cancer
Enfermedad_Cardiovascular
Situacion_Convivencia
Enfermedad_Traumatologica
Escolarizacion
Tenencia_Covid
Enfermedad_Respiratoria
Enfermedad_Autoinmune
Enfermedad_Cronica_Familiar
Medicacion_Familiar_por_Salud_Mental
Situacion_Laboral
Tenencia_de_Hijos
0.0 2.5 5.0 7.5 10.0 12.5 15.0 17.5 20.0
K-Best scores


K-Best scores para la columna objetivo: psic_despues
Ataques_Ansiedad
Edad
Sanitario
Hospitalizacion_Covid
Sexo
Religion
Tenencia_de_Hijos
Otra_Enfermedad_Cronica
Tenencia_Covid
Medicacion_Familiar_por_Salud_Mental
Diabetes
Enfermedad_Traumatologica
Clase_Social
Fallecimiento_Familiar_por_Covid
Enfermedad_Cronica
Trastorno_Sueño
Cancer
Enfermedad_Cardiovascular
Enfermedad_Cronica_Familiar
Escolarizacion
Situacion_Laboral
Situacion_Convivencia
Estudios
Enfermedad_Respiratoria
Enfermedad_Autoinmune
0 5 10 15 20 25 30 35 40
K-Best scores


K-Best scores para la columna objetivo: farm_antes
Edad
Estudios
Sexo
Otra_Enfermedad_Cronica
Religion
Trastorno_Sueño
Enfermedad_Traumatologica
Situacion_Laboral
Enfermedad_Cronica
Diabetes
Tenencia_de_Hijos
Enfermedad_Cronica_Familiar
Situacion_Convivencia
Enfermedad_Respiratoria
Tenencia_Covid
Escolarizacion
Hospitalizacion_Covid
Ataques_Ansiedad
Clase_Social
Sanitario
Enfermedad_Autoinmune
Medicacion_Familiar_por_Salud_Mental
Enfermedad_Cardiovascular
Fallecimiento_Familiar_por_Covid
Cancer
0 2 4 6 8 10 12 14 16
K-Best scores


K-Best scores para la columna objetivo: farm_despues
Ataques_Ansiedad
Sexo
Fallecimiento_Familiar_por_Covid
Enfermedad_Respiratoria
Enfermedad_Cronica_Familiar
Enfermedad_Traumatologica
Clase_Social
Estudios
Religion
Hospitalizacion_Covid
Cancer
Diabetes
Enfermedad_Cronica
Escolarizacion
Enfermedad_Cardiovascular
Situacion_Laboral
Tenencia_de_Hijos
Tenencia_Covid
Sanitario
Situacion_Convivencia
Edad
Medicacion_Familiar_por_Salud_Mental
Enfermedad_Autoinmune
Trastorno_Sueño
Otra_Enfermedad_Cronica
0 10 20 30 40 50
K-Best scores

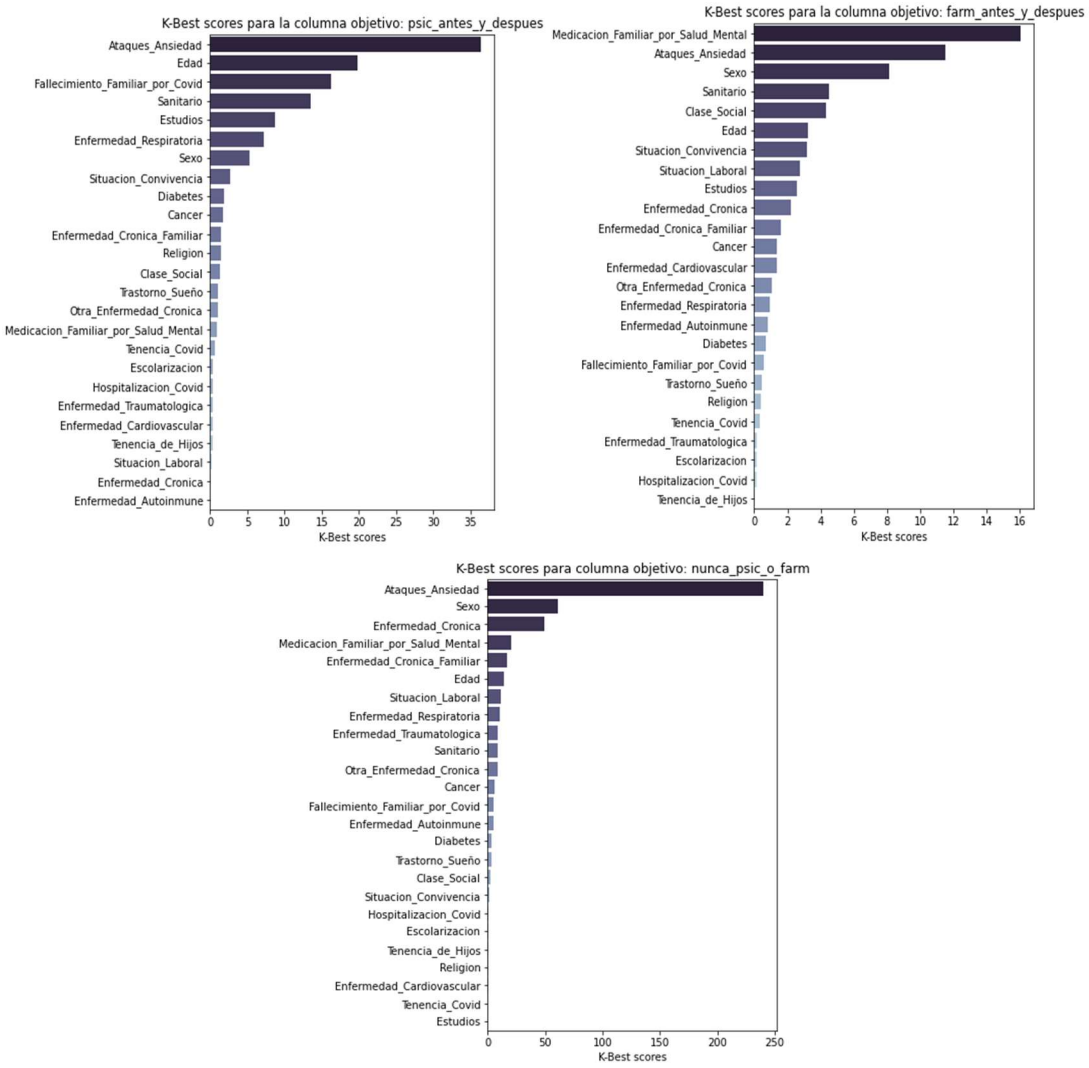


Ilustración 5. Scores de relevancia de cada variable de encuesta, para cada perfil de tratamiento

Tras estas últimas representaciones (*Ilustración 6*) poseemos una visión clara de como se ordenan las características en función de sus relevancias frente a cada tipo de tratamiento.

5. Conclusión y trabajos futuros

Conclusiones

En este trabajo hemos realizado una caracterización de perfiles asociados a distintos grados de tratamientos, psicológicos y/o farmacológicos con el objetivo de determinar el impacto de la pandemia sobre la salud mental de la población española. A continuación hemos comparado estos perfiles de tratamiento para obtener las similitudes entre ellos, y por último realizamos una selección de características para determinar cuales eran las más relevantes para cada tipo de tratamiento.

En nuestro resultados confirmamos la presencia de diferencias significativas entre las características de aquellas personas sometidas a tratamientos antes de la pandemia y después de esta, donde sobre todo es destacable esta diferencia para el caso de aquellas personas que estaban sometidas a tratamiento farmacológico antes de la pandemia y aquellas lo estuvieron después de la pandemia. Comprobamos como para estos casos, la edad de los encuestados que consumían los fármacos pasaba de ser en general de edades avanzadas (<65 años), a una distribuciones de edad más centradas en edades medias (en torno a los 50 años), demostrando como el estado de confinamiento y todas las consecuencias del Covid provocaron un acceso a los fármacos para tratar problemas de salud mental sin ser tan relevante la edad. También se comprobó como factores como el sexo, o la muerte de familiares cercanos podía ser un factor influyente a la hora de consumir estos medicamentos, y como los trabajadores de la salud, los sanitarios, poseen una mayor vulnerabilidad, provocando que posean índices de tratamiento psicológicos superiores al resto de la población.

Una conclusión interesante alejada de los problemas ocasionados por la pandemia, es la mayor tendencia de la población encuestada a consumir fármacos (antes y después del covid) si algún familiar cercano los consume, probando que el ambiente familiar puede ser un factor relevante ante este tipo de tratamientos. También fue probado como el hecho de sufrir trastornos ansiosos o depresivos antes y/o después de la pandemia estaba directamente relacionado con la recepción de tratamientos psicológicos.

Un factor a destacar también es la enorme diferencia (encontrada a través de los métodos de similitudes) de perfiles entre aquellos que han recibido algún tipo de tratamiento en cualquier momento, y aquellos que no lo habían recibido nunca, sobre todo en cuanto a como se distribuyen sus respuestas para cada pregunta (analizado a través de la distancia euclídea), indicándonos que el echo de que cierta parte de la población este sometidas a tratamientos y otra parte no, no se debe a la casualidad, sino a ciertos factores sociodemográficos y de otros tipos.

Como conclusión más relevante, y preocupante, de este estudio podemos sacar que los tratamientos psicológicos y farmacológicos han sufrido un cambio debido a la pandemia en cuanto a que tipo de población los solicita, siendo la actual significativamente más joven que antes del Covid, sobre todo para el caso del tratamiento farmacológico.

Limitaciones y trabajos futuros

Una de las principales limitaciones de este proyecto es el tamaño de los datos utilizados, dado que al ser una muestra de en torno a 3000 personas, muchas de las respuestas a las preguntas no tienen mucha representación, provocando que los estudios aplicados sobre ellas puedan no ser representativos de la realidad de una población como la española. Un ejemplo de esto pueden ser el bajo porcentaje de sanitarios presente en la encuesta, o los muchos trastornos que tuvieron que descartarse en el preproceso por no suponer su representación más del 1% del total de los registros.

Como trabajo futuro queda la aplicación de árbol de decisión para comprobar si las características más relevantes obtenidas por este algoritmo de Machine Learning se corresponden con las resultantes del análisis estadístico que hemos llevado a cabo.

6. Referencia

1. Organización Mundial de la Salud (s.f). WHO Coronavirus (COVID-19) Dashboard. Recuperado el 03 de octubre de 2022 de https://covid19.who.int

2. Holmes, E. A., O'Connor, R. C., Perry, V. H., Tracey, I., Wessely, S., Arseneault, L., Ballard, C., Christensen, H., Cohen Silver, R., Everall, I., Ford, T., John, A., Kabir, T., King, K., Madan, I., Michie, S., Przybylski, A. K., Shafran, R., Sweeney, A., Worthman, C. M., … Bullmore, E. (2020). Multidisciplinary research priorities for the COVID-19 pandemic: a call for action for mental health science. The lancet. Psychiatry, 7(6), 547–560. doi: 10.1016/S2215-0366(20)30168-11

3. Elovainio, M., Hakulinen, C., Pulkki-Råback, L., Virtanen, M., Josefsson, K., Jokela, M., Vahtera, J., & Kivimäki, M. (2017). Contribution of risk factors to excess mortality in isolated and lonely individuals: an analysis of data from the UK Biobank cohort study. The Lancet. Public health, 2(6), e260–e266. doi: 10.1016/S2468-2667(17)30075-0

4. Rodgers, R. F., Lombardo, C., Cerolini, S., Franko, D. L., Omori, M., Fuller-Tyszkiewicz, M., Linardon, J., Courtet, P., & Guillaume, S. (2020). The impact of the COVID-19 pandemic on eating disorder risk and symptoms. The International journal of eating disorders, 53(7), 1166–1170. doi: 10.1002/eat.23318

5. Zhu, H., Xie, S., Liu, X., Yang, X., & Zhou, J. (2022). Influencing factors of burnout and its dimensions among mental health workers during the COVID-19 pandemic. Nursing Open, 9(4), 2013-2023. doi: 10.1002/nop2.1211

6. Wang, C., Pan, R., Wan, X., Tan, Y., Xu, L., Ho, C. S., & Ho, R. C. (2020). Immediate Psychological Responses and Associated Factors during the Initial Stage of the 2019 Coronavirus Disease (COVID-19) Epidemic among the General Population in China. International journal of environmental research and public health, 17(5), 1729. doi: 10.3390/ijerph17051729

7. Hawryluck, L., Gold, W. L., Robinson, S., Pogorski, S., Galea, S., & Styra, R. (2004). SARS control and psychological effects of quarantine, Toronto, Canada. Emerging infectious diseases, 10(7), 1206–1212. doi: 10.3201/eid1007.030703

8. Fayyad, U., Piatetsky-Shapiro, G., & Smyth P. (1996). The KDD process for extracting useful knowledge from volumes of data. Commun. ACM 39, 11 (Nov. 1996), 27–34. doi: 10.1145/240455.240464

9. Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. AI Magazine, 17(3), 37. doi: 10.1609/aimag.v17i3.1230

10. Clifton, C. (2022). data mining. Encyclopedia Britannica. https://www.britannica.com/technology/data-mining

11. Schervish, M. J. (1996). P Values: What They Are and What They Are Not. The American Statistician, 50(3), 203–206. doi: 10.2307/2684655

Apéndice I

Añadimos en esta sección la versión completa de las tablas expuestas en los resultados, que no pudieron incluirse en el cuerpo de la memoria principal porque el elevado número de columnas dificultaba su visibilidad. También añadiremos aquí las gráficas con los valores de las relevancias de la selección de características para las variables objetivo que no incluimos en la memoria principal.

Variables	psic_antes	psic_despues	psic_antes_y_despues	farm_antes	farm_despues	farm_antes_y_despues	psic_y_farm_antes	psic_y_farm_despues	psic_y_farm_antes_y_despues	nunca_psic_o_farm
Sexo	0.257	0.098	0.033	0.001	0.024	0.01	0.022	0.036	0.014	0.0
Sanitario	0.171	0.001	0.001	0.886	1.0	0.11	0.768	0.394	0.192	0.004
Tenencia_Covid	0.752	0.256	0.589	0.311	0.982	0.901	0.896	1.0	0.977	0.973
Hospitalizacion_Covid	0.562	0.019	1.0	1.0	0.817	1.0	1.0	1.0	0.803	0.472
Fallecimiento_Familiar_por_Covid	0.552	0.561	0.0	1.0	0.074	0.948	1.0	0.46	0.92	0.029
Ataques_Ansiedad	0.354	0.0	0.0	0.73	0.0	0.002	0.007	0.0	0.0	0.0
Enfermedad_Cronica	0.133	0.431	1.0	0.142	0.547	0.228	0.0	0.053	0.0	0.0
Enfermedad_Cronica_Familiar	0.753	0.733	0.301	0.228	0.096	0.326	0.947	0.084	0.031	0.0
Enfermedad_Cardiovascular	0.604	0.675	0.767	1.0	0.727	0.452	0.003	1.0	0.816	0.783
Diabetes	0.339	0.443	0.333	0.222	0.581	0.818	1.0	1.0	0.852	0.09
Cancer	0.681	0.837	0.465	1.0	0.69	0.768	0.004	1.0	0.704	0.026
Enfermedad_Respiratoria	0.777	1.0	0.02	0.32	0.118	0.686	0.64	0.29	0.105	0.001
Enfermedad_Traumatologica	0.766	0.447	0.86	0.126	0.18	1.0	0.001	1.0	0.164	0.004
Enfermedad_Autoinmune	0.849	1.0	1.0	1.0	1.0	0.898	0.368	0.275	0.015	0.038
Otra_Enfermedad_Cronica	0.445	0.378	0.637	0.003	1.0	0.836	0.065	1.0	0.105	0.007
Trastorno_Sueño	0.15	0.715	0.586	0.064	1.0	1.0	0.02	1.0	1.0	0.109
Tenencia_de_Hijos	0.993	0.184	0.707	0.172	0.786	1.0	0.276	0.695	0.723	0.658
Medicacion_Familiar_por_Salud_Menta	0.871	0.354	0.594	1.0	1.0	0.001	1.0	0.0	0.0	0.0
Situacion_Laboral	0.015	0.085	0.087	0.023	0.479	0.0	0.001	0.031	0.0	0.0
Situacion_Convivencia	0.06	0.844	0.02	0.52	0.571	0.592	0.121	0.795	0.403	0.001
Escolarizacion	0.805	0.853	0.817	0.766	0.813	0.924	0.088	0.886	0.817	0.221
Estudios	0.052	0.294	0.128	0.005	0.258	0.0	0.565	0.73	0.01	0.823
Religion	0.017	0.328	0.455	0.014	0.651	0.634	0.107	0.256	0.068	0.502
Clase_Social	0.271	0.215	0.189	0.353	0.683	0.159	0.221	0.148	0.776	0.033
Trastorno_Depresivo_Antes	0.0	0.264	0.056	0.102	0.17	0.597	0.0	0.379	0.0	0.0
Trastorno_Ansioso_Antes	0.0	0.334	0.0	0.145	0.227	0.686	0.0	0.459	0.0	0.0
Trastorno_Depresivo_Despues	0.02	0.0	0.0	0.539	0.672	1.0	0.318	0.0	0.0	0.0
Trastorno_Ansioso_Despues	0.009	0.0	0.0	0.445	0.572	1.0	0.243	0.0	0.0	0.0

Tabla 15. P-Values Test de independencia estadística Chi Square (completa).

Variables	psic_antes	psic_despues	psic_antes_y_despues	farm_antes	farm_despues	farm_antes_y_despues	psic_y_farm_antes	psic_y_farm_despues	psic_y_farm_antes_y_despues	nunca_psic_o_farm
Sexo	0.257	0.098	0.033	0.001	0.024	0.01	0.022	0.036	0.014	0.0
Sanitario	0.171	0.001	0.001	0.886	1.0	0.11	0.768	0.394	0.192	0.004
Tenencia_Covid	0.752	0.256	0.589	0.311	0.982	0.901	0.896	1.0	0.977	0.973
Hospitalizacion_Covid	0.562	0.019	1.0	1.0	0.817	1.0	1.0	1.0	0.803	0.472
Fallecimiento_Familiar_por_Covid	0.552	0.561	0.0	1.0	0.074	0.948	1.0	0.46	0.92	0.029
Ataques_Ansiedad	0.354	0.0	0.0	0.73	0.0	0.002	0.007	0.0	0.0	0.0
Enfermedad_Cronica	0.133	0.431	1.0	0.142	0.547	0.228	0.0	0.053	0.0	0.0
Enfermedad_Cronica_Familiar	0.753	0.733	0.301	0.228	0.096	0.326	0.947	0.084	0.031	0.0
Enfermedad_Cardiovascular	0.604	0.675	0.767	1.0	0.727	0.452	0.003	1.0	0.816	0.783
Diabetes	0.339	0.443	0.333	0.222	0.581	0.818	1.0	1.0	0.852	0.09
Cancer	0.681	0.837	0.465	1.0	0.69	0.768	0.004	1.0	0.704	0.026
Enfermedad_Respiratoria	0.777	1.0	0.02	0.32	0.118	0.686	0.64	0.29	0.105	0.001
Enfermedad_Traumatologica	0.766	0.447	0.86	0.126	0.18	1.0	0.001	1.0	0.164	0.004
Enfermedad_Autoinmune	0.849	1.0	1.0	1.0	1.0	0.898	0.368	0.275	0.015	0.038
Otra_Enfermedad_Cronica	0.445	0.378	0.637	0.003	1.0	0.836	0.065	1.0	0.105	0.007
Trastorno_Sueño	0.15	0.715	0.586	0.064	1.0	1.0	0.02	1.0	1.0	0.109
Tenencia_de_Hijos	0.993	0.184	0.707	0.172	0.786	1.0	0.276	0.695	0.723	0.658
Medicacion_Familiar_por_Salud_Menta	0.871	0.354	0.594	1.0	1.0	0.001	1.0	0.0	0.0	0.0
Situacion_Laboral	0.015	0.085	0.087	0.023	0.479	0.0	0.001	0.031	0.0	0.0
Situacion_Convivencia	0.06	0.844	0.02	0.52	0.571	0.592	0.121	0.795	0.403	0.001
Escolarizacion	0.805	0.853	0.817	0.766	0.813	0.924	0.088	0.886	0.817	0.221
Estudios	0.052	0.294	0.128	0.005	0.258	0.0	0.565	0.73	0.01	0.823
Religion	0.017	0.328	0.455	0.014	0.651	0.634	0.107	0.256	0.068	0.502
Clase_Social	0.271	0.215	0.189	0.353	0.683	0.159	0.221	0.148	0.776	0.033
Trastorno_Depresivo_Antes	0.0	0.264	0.056	0.102	0.17	0.597	0.0	0.379	0.0	0.0
Trastorno_Ansioso_Antes	0.0	0.334	0.0	0.145	0.227	0.686	0.0	0.459	0.0	0.0
Trastorno_Depresivo_Despues	0.02	0.0	0.0	0.539	0.672	1.0	0.318	0.0	0.0	0.0
Trastorno_Ansioso_Despues	0.009	0.0	0.0	0.445	0.572	1.0	0.243	0.0	0.0	0.0

Tabla 16. Variable de encuesta categóricas relevantes para cada perfil de tratamiento (completa).

Variables	psic_antes	psic_despues	psic_antes_y_despues	farm_antes	farm_despues	farm_antes_y_despues	psic_y_farm_antes	psic_y_farm_despues	psic_y_farm_antes_y_despues	nunca_psic_o_farm
Edad	0.0	0.0	0.0	0.0	0.778	0.072	0.022	0.197	0.795	0.0

Tabla 17. P-Value test de independencia T-Test entre la edad y los diferentes perfiles de tratamiento (completa).

Variables	psic_antes	psic_despues	psic_antes_y_despues	farm_antes	farm_despues	farm_antes_y_despues	psic_y_farm_antes	psic_y_farm_despues	psic_y_farm_antes_y_despues	nunca_psic_o_farm
Edad	-4.568	-4.14	-4.457	3.993	-0.282	1.798	2.289	-1.289	-0.26	3.691

Tabla 18. T-test scores de los perfiles de tratamientos frente a la edad (completa).

Variables	psic_antes	psic_despues	psic_antes_y_despues	farm_antes	farm_despues	farm_antes_y_despues	psic_y_farm_antes	psic_y_farm_despues	psic_y_farm_antes_y_despues	nunca_psic_o_farm
psic_antes	1	0.365	0.408	0.365	0	0.183	0.387	0.5	0.589	0.468
psic_despues	0.365	1	0.596	0	0.316	0.2	0.141	0.548	0.387	0.41
psic_antes_y_despues	0.408	0.596	1	0.149	0.471	0.298	0.316	0.544	0.481	0.688
farm_antes	0.365	0	0.149	1	0.316	0.6	0.283	0.365	0.387	0.308
farm_despues	0	0.316	0.471	0.316	1	0.632	0.447	0.577	0.408	0.324
farm_antes_y_despues	0.183	0.2	0.298	0.6	0.632	1	0.424	0.73	0.645	0.41
psic_y_farm_antes	0.387	0.141	0.316	0.283	0.447	0.424	1	0.387	0.548	0.58
psic_y_farm_despues	0.5	0.548	0.544	0.365	0.577	0.73	0.387	1	0.707	0.562
psic_y_farm_antes_y_despues	0.589	0.387	0.481	0.387	0.408	0.645	0.548	0.707	1	0.728
nunca_psic_o_farm	0.468	0.41	0.688	0.308	0.324	0.41	0.58	0.562	0.728	1

Tabla 19. Similitudes coseno entre las características relevantes para los distintos perfiles de tratamiento (completa).

Variables	psic_antes	psic_despues	psic_antes_y_despues	farm_antes	farm_despues	farm_antes_y_despues	psic_y_farm_antes	psic_y_farm_despues	psic_y_farm_antes_y_despues	nunca_psic_o_farm
psic_antes	1	0.365	0.408	0.365	0	0.183	0.387	0.5	0.589	0.468
psic_despues	0.365	1	0.596	0	0.316	0.2	0.141	0.548	0.387	0.41
psic_antes_y_despues	0.408	0.596	1	0.149	0.471	0.298	0.316	0.544	0.481	0.688
farm_antes	0.365	0	0.149	1	0.316	0.6	0.283	0.365	0.387	0.308
farm_despues	0	0.316	0.471	0.316	1	0.632	0.447	0.577	0.408	0.324
farm_antes_y_despues	0.183	0.2	0.298	0.6	0.632	1	0.424	0.73	0.645	0.41
psic_y_farm_antes	0.387	0.141	0.316	0.283	0.447	0.424	1	0.387	0.548	0.58
psic_y_farm_despues	0.5	0.548	0.544	0.365	0.577	0.73	0.387	1	0.707	0.562
psic_y_farm_antes_y_despues	0.589	0.387	0.481	0.387	0.408	0.645	0.548	0.707	1	0.728
nunca_psic_o_farm	0.468	0.41	0.688	0.308	0.324	0.41	0.58	0.562	0.728	1

Tabla 20. Distancia euclídea entre la distribución de categorías de las variables de encuesta para cada perfil de tratamiento (completa).

index	psic_antes	psic_despues	psic_antes_y_despues	farm_antes	farm_despues	farm_antes_y_despues	psic_y_farm_antes	psic_y_farm_despues	psic_y_farm_antes_y_despues	nunca_psic_o_farm
psic_antes	0.0	51.195	49.562	46.436	49.014	54.427	40.42	52.922	49.931	449.855
psic_despues	51.195	0.0	2.448	5.416	2.788	3.657	11.533	2.073	2.969	500.189
psic_antes_y_despues	49.562	2.448	0.0	4.551	2.612	5.366	10.005	3.768	2.655	498.753
farm_antes	46.436	5.416	4.551	0.0	3.164	8.259	6.924	6.881	4.703	495.314
farm_despues	49.014	2.788	2.612	3.164	0.0	5.631	9.326	4.201	2.908	497.976
farm_antes_y_despues	54.427	3.657	5.366	8.259	5.631	0.0	14.392	2.03	5.151	503.466
psic_y_farm_antes	40.42	11.533	10.005	6.924	9.326	14.392	0.0	13.004	9.941	489.731
psic_y_farm_despues	52.922	2.073	3.768	6.881	4.201	2.03	13.004	0.0	3.613	501.935
psic_y_farm_antes_y_despues	49.931	2.969	2.655	4.703	2.908	5.151	9.941	3.613	0.0	499.107
nunca_psic_o_farm	449.855	500.189	498.753	495.314	497.976	503.466	489.731	501.935	499.107	0.0

Tabla 21. Distancia euclídea entre la distribución de categorías de las variables de encuesta para cada perfil de tratamiento (completa).

Variables	Combinacion categorias	farm_antes	farmacos_despues	psic_antes	psic_despues	psic_y_farm_antes	psic_y_farm_despues	psic_y_farm_antes_y_despues
Sexo	(0, 0)	48.72	48.68	45.37	48.87	48.06	49.01	48.72
Sexo	(0, 1)	1.04	1.09	4.4	0.9	1.7	0.76	1.04
Sexo	(1, 0)	49.81	49.91	45.98	49.81	49.15	50.0	49.81
Sexo	(1, 1)	0.43	0.33	4.25	0.43	1.09	0.24	0.43
Sanitario	(0, 0)	91.82	91.87	85.3	92.16	90.6	92.34	91.73
Sanitario	(0, 1)	1.37	1.32	7.89	1.04	2.6	0.85	1.47
Sanitario	(1, 0)	6.71	6.71	6.05	6.52	6.62	6.66	6.81
Sanitario	(1, 1)	0.09	0.09	0.76	0.28	0.19	0.14	0.0
Tenencia_Covid	(0, 0)	90.17	90.26	83.46	90.45	89.04	90.64	90.22
Tenencia_Covid	(0, 1)	1.37	1.28	8.08	1.09	2.5	0.9	1.32
Tenencia_Covid	(1, 0)	8.36	8.32	7.89	8.22	8.18	8.36	8.32
Tenencia_Covid	(1, 1)	0.09	0.14	0.57	0.24	0.28	0.09	0.14
Hospitalizacion_Covid	(0, 0)	97.78	97.87	90.6	98.02	96.46	98.25	97.83
Hospitalizacion_Covid	(0, 1)	1.47	1.37	8.65	1.23	2.79	0.99	1.42
Hospitalizacion_Covid	(1, 0)	0.76	0.71	0.76	0.66	0.76	0.76	0.71
Hospitalizacion_Covid	(1, 1)	0.0	0.05	0.0	0.09	0.0	0.0	0.05
Fallecimiento_Familiar_por_Covid	(0, 0)	95.27	95.46	88.23	95.37	94.0	95.79	95.32
Fallecimiento_Familiar_por_Covid	(0, 1)	1.42	1.23	8.46	1.32	2.69	0.9	1.37
Fallecimiento_Familiar_por_Covid	(1, 0)	3.26	3.12	3.12	3.31	3.21	3.21	3.21
Fallecimiento_Familiar_por_Covid	(1, 1)	0.05	0.19	0.19	0.0	0.09	0.09	0.09
Ataques_Ansiedad	(0, 0)	83.27	83.65	77.36	83.7	82.33	84.07	83.6
Ataques_Ansiedad	(0, 1)	1.04	0.66	6.95	0.61	1.98	0.24	0.71
Ataques_Ansiedad	(1, 0)	15.26	14.93	13.99	14.98	14.89	14.93	14.93
Ataques_Ansiedad	(1, 1)	0.43	0.76	1.7	0.71	0.8	0.76	0.76
Enfermedad_Cronica	(0, 0)	63.85	63.89	59.5	63.75	63.66	64.22	64.37
Enfermedad_Cronica	(0, 1)	0.85	0.8	5.2	0.95	1.04	0.47	0.33
Enfermedad_Cronica	(1, 0)	34.69	34.69	31.85	34.92	33.55	34.78	34.17
Enfermedad_Cronica	(1, 1)	0.61	0.61	3.45	0.38	1.75	0.52	1.13
Enfermedad_Cronica_Familiar	(0, 0)	72.59	72.83	67.25	72.68	71.6	73.06	72.78
Enfermedad_Cronica_Familiar	(0, 1)	1.04	0.8	6.38	0.95	2.03	0.57	0.85
Enfermedad_Cronica_Familiar	(1, 0)	25.95	25.76	24.1	25.99	25.61	25.95	25.76
Enfermedad_Cronica_Familiar	(1, 1)	0.43	0.61	2.27	0.38	0.76	0.43	0.61
Enfermedad_Cardiovascular	(0, 0)	87.9	88.0	81.38	88.04	87.1	88.33	88.04
Enfermedad_Cardiovascular	(0, 1)	1.37	1.28	7.89	1.23	2.17	0.95	1.23
Enfermedad_Cardiovascular	(1, 0)	10.63	10.59	9.97	10.63	10.11	10.68	10.49
Enfermedad_Cardiovascular	(1, 1)	0.09	0.14	0.76	0.09	0.61	0.05	0.24
Diabetes	(0, 0)	93.81	93.95	86.81	93.95	92.53	94.33	93.86
Diabetes	(0, 1)	1.47	1.32	8.46	1.32	2.74	0.95	1.42
Diabetes	(1, 0)	4.73	4.63	4.54	4.73	4.68	4.68	4.68
Diabetes	(1, 1)	0.0	0.09	0.19	0.0	0.05	0.05	0.05
Cancer	(0, 0)	96.6	96.64	89.65	96.74	95.51	97.12	96.6
Cancer	(0, 1)	1.47	1.42	8.41	1.32	2.55	0.95	1.47
Cancer	(1, 0)	1.94	1.94	1.7	1.94	1.7	1.89	1.94
Cancer	(1, 1)	0.0	0.0	0.24	0.0	0.24	0.05	0.0
Enfermedad_Respiratoria	(0, 0)	93.1	93.19	86.34	93.24	91.78	93.57	93.24
Enfermedad_Respiratoria	(0, 1)	1.37	1.28	8.13	1.23	2.69	0.9	1.23
Enfermedad_Respiratoria	(1, 0)	5.43	5.39	5.01	5.43	5.43	5.43	5.29
Enfermedad_Respiratoria	(1, 1)	0.09	0.14	0.52	0.09	0.09	0.09	0.24
Enfermedad_Traumatologica	(0, 0)	94.23	94.38	87.33	94.28	93.1	94.66	94.28
Enfermedad_Traumatologica	(0, 1)	1.37	1.23	8.27	1.32	2.5	0.95	1.32
Enfermedad_Traumatologica	(1, 0)	4.3	4.21	4.02	4.4	4.11	4.35	4.25
Enfermedad_Traumatologica	(1, 1)	0.09	0.19	0.38	0.0	0.28	0.05	0.14
Enfermedad_Autoinmune	(0, 0)	95.84	95.89	88.94	96.03	94.71	96.41	96.03
Enfermedad_Autoinmune	(0, 1)	1.47	1.42	8.36	1.28	2.6	0.9	1.28
Enfermedad_Autoinmune	(1, 0)	2.69	2.69	2.41	2.65	2.5	2.6	2.5
Enfermedad_Autoinmune	(1, 1)	0.0	0.0	0.28	0.05	0.19	0.09	0.19
Otra_Enfermedad_Cronica	(0, 0)	96.08	95.94	88.85	96.12	94.71	96.41	96.03
Otra_Enfermedad_Cronica	(0, 1)	1.28	1.42	8.51	1.23	2.65	0.95	1.32
Otra_Enfermedad_Cronica	(1, 0)	2.46	2.65	2.5	2.55	2.5	2.6	2.5
Otra_Enfermedad_Cronica	(1, 1)	0.19	0.0	0.14	0.09	0.14	0.05	0.14
Trastorno_Sueño	(0, 0)	95.65	95.65	88.8	95.7	94.47	96.08	95.6
Trastorno_Sueño	(0, 1)	1.37	1.37	8.22	1.32	2.55	0.95	1.42
Trastorno_Sueño	(1, 0)	2.88	2.93	2.55	2.98	2.74	2.93	2.93
Trastorno_Sueño	(1, 1)	0.09	0.05	0.43	0.0	0.24	0.05	0.05
Tenencia_de_Hijos	(0, 0)	71.69	71.88	66.54	72.02	70.7	72.02	71.6
Tenencia_de_Hijos	(0, 1)	1.09	0.9	6.24	0.76	2.08	0.76	1.18

*Tabla 22. Distribución de categorías de variables de encuesta para cada perfil de tratamiento.**

* Dado el elevado número de columnas y de registros, incluso en el anexo hemos añadido una versión reducida de la tala, la versión completa se puede obtener ejecutando el código presente en el repositorio y descargándola.

Variables	psic_antes	psic_despues	psic_antes_y_despues	farm_antes	farm_despues	farm_antes_y_despues	psic_y_farm_antes	psic_y_farm_despues	psic_y_farm_antes_y_despues	nunca_psic_o_farm
Sexo	1.441	3.386	5.255	11.806	5.881	8.144	5.783	5.327	6.865	61.399
Edad	20.866	17.144	19.864	15.943	0.08	3.233	5.238	1.662	0.068	13.623
Sanitario	2.256	14.183	13.461	0.187	0.092	4.52	0.278	1.63	2.681	8.788
Tenencia_Covid	0.186	2.116	0.679	1.59	0.092	0.331	0.108	0.004	0.098	0.013
Hospitalizacion_Covid	0.852	10.878	0.361	0.477	1.148	0.141	0.1	0.217	1.212	0.908
Fallecimiento_Familiar_por_Covid	0.593	1.147	16.257	0.018	4.926	0.57	0.005	1.684	0.287	5.306
Ataques_Ansiedad	1.036	40.76	36.392	0.295	52.726	11.541	8.075	53.863	48.46	240.312
Enfermedad_Cronica	2.476	0.961	0.015	2.62	0.595	2.183	25.939	4.642	25.274	49.674
Enfermedad_Cronica_Familiar	0.152	0.304	1.489	1.873	3.421	1.635	0.039	3.881	5.496	16.622
Enfermedad_Cardiovascular	0.392	0.513	0.314	0.025	0.371	1.37	10.024	0.096	0.246	0.121
Diabetes	1.243	1.454	1.847	2.441	0.88	0.723	0.044	0.003	0.333	3.245
Cancer	0.403	0.687	1.651	0.011	0.896	1.378	10.89	0.486	0.873	5.646
Enfermedad_Respiratoria	0.181	0.047	7.167	1.692	3.644	0.92	0.508	2.272	3.897	10.881
Enfermedad_Traumatologica	0.208	1.442	0.323	3.506	2.985	0.157	12.415	0.002	3.189	8.984
Enfermedad_Autoinmune	0.153	0.033	0.003	0.116	0.007	0.837	1.553	2.988	8.533	4.903
Otra_Enfermedad_Cronica	0.949	2.157	0.988	11.776	0.001	1.072	4.887	0.324	4.598	8.204
Trastorno_Sueño	2.693	0.858	1.09	5.245	0.005	0.427	7.138	0.221	0.002	3.03
Tenencia_de_Hijos	0.007	2.353	0.312	2.329	0.205	0.006	1.484	0.393	0.289	0.244
Medicacion_Familiar_por_Salud_Menta	0.1	1.844	0.839	0.064	0.012	16.076	0.017	23.491	22.773	20.745
Situacion_Laboral	0.088	0.211	0.274	2.748	0.218	2.767	5.926	2.962	8.062	11.455
Situacion_Convivencia	0.22	0.191	2.646	1.706	0.083	3.209	2.16	0.078	0.261	1.525
Escolarizacion	0.188	0.285	0.362	0.478	0.372	0.142	0.801	0.218	0.362	0.501
Estudios	10.061	0.1	8.639	11.88	1.425	2.595	3.58	0.015	3.252	0.0
Religion	10.082	3.187	1.477	6.671	1.152	0.375	4.254	0.192	0.347	0.182
Clase_Social	1.204	1.17	1.353	0.215	2.724	4.325	0.947	0.694	0.787	2.058
Trastorno_Depresivo_Antes	315.12	2.197	4.998	3.69	2.866	1.092	411.539	1.68	119.172	898.88
Trastorno_Ansioso_Antes	481.571	1.851	207.293	3.109	2.415	0.92	104.163	1.416	48.157	721.139
Trastorno_Depresivo_Despues	6.493	43.195	31.757	1.191	0.925	0.353	1.903	113.948	738.492	238.556
Trastorno_Ansioso_Despues	7.853	277.9	410.408	1.44	1.119	0.427	2.301	274.975	82.218	293.617

Tabla 23. Scores de las relevancias de cada característica normalizados (completa).

Variables	psic_antes	psic_despues	psic_antes_y_despues	farm_antes	farm_despues	farm_antes_y_despues	psic_y_farm_antes	psic_y_farm_despues	psic_y_farm_antes_y_despues	nunca_psic_o_farm
Sexo	0.003	0.012	0.013	0.741	0.112	0.507	0.014	0.019	0.009	0.068
Edad	0.043	0.062	0.048	1.0	0.002	0.201	0.013	0.006	0.0	0.015
Sanitario	0.005	0.051	0.033	0.012	0.002	0.281	0.001	0.006	0.004	0.01
Tenencia_Covid	0.0	0.008	0.002	0.1	0.002	0.021	0.0	0.0	0.0	0.0
Hospitalizacion_Covid	0.002	0.039	0.001	0.03	0.022	0.009	0.0	0.001	0.002	0.001
Fallecimiento_Familiar_por_Covid	0.001	0.004	0.04	0.001	0.093	0.035	0.0	0.006	0.0	0.006
Ataques_Ansiedad	0.002	0.147	0.089	0.019	1.0	0.718	0.02	0.196	0.066	0.267
Enfermedad_Cronica	0.005	0.003	0.0	0.164	0.011	0.136	0.063	0.017	0.034	0.055
Enfermedad_Cronica_Familiar	0.0	0.001	0.004	0.117	0.065	0.102	0.0	0.014	0.007	0.018
Enfermedad_Cardiovascular	0.001	0.002	0.001	0.002	0.007	0.085	0.024	0.0	0.0	0.0
Diabetes	0.003	0.005	0.005	0.153	0.017	0.045	0.0	0.0	0.0	0.004
Cancer	0.001	0.002	0.004	0.001	0.017	0.086	0.026	0.002	0.001	0.006
Enfermedad_Respiratoria	0.0	0.0	0.017	0.106	0.069	0.057	0.001	0.008	0.005	0.012
Enfermedad_Traumatologica	0.0	0.005	0.001	0.22	0.057	0.01	0.03	0.0	0.004	0.01
Enfermedad_Autoinmune	0.0	0.0	0.0	0.007	0.0	0.052	0.004	0.011	0.012	0.005
Otra_Enfermedad_Cronica	0.002	0.008	0.002	0.739	0.0	0.067	0.012	0.001	0.006	0.009
Trastorno_Sueño	0.006	0.003	0.003	0.329	0.0	0.027	0.017	0.001	0.0	0.003
Tenencia_de_Hijos	0.0	0.008	0.001	0.146	0.004	0.0	0.004	0.001	0.0	0.0
Medicacion_Familiar_por_Salud_Menta	0.0	0.007	0.002	0.004	0.0	1.0	0.0	0.085	0.031	0.023
Situacion_Laboral	0.0	0.001	0.001	0.172	0.004	0.172	0.014	0.011	0.011	0.013
Situacion_Convivencia	0.0	0.001	0.006	0.107	0.002	0.2	0.005	0.0	0.0	0.002
Escolarizacion	0.0	0.001	0.001	0.03	0.007	0.009	0.002	0.001	0.0	0.001
Estudios	0.021	0.0	0.021	0.745	0.027	0.161	0.009	0.0	0.004	0.0
Religion	0.021	0.011	0.004	0.418	0.022	0.023	0.01	0.001	0.0	0.0
Clase_Social	0.003	0.004	0.003	0.013	0.052	0.269	0.002	0.003	0.001	0.002
Trastorno_Depresivo_Antes	0.654	0.008	0.012	0.231	0.054	0.068	1.0	0.006	0.161	1.0
Trastorno_Ansioso_Antes	1.0	0.007	0.505	0.195	0.046	0.057	0.253	0.005	0.065	0.802
Trastorno_Depresivo_Despues	0.013	0.155	0.077	0.075	0.018	0.022	0.005	0.414	1.0	0.265
Trastorno_Ansioso_Despues	0.016	1.0	1.0	0.09	0.021	0.027	0.006	1.0	0.111	0.327

Tabla 24. Scores de las relevancias de cada característica normalizados (completa).

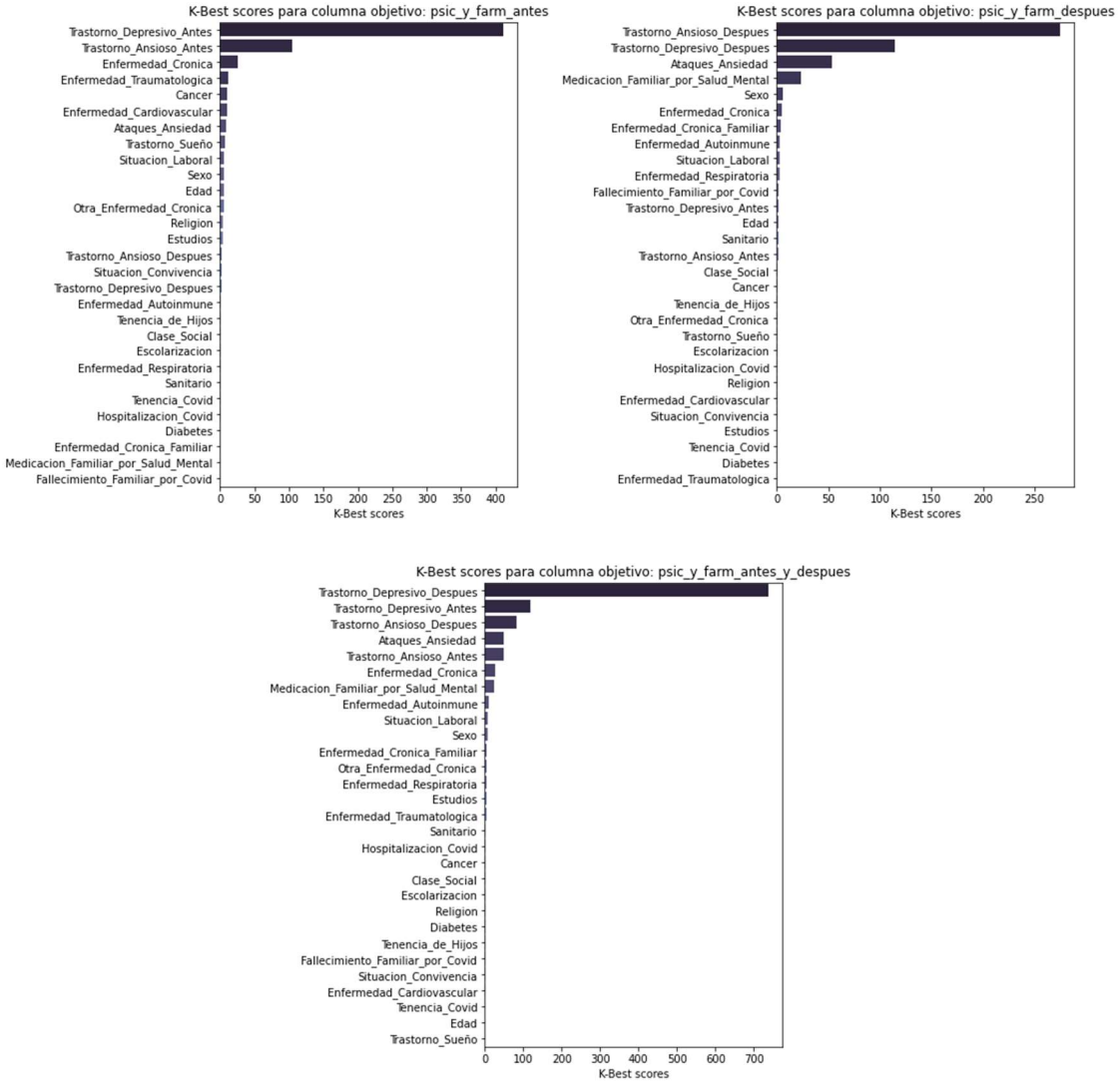


Ilustración 6. Scores de relevancia de cada variable de encuesta para los perfiles de tratamiento no incluidos en la memoria principal.