Descubrimiento de sinergias en la composición química del cannabis a partir de grandes volúmenes de datos sobre efectos subjetivos auto-reportados

Lucía Evangelista Gallo

Tesis de Licenciatura en Ciencias Físicas

Facultad de Ciencias Exactas y Naturales

Universidad de Buenos Aires

Noviembre de 2022

Tema: Descubrimiento de sinergias en la composición química del cannabis a partir de grandes volúmenes de datos sobre efectos subjetivos auto-reportados.

Alumno: Lucía Evangelista Gallo

L.U. N°: 639/14

Lugar de Trabajo: Departamento de Física, Facultad de Ciencias Exactas y Naturales, UBA.

Director del Trabajo: Dr. Enzo Tagliazucchi

Co-Director del Trabajo:

Fecha de Iniciación: Marzo 2021

Fecha de Finalización: Octubre 2022

Fecha del Examen Final:

Informe Final Aprobado por:

Autor	Jurado
Director	Jurado
Profesor de Tesis de Licenciatura	Jurado

Resumen

Prohibido durante décadas, tanto para uso recreativo como para investigaciones científicas, en los últimos años el consumo de cannabis fue paulatinamente admitido bajo ciertas condiciones, y empezaron a aparecer estudios científicos que muestran su eficacia a la hora de tratar condiciones médicas que abarcan desde la ansiedad, insomnio y dolor crónico hasta cáncer y trastornos neurológicos.

Estas investigaciones, además, sugieren que el cannabis no funciona de igual manera que un fármaco tradicional, pues no tiene un único principio activo. En un primer momento, el cannabinoide THC fue el que más atención recibió, dado que es el responsable de los efectos euforizantes, pero luego se empezaron a estudiar los otros cannabinoides presentes en la planta. Asimismo, se observó evidencia apoyando que el efecto de los cannabinoides por separado sería menos efectiva que su uso combinado, lo que se conoce como *efecto entourage*. Inicialmente se evaluaron combinaciones entre cannabinoides, pero más tarde el interés se volcó a las combinaciones entre cannabinoides y terpenos, moléculas aromáticas que también pueden producir efectos psicoactivos separadamente. Dada la enorme cantidad de combinaciones de moléculas posibles, la variedad de condiciones médicas y los costos que conlleva un ensayo clínico, se vuelve de especial importancia utilizar vías alternativas que puedan proponer candidatos a ser estudiados en un entorno experimental. El objetivo de este trabajo, entonces, fue utilizar técnicas computacionales para analizar datos de usuarios que consumieron cannabis para mitigar una condición médica y el perfil químico de las cepas disponibles para obtener todas las posibles sinergias (entre dos moléculas) candidatas a ser estudiadas en profundidad.

Los datos utilizados fueron provistos por Leafly (`https://www.leafly.com`), y contienen información sobre para qué condiciones médicas se consumió cada cepa, y cuáles fueron los efectos y sabores percibidos por los usuarios, en forma de suma total de votos de cada elemento. Además, compartieron una tabla con el perfil de cannabinoides y terpenos de las cepas cuantificado mediante análisis químicos. Este trabajo comenzó con un análisis exploratorio de estos datos y su posterior procesamiento, para llevarlos a un formato apto para los análisis siguientes.

Luego se realizó un estudio más en profundidad de los patrones en los datos, calculando las correlaciones de Spearman y buscando *clusters* y comunidades. Estos análisis resultaron útiles para la última parte del trabajo, en la que se buscaron las

sinergias y se utilizó este conocimiento para evaluar la coherencia de lo obtenido. Los resultados conseguidos proponen nuevas sinergias candidatas a ser estudiadas y, además, respaldan el interés actual en el estudio de combinaciones cannabinoide–terpeno, pues las selecciones con mayor cantidad de ocurrencias involucran este tipo de combinación. Más generalmente, los desarrollos de esta tesis muestran cómo un gran volumen de datos sobre experiencias subjetivas inducidas por drogas puede combinarse con información química para buscar posibles usos terapéuticos y/ó combinaciones de drogas útiles para distintos fines.

Índice general

Agradecimientos

Esta tesis existe porque mis amigos no me dejaron largar todo y poner un carrito en Costanera. Va dedicada a todos ustedes.

Quiero empezar por agradecerle a Enzo por acompañarme en este trabajo, por las discusiones y los intercambios que tuvimos durante estos meses.

A mis viejos, que me apoyaron todos estos años y siempre tuvieron palabras de aliento. A Meli, que soportó varias de mis crisis, mis broncas, mis frustraciones y mis ansiedades. A Vicky, Cami y Juli, que siempre estuvieron y nunca dejaron de animarme. A Mati, que tuvo la mala suerte de tener que soportarme en mis últimas materias y sobre todo en el proceso de esta tesis.

A Pili, Cami, Juli, Manu, Juli y Valen. Fueron un montón de años juntos empujando para adelante. Gracias por siempre apoyarme, escucharme y ayudarme, pero sobre todo por acompañarme. Cami y Pili, sobre todo a ustedes, que me bancaron a diario y en los momentos en los que quería largar todo o en los que estaba completamente sobrepasada.

A todas las personas que me fui cruzando en estos años y con quienes compartí parte del camino, gracias.

Introducción

El cannabis ha sido, por milenios, una planta con gran versatilidad, y no solo desde el punto de vista medicinal: además de su uso recreativo, también se utiliza para fabricar papel e indumentaria. Presuntamente oriunda de la estepa de Asia Central, pinturas en cerámicas sugieren que hace por lo menos unos 8000 años que el humano hace uso de esta planta. No obstante, como sucedió con muchas otras sustancias psicotrópicas, a lo largo de la historia fue prohibida por cuestiones religiosas, políticas y sociales[1], siendo la última vez la del prohibicionismo estadounidense. Esto provocó que, durante mucho tiempo, la evidencia fuera más bien anecdótica, hecho que fue cambiando en los últimos años, en los cuales se publicaron estudios científicos que proporcionan evidencia sólida sobre la eficacia del cannabis en trastornos variados, entre los que se encuentran ansiedad, insomnio, dolor crónico y la epilepsia refractaria[2]. A diferencia de otros fármacos, el cannabis no tiene un único principio activo[3]. Históricamente, el THC (tetrahidrocannabinol), responsable de los efectos euforizantes, fue el cannabinoide al que más atención se le dio, pero otros cannabinoides como CBD (cannabidiol), CBC (cannabichromene) y CBN (cannabinol), entre otros, también presentan propiedades interesantes, que van desde eficacia anti–tumoral en estudios sobre cáncer de mama[4], hasta antidepresivas[5].

Un fenómeno observado, que diferencia al cannabis (utilizado con fines terapéuticos) de otros fármacos, es que la eficacia de los cannabinoides por separado podría ser menor que su aplicación combinada, lo que se conoce como *efecto entourage* o *sinergia*[6]. Este fenómeno no se limita únicamente a combinaciones cannabinoide–cannabinoide, sino también a combinaciones entre cannabinoides y terpenos (las moléculas que dan sabor y aroma), en las cuales los terpenos modulan los efectos del cannabis[7]. Bajo esta premisa, lo ideal sería encontrar, en primer lugar, aquellas combinaciones con mayor probabilidad de generar cierto efecto deseado y, en segundo, cuál es la proporción en que debe estar presente cada compuesto. El ob-

jetivo principal de este trabajo fue, a partir del cruce entre el perfil químico de muchas cepas de cannabis y el "perfil de sensaciones" de los usuarios, obtener pares de moléculas (cannabinoide–cannabinoide, terpeno–terpeno, cannabinoide–terpeno) que fueran posibles candidatas a ser estudiadas en mayor detalle en función de un dado efecto terapéutico deseado.

Los datos que se utilizaron en este trabajo fueron provistos por Leafly (`https://www.leafly.com`), el mayor sitio web de internet especializado en cannabis que, además de sugerir vendedores (en estados de EE. UU. donde el cannabis es legal), tiene mucha información sobre cepas y sobre cannabinoides y terpenos. Lo relevante para este trabajo es que los usuarios pueden completar una encuesta con la experiencia vivida al consumir. En ella se puede describir la experiencia con texto libre, y también seleccionando etiquetas predefinidas por Leafly en tres "áreas": efectos sentidos (19 etiquetas), aromas y/o gustos percibidos (47) y la condición médica por la cual estaban consumiendo (40). Esta información es recolectada por Leafly y condensada en forma de "votos", y en cada cepa se puede ver una sección que indica qué efectos y gustos reportaron los usuarios, y para qué condiciones es más efectiva (entendiendo efectivo como las condiciones que más veces fueron seleccionadas). En la figura 1.1 se ve un ejemplo para la cepa *Strawberry Cough*.

Más en concreto, se utilizaron dos tablas, una con los votos, y otra con el perfil de cannabinoides y terpenos de cada cepa. En el capítulo "Análisis exploratorio de los datos" se exploran ambas tablas y se describen las transformaciones que se le hicieron a los datos para llevarlos a la forma final, con la cual se realizaron los análisis de los capítulos siguientes. El objetivo principal del proyecto fue investigar sinergias entre compuestos y proponer nuevos candidatos. Este proceso y los resultados obtenidos se discuten en el capítulo "Identificación de sinergias". Previo a la aplicación del modelo, en el capítulo "Identificación de relaciones y detección de comunidades", partiendo del conocimiento de que existen correlaciones entre efectos y sabores percibidos, entre efectos y el perfil de cannabinoides, y entre sabores y el perfil de terpenos[7], se estudiaron las correlaciones entre los distintos "grupos" de datos (i.e. condiciones, efectos subjetivos, sabores, cannabinoides, terpenos). Este proceso sirvió, por un lado, para reforzar los resultados ya conocidos y, por el otro, para relacionarlos con las condiciones. Con este análisis se consiguió una noción general de qué buscaban los usuarios al consumir[1] que fue útil al momento de dis-

[1]Por ejemplo, se encontró que la condición *depresión* se conectaba con efectos energizantes y con sabores cítricos, es decir, cepas que produjeran esos efectos eran preferidas por usuarios con

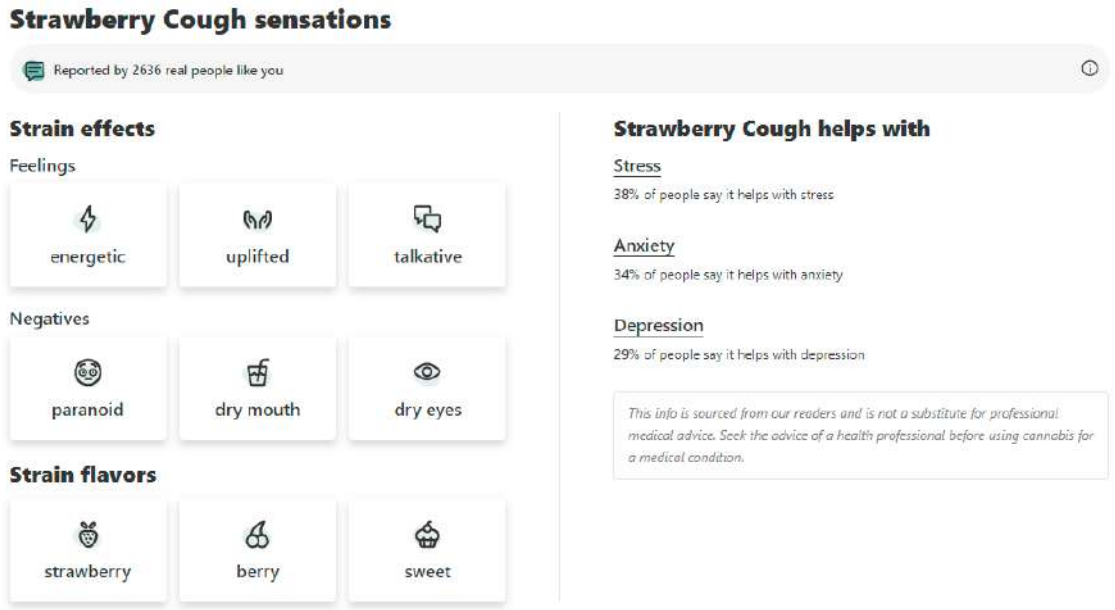


Figura 1.1: Resumen de las características principales de la cepa *Strawberry Cough*. Debajo del título muestran cuánta gente dio su opinión sobre esta cepa; le siguen, a izquierda, los efectos reportados, comenzando por los tres sentimientos más votados y seguido de los efectos negativos. Debajo se reportan los sabores más votados. A la derecha se muestran las tres condiciones más elegidas para las cuales los usuarios reportaron haber utilizado esta cepa. Imagen tomada de `www.leafly.com`.

cutir si las sinergias identificadas por el modelo eran coherentes o no. Además, con la información de correlaciones, se obtuvieron *clusters* por cada grupo de los antes mencionados (para conocer, por ejemplo, subgrupos de condiciones que estén relacionadas[2]) y también se construyeron grafos no dirigidos para identificar las distintas comunidades en los grupos. Finalmente, en el capítulo "Discusión", se recapitulan los resultados obtenidos y se hace un comentario general sobre ellos.

depresión.

[2]No necesariamente desde el punto de vista médico.

Análisis exploratorio de los datos

En este capítulo se presentan los datos provistos por Leafly, con los cuales se desarrolló este trabajo. El objetivo de esta parte es entender qué información contienen los datos, cuánta información falta, cómo se puede pulir el *dataset* de manera que se pueda hacer un buen análisis mientras se mantiene un compromiso con la cantidad de datos que hay que descartar y, en última instancia, llevar los datos crudos a un formato apto para alimentar modelos de *machine learning*.

El capítulo consta de tres secciones: en las primeras dos se hace una descripción de los dos conjuntos de datos utilizados, de la información que contienen, de las distribuciones que se empiezan a ver, y se hace una primera limpieza e interpretación inicial de ellos. Una vez comentados ambos conjuntos, en la última sección se describe el proceso que llevó a la forma final de los datos, en una única tabla, con la que se realizaron los análisis y los aplicaron los modelos explicados en los capítulos 3 y 4.

2.1. Los datos

De los archivos provistos por Leafly, en este trabajo se utilizaron dos de ellos[1]:

> `standardized_lab_data_210311.csv`: contiene información sobre el contenido de cannabinoides y terpenos en cada cepa. Para cada una, hay varias mediciones, realizadas en fechas distintas y/o en laboratorios distintos. Para más información al respecto, se puede consultar la web de Leafly en la sección Lab partners o la web de uno de ellos, Psilabs. En adelante, se referirá a este archivo como "Tabla químicos".

> `strain_page_info_20210325.csv`: por cada cepa, contiene la especie a la que

[1]Pueden consultarse en el repositorio `https://github.com/luuevang/datos_tesis`.

Cannabinoides				
cbc	cbd	cbda	cbdv	cbg
cbga	cbn	cbt	d8_thc	thc
thca	thcv	thcva		

Cuadro 2.1: Nombre de los cannabinoides en la tabla químicos.

pertenece, algunos datos informativos como el nombre, el alias, el *id* y la popularidad de la cepa, y la cantidad de votos recibidos respecto de la condición [médica][2] para la que fue utilizada, los efectos y los sabores percibidos por los usuarios. En adelante, se referirá a este archivo como "Tabla votos".

Por una cuestión de conveniencia, se decidió mantener los nombres de las condiciones, los efectos, los sabores, los cannabinoides y los terpenos como vienen de la base de datos de Leafly, en inglés. En castellano los nombres pueden ser ligeramente distintos.

2.1.1. Tabla químicos

Esta tabla contiene mediciones del contenido de una serie de cannabinoides y de terpenos presente en varias cepas de cannabis. Para cada cepa hay distintos registros, por haber sido generados en fechas o en laboratorios distintos. El archivo original tiene 90607 filas y 90 columnas, de las cuales 13 corresponden a cannabinoides, 57 a terpenos, 9 a recuento del total de cannabinoides (e.g. `tot_cbc`) y las 11 restantes tienen que ver con el *id* del laboratorio, la denominación de la muestra, la fecha, si se le hicieron estudios de terpenos, el nombre de la cepa, etc. Los cannabinoides y los terpenos analizados se detallan en las tablas 2.1 y 2.2.

De las 90607 mediciones en la tabla, hay 16764 que no tienen identificada la cepa, por lo que se decidió eliminarlas, quedando así 73843 filas, correspondientes a 3090 cepas. Hay tres tipos de quemotipos, THC-dominante, CBD/THC balanceado y CBD-dominante, a los cuales les corresponden, respectivamente, 2934, 97 y 56 cepas, y hay tres para las que no está el dato. Por otro lado, no todas las cepas fueron estudiadas la misma cantidad de veces: hay 3004 cepas con menos de 150 mediciones (de las cuales 2887 tienen menos de 80 mediciones) y apenas 86 cepas

[2]No se tiene constancia de que haya sido indicado por un médico, solo la declaración de los usuarios de qué condición se buscó alivianar al consumir cannabis.

Terpenos				
α_cedrene	α_ocimene	α_phellandrene	α_pinene	α_terpinene
α_terpineol	β_maaliene	β_nerolidol	β_ocimene	β_pinene
bisabolol	borneol	borneol_isomers	camphene	camphor
carene	caryophyllene	caryophyllene_oxide	cedrol	cis_nerolidol
cis_ocimene	cis_phytol	citronellol	eucalyptol	farnesene
fenchol	fenchone	fenchyl_alcohol	g_terpinene	g_terpineol
geraniol	geranyl_acetate	guaiol	humulene	iso_borneol
iso_pulegol	limonene	linalool	menthol	myrcene
nerol	ocimene	p_cymene	phytol	pulegone
sabinene	sabinene_hydrate	selinadiene	terpineol	terpinolene
thujene	thymol	trans_nerolidol	trans_ocimene	trans_phytol
valencene				

Cuadro 2.2: Nombre de los terpenos en la tabla químicos.

con más de 150 mediciones, y 203 cepas con 80 o más (correspondiente al 7 % de las cepas).

En la figura 2.1 se ve la zona hasta 80 mediciones, y la tendencia a angostarse que se aprecia desde aproximadamente 40 se mantiene hasta el final del gráfico. Finalmente, para los terpenos *β_maaliene*, *borneol_isomers*, *selinadiene* y *thujene* no había datos, por lo que fueron eliminados de la tabla. De esta forma, el dataset final contenía 3090 filas y 78 columnas (las variables categóricas se perdieron al hacer el promedio de las filas, pero no eran relevantes).

2.1.2. Tabla votos

Antes de comenzar, es importante remarcar que esta tabla contiene información **subjetiva**. Los datos son provistos por usuarios (figura 1.1 en "Introducción"), no son recolectados en un ambiente controlado ni se tiene grupo de control, así como tampoco es posible determinar exactamente qué consumieron. Por otro lado, no se tiene constancia de indicaciones médicas de consumo para aliviar dolencias, de modo que se interpretará como "automedicación". Por último, dado que la información es pública, no se puede descartar que las experiencias de los usuarios no puedan ser influenciadas por lo leído. Un ejemplo podría ser la cepa llamada *"strawberry cough"*: su nombre podría ser producto de opinión popular (con orígenes variados) o bien, si los análisis de composición química hallaran que esta cepa contiene terpenos con aroma y/o gusto a frutilla en una concentración alta, podría ser que los usuarios

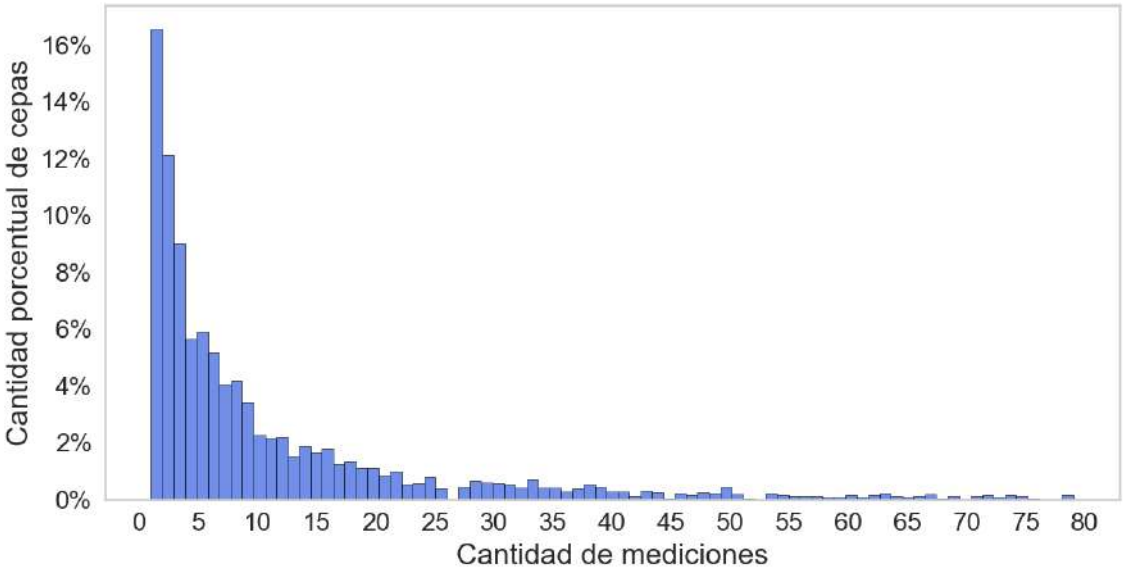


Figura 2.1: Distribución de la cantidad de mediciones hecha por cada cepa en la tabla químicos. En la figura se muestra la zona entre 0 y 80 mediciones, que concentra el 93 % de las cepas.

efectivamente hubieran percibido la frutilla y esa fuera la razón del nombre.

La cantidad total de votos[3] que cada cepa recibió respecto de los sabores y efectos percibidos, y las condiciones médicas para las cuales se consumió, se encuentra en el archivo `strain_page_info_20210325.csv`. Leafly tiene, en su base de datos, 3439 cepas, 40 condiciones médicas (se llamará "condiciones" a lo que en Leafly catalogan como *conditions* y como *symptom*), 19 efectos (engloban lo que en Leafly separan en *effect* y *negative*, e.g. *negative_dizzy_votes*) y 47 sabores. El dataset tiene 3439 filas y 117 columnas, de las cuales 6 corresponden a nombre, *id*, popularidad, especie y alias de cada cepa, y 5 a votos totales por cada sección de la tabla (*ReviewTotalConditionVotes*, *ReviewTotalEffectVotes*, *ReviewTotalFlavorVotes*, *ReviewTotalNegativeVotes*, *ReviewTotSymptomVotes*). Las 106 columnas restantes se detallan en las tablas 2.3, 2.4 y 2.5.

En el dataset hay solo dos columnas con `NaN`, la que tiene los alias (2063 `NaN`), que no fue relevante para el trabajo, y la que tiene el nombre de la especie a la que pertenece la cepa (127). Dado que es un dato importante, se eliminaron las filas sin datos. En la figura 2.3 se puede ver la distribución de la cantidad de votos totales respecto de las condiciones (variable *ReviewTotalConditionVotes*) que recibieron las cepas. Por cuestiones de escala, el gráfico se separó en tres partes: en la primera,

[3]Al 25 de marzo de 2021.

Condiciones médicas			
addAdhd	alzheimers	anorexia	anxiety
arthritis	asthma	bipolarDisorder	cachexia
cancer	crohnsDisease	epilepsy	fibromyalgia
gastrointestinalDisorder	glaucoma	hivAids	hypertension
migraines	multipleSclerosis	muscularDystrophy	parkinsons
phantomLimb	pms	ptsd	spinalCord
tinnitus	tourettesSyndrome	cramps	depression
eyePressure	fatigue	headaches	inflammation
insomnia	lackOfAppetite	muscleSpasms	nausea
pain	seizures	spasticity	stress

Cuadro 2.3: Nombre de las condiciones médicas en la tabla votos.

Efectos (positivos y negativos)						
aroused	creative	energetic	euphoric	focused	giggly	happy
hungry	relaxed	sleepy	talkative	tingly	uplifted	anxious
dizzy	dryEyes	dryMouth	headache	paranoid		

Cuadro 2.4: Nombre de los efectos positivos y negativos en la tabla votos.

Sabores						
ammonia	apple	apricot	berry	blueCheese	blueberry	butter
cheese	chemical	chestnut	citrus	coffee	diesel	earthy
flowery	grape	grapefruit	honey	lavender	lemon	lime
mango	menthol	mint	nutty	orange	peach	pear
pepper	pine	pineapple	plum	pungent	rose	sage
skunk	spicyHerbal	strawberry	sweet	tar	tea	tobacco
treeFruit	tropical	vanilla	violet	woody		

Cuadro 2.5: Nombre de los sabores en la tabla votos.

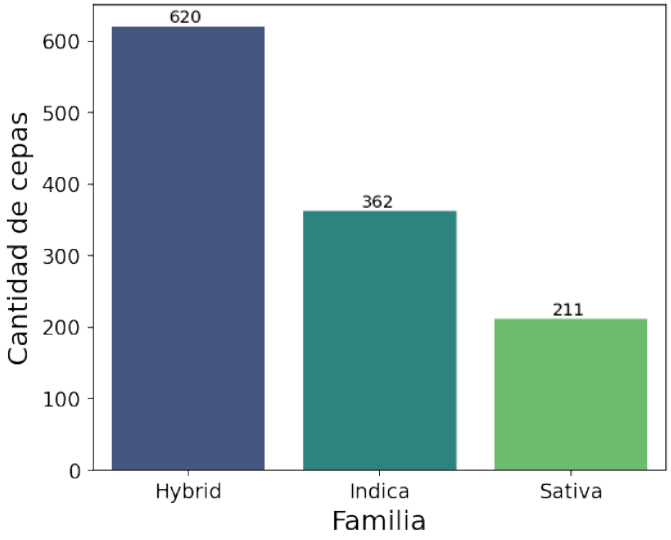


Figura 2.2: Cantidad de cepas pertenecientes a las especies *Sativa, Indica* e *Hybrid*.

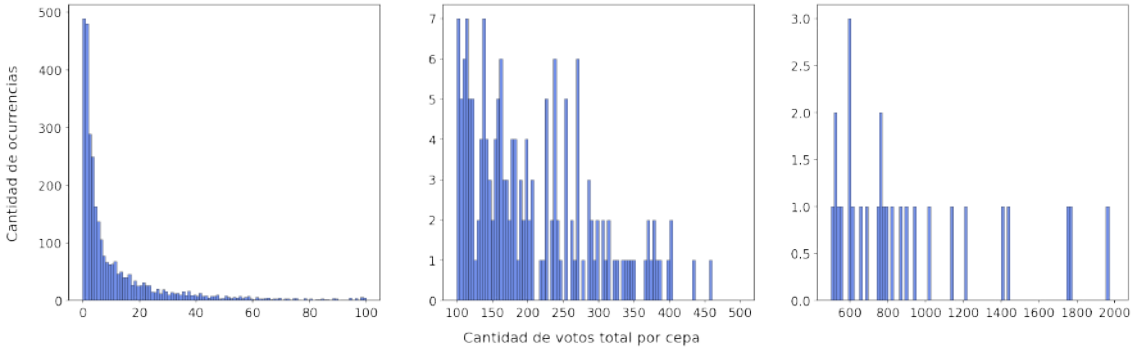


Figura 2.3: Cantidad total de votos por cepa, dividido en tres zonas: a la izquierda, entre 0 y 100 votos, en el centro, entre 100 y 500, a la derecha entre 500 y 200.

la cantidad de cepas con votos entre 0 y 100, en la segunda, de 100 a 500 votos, y la tercera de 500 a 2000 votos. Se puede ver que hay alrededor de 2100 cepas que fueron elegidas menos de 10 veces, de las cuales cerca de 1000 tienen entre 0 y 1 votos. Hubo 994 cepas con votos entre 10 y 100, 170 con votos entre 100 y 500, y 28 cepas con votos entre 500 y 2000, de las cuales tan solo 8 tienen más de 1000 votos. De los histogramas de la figura 2.3 se sigue que hay algunas pocas cepas que son muy votadas, es decir, altamente consumidas y reseñadas. Esto podría deberse a ser más conocidas, o el lugar donde son vendidas, u otros factores que escapan nuestro conocimiento. Dado que las cepas con poca cantidad de votos no aportarían información significativa al análisis y modelo que se realizó, se decidió trabajar únicamente con cepas que tuvieran 10 o más votos. Con este filtrado y las filas anteriormente eliminadas, queda un *dataset* de 1193 filas y 117 columnas.

Leafly clasifica a las cepas en tres grupos: *Sativa*, *Indica* e *Hybrid*, de las cuales había, respectivamente, 211 (18 % del total), 362 (30 %) y 620 (52 %) ejemplares (figura 2.2).

En la figura 2.4 está la distribución de condiciones en base a la cantidad de veces que fueron votadas, es decir, el reporte de la dolencia o condición que se buscó mitigar. La más elegida fue *stress*, con ∼85.000 votos, seguida por *anxiety* (∼66K), *pain* (∼65K), *depression* (∼60K) e *insomnia* (∼41K) (figura 2.4b)

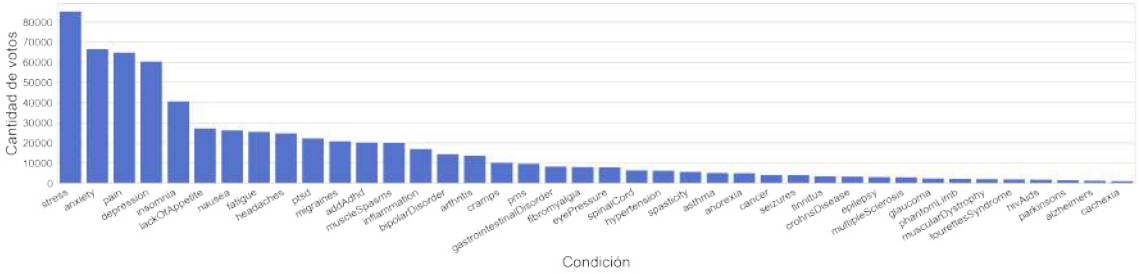
Respecto de los efectos producidos al consumir (figura 2.5), hay cuatro con más de 100K votos: *happy, relaxed, euphoric* y *uplifted*, y es interesante notar que son todos efectos positivos. Por el otro lado, los cuatro efectos menos experimentados por la gente son negativos: *dizzy, paranoid, anxious* y *headache*. En el rango intermedio de votos, en general son positivos y pocos negativos.

En "Introducción" se mencionó la relación entre efectos y sabores percibidos por la acción conjunta de cannabinoides y terpenos. Por otro lado, hay cierta predominancia de la especie *Indica* en generar efectos de relajación, relacionado con sabores como *earthy, berry* y de la especie *Sativa* de efectos de creatividad, euforia y mejora de ánimo, asociados a gustos *citrus, flowery, lemon*. En base a esto, y teniendo en cuenta que alrededor del 50 % de los datos corresponden a *Hybrid*, tiene sentido la distribución de votos por cada sabor en la figura 2.6: con más de 40K, están *earthy* y *sweet*, dos sabores con los que se puede identificar rápidamente el cannabis; con entre 15K y 26K están *pungent, citrus, pine, woody, berry, flowery*, sabores que están correlacionados con los efectos de relajación/mejora de ánimo[7], y luego hay varios otros sabores con menor cantidad de votos, hasta llegar al menos votado, *pear*, con 721 votos.

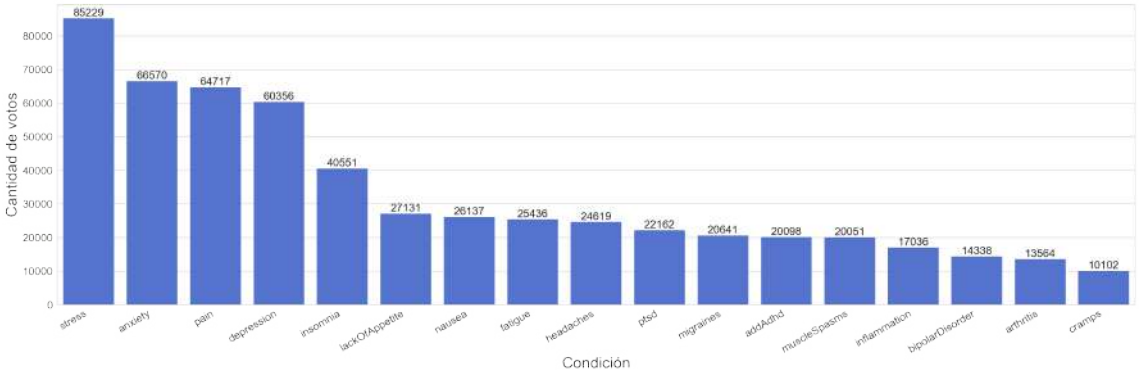
La distinción entre *Sativa*, *Indica* e *Hybrid*, si bien cuestionada por botánicos[8], continúa siendo la más común hoy en día. Para corroborar si los datos respondían a esa clasificación, se tomó el subconjunto de columnas correspondientes a las condiciones médicas, se normalizó, y se utilizó el algoritmo de `K-Means`[4] para obtener los grupos. El resultado se muestra en la figura 2.7[5], en la cual hay dos grupos bien distinguibles (imagen superior derecha), correspondientes a las especies *Sativa* e *Indica* y, superpuesta, la *Hybrid*. `K-Means` requiere que el número k de clusters a

[4] Ver K–Means.

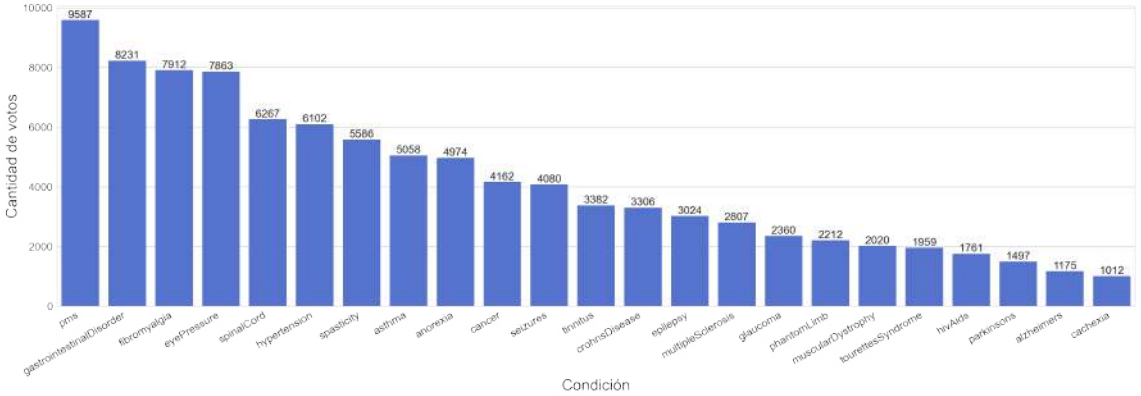
[5] Para graficar, se aplicó PCA a los resultados, de modo de poder proyectar los patrones a un espacio bidimensional (ver *Principal Component Analysis* en la sección "Metodologías" de este capítulo).



(a) Distribución completa.



(b) Zoom en las condiciones con más de diez mil votos.



(c) Zoom en las condiciones con menos de diez mil votos.

Figura 2.4: Distribución de las condiciones en base a la cantidad de votos recibidos.

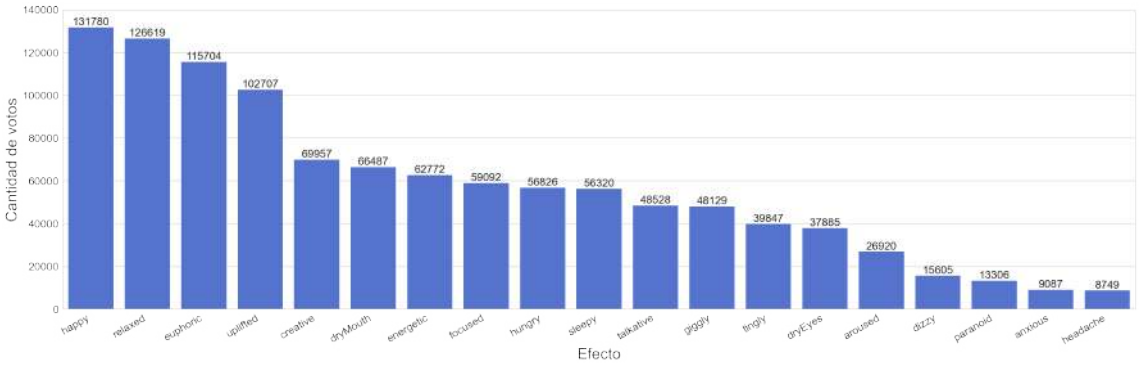
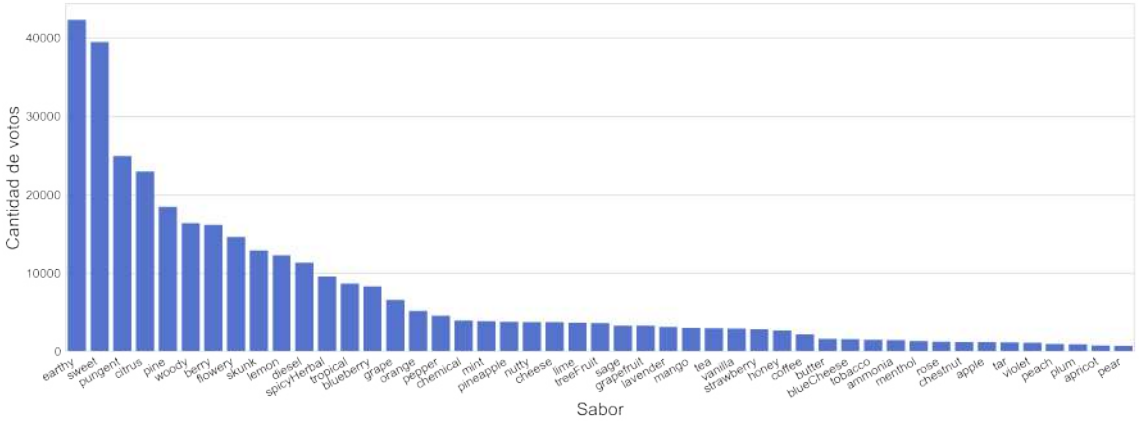
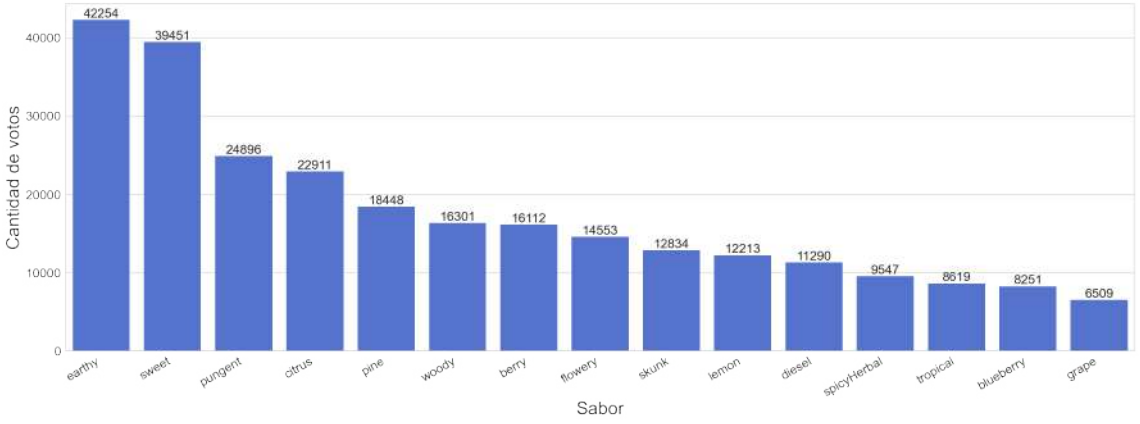


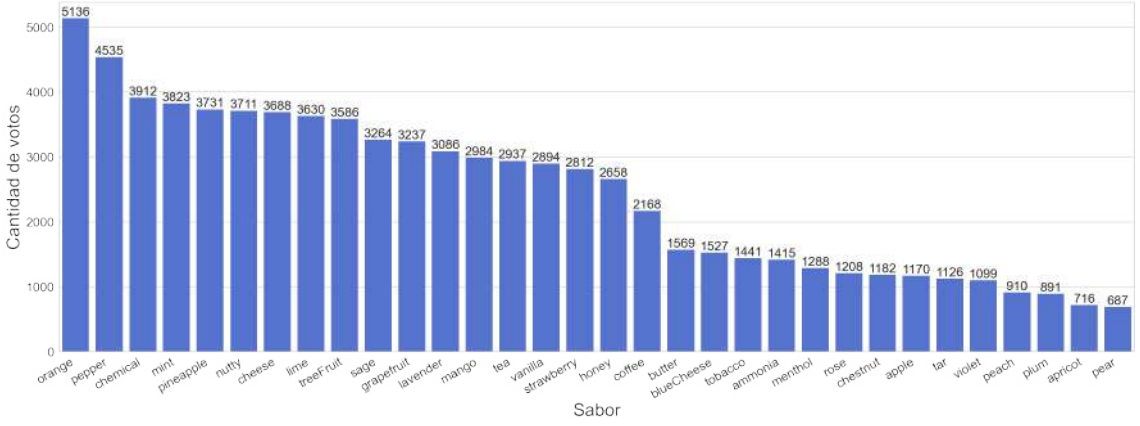
Figura 2.5: Distribución de los efectos en base a la cantidad de votos recibidos.



(a) Distribución completa.



(b) Zoom de los quince sabores más votados.



(c) Zoom del resto de los sabores.

Figura 2.6: Distribución de los sabores en base a la cantidad de votos recibidos.

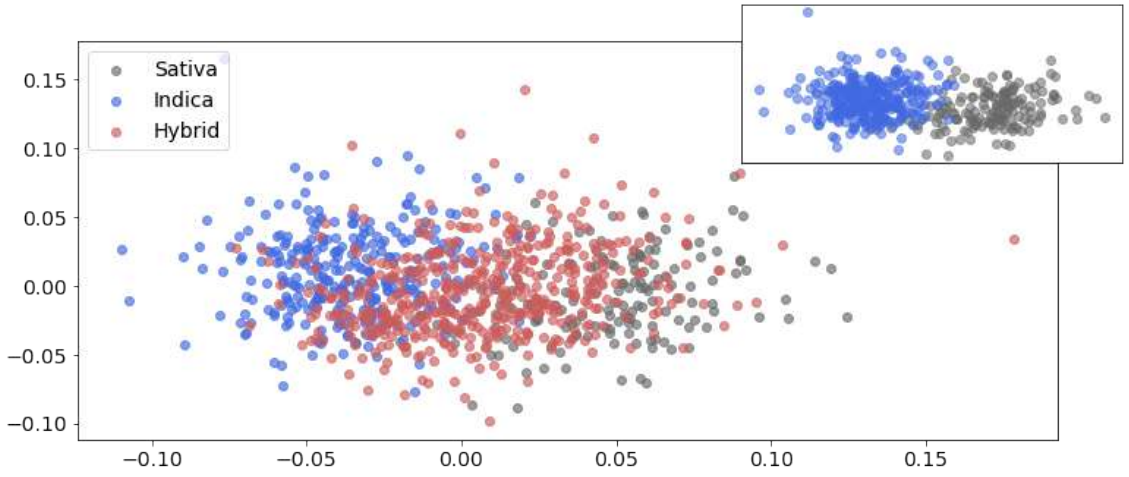


Figura 2.7: Representación bidimensional de los *clusters* de especies hallados por `K-Means`, en el espacio generado por PCA. Se observan tres *clusters*, dos bien distinguibles (figura superior derecha) correspondientes a *Indica* y *Sativa*, y el tercero *Hybrid* superpuesto.

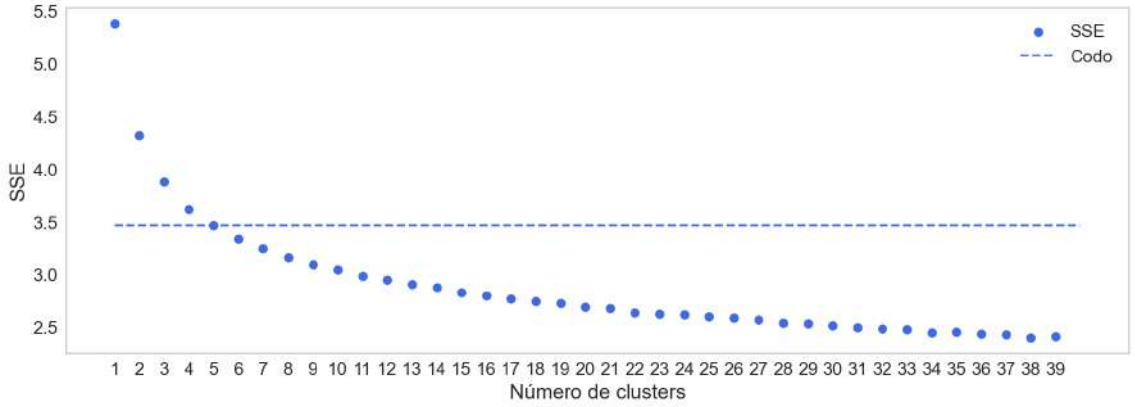


Figura 2.8: Método del codo. La línea punteada marca el comienzo del codo.

generar sea fijado de antemano, y una forma de determinar k es el método del codo (figura 2.8). Si bien pedir tres clusters arrojó un resultado esperable, de la figura 2.8 se observa que no sería el número óptimo según estos criterios; basándose en el método del codo, habría que buscar cinco o seis grupos. Esta discrepancia no es sorprendente dado que, por un lado, existe una discusión alrededor de la división en especies[8] y, por el otro, los distintos cruces genéticos que ocurrieron con el tiempo[9] pudieron dar lugar a nuevas especies. No obstante, es interesante que, a pesar de estas observaciones, los tres clusters obtenidos de K-Means sean coherentes.

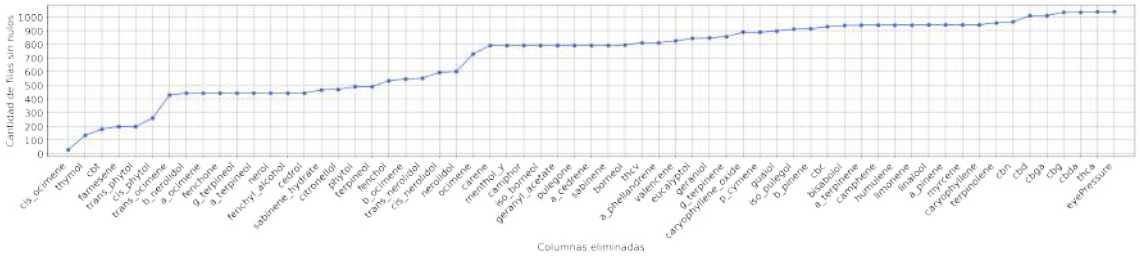


Figura 2.9: Cantidad de filas sin valores nulos en función de las variables eliminadas. A medida que se avanza hacia la derecha, las variables a izquierda fueron eliminadas de a una por vez. A la altura de *menthol_y* hay 790 filas y 143 columnas en la tabla.

2.1.3. El dataset final

Una vez limpias las tablas de químicos y de votos, el paso siguiente fue asegurarse de que las dos tablas tenían datos sobre las mismas cepas. Esto es relevante pues, como se verá en el capítulo 3, se analizaron todas las combinaciones de a pares entre condiciones, efectos, sabores, terpenos y cannabinoides y, para ser consistentes, es necesario que estén las mismas cepas en todos los casos. Combinando las tablas en base a la columna de cepas, *strain_slug*, quedan 1038 cepas comunes.

Por último, dado que se iba a utilizar un algoritmo de árboles de decisión más adelante (ver capítulo 4), era importante que la tabla no tuviera valores nulos en sus filas. Las columnas de condiciones, efectos y sabores no tenían, pero las de cannabinoides y terpenos sí, y en cantidad variable. Para decidir con qué variables quedarse, primero se ordenaron de mayor a menor en cantidad de nulos, y luego se hizo un gráfico en el cual se mostraba la cantidad total de filas sin valores nulos en función de las variables eliminadas. Es decir, se fueron eliminando de a una las columnas de la tabla, en forma aditiva (la primera, la primera y la segunda, etc.) y, por cada caso, se guardó el número total filas sin nulos en la tabla. Habiendo 169 columnas en la tabla completa, no fue posible mostrar la curva entera; en la figura 2.9 se presenta el tramo inicial, hasta que comienza a estabilizarse: a partir de *cbg* en adelante, la cantidad de filas no nulas es la misma. Buscando un compromiso entre cantidad de cepas (i.e. filas en la tabla) y de características (las columnas), se eligió como punto de corte *menthol_y*[6], quedando así una tabla final de 790 filas y 143 columnas como punto de partida para los análisis.

[6]Corresponde al terpeno *menthol*. Como hay un sabor *menthol*, al hacer la unión de las tablas, `python` le asigna _x y _y al final del nombre.

2.2. Métodos

2.2.1. *Principal Component Analysis* (PCA)

Principal component analysis (PCA) es un algoritmo que descompone los datos en vectores (ortogonales) que explican, de mayor a menor, la varianza de los datos. Es decir, el primer vector contendrá datos que expliquen una mayor proporción de la varianza que el segundo, y así siguiendo. Con esta técnica se generan nuevas variables y, al conocer cuanto de la varianza explica cada una, se puede reducir la cantidad de variables sin perder información importante y, además, al ser ortogonales entre sí, se asegura que son independientes. Estas dos características son particularmente útiles pues hay modelos de *machine learning* que presuponen que las variables son independientes (las regresiones lineales, por ejemplo) o la cantidad de variables aporta ruido al modelo en vez de información. PCA es una de las tantas técnicas existentes para identificar las variables más importantes de un conjunto de datos. Por su capacidad de reducción de dimensionalidad, es también utilizado para proyectar datos multidimensionales (3+) a dos o tres dimensiones (en el espacio de componentes de PCA) para poder visualizar la forma de los datos, que fue lo que se hizo en la figura 2.7. Una desventaja que tiene el uso de este algoritmo es que las variables dejan de ser explicativas, pues se pasa a un espacio de componentes en el cual las variables originales están mezcladas, y ello complica la interpretación de los modelos.

2.2.2. *K–Means*

Ver sección Métodos/K–Means del capítulo “Identificación de relaciones y detección de comunidades”.

Identificación de relaciones y detección de comunidades

Una vez definido el conjunto de datos a utilizar, que contenía información respecto de las condiciones médicas para las cuales los usuarios habían declarado consumir cannabis, los efectos y sabores percibidos, y las mediciones de concentraciones de cannabinoides y terpenos por cada cepa, se procedió a analizar las relaciones existentes entre los distintos "pares de variables". Se sabe que los efectos están conectados con la presencia de cannabinoides, así como los sabores con los terpenos[7]; el objetivo fue reproducir los resultados conocidos[7], y aplicar el método a combinaciones no estudiadas previamente. Del *dataset* completo se generan los siguientes subconjuntos:

Condiciones – Efectos
Condiciones – Sabores
Condiciones – Cannabinoides
Condiciones – Terpenos
Efectos – Sabores
Efectos – Cannabinoides
Efectos – Terpenos
Sabores – Cannabinoides
Sabores – Terpenos
Cannabinoides – Terpenos

Por cada uno de estos conjuntos se repitió el mismo proceso. En primer lugar, se calculó la correlación de Spearman entre los elementos de cada conjunto (por ejemplo, para Condiciones–Efectos, se computaron las correlaciones de todas las condiciones con todos los efectos) y el resultado se plasmó en un mapa de calor para poder identificar rápidamente qué elementos están correlacionado (positiva o negativamente) y cuáles no. A continuación, con la matriz de distancias basadas en correlaciones se entrenó el modelo de `K-Means` para obtener grupos de elementos similares. En paralelo, a partir de la matriz de correlaciones, se hallaron comunidades, utilizando el algoritmo de Louvain, y se visualizaron en forma de grafos. Para más

detalle de los pasos seguidos en cada etapa y explicaciones de los métodos empleados, ver la sección Métodos (3.1) a continuación. En la sección Resultados (3.2) se presentan y comentan los resultados obtenidos.

3.1. Métodos

3.1.1. Correlación de Spearman

Spearman[10][1] es un método no paramétrico para calcular las correlaciones entre dos o más variables, es decir, cómo cambia una en relación a la otra. En particular, este método se utiliza para comparar variables que sean ordinales u ordenadas por rango. El coeficiente de correlación de una población se identifica con la letra ρ, y hay dos tipos de relaciones entre variables:

- Directa: el valor se encuentra entre 0 y 1. Es decir, si una variable crece, la otra crece
- Inversa: el valor se encuentra entre -1 y 0. Es decir, si una variable crece, la otra decrece

El método de Spearman no utiliza los valores de las variables, sino su rango, para computar las correlaciones. Dada una variable, se ordenan sus valores de menor a mayor y se construye una columna cuyos valores están entre $(1, \infty)$. En caso de que haya valores repetidos ("un empate"), se calcula el promedio entre el rango que ocuparían y la cantidad de elementos que sean, y ese será el valor del rango[2] (figura 3.1)

Para calcular las correlaciones se utiliza la fórmula 3.1 si no hay empates y la fórmula 3.2 si los hay

$$\rho = 1 - \frac{6 \sum D_i^2}{n(n^2 - 1)} \tag{3.1}$$

$$\rho = \frac{(n^3 - n) - 6 \sum D_i^2 - [(T_x + T_y)/2]}{\sqrt{(n^3 - n)^2 - (T_x + T_y)(n^3 - n) + T_x T_y}} \tag{3.2}$$

[1]Se utilizó el algoritmo implementado en la librería `scipy` de python (`https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.spearmanr.html`)

[2]Por ejemplo, sean dos elementos iguales que debieran ocupar los rangos 6 y 7. Se calcula $\frac{6+7}{2} = 6{,}5$, y ese será el valor que aparezca en la columna de rangos.

TABLE 1.5

Students Who Ate Breakfast	Students Who Skipped Breakfast
90	75
85	80
95	55
70	**90**

(a) Tabla con dos columnas. En ambas hay un valor que se repite.

TABLE 1.6

Value	Rank Ignoring Tied Values	Rank Accounting for Tied Values
55	1	1
70	2	2
75	3	3
80	4	4
85	5	5
90	**6**	**6.5**
90	**7**	**6.5**
95	8	8

(b) Resultado de calcular los rangos teniendo en cuenta, o no, los valores "empatados".

Figura 3.1: Ejemplo del cálculo de los rangos cuando se tienen valores "empatados" en las columnas. Tablas tomadas del libro *Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach*[10].

donde n es el tamaño de la muestra,

$$T_x, T_y = \sum_{i=1}^{g} (t_i^3 - t_i), \tag{3.3}$$

g es el número de grupos empatados en la variable y t_i es el número de valores empatados dentro de cada grupo. Si no hay empates en una variable, $T = 0$.

Previo a calcular las correlaciones, se establecen las hipótesis nula y alterna:

- Hipótesis nula H_0: no existen diferencias entre las variables,
- Hipótesis alterna H_A: existen diferencias o hay relación entre variables.

Un valor de ρ distinto de 0 no significa automáticamente que exista una correlación no nula entre las variables (y que por lo tanto pueda rechazarse la hipótesis nula). Esto es porque incluso para dos variables estadísticamente independientes, se observarán fluctuaciones en el valor de ρ. Se trata, entonces, de estimar la probabilidad de haber observado un tal ρ asumiendo que la hipótesis nula es verdadera (p–valor). Si esa probabilidad es baja (típicamente se usa un umbral de $\alpha = 0.05$) entonces se dice que la correlación es significativa o bien que se rechaza la hipótesis nula con dicho valor de α.

Dadas dos variables, es posible que se encuentren correlacionadas debido al efecto de una tercera variable de no interés. Incluso si el efecto de esa tercera variable es muy pequeño, si las variables que se estudian tienen un número de muestras lo suficientemente grande, entonces se podrá detectar como significativa esa correlación de no interés, dado que el poder estadístico del test de Spearman incrementa con el número de muestras. En otras palabras: a medida que el test estadístico que usamos es mas potente, tiene más capacidad de detectar correlaciones débiles entre las variables, las cuales pueden carecer de interés para el análisis que se lleva a cabo. Una forma de evitar este problema es no enfocarse únicamente en el p–valor (significancia estadística), sino también en el valor de ρ (tamaño de efecto).

Interpretación de los resultados

Las 143 columnas del conjunto de datos (ver final del Capítulo 2) se dividen en 5 grupos de variables: condiciones, efectos, sabores, cannabinoides y terpenos. La información de los primeros tres viene de las opiniones de los usuarios y los restantes

son la concentración de cada químico en cada cepa, medidos en un laboratorio. Cada fila de la tabla corresponde a una cepa de cannabis. Al calcular las correlaciones entre dos subgrupos, subyacentes están las cepas, de modo que si, por ejemplo, una condición está muy correlacionada con un efecto, entonces hay una cepa (o algunas pocas) que recibió muchos más votos en esa condición y en ese efecto que las demás. Del mismo modo, las correlaciones negativas significan que la condición pudo haber sido muy votada, pero el efecto no. En la misma línea, para las correlaciones entre condiciones/efectos/sabores y cannabinoides/terpenos, los atributos más votados se asocian a concentraciones químicas más altas[3] y viceversa.

El resultado de computar las correlaciones es una matriz de dimensión

$$\texttt{dim}(matriz) = [\texttt{dim}(var_1) + \texttt{dim}(var_2)] \times [\texttt{dim}(var_1) + \texttt{dim}(var_2)],$$

donde "var" hace referencia al elemento del par (e.g. $var_1 \equiv$ condiciones, $var_2 \equiv$ efectos). En ella hay una región de dimensión

$$\texttt{dim}(region) = \texttt{dim}(var_1) \times \texttt{dim}(var_1)$$

y otra de

$$\texttt{dim}(region) = \texttt{dim}(var_2) \times \texttt{dim}(var_2),$$

que corresponden a las correlaciones de cada variable consigo misma, y dos regiones de

$$\texttt{dim}(region) = \texttt{dim}(var_1) \times \texttt{dim}(var_2)$$

que contienen las correlaciones entre ambas variables, que es lo que se analiza en este capítulo. Para visualizarlas se optó por un mapa de calor y una asignación de color por cada decil, de modo de poder distinguir mejor las intensidades. Además, dado que en la mayoría de las figuras los ejes x e y contienen muchas variables, para facilitar la lectura se decidió incluir la anotación de cada valor en la celda. Por último, en base a lo discutido al final de la sección "Correlación de Spearman", se decidió incorporar un filtro de módulo 0.1 para el valor de la correlación, con el objeto de evitar discutir correlaciones significativas pero con tamaño de efecto muy pequeño (corresponde a las celdas blancas en las figuras).

[3] En el caso ideal. Se verá que no se cumple en todos los casos.

3.1.2. *K–Means* y el método del codo

`K-Means`[4][11] es un algoritmo que tiene por objetivo agrupar los datos en k clusters distintos y no solapantes (es decir, no existen elementos pertenecientes a más de un *cluster*). Dado un *cluster* C_k, se considera bueno cuando las variaciones intra–*cluster*[5] son lo más pequeñas posibles. En el caso de `K-Means`, la variación intra–*cluster* se define como distancia euclidiana cuadrada, es decir

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \tag{3.4}$$

donde $|C_k|$ es el número de elementos en el *cluster* k. El proceso de determinación de *clusters* que hace `K-Means` es iterativa. En el paso inicial, `K-Means` selecciona aleatoriamente algunos puntos en los datos y los utiliza como centroides. Luego mide la distancia (euclidiana en general) de todos los puntos a esos centroides y los corrige, y continúa así hasta que la variación intra–*cluster* $W(C_k)$ cae por debajo de un umbral (habitualmente se utiliza el establecido por el algoritmo) o hasta que se llegue al número máximo de iteraciones permitidas por el usuario.

Uno de los requerimientos de `K-Means` es que se fije el número de *clusters* k de antemano. Para hallar el k óptimo, en este trabajo se utilizó el "método del codo", que consiste en correr el algoritmo para k en un cierto rango y guardar el valor de una métrica[6] obtenida en cada corrida, para luego graficar esa métrica en función del valor de k (figura 3.2). En el gráfico hay que identificar el punto de inflexión, el punto donde "comienza" el codo, que será el valor de k óptimo. Este método es el más simple y explicable, pero tiene algunas desventajas: no siempre es bien identificable el lugar donde comienza el codo y la elección, si bien existen algoritmos que lo detectan[7], es más bien manual.

Con los perfiles de correlación obtenidos con Spearman, el siguiente paso fue investigar si un algoritmo de *clusterización* como `K-Means` lograba generar grupos de elementos en función de sus correlaciones. Debido a que `K-Means` agrupa elementos

[4]Se utilizó el algoritmo implementado en la librería `scikit--learn` de python (`https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html`)

[5]Las variaciones intra–*cluster* son una medida de qué tan distintas son las observaciones pertenecientes al *cluster*.

[6]En este caso, la métrica es la que implementa el algoritmo de `K-Means` de `scikit-learn`, `inertia_`, que es la suma de las distancias cuadradas al centroide(centro de un *cluster*) más cercano (pesado por los pesos de los elementos, si los tuvieran).

[7]El utilizado en este trabajo es el método `KneeLocator` de la librería `kneed`, `https://kneed.readthedocs.io/en/stable/parameters.html`

en base a su distancia euclidiana, los datos requieren de cierto tratamiento previo a ser enviados al modelo. En primer lugar se computó la matriz de distancias basadas en correlación haciendo la resta de todos los pares *(i, j)* de la matriz y, en segundo, se utilizó el método `MinMaxScaler`[8] para llevar los valores a un rango entre -1 y 1 (elección arbitraria). Una vez hechas estas transformaciones, se entrenó el modelo y se predijeron los *clusters*.

Los algoritmos de aprendizaje no supervisado son ampliamente utilizados, tanto en aplicaciones académicas como comerciales, por su capacidad de detectar patrones en los datos que el análisis humano puede pasar por alto con los métodos clásicos de análisis exploratorio. No obstante, a pesar de su utilidad, tienen la desventaja de que, habitualmente, no es posible contrastar los resultados con datos. Es decir, no suele existir un *target* con el cual los modelos vayan ajustando sus parámetros en la fase de entrenamiento como hacen en el caso de aprendizaje supervisado. Este factor obligó a que se desarrollaran métricas específicas para evaluar la calidad de los *clusters* hallados por los algoritmos, y aún hoy no hay consenso sobre una métrica que sea apta para cualquier situación y las formas de evaluar los resultados dependen de cada problema en particular. Varias de las métricas disponibles se basan en el cálculo de distancias (por ejemplo, Silhouette[9]) y otras aprovechan que a veces se dispone de conocimiento previo del problema, del cual se obtuvieron *targets*, para medir similitud entre los *clusters* hallados y los reales (las llamadas *true labels*). Siempre que sea posible, es una buena idea revisar los resultados manualmente para interpretarlos y, en última instancia, aceptarlos o rechazarlos. En el caso de este trabajo, al buscar clusters a partir de correlaciones (que no son variables sencillas de interpretar), al revisar los resultados surgió la duda de si los *clusters* eran verdaderos o si eran producto de la obligación de `K-Means` de devolver *k clusters*. Para resolverlo, se empleó la métrica de similitud `adjusted rand index (ARI)`[10], que evalúa qué tan similares son los *clusters* fijándose todos los elementos que fueron asignados a los mismos *clusters* en el caso real (las *true labels*) y en el que se desea corroborar. La propuesta fue tomar los resultados como válidos si los *clusters* eran estables, midiendo la estabilidad con `ARI`. Para ello, se tomaron como *true labels* las asignaciones de `K-Means` con el conjunto de datos original entero (la matriz de

[8] `https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html`

[9] `https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html?highlight=silhouette#sklearn.metrics.silhouette_score`

[10] `https://scikit-learn.org/stable/modules/generated/sklearn.metrics.adjusted_rand_score.html`

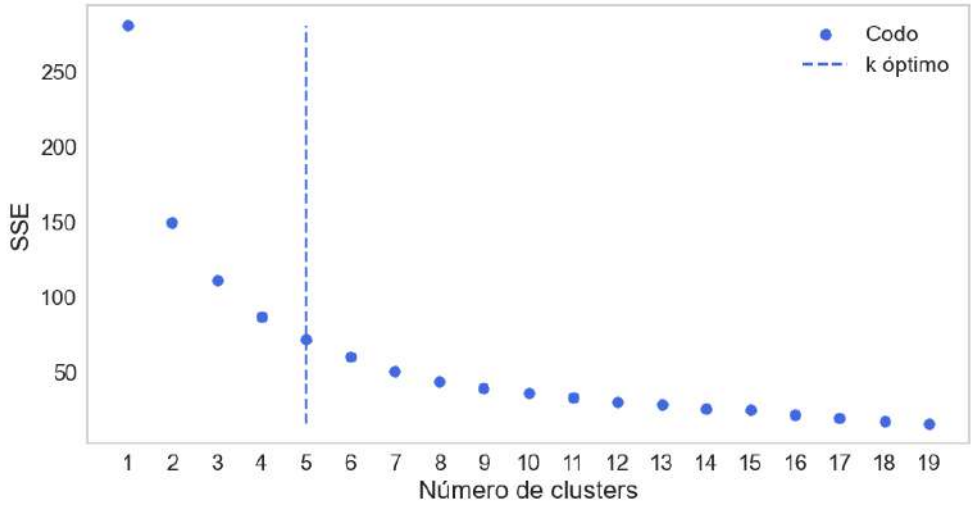


Figura 3.2: Método del codo: se grafica el SSE en función del número de *clusters* k y se toma el punto de inflexión como el k óptimo. En este caso, el algoritmo utilizado para detectar ese punto, lo ubicó en $k = 5$.

correlaciones), y se generaron nuevos resultados con el siguiente procedimiento, repetido un total de 1000 veces: se seleccionó en forma aleatoria un subconjunto que tuviera el 75 % de los datos originales; con ellos se entrenó `K-Means`, y se predijo la pertenencia del 25 % restante, de manera de reobtener el 100 % de los datos. Luego de reordenar las *true labels* para que coincidieran los índices con los de las nuevas etiquetas, se aplicó `ARI` y se obtuvieron las similitudes en cada uno de los 1000 casos. Para `ARI`, un valor cercano al 1 indica que los *clusters* son muy similares, mientras que un valor cercano al 0 da cuenta de resultados distintos.

3.1.3. Detección de comunidades

Dado un sistema complejo, una forma de entenderlo y de identificar sus partes es a través del estudio de comunidades, es decir, cómo los nodos (o vértices) se agrupan en *clusters*, con muchas aristas uniendo los nodos que pertenecen al mismo *cluster* y pocos que conectan con otros. La idea detrás de estas estructuras es que cada comunidad podría considerarse relativamente independiente, cada una presentando características particulares. Una ventaja de esta representación es que es intuitiva y fácil de visualizar; además, parte únicamente de relaciones entre elementos y no requiere información adicional sobre cada uno de ellos. En particular, es fácil construir modelos nulos aleatorizando las conexiones de la red. Por último, y a diferencia

de `K-Means`, la búsqueda de comunidades no suele requerir especificar un número previo, y el algoritmo puede detectar cuantas parezca haber en los datos dada una resolución determinada.

Las comunidades que se verán en la sección Resultados (3.2) se obtuvieron para ambos subconjuntos del par (salvo para terpenos y cannabinoides, por falta de interpretabilidad). Por ejemplo, del par Condiciones–Efectos se generó un grafo[11] no dirigido de condiciones y uno de efectos. En ellos, los nodos eran cada una de las condiciones (o efectos) y las aristas se calcularon con la matriz de correlaciones de Spearman. Más en detalle, se computaron las distancias euclidianas entre todos los elementos *i, j* de la matriz, se guardaron los valores en una matriz d, y se tomó la inversa de esa matriz, dado que las aristas son conexiones pesadas (los valores en la diagonal de d, que eran 0, al tomar inversa se vuelven ∞; se reemplazaron por 0, en tanto serían conexiones que salen y entran al mismo nodo, y no contribuyen al análisis que se quiere hacer).

Para detectar las comunidades se empleó el algoritmo de Louvain[12], el cual busca maximizar la diferencia entre la cantidad real de aristas presentes en la comunidad y el número esperado asumiendo que las aristas están distribuidas al azar preservando el grado (cantidad de aristas conectadas a cada nodo). Esta diferencia se expresa en la modularidad (ecuación 3.5). En ella, m es la suma de los pesos de todas las aristas del grafo, A_{ij} es la entrada ij de la matriz de adyacencia A, que representa el peso de la arista que conecta los nodos i y j, k_i (k_j) es el grado del nodo i (j), c_i (c_j) es la comunidad a la que pertenece.

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j) \tag{3.5}$$

El algoritmo de Louvain[12] se divide en dos etapas que se alternan en forma iterativa. La primera corresponde a la optimización de la modularidad, y la segunda a la agregación de comunidades para formar super–comunidades que las contengan. En la primera etapa, cada nodo constituye su propia comunidad, y el algoritmo calcula cual es el cambio de la modularidad Q (ΔQ) cuando se remueve el nodo i de su comunidad y se lo agrega a la de cada uno de sus vecinos j. Una vez que hizo este proceso con todos los vecinos j, el nodo i es asignado a la comunidad cuyo ΔQ sea mayor. En la segunda etapa, los nodos de la misma comunidad son agrupados

[11] `https://networkx.org`
[12] `https://github.com/taynaud/python-louvain`

y la red ahora es una donde los nodos son las comunidades de la etapa anterior. Aquí, las conexiones entre nodos de la misma comunidad se representan en forma de auto–bucles, y las conexiones con nodos de otras comunidades se expresan con aristas ponderadas. Con esta nueva red creada, el proceso puede volver a comenzar, aplicado a esta red.

Al igual que con `K-Means`, para las comunidades también se realizó una corroboración de la validez de los resultados. Para ello, se *randomizó* la red no dirigida 1000 veces y en cada una se calculó la modularidad Q. Si la modularidad de los datos cuya matriz de adyacencia no fue *randomizada* es mayor que las demás, entonces esos resultados se aceptan como válidos, en tanto se interpreta que no fueron producto del azar.

Por último, para visualizar las redes, se aplicó un valor de corte para incluir únicamente las aristas con peso mayor. Este valor fue definido en cada caso en base a criterio estético, siendo habitualmente uno que filtrara entre el 70 % y el 90 % de las aristas.

3.2. Resultados

El algoritmo descripto en la introducción del capítulo se aplicó a los diez pares de subconjuntos posibles. Sin embargo, no todos arrojaron resultados de los que se pudiera extraer información relevante: en algunos casos, los clusters no identificaron correctamente grupos de elementos, en otros, los grafos no mostraron comunidades separadas, sino que se vio un único grupo de nodos. En primer lugar se mostrarán los pares de subconjuntos de condiciones, efectos y sabores, y luego las combinaciones con los subconjuntos de cannabinoides y terpenos.

Nota 1. En los pares con cannabinoides y terpenos no se presentan los resultados de `K-Means` *ni de comunidades por imposibilidad de interpretación de los mismos. Por ese mismo motivo no se estudió el par cannabinoides–terpenos.*

Nota 2. Dado que las magnitudes de correlación son muy pequeñas y no dan lugar a observaciones relevantes, los resultados de los pares sabores–cannabinoides, efectos–terpenos, condiciones–cannabinoides y condiciones–terpenos no serán presentados.

3.2.1. Condiciones y efectos

Partiendo del cálculo de correlación de Spearman entre condiciones (matriz de 790×40) y efectos (790×19) se obtuvo una matriz de 59×59. En ella se pueden identificar cuatro regiones: una de 40×40 donde se encuentran las correlaciones entre condiciones, una de 19×19 correspondiente a los efectos, y las dos restantes son una de 19×40 y otra de 40×19, ambas con la información de las correlaciones entre condiciones y efectos (una es la traspuesta de la otra). En la figura 3.5 se muestra el mapa de calor de esta última. Lo primero que se nota es que el rango de valores es amplio, permitiendo identificar, sin lugar a dudas, correlaciones y anticorrelaciones entre pares. Lo segundo que se ve es que hay algunas condiciones cuyos valores son, en su mayoría, menores a módulo de 0,1, tales como *eye pressure*, *headaches*, *pms*[13], *tinnitus*, *Tourettes syndrome*, *Parkinsons*, *asthma*, *cachexia*, *chrons disease*, *epilepsy*, *glaucoma*, *hiv aids*, y algunos efectos, como *euphoric* y *aroused*.

Los datos de condiciones y efectos (y sabores) provienen de los votos de los usuarios. En ese contexto, hay que entender las condiciones como "el consumo fue para aliviar la condición *c*" y los efectos como "luego de consumir, se sintieron *e* efectos". Por lo tanto, las correlaciones positivas indican los efectos percibidos con mayor frecuencia, mientras que las negativas hacen referencia a los que menos se sienten. En esa línea, el valor 0,66 entre *depression* y *energetic* significaría que muchos usuarios se sintieron energizados luego de consumir, y el valor $-0{,}7$ entre *depression* y *sleepy* apuntaría a que no suelen sentirse adormilados.

Continuando con los resultados de *depression*, las correlaciones positivas más significativas están dadas por *energetic* (0,66), *creative* (0,6), *uplifted* (0,6), *talkative* (0,55) y *focused* (0,43), y las negativas por *sleepy* ($-0{,}7$), *relaxed* ($-0{,}6$) y *hungry* ($-0{,}47$). La condición *insomnia*, por su parte, correlaciona con *sleepy* (0,83), *hungry* (0,61) y *relaxed* (0,52), y anticorrelaciona con *energetic* ($-0{,}78$), *uplifted* ($-0{,}73$), *focused* ($-0{,}66$), *creative* ($-0{,}63$) y *talkative* ($-0{,}59$). *Fatigue* es otra de las condiciones en las que se encontraron correlaciones significativas: 0,73 con *energetic*, 0,59 con *uplifted* y *creative*, 0,55 con *talkative*, 0,53 con *focused* y $-0{,}68$ con *sleepy*, $-0{,}61$ con *relaxed*, $-0{,}5$ con *hungry*. *Stress* anticorrelaciona con *relaxed* en $-0{,}51$ y está levemente por encima de 0,3 con *creative* y *energetic*. En condiciones relacionadas con dolor (*pain*, *headaches*, *inflammation*, *muscle spasms*) se encuentran correlaciones, en general, por debajo de módulo de 0,3, con la excepción de *relaxed*, cuyo valor

[13]Síndrome premenstual

con *inflammation* es 0,42 y 0,39 con *muscle spasms*, y anticorrelacionan con efectos positivos como *uplifted*, *creative* y *energetic*. Para las condiciones restantes las correlaciones resultaron menores y, con estos últimos dos casos, se está nuevamente en la situación de valores significativos pero asociaciones débiles, que no permiten que se tomen para el análisis esos pares {condición, efecto}.

Como se verá en los resultados de Efectos y sabores (3.2.2) y Condiciones y sabores (3.2.3), los sabores (y por consiguiente, los terpenos presentes en las cepas), están relacionados con los efectos percibidos, entonces las cepas que los usuarios elijan para aliviar la depresión no serán las mismas que para el insomnio.

Los *clusters* identificados por `K-Means` para las condiciones son los que se muestran en la tabla 3.1. En ella se pueden ver que hay 5 grupos: el primero y el último engloban más de diez condiciones sin coherencia aparente; el segundo y el tercero reúnen elementos que podrían estar relacionados (sobre todo *depression* y *fatigue*), y el cuarto mezcla condiciones que, si bien son de salud mental, no necesariamente son agrupables. Por otro lado, los *clusters* de efectos (tabla 3.2) resultaron en cuatro grupos bien definidos. El grupo 3 engloba sensaciones negativas como *paranoid*, *dizzy* y *headache* y el grupo 4 positivas como *energetic*, *creative* y *uplifted*; los efectos del grupo 2 son más bien de sedación, con *relaxed* y *sleepy*, *hungry*, que vendría a ser un efecto neutro, y el grupo 1 es el único en el cual hubo elementos que no pertenecían al *cluster*, pues *anxious* y *tingly* (que deberían pertenecer al grupo 3) están junto a efectos de bienestar y buen ánimo como son *euphoric*, *aroused*, *giggly* y *happy*. Omitiendo las asignaciones de *anxious* y *tingly*, en este caso `K-Means` logró identificar grupos bien definidos, no solo desde el punto de vista positivo–negativo, sino también del tipo de sensaciones, en tanto un grupo son malestares, otro es energía creativa, otro es relajación y el otro es excitación–felicidad.

En el párrafo anterior se presentaron los resultados de aplicar `K-Means` sobre la matriz de distancias basadas en correlaciones de condiciones y efectos; esa matriz tenía dimensión 40×19. Con ella se obtuvieron los *clusters* de efectos, y con la traspuesta, de 19×40, los de condiciones. En el caso de efectos, los grupos son satisfactorios, pero no ocurre lo mismo con los de condiciones. Tal como se detalló en la sección "K–Means y el método del codo", en todos los casos se realizó un análisis de estabilidad de los *clusters* para determinar si los resultados habían sido producto del azar o no, pues podría ocurrir que los datos no tuvieran estructura *clusterizable* y `K-Means` generara *clusters* porque así es como funciona el algoritmo. Para

complementar, se calculó la media, la mediana y la moda del vector de coeficientes `ARI`. En la figura 3.3 se muestran los valores de `ARI` en cada iteración, los cuales se distribuyen en torno a -0.03, indicando que los *clusters* en cada iteración difieren considerablemente de los originales, tomados como "verdaderos". En consecuencia, en este caso no se puede afirmar que los datos tengan una estructura *clusterizable* (lo cual explicaría los grupos indefinidos que se observan en la tabla 3.1) y hay que descartar los resultados hallados. Por el contrario, en la figura 3.4 se encuentran los coeficientes para los *clusters* de efectos, que se ubican entre 0.5 y 1, con la media, mediana y moda entre 0.9 y 1, indicando así que los *clusters* de efectos son estables y se tomarán por buenos. Respecto de los resultados de condiciones, el análisis de estabilidad descarta la certeza de que los datos sean *clusterizables*, pero no afirma que no lo sean. Podría haber otros motivos que hicieran que el modelo no fuera el correcto para estos datos. Podría ocurrir que los perfiles de correlación no fueran suficientemente distintos para ser separados en grupos interesantes aunque, si fuera ese el caso, se debería ver un comportamiento similar en los *clusters* de efectos. Otra opción, de tipo algorítmica, podría ser lo que se conoce como "maldición de la dimensionalidad": `K-Means` es un algoritmo que es sensible a la cantidad de dimensiones del problema, dado que, cuando hay alta dimensionalidad, todos los puntos parecen aproximadamente equidistantes, lo cual no permite que un algoritmo que distingue en base a distancias euclidianas pueda diferenciar puntos entre sí y arme grupos relevantes. Una regla empírica apunta a que habría que tener mínimo 5 registros por cada *feature* o, como mínimo, que los registros superen en cantidad a las *features*, que es lo que ocurre en el caso de efectos, donde hay 40 registros y 19 *features* y no en el de condiciones, donde hay tan solo 19 registros para 40 *features*.

Los resultados de aplicar el algoritmo de Louvain (figuras 3.6 y 3.7) siguieron la línea de los *clusters*: para las condiciones se encontraron tres comunidades, dos que se dividen todos los elementos salvo dos (*fatigue* y *depression*, en gris) y que mezclan condiciones que no están relacionadas. En el caso de los efectos, lo obtenido es más interesante. El algoritmo halló cuatro comunidades bien distinguibles, una correspondiente a estados de ánimo de energía y creatividad (en gris), una de euforia (en celeste), una de estados de relajación (en amarillo) y una de efectos negativos (en rosa). Al igual que con los *clusters*, los efectos negativos *tingly* y *anxious* están ubicados en comunidades a las que no deberían pertenecer pero, a pesar de ello, se considera que los resultados son aceptables. Como se explicó en "Detección de comunidades", se realizó una validación de las comunidades mediante un proceso en

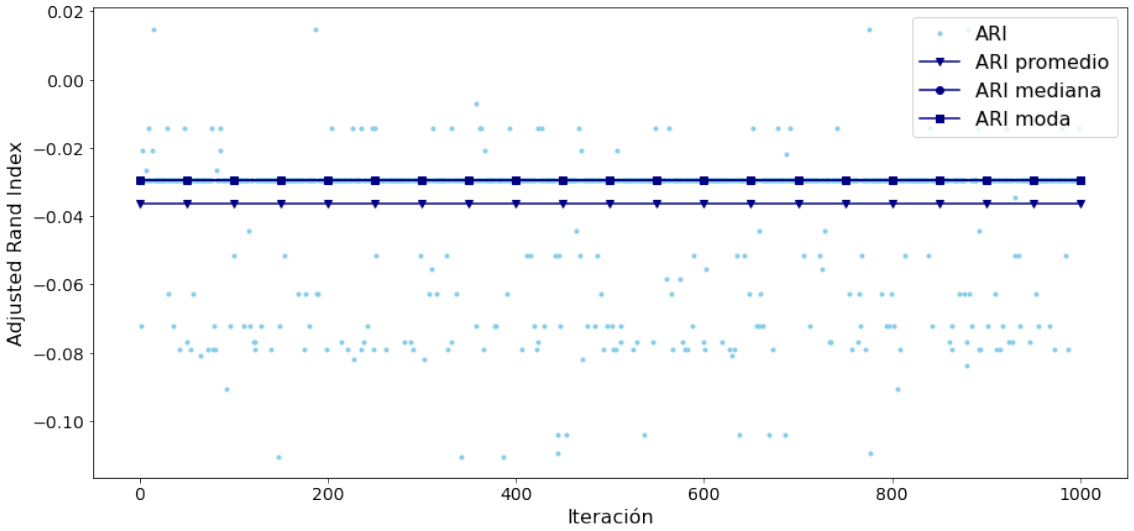


Figura 3.3: Adjusted Rand Index (ARI) para los *clusters* de condiciones en base a las correlaciones con efectos.

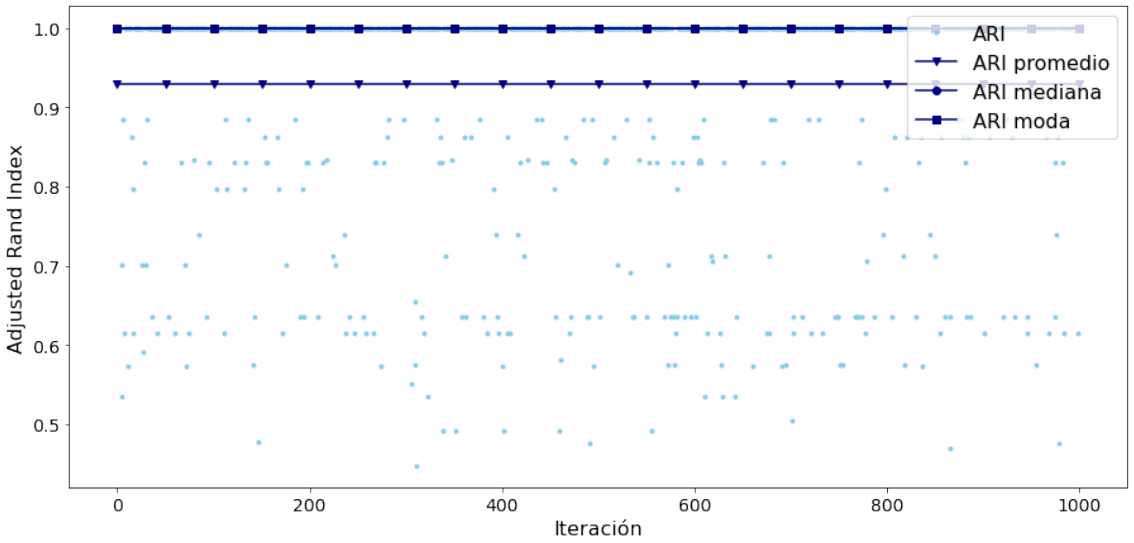


Figura 3.4: Adjusted Rand Index (ARI) para los *clusters* de efectos en base a las correlaciones con condiciones.

el cual, en cada corrida, las aristas eran intercambiadas con otras en forma aleatoria, y en cada una se calculó la modularidad Q de la red, de modo de descubrir si la modularidad de la red original era mayor a las demás y así validar los resultados. En la figura 3.8 está el resultado de reordenar aleatoriamente los pesos de la matriz de adyacencia. En el cuadro superior se muestra para los efectos, y se observa que la modularidad original no es siempre superior a los casos aleatorios. No obstante, tan solo un 6.3 % de los casos tienen Q mayor a la original. Aceptando un margen de error de hasta un 10 %, se acepta el resultado. En el cuadro inferior, por el otro lado, se ve que las modularidades de los casos aleatorios superan habitualmente a la original, de modo que no se pueden considerar válidas las comunidades de condiciones encontradas.

En base a lo obtenido, se puede afirmar el perfil de correlaciones entre las condiciones y los efectos brinda información suficiente para generar clusters y comunidades de efectos con sentido. El contrario no es cierto y, si bien en primera instancia se podría decir que la cantidad de efectos no es suficiente para separar condiciones (pues muchas comparten sintomatología), desde el punto de vista conceptual también es sensato. Por cómo está armada la obtención de los datos, hay una dirección desde condiciones hacia efectos. Es decir, la encuesta pregunta al usuario cuál es la condición que busca "tratar" y cuáles fueron los efectos (y los sabores) sentidos. Es importante recordar que, en el fondo, lo que se tiene son cepas: cepas que "funcionan" mejor o peor frente a una dada situación.

3.2.2. Efectos y sabores

El la figura 3.9 se muestra la matriz de correlaciones de 19×47 generada a partir de los vectores de efectos (790×19) y sabores (790×47). En ella se observa que los efectos *headache*, *anxious*, *euphoric* y *aroused* y los sabores *apple*, *ammonia*, *apricot*, *rose*, *pear*, *peach*, *pepper*, *sage*, *strawberry*, *sweet*, *tar*, *tea*, *mint*, *menthol*, *blue cheese*, *butter*, *cheese*, *chemical*, *diesel*, *honey* y *vanilla* tienen prácticamente todas las celdas en blanco. En general, los valores son bajos, la mayoría por debajo de módulo de 0,2, algunos menores a módulo de 0,4.

En los efectos negativos, *paranoid*, *headache*, *dry mouth*, *dry eyes* y *dizzy*, se ven correlaciones positivas con sabores como *woody*, *ammonia*, *pungent*, *skunk*, *spacy herbal*, *tobacco*, *chestnut* y *earthy*, sabores de tipo fuerte y con cuerpo. En cambio, con sabores como *citric*, *lemon* y *tree fruit* se ven correlaciones negativas. El resto

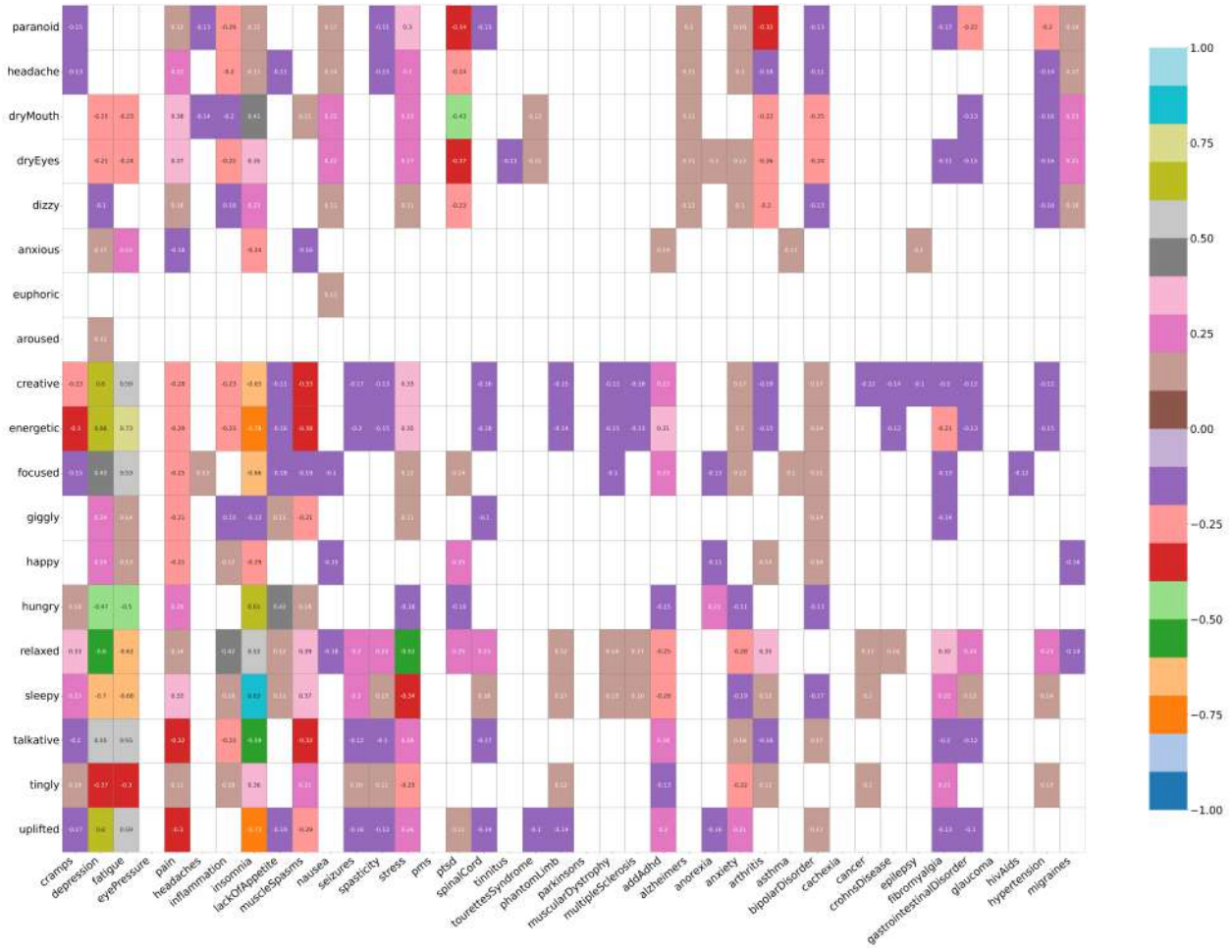


Figura 3.5: Mapa de correlaciones entre condiciones (eje x) y efectos (eje y).

Grupos de condiciones			
Grupo 1	eye pressure	headaches	lack of appetite
	nausea	pms	tinnitus
	Tourettes Syndrome	Alzheimers	anorexia
	asthma	cachexia	epilepsy
	glaucoma	hiv Aids	migraines
Grupo 2	pain	insomnia	muscle spasms
Grupo 3	depression	fatigue	
Grupo 4	stress	add adhd	anxiety
	bipolar disorder		
Grupo 5	cramps	inflammation	spasticity
	seizures	ptsd	spinal cord
	phantom limb	parkinsons	muscular dystrophy
	multiple sclerosis	arthritis	cancer
	crohns disease	fibromyalgia	gastrointestinal disorder
	hypertension		

Cuadro 3.1: Grupos de condiciones identificados por `K-Means` a partir de la matriz de correlaciones.

Grupos de efectos					
Grupo 1	anxious	euphoric	aroused	giggly	happy
	tingly				
Grupo 2	hungry	relaxed	sleepy		
Grupo 3	paranoid	headache	dryMouth	dryEyes	dizzy
Grupo 4	creative	energetic	focused	talkative	uplifted

Cuadro 3.2: Grupos de efectos identificados por `K-Means` a partir de la matriz de correlaciones.

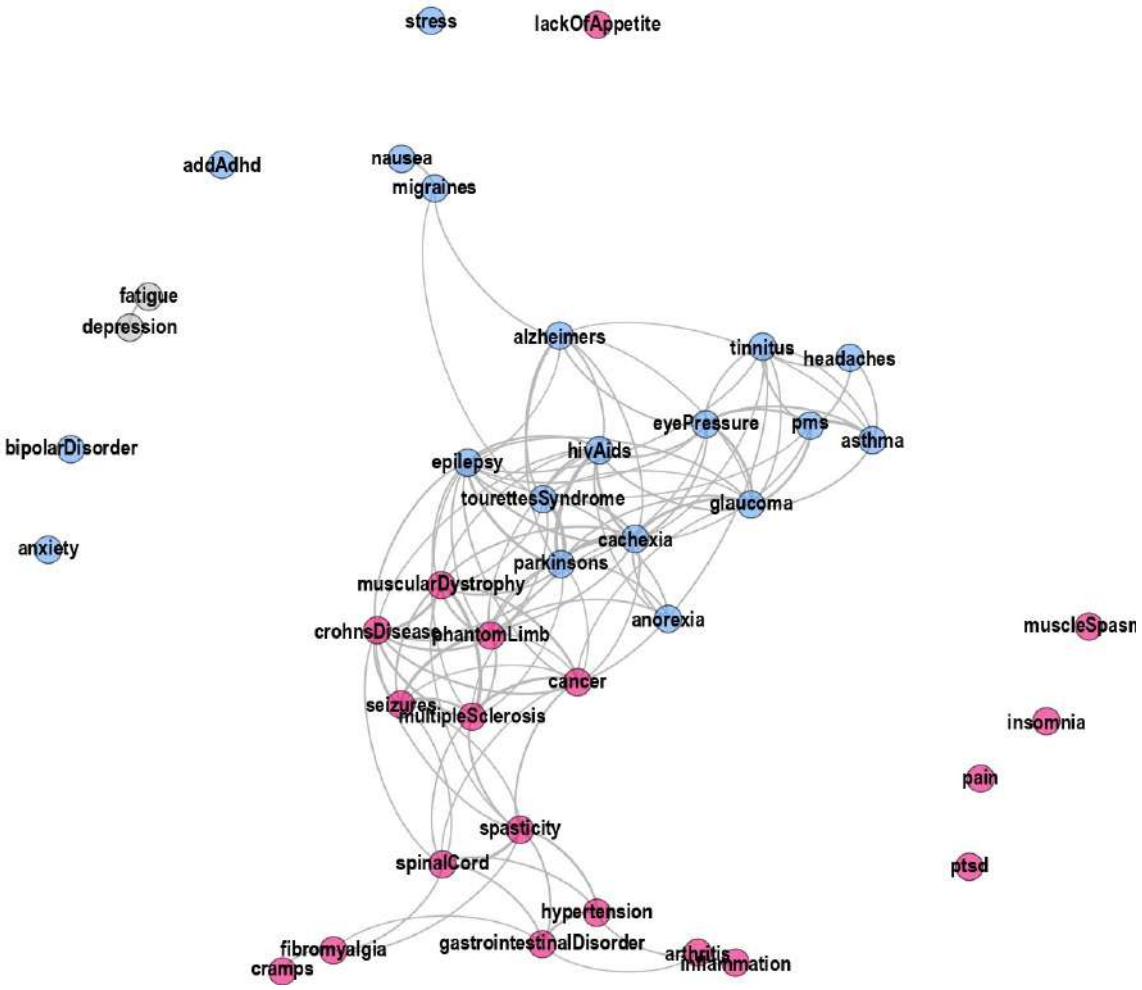


Figura 3.6: Grafo de condiciones calculado en base a las correlaciones entre condiciones y efectos. Se identifican tres comunidades, diferenciadas por color.

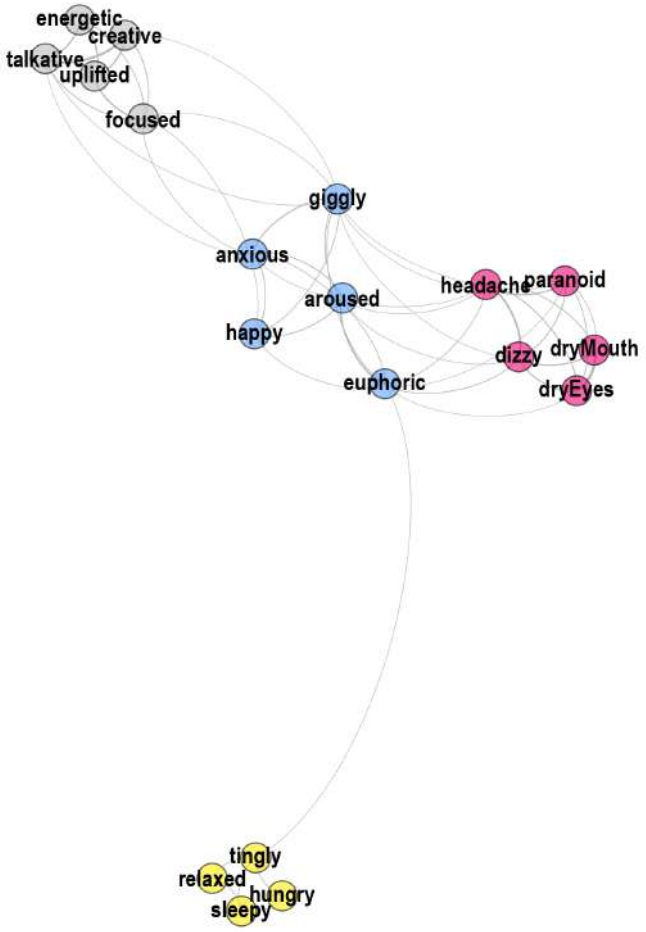


Figura 3.7: Grafo de efectos calculado en base a las correlaciones entre condiciones y efectos. Se identifican cuatro comunidades, diferenciadas por color.

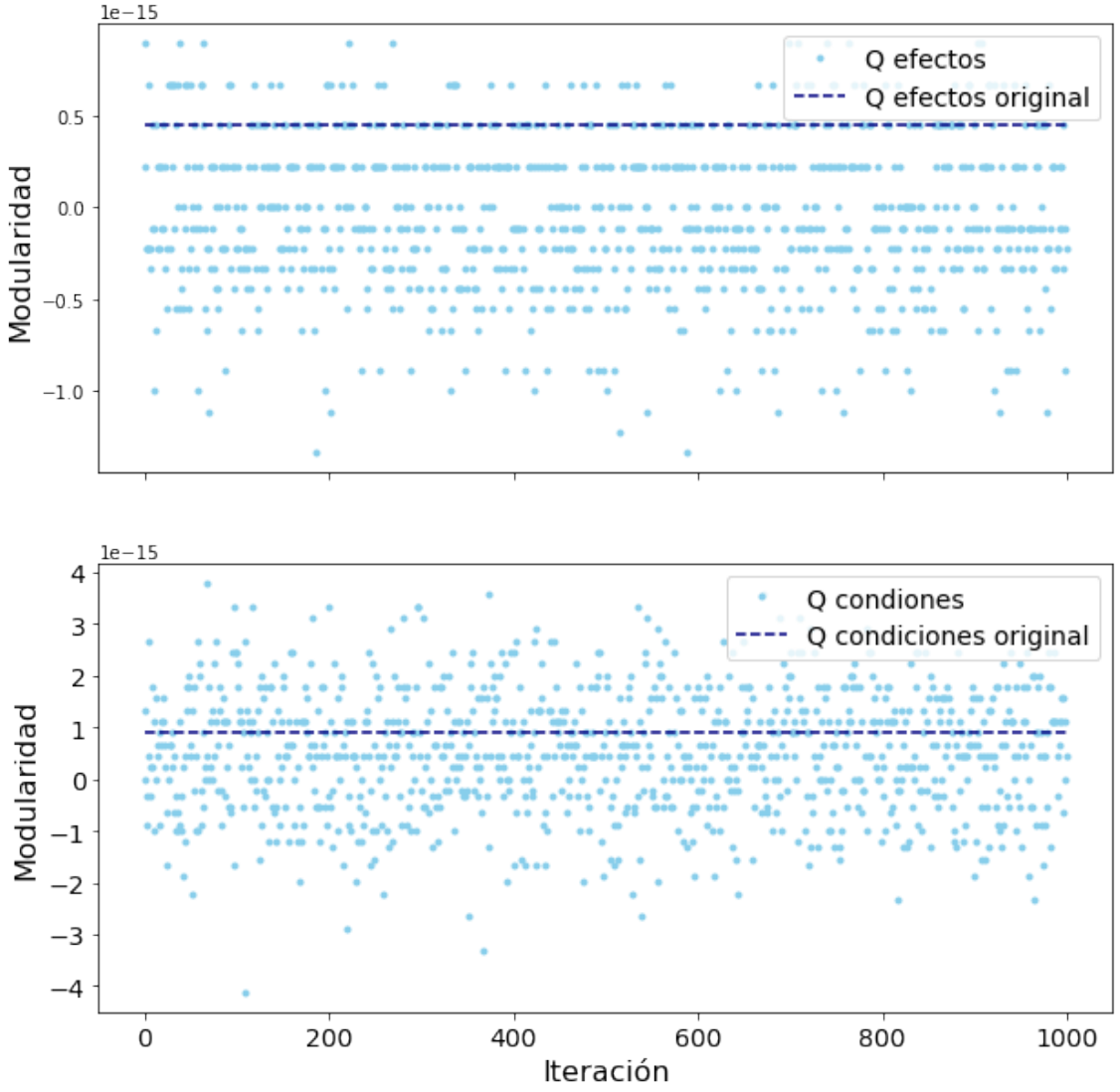


Figura 3.8: Validación de las comunidades obtenidas para las condiciones (cuadro superior) y efectos (cuadro inferior).

de los efectos se pueden dividir en efectos de felicidad/energía y de relajación. Los de relajación (*relaxed*, *sleepy*) tienen valores positivos con *violet*, *woody*, *pine*, *berry* (0,27 y 0,28 respectivamente), *nutty*, *blueberry*, *coffee*, *earthy* (0,29 y 0,32), *flowery*, *grape* (0,25 y 0,33) y *lavender*, y valores negativos con sabores frutales y cítricos, como *tropical* (−0,3, −0,32), *orange* (−0,23, −0,27), *pineapple*, *tree fruit*, *mango*, *citrus* (−0,25, −0,39), *grapefruit*, *lemon* y *lime*. *Hungry*, si bien no sería un efecto de relajación, comparte las correlaciones de *relaxed* y *sleepy*. Los de felicidad/energía (*creative*, *energetic*, *focused*, *happy*, *talkative*, *uplifted*), en cambio, tienen sus correlaciones opuestas al grupo anterior, con valores de hasta 0,4 con *citrus* y 0,34 con *tropical* y de hasta −0,32 con *grape* y −0,3 con *berry*.

En la tabla 3.4 se ven los *clusters* de efectos identificados por `K-Means`. Una vez más, salvo dos elementos fuera de lugar (*paranoid* y *anxious* en el grupo 3), los grupos son coherentes. En el primero se encuentran los estados enérgicos, creativos; en el segundo, los efectos negativos; en el tercero los efectos de excitación (que podrían conformar un único grupo con el 1) y en el último los efectos de relajación. Los *clusters* de sabores se muestran en la tabla 3.3. Los grupos no están bien definidos, pero se puede ver que el grupo 1 contiene varios de los sabores asociados con los efectos negativos, mientras que en el segundo están los asociados con los estados de felicidad/creatividad. En el cuarto se encuentran los sabores asociados con los efectos de relajación, y el tercero es un *cluster* misceláneo. Los análisis de estabilidad se muestran en las figuras 3.10 y 3.11. Al igual que para condiciones y efectos, del primer gráfico se ven los coeficientes `ARI` muy cercanos a cero, lo cual concluye que hay que desestimar los *clusters* de sabores. Para los efectos, los `ARI` se distribuyen alrededor de 0.7, y se aceptan los *clusters*.

Las comunidades halladas se muestran en las figuras 3.12 y 3.13. En la primera, el grafo de efectos muestra tres comunidades, una que comprende las sensaciones de bienestar y energía (en gris), otra con los estados de felicidad y euforia (en celeste) y otra en la que están mayormente los efectos negativos y algunos positivos mal ubicados (en rosa). En la segunda, los sabores están separados en cuatro comunidades, de las cuales tres son mixtas (rosa, amarillo y gris) y en la cuarta se ven gustos frutales tropicales y cítricos. Las simulaciones de validación de los resultados se muestran en la figura 3.14, en la cual se ve que no se pueden aceptar las comunidades de efectos (cuadro superior) pero sí las de sabores (cuadro inferior), dado que apenas un 1 % de los casos aleatorios tuvo una modularidad mayor a la original.

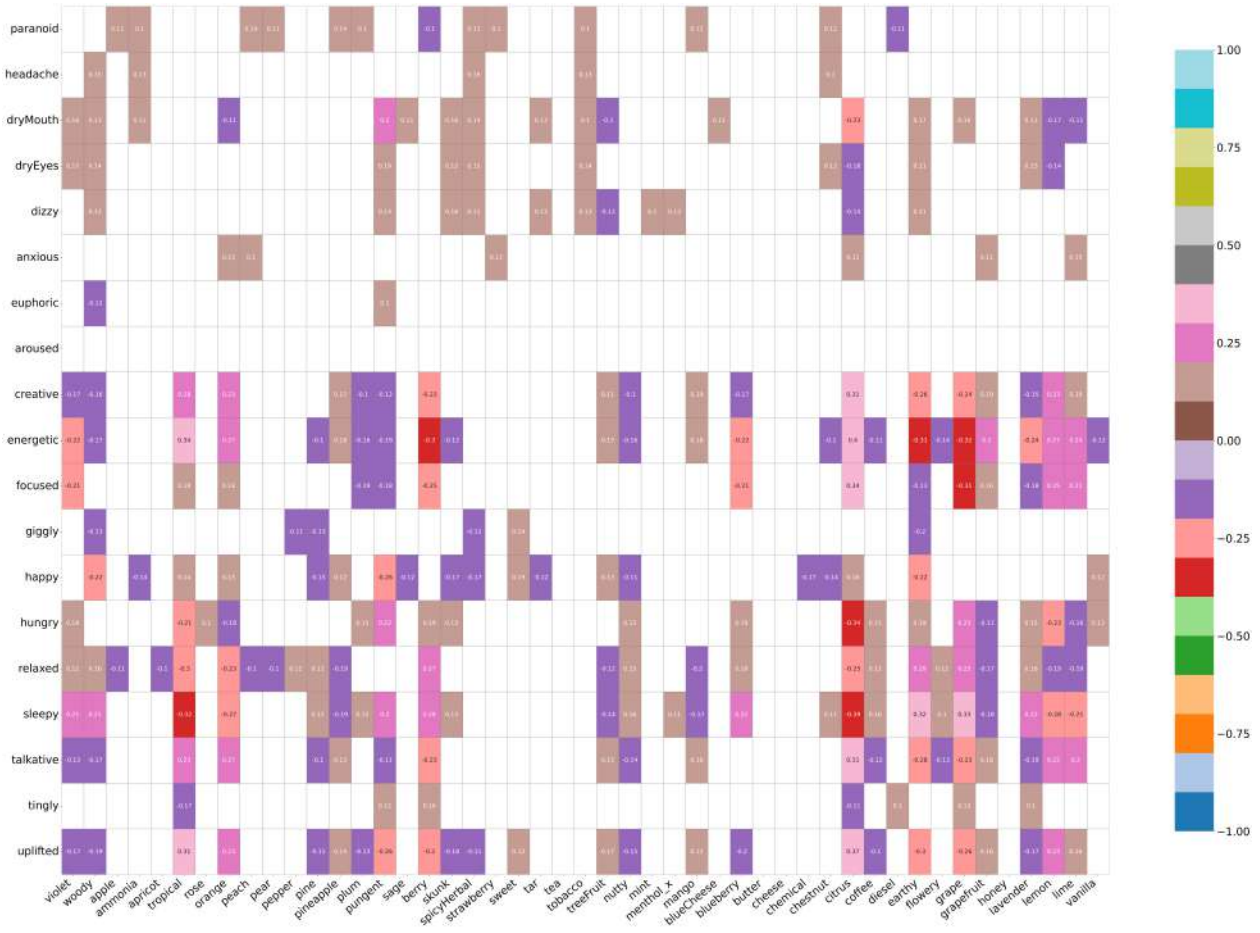


Figura 3.9: Mapa de correlaciones entre sabores (eje x) y efectos (eje y).

Grupos de sabores					
Grupo 1	ammonia skunk tobacco chemical	rose spicy herbal menthol chestnut	pepper tar blue cheese coffee	plum tar butter flowery	sage tea cheese
Grupo 2	tropical tree fruit lemon	orange mango	pineapple citrus	treeFruit grapefruit	sweet lime
Grupo 3	apple mint	apricot diesel	peach honey	pear vanilla	strawberry
Grupo 4	violet nutty	woody blueberry	pine earthy	pungent grape	berry lavender

Cuadro 3.3: Grupos de sabores identificados por `K-Means` a partir de la matriz de correlaciones.

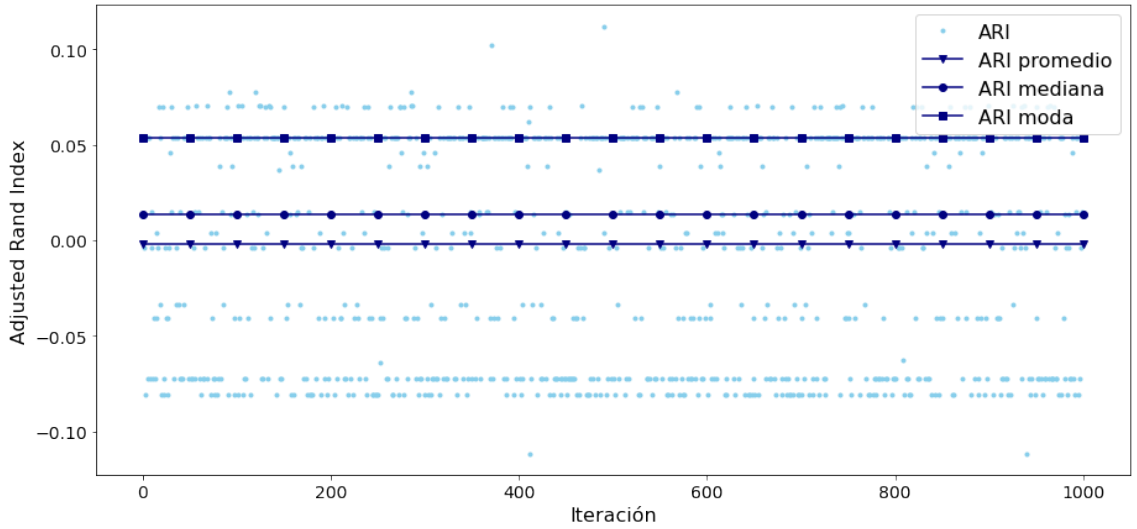


Figura 3.10: Adjusted Rand Index (ARI) para los *clusters* de sabores en base a las correlaciones con efectos.

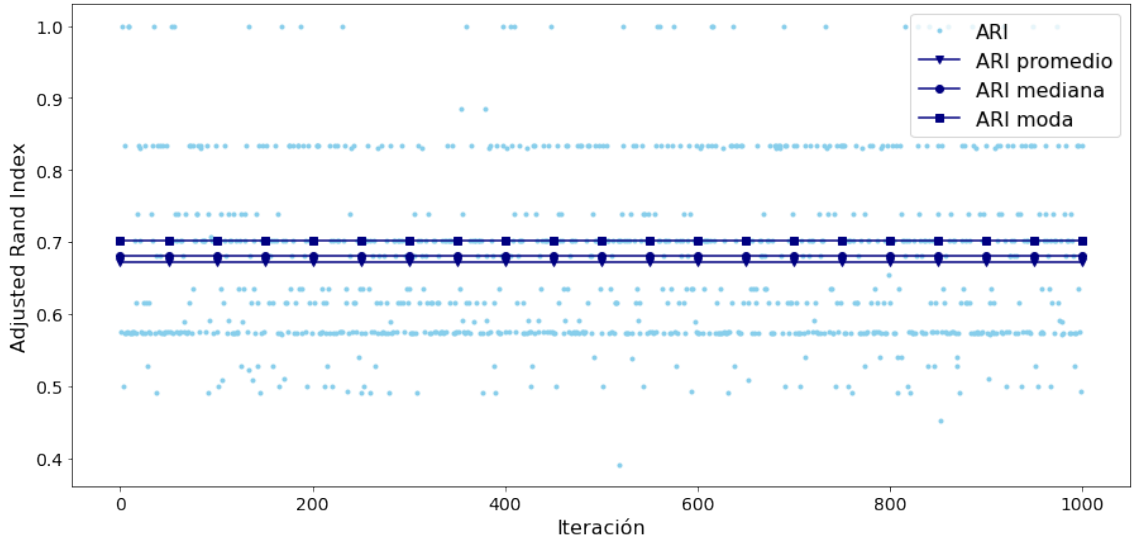


Figura 3.11: Adjusted Rand Index (ARI) para los *clusters* de efectos en base a las correlaciones con sabores.

Grupos de efectos					
Grupo 1	creative	energetic	focused	talkative	uplifted
Grupo 2	tingly	headache	dry mouth	dry eyes	dizzy
Grupo 3	euphoric anxious	aroused	giggly	happy	paranoid
Grupo 4	hungry	relaxed	sleepy		

Cuadro 3.4: Grupos de efectos identificados por `K-Means` a partir de la matriz de correlaciones.

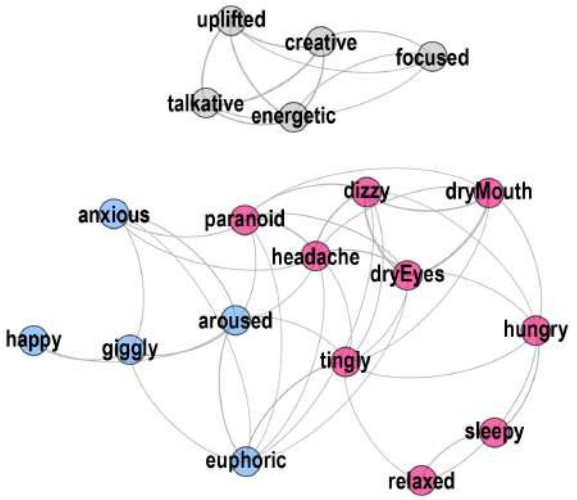


Figura 3.12: Grafo de efectos calculado en base a las correlaciones entre efectos y sabores. Se identifican tres comunidades, diferenciadas por color.

3.2.3. Condiciones y sabores

Del par condiciones (790×40) y sabores (790×47) se obtuvo una matriz de correlaciones (cruzadas) de 47×40 (figura 3.15). Los valores de correlación son bajos, en su gran mayoría menores a módulo de 0,2. Las excepciones están en las condiciones *depression*, *fatigue*, *pain* e *insomnia*. *Depression* y *fatigue* correlacionan positivamente con *tropical*, *orange*, *citrus*, *lemon* y *lime* (0,25, 0,23, 0,31, 0,2, 0,2 y 0,3, 0,23, 0,34, 0,22, 0,23 respectivamente), es decir, sabores frutales y esencialmente cítricos, y negativamente con *pungent*, *berry*, *earthy*, *grape* y *lavender* ($-0{,}2$, $-0{,}23$, $-0{,}32$, $-0{,}23$, $-0{,}19$ y $-0{,}14$, $-0{,}23$, $-0{,}26$, $-0{,}24$, $-0{,}2$ respectivamente). En el caso de *pain*, el valor más alto es de $-0{,}23$ con *citrus* y 0,2 con *earthy*. Las correlaciones en la condición *insomnia* son parecidas en módulo a las de *depression* y *fatigue*, pero con signo opuesto: valores negativos con *tropical*, *orange*, *citrus*, *lemon* y *lime* ($-0{,}28$, $-0{,}21$, $-0{,}37$, $-0{,}26$ y $-0{,}19$); positivos con *pungent*, *berry*, *blueberry*, *earthy*, *grape* y *lavender* (0,22, 0,25, 0,22, 0,26, 0,29 y 0,21). Resulta interesante que condiciones que tienen algunos síntomas opuestos muestren correlaciones opuestas con ciertos sabores. A la vez, si bien los valores son bajos, parecería haber indicios de una relación entre los efectos deseados y los sabores: en los resultados condiciones–efectos[14] se observaron correlaciones positivas entre *depression* y los efectos *energetic*, *creative*, *uplifted*, *talkative* y *focused*, y negativas con *insomnia*, tal como ocurrió con los sabores.

[14]Sección 3.2.1

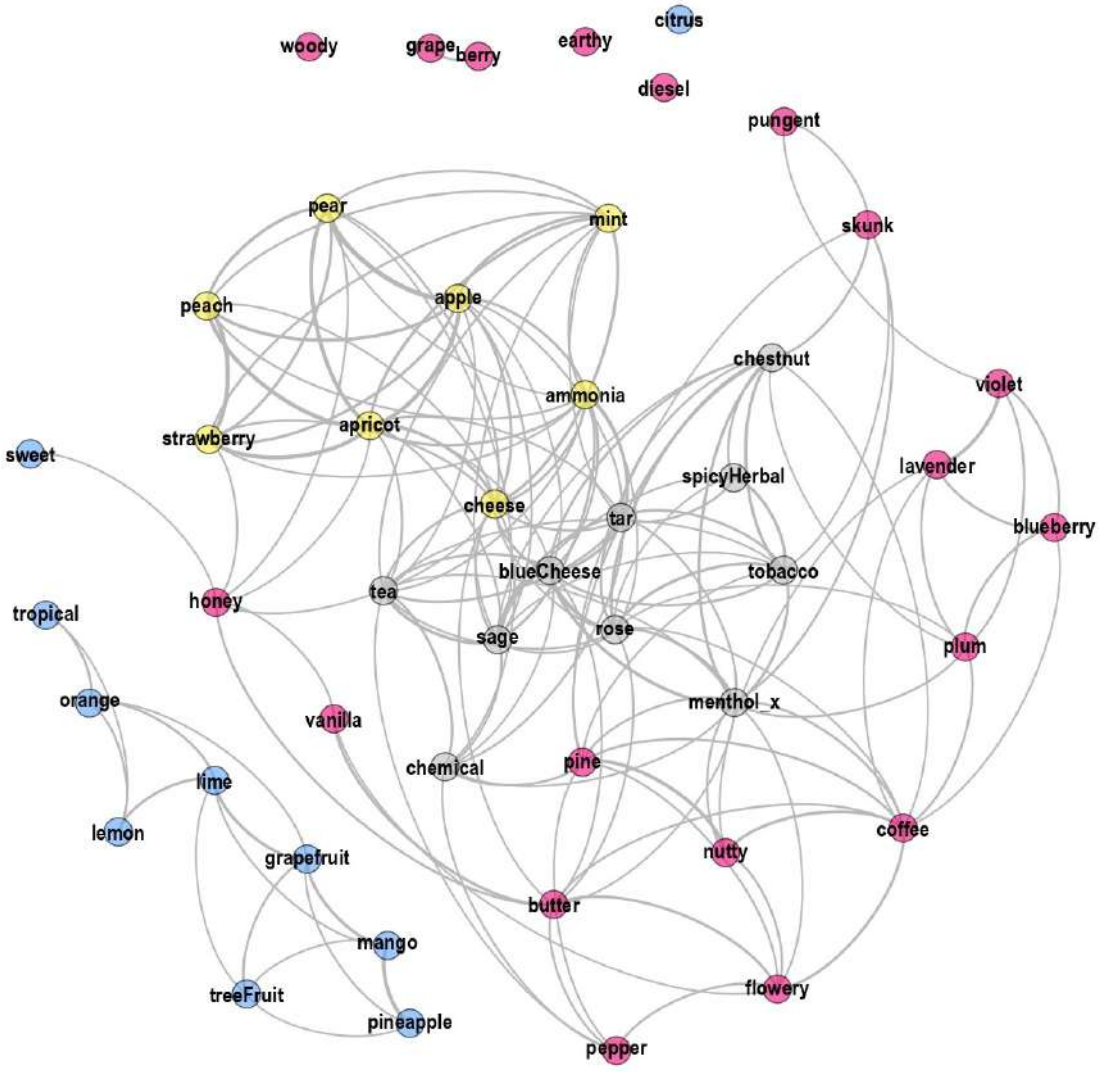


Figura 3.13: Grafo de sabores calculado en base a las correlaciones entre efectos y sabores. Se identifican cuatro comunidades, diferenciadas por color.

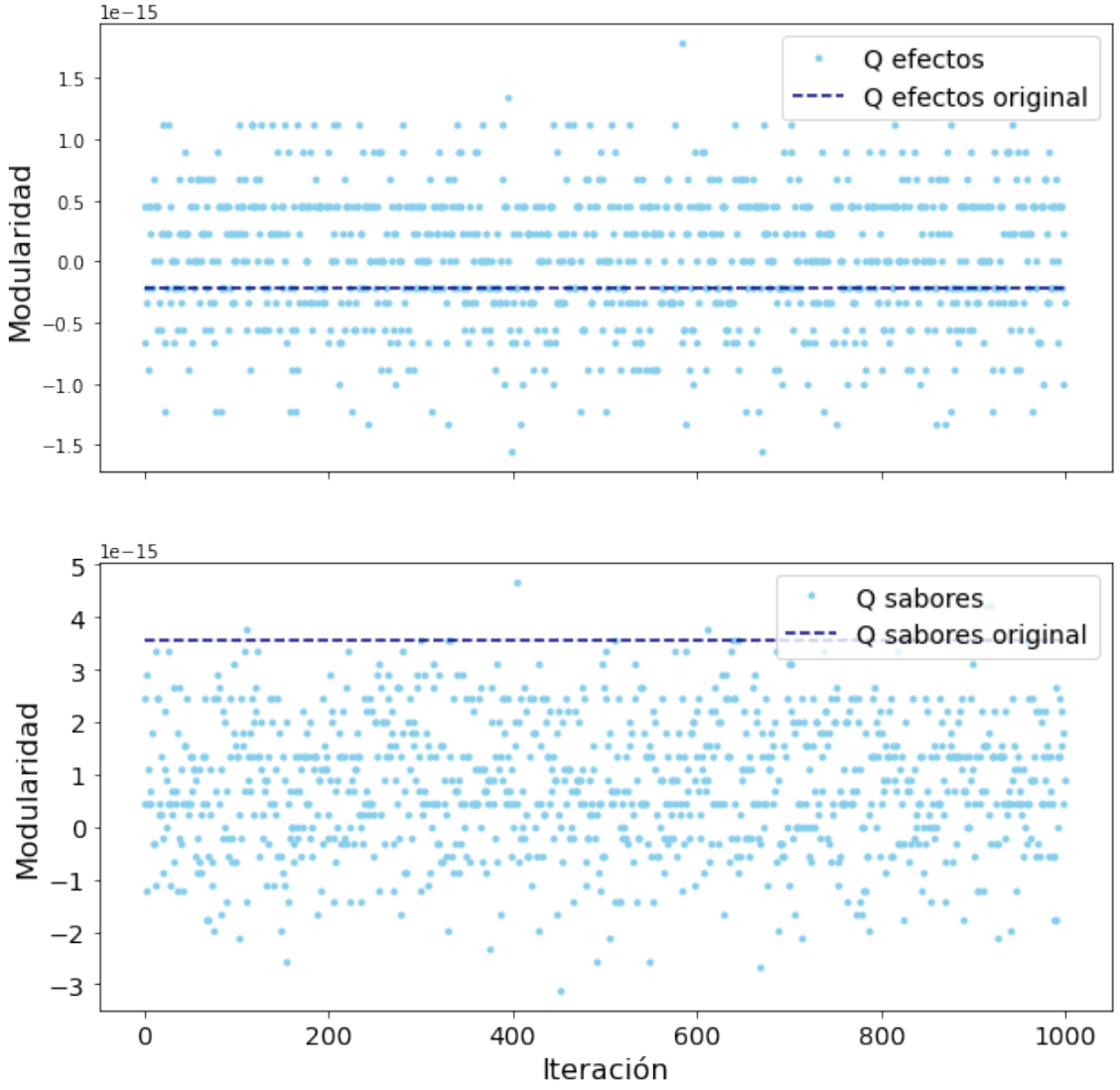


Figura 3.14: Validación de las comunidades obtenidas para las efectos (cuadro superior) y sabores (cuadro inferior).

No obstante la presencia de algunas correlaciones más grandes (aquellas $> 0{,}2$), es importante remarcar que, al haber sido calculados a partir de vectores muy grandes (790×40 y 790×47), varios valores de correlación podrían haber resultado significativos por el elevado poder estadístico. Si bien es positivo que los valores sean significativos, en el caso de este análisis, sería mejor que se evidenciara una asociación relevante entre los datos en los valores de correlación y, si el tamaño de efecto es muy pequeño, resulta que la asociación es débil como para ser interesante, ya que podría surgir por ruido o por factores de confusión que no están siendo tenidos en cuenta en el análisis.

Los resultados de aplicar `K-Means` se pueden ver en las tablas 3.5 y 3.6. Los grupos hallados por el algoritmo son nuevamente misceláneos, tanto los de condiciones como los de sabores. En el caso de condiciones, el grupo 2 reúne cuatro condiciones de salud mental (*depression*, *fatigue*, *stress* y *anxiety*) que tienen elementos en común. En el de sabores, los elementos del grupo 4 son frutas, aunque no están todas, el grupo 5 tiene cuatro sabores asociados con relajación, el grupo 2 junta cítricos con gustos dulces y tropicales, y los grupos 1 y 3 son una mezcla de sabores distintos. Respecto de los análisis de estabilidad, una vez más se tiene que el caso en el que el número de features supera al de registros (los *clusters* de sabores), el `ARI` promedio es cercano a 0 (figura 3.16), indicando que no se puede confiar en los resultados obtenidos. Por otro lado, el `ARI` promedio de los *clusters* de condiciones (figura 3.17) es aproximadamente 0.85, lo cual permitiría concluir que los *clusters* son estables. Este es un buen ejemplo de que estabilidad no implica que los grupos sean significativos en la realidad. Los *clusters* de condiciones, salvo el grupo 2, no parecerían tener un sentido claro; podría ocurrir que el perfil de correlaciones con sabores percibidos no fuera suficiente para distinguir condiciones similares, o también podría ser que, si bien el `ARI` no lo detecta, hubiera alguna influencia de la "maldición de la dimensionalidad", ya que la matriz tiene dimensión 47×40.

En las figuras 3.18 y 3.19 se muestran las comunidades de condiciones y sabores, respectivamente. En el primero hay tres comunidades, dos que se dividen la mayoría de los elementos (rosa y celeste), y una que se encuentra partida (gris): en la zona inferior hay varias condiciones relacionadas con desórdenes y trastornos mentales, pero cuyas conexiones son más débiles que el umbral establecido, y en la zona superior, hacia el centro, hay más elementos, pero de caracter variado. Respecto de los sabores, se hallaron dos comunidades que se reparten aproximadamente la mitad de los elementos cada una, y son ambas un mix indefinido de sabores. La validación

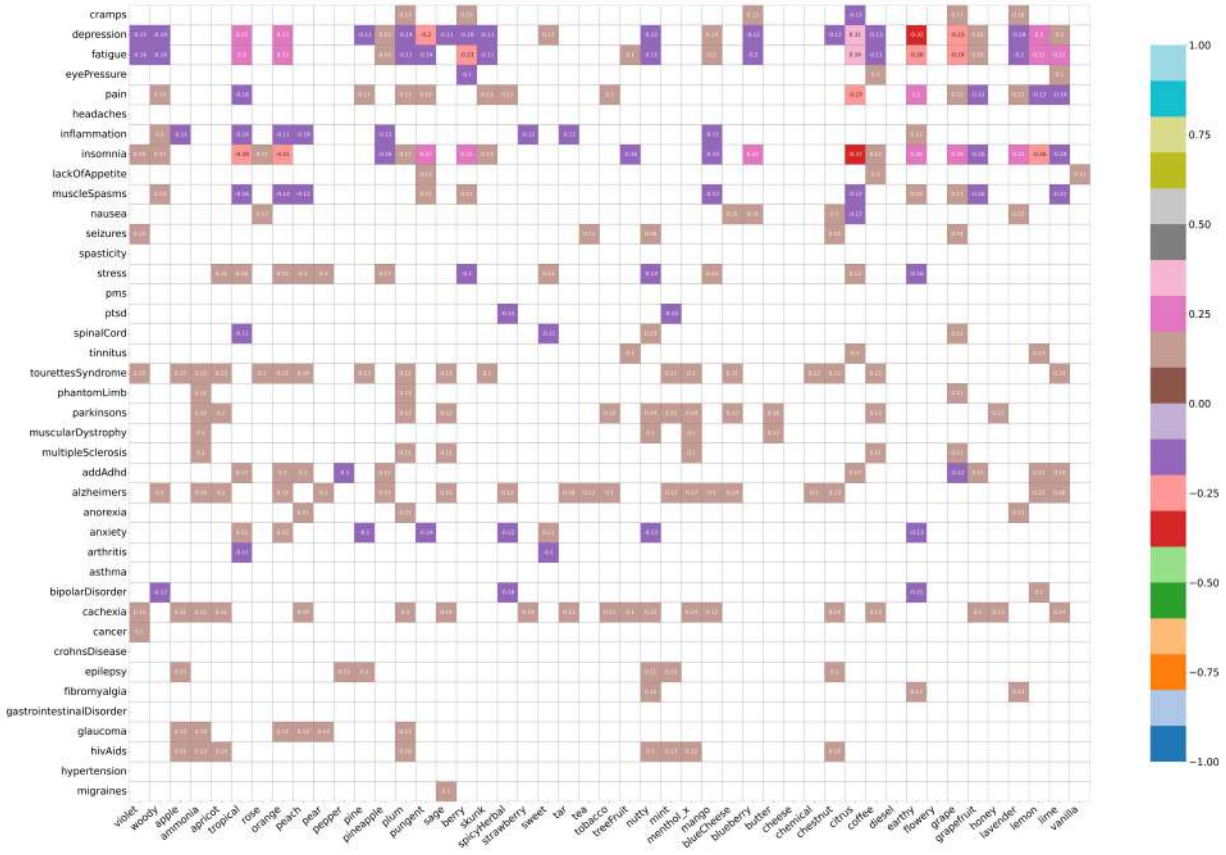


Figura 3.15: Mapa de correlaciones entre sabores (eje x) y condiciones (eje y).

de los resultados (figura 3.20 indica que las comunidades de sabores no pueden ser tenidas en consideración (cuadro inferior), mientras que en las de comunidades hay un 5.9 % de casos aleatorios con modularidad por encima de la original. Dado que este valor se encuentra dentro del margen aceptado (10 %), no se descartan las comunidades de condiciones desde el punto de vista algorítmico, pero en la realidad no aportan información relevante.

3.2.4. Sabores y terpenos

Para el par sabores–terpenos (matriz de 790×47 y 790×28 respectivamente) se obtuvo una matriz de correlaciones de Spearman de 28×47, cuyo mapa de calor se muestra en la figura 3.21. En él se observa que los valores se mueven entre ±0,2 aproximadamente, siendo el más alto 0,26. En función de lo que se ve en el gráfico (y recordando que los datos respectivos a sabores son subjetivos), cada sabor estaría dado por una combinación de terpenos, algunos con mayor presencia que otros, pero sin un predominante en ningún caso.

Grupos de condiciones			
Grupo 1	cramps	seizures	spasticity
	ptsd	spinal cord	phantom limb
	Parkinsons	muscular dystrophy	multiple sclerosis
	cancer	Crohns disease	gastrointestinal disorder
	hiv aids	hypertension	
Grupo 2	depression	fatigue	stress
	anxiety		
Grupo 3	eye pressure	headaches	lack of appetite
	nausea	pms	tinnitus
	Tourettes syndrome	add adhd	anorexia
	Alzheimers	asthma	bipolar disorder
	cachexia	epilepsy	glaucoma
	migraines		
Grupo 4	pain	insomnia	inflammation
	muscle spasms	arthritis	fibromyalgia

Cuadro 3.5: Grupos de condiciones identificados por `K-Means` a partir de la matriz de correlaciones.

Grupos de sabores					
Grupo 1	violet	rose	pine	plum	skunk
	spicy herbal	blueberry	grape	lavender	
Grupo 2	tropical	orange	sweet	citrus	
Grupo 3	ammonia	pepper	sage	tar	tea
	tobacco	nutty	menthol	mint	blue cheese
	butter	cheese	chemical	chestnut	coffee
	diesel	flowery	honey	vanilla	
Grupo 4	apple	apricot	peach	pear	pineapple
	strawberry	tree fruit	mango	grapefruit	lemon
	lime				
Grupo 5	woody	pungent	berry	earthy	

Cuadro 3.6: Grupos de sabores identificados por `K-Means` a partir de la matriz de correlaciones.

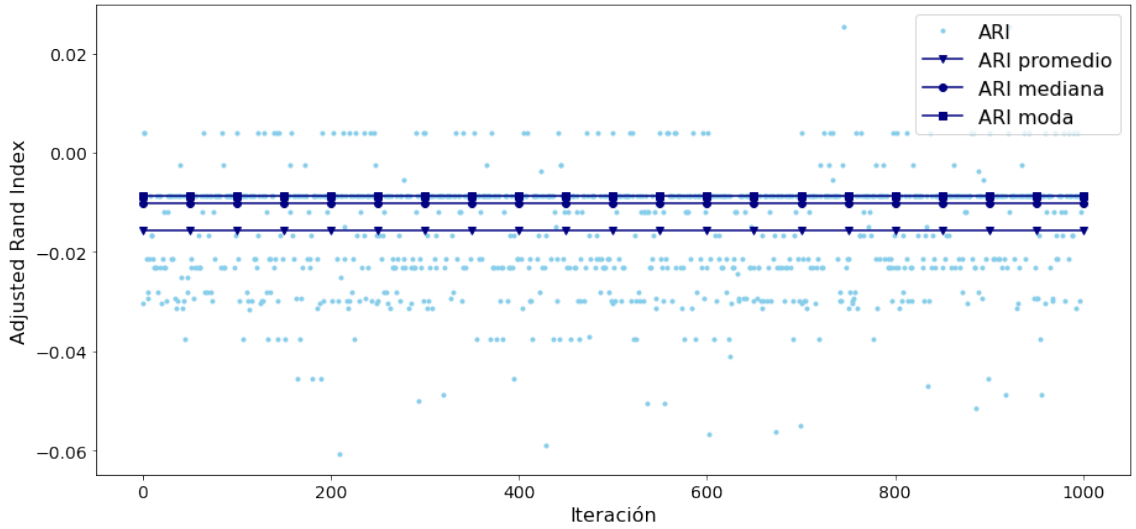


Figura 3.16: Adjusted Rand Index (ARI) para los *clusters* de sabores en base a las correlaciones con condiciones.

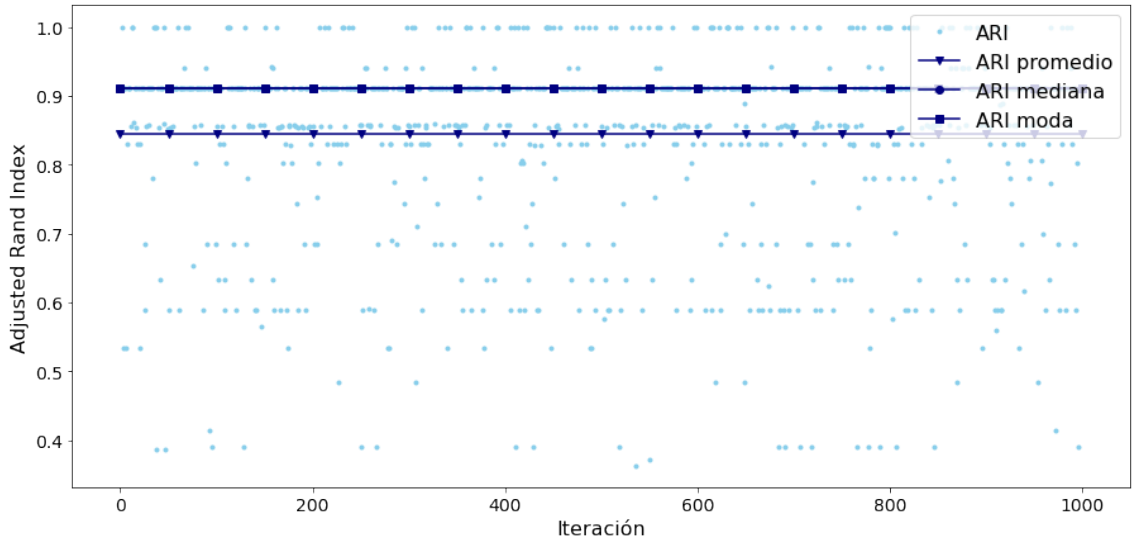


Figura 3.17: Adjusted Rand Index (ARI) para los *clusters* de efectos en base a las correlaciones con sabores.

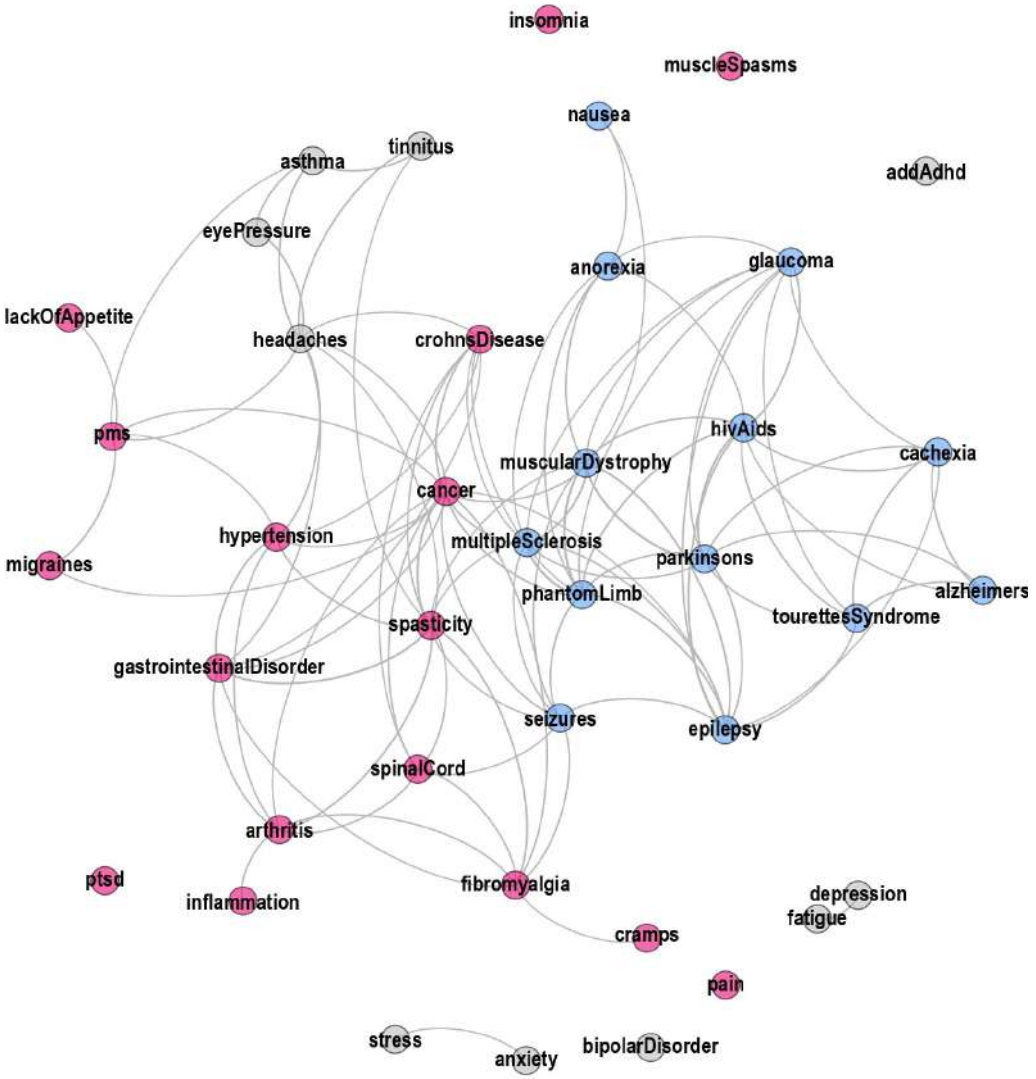


Figura 3.18: Grafo de condiciones calculado en base a las correlaciones entre condiciones y sabores. Se identifican tres comunidades, diferenciadas por color.

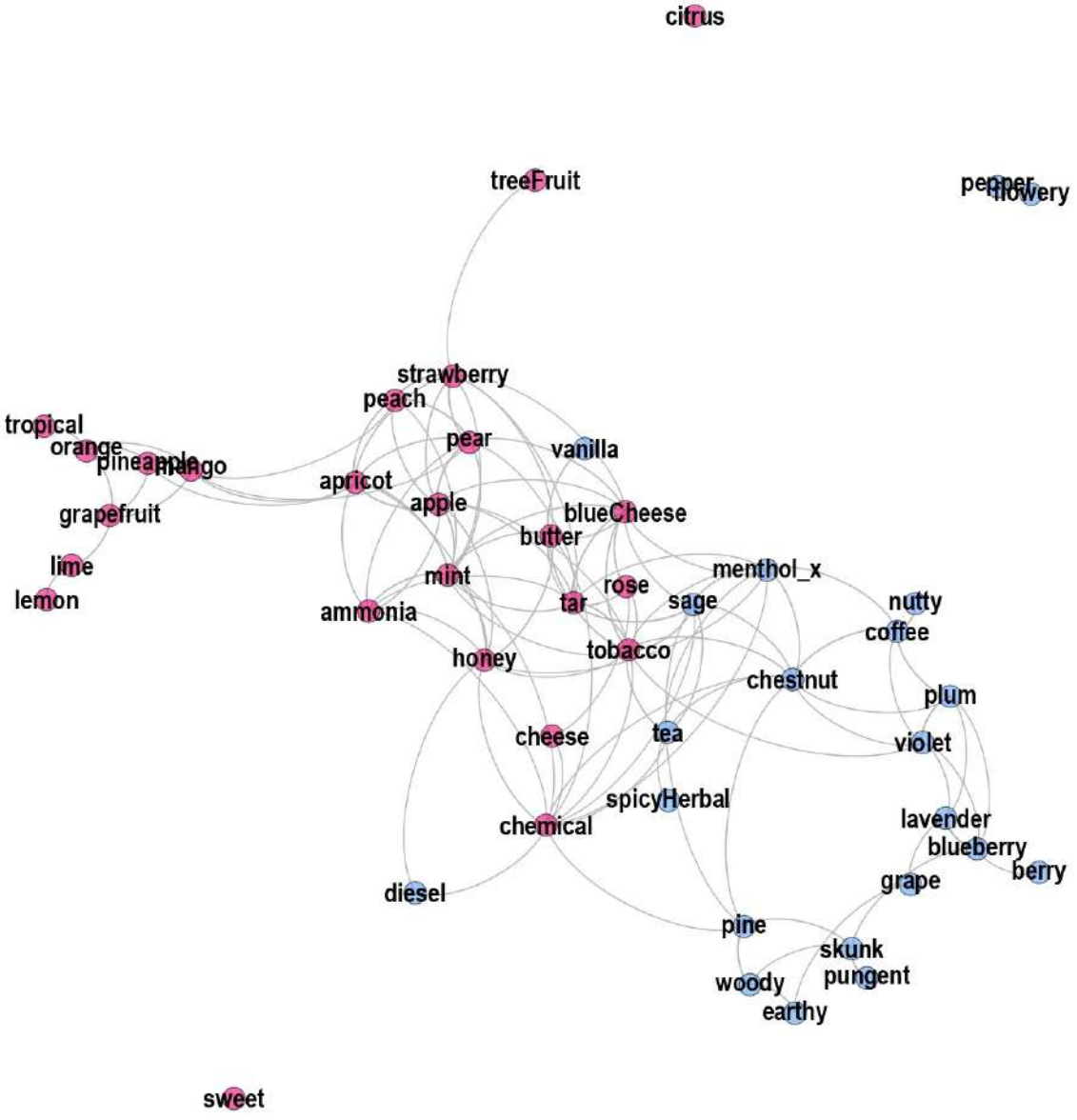


Figura 3.19: Grafo de sabores calculado en base a las correlaciones entre condiciones y sabores. Se identifican dos comunidades, diferenciadas por color.

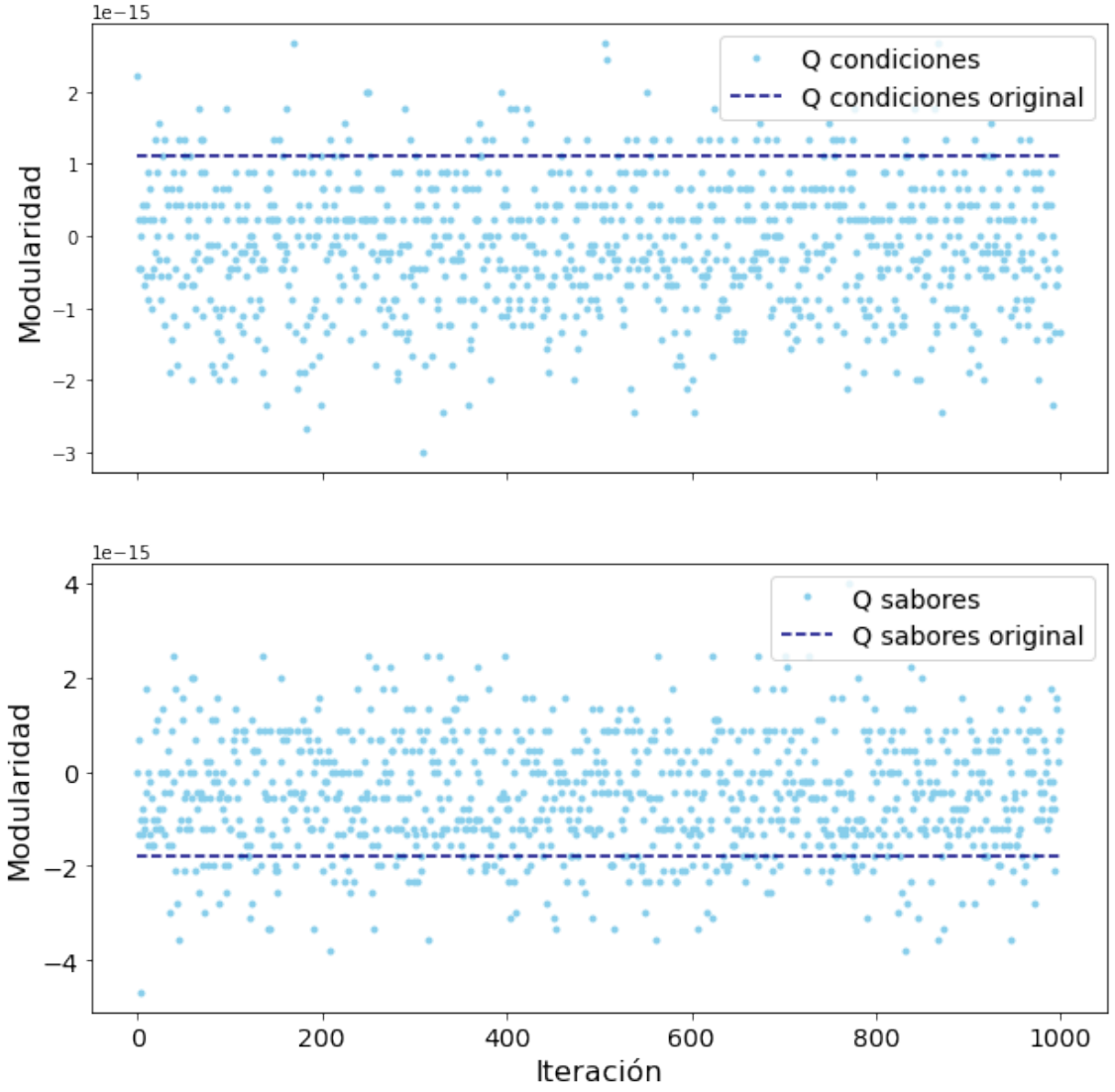


Figura 3.20: Validación de las comunidades obtenidas para las condiciones (cuadro superior) y sabores (cuadro inferior).

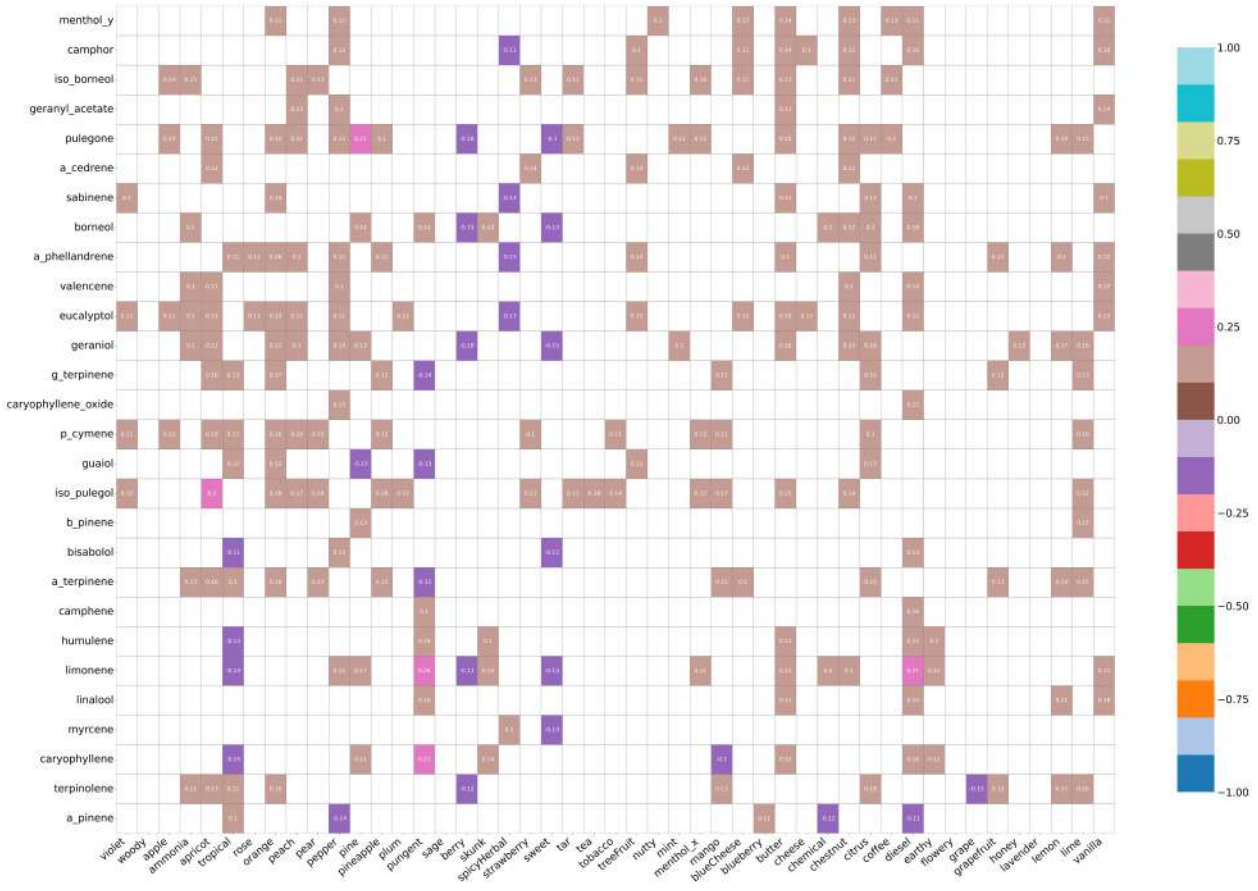


Figura 3.21: Mapa de correlaciones entre sabores (eje x) y terpenos (eje y).

Adicionalmente a que los valores son bajos, las relaciones que podrían esperarse entre terpenos y sabores no siempre se observaron. El ejemplo más claro es el del *limonene*, para el cual las correlaciones con gustos cítricos (*citrus*, *lemon* y *lime*) son menores a 0,1 (no esperable), las anticorrelaciones son con *tropical*, *berry*[15] y *sweet* (esperable), y las correlaciones más fuertes son de 0,26 con *pungent*[16] (esperable) y de 0,21 con *diesel*[17] (aunque no hay evidencia de que el sabor y el terpeno estén relacionados). En el caso del *isopulegol*, un químico utilizado en perfumes para reproducir gustos como damasco, durazno, ciruela, etc, la correlación más fuerte es con *apricot* (0,2), seguido de *peach* y *mango* (0,17) (esperable) y, con 0,18, *pineapple* y *tea*. El *caryophyllene*, que es una de las moléculas que le dan sabor a la pimienta negra, presenta valores de 0,21 con *pungent*, aunque la correlación con *pepper* es menor a 0,1; por otro lado, anticorrelaciona con *tropical* y con *mango* (esperables). El *pulegone* es un líquido aceitoso que, entre otros, tiene olor mentolado. Sin embargo, las correlaciones con *mint* y *menthol* son de 0,13 (estos sabores tampoco presentaron correlaciones relevantes con el terpeno *eucalyptol*), y la más fuerte, de 0,21, es con el sabor *pine*, seguida de *lemon* con 0,19. Respecto de las anticorrelaciones, presenta un valor de $-0{,}18$ con *berry* y $-0{,}1$ con *sweet*.

`K-Means` devolvió cuatro grupos de sabores; los primeros dos tienen muchos elementos y no están bien definidos, aunque el segundo contiene casi todas las frutas disponibles y algunas flores. El tercer grupo reúne los frutos *berry*, *blueberry* y *grape*, el sabor *sweet*, y el sabor *tropical*, que debería pertenecer al segundo. Por último, el cuarto grupo tiene elementos relacionados con tierra, como *woody*, *spicy herbal* y *earthy*. Trabajar con sabores permite determinar rápidamente si los resultados fueron satisfactorios o no, en cuanto se pueden agrupar mentalmente sin problema. En este caso no se considera que los resultados lo sean. En la figura 3.22 se muestran los valores de `ARI` en cada iteración, los cuales se distribuyen en torno a -0.02, indicando que los *clusters* en cada iteración difieren considerablemente de los originales. En consecuencia, en este caso no se puede afirmar que los datos tengan una estructura *clusterizable* y hay que descartar los resultados hallados.

Por último, en la figura 3.23 se muestran las comunidades de sabores. En ella se pueden identificar tres comunidades diferentes pero con elementos mezclados que no definen al conjunto. De la validación (figura 3.24) se desprende que los resultados no

[15]Fruta del bosque
[16]Sabor: fuerte, intenso, punzante
[17]Sabor: combustible

Grupos de sabores					
Grupo 1	ammonia mint chemical vanilla	pepper menthol chestnut	tar blueCheese coffee	tobacco butter diesel	nutty cheese lime
Grupo 2	violet peach strawberry flowery	apple pear tea grapefruit	apricot pineapple treeFruit honey	rose plum mango lavender	orange sage citrus lemon
Grupo 3	tropical	berry	sweet	blueberry	grape
Grupo 4	woody earthy	pine	pungent	skunk	spicy herbal

Cuadro 3.7: Grupos de sabores identificados por `K-Means` a partir de la matriz de correlaciones.

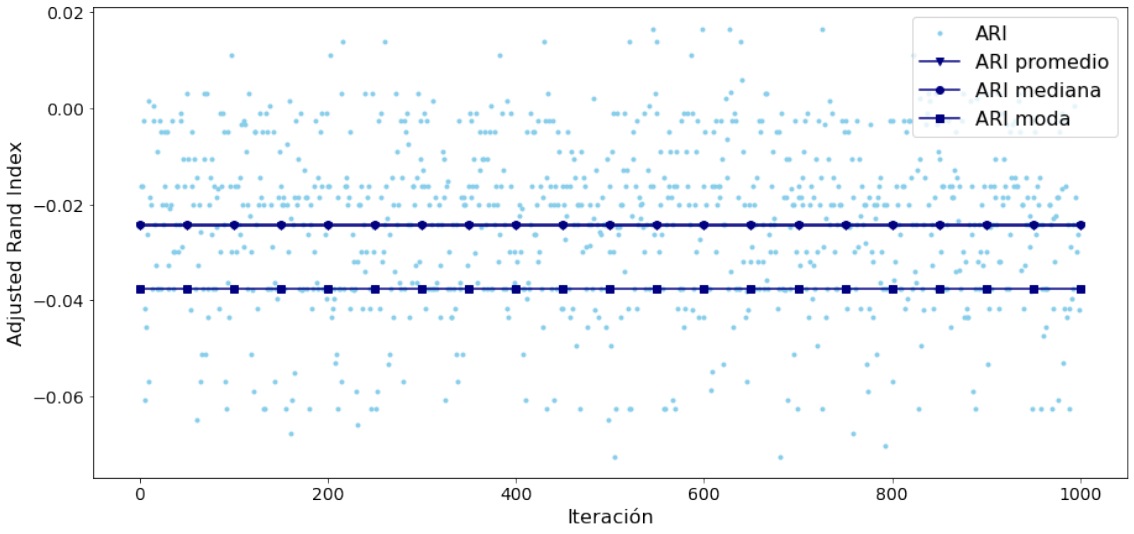


Figura 3.22: Adjusted Rand Index (ARI) para los *clusters* de sabores en base a las correlaciones con terpenos.

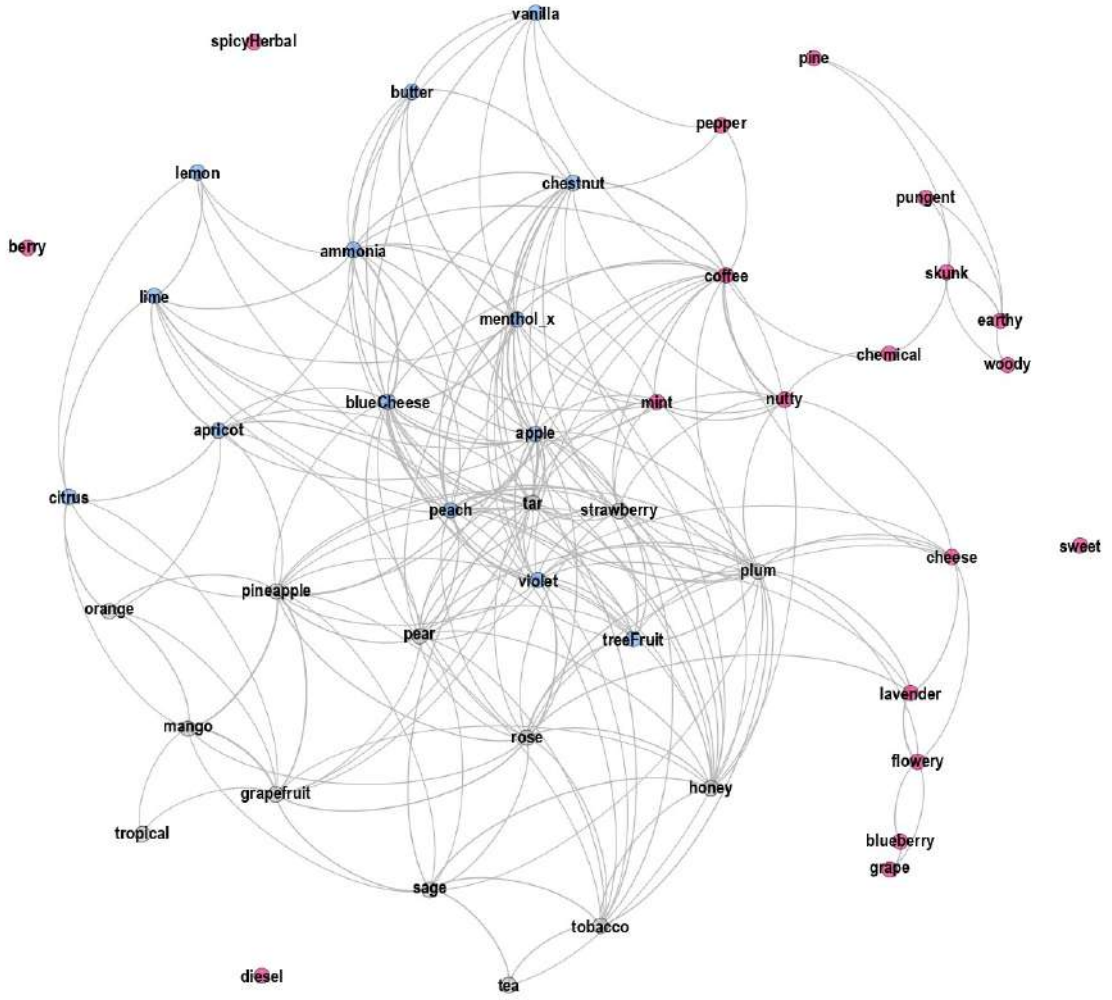


Figura 3.23: Grafo de sabores calculado en base a las correlaciones entre sabores y terpenos. Se identifican tres comunidades, diferenciadas por color.

son confiables, dado que aproximadamente la mitad de los casos aleatorios resultaron en un valor de modularidad mayor al original.

3.2.5. Efectos y cannabinoides

El par efectos–cannabinoides dio lugar a una matriz de correlaciones de 9×19. En la figura 3.25 se observa un mapa de calor con casi todos sus casilleros enmascarados y, los que no lo están, muestran correlaciones muy pequeñas como para dar cuenta de relaciones interesantes entre efectos y cannabinoides. Sabiendo que hay una relación directa entre cannabinoides y efectos percibidos[18], resultó sorprendente no observar correlaciones más definidas: los valores positivos tienen un máximo de 0.19, y los

[18]Ver todas las referencias del capítulo Identificación de sinergias, o [7].

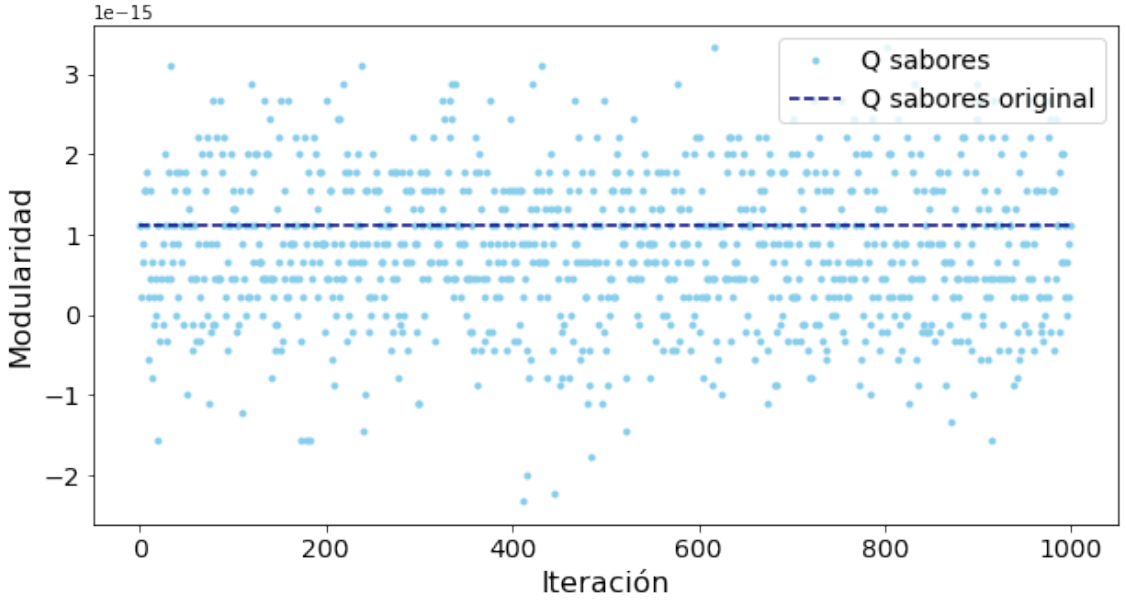


Figura 3.24: Validación de las comunidades obtenidas para los sabores.

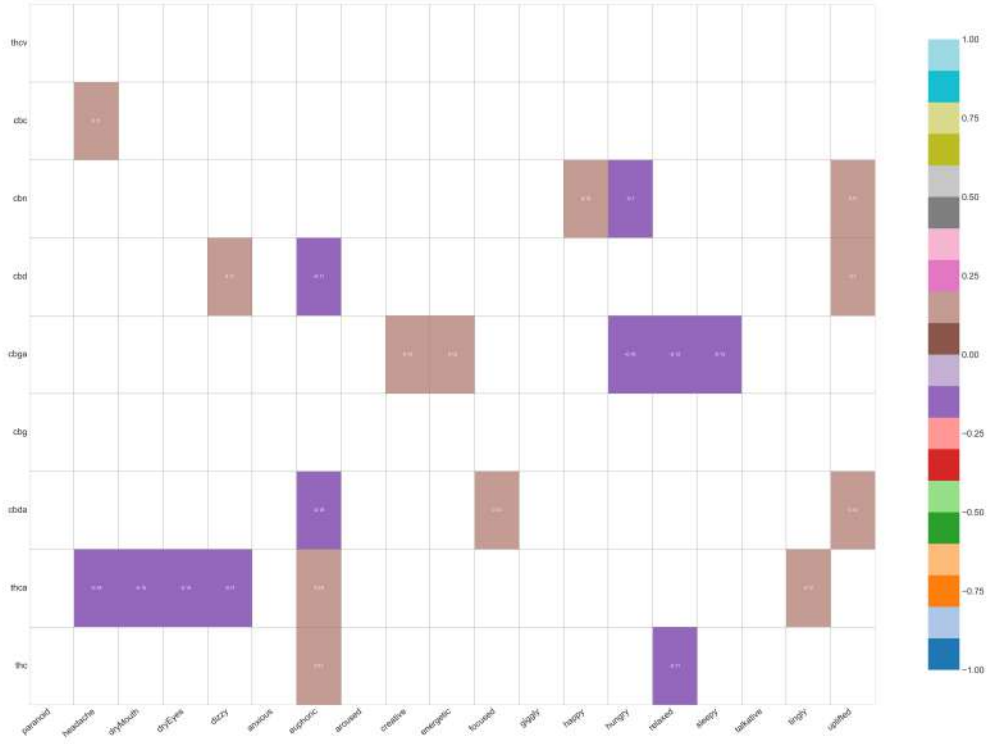


Figura 3.25: Mapa de correlaciones entre efectos (eje x) y cannabinoides (eje y).

negativos de -0.18 aunque, en general, son más cercanos a ± 0.1.

Grupos de efectos					
Grupo 1	paranoid aroused	headache giggly	dizzy happy	anxious	euphoric
Grupo 2	dryMouth tingly	dryEyes uplifted	creative	focused	talkative
Grupo 3	energetic	hungry	relaxed	sleepy	

Cuadro 3.8: Grupos de efectos identificados por `K-Means` a partir de la matriz de correlaciones con cannabinoides.

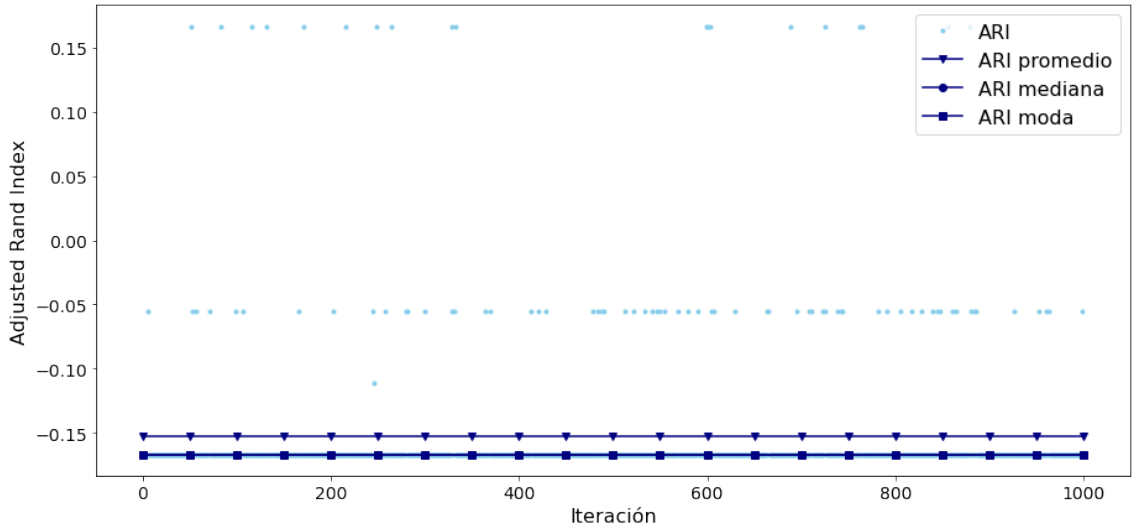


Figura 3.26: Adjusted Rand Index (ARI) para los *clusters* de efectos en base a las correlaciones con cannabinoides.

Por otra parte, `K-Means` identificó tres grupos que mezclan efectos positivos y negativos, y de euforia con de relajación, como se ve en la tabla 3.8. Una vez más se está en la situación en la que la cantidad de *features* supera a los registros y `K-Means` no tiene suficiente información para generar grupos relevantes, y los coeficientes `ARI` son muy cercanos a 0 (figura 3.26), respaldando el descarte de los *clusters* hallados.

En la figura 3.27 se ven las tres comunidades de efectos encontradas por el algoritmo. La que se observa en celeste engloba a los efectos de tipo eufóricos y de bienestar, con *anxious* fuera de lugar. Respecto de las otras dos, el conjunto en rosa mezcla efectos negativos con algunos positivos, y en el gris estarían los efectos de relajación y *tingly*. El análisis de validez de los resultados, como se puede ver en la figura 3.28, muestra que la modularidad original es en todos los casos mayor que las aleatorias, indicando que los resultados son aceptables. De todos modos, los resultados no son satisfactorios, en tanto no presentan grupos bien definidos.

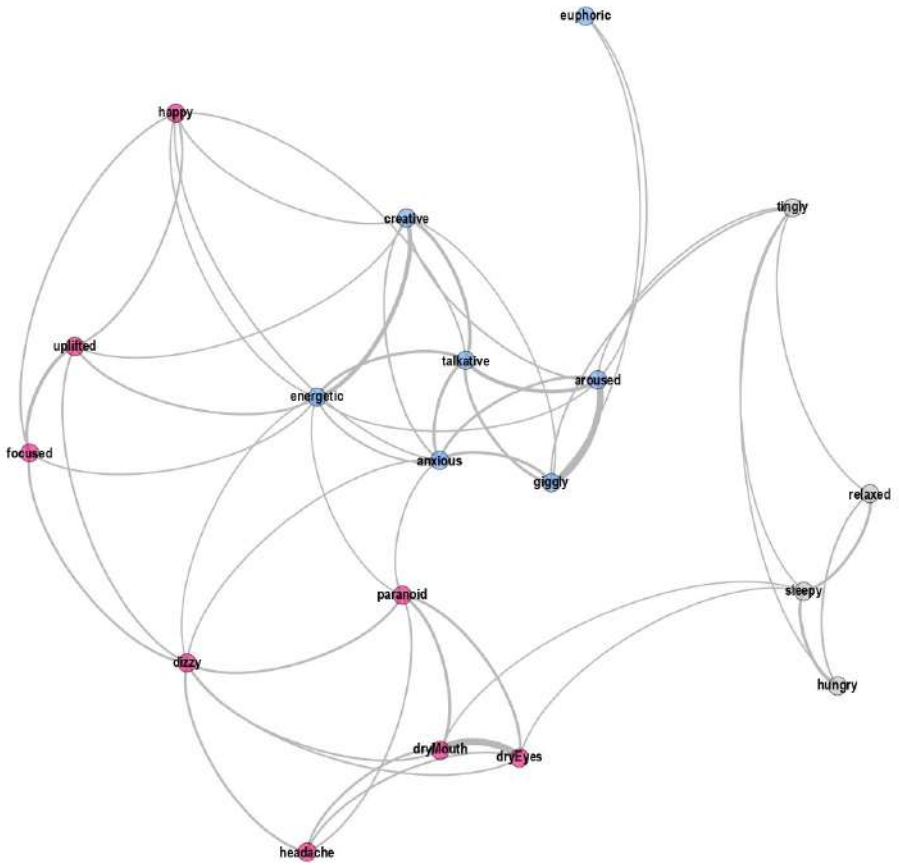


Figura 3.27: Grafo de sabores calculado en base a las correlaciones entre efectos y cannabinoides. Se identifican tres comunidades, diferenciadas por color.

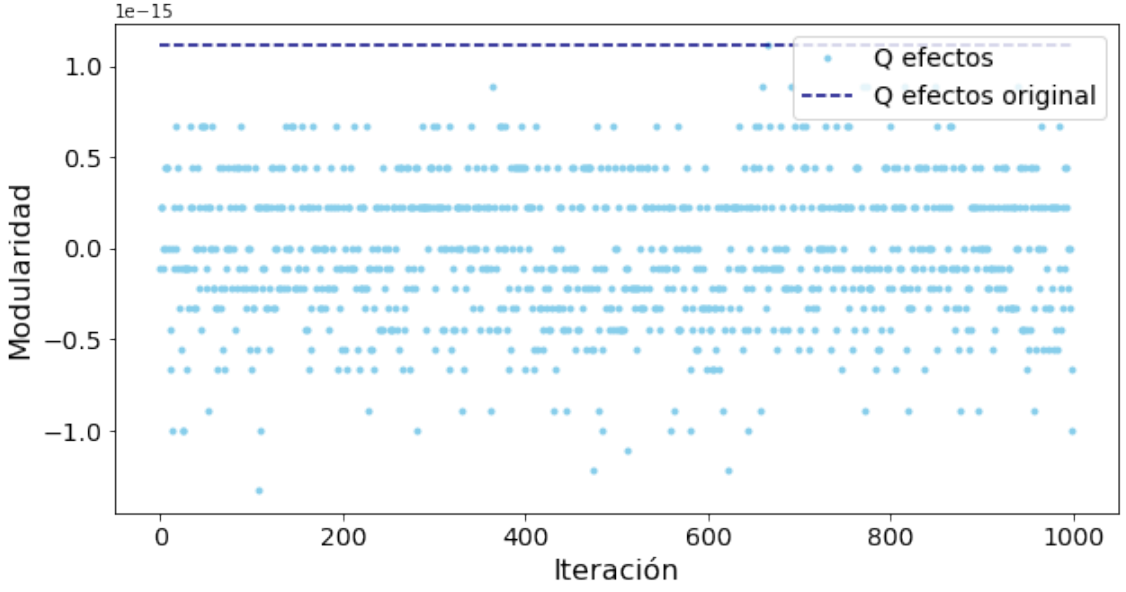


Figura 3.28: Validación de las comunidades obtenidas para los efectos.

3.3. Resultados generales

En este capítulo se presentaron los resultados de estudiar las correlaciones entre los distintos subconjuntos de datos y lo obtenido de aplicar los algoritmos `K-Means` y de Louvain, a partir de estas correlaciones, para hallar patrones en los datos en forma de *clusters* y comunidades.

Contrario a lo que se hubiera esperado en base a la intuición y a los resultados previos[7], los perfiles de correlaciones hallados fueron más bien poco informativos, con valores en general por debajo de ± 0.2 o ± 0.3, con excepción del caso Condiciones–Efectos, en el cual se evidenciaron relaciones más fuertes. No obstante, se observaron algunas correlaciones interesantes, como se vio en los resultados.

Los *clusters* y las comunidades halladas por los algoritmos en general no mostraron poder distinguir los elementos en forma efectiva. En la mayoría de los casos se obtuvieron grupos grandes con elementos mezclados. Los mejores resultados se tuvieron para los efectos, especialmente en el par Condiciones–Efectos. Es de recordar que los *clusters* y las comunidades se buscaron a partir de perfiles de correlaciones sin mucha variabilidad, por lo cual, al generar matrices de distancias con ellos, los vectores pudieron verse aún más homogéneos, lo cual dificulta que los algoritmos encuentren patrones. Además, las categorías son genéricas y aplican a varios casos. Por ejemplo, muchas de las condiciones involucran dolor crónico y, en consecuencia, los efectos buscados pueden ser de sedación, de mejora de ánimo, de concentración, los cuales aplicarían también a una condición como *anxiety*, de modo que agrupar condiciones a partir de cómo correlacionan con efectos podría volverse una tarea compleja incluso con valores más altos, únicamente porque las categorías no alcanzan para distinguir condiciones. Otro ejemplo es el caso de sabores y terpenos: los sabores no están compuestos por uno o unos pocos terpenos, sino por varios, lo cual genera que los perfiles de correlación entre dos sabores distintos puedan resultar parecidos, lo cual deviene en una agrupación ineficaz de sabores por parte de los algoritmos.

Además de la motivación conceptual, existe una razón de tipo algorítmica que tiene que ver con el tamaño de los vectores. En el caso de las correlaciones, estas se calculan a partir de dos matrices de dimensión $790 \times Z$, donde Z puede tomar valores entre 9 y 47. Si bien las correlaciones fueron corregidas con el método de Bonferroni, existe la posibilidad de que los valores hayan resultado significativos por

el elevado poder estadístico; al no observar asociaciones relevantes entre los datos, no se puede descartar que estos hayan sido producidos por ruido o por factores de confusión que no fueron considerados en el análisis, y entonces no son resultados del todo confiables. En el caso de `K-Means`, el modelo era alimentado con matrices de dimensión $\texttt{dim(subconjunto}_1\texttt{)} \times \texttt{dim(subconjunto}_2\texttt{)}$, en las que no se cumplía que la cantidad de registros (filas) fuera mucho mayor que la de *features* (columnas). En estas situaciones, algoritmos que se basan en el cálculo de distancias euclidianas sufren lo que se conoce como la "maldición de la dimensionalidad", que es la situación en la cual el espacio se vuelve tan grande que las distancias parecen todas aproximadamente iguales, imposibilitando que los algoritmos como `K-Means` puedan separar correctamente los elementos.

Si bien el volumen de datos con el que se trabajó es grande respecto de la cantidad de datos que se podrían obtener en un ensayo clínico, resultaron pocos para alimentar modelos de *machine learning*, y sobre todo la cantidad de *features* era muy grande respecto de la cantidad de registros. Una primera solución sería conseguir más datos, si fuera posible. Otras dos opciones incluyen hacer *feature engineering*, buscando *features* que sean más explicativas mediante la combinación y/ó eliminación de algunas de ellas, utilizar algoritmos de *feature importance* para identificar aquellas con mayor peso, y generación de datos sinteticos que puedan aumentar el volumen de datos disponibles para modelar.

Identificación de sinergias

El último objetivo del trabajo fue buscar sinergias entre las moléculas, principalmente entre cannabinoides y terpenos.

Del conjunto de datos preprocesados[1] se extrajeron las columnas de químicos, que pasaron a constituir la matriz de *features* o atributos (dimensión 790×37), y las de condiciones[2], los *targets*. Con las *features* y los *targets* definidos, se empezó la búsqueda de aquellas *features* con mayor peso en el modelo, las más "importantes" y, por lo tanto, buenas candidatas a formar sinergias y ser eventualmente estudiadas en un ensayo clínico.

Si bien las condiciones forman una matriz de 790×40, se estudió cada una de ellas por separado, dando lugar a 40 repeticiones del mismo proceso. Para hallar las *features* más importantes, se utilizó el selector `Recursive Feature Elimination (RFE)`[3], el cual requiere que se utilice en combinación con un algoritmo clasificador o regresor que guarde información sobre *feature importances*. En este caso se eligió el clasificador `Gradient Boosting Classifier`[4]. La información disponible para las condiciones provenía de los votos de los usuarios por cada cepa, y son valores numéricos, discretos y mayores o iguales a cero. Para llevar el problema a uno de clasificación binaria, los valores del *target* se transformaron en 0 o 1 dependiendo de si estaban por debajo o por encima de la mediana. De esta forma, los datos de entrada al proceso fueron la matriz de *features* con la información química y un vector *target* con los datos binarizados de una condición.

El algoritmo se puede dividir en dos partes. En la primera, en un proceso de *cross validation*, se alimentan `RFE` y `Gradient Boosting Classifier` con los da-

[1] Ver final del Capítulo 2.
[2] En esta parte del trabajo no se normalizaron los datos, pues el algoritmo elegido no lo requirió.
[3] Ver sección 4.1.2.
[4] Ver sección 4.1.3.

tos de entrenamiento, de lo cual se obtienen las dos *features* más importantes, y con la predicción hecha sobre los datos de validación se calcula la curva ROC. En la segunda parte, se genera un *target* aleatorio, se entrena y generan predicciones con `Gradient Boosting Classifier`, dentro del esquema de *cross validation*, y se obtiene la curva ROC. Por cada *target* se repite el proceso un número definido de iteraciones. Finalmente, con los valores ROC "real" y "aleatorio" se construye un p–valor contando la cantidad de veces que el modelo aleatorizado excedió en performance al modelo real, dividido por la cantidad de iteraciones (ver sección 4.1.2, ecuación 4.1), con el objetivo de estimar la probabilidad de haber obtenido la performance que se obtuvo, bajo la hipótesis de que no hay relación entre las *features* y el *target*.

Con la información obtenida se generaron visualizaciones y se identificaron las posibles sinergias.

4.1. Métodos

4.1.1. *Cross Validation* (CV)

A la hora de entrenar un modelo de tipo supervisado[5], una buena práctica para evitar que el modelo haga *overfitting*, es decir, que prediga muy bien sobre los datos de entrenamiento y mal sobre datos nunca antes vistos, es separarlos en un conjunto de entrenamiento (*train set*) y uno de prueba (*test set*). Con este esquema, dado que habitualmente se entrena el modelo, se hacen predicciones sobre el conjunto de *test*, y luego se retocan los parámetros y reentrena el algoritmo, es posible que haya filtración de información que está en el conjunto de *test* hacia el de *train*. Una forma de superar esta situación es separar un tercer conjunto, denominado de validación, de manera que el proceso de entrenamiento y testeo se haga con los datos de entrenamiento y de validación, y que una vez que se está satisfecho con el resultado, se hace la predicción sobre el conjunto de *test*. La desventaja de separar un conjunto de validación es que se achica la cantidad de datos disponibles para entrenar, y eso afecta la capacidad de aprendizaje del modelo, derivando en predicciones menos acertadas.

[5]Son aquellos modelos a los que se provee un vector *target*, con etiquetas correctas, para que el modelo prediga situaciones habiendo aprendido sobre la base de las *features* y los *targets*. En aprendizaje no supervisado, no se cuenta con un vector *target*, "no se conocen las respuestas", y los algoritmos, en este caso, buscan patrones en los datos.

El método de validación cruzada[6] (CV) surge para evitar este problema. Manteniendo la separación de *train* y *test sets*[7], se provee el conjunto de entrenamiento al algoritmo de CV, el cual lo divide en k partes (k definido por el usuario) y procede a entrenar el modelo con *k-1* partes y testear con la restante. Las partes se van rotando de manera que todas sean una vez el conjunto de *test*. El proceso se repite k veces, hasta que todas las partes hayan rotado, y el algoritmo de CV entrega una métrica que será el promedio de las métricas de cada iteración. Este método puede ser caro computacionalmente (dependiendo de la cantidad de datos), pero no se "pierden" datos por separar en tres conjuntos, y es especialmente útil cuando se dispone de pocos datos. Existen varios algoritmos disponibles en la librería `scikit-learn` de python; en este trabajo se utilizó `StratifiedKFold`, que tiene en cuenta el porcentaje de cada etiqueta del *target* y arma las partes manteniendo esas proporciones.

4.1.2. *Recursive Feature Elimination* (RFE)

`Recursive Feature Elimination (RFE)`[8] es uno de varios métodos que existen para hallar las *features* más determinantes en un conjunto de datos. Este selector elabora un rango de *features* via eliminación recursiva. A diferencia de otros, `RFE` requiere que se le indique un estimador de aprendizaje supervisado (regresor o clasificador) cuyo método `fit` guarde información sobre *feature importances*. `RFE` trabaja eliminando progresivamente *features* en función de los puntajes que asigna el regresor/clasificador, hasta llegar al número deseado por el usuario.

El proceso de búsqueda de las *features* se realizó en dos etapas iterativas. Dentro de un ciclo `for` de 50 iteraciones[9] se aplicó un proceso de *cross-validation*[10] de 5 *folds* con dentro `RFE` y `Gradient Boosting Classifier`. Por cada iteración del `CV`, el subproceso de `RFE` corre 35 veces, porque se pide que elija 2 *features* de las 37 que son. Además, con los resultados del clasificador, se obtiene la curva ROC (y, con ello, el área bajo la curva, AUC). Por cada una de las 50 iteraciones, se

[6] `https://scikit-learn.org/stable/modules/cross_validation.html`

[7] Idealmente. Si la cantidad de datos es pequeña, se puede hacer únicamente CV, sin separar un conjunto de test previamente.

[8] `https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html`

[9] Valor arbitrario elegido para balancear un buen número de iteraciones y un tiempo de ejecución razonable.

[10] `StratifiedKFold`; `https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html`

corre lo descripto sobre los datos reales, y se genera un conjunto de datos aleatorios sobre los que también se hace `CV` y una clasificación de la cual se obtiene la curva ROC. No se hace `RFE` sobre los datos aleatorios porque lo que se busca es obtener la *performance* del modelo nulo para asegurarse de que los resultados con datos reales no son aleatorios, a través de la comparación de los p–valores de cada caso. En la ecuación 4.1, el p–valor es la suma del valor de AUC más grande entre el resultado sobre los datos reales y sobre los aleatorios (elemento a elemento del vector AUC con los 50 valores de AUC y su correspondiente aleatorio), dividido la cantidad de iteraciones.

$$pvalue = \frac{\sum_i max\{AUC^i,\ AUC^i_{random}\}}{Cantidad\ de\ iteraciones} \tag{4.1}$$

Por último, como se comentó en la introducción del capítulo, dado que se considera que las condiciones son independientes, se corrieron 40 procesos paralelos, en los cuales cada condición era el `target`. El resultado final fueron 40 p–valores y 40 archivos con las *features* elegidas en cada una de las 50 iteraciones.

4.1.3. Clasificador: *Gradient Boosting Classifier*

Los árboles de decisión[11] son un tipo de modelo de aprendizaje supervisado, no paramétrico, que se utiliza en problemas de clasificación y regresión. A partir de reglas de decisión que arma en base a las *features*, el algoritmo hace predicciones. Este modelo tiene a favor que es fácilmente interpretable, se puede visualizar, y es relativamente rápido, entre otras propiedades, pero sus mayores desventajas son que es propenso a hacer *overfitting* y que es sensible a pequeños cambios en los datos. Una primera solución a este problema es usar ensambles de árboles de decisión, con un modelo como `RandomForestClassifier/Regressor`. Para mejorar ulteriormente los resultados, surgen los algoritmos de *boosting*, los cuales consisten en combinar una serie de modelos "débiles" (*weak learners* en inglés), es decir, modelos cuyas predicciones no sean aceptablemente acertadas, para obtener un único predictor confiable. Un paso más allá se encuentran los algoritmos de *gradient boosting*, en los cuales se busca optimizar una función de costo (diferenciable), habitualmente *mean square error* (MSE) si es un regresor y *log loss* si es un clasificador. En estos modelos, en vez de entrenar cada modelo sobre los datos en cada iteración, lo hace sobre los errores residuales cometidos en la iteración anterior; es decir, trata de minimizar el

[11] `https://scikit-learn.org/stable/modules/tree.html#tree`

residuo. En este trabajo se eligió el algoritmo `Gradient Boosting Classifier`[12].

Función *log loss*

Mientras que los algoritmos de regresión suelen optimizar en base a medidas de distancia como MSE, los clasificadores se basan en otras métricas, como *log loss*, la utilizada por `Gradient Boosting Classifier`. Al momento de asignar un punto a una clase, el clasificador calcula la probabilidad de pertenencia a esa clase y luego lo etiqueta como 0 o 1 (en el caso binario). La función *log loss* da cuenta de qué tan cerca está la probabilidad del valor real, que sería 0 o 1. Cuanto más grande sea la diferencia entre probabilidad y valor real, mayor será el valor de *log loss*. De esta forma, lo que se buscará será minimizar la *log loss*. Matemáticamente:

$$logloss_i = -\left[y_i \ \ln p_i + (1 - y_i) \ln(1 - p_i)\right],$$

donde y_i es el valor real del punto i y p_i es la probabilidad. Al finalizar la ejecución, el algoritmo reporta el valor medio de todos los *log loss*.

4.1.4. Curva ROC

La curva ROC[13] es una forma gráfica de ver la *performance* del modelo de clasificación cuando el umbral que utiliza para asignar las clases (0 y 1) es variado. El gráfico que genera tiene en el eje x la fracción de falsos positivos y en el eje y la fracción de verdaderos positivos, calculados en distintos umbrales. Dicho de otra manera, lo que se grafica es la sensibilidad[14] *vs.* (1 - especificidad[15]).

$$sensibilidad = \frac{verdaderos\ positivos}{verdaderos\ positivos + falsos\ negativos} \tag{4.2}$$

$$especificidad = \frac{verdaderos\ negativos}{verdaderos\ negativos + falsos\ positivos} \tag{4.3}$$

Con estos resultados, se calcula el área bajo la curva (AUC). En el peor escenario, el valor mínimo de AUC es 0.5, y se ve como una línea diagonal que atraviesa el

[12] `https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html`

[13] *Reicever operating characteristic* en inglés, `https://scikit-learn.org/stable/modules/model_evaluation.html#roc-metrics`

[14] Es la capacidad del estimador de identificar los elementos realmente positivos

[15] Es la capacidad del estimador de identificar los elementos realmente negativos

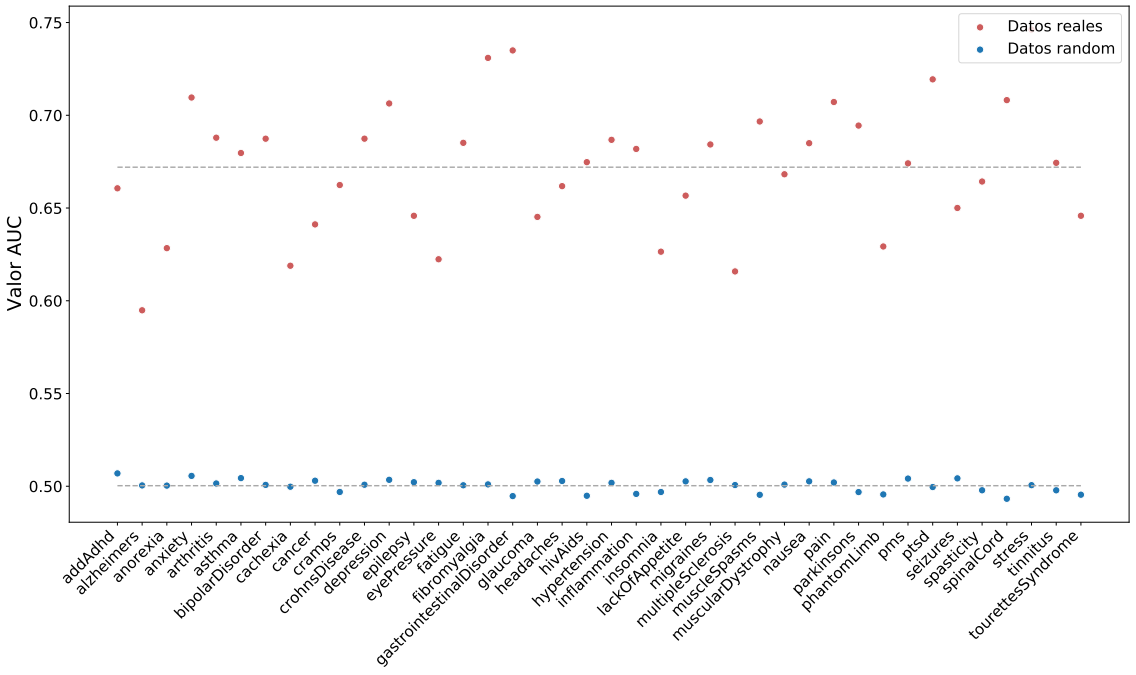


Figura 4.1: AUC obtenidas del modelo por cada condición, para el caso con los datos reales (rojo) y aleatorios (azul). En gris se muestran las medias de las AUC aleatorias y reales.

cuadro desde la esquina inferior izquierda hacia la superior derecha. En el caso perfecto, la AUC es igual a 1 y se ve como una línea vertical que va del (0, 0) al (0, 1) y una horizonal que va del (0, 1) al (1, 1). En los casos reales, dependerá del experimento cuándo el valor de la AUC es aceptable o no. Las AUC obtenidas para cada condición y su respectivo "caso aleatorio" se ven en la figura 4.1. Además de los valores de AUC, se graficaron, en gris punteado, los valores medios. Se ve que el caso aleatorio tiene una media de 0.5, como esperado; el AUC medio de los datos reales ronda 0.67, con aproximadamente la mitad de los valores por encima de la media.

4.2. Resultados y discusión

Partiendo de los 40 archivos con las *features* seleccionadas en cada iteración, por cada uno de ellos se contó la cantidad de pares distintos (no hay diferencia entre los pares A–B y B–A) y los resultados se guardaron en una matriz cuadrada de dimensión 37×37 (la cantidad de *features*).

A continuación se definieron grupos de condiciones, que no necesariamente se

corresponden con condiciones similares desde el punto de vista clínico, sino con el tipo de malestar, dolencia o sintomatología que los usuarios pudieran estar buscando paliar al consumir. Esta división se armó en base a la búsqueda de los efectos que cada condición produce y criterio personal, y no tiene validez científica desde el punto de vista médico. Se armaron cinco grupos, denominados informalmente de condiciones mentales, de dolor crónico, de condiciones que producen deterioro físico, neurológicas y gastrointestinales. En cada sección se detallan los elementos pertenecientes a cada grupo. Se excluyeron de los análisis las condiciones *eye pressure*, *glaucoma*, *ashma*, *hypertension*, *spasticity* y *spinal cord* por no poder ser categorizadas en los grupos mencionados.

Por cada grupo se sumaron los pares de cada condición, se visualizaron en forma de gráfico de barras y se analizan los resultados. Al final de la sección se muestra un gráfico resumen con los resultados de todas las condiciones juntas y se recapitulan los resultados en forma general.

4.2.1. Grupo de condiciones mentales

En este grupo se incluyeron condiciones que tienen que ver con enfermedades y síndromes psicológicos y psiquiátricos (no neurológicos): *anorexia*, *anxiety*, *bipolar disorder*, *depression*, *fatigue*, *insomnia*, *ptsd*[16], *stress*. En la figura 4.2 se ven los resultados del grupo, obtenidos sumando los resultados de cada condición. Se destacan sobre los demás los pares *iso pulegol/cbc* e *isopulegol/thc*. El terpeno *isopulegol* fue seleccionado un total de 158 veces con el cannabinoide *cbc*, 60 con *thc*, 17 con *cbn* y *cbd*, y 14 con *thc*; y por otro lado, fue seleccionado 26 veces el par *pcymene/thc*.

Dado que el grupo contiene condiciones cuya sintomatología puede resultar opuesta, o que los efectos buscados al consumir lo son (como *depression* y *stress* o *insomnia*), se miraron en detalle los pares de algunas de ellas. De las 50 elecciones hechas para *anxiety*, 40 corresponden a *isopulegol/cbc*; similarmente, en *bipolar disorder*, esa combinación se obtuvo 33 veces. En el caso de *depression*, está repartido entre 19 veces *isopulegol/cbc*, 18 *isopulegol/thc*. *Insomnia* tiene sus pares repartidos, con 9 elecciones de *isopulegol/cbc*, 7 con *thc*, 6 con *terpinolene*; para *ptsd* es similar, 10 veces *isopulegol-cbc*, 17 *cbd*, y para *stress*, 44 pares corresponden a *isopulegol/cbc* y no salió en ninguna oportunidad *thc*.

[16] *Síndrome post traumático*

Si bien estas condiciones son complejas y no todas las personas las atraviesan de igual forma, es interesante que el modelo haya seleccionado tantas veces la combinación *isopulegol/cbc*, dado que el *isopulegol* está siendo estudiado por su potencial uso como antidepresivo[17][13] y por sus propiedades antiinflamatorias[18][14], y el *cbc* también tendría propiedades antidepresivas (en principio en combinación con *thc* y *cbd*[15], queda por determinar su eficacia aislado), además de antiinflamatorio, características que intuitivamente se puede suponer que buscan los usuarios que sufren varias de las condiciones del grupo al consumir. Respecto del par *isopulegol/thc*, el *thc*, además de ser el cannabinoide que provoca efectos psicoactivos, tiene también propiedades antiinflamatorias, relajantes (a veces depende de las combinaciones de cannabinoides y terpenos presentes en la cepa), de alteración de memoria y de estimulación del apetito, deseables en mayor o menor medida según la condición. El *cbn* es un subproducto del *thc* y es levemente psicoactivo; combinado con el *thc*, potencia el efecto de "estar pegado al sillón", también combinado con los terpenos más estables (los sesquiterpenos, los que continúan presentes en la hierba vieja una vez que otros terpenos se evaporan y descomponen). Si bien es un cannabinoide en estudio, se sabe que tiene propiedades antiinflamatorias, es un relajante muscular, un estimulante de apetito y tiene algunos beneficios neuroprotectores y de alivio de artritis (ver grupo de dolor crónico). El *cbd*[16] es un cannabinoide que no produce efectos psicoactivos, y que tiene beneficios terapéuticos tales como antiinflamatorio, analgésico, neuroprotector, antiartrítico, entre otros. Finalmente, del par *thc/pcymene*, el terpeno *pcymene*[17] tiene propiedades antiinflamatorias, antioxidantes, analgésicas, vasorelajantes y neuroprotectoras, entre otras. Es decir, psicoactivas o no, las combinaciones de químicos seleccionadas por el algoritmo resultaron ser aquellas que potenciaban los efectos sedativos y antiinflamatorios, resultados intuitivamente esperados para estas condiciones.

4.2.2. Grupo de dolor crónico

En segundo lugar se identificó un grupo de condiciones que involucran dolores crónicos: *arthritis*, *cramps*, *headaches*, *inflammation*, *migraines*, *muscle spasms*, *pain*, *pms*[19], *fibromyalgia* y *Chrons disease*. Una vez más, los resultados (figura 4.3) muestran un claro predominio del par *isopulegol/cbc*, con 201 ocurrencias, 30 para

[17]Estudio realizado en animales por el momento.
[18]Ídem anterior.
[19]Síndrome pre menstrual

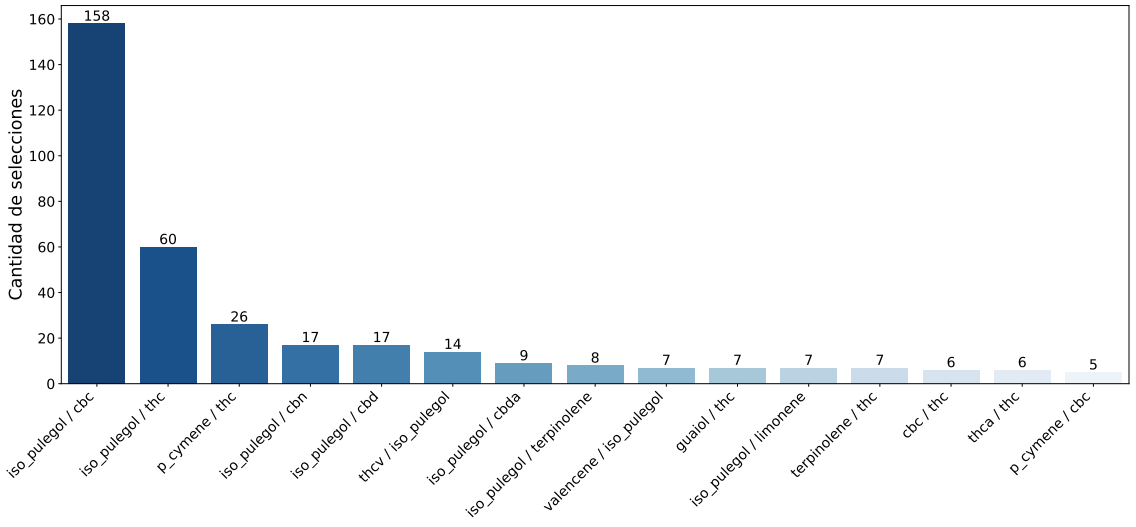


Figura 4.2: Pares del grupo de condiciones mentales con mayor cantidad de selecciones. Se muestran únicamente aquellos pares elegidos 5 o más veces. En total son 43 combinaciones.

isopulegol/cbn y tan solo 15 para *isopulegol/thc*. Un par nuevo respecto del grupo anterior es *isopulegol/guaiol*, seleccionado 74 veces. El *guaiol* es un terpeno con propiedades antiinflamatorias[18] y podría reducir el tamaño de tumores y potenciar los efectos de la quimioterapia[19] (ensayado en ratones por el momento).

Las elecciones del selector para estas condiciones muestran que el objetivo principal es conseguir efectos de tipo sedativo, a diferencia del grupo anterior, en los que había buena presencia de pares con *thc*, el cual, si bien tenía propiedades relajantes, su mayor característica son los efectos psicoactivos que genera. En estos resultados, la presencia de *cbn* es mayor que en los anteriores, y podría apoyar el incentivo a la investigación clínica que está recibiendo este cannabinoide actualmente. Es importante resaltar que, tanto en los pares de condiciones mentales como en los de dolor crónico, los pares más elegidos combinan un cannabinoide y un terpeno (en particular, casi siempre el mismo terpeno), fomentando así la idea de que los efectos producidos por el cannabis no provienen únicamente de los cannabinoides[6, 7], sino que existe una sinergia entre cannabinoides y terpenos, y que es esa sinergia la que genera los efectos deseados por el usuario, y que tienen potencial para ser estudiados clínicamente.

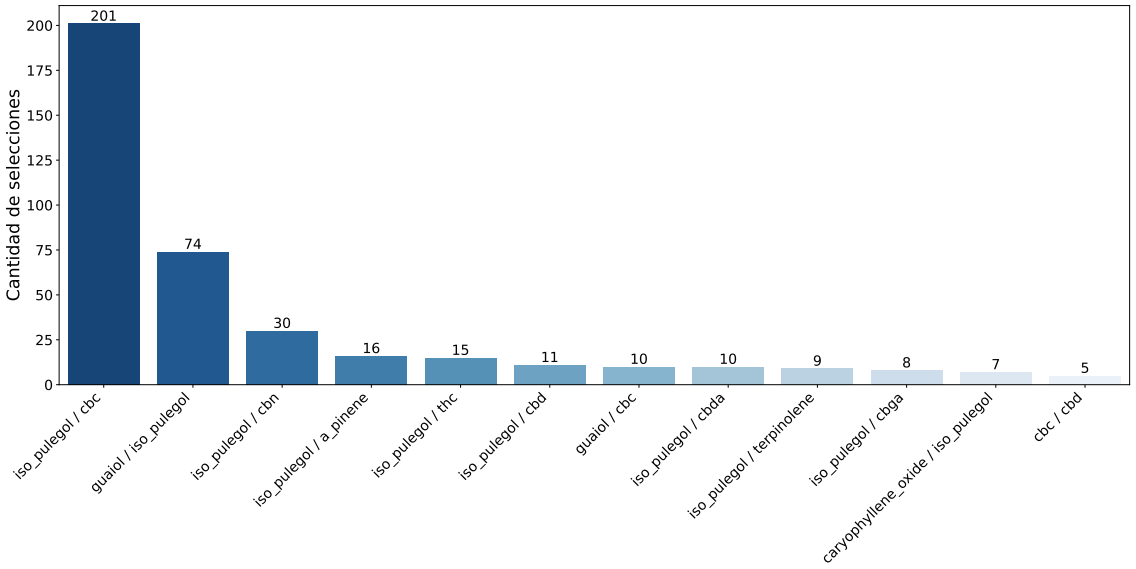


Figura 4.3: Pares del grupo de dolor crónico con mayor cantidad de selecciones. Se muestran únicamente aquellos pares elegidos 5 o más veces. En total son 42 combinaciones.

4.2.3. Grupo de condiciones que producen deterioro físico

Este grupo se pensó para enfermedades que produjeran algún tipo de deterioro físico que el usuario buscara relajar o mitigar, como *multiple sclerosis*, *muscular dystrophy*, *Parkinson*, *cancer*, *cachexia*, *HIV-AIDS*. En los resultados (figura 4.4) se observan algunas diferencias respecto de los casos anteriores: en primer lugar, el *isopulegol* no fue seleccionado tantas veces como antes y, en segundo, las combinaciones fueron mucho más variadas, si bien hay dos que tienen más votos que las demás. Los dos pares más elegidos son *thc/pcymene* (59 veces) y *cbc/pcymene* (42 veces). El *pcymene*, como se vio anteriormente, es un terpeno con propiedades antiinflamatorias, analgésicas y vasorelajantes, entre otras, y tanto el *cbc* como el *thc* tenían también características antiinflamatorias. Otros pares con menor cantidad de selecciones fueron *thc/eucalyptol* (15), *cbd/pulegone* (12) y *cbda/pulegone* (12). El *eucalyptol* es efectivo para tratar el dolor[20], y el *pulegone* es otro terpeno con aroma mentolado, con propiedades de alivio de ansiedad[21] y sedativas[22]. El *cbda* es un cannabinoide precursor del *cbd* que está recibiendo más atención en el último tiempo. Entre las características descubiertas recientemente, se destacan su potencial uso como agente antiinflamatorio[23] y como supresor de convulsiones (`https://patents.google.com/patent/WO2017025712A1/en`).

Es decir, para las condiciones en este grupo, el selector eligió combinaciones que

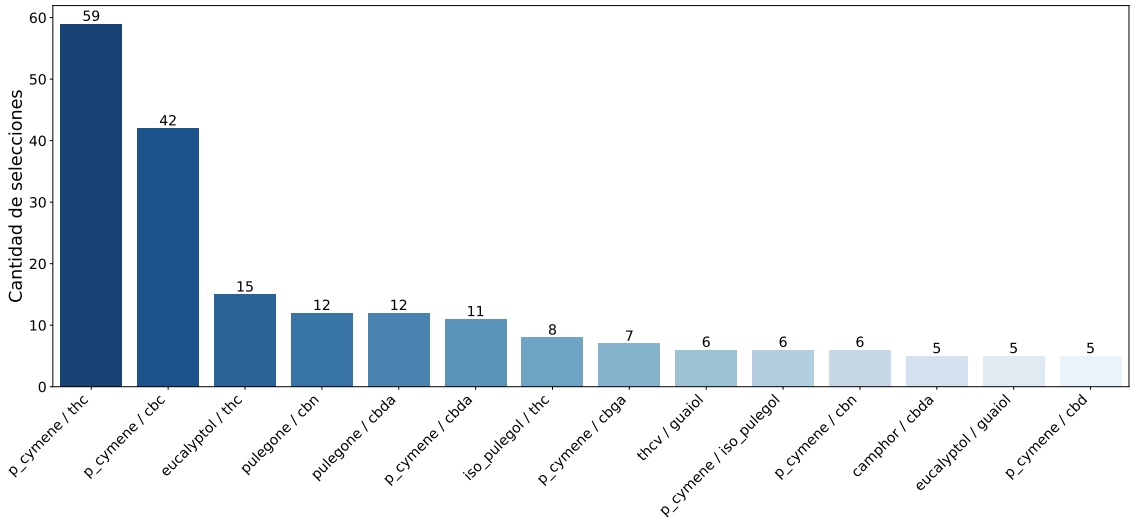


Figura 4.4: Pares del grupo de condiciones que producen deterioro físico con mayor cantidad de selecciones. Se muestran únicamente aquellos pares elegidos 5 o más veces. En total son 72 combinaciones.

generaran efectos antiinflamatorios y de relajación y, además, psicoactivos en muchos casos (los pares con *thc* y *cbn*). Si bien el efecto final no difiere en gran medida de los de los grupos anteriores, pues en los tres las combinaciones apuntan a relajar a los usuarios y las propiedades en general son compartidas, es relevante que los pares sean distintos, y que los terpenos seleccionados sean más específicos a las causas de los malestares, por más que el efecto obtenido pueda ser similar en todos los casos, pues podría indicar que hay sinergias más efectivas para ciertas condiciones que para otras.

4.2.4. Grupo de condiciones neurológicas

Este grupo está conformado por las condiciones *add adhd*, *epilepsy*, *seizures*, *Alzheimer*, *Parkinson*, *Tourette syndrome*, *tinnitus*, *phantom limb* y *headaches*. Los elementos son variados en su naturaleza, por lo que es esperable que los resultados lo sean, como se ve en la figura 4.5. En ellas se ven varias combinaciones distintas, con más o menos ocurrencias, pero sin ninguna que predomine por sobre las demás. Vuelven a aparecer los pares *isopulegol/cbc* (37 veces), *isopulegol/thc* (15), *pcymene/thc* (18), *pcymene/cbc* (16), *isopulegol/cbd* (8), *eucalyptol/cbc* (9). La novedad son las combinaciones de cannabinoides *cbc/thc*, el más elegido por el selector, 39 veces, de las cuales 28 fueron en *Tourette syndrome* (10 en *Alzheimer*), y *cbc/cbd*, 17 veces, de las cuales 14 fueron en *seizures*. La sinergia *thc/cbd* está documentada[6], pero en la

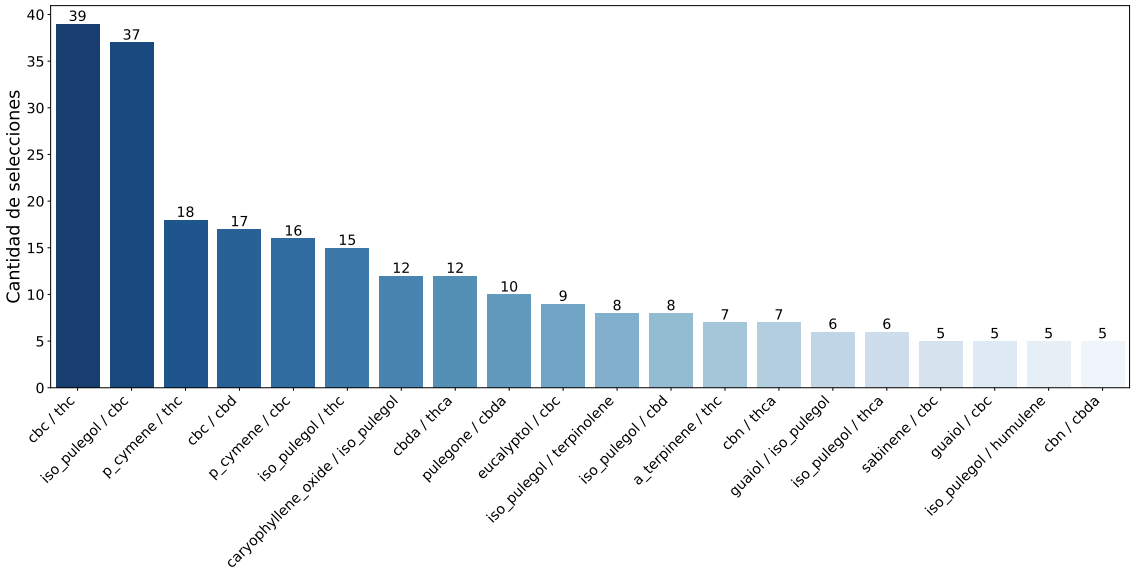


Figura 4.5: Pares del grupo de condiciones neurológicas con mayor cantidad de selecciones. Se muestran únicamente aquellos pares elegidos 5 o más veces. En total son 82 combinaciones.

actualidad se está estudiando, también, la combinación *cbc/thc/cbd*[24], como poderoso antidepresivo[15]. Existe la posibilidad que, de haberle pedido combinaciones de tres químicos, el selector eligiera *cbc/thc/cbd*. Repasando estos tres cannabinoides, tanto *cbc* como *thc* son antiinflamatorios, pero un estudio (en animales, por ahora) demostró que, combinados, el efecto era mayor[15]; además, estudios indicarían que el *cbc* es neuroprotector[25] (y tendría sentido que fuera elegido acá). Por otro lado, como se mencionó previamente, el *cbd* está siendo estudiado como supresor de convulsiones.

Entre las combinaciones que ya habían aparecido y las nuevas, se observa la tendencia del selector a elegir pares que sean antiinflamatorios y relajantes, pero también que tengan, potencialmente, efectos neurológicos. El efecto de base que podrían provocar estas combinaciones es de calma de síntomas, lo que podría darle una mejor calidad de vida al usuario.

4.2.5. Grupo de condiciones gastrointestinales

Este grupo, denominado "condiciones gastrointestinales" a falta de un nombre mejor, está integrado por *gastrointestinal disorder*, *lack of appetite* y *nausea*. Este grupo muesta (figura 4.6) nuevamente un predominio del par *isopulegol/cbc*, con 88 ocurrencias, y muy poca presencia de combinaciones con *thc*, lo cual tiene sentido,

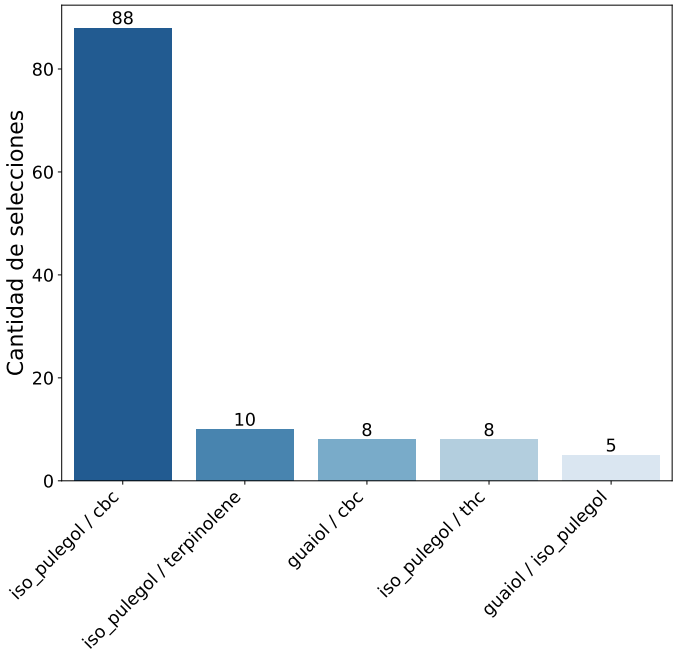


Figura 4.6: Pares del grupo de condiciones gastrointestinales con mayor cantidad de selecciones. Se muestran únicamente aquellos pares elegidos 5 o más veces. En total son 24 combinaciones.

considerando que uno de los efectos adversos de este cannabinoide es el mareo[20]. Con 10 votos, aparece la combinación *isopulegol/terpinolene*; el *terpinolene* es un terpeno con sabores de tipo cítrico, florales y boscosos, que suele estar presente en cepas que producen un efecto de mejora de ánimo[21]. En este conjunto pequeño de condiciones que se asocian a lo gastrointestinal pero que no necesariamente están relacionadas en sus causas, el algoritmo le dio mayor importancia a combinaciones que producen "alivio corporal", no euforia.

4.2.6. Resultados generales

A modo de resumen y recapitulación de lo comentado en las secciones anteriores, en la figura 4.7 se muestra un mapa de calor con la síntesis de los resultados de las 40 condiciones, en el que se ve que hay un par bien marcado en el centro y otros pocos menos intensos. Más en detalle, en el gráfico de barras (figura 4.8) se observa que el par más seleccionado fue *isopulegol/cbc*, en un total de 603 veces, seguido de *pcymene/thc*, *isopulegol/thc* y *pcymene/cbc* con apenas un poco más de 100 votos, y *guaiol/isopulegol* con 88. Las componentes de estos pares tienen características

[20] `https://www.leafly.com/learn/cannabis-glossary/thc`
[21] `https://www.leafly.com/learn/cannabis-glossary/terpinolene`

variadas, pero coinciden en sus propiedades relajantes, sedativas y antidepresivas, lo cual podría explicar porqué fueron seleccionadas.

A pesar de que los grupos de condiciones fueron, en cierta medida, arbitrarios, en líneas generales se puede decir que los resultados fueron coherentes. En el grupo de condiciones mentales hubo predominancia de pares cuyos efectos podían ser relajación, sedación, y también psicoactividad; es especialmente notorio, por ejemplo, el caso de *stress*, en el cual 44 de las 50 elecciones fueron *isopulegol/cbc*, y de *depression*, que fue repartido entre *isopulegol/cbc* e *isopulegol/thc*. Se sabe que la sintomatología de la depresión no es igual para todos, y eso podría verse reflejado en esas elecciones. En el grupo de dolor crónico, el par *isopulegol/cbc* fue de nuevo el más seleccionado, pero fue seguido de otros pares cuyos cannabinoides no eran psicoactivos o lo eran en menor medida (*cbn*), mostrando claramente que el objetivo era conseguir efectos de tipo antiinflamatorio y sedativos, claves a la hora de mitigar el dolor. En el grupo de condiciones que producen deterioro físico, el *isopulegol* ya no fue el terpeno favorito; en este caso, el protagonista fue el *pcymene*. Las selecciones para estas condiciones apuntaron sobre todo a efectos antiinflamatorios, analgésicos, vasorelajantes, de mitigación del dolor y psicoactivos. En muchos casos las condiciones de este grupo tienen que ver con enfermedades terminales, lo cual podría fomentar la búsqueda no solo del no–dolor, sino de la psicoactividad. Otro grupo heterogéneo fue el de condiciones neurológicas, y eso se reflejó en los resultados. En este caso, no hubo pares predominantes, y aparecen con cierto peso combinaciones de dos cannabinoides. Los pares seleccionados generarían efectos antiinflamatorios y de relajación, pero también habría potenciales efectos neurológicos positivos, actualmente en estudio. Finalmente, en el grupo de condiciones gastrointestinales, las elecciones apuntaron principalmente a efectos de tipo sedativo y de mejora anímica, evitando los cannabinoides psicoactivos, que suelen tener asociado el mareo como efecto negativo.

En los últimos años, además del estudio del efecto *entourage* entre cannabinoides, comenzó a tomar fuerza la idea de que las sinergias entre cannabinoides y terpenos podían ser responsables de los efectos del cannabis. En la figura 4.8 se ve claramente cómo los pares más veces seleccionados por el algoritmo consistieron en la combinación terpeno–cannabinoide, reforzando ulteriormente la idea de que las sinergias entre terpenos y cannabinoides podrían ser las principales responsables de los efectos buscados por los consumidores, a la vez que propone combinaciones posibles a ser estudiadas en un entorno clínico.

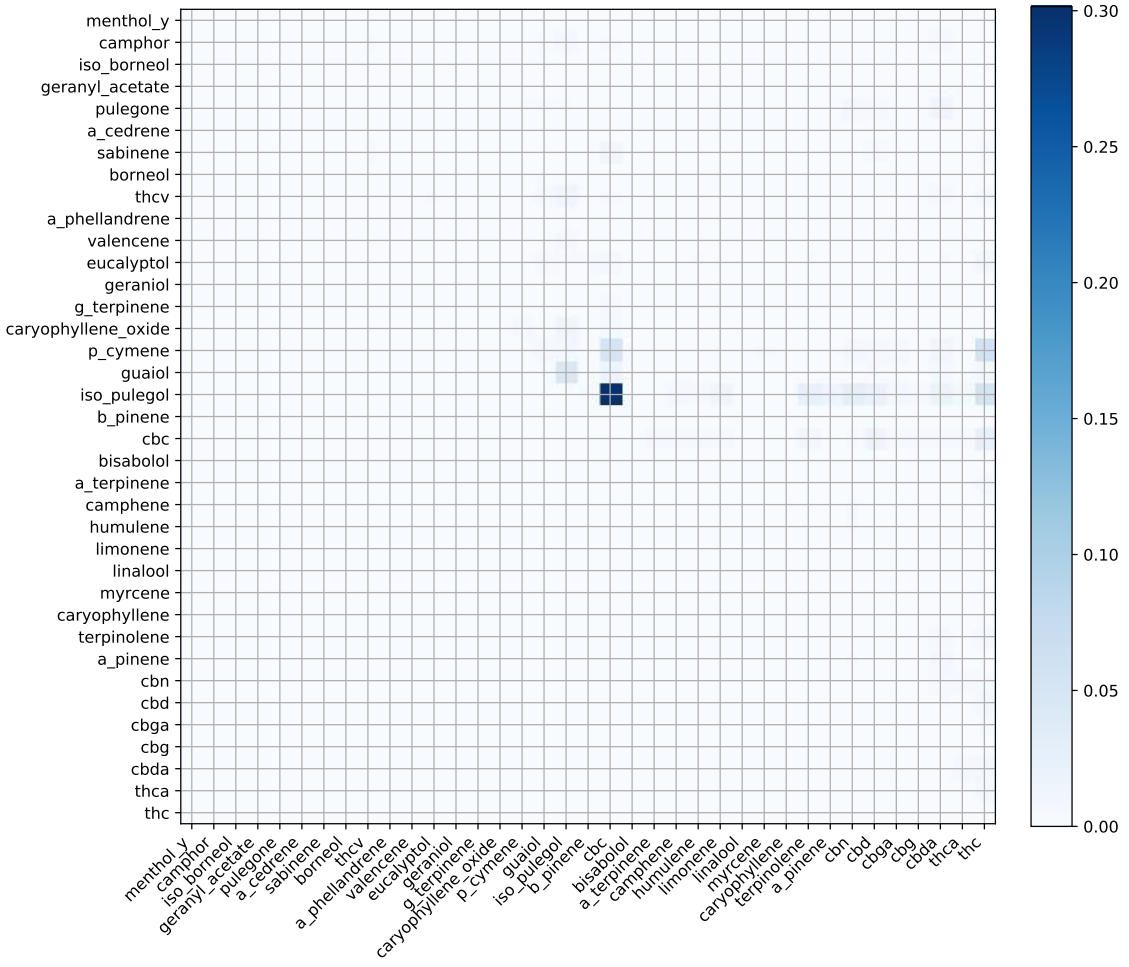


Figura 4.7: Mapa de calor con todas las selecciones hechas por `RFE`.

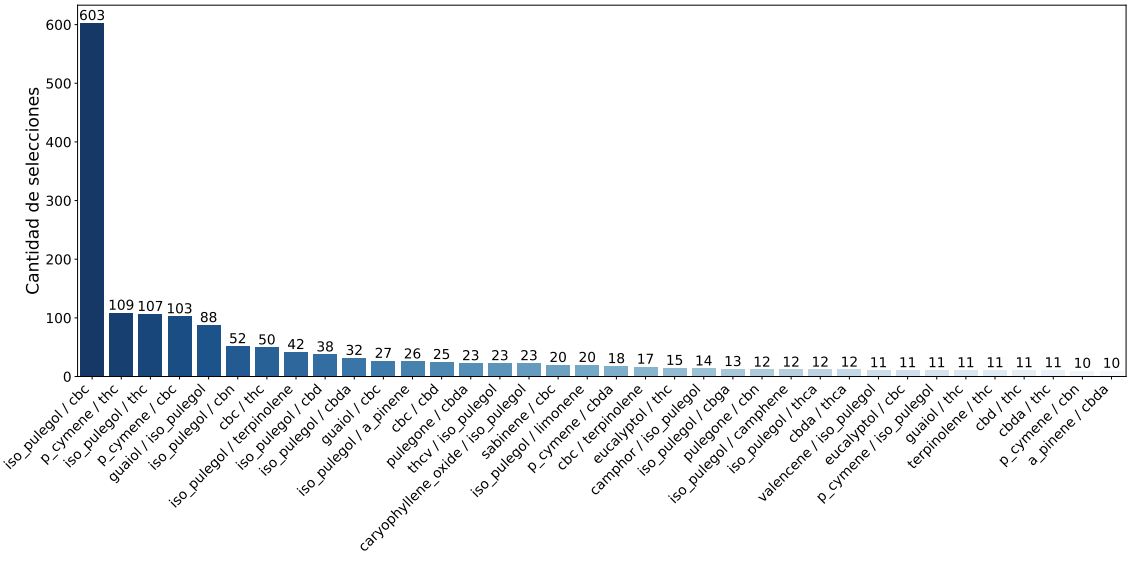


Figura 4.8: Resumen de los pares con mayor cantidad de selecciones. Se muestran únicamente aquellos pares elegidos 10 o más veces. En total son 172 combinaciones.

Discusión

Partiendo de una tabla con información sobre la cantidad de veces que las cepas fueron seleccionadas por su utilidad para mitigar condiciones y por los efectos y sabores percibidos, y otra tabla con el perfil químico de cannabinoides y terpenos de las cepas, se inició el trabajo con un análisis exploratorio y un proceso de preparación de los datos para llevarlos a la forma necesaria para los estudios siguientes.

Con ellos se inició la primera parte del trabajo, en el cual se hizo un análisis de las relaciones entre los datos mediante el cálculo de las correlaciones de Spearman y de búsqueda de patrones con `K-Means` y las comunidades de Louvain. Como se comentó al final de ese capítulo, los resultados obtenidos no fueron satisfactorios. Las correlaciones halladas fueron de poca magnitud y los *clusters* y comunidades no lograron agrupar elementos en conjuntos relevantes. Sabiendo que existían relaciones entre todos los pares estudiados[7], se esperaba hallar correlaciones más fuertes. No obstante, las relaciones que sí se pudieron ver, fueron en línea con lo conocido; en ningún caso de encontraron correlaciones que contradijeran resultados previos. Es posible que el hecho de que las correlaciones fueran débiles y que los patrones no se hayan podido observar se debiera a la cantidad de datos disponibles: si bien son muchos en comparación con un estudio clínico, los modelos suelen requerir un volumen bastante mayor. Como métodos de mitigación en caso de no poder acceder a más datos y/ó en conjunto, se propuso utilizar técnicas de *feature engineering* para descartar las variables con menor peso y así reducir la dimensionalidad de espacio.

A pesar de los problemas comentados, el análisis de correlaciones fue útil al momento de interpretar los resultados en la segunda etapa del trabajo. En ella se buscaron las sinergias entre compuestos, y los resultados no solo arrojaron combinaciones con moléculas actualmente en estudio, sino que también respaldaron la hipótesis de que las sinergias entre cannabinoides y terpenos eran buenos candidatos

a ser estudiados, pues de las 15 primeras combinaciones seleccionadas por el algoritmo, hay tan solo una que combina dos cannabinoides. Con el objetivo de visualizar más fácilmente las combinaciones, se armaron grupos de condiciones que podrían compartir de alguna manera la sintomatología. La selección fue arbitraria y se hizo tratando de ponerse en el lugar del usuario ("¿qué quiere sosegar al consumir?"), pero las combinaciones de cada una de las condiciones dentro de los grupos fueron coherentes entre sí. En líneas generales, los pares apuntaron a generar efectos de relajación y antiinflamatorios, aunque no en todos los casos fue del mismo "tipo".

Realizar ensayos clínicos requiere de mucho tiempo, dinero, personas y la evaluación de las distintas combinaciones posibles de principios activos. En este contexto, cuando se cuenta con datos como los de este trabajo, en los que se tiene la experiencia de los usuarios con distintas cepas (y por ende, combinaciones de principios activos) y el perfil químico de estas, es útil utilizar métodos computacionales para complementar las pruebas clínicas y por qué no, también buscar las opciones con mejores perspectivas, para optimizar tiempo y dinero. Los datos trabajados evidenciaron, una vez mas, lo que está en estudio hace años: la presencia del efecto *entourage* y que las combinaciones cannabinoide-terpeno podrían ser más relevantes de lo que se creía.

Bibliografía

[1] B. Warf. "High Points: A Historical Geography of Cannabis." *Geog Rev, 104(4): 414-438* (2014).

[2] Whittle B. A. Guy G. W. y P. Robson. *The medicinal uses of cannabis and cannabinoids (pp. 74-76).* Pharmaceutical Press, 2004.

[3] A. Hazekamp y J. T. Fischedick. "Cannabis-from cultivar to chemovar." *Drug testing and analysis, 4(7-8), 660-667.* (2012).

[4] Alessia et al. Ligresti. "Antitumor Activity of Plant Cannabinoids with Emphasis on the Effect of Cannabidiol on Human Breast Carcinoma." *Journal of Pharmacology and Experimental Therapeutics 318.3: 1375-1387* (2006).

[5] Abir T. El-Alfy et al. "Antidepressant-like effect of 9-tetrahydrocannabinol and other cannabinoids isolated from Cannabis sativa L." *Pharmacology Biochemistry and Behavior* 95.4 (2010), págs. 434-442.

[6] E. B. Russo. "Taming THC: potential cannabis synergy and phytocannabinoid-terpenoid entourage effects." *British journal of pharmacology, 163(7), 1344-1364.* (2011).

[7] A. de la Fuente y col. "Relationship among subjective responses, flavor, and chemical composition across more than 800 commercial cannabis varieties." *Journal of Cannabis Research, 2(1), 1-18* (2020).

[8] J. L. Erkelens y A. Hazekamp. "That which we call Indica, by any other name would smell as sweet." *Cannabinoids, 9(1), 9-15* (2014).

[9] A. L. Schwabe y M. E. McGlaughlin. "Genetic tools weed out misconceptions of strain reliability in Cannabis sativa: implications for a budding industry." *Journal of cannabis research, 1(1), 1-16.* (2019).

[10] G. W. Corder y D. I. Foreman. *Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach.* Wiley, 2009.

[11] G. James y col. *An introduction to statistical learning.* Springer, 2021.

[12] Vincent D Blondel et al. "Fast unfolding of communities in large networks." *J. Stat. Mech.* (2008).

[13] MI et al. Silva. "Central nervous system activity of acute administration of isopulegol in mice." *Pharmacol Biochem Behav, 88(2):141-7* (2007).

[14] Bounihi A et al. "In Vivo Potential Anti-Inflammatory Activity of Melissa officinalis L. Essential Oil." *Adv Pharmacol Sci.* (2013).

[15] El-Alfy AT et al. "Antidepressant-like effect of delta9-tetrahydrocannabinol and other cannabinoids isolated from Cannabis sativa L." *Pharmacol Biochem Behav., 95(4):434-42* (2010).

[16] R. Khan, S. Naveed y N. et al. Mian. "The therapeutic role of Cannabidiol in mental health: a systematic review." *J Cannabis Res 2, 2* (2020).

[17] Abdelaali B. et al. "Health beneficial and pharmacological properties of p-cymene." *Food and Chemical Toxicology, 153* (2021).

[18] Apel M.A. et al. "Anti-inflammatory activity of essential oil from leaves of Myrciaria tenella and Calycorectes sellowianus." *Pharm Biol, 48(4):433-8* (2010).

[19] Q. Yang. "(-)-Guaiol regulates RAD51 stability via autophagy to induce cell apoptosis in non-small cell lung cancer." *Oncotarget, 7(38), 62585–62597.* (2016).

[20] Jun YS et al. "Effect of eucalyptus oil inhalation on pain and inflammatory responses after total knee replacement: a randomized clinical trial." *Evid Based Complement Alternat Med.* (2013).

[21] Nayara Santos da Silveira et al. "The Aversive, Anxiolytic-Like, and Verapamil-Sensitive Psychostimulant Effects of Pulegone." *Biological and Pharmaceutical Bulletin.* (2014).

[22] L.C. Di Stasi et al. "Medicinal plants popularly used in the Brazilian Tropical Atlantic Forest." *Fitoterapia* (2002).

[23] Shuso Takeda y col. "Cannabidiolic Acid as a Selective Cyclooxygenase-2 Inhibitory Component in Cannabis." *Drug Metabolism and Disposition* (2008).

[24] Udoh M et al. "Cannabichromene is a cannabinoid CB2 receptor agonist." *Br J Pharmacol* (2019).

[25] Shinjyo N y Di Marzo V. "The effect of cannabichromene on adult neural stem/progenitor cells." *Neurochem Int.* (2013).