ReciPy, un proyecto para analizar recetas de comida

Jazmín López Chacón
Licenciatura Tecnologías para la Información en Ciencias
Escuela Nacional de Estudios Superiores
jlc1839@gmail.com

Figura 1: Logo creado para el proyecto

RESUMEN

En este reporte abordaremos el proceso y los resultados de hacer un análisis exploratorio sobre recetas de cocina recolectadas de *Food.com* .

KEYWORDS

agrupamiento, patrones frecuentes, evaluación intercluster

ACM Reference Format:
Jazmín López Chacón. 2022. ReciPy, un proyecto para analizar recetas de comida. In *Proceedings of Minería de Datos.* ACM, New York, NY, USA, 7 pages.

1. INTRODUCCIÓN

La comida es algo que forma parte de nuestra vida diaria, por lo que no es de extrañarse que en Internet nos encontremos con páginas donde las personas pueden compartir sus platillos favoritos y cómo prepararlos. Una de esas páginas es *Food.com*, que tiene más de 10 años existiendo, por lo que nos puede proveer una gran variedad de recetas a partir de las cuales se puede obtener diversa información de las costumbres culinarias.

En esta ocasión el análisis estará enfocado en el agrupamiento y patrones frecuentes de las comidas resultantes después de seguir las recetas.

Minería de Datos, Enero 2022, Morelia, Michoacán

2. DATOS

Los datos que se utilizaron provienen del usuario *Shuyang Li* en la plataforma Kaggle. El conjunto de datos en cuestión es *Food.com Recipes and Interactions*, y fueron usados por el usuario y compañía para la publicación *Generating Personalized Recipes from Historical User Preferences*[Majumder et al. 2019].

Dentro de todos los conjuntos que se nos ofrecen, solamente se tomaron tres:

- `RAW_recipes.csv`: Contiene todas las recetas que se encontraban en *Food.com* dos años atrás, tal y como los usuarios las publicaron en la página.
- `PP_recipes.csv`: Contiene la mayoría de las recetas que se encontraban en la página dos años atrás, solo que están preprocesadas para que sea más sencillo encontrar coincidencias con los ingredientes.
- `ingr_map.pkl`: Contiene la conversión de los ingredientes del nombre como tal a un código de identificación que se encuentra en las recetas preprocesadas.

Ahora que ya conocemos cómo se obtuvieron los datos, lo siguiente es hablar de su contexto y contenido.

2.1. Contexto

Food.com es una página procedente de Estados Unidos, por lo que a pesar de que se pueda subir recetas de cualquier parte del mundo, es de esperarse que las mayorías de las recetas correspondan a la gastronomía estadounidense, o bien, a su interpretación de otras comidas del mundo.

El mecanismo de la página es muy sencillo, el usuario se registra, selecciona la opción *ADD RECIPE* desde el menú de su perfil y simplemente empieza a llenar los campos necesarios para que su

receta sea guardada, y si lo desee pueda publicarla.

Los campos que se rellenan son:

- Nombre de la receta.
- Descripción.
- Categorías.
- Fotografía.
- Tiempo de preparación (en minutos).
- Tiempo de cocción (en minutos).
- Porciones.
- Ingredientes.
- Instrucciones.

El usuario ve todos los campos de la receta, además de la información nutrimental, en gramos y en porcentaje de valor diario, que la página calcula automáticamente a partir de los ingredientes de la receta y la cantidad de porciones que salen de ella.

2.2. Contenido

Como se mencionó en la breve descripción de los conjuntos de datos usados, es en `RAW_recipes.csv` donde se encuentra toda la información referente a las recetas, es decir, aquí se encuentran los atributos los atributos que el usuario llena al momento de crear una receta, a excepción de la fotografía y la cantidad de porciones. Además se agrega la fecha en la que se subió la receta y el identificador del usuario que lo hizo.

Algo muy importante de mencionar es que dentro de la información nutrimental que se encuentra en el conjunto de datos, solamente se encuentra el valor calórico, las grasas totales (PVD[1]), los azúcares (PVD), el sodio (PVD), la proteína (PVD),las grasas saturadas (PVD) y el total de carbohidratos (PVD). Como podemos ver la mayoría de la información se presenta en PVD, lo cual nos deja con menos información a comparación de tener todo en gramos. Sin embargo, esto resulta beneficiario, pues podemos considerar que los datos ya están normalizados.

3. LIMPIEZA DE DATOS

Una vez que conocemos con qué estamos trabajando, es momento de empezar a ver si existe ruido y tomar una decisión sobre lo que se hará con él.

Lo primero es que si la información nutrimental depende de la cantidad de porciones que el usuario haya proporcionado, es altamente probable que nos topemos con valores muy altos ya que por defecto se queda en una porción, para ver qué pasa realmente con estos valores, podemos revisar algunos valores estadísticos como lo son la media, a desviación estándar, los cuantiles y los valores mínimos y máximos.

[1]Porcentaje de Valor Diario

	Calorías	Grasas Totales (PVD)	Azúcares (PVD)	Sodio (PVD)	Proteína (PVD)	Grasas Saturadas (PVD)	Carbohidratos (PVD)
μ	459.87	34.80	80.83	28.35	34.43	43.93	15.10
σ	1247.49	64.08	894.13	99.72	54.29	89.47	90.95
mín	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Q_1	177.30	9.00	9.00	5.00	7.00	7.00	4.00
Q_2	310.70	20.00	24.00	15.00	19.00	22.00	9.00
Q_3	504.90	40.00	65.00	32.00	51.00	51.00	16.00
máx	434360.20	4331.00	362729.00	14664.00	6552.00	6875.00	36098.00

Tabla 1: Valores descriptivos de la información nutrimental disponible

A partir de estos valores podemos notar que en efecto, hay recetas que sobrepasan por mucho los valores esperados y estas son menos del 25 %, por lo que tiene sentido considerarlas como ruido. Sin embargo, no nos vamos a quedar solo con recetas que no sobrepasen el 100 % en porcentaje de valor diario, pues esto nos puede quitar información relevante sobre la alimentación en Estados Unidos, además que al momento de cocinar no se suele tener consciencia del valor nutritivo que puedan tener los ingredientes y las porciones usadas.

Para hacer el corte se consideró los valores que había en el tercer cuantil, y se decidieron de manera manual que una receta se consideraría como ruido si sobrepasaba alguno de los siguientes valores:

- Grasas totales (PVD) ≥ 150
- Azúcares (PVD) ≥ 200
- Sodio (PVD) ≥ 125
- Proteína (PVD) ≥ 175
- Grasas Saturadas (PVD) ≥ 175
- Carbohidratos (PVD) ≥ 100

Es importante mencionar que no se puso un máximo para las calorías porque si bien existe la idea de que una persona "normal" solo debe consumir 2,000 cal al día, la realidad es que no estan sencillo saber la cantidad máxima que se debe consumir.

Una vez detectadas las recetas que son "ruido", podemos var su descripción estadística para ver lo que está pasando con estos valores fuera de lo esperado.

	Calorías	Grasas Totales (PVD)	Azúcares (PVD)	Sodio (PVD)	Proteína (PVD)	Grasas Saturadas (PVD)	Carbohidratos (PVD)
μ	1596.35	122.64	456.01	96.43	72.84	160.75	62.26
σ	3621.20	162.81	2742.70	294.41	136.48	230.57	276.60
mín	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Q_1	622.03	27.00	67.00	13.00	13.00	29.00	20.00
Q_2	1055.10	78.50	244.00	40.00	38.00	92.00	33.00
Q_3	1913.10	164.00	495.00	114.00	99.00	207.00	74.00
máx	434360.20	4331.00	362729.00	14664.00	6552.00	6875.00	36098.00

Tabla 2: Valores descriptivos de la información nutrimental de recetas que se pueden considerar como ruido.

De esta tabla podemos destacar que al menos el 50 % de las recetas consideradas como ruido tienen el doble de azúcar de lo que se debería de consumir en un día, esto es relevante ya que al notar que en los azúcares los valores eran mayores que para el resto de atributos, se decidió tener una cota mucho más relajada. También

podemos notar que para todos los atributos el valor mínimo sigue siendo cero, lo cual nos indica que en este conjunto de recetas hay algunas que solo tienen exceso en un atributo.

Se podría hacer un análisis más completo sobre estas recetas, pero al no ser nuestro interés principal, lo dejaremos hasta aquí.

3.1. Creación de una muestra

Dado el poder computacional con el se cuenta, no es posible trabajar con todas las recetas que quedaron después de la limpieza, por lo que se obtuvo una muestra del 10 % de manera aleatoria.

Una vez se tiene la muestra con los datos limpios ya es posible empezar a hacer el análisis.

4. ANÁLISIS SOBRE LA INFORMACIÓN NUTRIMENTAL

Lo primero que podemos hacer es revisar los valores estadísticos que tenemos para nuestra muestra.

	Calorías	Grasas Totales (PVD)	Azúcares (PVD)	Sodio (PVD)	Proteína (PVD)	Grasas Saturadas (PVD)	Carbohidratos (PVD)
μ	326.510	24.50	37.02	20.39	29.71	30.26	9.59
σ	218.90	22.59	42.05	21.08	30.48	31.81	8.31
mín	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Q_1	164.80	8.00	9.00	5.00	7.00	6.00	3.00
Q_2	283.20	18.00	21.00	14.00	17.50	20.00	8.00
Q_3	437.10	34.00	50.00	29.00	47.00	44.00	14.00
máx	1822.60	149.00	200.00	125.00	175.00	175.00	99.00

Tabla 3: Valores descriptivos de la información nutrimental de la muestra seleccionada.

Podemos destacar que para todos los atributos el valor mínimo es cero, lo que nos indica la presencia de recetas que estén bajo la etiqueta de "saludable". Otra observación es que tanto la media como la desviación estándar disminuyeron bastante a comparación de los valores que se tenían antes de dejar a un lado aquellas recetas con excesos. Vemos que para todos los atributos se cumple que la mediana es menor que la media, y dada la naturaleza de este conjunto de datos, nos vamos a enfocar más en la mediana. Al revisarla vemos que para todos los atributos se tienen valores bastante razonables para una sola porción de comida, si consideramos que una persona normalmente come entre tres y cuatro porciones en un día. En cuanto a los niveles de azúcares, vemos que la diferencia con los otros atributos disminuyó a pesar de haber decidido conservar la "ventaja" que tenía, por lo que podemos decir que la limpieza fue buena.

Lo siguiente es empezar el agrupamiento, el cual lo haremos usando el algoritmo de *k-medias*, así que lo primero es decidir la cantidad de grupos que vamos a considerar.

4.1. Elección de grupos

Para elegir la cantidad de grupos adecuada, primero vamos a revisar el valor que nos sugiere el método del codo. La gráfica resultante es la siguiente:

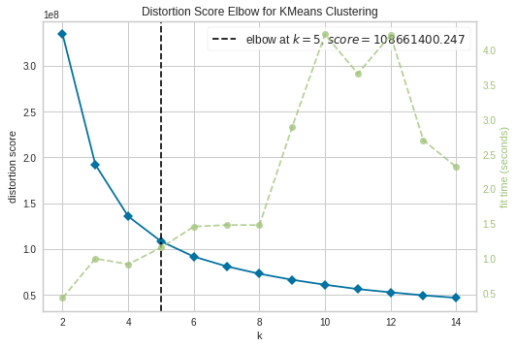

Figura 2: Gráfica de codo para el agrupamiento de las recetas según su valor nutrimental.

Aunque en la figura se nos sugiere que cinco grupos es lo ideal, podemos ver que hay otro codo considerable cuando $k = 4$, así que vale la penar revisar cuál de estos dos valores nos conviene usar más, para ello usaremos la método de *Silhouette*, esto nos genera las siguientes dos gráficas.

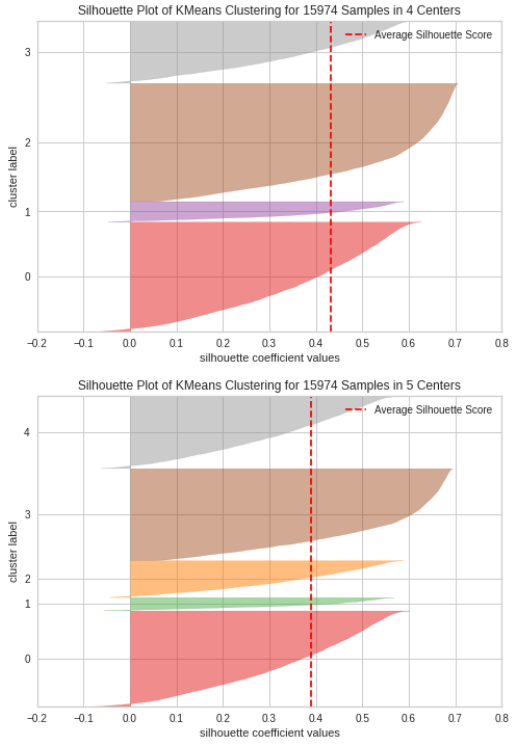

Figura 3: Gráficas de Silhouette para el agrupamiento de las recetas según su valor nutrimental cuando $k = 4$ y $k = 5$.

Vemos que para $k = 4$ obtenemos un valor más alto y las siluetas resultantes son mejores, por lo que usaremos cuatro grupos para este agrupamiento.

4.2. Resultados del agrupamiento

Una vez hecho el agrupamiento en cuatro grupos, podemos revisar la cantidad de recetas en cada grupo, eso nos resulta en la siguiente gráfica:

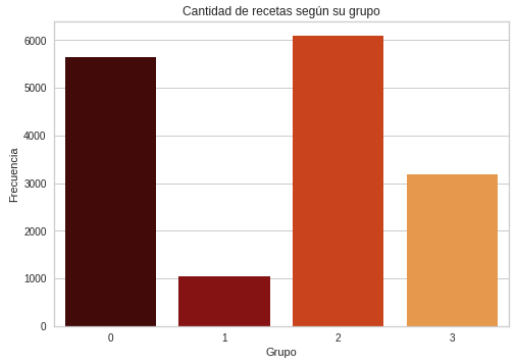

Figura 4: Cantidad de recetas en cada grupo.

Notamos que hay dos grupos dominantes, pero para saber lo que realmente significa, podemos hacer una reducción de dimensiones usando una descomposición de valores singulares (SVD).

Al hacerla y poner un color diferente a cada grupo obtenemos el siguiente resultado:

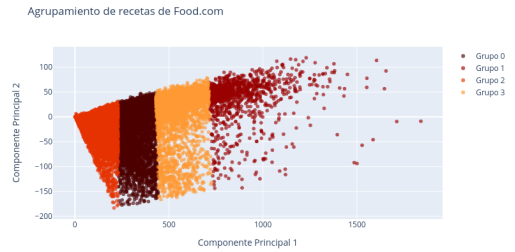

Figura 5: Representación en dos dimensiones de las recetas según su información nutrimental, separadas por grupo.

Hay dos cosas que resaltar de esta gráfica, la primera es que el resultado parece que las recetas tienen un mismo rango para los atributos, es decir, si tienen un nivel bajo de azúcares, también lo tienen de carbohidratos. La segunda es que parecer que la agrupación se hizo a partir de estos rangos, creando una nueva "escala" que sería la siguiente:

1. Grupo 2 (valores bajos)
2. Grupo 0 (valores medio bajos)
3. Grupo 3 (valores medio altos)
4. Grupo 1 (valores altos)

Para comprobar esta teoría, tendríamos que revisar qué pasa con cada atributo, la forma más sencilla de hacerlo es usando gráficas de caja como las que se presentan a continuación.

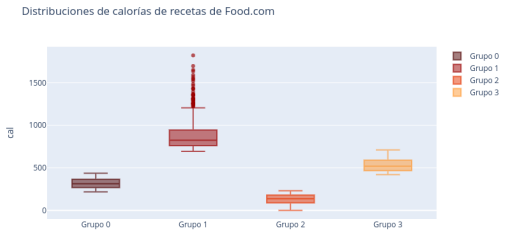

Figura 6: Distribución de calorías según el grupo al que se pertenece.

Para las calorías se mantiene la escala que se mencionó anteriormente.

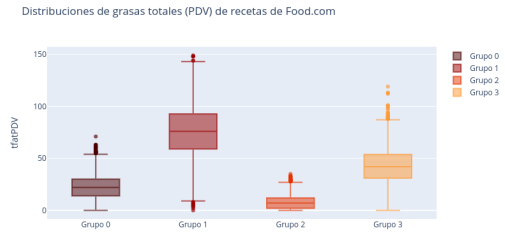

Figura 7: Distribución de grasas totales (PVD) según el grupo al que se pertenece.

Para las grasas totales se mantiene la escala mencionada, sin embargo, podemos notar una mayor variabilidad para el grupo 1.

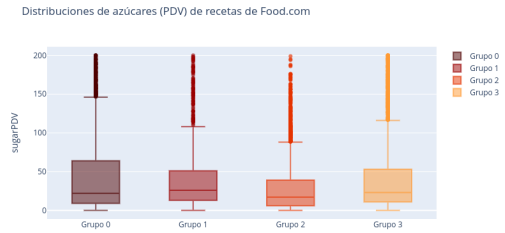

Figura 8: Distribución de azúcares (PVD) según el grupo al que se pertenece.

Con los azúcares pasa algo interesante y es que las diferencias entre los grupos no son tan drásticas como en los anteriores, pues todos los grupos tienen valores similares, pero sobretodo, recetas que contienen más azúcar de la recomendada en un día. Además, en este atributo se rompe la escala, ya que es el grupo 0 el que tiene valores más altos.

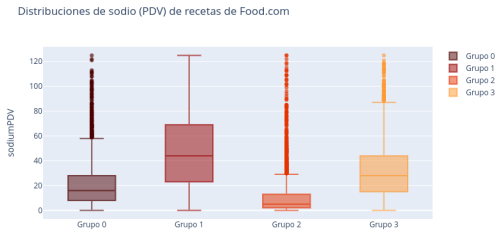

Figura 9: Distribución de sodio(PVD) según el grupo al que se pertenece.

Con el sodio pasa lo mismo que con los azúcares y es que todos los grupos tienen valores que se salen de lo recomendado, sin embargo, regresamos a la escala propuesta en un inicio.

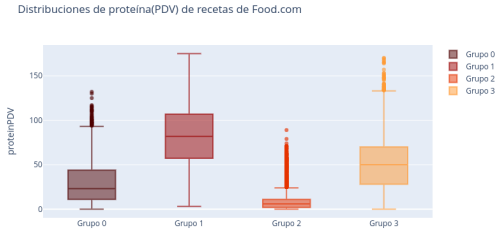

Figura 10: Distribución de proteína(PVD) según el grupo al que se pertenece.

Para la proteína se conserva la escala propuesta y podemos notar menos valores atípicos.

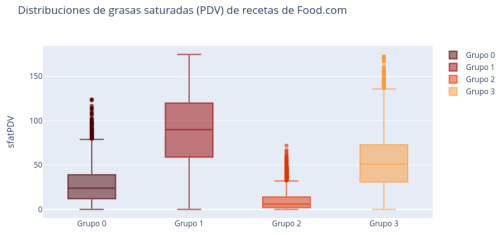

Figura 11: Distribución de grasas saturadas (PVD) según el grupo al que se pertenece.

Para las grasas saturadas se mantiene la escala que se había mencionado anteriormente, además el grupo 1 presenta de nueva una gran variabilidad que lo hace tener elementos que no tienen grasas saturadas y elementos con el 175 % de grasas saturadas, dentro de sus valores esperables, es decir, tanto los valores altos como bajos no son considerados como atípicos.

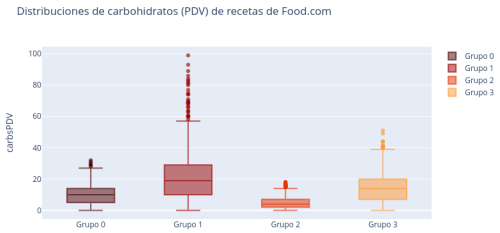

Figura 12: Distribución de carbohidratos (PVD) según el grupo al que se pertenece.

Podemos ver que de manera general, los carbohidratos se mantienen bajos para todos los grupos, solo que el grupo 1 tiene valores atípicos que sobresalen del resto.

De manera general, la escala propuesta se conserva a través de los diferentes grupos, y al ser los grupos 2 y 0 los que tienen más recetas, podemos decir que las recetas de la muestra aportan una cantidad de nutrimentos razonable para una sola porción de comida. También debemos destacar los niveles de azúcar ya que tienden a ser altos para que se trate de una sola porción de alimentos.

5. ANÁLISIS SOBRE LOS INGREDIENTES

Dentro de nuestra muestra contamos con 4,877 ingredientes de los 8,023 que hay disponibles en el conjunto de datos completo, por lo que tenemos una gran variedad que nos permite hacer principalmente dos exploraciones.

5.1. Patrones frecuentes

La búsqueda de patrones frecuentes en las recetas nos sirve para entender las combinaciones de ingredientes más comunes y a partir de estas poder identificar distintos tipos de comida.

En este caso particular, podríamos encontrar alguna combinación que nos ayude a comprender lo que pasa con los azúcares para esta muestra.

Para la búsqueda de patrones se necesita considerar un umbral, para este conjunto de datos no se contaba con un valor fijo, por lo que se fue probando con diferentes valores hasta que se pudo obtener patrones que incluyeran a más de un elemento. Lamentablemente, el umbral que nos permitía esto fue de 0.075, lo cual es muy bajo, así que podríamos decir que como tal no hay patrones frecuentes dentro de la muestra. Sin embargo, valdría la pena revisar qué combinaciones son las comunes.

En la tabla 4 se muestran los conjuntos que nos lanzó la búsqueda de patrones frecuentes. En esos resultados vemos que la sal es el ingrediente más usado y se encuentra en más de la tercera parte de las recetas, otros ingredientes que destacan son la mantequilla y la cebolla.

Soporte	Ingredientes
0.3700	{salt}
0.2380	{butter}
0.2347	{onion}
0.2058	{egg}
0.1901	{olive oil}
0.1673	{sugar}
0.1611	{garlic clove}
0.1392	{water}
0.1246	{milk}
0.1122	{pepper}
0.1114	{salt, egg}
0.1083	{flour}
0.1081	{butter, salt}
0.0976	{onion, salt}
0.0884	{pepper, salt}
0.0882	{sugar, salt}
0.0829	{scallion}
0.0818	{butter, egg}
0.0814	{garlic}
0.0762	{salt and pepper}

Tabla 4: Patrones frecuentes encontrados en la muestra con un umbral de 0.075.

Siguiendo con la línea de ingredientes que aparecen con mayor frecuencia que otros, tenemos al azúcar, que aparece en el 16.73 % de la muestra, lo que tiene relación con los valores nutrimentales que encontramos para los azúcares. Por otro lado, al revisar las combinaciones, seguimos notando una importante presencia de sal y pasa algo curioso con la mezcla de sal y pimienta, ya que aparece dos veces, una como un conjunto de dos ingredientes y otra como uno solo, esto se debe a la propia naturaleza de las recetas, que son creadas por cualquier persona con una cuenta en el sitio.

Otra observación interesante es que salvo al huevo, los ingredientes que más se repiten suelen ser secundarios o complementarios, es decir, no se suelen usar como base de un platillo.

Finalmente, la falta de patrones frecuentes, con más de un ingrediente, nos dice que en la cocina estadounidense hay una gran variedad de platillos provenientes de diferentes partes del mundo, así como su población.

5.2. Agrupamiento

Si seguimos deseando encontrar estilos de comida, el agrupamiento usando los ingredientes es una opción. Para hacerlo, es necesario ubicar a cada receta en un espacio con 4,877 dimensiones, una por cada ingrediente, sus c´ordenadas estarán dadas de la siguiente manera:

$$receta = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{4877} \end{pmatrix}, \quad x_i = \begin{cases} 1 & \text{si el ingrediente } i \text{ está en la receta} \\ 0 & \text{si no} \end{cases}$$

Ahora que ya tenemos a las recetas representadas como vectores, podemos hacer las comparaciones necesarias para hacer un agrupamiento.

5.2.1. Elección de grupos. Al igual que con la información nutrimental, usaremos primero el método del código, el cual nos regresa la siguiente gráfica:

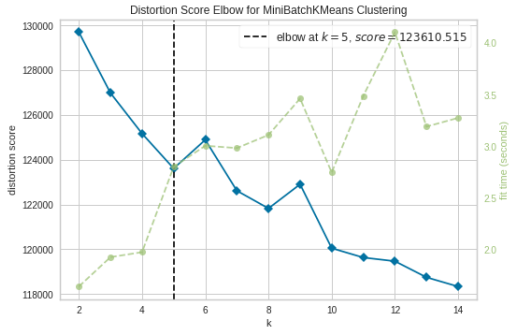

Figura 13: Gráfica de codo para el agrupamiento de recetas según sus ingredientes.

De nuevo tenemos que elegir entre $k = 4$ o $k = 5$, y esto lo haremos usando el método Silhouette, que nos resulta en las siguientes gráficas:

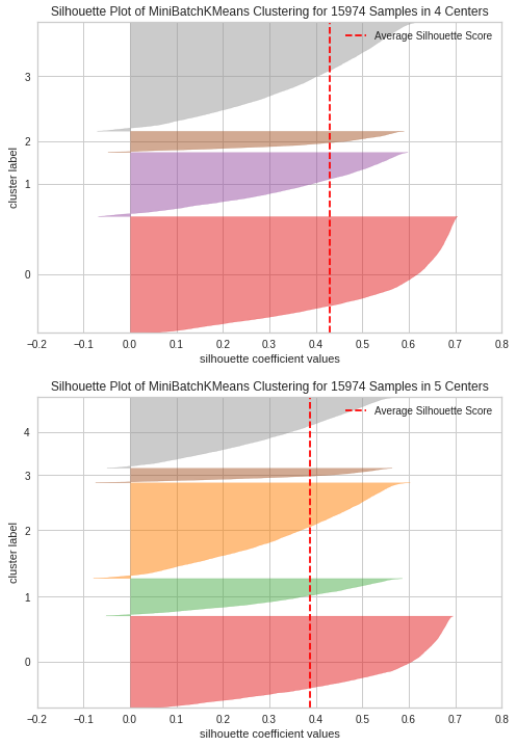

Figura 14: Gráficas de Silhouette para el agrupamiento de las recetas según sus ingredientes cuando $k = 4$ y $k = 5$.

De nuevo vemos que nos es conveniente usar cuatro grupos.

5.2.2. Resultados del agrupamiento. Podemos revisar la distribución de los grupos encontrados.

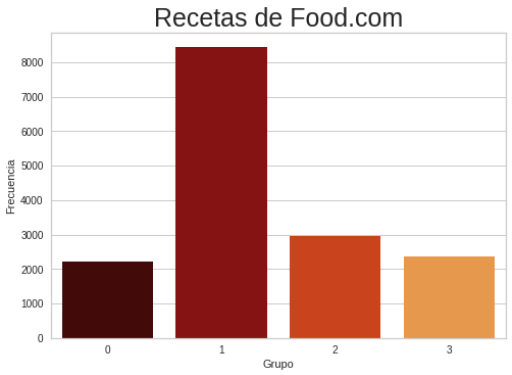

Figura 15: Cantidad de recetas en cada grupo según los ingredientes.

Vemos que hay un grupo que destaca más que los demás por la cantidad de recetas que se encuentran en él.

Al igual que con la información nutrimental, podemos hacer una reducción de dimensiones para ver cómo se ve el agrupamiento, sin embargo, es más complicado de interpretar para los ingredientes.

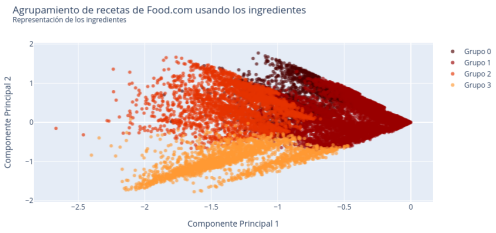

Figura 16: Representación de dos dimensiones de las recetas según sus ingredientes, separadas por grupo.

En esta representación, vemos que hay una notoria separación entre los grupos y que el grupo 2 parece ser el más disperso.

6. COMPARACIÓN ENTRE LOS DOS ANÁLISIS

Ambos análisis comparten un agrupamiento de recetas en cuatro grupos, así que sería interesante ver qué tan parecidos son estos agrupamientos entre sí.

Una primera comparación podría darse al revisar el agrupamiento de los ingredientes en la representación de la información nutrimental.

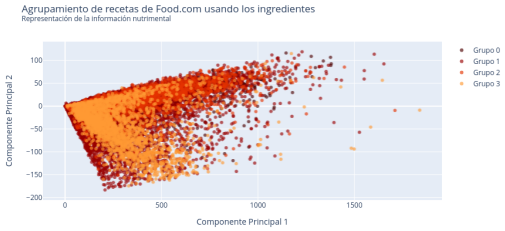

Figura 17: Representación de la información nutrimental en dos dimensiones con el agrupamiento por sus ingredientes.

En esta primera comparación no podemos apreciar un claro empalme, parece que todos los grupos están muy dispersos en cuanto a información nutrimental, por lo que será necesario usar métricas externas de evaluación.

Rand Index Al comparar estos dos agrupamientos obtenemos un valor de 0.5614, lo que nos indica que son un poco más parecidos que diferentes entre sí, a penas y salen del punto intermedio entre iguales y completamente diferentes.

Información Mutua Normalizada El valor para estos dos agrupamientos es de 0.0115, lo cual es muy bajo y nos indica que estos dos agrupamientos no comparten mucha información.

Con estas métricas podemos darnos cuenta que los agrupamientos no son parecidos entre sí, cada uno nos revela información diferentes.

7. CONCLUSIONES

En esta exploración de datos pudimos apreciar la gran variedad de estilos de cocina que existe en Estados Unidos, y sin embargo, hay ingredientes que son constantes como la sal y mantequilla.

Si nos vamos al lado de qué tan saludable es su cocina, podríamos decir que la mayoría de las recetas se encuentran por debajo del límite de lo adecuado para ser una sola comida. Sin embargo, el nivel de azúcar puede ser algo preocupante a pesar de que solo el 11.3 % de la población estadounidense padece algún tipo de diabetes [for Disease Control et al. 2020].

Finalmente, el haber conocido el contexto en el cual fueron creadas estas recetas fue de suman importancia para la realización del análisis, pues nos ayudó a tomar decisiones importantes.

REFERENCIAS

Centers for Disease Control, Prevention, et al. 2020. National diabetes statistics report, 2020. *Atlanta, GA: Centers for Disease Control and Prevention, US Department of Health and Human Services* (2020), 12–15.

Bodhisattwa Prasad Majumder, Shuyang Li, Jianmo Ni, and Julian McAuley. 2019. Generating Personalized Recipes from Historical User Preferences. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 5976–5982. https://doi.org/10.18653/v1/D19-1613