IBM Data Science Proyecto Capstone - La Batalla de los Vecindarios

‘Los Pollos Hermanos’ en Madrid

Juan Carlos Rua

Introducción y motivación

- Los Pollos Hermanos es una cadena de comida rápida extranjera que como parte de su expansión a Europa ha abierto el año pasado un restaurante piloto en Madrid.
- Los resultados obtenidos por el restaurante han superado las expectativas de la empresa por lo que han decidido continuar su expansión. Para la primera fase han decidido **abrir cinco restaurantes más en la ciudad de Madrid**.
- La dirección de la empresa no está familiarizada con la ciudad por lo que le ha encargado a una consultora, la misma que eligió la ubicación del primer restaurante, que **busque cuales son los barrios más parecidos al actual**. Ya que es donde la dirección de la empresa espera que se encuentre en mayor proporción su público objetivo.
- El propósito del estudio es **determinar cuáles, de los 131 barrios de la ciudad de Madrid, son más parecidos al barrio en el que se encuentra el restaurante piloto**.
- El restaurante piloto se encuentra en el barrio de Portazgo en el distrito de Puente de Vallecas

Material y Métodos

K-means

Para ello la empresa consultora ha decidido agrupar los barrios en clústeres utilizando K-means, una técnica de aprendizaje no supervisado. Esta técnica utiliza las características de los barrios, en este caso los tipos de negocios y lugares de interés que hay en cada barrio, para agruparlos en K grupos. Este algoritmo busca maximizar la distancia entre grupos y minimizar la distancia intra-grupo.

Gracias a un primer análisis realizado empleando información demográfica se ha determinado que en la ciudad de Madrid hay 7 tipos de barrios. Por lo que en este caso se agruparan los barrios en 7 clústeres.

1) *k* centroides iniciales (en este caso *k*=3) son generados aleatoriamente dentro de un conjunto de datos (mostrados en color).

2) *k* grupos son generados asociándole el punto con la media más cercana. La partición aquí representa el diagrama de Voronoi generado por los centroides.

3) EL centroide de cada uno de los *k* grupos se recalcula.

4) Pasos 2 y 3 se repiten hasta que se logre la convergencia.

[Imagen de Wikipedia: https://es.wikipedia.org/wiki/K-medias]

Material y Métodos

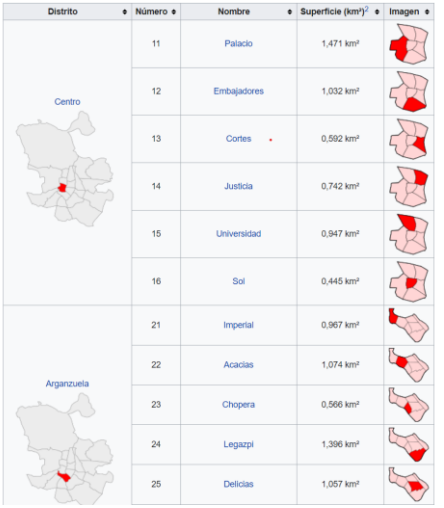
Datos

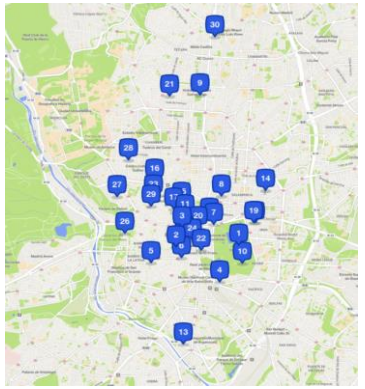
Wikipedia:

Se utilizará Wikipedia para obtener la lista de barrios y distritos de la Ciudad de Madrid.

Foursquare:

Se utilizará Foursquare para conseguir información sobre los barrios y distritos, concretamente sobre el tipo de negocios y lugares de interés que hay en cada zona accederemos a los datos de Foursquare mediante su REST API. Esta información se utilizará para agrupar los barrios de Madrid en clústeres en función del tipo de negocios y lugares de interés que hay en cada barrio.





Material y Métodos

Datos

Nominatim:

Nominatim se utilizará para obtener las coordenadas de cada barrio de Madrid. Nominatim utiliza los datos de OpenStreetMap para encontrar las coordenadas de una dirección (geocoding). La API de Nominatim ha dado error en 2 de los 131 barrios de Madrid.

GeoHack:

Al tratarse únicamente de dos errores las coordenadas de estos dos barrios se han obtenido manualmente buscándolas en geohack.

Material y Métodos

Datos

Etapas de la preparación de los datos:

1. Descargar la lista de barrios de Madrid de Wikipedia y darle el formato adecuado para las etapas posteriores.

2. El siguiente paso es añadir al dataframe las coordenadas de cada barrio de Madrid que obtenemos con Nominatim.

3. El último paso de la obtención de datos es la obtención de los lugares de interés en cada distrito utilizando la Rest API de Foursquare. Los datos se han estructurado en el siguiente dataframe.

	Distrito	Número	Nombre	Superficie (km²)[2]	Imagen
0	Centro	11	Palacio	1,471 km²	NaN
1	Centro	12	Embajadores	1,032 km²	NaN
2	Centro	13	Cortes	0,592 km²	NaN
3	Centro	14	Justicia	0,742 km²	NaN
4	Centro	15	Universidad	0,947 km²	NaN

	Distrito	Código	Barrio	Tamaño
0	Centro	11	Palacio	1.471
1	Centro	12	Embajadores	1.032
2	Centro	13	Cortes	0.592
3	Centro	14	Justicia	0.742
4	Centro	15	Universidad	0.947

	Distrito	Código	Barrio	Tamaño	Latitud	Longitud
0	Centro	11	Palacio	1.471	40.415129	-3.715618
1	Centro	12	Embajadores	1.032	40.409681	-3.701644
2	Centro	13	Cortes	0.592	40.414348	-3.698525
3	Centro	14	Justicia	0.742	40.423957	-3.695747
4	Centro	15	Universidad	0.947	40.425310	-3.706630

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Palacio	40.415129	-3.715618	Santa Iglesia Catedral de Santa María la Real ...	40.415767	-3.714516	Church
1	Palacio	40.415129	-3.715618	Plaza de La Almudena	40.416320	-3.713777	Plaza
2	Palacio	40.415129	-3.715618	Taberna Rayuela	40.413179	-3.713496	Tapas Restaurant
3	Palacio	40.415129	-3.715618	Corral de la Morería	40.412619	-3.714249	Performing Arts Venue
4	Palacio	40.415129	-3.715618	Palacio Real de Madrid	40.417940	-3.714259	Palace

Metodología del análisis

1. Codificación **onehot** para analizar los sitios más frecuentes en cada barrio.

2. El siguiente paso es **agrupar por barrios** utilizando la frecuencia de la ocurrencia de cada categoría. Es decir, con qué frecuencia aparece cada tipo de lugar en cada barrio. Hay 3 barrios para los que no hay sitios disponibles en la base de datos de Foursquare y serán eliminados, ya que no hay información para su clasificación.

3. Obtenemos los sitios más frecuentes de cada barrio

4. Agrupamos los barrios en 7 clústeres utilizando **K-means**. El criterio para la evaluación la proporción de cada tipo de lugar en cada barrio.

	Accessories Store	Adult Boutique	Airport	American Restaurant	Arcade	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Art Museum	Art Studio	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Automotive Shop	BBQ Joint	Bagel Shop	Bakery	Bar	Basketball Court	Basketball Stadium
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

	Neighborhood	Accessories Store	Adult Boutique	Airport	American Restaurant	Arcade	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Art Museum	Art Studio	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Automotive Shop	BBQ Joint	Bagel Shop	Bakery	Bar
0	Abrantes	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.1	0.0	0.0	0.0	0.100000	0.000000
1	Acacias	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.046512	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.023256	0.093023
2	Adelfas	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.027027	0.0	0.0	0.0	0.0	0.054054	0.054054
3	Aeropuerto	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000
4	Alameda de Osuna	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.045455	0.045455

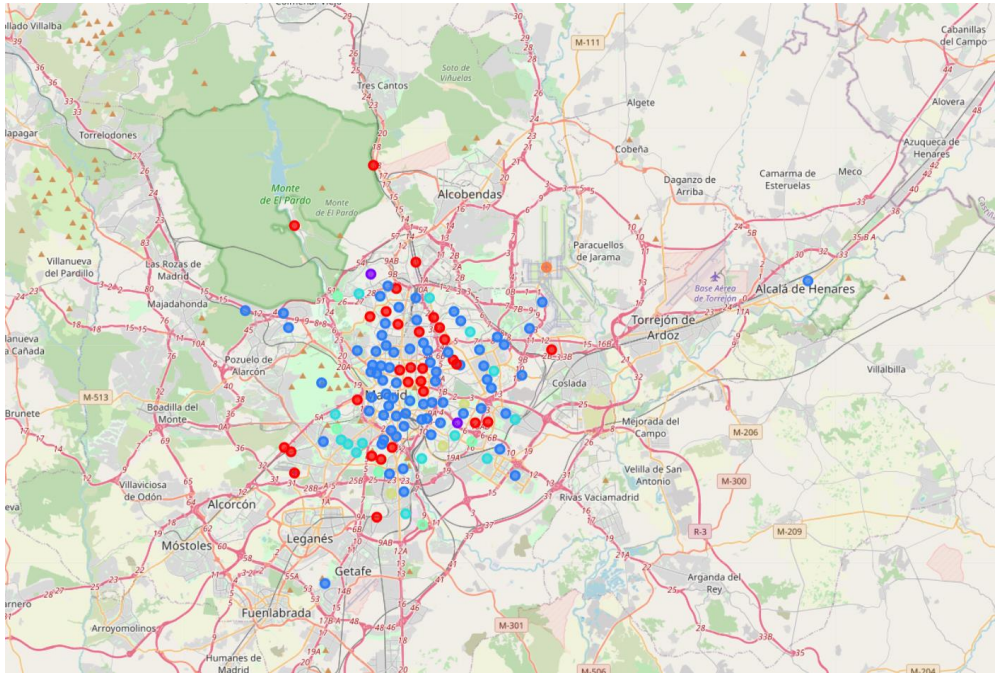
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue
0	Abrantes	Pizza Place	Restaurant	Grocery Store	Fast Food Restaurant	Bakery	Metro Station
1	Acacias	Bar	Pizza Place	Tapas Restaurant	Spanish Restaurant	Art Gallery	Park
2	Adelfas	Supermarket	Grocery Store	Fast Food Restaurant	Bar	Bakery	Tapas Restaurant
3	Aeropuerto	Business Service	Football Stadium	Fish Market	Flea Market	Flower Shop	Food
4	Alameda de Osuna	Tapas Restaurant	Hotel	Restaurant	Pizza Place	Smoke Shop	Shop & Service

	Distrito	Código	Barrio	Tamaño	Latitud	Longitud	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Centro	11	Palacio	1.471	40.415129	-3.715618	2.0	Tapas Restaurant	Spanish Restaurant	Plaza	Restaurant	Bar
1	Centro	12	Embajadores	1.032	40.409681	-3.701644	2.0	Bar	Café	Hostel	Plaza	Spanish Restaurant
2	Centro	13	Cortes	0.592	40.414348	-3.698525	2.0	Hotel	Plaza	Restaurant	Bar	Spanish Restaurant
3	Centro	14	Justicia	0.742	40.423957	-3.695747	2.0	Restaurant	Bakery	Vegetarian / Vegan Restaurant	Spanish Restaurant	Hotel
4	Centro	15	Universidad	0.947	40.425310	-3.706630	2.0	Café	Bookstore	Cocktail Bar	Coffee Shop	Plaza

Resultados

- Como resultado del análisis del apartado anterior hemos clasificado los barrios de la ciudad de Madrid en 7 grupos, en función del tipo de lugares que hay en cada barrio. En verde podemos ver todos los barrios que se encuentran en el mismo clúster que el restaurante piloto.

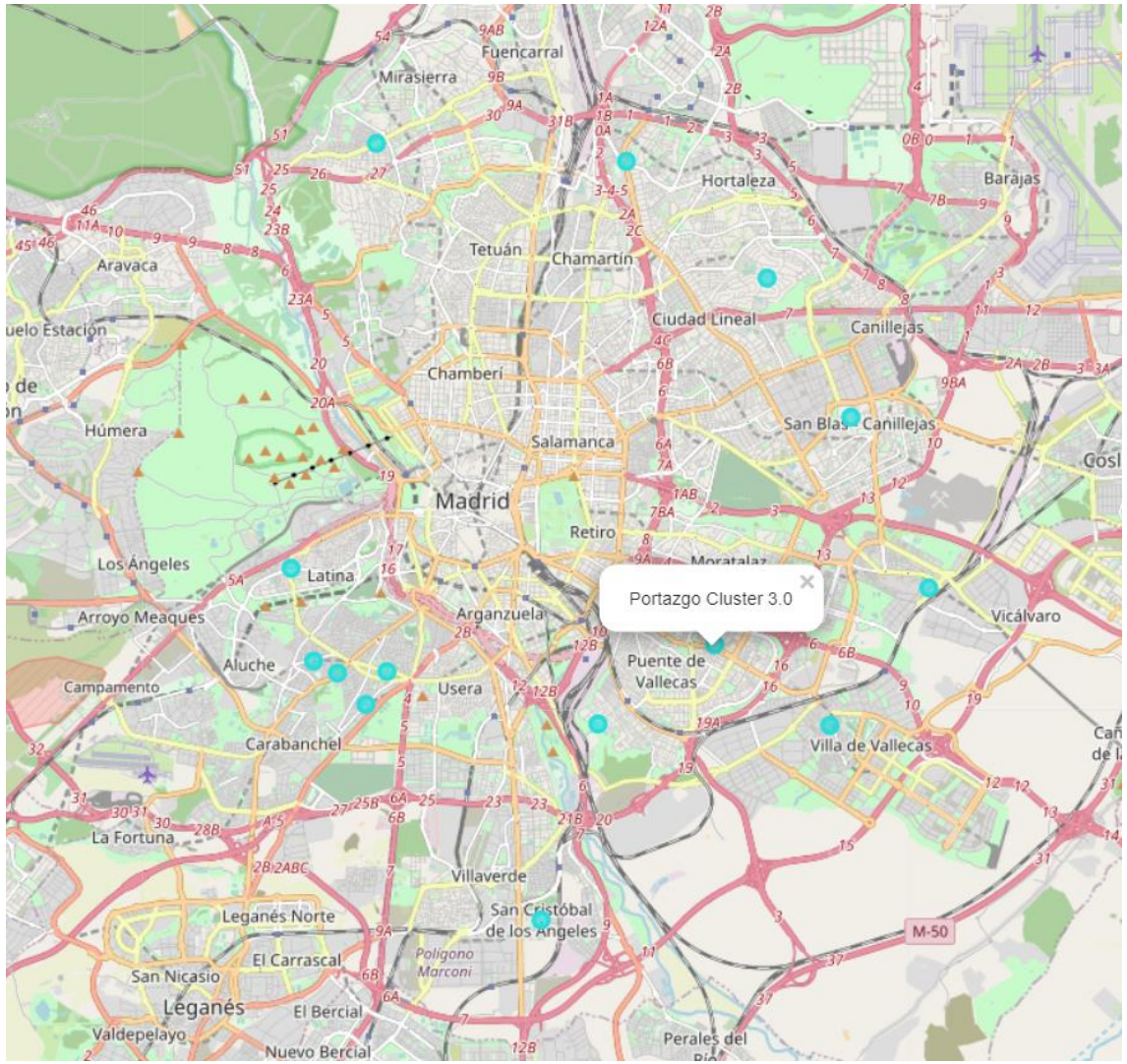
- El restaurante piloto se encuentra en el barrio de Portazgo en el distrito de Puente de Vallecas. En la tabla de la derecha se puede ver una lista con los demás barrios que están en el mismo clúster que el restaurante piloto, y los tipos de lugares más frecuentes en cada uno de ellos.



Distrito	Barrio	Longitud	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue
Fuencarral-El Pardo	Peñagrande	-3.725842	3.0	Soccer Field	Grocery Store	Argentinian Restaurant	Tapas Restaurant	Yoga Studio	Food & Drink Shop
Latina	Lucero	-3.745269	3.0	Pizza Place	Park	Snack Place	Supermarket	Fast Food Restaurant	Grocery Store
Carabanchel	Opañel	-3.723178	3.0	Bar	Burger Joint	Gym / Fitness Center	Plaza	Bakery	Nightclub
Carabanchel	Vista Alegre	-3.740044	3.0	Grocery Store	Fast Food Restaurant	Convenience Store	Café	Breakfast Spot	Spanish Restaurant
Carabanchel	Puerta Bonita	-3.734559	3.0	Pharmacy	Bar	Grocery Store	Fast Food Restaurant	Plaza	Pub
Carabanchel	Abrantes	-3.727985	3.0	Pizza Place	Restaurant	Grocery Store	Fast Food Restaurant	Bakery	Metro Station
Puente de Vallecas	Entrevías	-3.674761	3.0	Gym / Fitness Center	Park	Turkish Restaurant	Farmers Market	Fast Food Restaurant	Fish & Chips Shop
Puente de Vallecas	Portazgo	-3.648126	3.0	Bar	Tapas Restaurant	Grocery Store	Café	Pizza Place	Fried Chicken Joint
Ciudad Lineal	Costillares	-3.668512	3.0	Gym / Fitness Center	Grocery Store	Bar	Italian Restaurant	Asian Restaurant	Pet Store
Hortaleza	Piovera	-3.635960	3.0	Sporting Goods Shop	Sports Club	Stadium	Cheese Shop	Gym / Fitness Center	Athletics & Sports
Villaverde	San Cristóbal	-3.687817	3.0	Breakfast Spot	Metro Station	Athletics & Sports	Train Station	Park	Flea Market
Villa de Vallecas	Casco Histórico de Vallecas	-3.621712	3.0	Grocery Store	Soccer Field	Food	Park	Church	Sandwich Place
Vicálvaro	Valderrivas	-3.599094	3.0	Metro Station	Platform	Falafel Restaurant	Playground	Grocery Store	Café
San Blas-Canillejas	Hellín	-3.616769	3.0	Music Venue	Grocery Store	Snack Place	Park	Metro Station	Gym

Debate

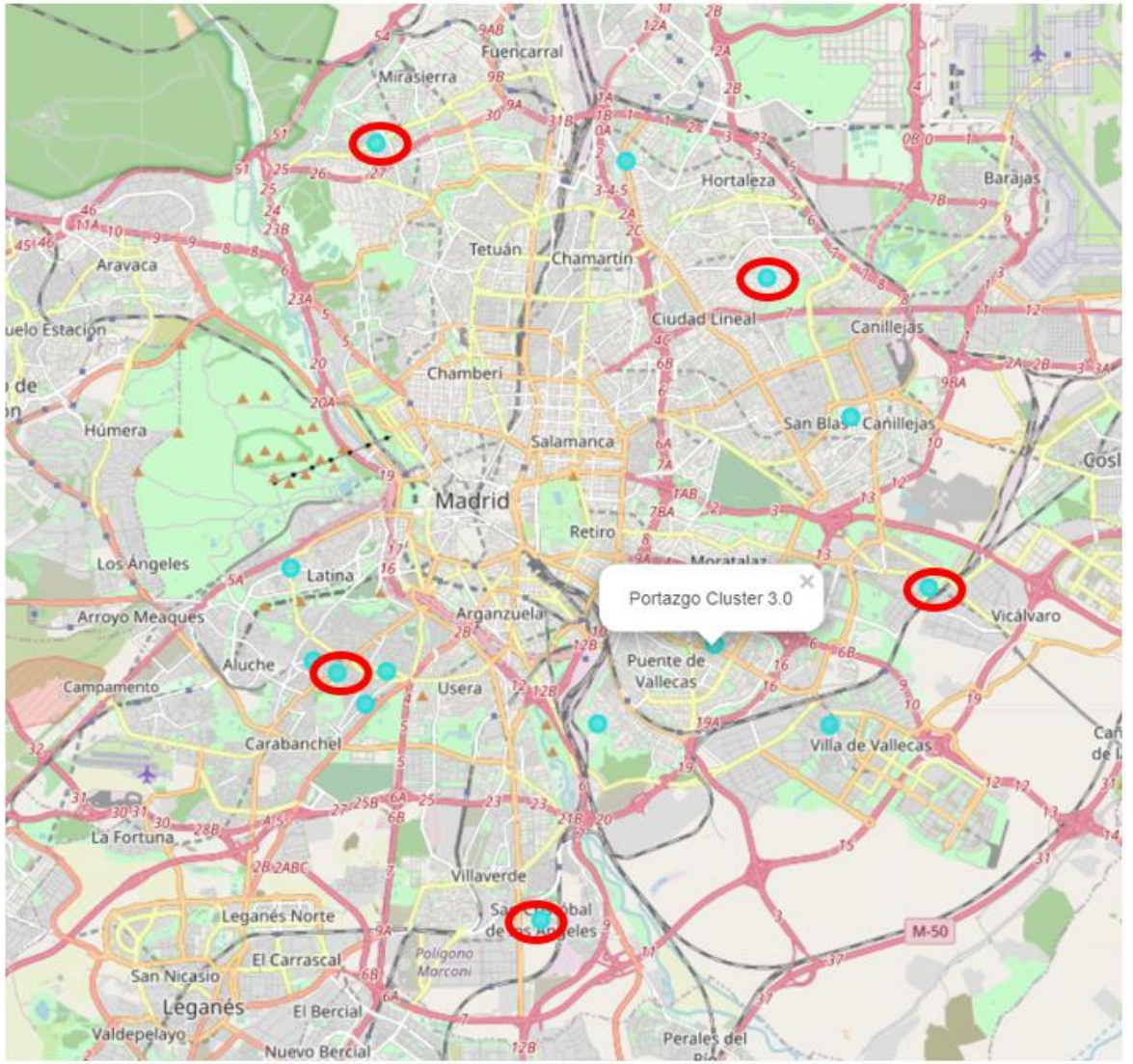
- Los resultados obtenidos nos indican que los barrios más similares al barrio objetivo se encuentran dispuestos de manera concéntrica respecto al centro de la ciudad de Madrid.
- Los barrios más similares al barrio objetivo se encuentran la periferia de la ciudad.
- Para conseguir un análisis más preciso se podría utilizar una base de datos con más entradas que la de Foursquare o dividir la ciudad en celdas más pequeñas que los barrios y del mismo tamaño.
- Los resultados que hemos obtenido parecen suficientemente precisos para elegir las ubicaciones más adecuadas.



Debate

Propuesta final de ubicaciones

Se ha optado por una disposición que permita reducir la distancia media al restaurante más cercano. Esta distribución evita los barrios adyacentes para cubrir mejor la ciudad.



Conclusiones

- Las fuentes de datos elegidas nos han permitido clasificar los barrios de Madrid en función del tipo de locales que hay en ellos.
- El algoritmo K-means es adecuado para el proceso de agrupación no supervisada con los datos que tenemos.
- Los resultados son coherentes con lo que se esperaba ya que el restaurante piloto se encuentra en un barrio periférico y las nuevas ubicaciones propuestas (barrios del mismo clúster) también son barrios periféricos.
- El nivel de detalle de los resultados es suficiente para elegir las ubicaciones de los nuevos restaurantes.