Prediciendo la mejor colonia para abrir una pizzería en Villahermosa, Tabasco

Emmanuel Rodriguez Zamora

21 /Mayo/2020

Introducción

Antecedentes

Actualmente en el mundo existe una enorme tendencia al emprendimiento, lo cual conlleva la creación de pequeñas empresas. Dentro del grupo de emprendedores siempre existen aquellos que desean abrir restaurantes con sus habilidades de cocina o recetas secretas de sus propias familias. Las pizzerías son sitios de comida en los que se pueden obtener muchas ganancias, esto debido a el alimento en sí mismo (considerado como el más adictivo del mundo), por lo que es una gran idea si se piensa en poner un restaurante. Sin embargo, existe una gran cantidad de cadenas multinacionales que es común que absorban la mayoría los clientes, por lo que es ideal estar alejado de estos lugares.

Problema

A través de los datos podemos elegir los lugares ideales para poner una pizzería en la ciudad, esto teniendo en cuenta la cercanía de las otras pizzerías y la densidad de población en el lugar. El fin del proyecto es identificar los mejores sitios para poner una pizzería en la ciudad de Villahermosa, Tabasco.

Interesados

Todas las personas interesadas en realizar un emprendimiento con un restaurante que su platillo principal sea la pizza.

Adquisición y limpieza de datos

Fuente de datos

Los datos de los nombres y códigos postales de las colonias de Villahermosa se obtuvieron de la página oficial del estado de Tabasco donde se enlista toda la información aquí. A partir de los código postales se obtuvieron la población y el área de cada uno de ellos en página de códigos postales nacional aquí. Después de tener la información básica de cada colonia se obtuvo la ubicación dada en latitud y longitud de cada una de las colonias a través de la librería geopy. Finalmente, la información concerniente a las pizzerías se obtuvo a través de la página Foursquare.

Limpieza de datos

Se realizaron una gran cantidad de acciones para lograr una limpieza de datos satisfactoria. Inicialmente los datos obtenidos a través de las APIs se encuentran en formato json, por lo que se tiene que filtrar la información deseada de cada petición, para crear las tablas deseadas.

Una vez obtenidos todo los datasets, se procedió a eliminar las colonias de las que no se logró obtener información útil como la ubicación o la cantidad de población. Además de ello, después de realizar diversas pruebas con la API de geopy se llegó a la conclusión de que el artículo “de” impedía a la librería diversas ciudades, por lo que se elimino de todas las colonias encontradas en el dataframe.

En las mediciones de área se eliminaron las letras representativas de la dimensión de los datos y se estandarizaron todos los datos ya que algunos se encontraban en metro mientras que otros estaban en kilómetros. Finalmente, los dos dataframe referentes a las colonias se unieron a partir de los códigos postales existentes.

Selección de características

A partir de los datos en el dataframe, debido a las variantes de área que posee cada código postal, se decidió unir al conjunto de datos una variable que relacionara el área y la población, por lo que se obtuvo la densidad de población de cada colonia. Para la realización del análisis es fue necesario seleccionar únicamente las características numéricas, debido a que no había variables categóricas en este dataframe. Además de ello, se realizó el cálculo de la cantidad de pizzerías cercanas a cada colonia en un radio de 700 metros para ayudar a escoger los lugares adecuados. Por lo tanto, se escogieron las siguientes características para pasar la siguiente fase:

- Posición en x de la colonia
- Posición en y de la colonia
- Población por código postal
- Área por código postal
- Densidad por código postal
- Cantidad de pizzerías 700 metros a las redondas
- Códigos postales

Análisis exploratorio de datos

A partir de los datos obtenidos se procedió a buscar las relaciones entre las características del dataframe para poder decidir cuales características son idóneas para usar en el algoritmo de aprendizaje. Inicialmente se utilizó el método propio de los dataframes para obtener las relaciones de toda la tabla. Como se puede observar en la figura 1, fue sumamente complicado encontrar grandes relaciones entre los datos. Las mejores relaciones fueron encontradas entre la población, área y la densidad como lo debe ser de manera natural.

	Código Postal	Latitude	Longitude	poblacion	area (km2)	densidad	posx	posy	pizzerias cerca
Código Postal	1.000000	-0.189758	-0.149556	-0.035159	0.298556	-0.059380	0.190343	-0.149016	-0.046484
Latitude	-0.189758	1.000000	0.384099	0.083244	0.125647	-0.114724	-0.999994	0.381566	0.002334
Longitude	-0.149556	0.384099	1.000000	0.101080	-0.032745	0.027660	-0.386768	0.999995	-0.007353
poblacion	-0.035159	0.083244	0.101080	1.000000	0.649322	-0.177112	-0.083397	0.100799	-0.003044
area (km2)	0.298556	0.125647	-0.032745	0.649322	1.000000	-0.303317	-0.125266	-0.033029	0.012103
densidad	-0.059380	-0.114724	0.027660	-0.177112	-0.303317	1.000000	0.114593	0.028013	0.002121
posx	0.190343	-0.999994	-0.386768	-0.083397	-0.125266	0.114593	1.000000	-0.384236	-0.002142
posy	-0.149016	0.381566	0.999995	0.100799	-0.033029	0.028013	-0.384236	1.000000	-0.007233
pizzerias cerca	-0.046484	0.002334	-0.007353	-0.003044	0.012103	0.002121	-0.002142	-0.007233	1.000000

Tabla 1. Relación de características

Fue difícil encontrar una relación proporcional entre los datos por lo que se procedió a analizar las características de manera individual con el fin de conocer cuáles podrían tener problemas de proporción en su mismas y por ello eliminarlas del algoritmo de aprendizaje para evitar anormalidades en el resultado del mismo.

La primera característica que se analizo fue la densidad, ya que esta fue una característica obtenida a partir de dos características originales del dataframe. Como se puede observar en la figura 2 el mayor porcentaje de la densidad se encuentra en un rango muy pequeño de valores, y existen una cantidad moderada de valores atípicos que pueden afectar el resultado de esos puntos en del dataframe en específico.

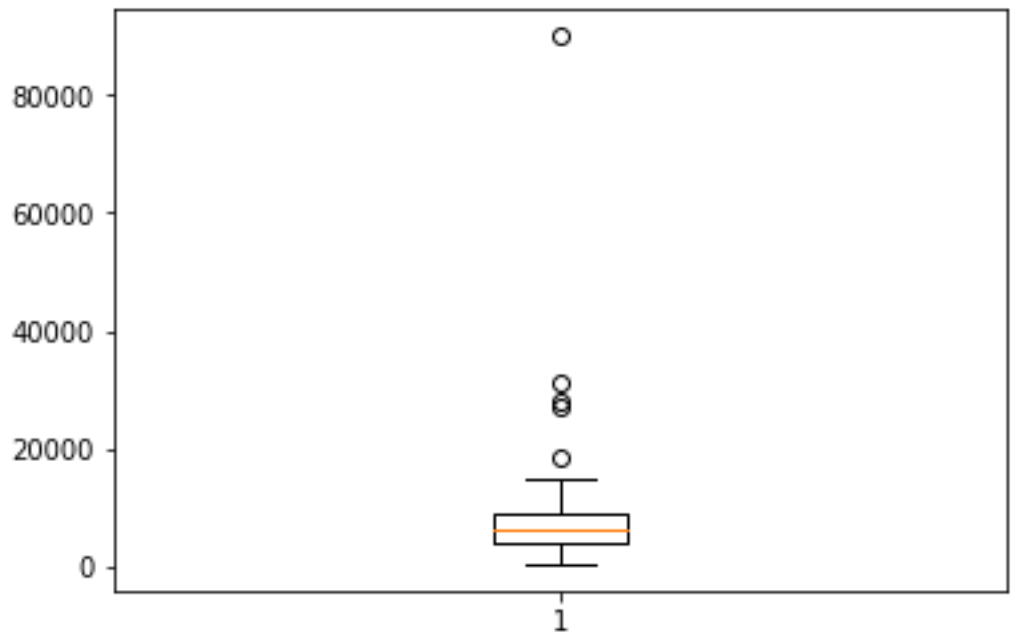

Figura 1. Diagrama de caja de la densidad.

Ya que las características mas importantes en este dataframe son la cantidad de pizzerías cercanas a cada colonia, la población, el área y la densidad. Se procedió a analizar el estado general de los datos en estos con el fin de encontrar las mejores características para el algoritmo de aprendizaje. En la figura 3 podemos observar el diagrama que se hizo con los datos del numero de habitantes en cada colonia.

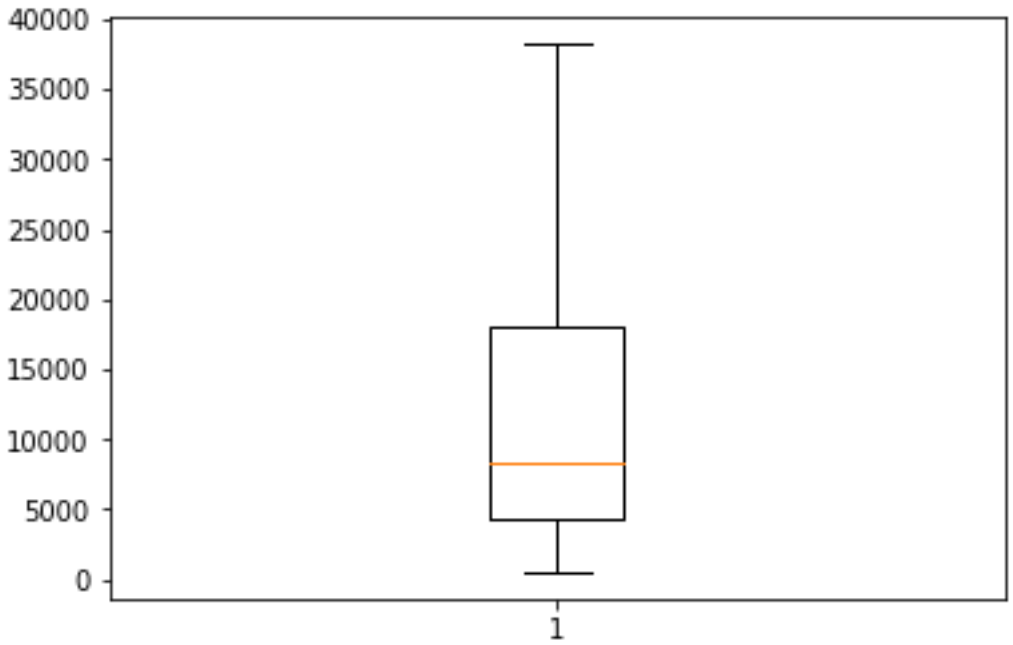

Figura 2. Diagrama de caja de la población.

Podemos observar que en el diagrama de la población existe una mejor distribución de los valores en el dataframe por lo que se hace una característica optima para utilizar en el algoritmo de aprendizaje.

Por ultimo se analizo el diagrama del área de cada colonia de acuerdo a su código postal. El resultado de esto es lo que se puede observar en la figura 4.

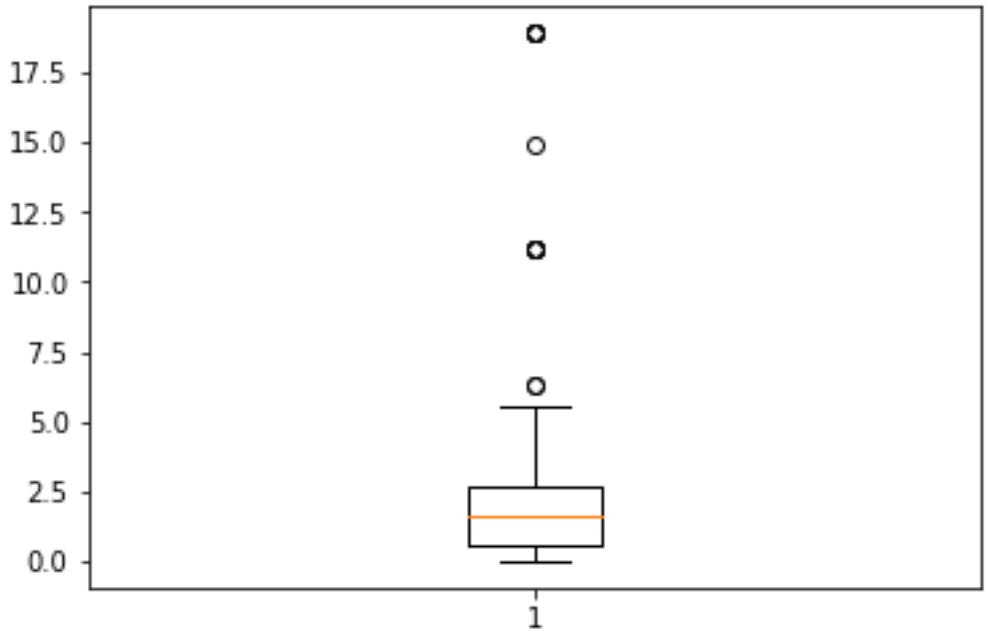

Figura 3. Diagrama de caja del área.

Después de este análisis de los datos principales del dataframe, decidí debido a la disparidad en la característica de la densidad, eliminarla del dataframe que será usado en el algoritmo de aprendizaje. No debería existir ninguna complicación debido a que es una unidad obtenida a partir de las dos que se quedaran en el mismo dataframe.

Uso de algoritmos de aprendizaje

Para obtener la separación de los diferentes grupos de colonias que se pueden encontrar en la ciudad y poder analizar cuál es la mejor área para colocar una pizzería fue necesario aplicar algoritmos de aprendizaje no supervisado. En esta ocasión se utilizó el algoritmo Kmeans. Se decidió usar este algoritmo para poder analizar al final del entrenamiento los diferentes grupos formados, a partir de la agrupación de los clústeres obtenidos y sus promedios.

Algoritmo de aprendizaje k-means

Se procedió introducir el dataframe con las características seleccionadas al algoritmo de kmeans para obtener las etiquetar correspondientes a cada colonia y poder separar estas de acuerdo a sus características.

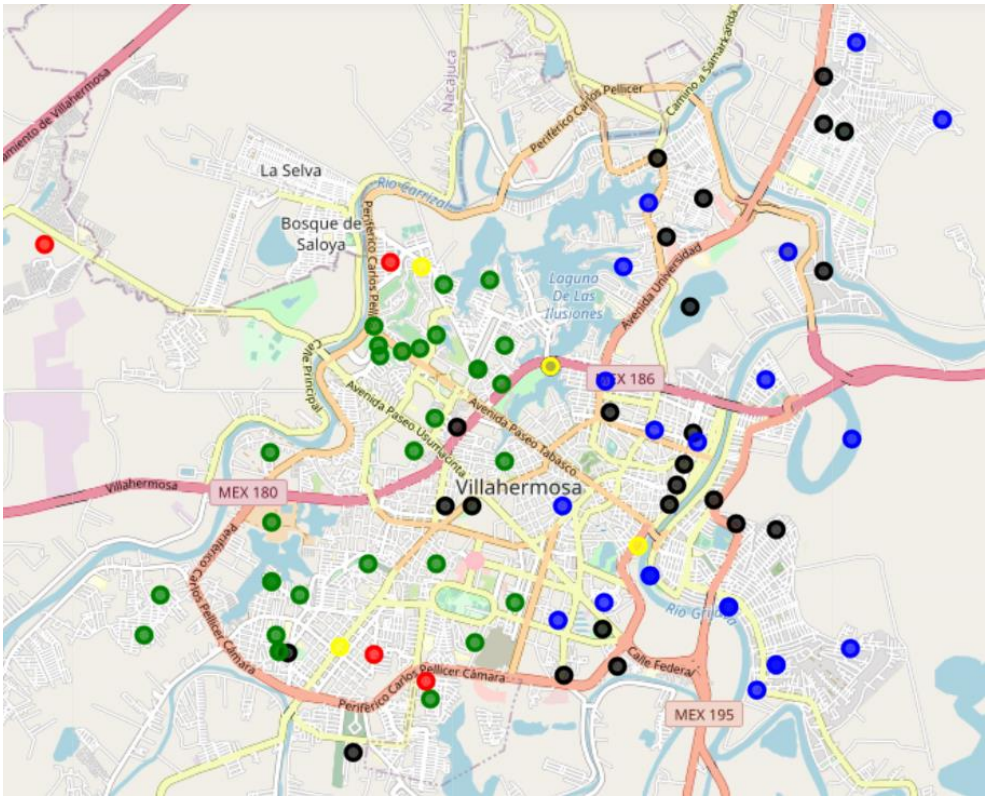

Figura 4. Mapa con colonias de Villahermosa etiquetadas.

Después de usar el algoritmo y decidir utilizar 6 etiquetas de clase para las diferentes colonias de Villahermosa, se realizó el análisis de los resultados a partir de las medias en cada una de las características de dataframe con fin de observar el comportamiento único de cada una de las clases que se obtuvieron.

El resultado antes mencionado se puede observar en la tabla 2 mostrada a continuación:

color	Código Postal	Latitude	Longitude	poblacion	area (km2)	densidad	posx	posy	pizzerias cerca	cluster
black	86067.880000	17.995670	-92.922932	20920.040000	3.195960	7732.158800	1.216706e+07	1.002366e+07	0.480000	2.0
blue	86088.615385	17.993217	-92.917396	4507.038462	1.209984	11591.521538	1.216791e+07	1.002551e+07	0.115385	0.0
green	86086.933333	17.989262	-92.952797	6528.933333	1.200307	7278.471333	1.216932e+07	1.001371e+07	0.366667	5.0
pink	86110.000000	17.980348	-92.941741	8985.000000	5.974000	5432.750000	1.217242e+07	1.001741e+07	6.000000	4.0
red	86136.800000	17.983879	-92.982892	9370.600000	10.128400	1479.146000	1.217120e+07	1.000366e+07	0.000000	1.0
yellow	86127.000000	17.996365	-92.933389	38224.000000	18.900000	2022.430000	1.216683e+07	1.002017e+07	0.000000	3.0

Tabla 2. Mapa con colonias de Villahermosa etiquetadas.

Resultados y discusión

Los resultados obtenidos nos indican que las mejores colonias para poner una pizzería en la ciudad de Villahermosa, Tabasco, son aquellas que se encuentran etiquetadas con el color amarillo y rojo en la figura 4. Esto debido a que posee dos características que las hacen ideales:

- No tienen ningunas pizzerías 700 metros a la redonda
- Tiene una población medianamente grande que puede traducirse en clientes potenciales.

Como se pudo notar, comúnmente los mejores lugares se encuentran casi en los alrededores de la ciudad, ya que en el centro existen una gran cantidad de lugares de este tipo que puede influir en las ganancias. Para ser mas precisos, los mejores lugares para poner una pizzería, se encuentran en el sur de la ciudad y al noroeste de la misma, en colonias como: Cumbres, Miguel Hidalgo, Real del Sur, entre otras.

Es importante tener en cuenta que, en este proyecto, existió un poco de deficiencia en los datos de la ciudad principalmente en la obtención de ubicación de absolutamente todas las colonias, por lo que el análisis se realizó con la mayoría de las colonias mas conocidas de la ciudad, dejando fuera pequeñas lugares que pudieran haberse analizado.

Conclusión

En este estudio analice todos las colonias y pizzerías encontradas en la ciudad de Villahermosa, Tabasco. Utilice los mejores datos para introducir al algoritmo de aprendizaje para obtener los diferentes tipos de colonia de Villahermosa y a partir de ello, elegir las mejores colonias para abrir una pizzería y obtener el mayor beneficio de acuerdo su posición en la ciudad. Este estudio puede servir para cualquier persona que piense poner un negocio de comida en la ciudad y quería evitar ser ignorado debido a las grandes cadenas de restaurante que existen en esta ciudad.