UNIVERSIDAD DEL NORTE

MINERIA DE DATOS

QUIZ No. 4

Cesar Pájaro Miranda

Abril 25 de 2021
Versión Final

UNIVERSIDAD DEL NORTE

CONTENIDO

1. CASO 1 3
2. CASO 2 3

TABLA DE FIGURAS

Figura 7. Ubicación de las coordenadas enviadas por los estudiantes. 4

TABLA DE TABLAS

Tabla 3. Coordenadas estudiantes. 3

1. CASO 1

El caso 1 del Quiz tiene como base de datos 1000 reseñas de restaurantes realizadas en la aplicación ***yelp*** la base de datos se encuentra compuesta por 500 reseñas positivas y 500 reseñas negativas, acompañadas por el texto asociado a cada reseña. La resolución de este caso se realizará en la libreta de Google Colab alojada en el repositorio público de GitHub del Link https://github.com/capajaro/Quiz--4/blob/main/Case1_Quiz_4.ipynb.

Las suposiciones realizadas para la solución de este caso son las siguientes:

A. Es posible generar un modelo que permita la predicción con un grado de precisión adecuado del sentimiento generado por la reseña (positivo o negativo) a partir del texto de esta.
B. Para el modelo a realizar los predictores corresponderán a las palabras del texto de la reseña y la variable será una variable binaria 0/1 representando si la reseña es negativa o positiva respectivamente.
C. Se hará uso de Python para apoyar la visualización de los datos y su limpieza.
D. Se hará uso de R para la generación del modelo, en este caso se probarán 3 alternativas modelo, las cuales serán ***random forest, lasso y SVM***.
E. Se recomendará el mejor modelo con base en los resultados obtenidos de precisión para los 3 modelos analizados.

Con el fin de presentar todos los resultados en un solo código, se presenta en el cuaderno de Google Colab los resultados de la modelación en Python y R. Además, en el código se presentan los comentarios necesarios para la compresión del código y la resolución de los puntos asignados por el docente.

2. CASO 2

El caso 2 del presente Quiz correspondía en hacer un análisis de clustering con base en las ubicaciones reportadas por los estudiantes de la asignatura. Para ello se realizo un archivo de Google Docs. en el cual los estudiantes rellenaban sus coordenadas geográficas en el sistema WGS-84, es decir, Lat/Lon. En la Tabla 1 se presentan las coordenadas reportadas por los estudiantes. Mientras que en la Figura 1 se presentan la ubicación geográfica de los estudiantes representados con un marcador naranja.

Tabla 1. Coordenadas estudiantes.

Tag	Nombre	Lat	Lon
1	Cesar Pajaro	11.007	-74.834
2	Danny Daniel Ortega	10.982	-74.809
3	Efrain Boom carcamo	10.992	-74.845
4	Jose Navarro De la Cruz	10.788	-74.755

Tag	Nombre	Lat	Lon
5	Joseph Soto	11.019	-74.815
6	Salvador Villamizar	10.983	-74.804
7	Maria Isabel Arrieta Escobar	11.026	-74.865
8	Nazhir Amaya	10.902	-74.787
9	Jhoan Castro	11.006	-74.835
10	Loraynne Amaya	10.977	-74.809
11	Valentina Mejia	10.963	-74.795
12	Ethel Garcia	11.017	-74.818
13	Kevin Palomino	10.993	-74.847
14	Carlos Ferreira	10.915	-74.801
15	Fernando Gonzalez	10.996	-74.822
16	Ana Luisa Cuello	10.997	-74.802

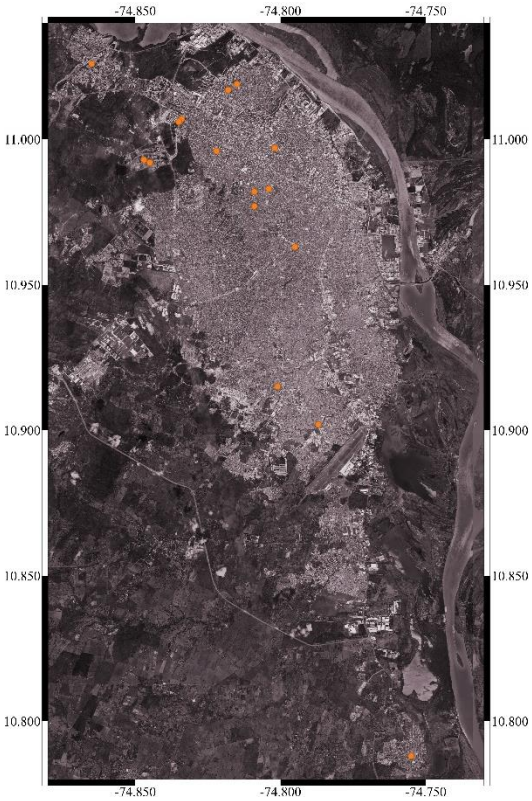

Figura 1. Ubicación de las coordenadas enviadas por los estudiantes.

Lo primero que debe realizarse para hacer el análisis de clusters, es el cálculo de las distancias entre los puntos, esto se realizado haciendo uso del software “R”. sin embargo, en este caso por requerimiento del docente no se utilizará la distancia euclidiana sino la distancia de “manhattan”.

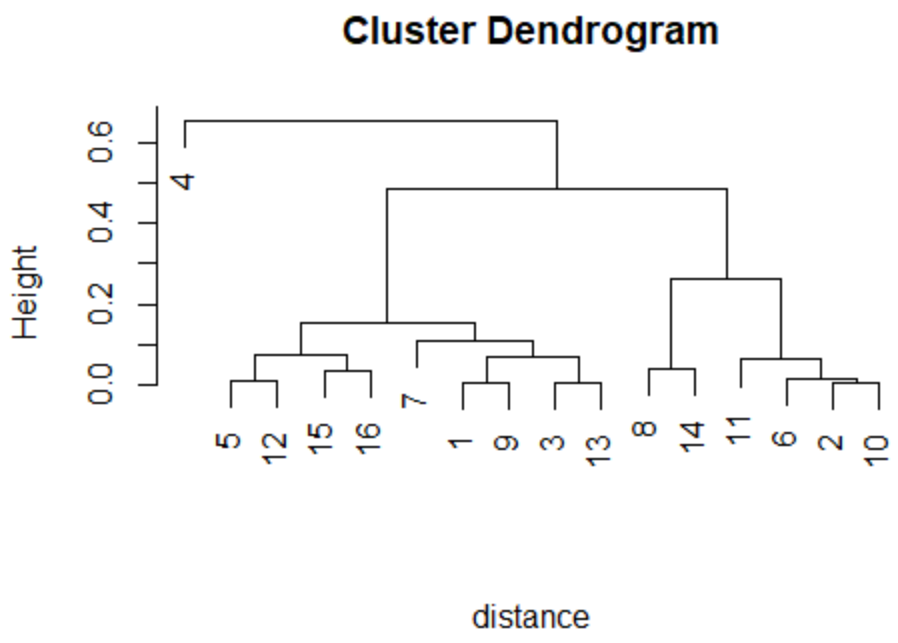

Figura 2. Dendograma de representación del problema.

Teniendo en cuenta el dendograma presentado en la Figura 2 se seleccionaran 5 cluster par agrupar los datos ya que permiten una separación adecuada de los datos. Lo anterior es presentado en la

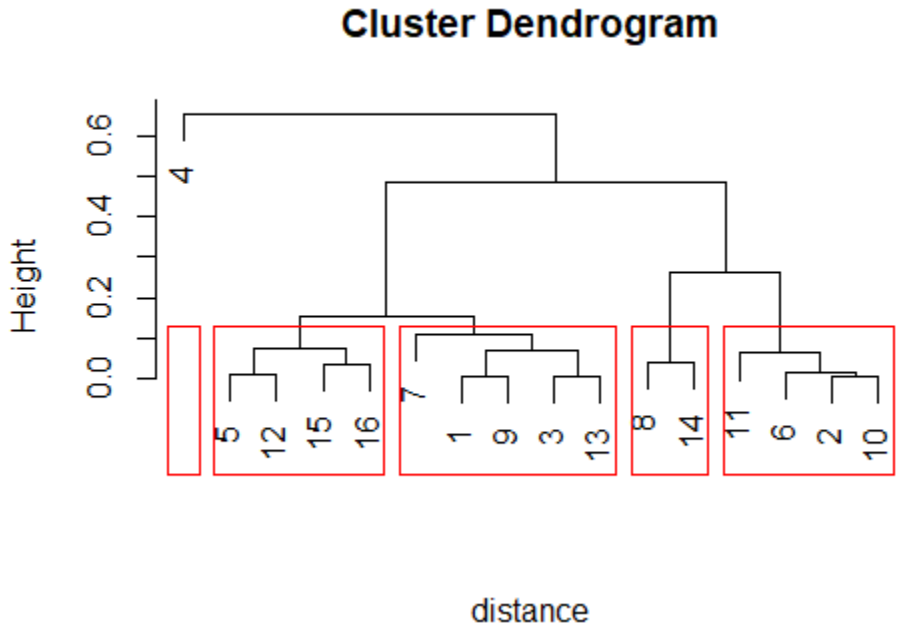

Figura 3. Clusters seleccionados para la agrupación de la localización de los estudiantes.

Una vez seleccionado el número de clúster se utilizó un análisis del tipo k-Means Clustering con 1000 realizaciones en el software "R". Los resultados de este análisis son presentados en la Tabla 2 y en la Figura 4. En la tabla se presenta el cluster al que se ha asignado cada estudiante, las coordenadas del centro del cluster y la distancia de la ubicación de los estudiantes hasta el centro

de su respectivo cluster. Como podemos notar los estudiantes José Navarro De la Cruz y María Isabel Arrieta Escobar tienen una distancia de "0" al centro de cada cluster lo que significa que el cluster al que pertenecen está compuesto por ellos exclusivamente, esto debido a que su ubicación es bastante distante respecto a los demás estudiantes. Lo anterior puede corroborarse en la Figura 4 donde los centros de cada cluster se encuentran representados por un triangulo invertido. El cluster con mayor número de estudiantes es el número 1.

Tabla 2. Resultados análisis k-Means Clustering para la localización de los estudiantes.

Tag	Nombre	Lat.	Lon	Cluster	Lat_Center	Long_Center	DistCenter
1	Cesar Pajaro	11.007	-74.834	1	11.004	-74.831	19.38
2	Danny Daniel Ortega	10.982	-74.809	4	10.980	-74.804	8.42
3	Efrain Boom carcamo	10.992	-74.845	1	11.004	-74.831	76.90
4	Jose Navarro De la Cruz	10.788	-74.755	2	10.788	-74.755	0.00
5	Joseph Soto	11.019	-74.815	1	11.004	-74.831	94.34
6	Salvador Villamizar	10.983	-74.804	4	10.980	-74.804	15.58
7	Maria Isabel Arrieta Escobar	11.026	-74.865	3	11.026	-74.865	0.00
8	Nazhir Amaya	10.902	-74.787	5	10.909	-74.794	41.57
9	Jhoan Castro	11.006	-74.835	1	11.004	-74.831	8.38
10	Loraynne Amaya	10.977	-74.809	4	10.980	-74.804	19.83
11	Valentina Mejia	10.963	-74.795	4	10.980	-74.804	108.60
12	Ethel Garcia	11.017	-74.818	1	11.004	-74.831	78.49
13	Kevin Palomino	10.993	-74.847	1	11.004	-74.831	69.06
14	Carlos Ferreira	10.915	-74.801	5	10.909	-74.794	41.57
15	Fernando Gonzalez	10.996	-74.822	1	11.004	-74.831	54.64
16	Ana Luisa Cuello	10.997	-74.802	4	10.980	-74.804	104.43

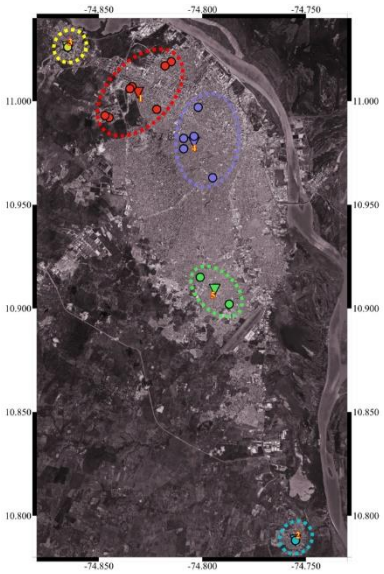

Figura 4. Resultados agrupación de los estudiantes en 5 grupos(clusters).