Herramientas computacionales: el arte de la analítica (Gpo 570)

Actividad Evaluable: Patrones con K-means

Diego Andrés Moreno Molina

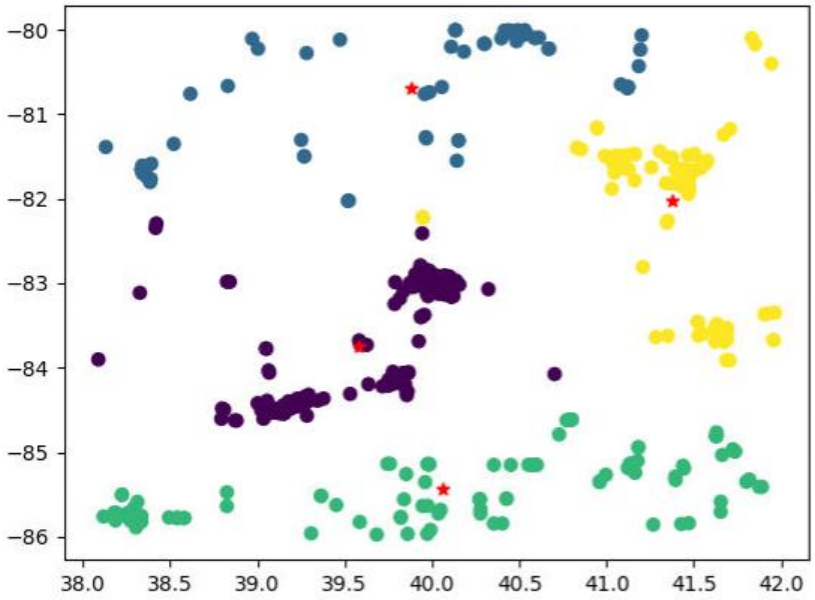
A01283790

Introducción

Para esta etapa nos enfocaremos en encontrar patrones con K-Means, los cuales son útiles para encontrar grupos de datos que están relacionados de manera no explicita. En nuestro caso estamos analizando accidentes de carro dentro de una ciudad, y unas de las variables que se nos proporcionan son las coordenadas del choque. apoyándonos de este algoritmo podemos buscar patrones de grupos para encontrar que zonas de choque son las más comunes, lo cual podría ser útil para reducir la cantidad de choques en esas respectivas zonas.

Funcionamiento del algoritmo

La manera que este algoritmo funciona es buscar una cantidad "X" de centroides dentro de la dispersión de los datos. Estos centroides intentaran estar tan cerca de lo posible en promedio de diferentes grupos, donde cada centroide puede representar un grupo de datos.



Ahora bien, observando los resultados, podemos observar como se identifican 4 grupos principales, los cuales representan una zona geográfica dentro de la ciudad. Esta estadística es útil para saber cuales zonas tienen mayores tendencias de choques, y se podría investigar el porqué de esta razón para prevenir futuros choques. Cabe mencionar que se eligió una cantidad arbitraria de 4 centroides, ya que observando los datos gráficamente se pueden observar 4 grupos generales claramente.

Es importante entender que, si tenemos mayor cantidad de centroides, estos serían mas precisos y por lo tanto mas representativos de su zona. Sin embargo, ya que estamos hablando de zonas geográficas, no es tan útil dividir las zonas en grupos tan pequeños (ya que nos darían grupos diferentes para mismas calles). Por lo que podemos concluir que, aunque una mayor cantidad de K si nos representaría mejor los diferentes grupos, su interpretación no sería tan útil para este caso.

La distancia entre los centroides es en promedio cerca a 1.25 en la gráfica, donde los 4 centroides tienen una distancia similar entre ellas (lo cual nos indica que el algoritmo tuvo un resultado bueno).

Adicionalmente es importante recalcar el impacto que tienen los outliers en este algoritmo. Estos tienen un efecto negativo en el calculo de los centroides, ya que estos se basan en promedios de distancias entre los puntos, los “outliers” afectarían el cálculo de estas medidas. Por esto mismo es importante limpiar la información si queremos un centroide que sea más representativo de su grupo

Conclusión

Basado en los resultados de nuestra simulación, se pueden encontrar como están divididas las 4 zonas de choques principales, por lo que podríamos llevar estos datos al mundo real para investigar esas zonas geográficas e implementar medidas preventivas o una investigación en búsqueda de reducción de estos choques. El algoritmo KMeans fue una herramienta útil para encontrar estos grupos representativos, y fue importante jugar con los valores que recibe este algoritmo para encontrar un gráfico que nos permita hacer la interpretación más acertada.

Referencias

Gonzalez, L. (2021, September 08). Algoritmo KMeans - Teoría. Retrieved from https://aprendeia.com/algoritmo-kmeans-clustering-machine-learning/