Modelo de ML para el agrupamiento(clustering) de oficinas

Introducción

El siguiente algoritmo de machine learning (ML) tiene por objetivo realizar el agrupamiento de oficinas que prestan servicios A y C, basado en variables como el tiempo promedio de atención, el tiempo de espera, el volumen de casos atendidos y casos posiblemente en espera.

Luego de aplicar un modelo de clustering, K-Means, por su simplicidad y ajuste al tipo de datos: tamaño de la muestra, número de variables y densidades separables. Se lograron obtener tres clústeres que permiten agrupar las oficinas con base en las variables en mención.

Resultado de ello, se logra diferenciar que, hay por lo menos dos clústeres diferenciales a partir del comportamiento de las variables que son compatibles con lo esperado.

Proceso de clusterización

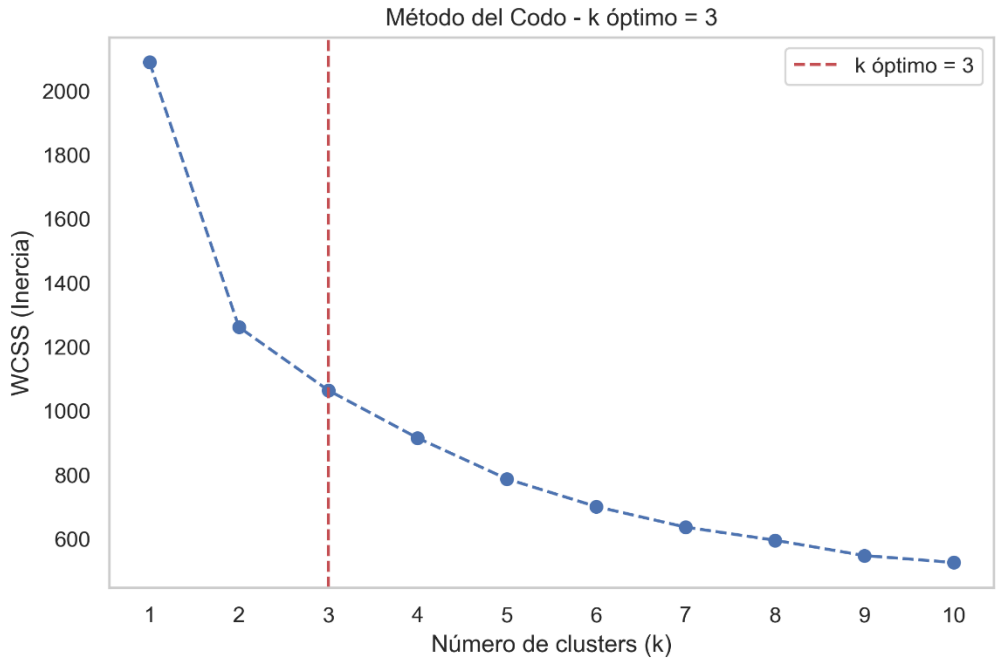
En total se tienen 2 tipos de servicios: A y C. Se tienen 7 conjuntos de datos que representan variables como recuento de demanda, tiempos de espera, tiempos de atención discriminados por servicio. Se encontraron 522 oficinas que prestan dichos servicios. También se tiene un registro de tres meses consecutivos a inicios y final de año.

Dado que el problema se podría abordar mezclando los servicios (A y C), se optó por abordar el problema por separado, es decir: realizar el proceso de clústering por cada servicio.

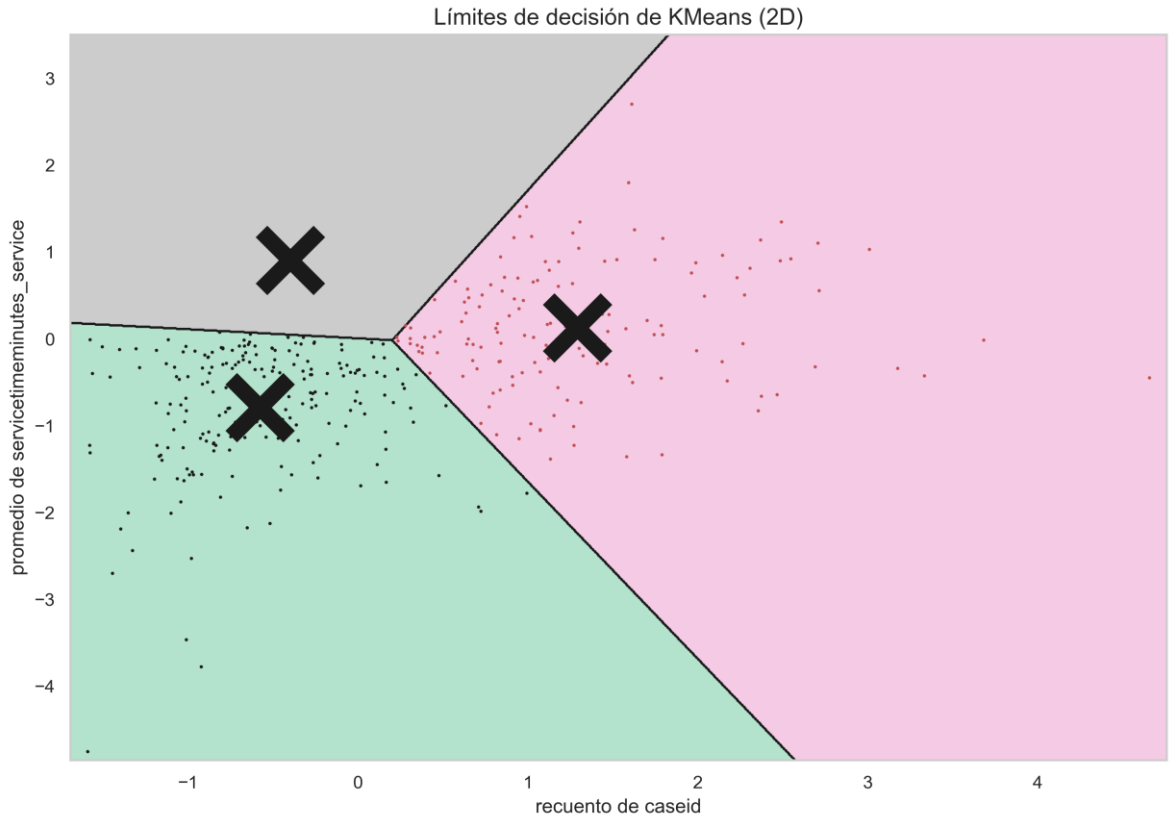
Una vez analizados los datos, se observa que, las variables numéricas para realizar el proceso de clústering son 4 (recuentos de demanda, de espera, tiempos de atención y de espera.) Como el volumen de variables es es pequeño (4 variables), se pudieron apreciar separabilidad de datos por cada variable. El algoritmo seleccionado por su sencillez, simplicidad y control de hiperparámetros es K-Means. Otros modelos hubiesen sido útiles, como DBSCAN o Cluster Jerárquico, pero el comportamiento de los datos no mostraba muchos outliers.

Ejecución del modelo.

Para la ejecución del modelo se realizó un proceso de agrupamiento por oficina a partir del promedio de las variables. Seguido de ello se realizó un proceso de escalado de los datos. Una vez obtenida la gráfica sobre el método del codo, para hallar el número de clúster necesarios para ejecutar el algoritmo, se encontró que k = 3 y se obtuvo el modelo de clúster por oficina.



Observando los tres clústeres formados para el caso del servicio A. Aquí se logra apreciar la separabilidad que hay en el conjunto de datos a partir de las variables de demanda y tiempos promedios de servicio.



De igual manera, a nivel de clúster, para el caso de la variable de demanda, se logran apreciar que los clústeres tienen coherencia, puesto que se aprecia separabilidad entre las oficinas que atienden dicha demanda. Es decir, las oficinas del clúster 2 tienden a concentrar más demanda que las del clúster 0 o 1, por mencionar solo un caso.

Luego, el set de datos queda conformado por las siguientes variables y el clúster asignado:

	unitnam	recuento de caseid	promedio de servicetimeminutes_service	promedio de waittimeminutes	recuento de countcaseid	cluster
0	(001) C	2519.000000	19.415353	17.123237	338.333333	0
1	(002) C	2916.000000	21.148990	16.860044	324.333333	0
2	(004) L	1895.000000	20.419270	14.399863	264.666667	2
3	(005) A	1235.333333	25.608706	11.294973	235.000000	2
4	(006) A	1860.000000	27.102240	21.335366	311.333333	2
...	...	...	...	...	...	...
517	(967) B	90.000000	20.129647	12.234889	66.000000	1
518	(972) C	1932.333333	18.072657	9.042128	354.666667	2
519	(991) V	771.000000	15.370869	7.572701	245.000000	1
520	(993) U	771.000000	14.980300	14.657404	229.333333	1
521	(996) F	1027.666667	20.855942	6.358991	266.000000	2