****

Instituto Tecnológico de Estudios Superiores Monterrey

CAMPUS QUERÉTARO

Herramientas computacionales: el arte de la analítica

Grupo 120

**Profesor:**

Pedro Oscar Pérez Murueta

Actividad

**Patrones con K-means**

**PRESENTA:**

Equipo 4

Jordana Betancourt Menchaca - A01707434

Leonardo Santiago Ramos Pérez A01707812

Stefano Herrejón - A00571720

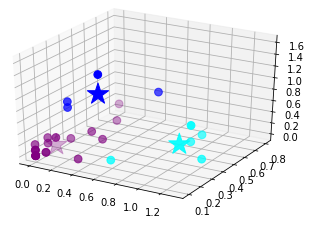
Fecha:

28 de octubre de 2021

La hipótesis planteada fue que el procesamiento, transporte y empaquetado de alimentos contaminan lo mismo, esto debido a que nos dimos cuenta que los valores en esas columnas eran demasiado parecidos unos con los otros, por lo cual decidimos comprobar si era cierto.

* **¿Crees que estos centros puedan ser representativos de los datos? ¿Por qué?**

Si, debido a que se observa una buena división de grupos, donde no quedan encimados los centros o los valores muy alejados a éste.



* **¿Cómo obtuviste el valor de k a usar?**

A partir de la creación de una curva de codo y sus agrupaciones que iban generando de acuerdo a cuanto más suave era la curva en ese momento.

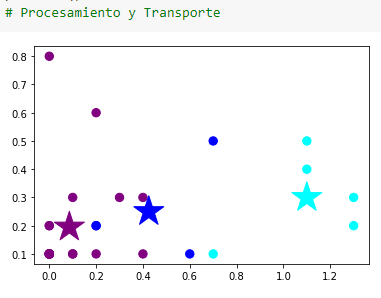
Por lo cual se optó con un K de 3 clusters.

* **¿Los centros serían más representativos si usaras un valor más alto? ¿Más bajo?**

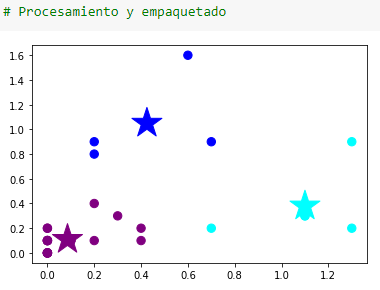
Al tener menos centros se pierden muchos datos, dejando los más básicos o un punto en el que ya ni se puedan distinguir correctamente,quitándoles representatividad ;por el otro lado tener más valores de centro toma en cuenta todos los datos, pero resultaría imposible encontrar una relación buena, ya que, más centros implica más separaciones de datos, puede resultar más representativo, pero la finalidad es encontrar aquellos que sí hagan sentido y no de más.

Particularmente en nuestro caso, al colocar dos clusters resultaba que entre ellos estaban muy separados y poner cuatro centroides hacía que quedaran unos encima de otros.

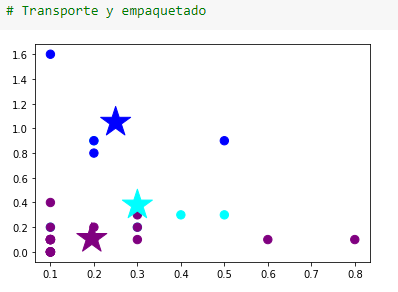
* **¿Qué distancia tienen los centros entre sí? ¿Hay alguno que esté muy cercano a otros?**

****

En procesamiento y transporte podemos observar que efectivamente tenemos dos centros que se encuentran cercanos.

****

En procesamiento y empaquetamiento los centros se encuentran distantes.

****

En procesamiento y empaquetamiento es la relación de centros que se encuentran más juntos, por lo cual podrían ser casi idénticos los valores.

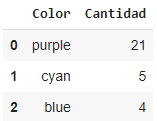
.

* **¿Qué pasaría con los centros si tuviéramos muchos outliers en el análisis de cajas y bigotes?**

Los centros más separados no serían tomados en cuenta, y los valores más alejados a ellos tampoco quedarían en el gráfico de caja y bigotes.

* **¿Qué puedes decir de los datos basándose en los centros?**

Que las tres columnas empleadas no tienen relación, pero que el transporte y empaquetado si pueden llegar a tener los mismos valores, ya que sus centros están más juntos, por tanto podrían agruparse en uno solo.



De hecho un 70% de los valores tienden a morado, por lo cual se podría decir que casi todos los valores de Procesamiento,Empaquetado y Transporte son los mismos.

Pero la hipótesis se descarta porque no todos son iguales.