Clusterizacion de casos positivos de COVID19 usando K-means

Resumen

El trabajo tiene como objetivo obtener clusters de personas que han dado positivo en la prueba de COVID-19, en base a criterios geográficos , demográficos y tipo de prueba aplicada , para lo cual se usó el algoritmo de K - means obteniéndose un total de 7 grupos, esta serie de grupos han sido agrupados ya que dichas personas comparten ciertas características en común sobre los cuales se puedan implementar medidas restrictivas del gobierno para mitigar el avance del COVID-19 en el país.

Descripción del dataset:

El dataset original está compuesto por un total de 1035184 registros y 9 column as, entre las cuales se encuentran: Fecha de corte,UUIID,Departamento, Provin cia, Distrito, Tipo de prueba, Sexo, Edad y Fecha de resultado de la prueba.

Preprocesamiento:

A) Eliminar variables que no son relevantes

Se eliminó las variables de formato de fecha , así como identificador dad o que no nos daban mayor información al momento de realizar el modelo

B) Variables numéricas

1. Se encontró un total de 58 registros que contaban con un valor nulo en e l campo Edad, por lo cual se procedió a eliminarlos.

2. Además para el campo edad se realizó un diagrama de cajas a modo de identificar los valores atípicos, para su posterior eliminación.

C) Variables categóricas

1. En el caso de la variable Sexo y Tipo de prueba ambas al ser de tipo cat egórica no ordinal, se realizó una conversión a variables dummy, para qu e puedan ser procesadas por el algoritmo kmeans.
2. Para las variables de tipo geográfico como Departamento, Provincia y Di strito se usó el ratio de aparición de cada uno estos , reemplazandolos e n el dataset los valores categóricos de dichas variables, esto se realizó y a que el número de departamentos, provincias y distritos es muy grande , lo cual al haberse aplicado One Hot encoding la dimensión del dataset or iginal hubiese crecido demasiado, lo cual dificultaría el procesamiento.

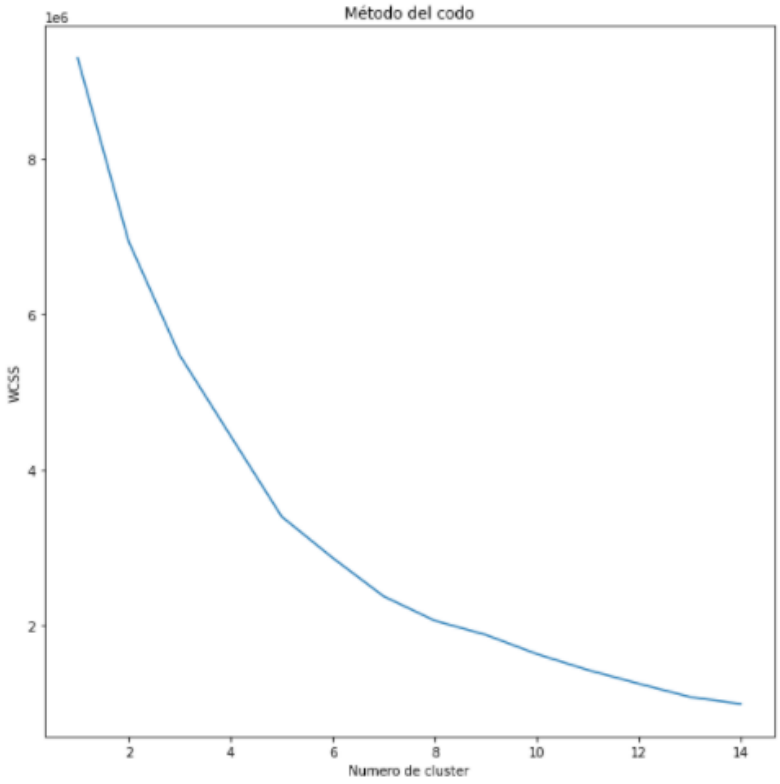
- **Dataset original:**

https://www.datosabiertos.gob.pe/dataset/casos-positivos-por-covid-19-ministerio-de-salud-minsa

- **Dataset limpio:**

Descripción de los resultados obtenidos:

Se usó el algoritmo de k-means y para la selección del número de cluster se usó el método del codo , tal como se muestra en la siguiente figura:



Se seleccionó un total de 7 clusters, además se encontró como reporte el número de instancias por clase fue el siguiente:

Cluster	Numero_observaciones
2	264232
0	252945
4	157957
3	136161
1	130988
5	88880
6	1534

Conclusiones:

1.En el cluster 2 se encontró que de los 10 departamentos que más aportan a dicho cluster en términos de cantidad; estos se encuentran actualmente categorizados por el gobierno como zonas de alto contagio o muy alto contagio, por lo que se puede mejorar aún la segmentación en este cluster a manera de identificar a que departamentos pertenezcan a las zonas dichas por el gobierno.

2.En el cluster 4 se puede notar que el departamento con mayor número de ocurrencias en dicho cluster es Lima , además que los sectores de las zonas 1,2,3 ,y 4 de Lima son las que mayor aporte dan a este cluster, donde se encuentran distritos como: San Juan de Lurigancho,San Martin de Porres,Los Olivos y Comas principalmente.

3. El uso de modelos que realicen clustering pueden ser de ayuda al momento de implementar acciones restictivas a ciertas zonas del pa ís , como puede ser el caso de las restricciones que están implemen tando el país actualmente en ciertas regiones del país, donde se ha

segmentado en 3 grandes grupos: MODERADO,ALTO y MUY ALT O, para mitigar el contagio del COVID-19.