UANL
UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FCFM
FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

PRACTICA 4: Agrupamiento.

APRENDIZAJE AUTOMATICO

CÉSAR ALEJANDRO HERNÁNDEZ OROZCO

Matricula: 1990010

GRUPO 003

MAESTRO: JOSÉ ANASTASIO HERNÁNDEZ SALDAÑA

INTRODUCCIÓN:

En esta tarea, se centra en la implementación y evaluación de un modelo de agrupamiento utilizando el algoritmo k-means. El objetivo principal es identificar el número óptimo de clusters en el conjunto de datos y comparar el rendimiento del modelo de agrupamiento con el modelo de clasificación previamente evaluado.

Este proyecto se realizo usando el conjunto de datos proporcinado por varios hospitales de distintas partes del mundo. Este es usado para tratar de encontrar si existe alguna enfermedad del corazon en los pacientes de estos hospitales.

En este conjunto de datos podemos encontrar bastante información de los distintos pacientes como lo son su edad, su sexo, si han tenido problemas cardiacos y como variable de respuesta se tiene si cuentan con un diagnostico de enfermedad del corazon.

En esta práctica se hizo uso del agrupamiento. El agrupamiento es una técnica no supervisada que busca identificar grupos o clústeres en un conjunto de datos sin tener etiquetas previas. Este análisis es útil para descubrir patrones y estructuras ocultas en los datos que pueden no ser evidentes a partir de la clasificación supervisada.

DESARROLLO:

Lo primero que hicimos para realizar esta practica fue empezar a conocer un poco mas como esta estructurado este conjunto de datos:

	name	role	type	demographic	description	units	missing_values
0	age	Feature	Integer	Age	None	years	no
1	sex	Feature	Categorical	Sex	None	None	no
2	cp	Feature	Categorical	None	None	None	no
3	trestbps	Feature	Integer	None	resting blood pressure (on admission to the ho...	mm Hg	no
4	chol	Feature	Integer	None	serum cholestoral	mg/dl	no
5	fbs	Feature	Categorical	None	fasting blood sugar > 120 mg/dl	None	no
6	restecg	Feature	Categorical	None	None	None	no
7	thalach	Feature	Integer	None	maximum heart rate achieved	None	no
8	exang	Feature	Categorical	None	exercise induced angina	None	no
9	oldpeak	Feature	Integer	None	ST depression induced by exercise relative to ...	None	no
10	slope	Feature	Categorical	None	None	None	no
11	ca	Feature	Integer	None	number of major vessels (0-3) colored by flour...	None	yes
12	thal	Feature	Categorical	None	None	None	yes
13	num	Target	Integer	None	diagnosis of heart disease	None	no

Podemos ver que este conjunto de datos esta bastante bien estructurado y nos presenta incluso como podemos trabajar con estos datos.

En esta practica se hizo uso de el algoritmo K-Means: Este es un algoritmo de agrupamiento que particiona el conjunto de datos en k clusters basándose en la minimización de la variación intra-cluster. K-means asigna cada punto de datos al cluster cuyo centroide está más cercano, y luego actualiza los centroides para minimizar la distancia total entre los puntos de datos y sus centroides.

La variable con la trabajaremos principalmente es “num”. Esta variable toma el valor 1 si se diagnostica a esta persona con una enfermedad del corazon y y 0 si esta no tiene una enfermedad del corazon diagnosticada.

Las demas variables con las cuales trabajaremos para obtener un buen agrupamiento son las siguientes:

Las variables con las que trabajaremos seran sex, cp, fbs y exang.

Sex =Sex (1=male,0=female)

cp = Chest Pain (1=typical, 2=atypical,3=non-anginal pain, 4=asymptomatic)

fbs= Fasting Blood Sugar (1=true,0=false)

exang=exercise induced angina (1=yes,0=no)

Para mejorar el rendimiento de nuestro codigo decidimos eliminar los demas datos no relevantes y se aseguro que este interpretara todas estas variables como categoricas.

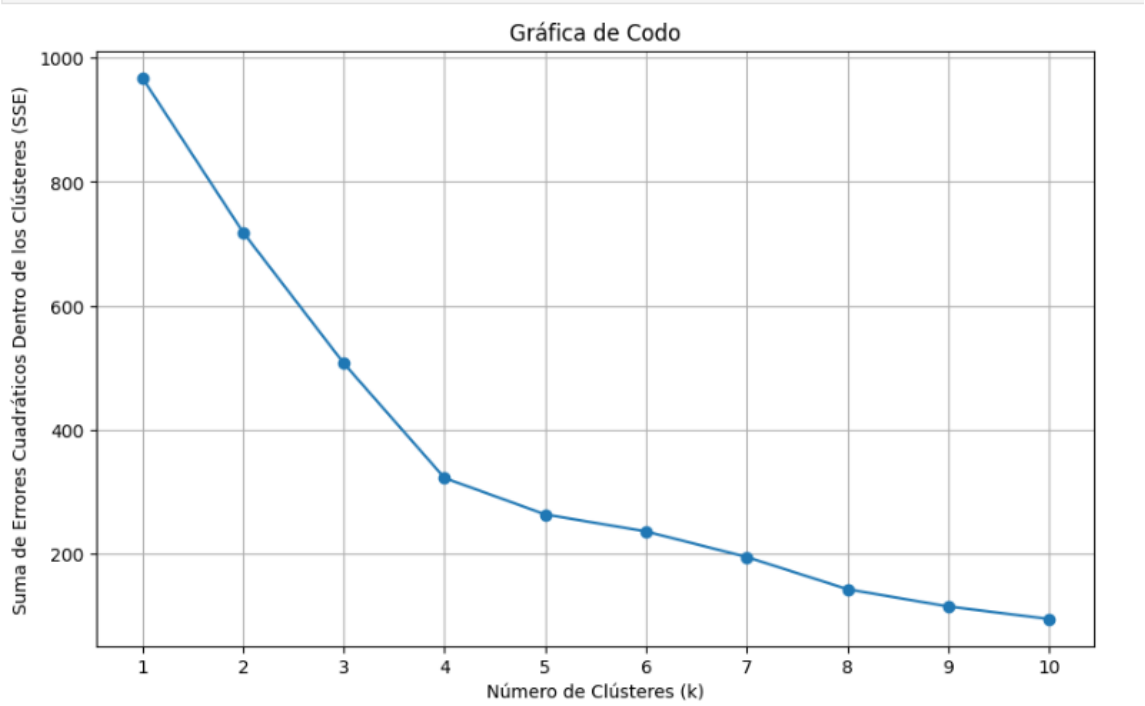
Durante el desarrollo de esta practica vimos que la gran mayoria de los datos estaban completos, aun asi se eliminaron 6 registros que estaban vacios.

Una vez limpiamos los datos vacios se dividio el conjunto de datos en una de entrenamiento y uno de pruebas dejando el 80% de los datos para entrenar y el 20% para probar.

Una vez se dividio el conjunto de datos en partes para poder trabajar se normalizaron las caracteristicas de cada uno de los nuevos datasets.

Para empezar a trabajar con el metodo K-means tenemos que decidir cuantos clusteres queremos implementar. Para obtener el numero mas optimo de clusteres podemos hacer uso de una grafica de codo para saber cual es el numero mas optimo para trabajar.

Empezaremos a calcular la suma de las distancias cuadradas de cada punto al centroide con el que cuenta cada numero de clusters. Al hacer esto podremos obtener una grafica de codo que nos permita ver cual es el mejor numero de clusteres con el cual trabajar.



Podemos ver con esta grafica que nos conviene mas usar 4 clusteres para trabajar con este conjunto de datos ya que cuenta con un buen balance entre el SSE y no hace uso de tantos clusteres.

Una vez decidimos que trabajaremos con 4 clusteres entonces podemos empezar a hacer prediciones y a compararlas con los datos del conjunto de pruebas haciendo usos de matrices de contingencia para apoyar en la comparación.

Al realizar todos estos metodos podemos ver que nuestro modelo de agrupamiento cuenta con una pureza o precisión del 0.4554 para predecir los datos de este conjunto de datos. Si recordamos en la practica anterior nuestro mejor modelo de clasificacion contaba con una precisión de 0.5708. Asi que al menos en esta practica y con estos metodos podemos decir que un algoritmo de clasificación es mejor que una de agrupamiento al momento de predecir si un paciente tiene diagnosticada una enfermedad del corazon.

Conclusión:

El análisis del método del codo permite identificar el número de clusters que mejor representa la estructura de los datos. La comparación con el modelo de clasificación proporciona una visión adicional sobre la utilidad del agrupamiento en la identificación de patrones y estructuras subyacentes en los datos. Los resultados obtenidos ayudarán a comprender mejor la segmentación de los datos y su relación con las etiquetas de clasificación.

Bibliografia:

UCI Machine Learning Repository. (2019). Uci.edu.

https://archive.ics.uci.edu/dataset/45/heart+disease