Universidad Autónoma de Nuevo León

Facultad de Ciencias Físico Matemáticas

Tarea 5

Profesor: José Alberto Benavides Vazquez
Alumno: Damián Atilano Martínez Alvarado 173552

1. Introducción

El aprendizaje no supervisado es una técnica de aprendizaje automático en la que se utilizan algoritmos para analizar datos sin la necesidad de tener una etiqueta o respuesta previa que se quiera predecir. En este tipo de aprendizaje, el modelo no recibe información específica sobre la salida deseada y es el encargado de encontrar patrones o estructuras en los datos por sí solo.

El objetivo principal del aprendizaje no supervisado es descubrir patrones interesantes en los datos y agruparlos en diferentes categorías o clusters, o bien reducir la dimensionalidad de los datos para visualizarlos de manera más sencilla.

El conjunto de datos que tomamos cuenta con variables que determinan si una persona tiene cardiopatía, las variables son las siguientes:

- Age: La edad del paciente.
- Sex: Género del paciente. 0 = M, 1= H.
- Chest pain type: Tipo de dolor en el pecho. Dolor 1-4 siendo el 4 mas fuerte.
- BP: Presión arterial del paciente.
- Cholesterol: Nivel de colesterol del paciente.
- FBS over 120: Prueba de azúcar.
- EKG results: Resultado electrocardiograma.
- Max HR: Frecuencia cardíaca.
- Exercise angina: Ejercicio de angina. 0 - 1, siendo 1 malo.
- ST depression: Depresión del segmento ST en el electrocardiograma.
- Slope of ST: La inclinación del segmento ST en el electrocardiograma.
- Number of vessels fluor: La cantidad de vasos que se ven en las imágenes de fluoroscopia.
- Thallium: Prueba de talio.
- Cardiopatía: Si el paciente cuenta con problemas cardíacos. 0 - Si 1 - No

El objetivo de este proyecto es determinar que variables estan mas relacionadas o cuales son las que influyen mas para determinar si una persona tiene cardiopatía. En esta ocasion usaremos un metodo para determinar el número indicado de clusters, esto es para saber para agrupar a los pacientes con características especificas y asi ver que grupo tiende a contar con esta enfermedad.

2. Técnicas

Una de las técnicas que empleamos es el feature scalling que es una técnica de preprocesamiento de datos utilizada en el aprendizaje automático para normalizar o estandarizar los valores de las características de los datos. El objetivo es ajustar las escalas de las características para que estén en un rango similar, de manera que el modelo de aprendizaje automático pueda interpretar mejor la importancia de cada característica en la predicción de la variable objetivo.

Las características pueden tener diferentes unidades de medida o escalas, lo que puede provocar que algunas características tengan un impacto desproporcionado en la predicción del modelo debido a su escala. La técnica de feature scaling permite transformar los valores de las características para que estén en una escala común, lo que facilita la comparación y el análisis de las características.

El método del codo es una técnica utilizada en el aprendizaje no supervisado, específicamente en el análisis de clusters, para determinar el número óptimo de clusters a utilizar en un conjunto de datos.

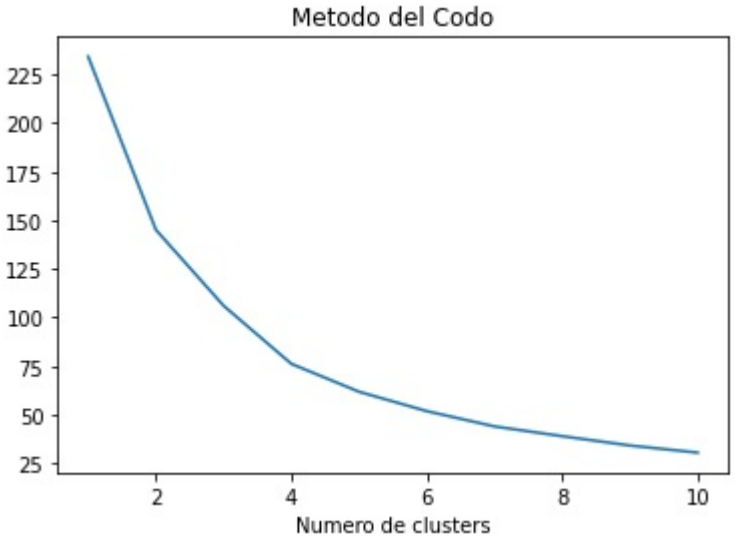
El método del codo se basa en el principio de que la adición de clusters adicionales mejorará la varianza explicada, pero llegará un punto en el que la adición de más clusters ya no proporcionará una mejora significativa. En ese punto, el gráfico de la varianza explicada en función del número de clusters forma un ángulo como un codo, lo que indica el número óptimo de clusters.

A continuación los resultados de los métodos:

Chest pain type	Sex	Cholesterol	Exercise angina	Max HR	Age	Cardiopatia
4	1	322	0	109	70	0
3	0	564	0	160	67	1
2	1	261	0	141	57	0
4	1	263	1	105	64	1
2	0	269	1	121	74	1

Chest pain type	Sex	Cholesterol	Exercise angina	Max HR	Age	Cardiopatia
1.00	1.0	0.447489	0.0	0.290076	0.854167	0.0
0.66	0.0	1.000000	0.0	0.679389	0.791667	1.0
0.33	1.0	0.308219	0.0	0.534351	0.583333	0.0
1.00	1.0	0.312785	1.0	0.259542	0.729167	1.0
0.33	0.0	0.326484	1.0	0.381679	0.937500	1.0

Notamos que la técnica Feature Scalling hace que los diferentes datos tengan la misma escala.



Observamos en la imagen que la curva o el codo se empieza a marcar desde 3 pasando por 4, 5, y 6. Realizaremos las pruebas para ver cual es la mejor opción.

- #1 Cluster: 120 de 270 muestras bien etiquetadas con una precisión del 44 %.
- #2 Cluster: 0 de 270 muestras bien etiquetadas con una precisión del 0 %.
- #3 Cluster: 0 de 270 muestras bien etiquetadas con una precisión del 0 %.
- #4 Cluster: 3 de 270 muestras bien etiquetadas con una precisión del 1 %.
- #5 Cluster: 0 de 270 muestras bien etiquetadas con una precisión del 0 %.
- #6 Cluster: 0 de 270 muestras bien etiquetadas con una precisión del 0 %.
- #7 Cluster: 112 de 270 muestras bien etiquetadas con una precisión del 41 %.
- #8 Cluster: 83 de 270 muestras bien etiquetadas con una precisión del 31 %.
- #9 Cluster: 103 de 270 muestras bien etiquetadas con una precisión del 38 %.

Los números de cluster 3, 4, 5 y 6 nos arrojan un resultado no deseado, realizamos las pruebas para los demas grupos y nos da como resultado que el #1 con una precisión del 44 % 120 muestras bien etiquetadas de 270.

3. Conclusiones

Podemos concluir que la técnica Feature Scalling y el Método del codo, nos ayudaron para buscar el numero de cluster ideal pero no nos dio la precision esperada para nuestro trabajo.