



Tecnológico de Monterrey

Campus Guadalajara

Análisis de datos médicos con aprendizaje no supervisado

Materia:

Modelación del aprendizaje con inteligencia artificial (Gpo 301)

Estudiantes

Juan Pablo Valenzuela Dorado A00227321

Juan Pablo Bernal Lafarga A01742342

Francelio Uriel Rodríguez García A01352663

Fecha

06/14/2023

Introducción

Este documento presenta los resultados de un análisis de agrupamiento realizado en una base de datos relacionada con la atención médica de pacientes diabéticos. La base de datos contiene diversos atributos, como el número de pacientes, raza, sexo, edad, tipo de admisión, tiempo en el hospital, especialidad médica del médico de admisión, número de pruebas de laboratorio realizadas, resultado de la prueba de HbA1c, diagnóstico, número de medicamentos, medicamentos para la diabetes y número de consultas ambulatorias, hospitalarias y de emergencia en el año anterior a la hospitalización.

Descripción de la base de datos

El conjunto de datos contiene atributos tales como número de pacientes, raza, sexo, edad, tipo de admisión, tiempo en el hospital, especialidad médica del médico de admisión, número de pruebas de laboratorio realizadas, resultado de la prueba de HbA1c, diagnóstico, número de medicamentos, medicamentos para la diabetes, número de consultas ambulatorias, hospitalarias y de emergencia en el año anterior a la hospitalización, etc.

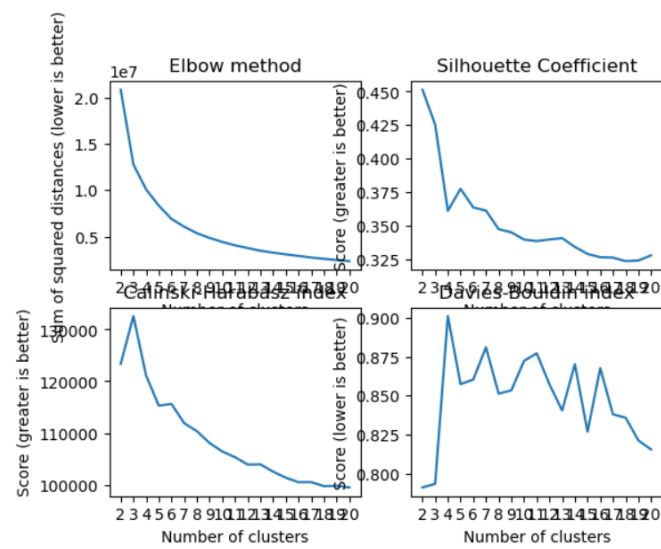
Existía un atributo que indica si el paciente estaba si el paciente recibía o no medicamentos para diabetes, sin embargo, como equipo decidimos no trabajar con el atributo dentro de nuestros modelos para evitar lo más posible tener una categorización en la base de datos, y permitir que los algoritmos de agrupamiento que encontramos realicen su trabajo de la mejor manera posible.

Los datos fueron registrados de parte de “Center for Clinical and Translational Research, Virginia Commonwealth University”.

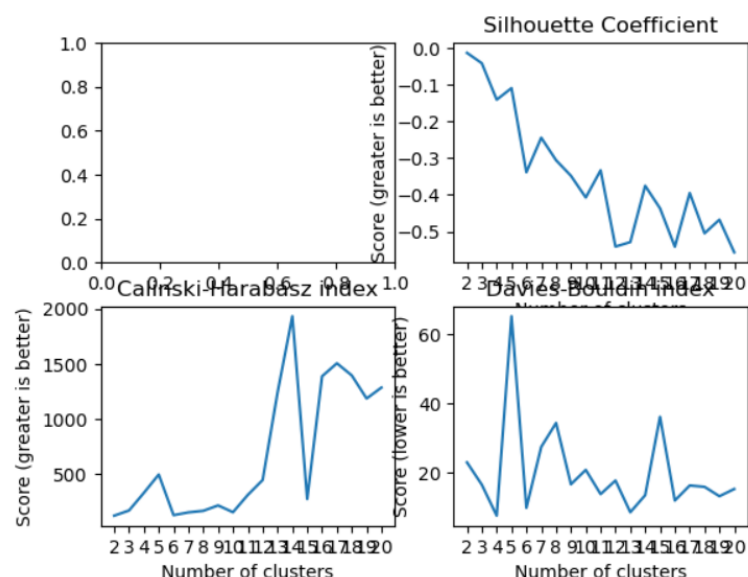
Resultados tras aplicar los métodos de agrupamiento

Comenzamos con la selección del número de clusters óptimos para nuestro modelo, para esto realizamos K-Means en la base de datos y utilizamos los métodos “elbow method”, “Silhouette coefficient”, “Calinski-Harabasz index”, y

“Davies-Bouldin index” para analizar cuál sería el número adecuado de clusters. Así, interpretando las siguientes gráficas encontramos que el número de clusters adecuado era 5.

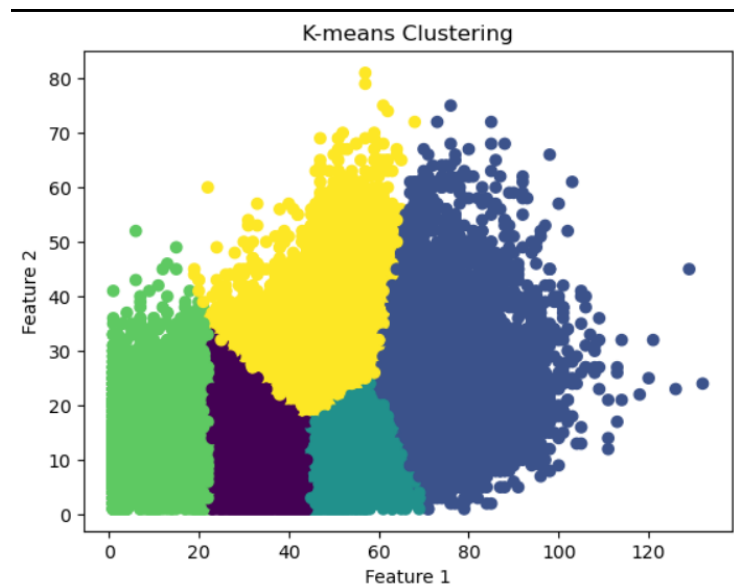


Continuando con los resultados de los modelos, cuando cambiamos los modelos para que coincidieran con el número de clusters adecuados que nos sugirieron los métodos anteriores, utilizamos k-means, spectral clustering y DBSCAN para probar cómo clasificarían los modelos nuestra base de datos, y tuvimos resultados bastante interesantes y diferentes para cada uno de los modelos.



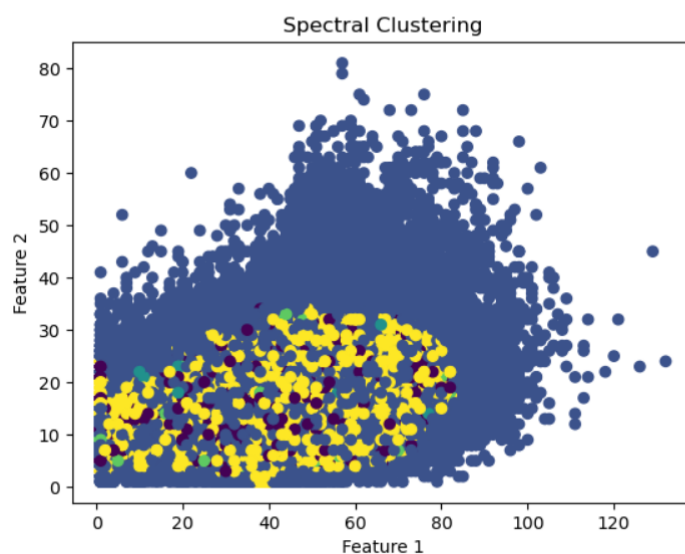
```
Hiperparámetros óptimos: {'eps': 0.1, 'min_samples': 3}
Coeficiente de silueta: 0.968716019130277
```

K-means clustering



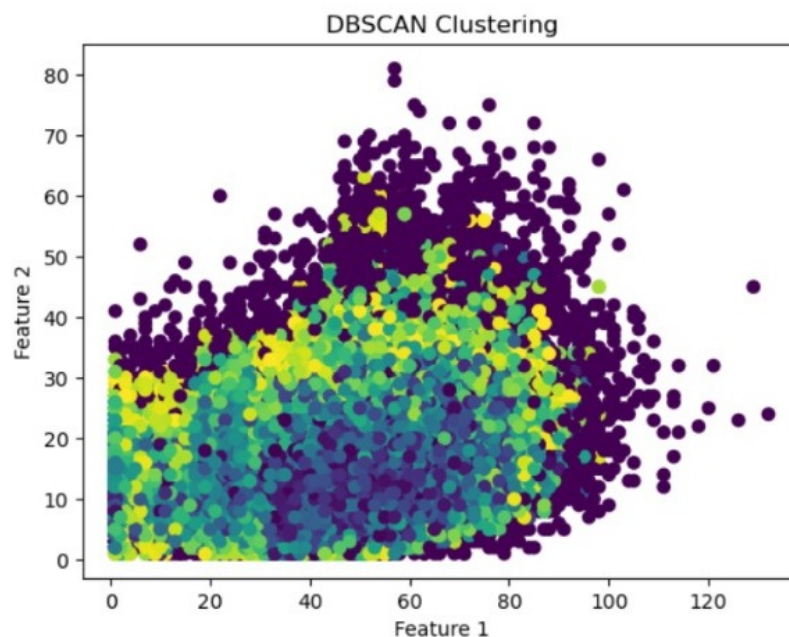
Citando a la página de “Aprende IA”, “El algoritmo funciona de manera iterativa para asignar cada punto de datos a uno de los grupos K en función de las características que se proporcionan. Los puntos de datos se agrupan en función de la similitud de las características”. Este algoritmo crea K centroides para etiquetar datos nuevos, y a cada dato se le asigna un solo centroide. El agrupamiento se ve bastante claro, el modelo clasificó por regiones los datos, y nos dio 5 clusters claramente visibles.

Spectral clustering



El spectral clustering se caracteriza por encontrar relaciones entre los datos sin importar la complejidad de las estructuras, es decir, no se basa en la forma de algún polinomio, ni centroides para agrupar los datos, en vez, se basa en las conexiones que existen entre los datos, y utilizando eigenvectores rompen las conexiones así formando clusters.

DBSCAN clustering



La otra cara de la moneda del método OPTICS, es otra aproximación basada en densidad, que sólo requiere dos parámetros para funcionar. Esto puede ser un poco difícil de encontrar, además el método no es capaz de encontrar clusters con distintas densidades.

Información descubierta

A través de los modelos previamente mencionamos, encontramos que pudimos realizar el agrupamiento de dos características que fueron relevantes para nuestro modelo, estas dos características son 'num_lab_procedures' y 'num_medications', a través del agrupamiento pudimos capturar la relación tanto sobre la atención médica recibida por los pacientes diabéticos como sobre las características de su tipo de condición diabética.

Dentro del algoritmo de K-Means podemos obtener información de los clusters creados para nuestro modelo, dentro de esta información pudimos notar que dentro del valor promedio de la variable llamada 'time_in_hospital' perteneciente al grupo 0 fue de 3.5 dando a entender que el tiempo en el que pasaban en el hospital fue relativamente corta dentro del promedio.

Asimismo, dentro a lo relacionado a la variabilidad de los valores dentro del grupo de características llamada 'num_lab_procedures' pudimos deducir que se obtuvo un resultado estándar bajo y con ello nos estaría indicando que la mayoría de los pacientes tienen un número similar de procedimiento de laboratorio dentro del grupo 0, en comparación al resto.

time_in_hospital	num_lab_procedures
30896.000000	30896.000000
3.501845	34.986018
2.467303	5.992805
1.000000	23.000000
2.000000	30.000000
3.000000	36.000000
4.000000	40.000000
14.000000	44.000000

Cuando realizamos la clusterización con el modelo de spectral clustering, podemos percatarnos que el modelo le dio un peso muy alto a ciertas peculiaridades que se repetían en varios datos, sin importar de ninguna manera la forma que tuvieran estos. Tomó datos que compartían ciertas características de manera casi idéntica, y los agrupó en clusters muy pequeños, lo que podemos interpretar como datos que estaban conectados de una manera muy fuerte, que cortaron la conexión con los otros clusters debido a que al algoritmo le pareció o irrelevante o muy distante a comparación.

Al terminar de compilar el modelo de spectral clustering, nos quedaron clusters de tamaños muy desproporcionados, y estos clusters entre más pequeños son, se vuelven más estrictos con las condiciones para formar parte de ellos, mientras que cuando los clusters se vuelven más grandes, las condiciones se vuelven más generales, por lo que las distancias entre sí crecen en gran medida.

time_in_hospital	num_lab_procedures	num_procedures	num_medications
30.000000	30.0	30.000000	30.0
4.300000	9.0	1.733333	19.0
2.793465	0.0	1.818171	0.0
2.000000	9.0	0.000000	19.0
2.000000	9.0	0.000000	19.0
3.000000	9.0	1.000000	19.0
5.000000	9.0	3.000000	19.0
12.000000	9.0	6.000000	19.0

Finalmente, en el algoritmo de DBSCAN, encontramos que el algoritmo creó una infinidad de clusters muy pequeños, y gracias al pequeño tamaño de los clusters, el algoritmo puede ser bastante específico con los requisitos para unirse a un cluster, sin embargo, seguramente la cantidad de clusters que se realizó de a pié a modelos sin utilidad, por lo que podemos suponer que este método no es el adecuado para clasificar este tipo de bases de datos.

Cluster 0

```

Xdbscan = data
Xdbscan['etiquetas_columna'] = clustering_labels
grupos = Xdbscan.groupby('etiquetas_columna')
grupo_0 = grupos.get_group(0)
grupo_0.describe()

```

admission_type_id	discharge_disposition_id	admission_source_id	time_in_hospital	num_lab_procedures	num_procedures	num_medications	number_outpatient	number_emergency	number_inpatient
81.000000	81.000000	81.000000	81.000000	81.0	81.000000	81.0	81.000000	81.000000	81.000000
2.037037	3.481481	5.703704	5.283951	59.0	1.246914	18.0	0.765432	0.259259	0.864198
1.676637	4.056202	3.465705	2.110892	0.0	1.520944	0.0	2.516305	0.997218	1.506263
1.000000	1.000000	1.000000	2.000000	59.0	0.000000	18.0	0.000000	0.000000	0.000000
1.000000	1.000000	1.000000	4.000000	59.0	0.000000	18.0	0.000000	0.000000	0.000000
1.000000	3.000000	7.000000	5.000000	59.0	1.000000	18.0	0.000000	0.000000	0.000000
2.000000	5.000000	7.000000	7.000000	59.0	2.000000	18.0	0.000000	0.000000	1.000000
8.000000	22.000000	17.000000	12.000000	59.0	6.000000	18.0	20.000000	8.000000	7.000000

Selección de videos para la creación de un tutorial basado en un modelo de agrupamiento

Juan Bernal:

Los video tutoriales en los que me inspiré para hacer mi video hablan de forma concisa e interesante acerca de sus temas. Fueron videos sobre los 3 paradigmas de aprendizaje, el aprendizaje por refuerzo (únicamente) y el

funcionamiento de algunos modelos de agrupamiento, siendo que ya estaba viendo el tema de aprendizaje no supervisado. Lo interesante de la secuencia de estos 3 videos es que uno de ellos te habla acerca de la caja negra en la que se encuentran actualmente los algoritmos de aprendizaje, pues son popularmente utilizados sin conocer la función matemática que se encuentra detrás de dichos algoritmos. Los otros videos te dan la teoría sobre los códigos, por lo que se nota un contraste entre la explicación que te pueden dar sobre un algoritmo a alguien que sabe su funcionamiento matemática, a alguien que únicamente importa una librería y corre 2 líneas de código. En mi video me inspiré en esta idea de abrir la caja negra y noté que existen algoritmos tanto complejos, como relativamente sencillos en el mundo de los algoritmos de aprendizaje.

Liga del video tutorial:

[¿Cómo funciona DBSCAN? - VideoTutorial](#)

Juan Valenzuela:

Lo que me hizo seleccionar estos videos sobre otros es que el nivel con el que se explicaron las cosas fueron adecuadas para un estudiante de universidad, además, los tres videos se complementan de una manera que uno te permite comprender el otro, y utilizándolos como herramientas, sentí la capacidad de crear un video explicativo más útil para aquellos que lo visualicen.

Liga del video tutorial:

[Resumen de Spectral Clustering en Español: Método de agrupamiento - YouTube](#)

Francelio Rodriguez:

Fue difícil encontrar videos adecuados e interesantes de acuerdo con el tema que quería escuchar hablar, al encontrarme estos 3 videos y verlos completos, entendí de mejor manera el funcionamiento del algoritmo de DBSCAN y todo lo que conlleva detrás para que en la implementación de este algoritmo a nuestra base de datos fuera correcta y con el mínimo de margen de errores. Gracias a estos videos,

se puede aplicar y mejorar lo aprendido de hoy en adelante en la utilización de este algoritmo para aplicar métodos de agrupamiento en algún proyecto que lo requiera.

Link del video tutorial:

[DBSCAN Clustering](#)

Conclusión

El análisis de agrupamiento aplicado a la base de datos de atención médica de pacientes diabéticos proporcionó insights valiosos sobre las características y patrones presentes en los datos. A través de los métodos de K-means clustering, Spectral clustering y DBSCAN clustering, se identificaron diferentes agrupamientos y relaciones entre los pacientes.

El algoritmo de K-means clustering clasificó los datos en cinco clusters claramente visibles, lo que permitió una segmentación efectiva de los pacientes diabéticos según características similares. Se destacaron diferencias en el tiempo de hospitalización y en el número de procedimientos de laboratorio entre los distintos grupos.


Por otro lado, Spectral clustering reveló conexiones fuertes entre los datos, independientemente de su forma, lo que llevó a la formación de clusters más pequeños pero más específicos. Esta metodología se centró en características compartidas casi idénticas y permitió identificar patrones más sutiles en los datos.

En contraste, DBSCAN clustering generó una gran cantidad de clusters muy pequeños, lo que dificulta su interpretación y utilidad práctica en este contexto particular. Aunque este método fue más específico en los requisitos para unirse a un cluster, su alta cantidad de clusters puede hacer que los resultados sean menos interpretables.

En resumen, el análisis de agrupamiento proporcionó una comprensión más profunda de la atención médica de los pacientes diabéticos, destacando la relación

entre el tiempo de hospitalización, el número de procedimientos de laboratorio y otras características relevantes. Estos hallazgos si se les dedica el tiempo suficiente para realizar una exploración de los resultados pueden tener implicaciones importantes para la planificación de la atención médica y el desarrollo de estrategias de tratamiento más personalizadas para los pacientes diabéticos. Sin embargo, es importante considerar las fortalezas y limitaciones de cada algoritmo de agrupamiento utilizado y seleccionar el enfoque más apropiado para los objetivos de análisis específicos.

Referencias

Gonzalez, L. (2020, August 19). *Aprendizaje no supervisado: K-means clustering*.  Aprende IA.

<https://aprendeia.com/aprendizaje-no-supervisado-k-means-clustering/#:~:text=K-Means%20es%20un%20tipo%20de%20aprendizaje%20no%20supervisado%2C,datos%20se%20agrupan%20seg%C3%BAn%20la%20similitud%20de%20caracter%C3%ADsticas>.

Humberto Brandão, Ph. D. (2017, October 31). *Diabetes 130 US hospitals for years 1999-2008*. Kaggle.

<https://www.kaggle.com/datasets/brandao/diabetes>