Universidad Internacional de la Rioja (UNIR)

ESIT

Máster Universitario en Inteligencia Artificial

Comparativa de técnicas de aprendizaje no supervisado para la identificación y caracterización de subpoblaciones de diabetes mellitus tipo II

Trabajo Fin de Máster

presentado por: Giorgio Enrique Hernández Pineda

Dirigido por: María Teresa García Ordás

Ciudad:Bucaramanga

Fecha: 1 de febrero de 2024

Índice de Contenidos

Índice de Ilustraciones

Índice de Tablas

Resumen

En el presente estudio, se propone la evaluación de tres técnicas de aprendizaje no supervisado en un conjunto de datos de diabetes mellitus, que recopila datos de aproximadamente 100.000 pacientes en más de 130 hospitales en Estados Unidos, con el objetivo de identificar y caracterizar distintas sub-poblaciones. La metodología implementada fue una versión de CRISP-DM acondicionada a los problemas de clasificación no supervisada. Fueron evaluados algoritmos con principios de funcionamiento diferentes, siendo el K-Means, DBSCAN, y agrupamiento jerárquico los seleccionados. Para evaluarlos, se seleccionaron 3 índices diferentes, sin embargo, el análisis de sub-poblaciones resultantes fue el más determinante en la evaluación. Finalmente, el agrupamiento jerárquico es una excelente opción, siempre y cuando la capacidad de cómputo permita utilizarle. Seguido, los algoritmos como K-Means Y DBSCAN requieren de representaciones adecuadas, siendo UMAP la preferida en cuanto a conjuntos de datos con alta cantidad de variables epidemiológicas y gran cantidad de datos.

Palabras Clave: Aprendizaje no Supervisado, DBSCAN, Agrupamiento Jerárquico, K-Means, Diabetes Mellitus.

Abstract

This study offers an assessment of unsupervised techniques on a diabetes mellitus dataset which comprises data on about 100.000 patients and 130 hospitals in the United States. The objective of this research is to identify sub-populations within the data. The methodology used is a CRISP-DM-adapted version for non-supervised algorithms. Three algorithms with different principles were evaluated: K-Means, DBSCAN, and hierarchical clustering. Three different metrics were used for evaluation; however, sub-population-based analysis was found to be the most pertinent. Additionally, when the processing capacity permits it, hierarchical clustering is an excellent option. Finally, K-Means and DBSCAN algorithms require proper representations of the data, making UMAP the most recommended for datasets including a significant number of epidemiological variables.

Key Words: Unsupervised Learning, DBSCAN, Hierarchical Clustering, K-Means, Diabetes Mellitus

Capítulo 1

Introducción

La diabetes mellitus es, y desde hace un tiempo viene siendo, una enfermedad cada vez más común que afecta actualmente a casi 500 millones de personas en todo el mundo, y se proyecta que para el año 2030, este número llegará hasta los 578 millones. Así, la diabetes mellitus es considerada actualmente como un problema de salud pública. Añadiendo, también se considera que al rededor de 193 millones de personas tienen la enfermedad sin estar enterados debido a su naturaleza asintomática (Atlas, 2019). La diabetes, es una enfermedad metabólica que se caracteriza por niveles altos de glucosa en la sangre debido a que el cuerpo no produce la suficiente insulina, o que éste es resistente a los efectos de la insulina. El cuerpo, necesita de insulina para poder utilizar proteínas, azúcar y grasa como energía (Temurtas et al., 2009). También, la diabetes es generalmente dividida en dos clases, tipo I y tipo II, de las cuales la diabetes tipo II es la más común. También, la diabetes es asociada con varias complicaciones, como el riesgo a la ceguera, presión arterial, y enfermedades cardiovasculares, entre otras. De esta manera, es cada vez más necesario realizar estudios que permitan su detección temprana, mejora de tratamiento, e inclusive poder evitarla. Su detección temprana es de extrema dificultad, por lo que se han intentado distintas soluciones utilizando tanto aprendizaje supervisado como no supervisado (Khashei et al., 2012).

Actualmente, se han utilizado gran cantidad y variedad de modelos de clasificación para poder explorar los datos de pacientes, y así, poder mejorar el diagnóstico, tratamiento utilizado o incluso una detección temprana de la enfermedad. En su mayoría, se han desarrollado modelos supervisados, sin embargo, en este documento se propone la utilización de técnicas no supervisadas. Estas técnicas, se diferencian de las supervisadas en que los

modelos no son entrenados con ninguna indicación sobre cual es la salida esperada, y por tal motivo, aprenden directamente de los datos proporcionados. Entonces, poseen la ventaja de detectar patrones complejos en los datos disponibles y da la posibilidad de encontrar relaciones entre ellos que no se podrían encontrar de otra forma. Por lo tanto, se planeó la comparación de diversas técnicas de aprendizaje no supervisado sobre un conjunto de datos recopilado por más de 130 hospitales en Estados Unidos durante 10 años (1999-2008). Este conjunto de datos, está disponible en el sitio web de la universidad de California, Irvine, escuela de información y ciencias de la computación *UCI Machine Learning Repository* y fue compartido por el Centro de investigación clínica y traslacional (*Center for Clinical and Translational Research*) en conjunto con la universidad Commonwealth en Virginia (Dua and Graff, 2017). Es de importancia mencionar que el conjunto de datos tiene limitaciones que pueden afectar directamente el resultado de los modelos a comparar, pues algunas características utilizadas para realizar los agrupamientos en la literatura no se encuentran presentes o tienen gran cantidad de valores nulos (HbA1c, BMI, HOMA2-B, entre otras). Este análisis exploratorio del conjunto de datos es abordado en detalle por el autor en (Strack et al., 2014). Así, se busca comparar y evaluar el rendimiento de diversos modelos no supervisados entre un grupo seleccionado previamente, utilizando el mismo conjunto de datos y distintas métricas de distancia.

Capítulo 2

Contexto y Estado del Arte

2.1. Diabetes Mellitus

Los autores Francisco Javier Tébar Massó y Mercedes Ferrer Gómez, mencionan la diabetes mellitus como "un grupo de enfermedades o síndromes metabólicos caracterizados por la aparición de hiperglucemia secundaria a defectos de la secreción de insulina, de la acción de la insulina, o de ambas" (Tebar Masso and Ferrer Gomez, 2009). Así mismo, en (Jameson et al., 2018), se la define como: "La diabetes mellitus comprende un grupo de trastornos metabólicos frecuentes que comparten el fenotipo de la hiperglucemia". Durante muchos años, han existido distintas clasificaciones para la enfermedad (Adame et al., 2002; Group, 1979; Nallaperumal et al., 2013), siendo la diabetes tipo I, II, gestacional, y otros tipos específicos, las categorías utilizados actualmente de la enfermedad (on the Diagnosis and of Diabetes Mellitus, 1998; Metzger et al., 1998; Tebar Masso and Ferrer Gomez, 2009; Jameson et al., 2018). También, la diabetes mellitus provoca afectaciones en todo el organismo, siendo las más notorias: pérdida de visión hasta ceguera, afectación renal que puede llegar hasta insuficiencia renal terminal, afectación de los grandes vasos que puede llevar a una insuficiencia arterial de extremidades inferiores, y cardiopatía isquémica. Es importante mencionar que mientras disminuye la acción insulínica y aumenta la hiperglicemia, el paciente notará lo que se conoce como síntomas cardinales de la diabetes mellitus: poliuria, polidipsia, polifagia. Hasta que esto ocurre, y especialmente en la diabetes mellitus tipo II, existe un periodo extenso (5 a 10 años) de hiperglucemia asintomática, en los cuales el paciente es expuesto a la aparición de complicaciones crónicas de la diabetes mellitus. En otras palabras, es común que en el momento de diagnóstico de un paciente de diabetes mellitus tipo 2, este ya sufra de complicaciones de distinta índole y que el diagnóstico se

Tabla 2.1: Criterios establecidos por la ADA para el diagnóstico de diabetes mellitus.

Test	**Prediabetes**	**Diabetes**
Hemoglobina A_{1C}	5.7−6.4 % (39−47 mmol/mol)	⩾6.5 % (48 mmol/mol)
FPG	100−125 mg/dL (5.6−6.9 mmol/L)	⩾126 mg/dL (7.0 mmol/L)
glucosa plasmática durante 75g OGTT	140−199 mg/dL (7.8−11.0 mmol/L)	⩾200 mg/dL (11.1 mmol/L)

realice a partir de otros procesos no relacionados con la diabetes mellitus (Tebar Masso and Ferrer Gomez, 2009; Jameson et al., 2018).

Para el diagnóstico de la diabetes mellitus, se deben conocer las tres clasificaciones de la tolerancia a la glucosa: homeostasis normal de la glucosa, diabetes mellitus, homeostasis alterada de la glucosa. La homeostasis de la glucosa, básicamente, es el equilibrio entre el consumo de energía proveniente de alimentos ingeridos, la producción hepática de glucosa, y la utilización de glucosa por parte de los tejidos periféricos, la insulina sería el regulador más importante de este equilibrio. Para valorar la tolerancia a la glucosa se puede emplear la glucosa plasmática en ayuno (FPG, *Fasting Plasma Glucose*), la respuesta a una carga oral de glucosa (OGTT *Oral Glucose Tolerance Test*), o la hemoglobina A_{1C} (HbA_{1C}). Así, una glucosa en plasma <140 mg/100 mL (7.9 mM/L) después de una reacción a una carga oral de glucosa y una HbA_{1C} <5.7 % se considera que definen la tolerancia normal a la glucosa (Jameson et al., 2018). En la Tabla 2.1, se pueden observar los criterios establecidos por la ADA (*American Diabetes Association*) para el diagnóstico de diabetes mellitus Adaptado de (Care, 2022).

2.1.1. Diabetes Mellitus Tipo I

La característica principal de la diabetes tipo I consiste en una destrucción inmunitaria de las células β del páncreas, las cuales, se encargan de producir insulina para posteriormente liberarla al torrente sanguíneo. Esta respuesta autoinmunitaria, ocasiona una deficiencia absoluta de insulina (Tebar Masso and Ferrer Gomez, 2009; Jameson et al., 2018). En fases tempranas de la enfermedad, cuando no hay criterios diagnósticos para

la diabetes mellitus, pero sí de otras anomalías del metabolismo de la glucosa, aparecen distintos tipos de anticuerpos en la sangre. Estos anticuerpos pueden estar dirigidos a: las propias células (anti-islotes o ICA), insulina, la descarboxilasa del ácido glutámico (GAD_{65}) o las tirosin-fosfatasas (IA-2 e IA-2β). La velocidad de aparición de la enfermedad es variable, y depende de la velocidad de destrucción de las células β, la cual es mayor en niños y adolescentes (Jameson et al., 2018).

Existen varias intervenciones realizadas que han evitado la diabetes mellitus tipo I en animales, sin embargo, pocas intervenciones se han realizado en humanos. No obstante, se ha probado que el suministro de insulina a pacientes con alto riesgo de diabetes tipo I no impide (ni retarda) que sufriesen la enfermedad (Sosenko et al., 2006).

2.1.2. Diabetes Mellitus Tipo II

La diabetes mellitus tipo II se caracteriza principalmente por la presencia de resistencia a la acción periférica de la insulina, secreción de insulina defectuosa, o ambas. Estos problemas, son causados por diversos defectos genéticos o metabólicos que llevan a la hiperglucemia. Sin embargo, en la actualidad existen diversos fármacos para corregir o modificar trastornos metabólicos específicos. Mientras que el páncreas mantenga una secreción de insulina suficiente para superar la resistencia a la insulina, el paciente de diabetes tipo II, se mantiene en una situación de no insulinodependencia. No obstante, con el paso de los años el páncreas empezará a ceder, y la secreción de insulina será insuficiente para compensar la glucemia, causando una insulinodependencia. En su mayoría, los diabéticos tipo II, pasan cerca de 10 años sin diagnóstico debido a que son asintomáticos durante ese tiempo, razón por la cual, la ADA considera los test de FPG y HbA_{1C} como pruebas de detección. Del mismo modo, para alentar evitar la diabetes, la ADA recomienda estudios de detección inicial a toda persona mayor a 45 años cada 3 años, o en sujetos con índice de masa corporal (BMI *Body mass index*) superior a los $25 kg/m^2$ (Tebar Masso and Ferrer Gomez, 2009; Jameson et al., 2018; Care, 2022).

Cambios de estilo de vida en el periodo de prediabetes pueden retrasar e incluso prevenir la diabetes mellitus tipo II. Específicamente, reducir el peso corporal y aumentar la actividad física evita o retrasa la aparición de diabetes mellitus II en el 58 % de los casos.

2.1.3. Diabetes Gestacional

En el caso de la diabetes gestacional, se diagnostica cuando se desarrolla hiperglucemia e intolerancia a la glucosa durante el segundo o tercer trimestre del embarazo y no debe haber sido conocida antes del mismo. Es causada por alteraciones metabólicas propias del embarazo, que llevan a insulinoresistencia. Luego del embarazo, la mayoría de mujeres recuperan una tolerancia normal a la glucosa, no obstante, tienen un riesgo del 35 a 60 % de padecer diabetes mellitus en los próximos 10 a 20 años. Igualmente, sus hijos poseen un incremento en el riesgo de desarrollar síndromes metabólicos y padecer diabetes mellitus tipo II. La ADA, recomienda a mujeres con antecedentes de diabetes gestacional, realizarse a pruebas de detección cada 3 años a lo largo de su vida (Tebar Masso and Ferrer Gomez, 2009; Jameson et al., 2018; Care, 2022).

2.1.4. Otros tipos de Diabetes Mellitus

Esta sección reúne situaciones clínicas con un diagnóstico de Diabetes Mellitus que no tienen relación entre sí. En general, son defectos genéticos específicos de la secreción o acción de la insulina (Tebar Masso and Ferrer Gomez, 2009; Jameson et al., 2018).

- Diabetes tipo MODY (*Maturity Onset Diabetes of the Young*): Son defectos genéticos en las células β, que inducen un déficit en la producción de insulina. Usualmente, se diagnostica antes de los 25 años de edad y su comportamiento es similar al de la Diabetes Mellitus tipo II.

- Defectos en la acción de la insulina de etiología genética: Son causas infrecuentes de Diabetes Mellitus, y se determinan por alteraciones del receptor de insulina que pueden llevar a hiperglucemias graves.

- Enfermedades del páncreas exocrino: son enfermedades que pueden afectar el páncreas endocrino y producir Diabetes Mellitus. Entre las más conocidas se encuentran la pancreatitis, pancreatectomía, neoplasia, pancreatopatía fibrocalculos, y cáncer de pancreas. En éste último, la alteración metabólica aparece como manifestación del cáncer.

- Endocrinopatías: es la aparición de hormonas contra-insulares a niveles anormales causados por una variedad de enfermedades que da lugar a hiperglucemia y Diabetes

Mellitus. El control de la enfermedad, hará desaparecer la Diabetes Mellitus.

- Inducidas por fármacos u otras sustancias químicas: En realidad, precipitan la aparición de alteraciones ya existentes, ya sean disfunción secretora o insulinoresistencia.
- Infecciones: Algunos virus se han asociado con la destrucción de las células β.
- Formas infrecuentes de diabetes mediada por inmunidad:
  - síndrome del hombre rígido (*stiff man*), el cual, sus pacientes suelen tener anticuerpos anti-GAD positivos, y un tercio desarrolla Diabetes Mellitus.
  - anticuerpos anti-receptor de insulina que bloquean la acción insulínica y causan hiperglucemia.
- Otros síndromes genéticos: son síndromes que a veces se acompañan con diabetes, como el síndrome de Down, síndrome de Wolfram, síndrome de Klinefelter, entre otros.

2.2. Aprendizaje Automático

El autor Ethem Alpaydin, en su libro (Alpaydin, 2016), describe el aprendizaje automático (*Machine Learning*) como un requerimiento de la inteligencia artificial, en el cual, un sistema dentro de un ambiente cambiante se adapta y adquiere la habilidad de aprender. De este modo, el diseñador no deberá prever soluciones a todas las posibles situaciones. Así, el aprendizaje automático es una rama de la inteligencia artificial, en donde un método (o conjunto de métodos) detecta o extrae patrones desde grandes conjuntos de datos, y más adelante, utiliza los patrones encontrados para predecir datos futuros o realizar diferentes tipos de decisiones (Murphy, 2012).

En el aprendizaje automático, es de gran importancia, además de la recolección de datos de entrada, analizar éstos para descubrir relaciones entre sí. Estas relaciones (o características) terminan siendo factores que afectan al resultado final del modelo, y deben representar suficientemente bien las características de la tarea a realizar. Cualquier modelo de aprendizaje automático debe tomar los datos recolectados y pasar por una etapa de entrenamiento, donde dependiendo de su desempeño, se considerará si ha aprendido o no una tarea. Existen muchos algoritmos que difieren mucho en sus criterios, funcionamiento, o los parámetros que optimizan internamente, sin embargo, el propósito principal del

aprendizaje automático es ya sea tomar la decisión o predicción correcta de casos que no conozca, que de otro modo solo sería una memorización de los datos de entrenamiento. A este concepto se le conoce como generalización. De este modo, la similitud entre los datos conocidos (de entrenamiento) y los nuevos datos de entrada se da en términos de los atributos de entrada escogidos previamente (Alpaydin, 2016).

Existen dos principales tipos de aprendizaje. En primer lugar, el aprendizaje supervisado, consiste en aprender una función a partir de ejemplos en los cuales se conocen sus entradas y sus salidas, de forma que, dependiendo de si la variable objetivo es numérica o categórica, pueda realizar una predicción o clasificación respectivamente. Por otro lado, el aprendizaje no supervisado, también conocido como descriptivo, detecta patrones a partir de ejemplos en los cuales solo se conocen sus entradas. Así, el principal objetivo del aprendizaje no supervisado es descubrir la estructura interna de los datos. De esta manera, el problema del aprendizaje no supervisado se podría considerar como discernir múltiples categorías en una colección de objetos no etiquetados, y se parte por entender el tipo de distribución de probabilidad que los datos pueden haber generado. Usualmente, se predefine un número de grupos o *clusters* (K) para luego estimar el grupo al que pertenece cada ejemplo dado en el conjunto de datos (Russell et al., 2004; Alpaydin, 2016; Murphy, 2012).

Este proyecto, se enfoca en distintas técnicas de aprendizaje no supervisado, por lo que se va a profundizar mucho más en este, explorando distintos algoritmos y técnicas que se utilizan para su optimización y personalización.

2.2.1. Algoritmos de Aprendizaje No Supervisado

Agrupamiento *K-Means*

El algoritmo K-Means, es uno de los métodos de agrupamiento más utilizados en diversas áreas, debido a su fácil implementación y baja complejidad computacional (Xu and Wunsch, 2005). Suponiendo un conjunto de datos $X = \{x_1, ..., x_N\}, x_n \in R_d$, el cual se desea agrupar en M grupos, $C_1, ..., C_M$, de acuerdo a un criterio de agrupamiento optimizado. Comúnmente, este criterio es la suma de las distancias euclideanas al cuadrado entre cada punto x_i y el centroide m_k (centro del grupo o *cluster*) del grupo C_k que contiene x_i. Este criterio, es llamado error de agrupamiento y depende de los centros de cada grupo $m_1, ..., m_M$ (Xu and Wunsch, 2005; Likas et al., 2003; Cam and Neyman, 1967). Así, el

algoritmo K-Means se definiría:

$$E(m_1, ..., m_M) = \sum_{i=1}^{N} \sum_{k=1}^{M} I(x_i \in C_k)||x_i - m_k||_2 \tag{2.1}$$

Sin embargo, el algoritmo K-Means presenta algunos problemas, como su gran sensibilidad a la posición inicial de los centroides, obligando a realizar distintas iteraciones con posiciones aleatorias de los centroides. Así mismo, tampoco le es posible manejar datos con una gran dimensionalidad (Xu and Wunsch, 2005; Likas et al., 2003). Por lo tanto, se han planteado distintos algoritmos que solucionan estos problemas, como el propuesto *global K-Means* por (Likas et al., 2003), que constituye un método de optimización global determinista que no depende de ningún valor inicial. Específicamente, inicializa el algoritmo con $k = 1$, cuya posición óptima corresponde con el centroide del dataset. Para un valor más alto de k, se realizan N iteraciones del algoritmo K-Means, en las cuales se ubica el segundo centroide en cada posición de los puntos de datos $x_n(n = 1, ..., N)$, siendo la mejor solución escogida como la respuesta al problema de agrupamiento.

Agrupamiento Jerárquico

Este agrupamiento, como su nombre lo indica, consiste en dividir o unir un conjunto de datos de forma jerárquica en una secuencia de particiones. Estas secuencias, se dividen en dos: Aglomerativas (*bottom-up*), o divisivas (*top-down*). En el tipo de jerarquía aglomerativa, se comienza con cantidad de grupos igual a la cantidad de objetos en el conjunto de datos. Donde cada objeto, es su propio grupo. Más adelante, se comienzan a unir los objetos más cercanos hasta que se termine teniendo un solo grupo con todos los objetos dentro. En el caso de la jerarquía divisiva, se realiza el proceso contrario, donde se comienza con un único grupo que contiene la totalidad del conjunto de datos para luego ir dividiéndolo en sub-grupos (Roux, 2018).
En el agrupamiento jerárquico, existen distintos métodos de conexión entre grupos formados. Entre los más conocidos y utilizados se encuentran:

- Enlace Simple (*Simple Linkage*):

$$d(u, v) = min(dist(u[i], v[j])) \tag{2.2}$$

  Para todos los puntos i en el grupo u, y j en el grupo v. También se conoce como el algoritmo del punto más cercano (*Nearest Point Algorithm*).

- Enlace Completo (*Complete Linkage*):

$$d(u, v) = max(dist(u[i], v[j])) \tag{2.3}$$

Para todos los puntos i en el grupo u, y j en el grupo v. También se conoce como el algoritmo del punto más lejano (*Farthest Point Algorithm*) o algoritmo Voor Hees.

- Enlace Promedio (*Average Linkage*):

$$d(u, v) = \sum_{ij} \frac{d(u[i], v[j])}{(|u| * |v|)} \tag{2.4}$$

Para todos los puntos i y j de los grupos u y v, donde $|u|$ y $|v|$ son las cardinalidades. También se conoce como el algoritmo UPGMA.

- Enlace Ponderado (*Weighted Linkage*):

$$d(u, v) = (dist(s, v) + dist(t, v))/2 \tag{2.5}$$

En donde el cluster u fue formado a partir de los clusters s y t, y el cluster v es remanente en el bosque. Este algoritmo, también se conoce como WPGMA.

- Enlace Centroide (*Centroid Linkage*):

$$d(u, v) = ||c_s - c_t||_2 \tag{2.6}$$

Donde c_s y c_t son los centroides de dos grupos s y t, que se unen para formar el nuevo grupo u. El nuevo centroide se computa sobre todos los miembros originales de los grupos s y t, y la distancia utilizada sería la euclidiana entre el nuevo centroide u y el centroide v de otro grupo presente en el bosque. Esto se conoce como el algoritmo UPGMC.

- Enlace de Mediana (*Median Linkage*): Las distancias se asignan igual que en el enlace centroide (ecuación 2.6), cuando se forma un nuevo grupo u a partir de los grupos s y t, el centroide se calcula mediante el promedio de los centroides de s y t. Esto se conoce como el algoritmo WPGMC.

- Enlace de Ward (*Ward Linkage*): propuesto por Joe H. Ward en el año 1963. En cada paso, el algoritmo crea un grupo nuevo a partir de la combinación de dos grupos existentes, el cual minimiza la varianza, medida en un índice E (conocido como el índice de la suma de cuadrados):

$$d(u,v) = \sqrt{\frac{|v|+|s|}{T}d(v,s)^2 + \frac{|v|+|t|}{T}d(v,t)^2 - \frac{|v|}{T}d(s,t)^2} \tag{2.7}$$

$$T = |v| + |s| + |t| \tag{2.8}$$

Para escoger los grupos que se combinarán en cada paso, debe ser explorada cada posible combinación, en la cual será escogida la combinación que menos afecte el índice E (Ward Jr, 1963; Murtagh and Legendre, 2014).
En el caso del agrupamiento de Ward, no es necesario ingresar un número de grupos predefinidos (como k, en el caso del *K-Means*), pues su funcionamiento es aglomerativo y se debe escoger el número óptimo de grupos de acuerdo al comportamiento de la función E en cada iteración del algoritmo (Ward Jr, 1963).

Agrupamiento DBSCAN

El algoritmo de agrupamiento DBSCAN es determinista, lo que significa que siempre va a generar el mismo número de grupos (a partir del mismo conjunto de datos en el mismo orden) asumiendo los grupos como áreas de alta densidad que se separan por áreas de baja densidad de muestras. Lo cual, permite al algoritmo encontrar clusters de cualquier forma (contrario a algoritmos como K-Means). El concepto principal detrás de DBSCAN son las muestras núcleo, que básicamente son muestras en áreas de alta densidad, de manera que un conjunto de muestras núcleo conformarían un cluster.
Existen dos parámetros básicos del algoritmo *min samples*, y *eps*, los cuales definen, en otras palabras, lo que se considera denso. Básicamente, una muestra se denomina muestra núcleo mientras existan *min samples* otras muestras dentro de una distancia *eps* (Ester et al., 1996).

2.2.2. Medidas de distancia

En los algoritmos de agrupamiento, medir la distancia entre dos elementos se puede interpretar como medir la similitud que tienen entre sí. De esta manera, la similitud afecta directamente a la forma de los grupos formados, y por tanto, distintas medidas de distancia son empleadas para distintas aplicaciones. La medida de distancia seleccionada, es optimizada y utilizada para la generación de los distintos grupos, por lo que elementos del

mismo grupo se consideran con una pequeña distancia, y elementos de grupos distintos tendrán una distancia mayor (Pandit et al., 2011).

Distancia Euclideana

Es la medida de distancia por defecto que utiliza el algoritmo de agrupamiento K-Means. Consiste en la distancia ordinaria entre dos puntos. En otras palabras, es la distancia en línea recta entre dos puntos (Pandit et al., 2011). En N dimensiones, la distancia euclideana entre dos puntos p y q, donde p_i, q_i son las coordenadas de los puntos en la dimensión i, sería:

$$D_{(p,q)} = \sqrt{\sum_{l=1}^{n}(p_i - q_i)^2}; n \in N \tag{2.9}$$

Distancia Manhattan

Se le conoce por distintos nombres, como distancia rectilínea, distancia de Minkowsky, o métrica de taxi. Consiste en la distancia entre dos puntos medida a través de los ejes en ángulos rectos. En un plano con p_1 en (x_1, y_1) y p_2 en (x_2, y_2), sería $|x_1 - x_2| + |y_1 - y_2|$. Esto, es fácilmente generalizable a múltiples dimensiones (Pandit et al., 2011):

$$D_{(p,q)} = \sum_{l=1}^{n}|p_i - q_i|; n \in N \tag{2.10}$$

Distancia Hamming

Comúnmente utilizada para características discretas, la distancia de Hamming entre dos cadenas de caracteres del mismo tamaño es el número de posiciones en las cuales los símbolos correspondientes son distintos. Así, se puede decir que la distancia de Hamming refleja el número de posiciones a cambiar para convertir una cadena de caracteres en otra (Pandit et al., 2011; Russell et al., 2004).

Índice de Jaccard

También conocido como el coeficiente de similitud de Jaccard, es utilizado para comparar la similitud y diversidad de conjuntos de muestras. Se define como el tamaño de la intersección dividido por el tamaño de la unión de los dos conjuntos de muestras (Pandit

et al., 2011).

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{2.11}$$

Índice de Dado

Es también encontrado como el coeficiente de dado, está altamente relacionado con el índice de Jaccard (Pandit et al., 2011). Por ejemplo, para conjuntos de palabras clave X y Y, se define como:

$$S = \frac{2|X \cap Y|}{|X| + |Y|} \tag{2.12}$$

Índice de Coseno

Es una medida de similitud entre dos vectores de n dimensiones, que calcula el ángulo que existe entre estos. Frecuentemente utilizada en comparar documentos en minería de textos. Dados los dos vectores A y B, el índice de Coseno (θ), sería:

$$\theta = arccos \frac{A.B}{||A||||B||} \tag{2.13}$$

El valor resultante del índice de Coseno es un valor entre 0 y π, indicando exactos opuestos cuando toma el valor de π, $\pi/2$ cuando son independientes, y 0 cuando son exactamente iguales (Pandit et al., 2011).

Distancia de Gower

Es una medida de distancia muy utilizada cuando el conjunto de datos posee tanto variables cuantitativas como variables categóricas. En sí, la distancia de Gower indica que tan distintas son dos observaciones mediante un número desde 0 (son idénticas) hasta 1 (totalmente disimilares). Sin embargo, para cada tipo de dato que posean las observaciones, utiliza una métrica distinta (D'Orazio, 2021):

- Para variables binarias simétricas, basta con un coeficiente de coincidencia simple (1 si coinciden, 0 si no coinciden).
- En variables binarias asimétricas, se utiliza el índice de Jaccard.
- Cuando existen variables categóricas nominales, se codifica la variable a one hot encoding para luego aplicar el Índice de Dado.

- Para variables numéricas, se utiliza la distancia Manhattan escalada de acuerdo al rango.

Distancia Canberra

La distancia Canberra es una versión ponderada de la distancia de Manhattan (Lance and Williams, 1967), en donde:

$$D_{(x,y)} = \sum_{i=1}^{n} \frac{|x_i - y_i|}{|x_i| + |y_i|} \tag{2.14}$$

2.2.3. Cálculo de número de grupos o k

Average Silhouette Width (ASW)

El método ASW, es un índice basado en distancia que muestra la calidad de un agrupamiento. Es una intuitiva medida de calidad de grupos que no depende de asunciones estadísticas. Así, ASW es ampliamente utilizado para comparar la calidad de distintos métodos de agrupamiento. Por tanto, es posible utilizarle iterativamente para calcular el valor óptimo de clusters.

Tomando cualquier objeto i en un conjunto de datos, asignado al cluster A, donde A tiene asignados otros elementos distintos a i, podremos decir que $a(i)$ = disimilitud promedio de i a todos los demás elementos de A. Teniendo en cuenta otro cluster C distinto a A, $d(i, C)$ = disimilitud promedio de i a todos los elementos de C. Luego de calcular $d(i, C)$ para todos los clusters $C \neq A$, se selecciona el más pequeño y se denotaría $b(i)$ = mínimo $d(i, C); C \neq A$. De esta manera, el cluster seleccionado sería el vecino de i, el cual sería su grupo de no poder hacer parte del cluster A. Así, en un principio se debe asumir que el número de clusters (k) deberá ser mayor a 1. Entonces, el valor del *Silhouette Width* para i sería:

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \tag{2.15}$$

Los valores de $s(i)$ pueden resultar $-1 > s(i) > 1$, en donde si $s(i)$ es muy cercano a 1, se considera que i ha sido muy bien clasificada. Por otro lado, un valor de $s(i)$ muy cercano a 0 indica incertidumbre sobre a cual cluster debería pertenecer i. Por último, un valor de $s(i)$ cercano a -1 indica que el objeto i ha sido mal clasificado. Calculando $s(i)$ para todos los demás elementos del conjunto de datos, finalmente podremos obtener el valor de ASW:

$$\bar{S} = \frac{1}{n}\sum_{i=1}^{n} s(i) \tag{2.16}$$

De esta manera, se selecciona previamente un rango de posibles clusters $K = \{1, ..., k\}$, con los cuales se calcula $\bar{S}$ para cada valor de K. El valor de $\bar{S}$ más cercano a 1 indica el número óptimo de clusters dentro del rango K seleccionado (Batool and Hennig, 2021; Rousseeuw, 1987).

En algunos casos, encontrar el valor óptimo de $\bar{S}$ puede ser computacionalmente imposible en conjuntos de datos grandes. Razón por la cual se han propuesto alternativas como el algoritmo OSil y FOSil (*Optimum Silhouette* y *Fast Optimum Silhouette*, respectivamente) (Batool and Hennig, 2021).

Elbow

El método de Elbow, es un método con el cual se detecta el número apropiado de clusters con ayuda de un gráfico. Para la generación del gráfico, es necesario definir un conjunto de posibles clusters $k = 1, ..., n$, para luego calcular la suma de los errores cuadráticos (SSE, *Sum Squared Error*) para cada valor de k. En la figura 2.1 , se puede observar un ejemplo de una gráfica generada. El valor óptimo de k, es en el cual el valor de SSE caiga con forma de "codo", en el caso de la figura 2.1 , sería el valor de $k = 5$ (Nainggolan et al., 2019; Humaira and Rasyidah, 2020).

Para el cálculo de SSE se debe:

$$SSE = \sum_{i=1}^{n} (d)^2 \tag{2.17}$$

Donde d es la distancia entre el elemento y el centro del cluster (Nainggolan et al., 2019).

2.2.4. Reducción de Dimensionalidad

Principal Component Analysis

El análisis del componente principal (PCA, o *Principal Component Analysis*), es una herramienta que se utiliza principalmente para la reducción de dimensionalidad, pues permite reflejar el comportamiento de distintas variables correlacionadas en una sola, manteniendo el tamaño de la variación de todas ellas mediante el uso de pesos o *weights*. Así mismo, es necesario que todos los datos de entrada estén normalizados, pues de no ser así, variables de gran tamaño van a provocar que el modelo PCA solo se enfoque en valores

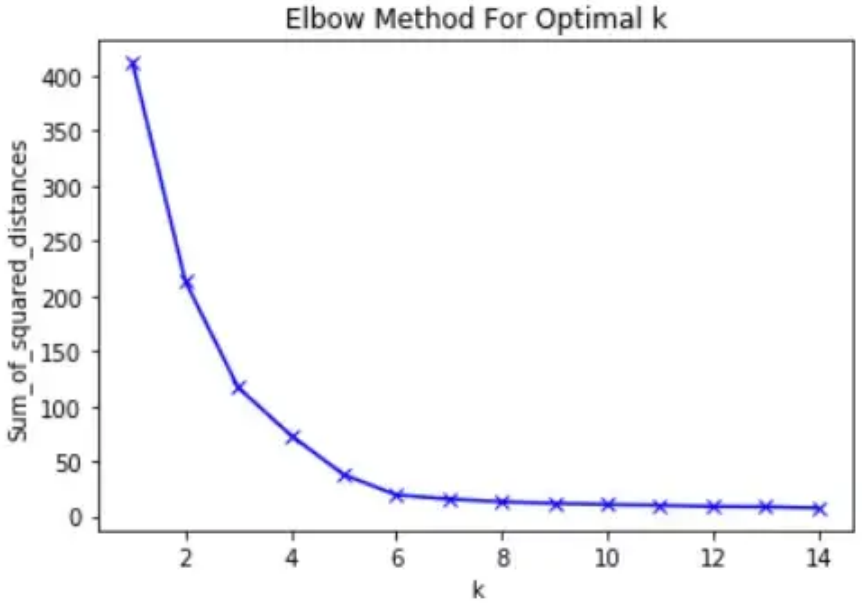


Figura 2.1: Suma de los errores cuadráticos contra k para la obtención del valor óptimo de k.

grandes. Esta normalización, indica que a cada variable se le sustrae el promedio y se divide por su desviación estándar.

El promedio ponderado que se utiliza para calcular el primer componente principal, no es más que una combinación lineal de las variables. En un conjunto de datos de una matríz X con i filas ($i = 1, ..., I$) (usualmente objetos/muestras) y j columnas ($j = 1, ..., J$) (usualmente características/variables) de tamaño $I * J$, cada variable individual de X se denota como $x_j (j = 1, ..., J)$ y son todas vectores en un espacio I-dimensional. Una combinación lineal de esas variables de x, se puede escribir: $t = w_1 x_1 + ... + w_J x_J$, donde t ahora es un vector en el mismo espacio que x. En notación de matrices, sería $t = Xw$ con w siendo el vector con los elementos $w_j (j = 1, ..., J)$. Ya que X posee toda la variación relevante, se desea tener la máxima variación posible también en t, y de ser apreciable, t funcionaría como una representación de las variables x. Cabe resaltar que los pesos presentes en w, también deben ser normalizados, pues multiplicando un valor óptimo de w con un número arbitrariamente grande, hará la varianza arbitrariamente grande. Esta normalización requiere que su SSE, sea siempre de 1.

Calculando qué tan representativo es t en términos de reemplazar X se obtiene una evaluación de las capacidades de t. Proyectando las columnas de X en t y calculando los residuos de esa proyección, donde p serían los coeficientes de regresión, y E los residuos.

$$X = tp^T + E \tag{2.18}$$

Normalmente, p iguala w, por lo que toda la regresión se puede utilizar para juzgar la calidad de t, en términos del porcentaje de variación explicada por t.

$$\frac{||X||^2 - ||E||^2}{||X||^2} 100\,\% \tag{2.19}$$

En ocasiones, el primer componente principal no describe la varianza como se espera, por lo que puede ser necesario tener componentes adicionales. Los vectores de p son ortogonales a otros vectores, los cuales tienen una participación independiente en la varianza total que todos los componentes explican (Bro and Smilde, 2014).

$$||X||^2 = ||t_1 p_1^T||^2 + ...||t_R p_R^T||^2 + ||E||^2 \tag{2.20}$$

Existen trabajos donde también se utiliza el análisis del componente principal como método de agrupamiento. Es el caso de (Pes et al., 2016), donde se preseleccionaron once características a las cuales se practicó PCA, dejando 4 componentes interpretables como grupos.

Unsupervised Feature Selection

Es un método diseñado especialmente para la selección de características en problemas de índole no supervisada basado en el agrupamiento espectral. Esto, consiste en 3 pasos (Solorio-Fernández et al., 2020):

- Análisis espectral: Este análisis se realiza para detectar la estructura de los grupos en los datos.
- *Coefficient Learning*: A través de los primeros k vectores propios de la matríz laplaciana, se mide la importancia de cada característica con un modelo de regresión con regularización $l_1 - norm$.
- Selección de características: Se seleccionan d características basándose en los valores absolutos más altos de los coeficientes obtenidos en el paso anterior.

Este método, tiene un solo parámetro, el cual es p, o la cantidad de vecinos más cercanos a tener en cuenta para la construcción del gráfico. La selección de características no su-

pervisada puede utilizarse al mismo tiempo como método de agrupamiento y reductor de dimensionalidad, como se menciona en (Cai et al., 2010).

Uniform Manifold Approximation and Projection

Es una técnica de reducción de dimensionalidad especializada en detectar estructuras no lineales en los datos (Comúnmente conocido como *Manifold Learning*) a diferencia de otras conocidas como PCA, que no presentan mucha sensibilidad a tales no-linealidades. UMAP (*Uniform Manifold Approximation and Projection*), está construido basado en la geometría Riemanniana y una topología algebraica, y es un algoritmo que permite ya sea pre-procesar o visualizar conjuntos de datos en aprendizaje automático.
El algoritmo trabaja en términos de conjuntos simpliciales difusos, que computacionalmente se traduciría en un gráfico ponderado. Por lo tanto, UMAP se clasificaría como un algoritmo de aprendizaje por medio de gráficos basado en *K-neighbours*. El proceso se podría dividir en dos fases distintas, cuya primera parte se enfoca en construir un gráfico ponderado basado en *K-neighbours*. Y la siguiente fase, computa una disposición de baja dimensionalidad de el gráfico construido. Sin embargo, el algoritmo UMAP realiza tres asunciones sobre los datos, sobre las cuales es posible modelar el conjunto de datos con una estructura topológica difusa:

- Los datos están distribuidos uniformemente en una colección Riemanniana.
- La métrica Riemanniana es localmente constante (o se puede considerar como tal).
- La colección está localmente conectada.

El software de UMAP está disponible para utilizarse con Python, y fue diseñado para ser compatible con scikit-learn, permitiéndole ser integrado a *pipelines* (elementos de procesamiento de datos ordenados de forma que la salida de uno es la entrada del siguiente) de sklearn(McInnes et al., 2018), siendo también dependiente de sus dependencias, como numpy y scipy. Mayormente, se utiliza como un reemplazo de t-SNE (otra herramienta de reducción de dimensionalidad),
Existen casos, como el de (Bej et al., 2022), en los cuales se utiliza UMAP como una técnica de agrupamiento (en conjunto con otras técnicas). Esta técnica presenta resultados efectivos en conjuntos de datos particularmente diversos, como por ejemplo, epidemiológicos.

2.3. Estado del Arte

En esta sección, se expondrán los principales trabajos previos que abordan el diagnóstico y la evaluación de la diabetes mellitus. Numerosos estudios se han realizado al rededor de la temática, siendo al rededor del 65 % implementaciones de algoritmos de aprendizaje supervisado, 25 % de una combinación de aprendizaje supervisado con no supervisado, y por último, un 10 % de aprendizaje no supervisado (Chauhan et al., 2021). De tal manera, se explorarán las más recientes soluciones existentes referentes al aprendizaje no supervisado.

En primer lugar, los autores en (Sarría-Santamera et al., 2020), muestran una división en grupos que representan los sub-tipos de la diabetes mellitus. El artículo es una recopilación de varios estudios con el mismo propósito, en donde se utilizan distintas técnicas para dividir los pacientes, también basándose en distintos criterios (como los niveles de HbA1c (hemoglobina glicosilada), GADA (Anticuerpos Anti-Glutamato Decarboxilasa), edad de diagnóstico o los síntomas reportados por los pacientes). Entre los métodos utilizados para determinar el número adecuado de grupos, están los métodos de Silhouette, Ward, y análisis del componente principal (PCA, o *Principal Component Analysis*). Ahora, el modelo más utilizado para la creación de los clusters fue el K-Means, usado en al rededor de la mitad de los estudios revisados. Los resultados mostrados indican cinco estudios que encontraron los mismos grupos: Diabetes autoinmune severa, diabetes insulinodeficiente severa, diabetes insulinoresistente, diabetes relacionada con obesidad leve, diabetes relacionada con edad temprana (Ahlqvist et al., 2018; Zaharia et al., 2019; Dennis et al., 2019; Safai et al., 2018; Ahlqvist et al., 2017). Además, otros estudios encontraron los mismos grupos exceptuando la diabetes autoinmune severa (Kahkoska et al., 2020; Zou et al., 2019). Cabe destacar, que los grupos con resultados similares, a pesar de tener distintos conjuntos de datos seleccionaron casi las mismas variables para el análisis de grupos, como el índice de masa corporal (del inglés *Body Mass Index*) y HOMA (del inglés *Homeostasis Model Assessment*, entre otras). Estas variables, fueron seleccionadas debido a su relevancia para el manejo de la diabetes(Safai et al., 2018). Además de técnicas comunes, también se revisaron técnicas de reducción de dimensionalidad utilizadas como algoritmos de agrupamiento. Tal es el caso de (Pes et al., 2016), quienes con once características, realizaron un agrupamiento de pacientes de diabetes latente autoinmune en la adultez (LADA, del inglés *Latent Autoimmune Diabetes in Adults*). Mediante un umbral de car-

ga factorial, se consideraron las variables representativas de cada componente principal. Así mismo, los pacientes se agruparon de acuerdo a su componente principal dominante y las curvas Kaplan-Meier relacionadas con la progresión de su insulinodependencia. Los distintos componentes principales resultantes describen características de la enfermedad, siendo el primer componente la característica metabólica, reflejando la acumulación de tejido adiposo que ocurre en LADA, que luego exacerba a insulinoresistencia. El segundo componente, estima la influencia de variantes genéticas. Los dos componentes principales restantes, colectan la variabilidad remanente asociada a otros parámetros metabólicos, como colesterol y triglicéridos. Luego de una interpretación más detallada de todos los estudios, se concluye que el análisis de agrupamiento puede ser aplicado para monitorear el progreso y control de la enfermedad en los pacientes. Mediante los grupos generados, se podrían aplicar medidas terapéuticas y preventivas, y, a su vez, encontrar con mayor facilidad los pacientes que se beneficiarían más de estas medidas.

En el siguiente estudio (Carrillo-Larco et al., 2021), los autores se propusieron separar una población de pacientes de diabetes mellitus mediante algoritmos de agrupamiento (del inglés *clustering*). En esta ocasión, entre los datos disponibles en el conjunto de datos, utilizaron variables como la edad, sexo, índice de masa corporal, historial de diabetes mellitus en la familia, entre otros, para el agrupamiento. Al haber analizado conjuntos de datos de más de trece países distintos, fue necesario el análisis del componente principal como reductor de dimensionalidad. El algoritmo utilizado para el agrupamiento fue el K-Means. De tal forma, se realizó primero el agrupamiento en cada conjunto de datos de cada país, para así encontrar diferencias con el agrupamiento global. la distancia euclidiana, el método de Elbow y Silhouette, fueron utilizados para la selección del número adecuado de grupos. Luego, mediante el coeficiente de Jaccard, se evaluó la estabilidad de cada grupo. Como resultado, se obtuvieron cuatro distintos grupos donde predominan algunos valores de las características usadas. Así, se concluyó que los grupos pueden separar los pacientes en 4 distintos perfiles que pueden ayudar a identificar una etapa temprana de la enfermedad o factores de riesgo no encontrados en el pasado sobre la enfermedad en la población en general. Igualmente, pacientes en distintos grupos pueden necesitar un tratamiento o prevención a medida.

En el caso de (Nedyalkova et al., 2021), se utilizó un conjunto de datos de 51 pacientes de diabetes mellitus tipo 2, en el cual, se contaba con al rededor de 26 características por cada paciente. En promedio, cada paciente tenía 3 variables con valores nulos, razón

por la cual, se dependía de 23 características para la descripción de cada paciente. La descripción del paciente consistía en clasificar con alta precisión pacientes de diabetes tipo 2 con ciertas enfermedades subyacentes (pudiendo ser desde 1 hasta 4 de las más comunes: Hipertonía Arterial, Isquémia Cardiaca, Polineuropatía Diabética, y Microangiopatía Diabética) con los descriptores clínicos más adecuados. Los autores, propusieron una variante del algoritmo K-Means que llamaron K-Means combinatorio, creado para encontrar los casos de separación utilizando solamente una pequeña porción de las características disponibles. El algoritmo propuesto, puede trabajar con distintos métodos de unión, como promedio, centroide, o Ward, sin embargo, se escogió el método del centroide para todos los agrupamientos realizados. El funcionamiento del K-Means combinatorio se basa en realizar todos los grupos posibles utilizando solamente 3 descriptores de los disponibles en el conjunto de datos. Luego, se califican los agrupamientos utilizando la varianza global. Así mismo, se realizó un agrupamiento jerárquico mediante el método de Ward, el cual, requirió normalización (con *z-standarization*), determinación de la distancia Euclidiana, graficar los grupos en un dendograma, y seleccionar los grupos que sean significantes estadísticamente (mediante el criterio de Sneath). En conclusión, el K-Means combinatorio muestra la posibilidad de clasificar correctamente pacientes con enfermedades subyacentes incluso sin conocer todas las variables. Así mismo, el agrupamiento jerárquico proporcionó una determinación preliminar del número apropiado de clusters para este estudio.

Además de la extracción de características, el algoritmo K-Means, también es utilizado para la detección de la enfermedad. En el siguiente estudio (Raihan et al., 2019), los autores utilizan métodos de agrupamiento, como el K-Means y el agrupamiento jerárquico (del inglés *Hierarchical Clustering*) para mejorar la detección de la enfermedad en la ciudad de Bangladesh. A pesar de tener un conjunto de datos reducido y con limitaciones, lograron hacer un análisis adecuado, concluyendo que los algoritmos no supervisados pueden utilizarse para mejorar los métodos de detección actuales. Similares estudios, como el de (Mujumdar and Vaidehi, 2019), también utilizan el algoritmo K-Means para la clasificación de pacientes que tienen o no diabetes mellitus. Utilizando un número reducido de características, siendo los niveles de glucosa, la presión arterial, el espesor de la piel, el índice de masa corporal, y la edad, las variables más importantes.

Así mismo, el algoritmo K-Means es utilizado como reductor de dimensionalidad o etapa de preprocesamiento en detectores supervisados a partir de máquinas de vectores de soporte o árboles de decisión que predicen la diabetes mellitus. Este proceso, es realizado en dos

etapas distintas. La primera etapa, remueve inconsistencias con una variante jerárquica del algoritmo K-Means (llamada k*-Means), que inicia el algoritmo K-Means con un valor de "K" más alto, para luego hacer un refinamiento con el algoritmo K-Means convencional. La segunda etapa, comprende el uso de un algoritmo genético para la selección de características. (Bhatia and Syal, 2017; Chen et al., 2017; Qi et al., 2016).

Continuando, el estudio realizado por (Ojugo and Otakore, 2018), muestra la comparación de modelos supervisados y no supervisados utilizados con el fin de mejorar la detección temprana de diabetes mellitus en Nigeria, sobretodo, en el metamorfismo consecuente de la diabetes gestacional en madres. El objetivo de los algoritmos a entrenar fue realizar una correcta clasificación de pacientes que tienen o no un diagnóstico positivo para diabetes. Entre los modelos utilizados en este estudio se encuentran el análisis lineal discriminatorio (*Linear Discriminant Analysis*), análisis cuadrático discriminatorio, *K-Nearest Neighbour* (K-NN), Máquinas de vectores de soporte, y por último, una red neuronal entrenada por un algoritmo genético difuso (no supervisado). El último mencionado, según los resultados mostrados, ha superado a todos los demás modelos en dos métricas definidas para poder comparar los métodos supervisados con los no supervisados (porcentaje de mejora y porcentaje de error en la clasificación).

En el siguiente estudio (Cho et al., 2019), se presenta la división de la diabetes mellitus tipo 2 en 6 distintos subgrupos basándose en los principales factores de riesgo conocidos de la enfermedad (Edad, sexo, índice de masa corporal, hipertensión, historial de diabetes en la familia). Para esto, en un conjunto de datos inicial, denominado *Discovery data*, se utiliza agrupamiento jerárquico junto con la distancia de Gower como medida de similitud, pues el conjunto de datos contiene variables tanto categóricas como continuas. El número de grupos a crear fue obtenido mediante la función *Cutree* en el programa R. Fueron 6 grupos resultantes con valores distintos de prevalencia de diabetes mellitus tipo 2 (desde 0.09 hasta 0.44). La reproducibilidad de los grupos fue probada con otros 3 cohortes distintos *(HEXA, CAVAS, KNHANES)*. Por último, fue aplicado un modelo basado en máquinas de vectores de soporte para predecir la pertenencia a cualquiera de los 6 grupos definidos previamente basándose en los 5 factores de riesgo. El entrenamiento del modelo se realizó con el conjunto de datos *Discovery data*, para luego validar la clasificación con los 3 cohortes restantes. Las diferencias entre la prevalencia de la enfermedad a través de los distintos grupos fue altamente reproducible en todos los conjuntos de datos utilizados. Los resultados obtenidos mostraron las diferencias en el desarrollo de la enfermedad. A pesar

de pequeñas discrepancias con la distribución de los factores de riesgo dentro de los grupos en *Discovery data* y el conjunto de validación, la tendencia en general fue consistente.

Es posible utilizar técnicas de reducción de dimensionalidad para realizar agrupamientos, como en el caso de (Bej et al., 2022), quienes proponen el uso de UMAP (*Uniform Manifold Approximation and Projection*) para separar grupos distintos de pacientes de diabetes mellitus tipo 2 basándose en información epidemiológica (información dietaria, historial de adicciones, patrones socio-económicos y de estilo de vida, etc.). Para conseguir grupos relevantes, se tuvo que diseñar un flujo de trabajo agrupando de forma distribuida variables continuas, ordinales, y nominales de forma separada. Esto, combinando distintas configuraciones de similitud de UMAP. Integrando las dimensiones reducidas de cada tipo de característica, se obtuvieron 4 grupos diferentes mediante un algoritmo de agrupamiento basado en densidad conocido como DBSCAN. Dos de los grupos, representan pacientes no obesos de diabetes mellitus tipo 2.

Capítulo 3

Objetivos y Metodología de trabajo

3.1. Objetivos

3.1.1. Objetivo General

Extracción de características y comparar y evaluar distintas técnicas de aprendizaje no supervisado para el agrupamiento de un conjunto de datos de pacientes de diabetes mellitus tipo II con prevalescencia en características nominales.

3.1.2. Objetivos Específicos

- Realizar un análisis exploratorio del conjunto de datos para comprender la distribución y naturaleza de las variables nominales.

- Generar un flujo de datos que permita el pre-procesamiento necesario para el ingreso de los datos a los modelos a comparar.

- Implementación y evaluación de las diferentes técnicas de clustering con los datos pre-procesados.

- Interpretar y evaluar los agrupamientos resultantes de cada algoritmo seleccionado.

3.2. Metodología CRISP-DM

En (Chapman et al., 2000), los autores describen la metodología CRISP-DM (*Cross-Industry Standard Process for Data Mining*) como un proceso jerárquico, donde hay grupos de tareas con hasta cuatro niveles de abstracción o profundidad. Cada nivel es representado de la siguiente manera, desde lo general hacia lo específico: Fase, tarea genérica, tarea especializada, e instancia de proceso. Así, cada fase estaría compuesta de varias tareas genéricas, las cuales deben ser lo suficientemente generales como para cubrir todas las situaciones posibles en la minería de datos. En el siguiente nivel de abstracción, las tareas especializadas indican cómo se deberían realizar las tareas genéricas en el caso específico que se aborda. Por último, la instancia de proceso, es un registro de las acciones, decisiones y resultados.

En la metodología CRISP-DM, se propone un modelo de referencia que cubre todos los posibles problemas de minería de datos. Allí, se describen las fases recomendadas, las cuales no tienen una secuencia rígida y requieren el movimiento entre fases continuamente. Las fases recomendadas son las siguientes:

- Entendimiento del negocio: Básicamente, se definen los requisitos y objetivos del proyecto de acuerdo a la necesidad existente. Luego, transformar este conocimiento en un problema de minería de datos.
- Entendimiento de los datos: Es la recolección inicial de los datos, junto con procesos de identificación de calidad de datos, características, o hipótesis que ayudarán en etapas futuras.
- Preparación de los datos: Son todas las actividades necesarias para transformar los datos recolectados en el conjunto de datos que va a alimentar las herramientas de modelado.
- Modelado: la aplicación de todas las técnicas de modelado, junto con una calibración de sus parámetros a valores óptimos.
- Evaluación: Aquí, se evalúa el comportamiento de la etapa anterior, donde se decide si cumple con los objetivos planteados en la primera fase.
- Despliegue: Esta etapa puede involucrar únicamente la presentación del conocimiento adquirido, o actividades más complejas de mantenimiento.

De esta manera, en (Rodriguez León, 2016), se presenta la adecuación de las fases del proceso de minería de datos para problemas de clasificación no supervisada. En la figura 3.1, se encuentra el gráfico que describe las distintas fases adaptadas, junto con las actividades relevantes para el desarrollo del proyecto. Así mismo, en conjunto con las actividades propuestas, en la figura 3.2, se muestra el diagrama de Gantt del desarrollo del proyecto adoptando la metodología propuesta.

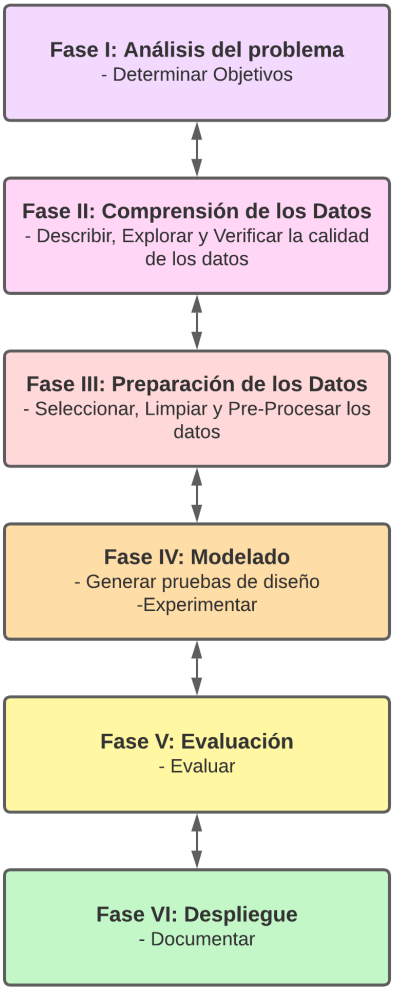


Figura 3.1: Metodología CRISP-DM Adaptada a problemas de clasificación no supervisados

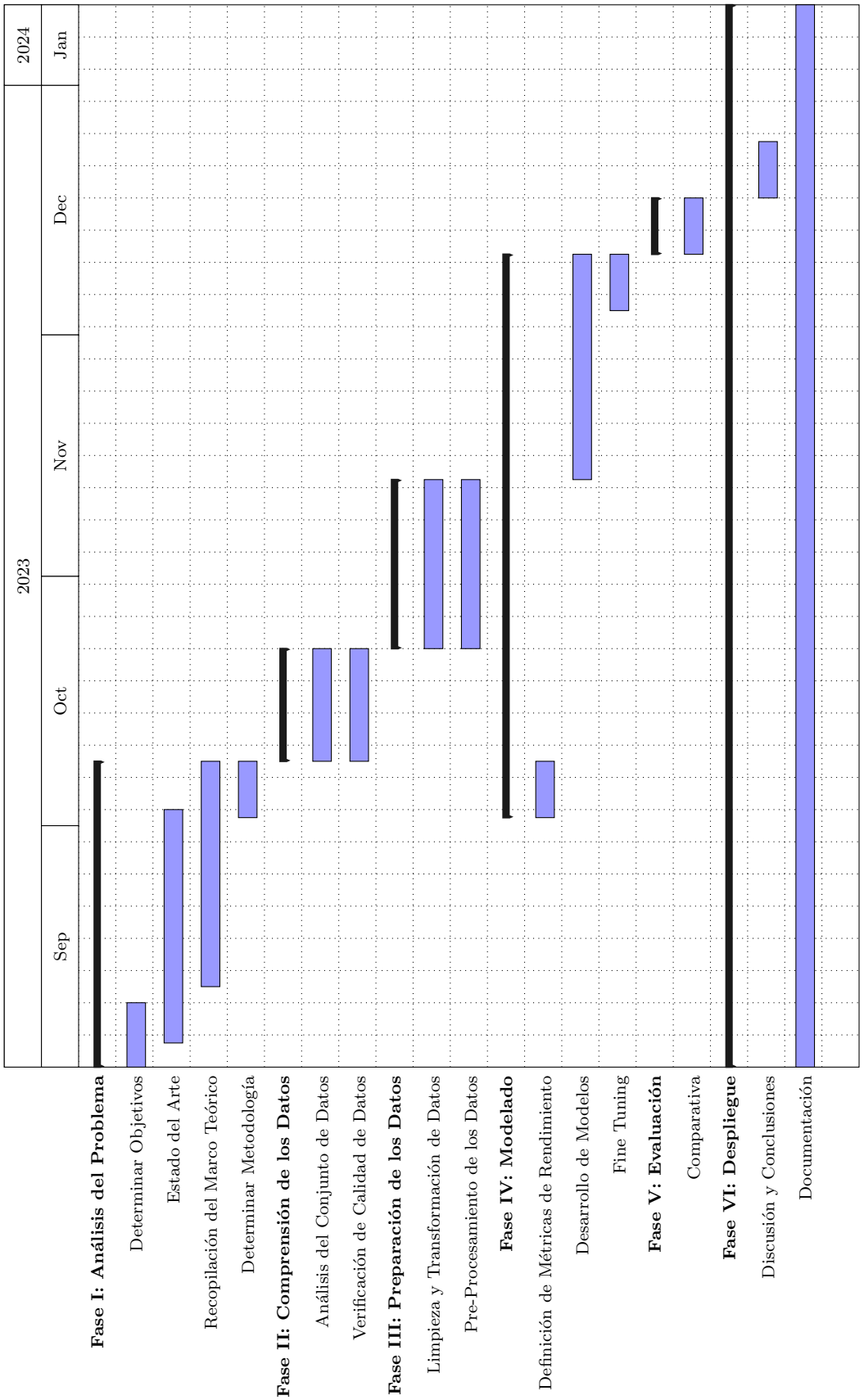


Figura 3.2: Diagrama de Gantt del proyecto

3.2.1. Fase I: Análisis del Problema

En esta fase se deben definir en los objetivos, teniendo en cuenta los beneficios que podrían generar la información obtenida de la extracción de características del conjunto de datos. Por esta razón, es bastante necesario el diálogo con un experto en Diabetes Mellitus, pues, ayudará en la interpretación de grupos resultantes y la determinación de qué se debe tener para cumplir con los objetivos.
Según el estado del arte analizado en la sección correspondiente, el número aproximado de grupos o clusters que se han obtenido con conjuntos de datos similares es de cuatro, claramente, con distintos acercamientos e interpretaciones.

3.2.2. Fase II: Comprensión de los Datos

En esta sección, se realizará un análisis exploratorio al conjunto de datos seleccionado, apoyándose en el trabajo de (Strack et al., 2014). En el cual, se describirán a detalle todas las variables del conjunto de datos con el propósito de seleccionar las más relevantes para cada uno de los distintos modelos de agrupamiento que se compararán. Por último, se verificará la calidad de los datos, comprobando que no tengan errores mediante herramientas de visualización y distintas representaciones (Histogramas, gráficos de dispersión, etc).

3.2.3. Fase III: Preparación de los Datos

Así mismo, la exploración de los datos indicará la calidad de cada una de las variables del conjunto de datos, y de ser necesario, se realizarán actividades de limpieza de valores nulos, transformación, y pre-procesamiento de los datos para que sean aptos de ingresar a cada uno de los modelos a comparar. Este pre-procesamiento, puede ir desde un cambio de sintaxis hasta un cambio completo de representación, incluyendo normalización o nivelación de algunas o todas las características encontradas en el conjunto de datos. Cabe mencionar que la etapa de pre-procesamiento puede repetirse varias veces, pues cada modelo puede necesitar que los datos sean ingresados en distintas representaciones.

3.2.4. Fase IV: Modelado

Esta fase comprende la aplicación de todos los modelos seleccionados al conjunto de datos ya tratado. Aquí, se definirán métricas de evaluación para comparar preliminarmente todos los modelos y poder validarlos para obtener los mejores resultados posibles con

cada uno. Puede ser necesario regresar a la etapa de preparación, pues algunas técnicas pueden tener requisitos de distintas representaciones de datos. Así mismo, de requerirse, se contempla el entrenamiento y *fine tunning* de los modelos seleccionados.

3.2.5. Fase V: Evaluación

Una vez se aplicaron todos los modelos en la fase anterior, se procede a compararlos de acuerdo a las métricas definidas previamente, y junto con un experto, se corrobora la utilidad de la información extraída por cada uno de los modelos. Así, no solo se comparan por una métrica en concreto sino también de la calidad cualitativa del agrupamiento, claramente teniendo en cuenta los resultados obtenidos por ese mismo modelo en el estado del arte.

3.2.6. Fase VI: Despliegue

Por último, comparados todos los modelos probados, se procede a realizar una lista ordenada de los resultados obtenidos, presentando la información relevante de cada agrupamiento y su lugar en el ranking. Así mismo, se documenta cada etapa del proceso junto con las conclusiones obtenidas durante el mismo y los resultados mostrados.

Capítulo 4

Planteamiento de la Comparativa

4.1. Conjunto de datos

El conjunto de datos a utilizar fue obtenido en el repositorio de machine learning de la universidad de California, Irvine, escuela de información y ciencias de la computación *UCI Machine Learning Repository* y fue compartido por el Centro de investigación clínica y traslacional (*Center for Clinical and Translational Research*) en conjunto con la universidad Commonwealth en Virginia Dua and Graff (2017).

Este conjunto de datos, fue resultado de una recopilación y pre-procesamiento realizado en Strack et al. (2014), donde inicialmente se contaba con una base de datos de 41 tablas con 117 características, correspondientes a 74'036.643 de visitas y 17'880,231 pacientes únicos. En primera instancia, encuentros de interés fueron seleccionados de la base de datos con 55 atributos. Seguido a esto, se aplicó análisis y pre-procesamiento buscando que los datos contengan suficiente información. Solamente los encuentros médicos que cumpliesen los siguientes criterios harían parte del conjunto de datos:

- Es un encuentro con un paciente hospitalizado (una admisión hospitalaria).
- Es un encuentro diabético, en el cual, cualquier tipo de diabetes sea ingresada en el sistema como un diagnóstico.
- La duración de la estadía fue como mínimo 1 día, y como máximo 14.
- Durante el encuentro se realizaron test de laboratorio.
- Durante el encuentro se suministraron medicamentos.

Como resultado, se obtuvo un conjunto de datos de 101.766 encuentros que cumplen con los criterios descritos. Así mismo, expertos clínicos realizaron la selección de características. Actualmente, el conjunto de datos cuenta con 50 columnas que se describen a detalle en la tabla 4.1.

Tabla 4.1: Descripción de cada característica del conjunto de datos. La tabla indica el nombre de la característica, su naturaleza (numérico u ordinal), una breve explicación de ella, y el porcentaje de datos faltantes que hay en la misma.(Strack et al., 2014)

Característica	**Tipo**	**Descripción**	**% Miss**
Encounter ID	Numérico	Identificador único de un encuentro hospitalario.	0 %
Patient Number	Numérico	Identificador único de un paciente.	0 %
Race	Nominal	Raza, valores: Caucasian, Asian, African American, Hispanic, y Other.	2 %
Gender	Nominal	Género, valores: male, female, y unknown/invalid.	0 %
Age	Nominal	Edades agrupadas en intervalos de 10 años: [0, 10), [10, 20), ..., [90, 100).	0 %
Weight	Numérico	Peso en libras.	97 %
Admission Type	Nominal	Número entero correspondiente a 8 distintos casos.	0 %
Discharge Disposition	Nominal	Número entero correspondiente a 29 distintos casos.	0 %
Admission Source	Nominal	Número entero correspondiente a 21 distintos casos.	0 %
Time in hospital	Numérico	Número de días entre la admisión y la descarga.	0 %
Payer code	Nominal	Número entero correspondiente a 23 distintos casos..	40 %
Medical Specialty	Nominal	Número entero correspondiente a 84 distintos casos dependiendo de la especialidad de quien realiza la admisión.	49 %

Tabla 4.1: (Continuación)

Característica	**Tipo**	**Descripción**	**% Miss**
Number of lab procedures	Número	Número de procedimientos de laboratorio realizados durante el encuentro.	0 %
Number of procedures	Número	Número de procedimientos (fuera de laboratorios) realizados durante el encuentro.	0 %
Number of medications	Número	Número de medicamentos genéricos administrados durante el encuentro.	0 %
Number of outpatient visits	Número	Número de visitas a consulta externa durante el último año anterior al encuentro.	0 %
Number of emergency visits	Número	Número de visitas a emergencias durante el último año anterior al encuentro.	0 %
Number of inpatient visits	Número	Número de hospitalizaciones durante el último año anterior al encuentro.	0 %
Diagnosis 1	Nominal	Diagnóstico primario (codificado como los tres primeros dígitos del código ICD9); 848 valores distintos.	>0.1 %
Diagnosis 2	Nominal	Diagnóstico secundario (codificado como los tres primeros dígitos del código ICD9); 923 valores distintos.	0.4 %
Diagnosis 3	Nominal	Diagnóstico secundario adicional (codificado como los tres primeros dígitos del código ICD9); 954 valores distintos.	1 %
Number of Diagnoses	Numérico	Número de diagnósticos ingresados al sistema.	0 %
Glucose serum test	Nominal	Rango del resultado del test de glucosa plasmática en la sangre, también indica si no fue tomado. Valores: “>200”, “>300”, “normal”, y “none” si no se midió.	0 %

Tabla 4.1: (Continuación)

Característica	Tipo	Descripción	% Miss
A1c test result	Nominal	Rango del resultado del test de hemoglobina A1c, también indica si no fue tomado. Valores: “>8 %”, “>7 %”, “normal”, y “none” si no se midió.	0 %
Change of medications	Nominal	Indica si hubo algún cambio en las medicaciones diabéticas, ya sea dosis o nombre genérico. Valores: “change” y “no change”.	0 %
Diabetes medications	Nominal	Indica si se prescribió algún medicamento diabético. Valores: “Yes” y “No”.	0 %
24 Columnas para medicamentos	Nominal	Indica si se prescribió el medicamento y si hubo cambio en la dosis. Valores: “no”, “steady” si la dosis no cambió, “up” si la dosis subió, y “down” si la osis bajó. En la tabla, se pueden encontrar cada una de estos medicamentos.	0 %
Readmitted	Nominal	Días que tardó la readmisión del paciente. Valores: “>30” si se readmitió en más de 30 días, “<30” si se readmitió en menos de 30 días, y “No” si no hay registro de readmisión.	0 %

Adicionalmente, y como se puede apreciar en la tabla 4.1, las variables categóricas ya han sido tratadas previamente añadiendo una codificación a cada uno de sus valores posibles. Así, en las tablas 4.2, 4.3, y 4.4, se puede encontrar el significado de cada valor a mayor detalle. También, se puede observar que existen algunas categorías faltantes (algún número faltante en la cronología de cada valor). Esto, debido a que se han removido categorías que, a pesar de estar estipuladas en la documentación del conjunto de datos, no están presentes en el mismo. Por ejemplo, en la tabla 4.4, no se encuentra la descripción para el valor “12”, pues ningún paciente presenta este valor.

Tabla 4.2: Posibles valores de la variable "Admission Type".

Valor	**Descripción**
1	Emergencia.
2	Urgencia.
3	Elección.
4	Recién Nacido.
5	No disponible.
6	Nulo.
7	Centro de Traumas.
8	No investigado.

Así como se realizó una codificación para cada valor categórico en el conjunto de datos, se utilizó el carácter "?" para representar un dato faltante en la base de datos. Así, el porcentaje representado en la tabla 4.1 para los valores faltantes hace referencia a los valores que entran en esta clase, y no a datos faltantes o nulos en el conjunto de datos. Esto, acelera en gran parte el pre-procesamiento de los datos, pues no se requiere ninguna técnica específica para el rellenado de valores nulos.

Tabla 4.3: Posibles valores de la variable "Discharge Disposition".

Valor	**Descripción**
1	Dado de alta a casa.
2	Dado de alta o transferido a otro hospital de corto plazo.
3	Dado de alta o transferido a un hospital especializado (*SNF, Skilled Nursing Facility*).
4	Dado de alta o transferido a un hospital intermedio (*ICF, Intermediate Care Facility*).
5	Dado de alta o transferido a otro tipo de institución de hospitalización.
6	Dado de alta o transferido a casa con servicio de salud.
7	Dado de alta en contra de recomendación médica (*AMA, Against Medical Advice*).

Tabla 4.3: (Continuación).

Valor	**Descripción**
8	Dado de alta o transferido a casa bajo cuidado de medicamento intravenoso autoaplicable (*IV, "In the Vein"*).
9	Admitido como interno en este hospital.
10	Neonato dado de alta a otro hospital para cuidado neonatal.
11	Fallecido.
12	Sigue siendo paciente o se espera regreso para servicios de consulta externa.
13	Cuidados de hospicio en casa.
14	Cuidados de hospicio en una institución médica.
15	Dado de alta o transferido dentro de esta institución a una cama aprobada por el seguro médico (*Medicare*).
16	Dado de alta, transferido, o referido a otra institución médica para servicios de consulta externa.
17	Dado de alta, transferido, o referido a esta institución médica para servicios de consulta externa.
18	Nulo
19	Fallecido en casa. Servicio subsidiado (*Medicaid*) solamente, hospicio.
20	Fallecido en un centro médico. Servicio subsidiado (*Medicaid*) solamente, hospicio.
22	Dado de alta o transferido a otra institución de rehabilitación incluyendo las unidades de un hospital.
23	Dado de alta o transferido a un hospital de cuidados a largo plazo.
24	Dado de alta o transferido a una institución de cuidados certificada por el servicio subsidiado (*Medicaid*), pero no por el seguro médico (*Medicare*).
25	No investigado.
27	Dado de alta o transferido a una institución médica federal.
28	Dado de alta o transferido a un hospital psiquiátrico o la unidad psiquiátrica de un hospital.

Tabla 4.4: Posibles valores de la variable "Admission Source".

Valor	**Descripción**
1	Remisión médica.
2	Remisión clínica.
3	Remisión por seguro médico (*HMO, Health Maintenance Organization*).
4	Transferencia desde un hospital.
5	Transferencia desde un hospital especializado (*SNF, Skilled Nursing Facility*).
6	Transferencia dese otra institución de cuidado médico.
7	Sala de emergencias.
8	Corte o cumplimiento de la ley.
9	No disponible.
10	Transferencia desde un hospital de acceso crítico (*CAH, Critical Access Hospital*).
11	Nacimiento normal.
13	Bebé enfermo.
14	Nacimiento extra-mural.
17	Nulo.
20	No investigado.
22	Transferencia desde hospitalización o la misma institución resultando en otra petición.
25	Transferencia desde un centro de servicio de cirugía ambulatoria.

4.2. Algoritmos a comparar

En esta sección, se describirán las implementaciones de distintos tipos de algoritmos de aprendizaje no supervisado que se llevaron a cabo en el presente estudio. Cabe destacar que los algoritmos mencionados aquí se utilizaron como una base, y se buscaron optimizar al máximo mediante cambios de parámetros y técnicas dentro de su proceso.

4.2.1. Agrupamiento Jerárquico

En (Cho et al., 2019), se ha desarrollado una versión del algoritmo especializada para datos con alta presencia de variables nominales, la cual consta de un agrupamiento jerárquico, cuya medida de distancia es la de Gower debido a su manejo de variables categóricas y numéricas. El método de enlace entre grupos es el enlace completo. En la implementación original de (Cho et al., 2019), el agrupamiento jerárquico fue utilizado como un identificador de clases dentro de un conjunto de datos similar al presente en este proyecto pero con un tamaño reducido, cuyos resultados fueron fácilmente reproducibles en distintos conjuntos de datos por una máquina de vectores de soporte. Se considera que el agrupamiento jerárquico diseñado obtuvo unos resultados bastante robustos.

4.2.2. Agrupamiento K-Means

El algoritmo K-Means es ampliamente utilizado de distintas maneras en la clasificación de diabetes mellitus. El caso de (Carrillo-Larco et al., 2021), presenta la implementación de un algoritmo K-Means cuyos resultados permiten la estratificación de factores de riesgo. Esta versión, muestra una transformación ortogonal mediante PCA antes de ingresar los datos al algoritmo para su posterior clasificación. Además, los autores se basaron en distintos métodos como Elbow, Silhouette, y el análisis de dendograma, para escoger la cantidad de grupos a crear.

4.2.3. Agrupamiento DBSCAN

El caso de (Bej et al., 2022), muestra un agrupamiento dividido en dos procesos distintos. En primer lugar el tratamiento, y creación de *embeddings* de cada una de las características con distintas medidas de distancia (Euclidiana, Hamming, y Canberra) mediante UMAP. Luego, se hace la unión de los distintos embeddings en un solo UMAP, el cual es dividido en distintas clases mediante el algoritmo DBSCAN. Este algoritmo, es ideal para encontrar las clases en el UMAP final, pues es un algoritmo basado en densidad de muestras, que han sido separadas previamente en un espacio bidimensional.

4.3. Evaluación de Algoritmos

La evaluación de los distintos métodos de agrupamiento se llevó a cabo en dos etapas separadas:

- La primera etapa, consiste en aplicar una serie de medidas de agrupamiento (definidas en la fase IV de la metodología, ver 5.3.1). Estas métricas, pueden dar una noción cuantitativa del desempeño de cada algoritmo mediante información sobre la composición de cada grupo creado. También, se pueden tener en cuenta factores medibles como la complejidad computacional de cada algoritmo y su respectiva escalabilidad a otros conjuntos de datos más grandes.

- La segunda fase, implica un análisis de la distribución e información del contenido de cada grupo. Esto, puede ser la explicación de cada uno de los grupos creados y de la relación de las variables en ellos.

La evaluación de los resultados de estas dos etapas concluye en consideraciones específicas para cada agrupamiento realizado, sus ventajas y desventajas para el tipo de problema planteado aquí, basándose en los criterios cuantitativos y cualitativos planteados.

Capítulo 5

Desarrollo de la Comparativa

5.1. Fase II: Comprensión de los Datos

5.1.1. Análisis del Conjunto de datos

En esta sección se llevaron a cabo distintas tareas de reconocimiento del conjunto de datos, ya sea la comprobación de la información ya estipulada por (Strack et al., 2014), como análisis adicionales de sus variables.

En primer lugar, se realizó la asignación de tipos de variable inicial para cada una de las variables del conjunto de datos, diferenciando entre numéricas y categóricas. Este paso es especialmente importante, pues existen variables codificadas previamente (ver tablas 4.4, 4.2, y 4.4) que pueden ser interpretadas como numéricas. Finalmente, se realizó la asignación de tipo de variable utilizando como referencia los tipos estipulados en la tabla 4.1. No obstante, la variable *Weight* se definió como categórica, pues sus valores no indican el peso en libras de cada paciente, sino si este se encuentra en un rango específico de peso (dentro de 9 distintas categorías).

Así mismo, se realizó una descripción básica de las variables numéricas del conjunto de datos, que aporta información general sobre los encuentros hospitalarios. En la tabla 5.1, se pueden observar algunas variables estadísticas básicas. También, en la figura 5.1, se aprecia una representación gráfica de la relación entre las variables numéricas presentes en el conjunto de datos junto con sus respectivas distribuciones, dando así una noción de su comportamiento. Por último, en la figura 5.2, se puede ver la matriz de correlación para las variables numéricas. Si bien, los valores no indican una alta correlación entre variables, sí se puede percibir cierta relación entre algunas de ellas. Tal es el caso del número

de medicaciones suministradas con el número de procedimientos realizados, siendo de los valores más altos en la matriz. Todas las representaciones mostradas aportan información cuando menos interesante para el desarrollo del agrupamiento, como la relación de tienen las variables de cantidad de medicamentos suministrados, tiempo en el hospital, procedimientos y laboratorios. Así mismo, la distribución del número de diagnósticos sugiere un promedio alto para todos los encuentros relacionados con al diabetes, lo que corrobora la teoría médica expuesta en (Nedyalkova et al., 2021) que indica la variedad de enfermedades subyacentes que los distintos pacientes de diabetes mellitus tipo 2 pueden presentar, además de que, es común que la diabetes no sea el primer diagnóstico.

Tabla 5.1: Información básica de las variables numéricas.

Variable	**Máximo**	**Mínimo**	**Desv. Estándar**	**Promedio**
Time in Hospital	14	1	2.98	4.39
Lab Procedures	132	1	19.67	43.1
Procedures	6	0	1.7	1.3
Medications	81	1	8.12	16
Outpatient	42	0	1.26	0.3
Emergency	76	0	0.93	0.19
Inpatient	21	0	1.26	0.63
Diagnoses	16	1	1.93	7.42

En el caso de las variables categóricas, se realizó la comprobación de cada uno de sus valores, donde se evidenció la cantidad de valores faltantes en las variables *Weight, Payer Code, y Medical Specialty*, las figuras 5.3 y 5.4, muestran la proporción de valores faltantes (representados con un símbolo "?" en el conjunto de datos) con respecto al resto de categorías. Sin embargo, a pesar de que las variables *A1Cresult y Glucose serum test* no presentan valores nulos o faltantes (tabla 4.1), si presentan una gran mayoría de datos "none". En otras palabras, la ausencia del test en los encuentros es bastante cercana al 100 %, punto que es fuertemente recalcado por (Strack et al., 2014), indicando la fuerte relación entre la posibilidad de readmisión y los test mencionados (ver los distintos test en la tabla 2.1). En la figura 5.5, se puede observar la proporción de las distintas categorías en las variables *A1Cresult y Glucose serum test.*

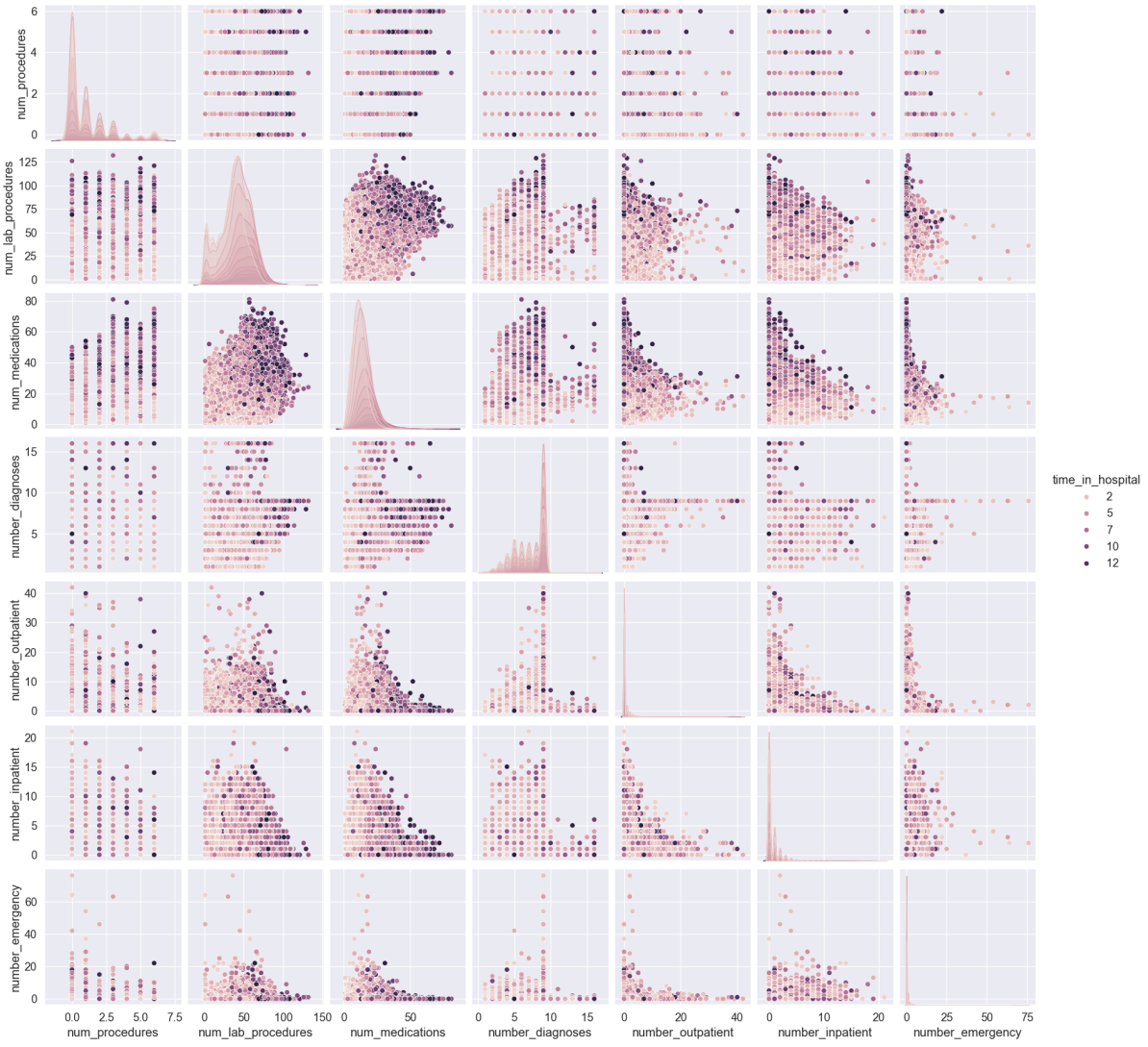


Figura 5.1: Relación entre variables numéricas

De igual forma, la exploración del balance de clases en cada una de las variables categóricas corrobora el criterio experto médico, como el caso de las variables de edad y género (*Age y Gender*), donde una mayoría de casos se concentran en pacientes con edades superiores a los 40 años, y así mismo, una proporción mayor del género femenino en los encuentros. En términos de raza, al ser un conjunto de datos extraído de hospitales en Estados Unidos, existe una clara dominancia de pacientes caucásicos (superior al 70 %) lo que dificulta encontrar efectos heterogéneos de los factores de riesgo de diabetes mellitus tipo 2 entre los distintos grupos étnicos, efecto que ha sido claramente expuesto en (Cho et al., 2019). En la figura 5.6, se puede observar la proporción de las variables de edad, raza, y género. En el caso de las variables correspondientes a los diagnósticos, existe una gran

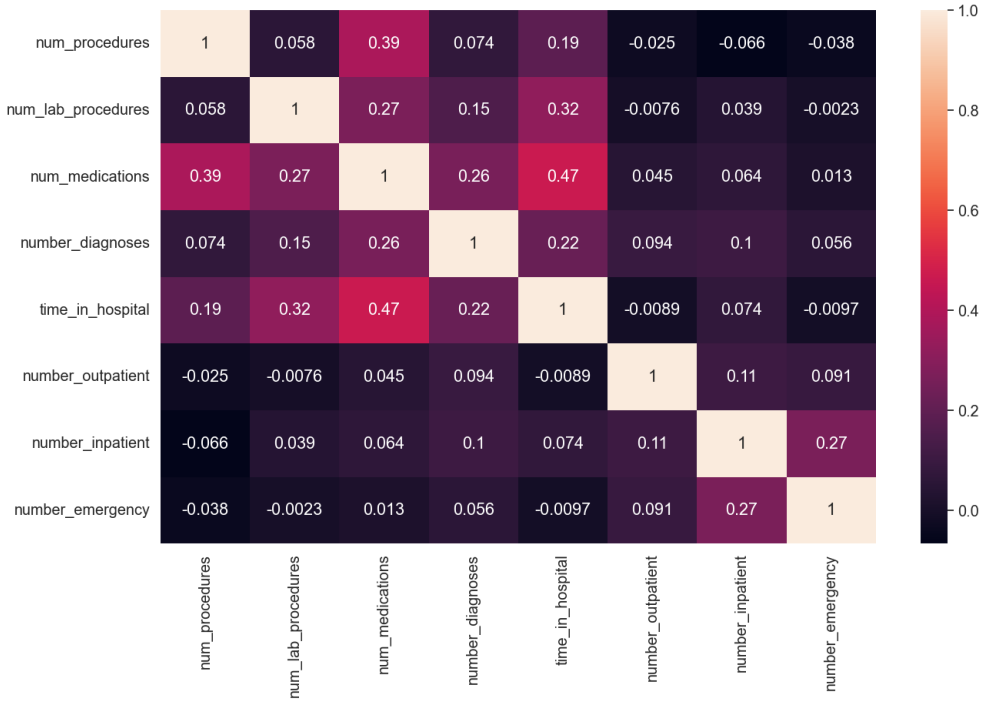


Figura 5.2: Correlación entre variables numéricas

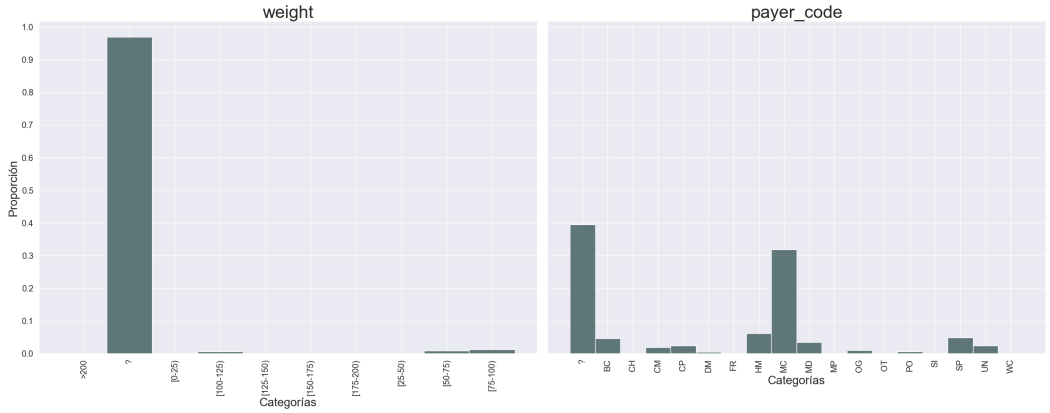


Figura 5.3: Proporción de las distintas categorías en las variables: *Weight y Payer Code*.

variedad dentro del conjunto de datos, razón por la cual se hablará únicamente de los diez diagnósticos más comunes en cada una de las columnas de diagnóstico. Estos diagnósticos, cubren cerca del 40 % de los encuentros dentro del conjunto de datos, y se han decodificado según el código ICD9 (Ministerio de Sanidad y Consumo). En la figura 5.7, se pueden observar los distintos diagnósticos comunes dentro de las variables de diagnóstico. Cabe resaltar que la variable de diagnóstico primario (*Diagnose 1*) no presenta comúnmente la diabetes mellitus (esta se ve más como un diagnóstico secundario, o secundario adicional), razón por la cual es poco frecuente que se efectúe el test de hemoglobina en los encuentros,

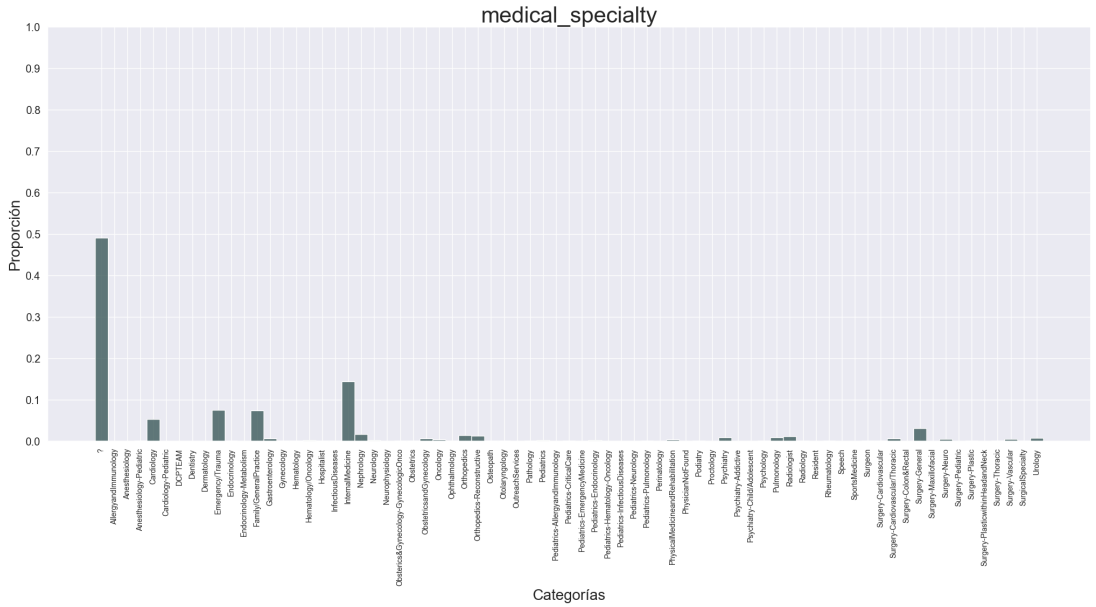


Figura 5.4: Proporción de las distintas categorías en la variable *Medical Specialty*.

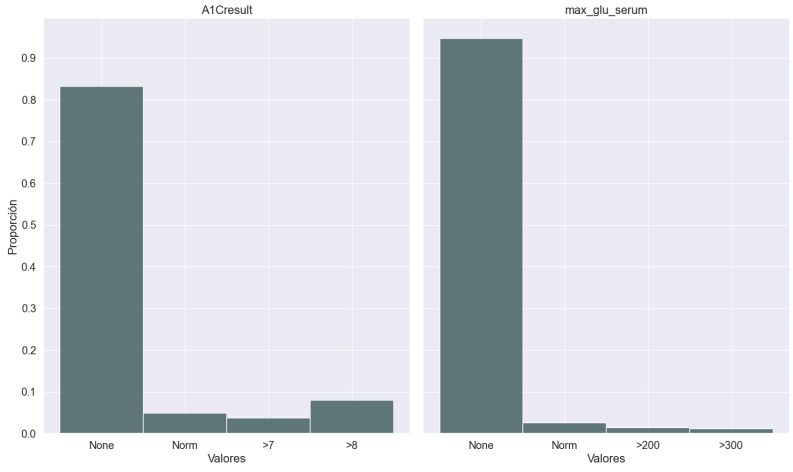


Figura 5.5: Proporción de las distintas categorías en las variables: *A1Cresult y Max glu serum*.

punto que es fuertemente resaltado por (Strack et al., 2014). Así mismo, el diagnóstico de diabetes mellitus no se hace muy común inclusive en el diagnóstico secundario (al rededor del 6 % de los encuentros), sino como un diagnóstico secundario adicional (con un porcentaje cercano al 12 %). Además del diagnóstico de diabetes, claramente se encuentran algunas de las enfermedades subyacentes expuestas por (Nedyalkova et al., 2021), como la falla cardiaca, hipertensión e isquémia.

Por último, se identifica que la mayoría de los encuentros fueron admisiones de tipo

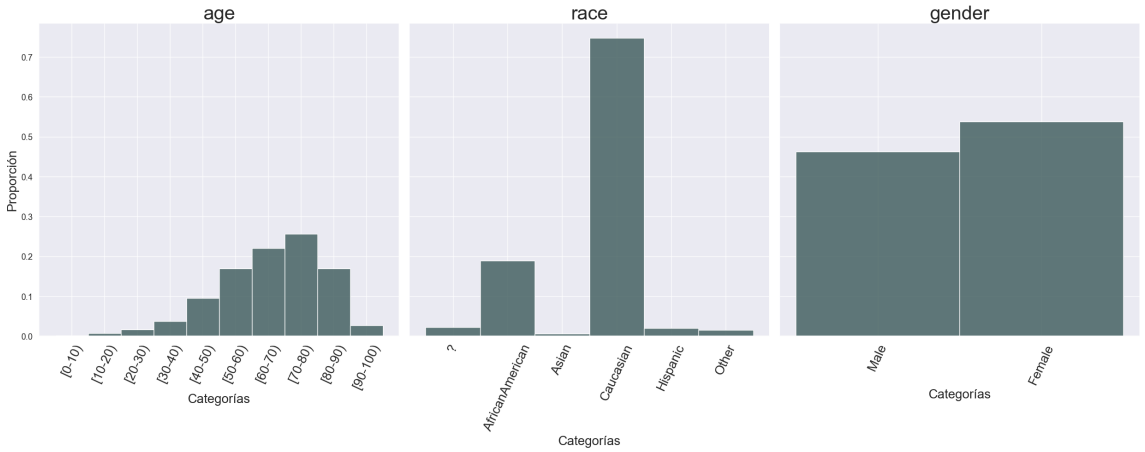


Figura 5.6: Proporción de las distintas categorías en las variables: *Age, Race, y Gender*.

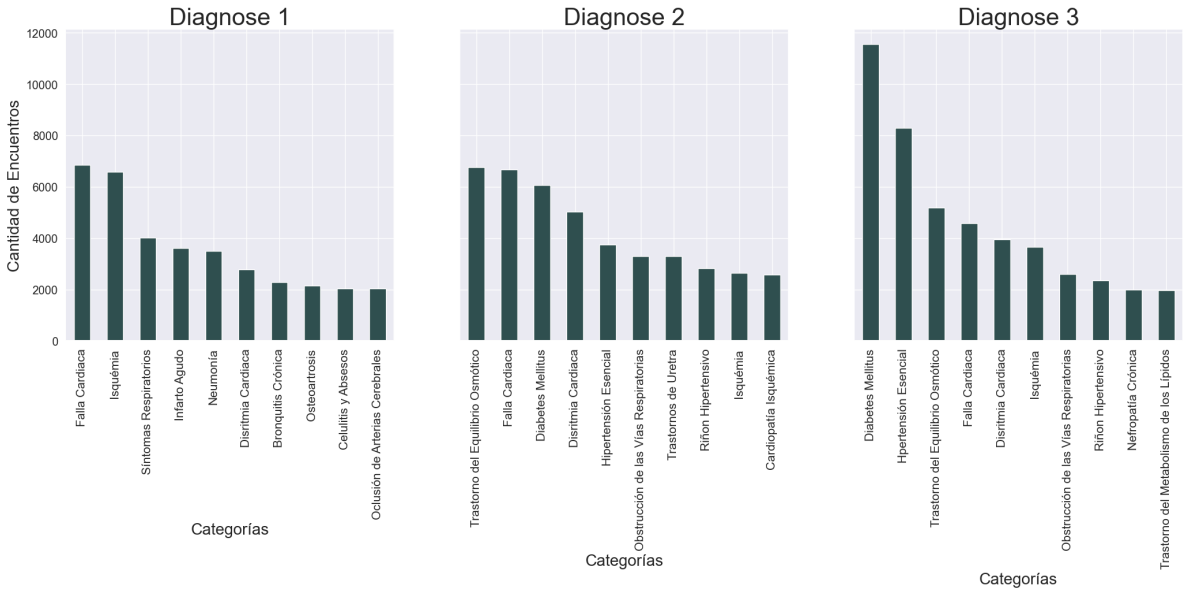


Figura 5.7: Diagnósticos primario, secundario, y secundario auxiliar más comunes dentro de los encuentros.

emergencia, dados de alta a casa, y la fuente de su admisión fue la sala de emergencias. Esta mayoría, corresponde a cerca del 60 % de los encuentros dentro del conjunto de datos.

5.1.2. Verificación de Calidad de Datos

Una vez realizado el análisis del conjunto de datos, se procedió a crear un sub-conjunto o subset con el fin de evitar el pre-procesamiento en datos que probablemente no se utilizarán en etapas posteriores. Así, por la cantidad de datos faltantes, información ya expuesta en las figuras 5.3 y 5.4, las variables *Payer Code, Weight, y Medical Specialty* se descarta-

ron. Complementando, variables como *Payer Code y Medical Specialty* no tienen mayor relevancia en el agrupamiento de sub-clases relacionadas diabetes mellitus, además de que fuera de los datos faltantes su valor no varía demasiado entre las clases disponibles, siendo pocas las categorías las que se reparten el porcentaje de observaciones restantes. A pesar de que el peso del paciente (*Weignt* en el conjunto de datos. Altamente relacionado con el índice de masa corporal) es una variable considerada relevante en el estudio, se decidió su descarte debido a la alta proporción de valores faltantes (cerca al 97 %).

Ya que el conjunto de datos presenta un pre-procesamiento previo, y como se aprecia en la tabla 4.1, no existía mucha presencia de datos nulos en la mayoría de variables, por lo que no fue necesaria una medida de llenado de valores faltantes. No obstante, se realizó una comprobación de valores duplicados. En la variable *Patient nbr*, se encontraban 30.248 valores duplicados, haciendo referencia a las múltiples veces que un paciente podría haber tenido un encuentro hospitalario relacionado con diabetes dentro del conjunto de datos. En este caso, existen pacientes que tuvieron encuentros hospitalarios hasta 40 veces. En la figura 5.8, se puede observar la gran cantidad de pacientes que solo han tenido 1 o pocos encuentros, en contraste, esta cantidad va disminuyendo a medida que aumentan los encuentros.

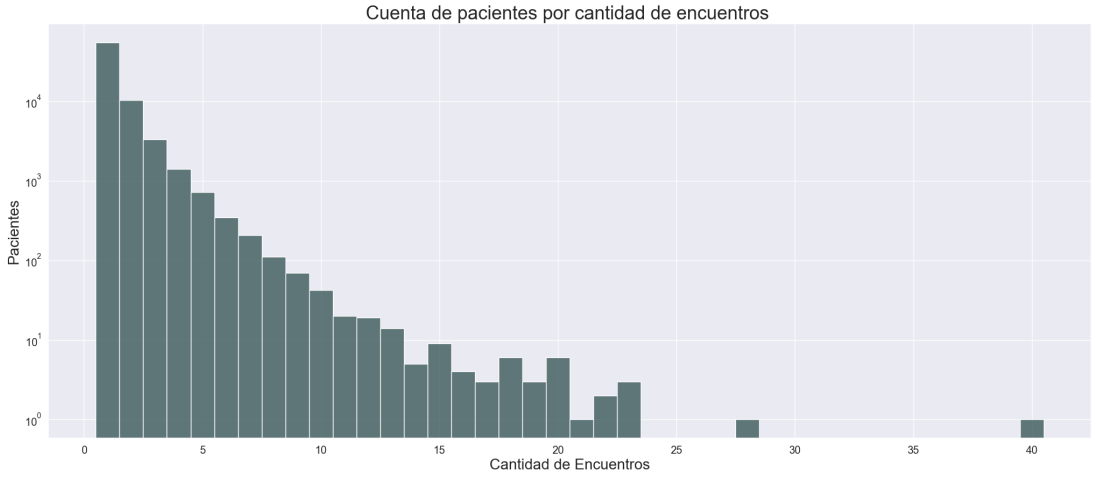


Figura 5.8: Pacientes vs cantidad de encuentros

5.2. Fase III: Preparación de los Datos

5.2.1. Limpieza y Transformación de los datos

En primer lugar, dentro de los distintos valores categóricos de las variables del subconjunto, se identificó un número muy reducido de encuentros (3) con el valor "Unknown/Invalid" en la variable *Gender*, razón por la cual se decidió eliminarlos y definir la variable como categórica binaria simétrica. Así mismo, las demás variables fueron definidas según su simetría, en la tabla 5.2, se puede observar la clasificación de cada variable.

Tabla 5.2: Clasificación de las variables categóricas según su simetría.

Variable	**Simetría**	**Binaria**
Race	Simétrica	No
Gender	Simétrica	Si
Age	Simétrica	No
Admission Type	Simétrica	No
Discharge Disposition	Simétrica	No
Admission Source	Simétrica	No
Diagnosis 1	Simétrica	No
Diagnosis 2	Simétrica	No
Diagnosis 3	Simétrica	No
Glucose serum text	Asimétrica	No
A1c test result	Asimétrica	No
Change of medications	Simétrica	Si
Diabetes medications	Asimétrica	Si
Readmitted	Asimétrica	No

Seguido, se debió realizar un codificado a las variables categóricas (similar al ya existente en las variables *Admission Type, Admission Source, y Discharge Disposition*) de acuerdo a su simetría. En las variables simétricas, cada valor tiene exactamente el mismo peso, caso contrario de las variables asimétricas, que a mayor valor más peso tendrá para el cálculo de distancia. Para la construcción del encoder necesario, se utilizó el paquete de *Scikit-Learn: Preprocessing - OrdinalEncoder*, sobre el cual se realizaron 3 tipos de enco-

ders distintos: Uno especial para las categorías de medicamentos, otro para las variables binarias asimétricas, y otro para el resto de categorías.

5.2.2. Pre-Procesamiento de los Datos

Una vez obtenido un conjunto de datos codificado, se considera preparado para el ingreso a las etapas de pre-procesamiento específicas de cada uno de los modelos propuestos. En otras palabras, los datos deben ser preparados de acuerdo a los requisitos de cada modelo.

Agrupamiento Jerárquico

En primer lugar, para realizar la implementación de agrupamiento jerárquico fue necesario definir dos parámetros: Una medida de distancia, y un método de enlace. Como establece el diseño previamente hecho por (Cho et al., 2019), se utilizó la distancia de Gower mediante el módulo de Python *Gower*.
El agrupamiento jerárquico es una técnica bastante costosa computacionalmente, razón por la cual se decidió hacer el agrupamiento con una muestra representativa del conjunto de datos (30.000 muestras, aproximadamente el 30 % del total del conjunto). Una vez creado el dataset más pequeño, se calculó la distancia de Gower y posteriormente se transformó su salida en una matriz de similaridad, como se muestra en la ecuación 5.1. Por último, la documentación de los módulos de *Scipy - Linkage y Dendogram* indica que el formato de entrada adecuado es una matriz condensada, por lo que se realizó esta última transformación a la matriz de similaridad mediante la función *Squareform* de *Scipy*.

$$SimilarityMatrix = 1 - GowerDistancesMatrix \tag{5.1}$$

Agrupamiento K-Means

Para el agrupamiento K-Means, que usualmente es un método que se beneficia de algún tipo de normalización de datos, se propuso una transformación ortogonal mediante el algoritmo PCA. Para el uso del reductor de dimensionalidad PCA se utilizó el paquete de *Scikit-Learn: Decomposition - PCA*, donde se activó el parámetro *whiten* que multiplica los vectores por la raíz cuadrada de las muestras, lo que garantiza salidas no correlacionadas con variaciones unitarias de los componentes. Esto, fue una recomendación del autor en (Carrillo-Larco et al., 2021). Inicialmente, se extrajeron los 2 primeros componentes

principales, los cuales explicaron el 91,18 % y el 8,82 % de la varianza total del conjunto de datos respectivamente, de manera que se decidió mantener ambos componentes.

Agrupamiento DBSCAN

Para el agrupamiento con DBSCAN, el autor (Bej et al., 2022), propone una reducción de dimensionalidad que a su vez actúa como un agrupamiento grafico conocida como UMAP. Ya que el conjunto de datos contiene diferentes tipos de variables (nominales, ordinales, y continuas) se realizó la extracción de embeddings de forma individual para cada tipo de variable. La razón principal de esto, es para poder utilizar distintas medidas de distancia con el conjunto de datos, así, para las variables continuas se utilizó la distancia Euclidiana, en el caso de las variables nominales se utilizó la distancia de Hamming, y por último para las variables ordinales se utilizó la distancia de Canberra. Además de las medidas de distancia, se escogieron las dos primeras dimensiones reducidas de UMAP para las variables continuas y ordinales, mientras que solamente se escogió la primera dimensión reducida para las variables nominales, por lo que con una concatenación de todos las respuestas deja una representación de 5 dimensiones de los datos. Los parámetros de *n neighbours* y *min distance* fueron de 30 y 0.1 respectivamente para todos los tipos de variables. En la figura **??**, se pueden observar las 3 distintas representaciones gráficas de UMAP.

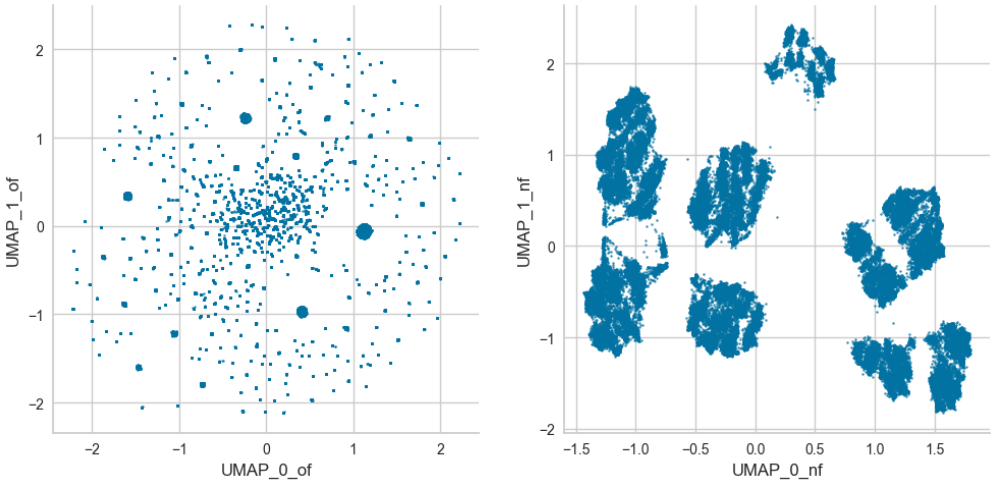
Se realizó una última reducción de dimensionalidad con UMAP sobre los datos, lo cual redujo la respuesta a una representación bidimensional. En la figura 5.10, se puede observar la representación final del conjunto de datos realizada con UMAP. Así, el conjunto de datos se considera preparado para ingresar al algoritmo DBSCAN.

5.3. Fase IV: Modelado

5.3.1. Definición de métricas de rendimiento

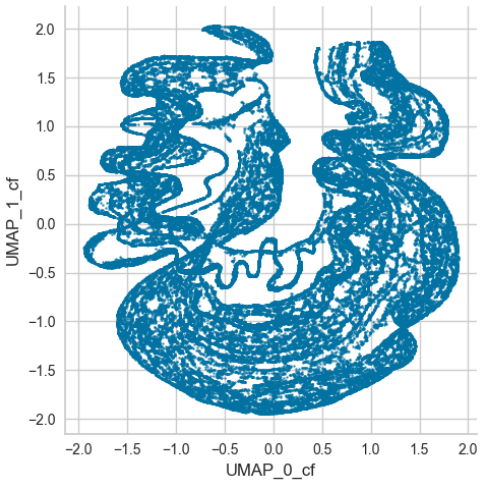
Average Silhouette Width

Tal y como se menciona en la sección 2.2.3, el Average Silhouette Width (ASW), o índice de Silhouette, es ampliamente utilizado para medir la calidad de un agrupamiento, pues se considera que su valor puede aportar información valiosa además de si un agrupamiento es bueno o no. Específicamente, si el valor de ASW es cercano a cero, indica que



(a) Representacion de dimensionalidad reducida de UMAP para variables ordinales.

(b) Representacion de dimensionalidad reducida de UMAP para variables nominales.



(c) Representacion de dimensionalidad reducida de UMAP para variables continuas

Figura 5.9: Representaciones de baja dimensionalidad de UMAP para cada tipo de variable.

hay grupos que se están superponiendo entre sí. Si en cambio, este valor es más cercano a -1, indica que la observación pertenece más a otro grupo que al actual. Por último, un valor cercano a 1 indica una buena clasificación. Esta información es tratada a más profundidad en la sección 2.2.3.Adicionalmente, es un índice que premia los grupos densos y bien separados. Debido a las características mencionadas, el ASW ha sido seleccionado como una de las métricas a utilizar.

Para realizar la evaluación del índice de Silhouette, se utilizó el método de *Scikit-Learn: Metrics - Silhouette Score*, el cual permite conocer tanto el valor promedio del índice para

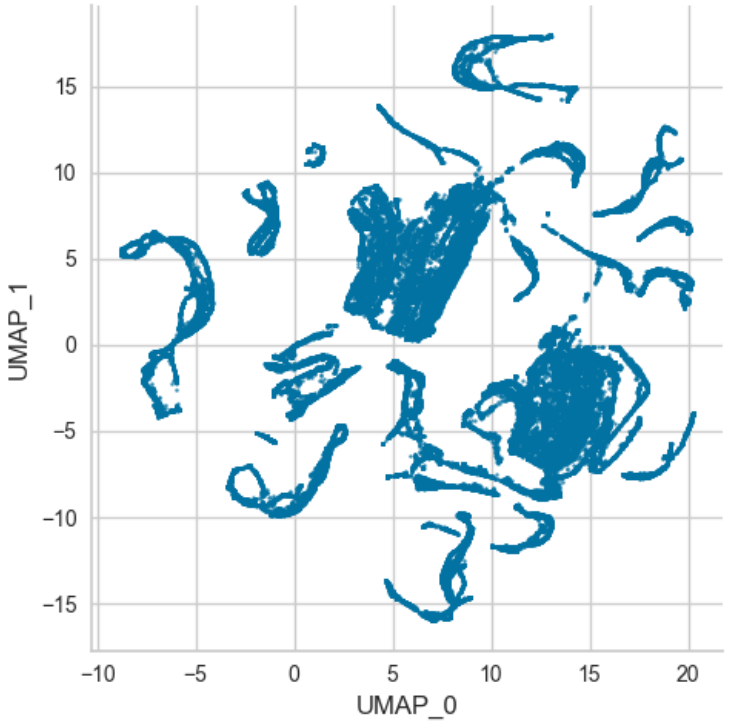


Figura 5.10: Representación de dimensionalidad reducida de UMAP para todo el conjunto de datos

todo el agrupamiento como el valor de cada observación. Estos valores se grafican y permiten tener una visión mucho más clara de la calidad de cada uno de los grupos generados por el clasificador.

Índice de Elbow

El índice de Elbow, al igual que el índice de Silhouette, permite medir la calidad de un agrupamiento. En este caso, método de Elbow se efectúa de manera visual graficando la suma de distancias cuadradas (*Squared Distances*) contra una serie de distintos agrupamientos. Luego, se ubica la cantidad de grupos que posee el clasificador actual para encontrar la suma de distancias cuadradas correspondiente. Este método es tratado a más detalle en la sección 2.2.3.

Para el cálculo del índice de Elbow se utilizó la librería de python *Yellow Brick* junto con el método *K Elbow Visualizer*, que proporciona una gráfica completa del método y de cual es el valor óptimo de k para el clasificador. Además, permite el uso de otras métricas de rendimiento para la selección de k, como el índice de Calinski-Harabasz y el índice de Silhouette. Para la evaluación de los modelos, se analizó el método de Elbow con las tres

métricas posibles, pero se dio prioridad a la suma de distancias cuadradas que corresponde al método original.

Índice de Davies-Bouldin

El índice de Davies-Bouldin se considera una medida de similaridad entre grupos, la cual, compara la distancia entre grupos con el tamaño de los mismos. Normalmente, el índice de Davies-Bouldin es más simple de calcular que el índice de Silhouette. Matemáticamente, este índice se define como la similaridad promedio entre cada cluster $C_i = 1, ..., k$ y su cluster más similar C_j. La similaridad R_{ij}, para éste índice, se mide como muestra la ecuación 5.2 (Davies and Bouldin, 1979):

$$R_{ij} = \frac{s_i + s_j}{d_{ij}} \tag{5.2}$$

- s_i es la distancia promedio entre cada elemento del cluster y el centroide del mismo cluster, también conocido como el diámetro del cluster.
- d_{ij} es la distancia entre los centroides de los clusters i y j.

Entonces, el índice de Davies-Bouldin sería como el mostrado en la ecuación 5.3:

$$DB = \frac{1}{k} \sum_{i=1}^{k} \underset{i \neq j}{max} R_{ij} \tag{5.3}$$

Así, el mejor valor posible del índice de Davies-Bouldin es cero, y valores cercanos a cero indican una buena partición.
Para el cálculo del Índice de Davies-Bouldin se utilizó el paquete de *Scikit-Learn: Metrics - davies bouldin score*.

Índice de Calinski-Harabasz

El índice de Calinski-Harabasz, también conocido como criterio de relación de varianza (del inglés *Variance Ratio Criterion*), contrario al índice de Davies-Bouldin, indica grupos mejor definidos entre más alto sea su valor. Para un conjunto de datos E de tamaño n_E que ha sido partido en k clusters, el índice de Calinski-Harabasz (s) se define como la relación entre la dispersión promedio entre los clustres y la dispersión dentro de los clusters. En la ecuación 5.4 se puede observar el índice de Calinski-Harabasz (Caliński and Harabasz, 1974).

$$s = \frac{tr(B_k)}{tr(W_k)} \times \frac{n_E - k}{k - 1} \tag{5.4}$$

Donde $tr(B_k)$ y $tr(W_k)$ representan la matriz de dispersión entre grupos y la matriz de dispersión entre elementos dentro del cluster respectivamente. Estos valores están definidos en las ecuaciones 5.5 y 5.6.

$$W_k = \sum_{q=1}^{k} \sum_{x \in C_q} (x - C_q)(x - C_q)^T \tag{5.5}$$

$$B_k = \sum_{q=1}^{k} n_q (c_q - c_E)(c_q - c_E)^T \tag{5.6}$$

Donde C_q es el conjunto de puntos en el cluster q, c_q es el centro del cluster q, c_E es el centro de E, y n_q es el número de puntos en el cluster q.
Para el cálculo del Índice de Calinski-Harabasz se utilizó el paquete de *Scikit-Learn: Metrics - calinski harabasz score*.

5.3.2. Desarrollo de modelos y Fine Tuning

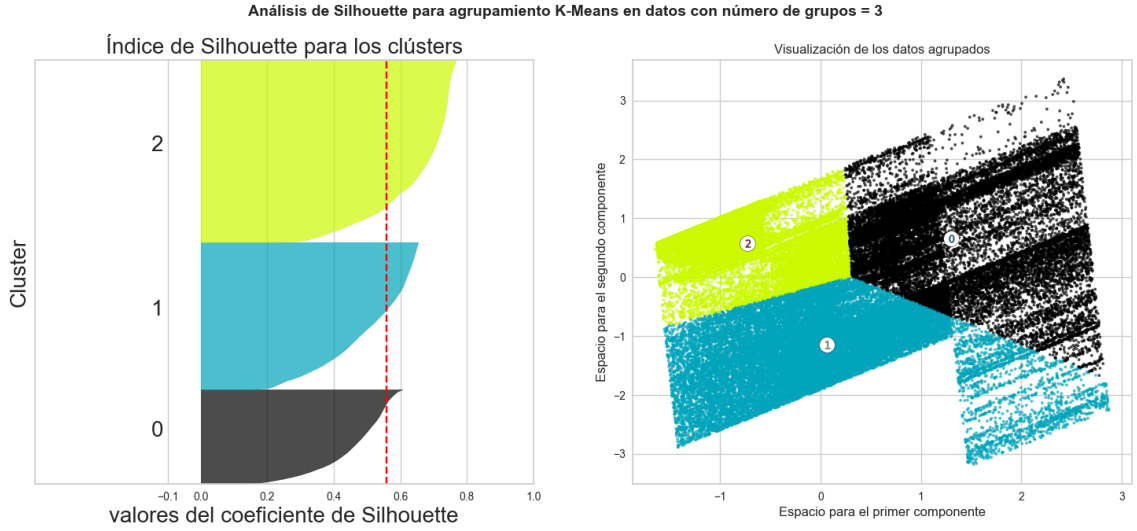
Agrupamiento Jerárquico

En el caso del agrupamiento jerárquico, se utilizaron los módulos de *Scipy - Linkage y Dendogram* para realizar el agrupamiento con el método de enlace completo y dibujar el dendograma respectivamente. Este último paso es importante para hacer la selección de grupos visualizando el dendograma. Sin embargo, al ser una muestra de datos grande, se dificulta encontrar un punto para el corte del árbol generado de manera visual, por lo que se decidió seguir el criterio de Elbow para la selección de grupos. Acorde al criterio de Elbow, el número óptimo de clusters fué de $k = 5$. Para facilitar el cálculo de métricas, se decidió hacer el agrupamiento con el módulo de *Scikit-Learn: Cluster - Agglomerative Clustering*, que posee un método *fit* para alimentar el modelo con los datos de entrenamiento, que permite ingresar el modelo alimentado a las funciones de cálculo de métricas.

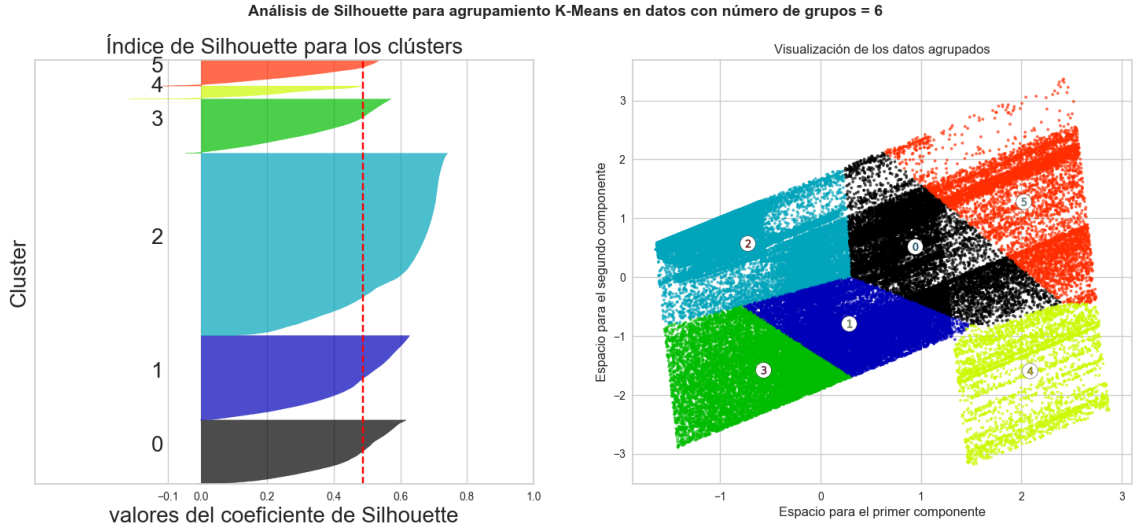
Agrupamiento K-Means

Para realizar el agrupamiento con el algoritmo K-Means, se utilizó el paquete de *Scikit-Learn: Cluster - KMeans*. En este caso, todos los parámetros del clasificador se dejaron en su valor por defecto y se escogió la cantidad de grupos siguiendo tanto el método de Elbow como el de Silhouette iterando el algoritmo con $k = 2, ..., 6$. No obstante, ambos métodos presentaban discrepancias, recomendando un $k = 6$ y un $k = 3$ respectivamente. En la figura 5.11, se pueden observar los resultados del índice de Silhouette junto con la gráfica

de los grupos creados por el algoritmo en ambos casos. En (Carrillo-Larco et al., 2021), el autor recomienda también realizar un análisis de dendograma, sin embargo, se prefirió no realizar considerando el tamaño del conjunto de datos y su alto coste computacional. Así mismo, los criterios de Elbow y Silhouette también presentan buenos resultados.



(a) Resultados del índice de silhouette y los agrupamientos realizados con $k = 3$.



(b) Resultados del índice de silhouette y los agrupamientos realizados con $k = 6$.

Figura 5.11: Índice de Silhouette para los dos mejores casos de k. Se escogió un valor final de $k = 3$ para evitar grupos con muy pocos elementos, además de que poseen un promedio del índice de Silhouette más bajo.

Agrupamiento DBSCAN

En el caso del agrupamiento DBSCAN se utilizó el paquete de *Scikit-Learn: Cluster - DBSCAN*. Su parámetro más importante (*eps*) fue establecido en 1.1, y representa la distancia máxima entre dos unidades para considerar una en el vecindario de la otra. Otro parámetro representativo fue el *min samples*, seteado en 200, y representa el número de muestras que debe tener un vecindario para considerarlo un núcleo. Estos parámetros fueron definidos a prueba y error, corroborando los resultados del agrupamiento. Una vez configurado el clasificador, detectó 13 distintos grupos en los datos de entrada, los cuales se pueden observar en la figura 5.12.

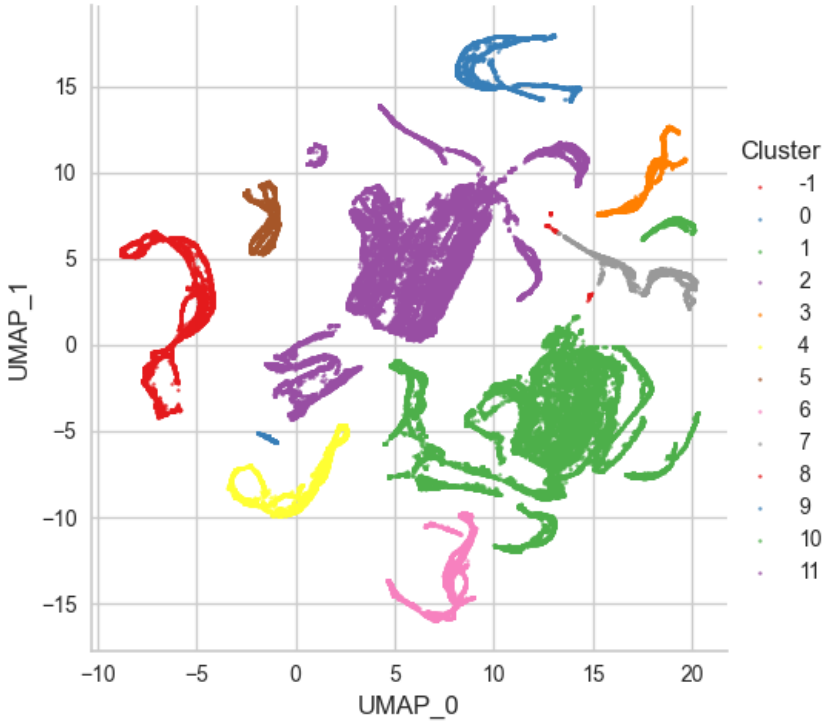


Figura 5.12: Grupos detectados por DBSCAN en la reducción de dimensionalidad UMAP del conjunto de datos.

5.4. Fase V: Evaluación

5.4.1. Agrupamiento Jerárquico

El agrupamiento jerárquico generó un total de $k = 5$ grupos, los cuales están distribuidos como se muestra en la figura 5.13. Así mismo, las métricas de rendimiento calculadas

dieron los resultados mostrados en la tabla 5.3. Complementando, el índice de Silhouette por cada cluster se muestra en la figura 5.14, y los distintos análisis del método de Elbow se observan en la figura 5.15.

Figura 5.13: Distribución de clusters del Agrupamiento Jerárquico

Tabla 5.3: Métricas de rendimiento del Agrupamiento Jerárquico

Métrica	**Resultado**
Índice promedio de Silhouette	-0.1889
índice de Davies-Bouldin	10.5643
Índice de Calinski-Harabasz	221.8803

Así mismo, se calcularon variables estadísticas para revisar la diferenciación de variables continuas entre los distintos clusters formados. En la tabla 5.4, se pueden observar los valores calculados.

Todos los grupos formados son muy similares en términos del número de diagnósticos (*Diagnoses*), diferenciándose el grupo 2, con un promedio más alto. En el caso del número de emergencias (*Emergency*), el grupo 3 incluye en su mayoría encuentros cuyos pacientes no tuvieron visitas a emergencias en el último año, los demás grupos son bastante simila-

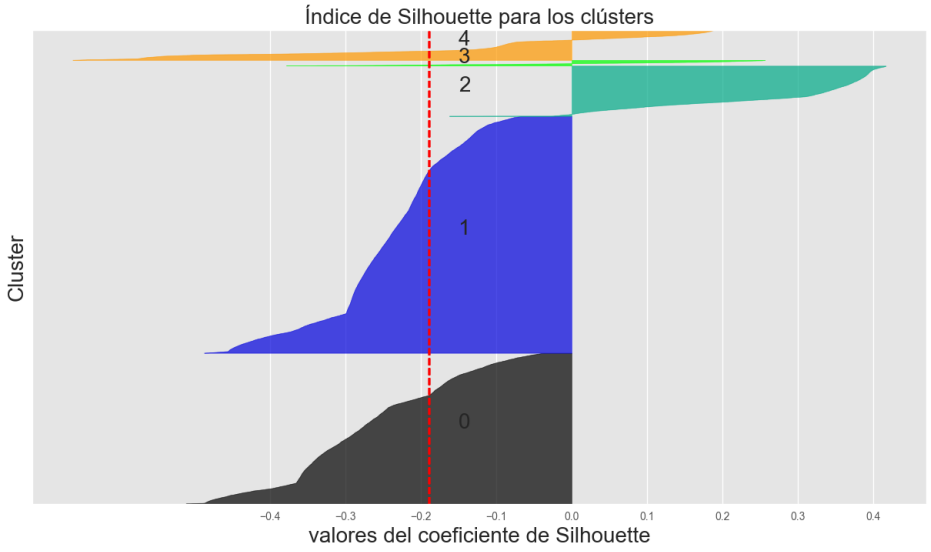
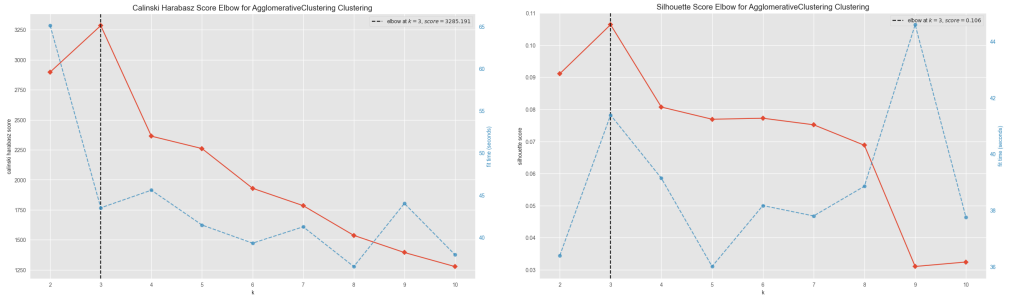
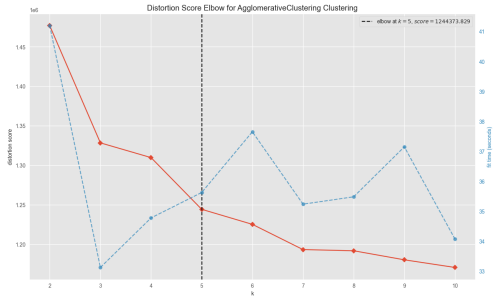


Figura 5.14: Índice de Silhouette para el Agrupamiento Jerárquico.



(a) Resultados del método de Elbow con Calinski-Harbasz como métrica.

(b) Resultados del método de Elbow con Silhouette como métrica.



(c) Resultados del método de Elbow con la suma de los errores cuadrados como métrica.

Figura 5.15: Método de Elbow para el Agrupamiento Jerárquico utilizando distintas métricas.

res en promedio y dispersión, exceptuando el grupo 2, que muestra valores más altos. Los niveles de hospitalizaciones (*Inpatient*) están distribuidos a través de los grupos, siendo el grupo 3 el más bajo, seguido de los grupos 4, 0, 1, y el grupo 2 con los valores más altos. Seguido, el grupo 2 se hayan encuentros cuyo valor de procedimientos de laboratorio (*Lab. Procedures*) es bastante más bajo que los demás grupos. Para los medicamentos administrados (*Medications*), los grupos 1 y 3, representan los valores más bajos, el grupo 0 y 4, muestran los valores más altos con encuentros de pacientes que recibieron hasta 70 medicamentos, y por último, el grupo 2 muestra valores intermedios. Para las visitas a consulta externa (*Outpatients*), los valores son ciertamente similares entre grupos, con los grupos 1 y 3 representando los valores más bajos, los grupos 0 y 4, valores intermedios, y el grupo 2 los más altos con una gran diferencia respecto a los demás. El caso de los procedimientos y el tiempo en el hospital (*Procedures y Time in Hospital*), los valores están repartidos proporcionalmente entre los grupos, siendo de más bajo a más alto los grupos 2, 3, 1, 4, y 0.

Tabla 5.4: Información básica de las variables numéricas del Agrupamiento Jerárquico.

Cl.	**Variable**	**Máximo**	**Mínimo**	**Desv. Est.**	**Promedio**
0	Time in Hospital	14	1	3.34	5.28
0	Lab Procedures	113	1	17.81	49.20
0	Procedures	6	0	1.93	1.73
0	Medications	70	1	9.24	17.76
0	Outpatient	20	0	0.75	0.14
0	Emergency	10	0	0.31	0.04
0	Inpatient	14	0	1.20	0.60
0	Diagnoses	9	1	2.01	6.87
1	Time in Hospital	14	1	3.01	4.39
1	Lab Procedures	129	1	16.70	46.41
1	Procedures	6	0	1.66	1.35
1	Medications	58	1	7.01	13.44
1	Outpatient	16	0	0.52	0.08
1	Emergency	22	0	0.33	0.03
1	Inpatient	17	0	1.23	0.61

Tabla 5.4: (Continuación).

Cl.	Variable	Máximo	Mínimo	Desv. Est.	Promedio
1	Diagnoses	9	1	2.10	6.67
2	Time in Hospital	14	1	2.75	4.07
2	Lab Procedures	80	1	5.77	19.86
2	Procedures	6	0	1.03	0.68
2	Medications	44	1	6.63	15.50
2	Outpatient	13	0	1.35	0.77
2	Emergency	25	0	1.01	0.38
2	Inpatient	16	0	1.31	0.62
2	Diagnoses	9	1	1.83	7.25
3	Time in Hospital	14	1	2.99	4.19
3	Lab Procedures	81	1	12.57	46.84
3	Procedures	6	0	1.34	0.81
3	Medications	43	2	5.47	12.72
3	Outpatient	3	0	0.21	0.02
3	Emergency	1	0	0.11	0.01
3	Inpatient	4	0	0.70	0.34
3	Diagnoses	9	2	2.01	6.27
4	Time in Hospital	14	1	3.16	5.07
4	Lab Procedures	105	1	18.50	51.64
4	Procedures	6	0	1.76	1.59
4	Medications	63	1	8.30	16.76
4	Outpatient	7	0	0.60	0.15
4	Emergency	4	0	0.27	0.04
4	Inpatient	8	0	1.11	0.56
4	Diagnoses	9	1	1.85	6.73

De igual manera, las categorías dominantes en cada una de las variables categóricas de los clusters se encuentran en la tabla 5.5.

En la variable de fuente de admisión (*Admission Source*), las categorías que más presencia

tienen son “Sala de Emergencias” y “Nulo”. El mismo caso sucede con la variable del tipo de admisión (*Admission Type*), cuyos casos más frecuentes son “Emergencia” y “No Disponible”. Todos los grupos poseen el rango de edad (*Age*) de 70-80 años el más común. El cambio en la medicación administrada (*Change*) permanece en todos los grupos, menos en el grupo 1, donde prevalecen los encuentros donde no se cambia la medicación del paciente. Ahora, la medicación relacionada con diabetes mellitus (*Diabetes Medication*) está en la mayoría de encuentros de todos los grupos, y se encuentra en todos los casos presentes en el grupo 3. Para los diagnósticos principales en los grupos (*Diagnosis 1*), se separan entre “Isquémia” y “Neoplasia Respiratoria”. Los diagnósticos auxiliares (*Diagnosis 2*) presentan una mayor variedad, con “Diabetes Mellitus” en los grupos 0, 1, y 3, “Enfermedad Hepática” en el grupo 2, y “Obstrucción de las vías respiratorias” en el grupo 4. Sin embargo, para el diagnóstico auxiliar secundario (*Diagnosis 3*), todos los grupos tienen “Riñón Hipertensivo” como categoría más común, y así mismo, la mayoría de casos son dados de alta a casa en la variable de disposición de descarga (*Discharge Disposition*). En los grupos 0, 1, y 2, la mayoría de encuentros son a pacientes mujeres, siendo entonces los grupos 3 y 4 los que poseen mayormente encuentros con pacientes hombres. En términos de raza (*Race*), todos los grupos tienen mayoría de pacientes caucásicos. Por último, todos los grupos muestran dominancia de encuentros sin readmisión (variable *Readmitted*), exceptuando el grupo 4, donde la mayoría de sus encuentros son pacientes readmitidos más de 30 días después.

Las variables de los resultados de los test de glucosa plasmática (*max glu serum*) y hemoglobina A1c (*A1Cresult*) muestran que no se realizaron en la mayoría de casos (ver figura 5.5), sin embargo, los encuentros donde sí se aplicaron los tests se tuvieron en cuenta en la distribución de los grupos. Para el test de hemoglobina, los pacientes con valores superiores al 8 % son mayoría dentro de los grupos 0, 1, 3, y 4. Así, en el grupo 3 solamente hay pacientes que no se les realizó el test. Similarmente, para el test de glucosa plasmática, los resultados dentro de los valores normales son mayoría en los grupos 0, 1, 2, y 4.

Además de las variables encontradas en la tabla 5.5, se observó la distribución de los distintos medicamentos genéricos para la diabetes mellitus. En el grupo 0 y 1, se encontró alta presencia de prescripción de insulina cuya dosis se mantuvo. En adición, el grupo 3 muestra alta prescripción de Metmorfina y Gliburida.

Tabla 5.5: Información básica de las variables categóricas del Agrupamiento Jerárquico.

Cluster	**Variable**	**Valor Dominante**
0	Race	Caucasian
0	Gender	Femenino
0	Age	70-80
0	Admission Type Id	Emergencia
0	Discharge Disposition Id	Dado de alta a casa
0	Admission Source Id	Sala de Emergencias
0	Diagnosis 1	Neoplasia Respiratoria y Digestiva
0	Diagnosis 2	Diabetes Mellitus
0	Diagnosis 3	Riñón Hipertensivo
0	Change	Cambio en medicación
0	Diabetes Medication	Si
0	Readmitted	No
1	Race	Caucasian
1	Gender	Femenino
1	Age	70-80
1	Admission Type Id	Emergencia
1	Discharge Disposition Id	Dado de alta a casa
1	Admission Source Id	Sala de Emergencias
1	Diagnosis 1	Neoplasia Respiratoria y Digestiva
1	Diagnosis 2	Diabetes Mellitus
1	Diagnosis 3	Riñón Hipertensivo
1	Change	No hay cambio en medicación
1	Diabetes Medication	Si
1	Readmitted	No
2	Race	Caucasian
2	Gender	Femenino
2	Age	70-80
2	Admission Type Id	No Disponible
2	Discharge Disposition Id	Dado de alta a casa
2	Admission Source Id	Nulo

Tabla 5.5: (Continuación).

Cluster	Variable	Valor Dominante
2	Diagnosis 1	Isquémia
2	Diagnosis 2	Enfermedad Hepática
2	Diagnosis 3	Riñón Hipertensivo
2	Change	Cambio en medicación
2	Diabetes Medication	Si
2	Readmitted	No
3	Race	Caucasian
3	Gender	Masculino
3	Age	70-80
3	Admission Type Id	Emergencia
3	Discharge Disposition Id	Dado de alta a casa
3	Admission Source Id	Sala de Emergencias
3	Diagnosis 1	Isquemia
3	Diagnosis 2	Diabetes Mellitus
3	Diagnosis 3	Riñón Hipertensivo
3	Change	Cambio en medicación
3	Diabetes Medication	Si (únicamente)
3	Readmitted	No
4	Race	Caucasian
4	Gender	Masculino
4	Age	70-80
4	Admission Type Id	Nulo
4	Discharge Disposition Id	Dado de alta a casa
4	Admission Source Id	Nulo
4	Diagnosis 1	Neoplasia Respiratoria y Digestiva
4	Diagnosis 2	Obstrucción Vías Respiratorias
4	Diagnosis 3	Riñón Hipertensivo
4	Change	Cambio en medicación
4	Diabetes Medication	Si

Tabla 5.5: (Continuación).

Cluster	Variable	Valor Dominante
4	Readmitted	>30 Días

5.4.2. Agrupamiento Kmeans

El agrupamiento K-Means generó un total de $k = 3$ grupos, los cuales están distribuidos como se muestra en la figura 5.16. Así mismo, las métricas de rendimiento calculadas dieron los resultados mostrados en la tabla 5.6. Complementando, el índice de Silhouette por cada cluster se muestra en la figura 5.17, y los distintos análisis del método de Elbow se observan en la figura 5.18.

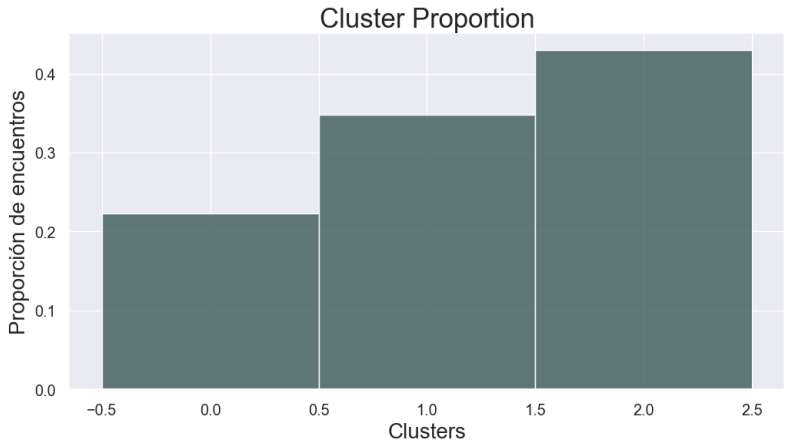


Figura 5.16: Distribución de clusters del Agrupamiento K-Means

Tabla 5.6: Métricas de rendimiento del Agrupamiento K-Means

Métrica	Resultado
Índice promedio de Silhouette	0.4690
índice de Davies-Bouldin	0.6830
Índice de Calinski-Harabasz	139025.0573

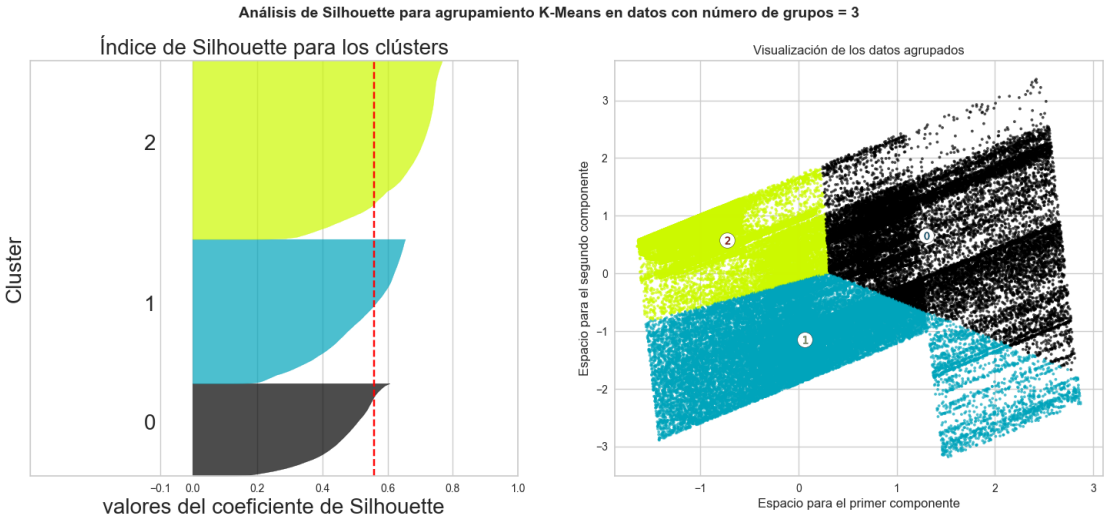
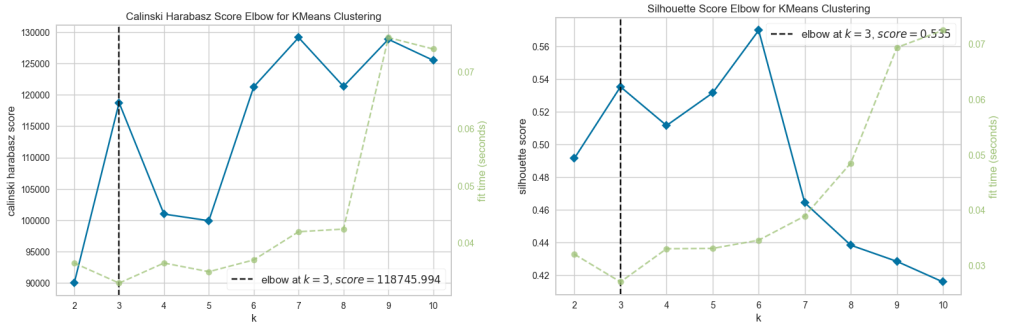
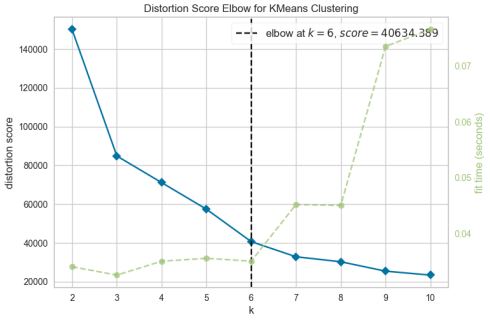


Figura 5.17: Índice de Silhouette para el agrupamiento K-Means junto con los agrupamientos realizados en las dimensiones reducidas por PCA.



(a) Resultados del método de Elbow con Calinski-Harbasz como métrica.

(b) Resultados del método de Elbow con Silhouette como métrica.



(c) Resultados del método de Elbow con la suma de los errores cuadrados como métrica.

Figura 5.18: Método de Elbow para el Agrupmiento K-Means utilizando distintas métricas.

Al igual que con el agrupamiento jerárquico, se realizaron los mismos cálculos estadísticos para las variables continuas en cada uno de los clusters. En la tabla 5.7, se pueden observar estos cálculos.

Para la variable diagnósticos, el grupo 2 representa en su mayoría valores bajos, teniendo un valor máximo de 9 diagnósticos en sus encuentros. Los grupos 1 y 0, representan valores intermedios y altos respectivamente incluyendo encuentros de hasta 16 diagnósticos. De la misma manera, el grupo 2 tiene el promedio más bajo en cuanto a emergencias, con los grupos 0 y 1 con un comportamiento muy similar, siendo el grupo 1 con una dispersión más baja y casos de hasta 25 emergencias en el último año. En cambio, el grupo 0 tiene los encuentros con hasta 75 casos de emergencias, y por ende, una dispersión más alta. Para los casos de hospitalizaciones y procedimientos de laboratorio, todos los grupos se comportan de forma muy similar. Añadiendo, los grupos son muy similares en cuanto a las medicaciones administradas, sin embargo, el grupo 0 presenta pequeñas diferencias con un promedio más alto. Para los casos de consulta externa, los grupos 0 y 1 tienen un comportamiento muy parecido, siendo el grupo 2 bastante más bajo con respecto a ellos. Por último, tanto los procedimientos y el tiempo en el hospital tienen un comportamiento muy similar entre todos los grupos.

Tabla 5.7: Información básica de las variables numéricas del Agrupamiento K-Means.

Cl.	**Variable**	**Máximo**	**Mínimo**	**Desv. Est.**	**Promedio**
0	Time in Hospital	14	1	2.86	4.14
0	Lab Procedures	126	1	20.99	42.66
0	Procedures	6	0	1.78	1.32
0	Medications	75	1	8.07	16.85
0	Outpatient	40	0	1.49	0.49
0	Emergency	76	0	1.34	0.27
0	Inpatient	16	0	1.34	0.69
0	Diagnoses	16	1	1.65	8.04
1	Time in Hospital	14	1	2.94	4.43
1	Lab Procedures	132	1	20.80	44.64
1	Procedures	6	0	1.65	1.31
1	Medications	66	1	7.59	15.81

Tabla 5.7: (Continuación).

Cl.	Variable	Máximo	Mínimo	Desv. Est.	Promedio
1	Outpatient	42	0	1.52	0.50
1	Emergency	25	0	0.90	0.23
1	Inpatient	19	0	1.27	0.67
1	Diagnoses	16	1	1.75	7.75
2	Time in Hospital	14	1	3.08	4.50
2	Lab Procedures	114	1	17.87	42.07
2	Procedures	6	0	1.71	1.37
2	Medications	81	1	8.54	15.77
2	Outpatient	35	0	0.82	0.20
2	Emergency	42	0	0.65	0.13
2	Inpatient	21	0	1.20	0.58
2	Diagnoses	9	1	2.05	6.84

También, fueron destacadas las variables categóricas más relevantes de cada agrupamiento. En la tabla 5.8 se pueden observar las categorías dominantes de cada cluster formado.

Todas las variables relacionadas con las admisiones y descargas de los hospitales se comportan de una manera muy similar en todos los grupos (*Admission Type, Discharge Disposition, y Admission Source*). Del mismo modo, la edad, género, raza, y la presencia de medicamentos de diabetes presentan valores muy parecidos entre los grupos. También, existe dominancia de casos que no fueron readmitidos en cada grupo. Sin embargo, en términos de cambios de medicación, los grupos 1 y 2 muestran en su mayoría casos donde no hubo cambio. Caso distinto al del grupo 0, cuya mayoría de casos muestra cambios en la medicación administrada. Para las variables de diagnóstico, los diagnósticos principales son en su mayoría isquémia para los grupos 0 y 1, siendo neoplasia respiratoria para el grupo 2. En cuanto al diagnóstico secundario, existe un caso similar, pues los grupos 0 y 2 mayormente muestran enfermedades hepáticas, con el grupo 1 dominando la obstrucción de vías respiratorias. Por último, todos los grupos muestran dominancia de riñón hipertensivo para el diagnóstico secundario auxiliar.

Para los tests de hemoglobina A1c y glucosa plasmática que se realizaron, el comportamiento de los grupos es el siguiente: Todos los grupos muestran una mayoría de pacientes con un resultado mayor al 8 % de hemoglobina A1c. Sin embargo, en los grupos 0 y 2, dominan resultados normales para el test de glucosa plasmática, para el grupo 1, también se encuentran pacientes con un resultado mayor a 300 mg/dL. En términos de insulina, todos los grupos presentan un número similar de pacientes con este medicamento.

Tabla 5.8: Información básica de las variables categóricas del Agrupamiento K-Means.

Cluster	**Variable**	**Valor Dominante**
0	Race	Caucasian
0	Gender	Femenino
0	Age	70-80
0	Admission Type Id	Emergencia
0	Discharge Disposition Id	Dado de alta a casa
0	Admission Source Id	Sala de Emergencias
0	Diagnosis 1	Isquémia
0	Diagnosis 2	Enfermedad Hepática
0	Diagnosis 3	Riñón Hipertensivo
0	Change	Cambio en medicación
0	Diabetes Medication	Si
0	Readmitted	No
1	Race	Caucasian
1	Gender	Femenino
1	Age	70-80
1	Admission Type Id	Emergencia
1	Discharge Disposition Id	Dado de alta a casa
1	Admission Source Id	Sala de Emergencias
1	Diagnosis 1	Isquémia
1	Diagnosis 2	Obstrucción Vías Respiratorias
1	Diagnosis 3	Riñón Hipertensivo
1	Change	No hay cambio en medicación
1	Diabetes Medication	Si

Tabla 5.8: (Continuación).

Cluster	Variable	Valor Dominante
1	Readmitted	No
2	Race	Caucasian
2	Gender	Femenino
2	Age	70-80
2	Admission Type Id	Emergencia
2	Discharge Disposition Id	Dado de alta a casa
2	Admission Source Id	Sala de Emergencias
2	Diagnosis 1	Neoplasia Respiratoria
2	Diagnosis 2	Enfermedad Hepática
2	Diagnosis 3	Riñón Hipertensivo
2	Change	No hay Cambio en medicación
2	Diabetes Medication	Si
2	Readmitted	No

5.4.3. Agrupamiento DBSCAN

El agrupamiento DBSCAN generó un total de $k = 13$ grupos, los cuales están distribuidos como se muestra en la figura 5.19. Así mismo, las métricas de rendimiento calculadas dieron los resultados mostrados en la tabla 5.9. Complementando, el índice de Silhouette por cada cluster se muestra en la figura 5.20. En el caso del método de Elbow, al ser un algoritmo que siempre arroja el mismo número de clusters, no se puede realizar la gráfica para distintos valores de k, por lo que no se aplicó el método en este algoritmo.

Las variables continuas, al igual que con los demás agrupamientos, fueron sometidas a varios cálculos estadísticos para diferenciar entre los clusters formados. En la tabla 5.10, se pueden observar los cálculos estadísticos para todos los grupos.

En términos de diagnósticos, varios grupos (1,2,4,5,7,8,9, y 10) tienen comportamiento similar en cuanto a promedio y dispersión, y representan los valores más altos. Similarmente, los grupos -1, 0, 6, y 11, muestran valores un poco más bajos y más dispersos. Ahora, el grupo 3, muestra los valores más bajos. Todos los grupos muestran valores máximos entre

Figura 5.19: Distribución de clusters del Agrupamiento DBSCAN

Tabla 5.9: Métricas de rendimiento del Agrupamiento DBSCAN

Métrica	**Resultado**
Índice promedio de Silhouette	-0.2526
índice de Davies-Bouldin	35.6171
Índice de Calinski-Harabasz	156.2110

3 y 9. El mismo caso ocurre con las emergencias, con los grupos 2, 3, 7, 9, y 10 con valores y dispersión altos. Existe un bloque de grupos que podría llamarse intermedio (1, 4, 5, 8, y 11), y por último, 3 grupos (-1, 0, 6) prácticamente no tienen casos de encuentros con emergencias durante el año anterior y una muy baja dispersión. Para las hospitalizaciones, los grupos 0 y 6 tienen el promedio y dispersión más bajos con hasta 6 y 12 visitas respectivamente. Los grupos 5, 8, y -1, tienen valores bajos, con encuentros de hasta 14 visitas. Continuando, los grupos 4, 11, 1, y 2, podrían considerarse con valores intermedios y una dispersión más amplia. Ahora, los grupos 9, 3, y 7, representan valores más altos que los anteriores grupos, esta vez con una brecha más amplia. Por último, el grupo 10 presenta los valores más altos de promedio y dispersión existiendo otra brecha bastante alta entre los anteriores.

Todos los grupos son bastante similares en términos de procedimientos de laboratorio, exceptuando los grupos 6 y 1, los cuales tienen un promedio un poco más bajo que los demás.

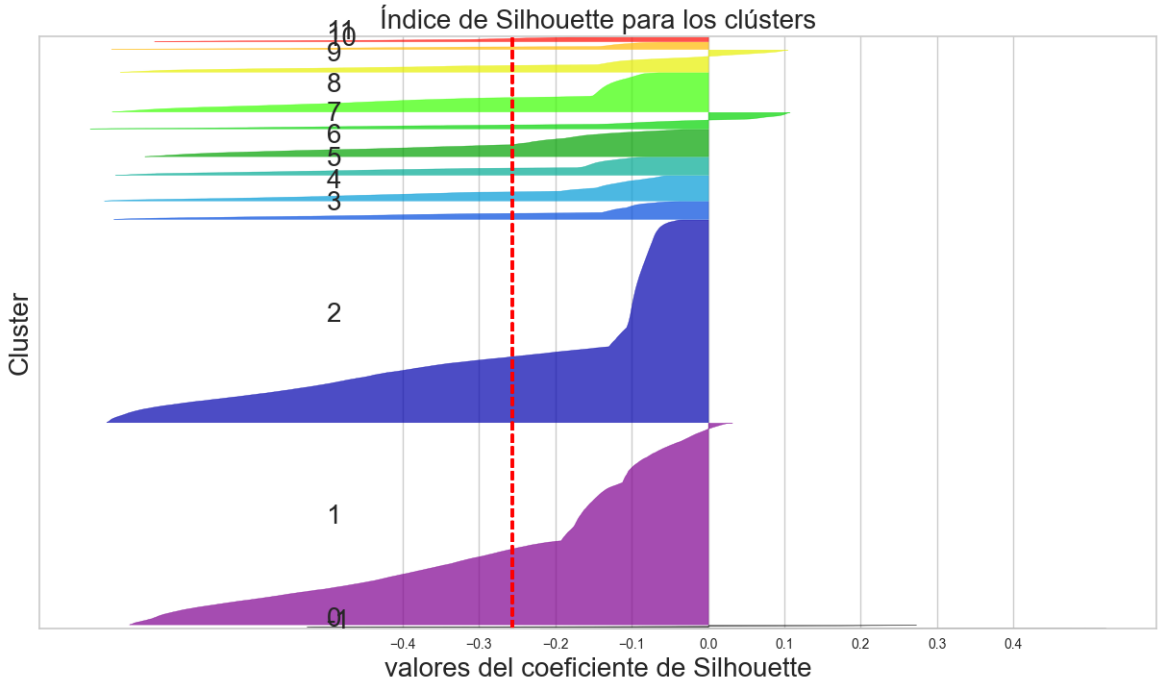


Figura 5.20: Índice de Silhouette para el Agrupamiento DBSCAN

Hablando de medicamentos administrados, existe un incremento gradual entre cada grupo, donde el grupo 8 posee los valores más bajos. Los grupos 11, 9, 5, y 0, serían los siguientes con valores similares entre sí, y se pueden considerar con valores bajos. Los grupos 4, 3, y 6, se consideran intermedios. A medida que incrementa el promedio de los grupos, la brecha que existe entre grupos también lo hace, siendo los grupos 10, -1, y 2, aquellos con valores altos. Los grupos 7 y 1 poseen los valores más altos. Es válido mencionar que el incremento mencionado en la variable de medicamentos es únicamente en el promedio, pues todos los grupos tienen una desviación muy similar. Así, el promedio en los grupos varía desde 11.95 hasta 17.65.

En el caso de las visitas de consulta externa, existen dos bloques de grupos que se comportan de forma similar entre sí. Los grupos -1, 5, 4, 6, 11, 8, y 0, muestran valores bajos de dispersión y promedio, mientras que los grupos 2, 7, 10, 3, 1, y 9, tienen valores altos de promedio, dispersión, y encuentros con hasta 42 visitas. No obstante, para el caso de los procedimientos, la mayoría de grupos (5, 11, 8, 3, 9, 7, 4, 10, y 2) tienen un comportamiento similar y representan los valores más bajos. Además, el grupo -1 posee valores un poco más altos que los mencionados, pero se consideran aún bajos. El grupo 1 se puede considerar un punto intermedio, pues los grupos 0 y 6 muestran los valores altos y más altos respectivamente. Por último, el tiempo en el hospital también se separa en distintos

bloques de grupos, con los grupos 0 y 6 siendo los valores de promedio más bajos, los grupos 8, 11, 5, 9, 4, y 3, con promedio intermedio, y los grupos 10, 7, 1, y 2, con valores altos. Adicionalmente, un único grupo (-1), muestra los valores más altos de promedio. Todos los grupos tienen una dispersión similar para esta variable.

Tabla 5.10: Información básica de las variables numéricas del Agrupamiento DBSCAN.

Cl.	**Variable**	**Máximo**	**Mínimo**	**Desv. Est.**	**Promedio**
-1	Time in Hospital	14	1	3.24	4.95
-1	Lab Procedures	101	1	18.77	48.48
-1	Procedures	6	0	1.63	1.44
-1	Medications	63	1	9.18	15.58
-1	Outpatient	4	0	0.65	0.18
-1	Emergency	2	0	0.20	0.02
-1	Inpatient	5	0	0.84	0.43
-1	Diagnoses	9	2	1.78	6.93
0	Time in Hospital	13	1	2.61	3.64
0	Lab Procedures	101	1	23.45	44.41
0	Procedures	6	0	1.80	1.97
0	Medications	33	1	6.17	13.67
0	Outpatient	8	0	0.98	0.35
0	Emergency	2	0	0.30	0.07
0	Inpatient	6	0	0.71	0.27
0	Diagnoses	9	1	2.04	6.79
1	Time in Hospital	14	1	3.10	4.52
1	Lab Procedures	129	1	19.53	37.19
1	Procedures	6	0	1.82	1.71
1	Medications	81	1	8.97	17.65
1	Outpatient	40	0	1.37	0.44
1	Emergency	76	0	0.89	0.17
1	Inpatient	17	0	1.18	0.60
1	Diagnoses	16	1	1.97	7.24
2	Time in Hospital	14	1	2.98	4.56
2	Lab Procedures	132	1	18.91	48.42

Tabla 5.10: (Continuación).

Cl.	Variable	Máximo	Mínimo	Desv. Est.	Promedio
2	Procedures	6	0	1.61	1.09
2	Medications	75	1	7.69	16.32
2	Outpatient	42	0	1.27	0.36
2	Emergency	64	0	1.03	0.25
2	Inpatient	19	0	1.34	0.68
2	Diagnoses	16	1	1.86	7.59
3	Time in Hospital	14	1	2.79	4.16
3	Lab Procedures	114	1	18.86	44.82
3	Procedures	6	0	1.48	1.03
3	Medications	54	1	6.48	14.57
3	Outpatient	38	0	1.53	0.41
3	Emergency	63	0	1.72	0.34
3	Inpatient	12	0	1.57	1.01
3	Diagnoses	16	1	1.77	1.80
4	Time in Hospital	14	1	2.86	4.13
4	Lab Procedures	106	1	18.81	45.62
4	Procedures	6	0	1.53	1.07
4	Medications	56	1	6.89	14.08
4	Outpatient	25	0	0.93	0.22
4	Emergency	20	0	0.53	0.12
4	Inpatient	13	0	1.04	0.52
4	Diagnoses	16	1	1.96	7.55
5	Time in Hospital	14	1	2.79	4.02
5	Lab Procedures	101	1	18.39	47.01
5	Procedures	6	0	1.49	0.94
5	Medications	47	1	6.43	13.32
5	Outpatient	10	0	0.78	0.21
5	Emergency	8	0	0.48	0.11
5	Inpatient	14	0	0.87	0.39

Tabla 5.10: (Continuación).

Cl.	Variable	Máximo	Mínimo	Desv. Est.	Promedio
5	Diagnoses	13	2	1.91	7.31
6	Time in Hospital	14	1	2.90	3.80
6	Lab Procedures	118	1	17.67	33.90
6	Procedures	6	0	1.83	2.13
6	Medications	58	1	8.02	14.65
6	Outpatient	20	0	0.94	0.22
6	Emergency	8	0	0.27	0.04
6	Inpatient	12	0	0.73	0.29
6	Diagnoses	15	1	2.08	6.76
7	Time in Hospital	14	1	2.90	4.50
7	Lab Procedures	96	1	19.00	46.69
7	Procedures	6	0	1.56	1.06
7	Medications	67	2	7.45	17.01
7	Outpatient	20	0	1.20	0.36
7	Emergency	24	0	1.17	0.34
7	Inpatient	13	0	1.65	1.01
7	Diagnoses	16	1	1.82	7.77
8	Time in Hospital	14	1	2.82	3.96
8	Lab Procedures	126	1	17.75	44.42
8	Procedures	6	0	1.46	1.02
8	Medications	60	1	6.49	11.95
8	Outpatient	20	0	0.99	0.25
8	Emergency	11	0	0.46	0.10
8	Inpatient	13	0	0.91	0.41
8	Diagnoses	16	1	1.90	7.46
9	Time in Hospital	14	1	2.77	4.11
9	Lab Procedures	104	1	18.55	44.25
9	Procedures	6	0	1.47	1.04
9	Medications	49	1	6.61	13.22
9	Outpatient	24	0	1.51	0.50

Tabla 5.10: (Continuación).

Cl.	Variable	Máximo	Mínimo	Desv. Est.	Promedio
9	Emergency	13	0	0.97	0.31
9	Inpatient	18	0	1.53	0.92
9	Diagnoses	16	1	1.70	7.86
10	Time in Hospital	14	1	2.87	4.44
10	Lab Procedures	97	1	18.20	45.85
10	Procedures	6	0	1.52	1.08
10	Medications	53	2	6.73	15.30
10	Outpatient	12	0	1.11	0.40
10	Emergency	13	0	1.06	0.37
10	Inpatient	21	0	2.13	1.47
10	Diagnoses	16	1	1.77	7.74
11	Time in Hospital	14	1	2.68	3.98
11	Lab Procedures	94	1	16.89	50.92
11	Procedures	6	0	1.58	0.95
11	Medications	57	1	7.09	13.11
11	Outpatient	10	0	0.88	0.22
11	Emergency	6	0	0.58	0.14
11	Inpatient	7	0	1.13	0.55
11	Diagnoses	11	1	2.64	6.49

Por último, las categorías más comunes dentro de las variables categóricas de cada cluster se encuentran en la tabla 5.11.

Las variables categóricas de raza, edad, y género, tienen los mismos valores dominantes en todos los grupos (caucásico, 70-80, y femenino respectivamente), exceptuando el grupo 11, el cual muestra mayores encuentros de pacientes de género masculino. Así mismo, todos los grupos tienen en su mayoría encuentros donde el paciente fue dado de alta a casa, con la excepción del grupo -1, donde no se investigó el destino final de los pacientes. Este mismo grupo, junto con el grupo 0, presentan una admisión nula en su mayoría. El grupo 0, también muestra una mayoría de categoría nula en cuanto a la fuente de admisión. Los

grupos 1 y 6, hacen referencia a pacientes que fueron enviados por el doctor (mediante remisión médica). Todos los demás grupos son similares en cuanto a las admisiones y descargas.

En cuanto a los diagnósticos, la mayoría de grupos (-1, 3, 4,, 7, 8, 9, 10, y 11) presentan una dominancia por isquémia como su diagnóstico principal. Otros grupos (0, 1, 2, y 6) muestran neoplasia respiratoria, y por último, el grupo 5, tiene en su mayoría encuentros con neoplasia de riñón como su diagnóstico principal. Para los diagnósticos secundarios, algunos grupos (0, 1, 5, y 6) presentan mayormente diabetes mellitus. Un bloque mayor de grupos (-1, 2, 3, 4, 7, 8, 10, y 11) tienen una dominancia de enfermedad hepática. El grupo 9, difiere de los demás presentando obstrucción de las vías respiratorias. En el caso del diagnóstico secundario auxiliar, todos los grupos presentan una mayoría de riñón hipertensivo.

Para la medicación relacionada con diabetes mellitus, los grupos 0, 6, 8, y 9, poseen en su mayoría encuentros en donde no se administró ningún medicamento de diabetes. Por otra parte, los grupos -1, 1, 2, 3, 4, 7, y 10) muestran lo contrario, y vale la pena mencionar que los grupos 4 y 11 no poseen encuentros donde no se haya administrado algún medicamento relacionado con diabetes. Similarmente, los grupos 0, 1, 2, 4, 5, 6, y 11, no cambiaron la medicación de los pacientes en su mayoría. Los grupos 8, 9, y 10, no mostraron encuentros que hiciesen cambios en la prescripción. Sin embargo, los grupos -1, 3, y 7, sí proporcionan una mayoría de encuentros donde se cambió la medicación.

En la variable de readmisión, la mayoría de grupos (0, 1, 2, 4, 5, 6, y 11) tienen una alta proporción de encuentros que no fueron readmitidos. Los grupos -1, 3, 7, y 9, en cambio, muestran una dominancia por encuentros de pacientes que fueron readmitidos más 30 días después. Así, el grupo 10, muestra en su mayoría casos donde los pacientes fueron re admitidos menos de 30 días después del encuentro.

Para el caso de los test de hemoglobina A1c y glucosa plasmática, la mayoría de grupos (-1, 0, 3, 4, 6, 7, 8, 9, y 10) presentan prácticamente solo casos donde no se realizó el test (existen verdaderamente pocos casos donde sí se tomaron los test). Los grupos 1 y 2, muestra una cantidad apreciable de pacientes cuyos resultados fueron: >8 % (Hemoglobina A1c) y Normal (glucosa plasmática). El grupo 5, presenta casos de pacientes donde el test de hemoglobina dio resultados normales y el grupo 11, tiene una mayoría de encuentros donde sí se realizó el test de hemoglobina y sus resultados fueron superiores al 8 %. En cuanto a los medicamentos relacionados con diabetes, los grupos 3, 4, 10, y 11, muestran

una gran cantidad de encuentros donde se prescribió insulina y se mantuvo la dosis, caso similar al del grupo 7, donde mayormente se prescribió insulina pero a una dosis más baja. Además, el grupo 7 también presenta un gran número de pacientes que se les administró metmorfina y su dosis se mantuvo estable. Similarmente, en el grupo 11, muestra una alta concentración de pacientes que se les administró glipizida y se mantuvo su dosis estable.

Tabla 5.11: Información básica de las variables categóricas del Agrupamiento DBSCAN.

Cluster	**Variable**	**Valor Dominante**
-1	Race	Caucasian
-1	Gender	Femenino
-1	Age	70-80
-1	Admission Type Id	Nulo
-1	Discharge Disposition Id	No Investigado
-1	Admission Source Id	Sala de Emergencias
-1	Diagnosis 1	Isquémia
-1	Diagnosis 2	Enfermedad Hepática
-1	Diagnosis 3	Riñón Hipertensivo
-1	Change	Cambio en medicación
-1	Diabetes Medication	Si
-1	Readmitted	>30 Días
0	Race	Caucasian
0	Gender	Femenino
0	Age	70-80
0	Admission Type Id	Nulo
0	Discharge Disposition Id	Dado de alta a casa
0	Admission Source Id	Nulo
0	Diagnosis 1	Neoplasia Respiratoria
0	Diagnosis 2	Diabetes Mellitus
0	Diagnosis 3	Riñón Hipertensivo
0	Change	No hay cambio en medicación
0	Diabetes Medication	No

Tabla 5.11: (Continuación).

Cluster	Variable	Valor Dominante
0	Readmitted	No
1	Race	Caucasian
1	Gender	Femenino
1	Age	70-80
1	Admission Type Id	Elección
1	Discharge Disposition Id	Dado de alta a casa
1	Admission Source Id	Remisión Médica
1	Diagnosis 1	Neoplasia Respiratoria
1	Diagnosis 2	Diabetes Mellitus
1	Diagnosis 3	Riñón Hipertensivo
1	Change	Cambio en medicación
1	Diabetes Medication	Si
1	Readmitted	No
2	Race	Caucasian
2	Gender	Femenino
2	Age	70-80
2	Admission Type Id	Emergencia
2	Discharge Disposition Id	Dado de alta a casa
2	Admission Source Id	Sala de Emergencias
2	Diagnosis 1	Neoplasia Respiratoria
2	Diagnosis 2	Enfermedad Hepática
2	Diagnosis 3	Riñón Hipertensivo
2	Change	Cambio en medicación
2	Diabetes Medication	Si
2	Readmitted	No
3	Race	Caucasian
3	Gender	Femenino
3	Age	70-80
3	Admission Type Id	Emergencia
3	Discharge Disposition Id	Dado de alta a casa

Tabla 5.11: (Continuación).

Cluster	Variable	Valor Dominante
3	Admission Source Id	Sala de Emergencias
3	Diagnosis 1	Isquémia
3	Diagnosis 2	Enfermedad Hepática
3	Diagnosis 3	Riñón Hipertensivo
3	Change	Cambio en medicación
3	Diabetes Medication	Si
3	Readmitted	>30 Días
4	Race	Caucasian
4	Gender	Femenino
4	Age	70-80
4	Admission Type Id	Emergencia
4	Discharge Disposition Id	Dado de alta a casa
4	Admission Source Id	Sala de Emergencias
4	Diagnosis 1	Isquémia
4	Diagnosis 2	Enfermedad Hepática
4	Diagnosis 3	Riñón Hipertensivo
4	Change	No hay cambio en medicación
4	Diabetes Medication	Si (únicamente)
4	Readmitted	No
5	Race	Caucasian
5	Gender	Femenino
5	Age	70-80
5	Admission Type Id	Emergencia
5	Discharge Disposition Id	Dado de alta a casa
5	Admission Source Id	Sala de Emergencias
5	Diagnosis 1	Neoplasia de Riñón
5	Diagnosis 2	Diabetes Mellitus
5	Diagnosis 3	Riñón Hipertensivo
5	Change	No hay Cambio en medicación
5	Diabetes Medication	Si

Tabla 5.11: (Continuación).

Cluster	Variable	Valor Dominante
5	Readmitted	No
6	Race	Caucasian
6	Gender	Femenino
6	Age	70-80
6	Admission Type Id	Elección
6	Discharge Disposition Id	Dado de alta a casa
6	Admission Source Id	Remisión Médica
6	Diagnosis 1	Neoplasia Respiratoria
6	Diagnosis 2	Diabetes Mellitus
6	Diagnosis 3	Riñón Hipertensivo
6	Change	No hay cambio en medicación
6	Diabetes Medication	No
6	Readmitted	No
7	Race	Caucasian
7	Gender	Femenino
7	Age	70-80
7	Admission Type Id	Emergencia
7	Discharge Disposition Id	Dado de alta a casa
7	Admission Source Id	Sala de Emergencias
7	Diagnosis 1	Isquémia
7	Diagnosis 2	Enfermedad Hepática
7	Diagnosis 3	Riñón Hipertensivo
7	Change	Cambio en medicación
7	Diabetes Medication	Si
7	Readmitted	>30 Días
8	Race	Caucasian
8	Gender	Femenino
8	Age	70-80
8	Admission Type Id	Emergencia
8	Discharge Disposition Id	Dado de alta a casa

Tabla 5.11: (Continuación).

Cluster	Variable	Valor Dominante
8	Admission Source Id	Sala de Emergencias
8	Diagnosis 1	Isquémia
8	Diagnosis 2	Enfermedad Hepática
8	Diagnosis 3	Riñón Hipertensivo
8	Change	No hay Cambio en medicación (únicamente)
8	Diabetes Medication	No
8	Readmitted	No
9	Race	Caucasian
9	Gender	Femenino
9	Age	70-80
9	Admission Type Id	Emergencia
9	Discharge Disposition Id	Dado de alta a casa
9	Admission Source Id	Sala de Emergencias
9	Diagnosis 1	Isquémia
9	Diagnosis 2	Obsturcción Vias Respiratorias
9	Diagnosis 3	Riñón Hipertensivo
9	Change	No hay Cambio en medicación (únicamente)
9	Diabetes Medication	No
9	Readmitted	>30 Días
10	Race	Caucasian
10	Gender	Femenino
10	Age	70-80
10	Admission Type Id	Emergencia
10	Discharge Disposition Id	Dado de alta a casa
10	Admission Source Id	Sala de Emergencias
10	Diagnosis 1	Isquémia
10	Diagnosis 2	Enfermedad Hepática
10	Diagnosis 3	Riñón Hipertensivo
10	Change	No hay cambio en medicación (únicamente)
10	Diabetes Medication	Si

Tabla 5.11: (Continuación).

Cluster	**Variable**	**Valor Dominante**
10	Readmitted	<30 Días
11	Race	Caucasian
11	Gender	Masculino
11	Age	70-80
11	Admission Type Id	Emergencia
11	Discharge Disposition Id	Dado de alta a casa
11	Admission Source Id	Sala de Emergencias
11	Diagnosis 1	Isquémia
11	Diagnosis 2	Enfermedad Hepática
11	Diagnosis 3	Sin Diagnóstico
11	Change	No hay cambio en medicación
11	Diabetes Medication	Si (únicamente)
11	Readmitted	No

Capítulo 6

Discusión y Análisis de resultados

En la tabla 6.1, se pueden observar todas las métricas calculadas para los distintos agrupamientos. Allí, se evidencia la clara superioridad que el algoritmo K-Means tiene sobre los otros métodos de agrupamiento seleccionados. Así mismo, se considera importante mencionar la eficiencia computacional de éste sobre los demás, mostrando los tiempos de procesamiento más cortos. El agrupamiento jerárquico, de igual manera, supera en métricas al algoritmo DBSCAN, con mejores resultados en los tres índices elegidos. No obstante, el agrupamiento jerárquico es una técnica bastante exigente computacionalmente, la cual requiere bastantes recursos para ser utilizada en la totalidad del conjunto de datos, por lo que el algoritmo DBSCAN le supera en este ámbito.

Más allá del rendimiento basado en métricas y eficiencia computacional, un análisis más profundo de los agrupamientos en cada variable sugiere:

- Los grupos generados por el agrupamiento K-Means destacan entre sí en pocas características, siendo muy similares en las restantes. Es decir, los grupos se comportan de manera muy similar en la mayoría de características, diferenciándose principalmente en las variables relacionadas con los diagnósticos variando ya sea el diagnóstico principal o secundario. Así mismo, el número de diagnósticos asociados a cada encuentro demarca la pertenencia a cada grupo. Fuera de esto, variables categóricas como el rango de edad, género, raza, relacionadas con admisiones y descargas, medicamentos, test de hemoglobina A1c, re-admisiones, medicamentos de diabetes, procedimientos, tiempo en hospital y hospitalizaciones previas, se comportan de forma muy similar entre cada grupo y no permiten la identificación de sub-poblaciones más allá de la clasificación por diagnósticos. Este comportamiento, es único del agrupamiento

Tabla 6.1: Métricas de rendimiento de todos los agrupamientos

Agrupamiento	**Resultado**
Índice promedio de Silhouette	
Jerárquico	-0.1889
K-Means	0.4690
DBSCAN	-0.2526
Índice de Davies-Bouldin	
Jerárquico	10.5643
K-Means	0.6830
DBSCAN	35.6171
Índice de Calinski-Harabasz	
Jerárquico	221.8803
K-Means	139025.0573
DBSCAN	156.2110

K-Means, pues los demás algoritmos demarcan diferencias en otras características más allá de las relacionadas a los diagnósticos. éste resultado, se podría atribuir a la reducción de dimensionalidad ortogonal sobre la cual se realizó el agrupamiento, con importante dificultad para explicar la variabilidad de las variables categóricas.

- No obstante, existen diferenciaciones de cada grupo dadas por distintos factores dentro del agrupamiento K-Means. El grupo 2 muestra encuentros con las más bajas tasas de emergencias, consultas externas, número de diagnósticos, además de diferenciarse en su diagnóstico primario "Neoplasia Respiratoria" sobre los demás grupos creados por el algoritmo. No obstante, no es un grupo muy congruente con la teoría médica, pues muestra un diagnóstico con un mal pronóstico en general combinado con bajas tasas en el historial del paciente. Este grupo, contiene cerca del 45 % de todos los encuentros del conjunto de datos. Siendo así, el grupo 0 es el que muestra mayores cantidades de diagnósticos, medicamentos administrados, y un consecuente cambio en la medicación de los pacientes, lo cual, muestra la naturaleza del diagnóstico, cerca del 25 % de los encuentros hacen parte de este grupo. Para finalizar, el grupo

1 destaca en el diagnóstico secundario referente a la obstrucción de vías respiratorias y un resultado de examen de glucosa plasmática superior a 300 mg/dL, lo cual concuerda con las características del diagnóstico, pues los medicamentos necesarios para el tratamiento suelen incrementar los niveles de glucosa en sangre. Un valor aproximado del 35 % de los encuentros, hacen parte de este grupo.

- Hablando del agrupamiento jerárquico, se denota una clara clasificación de encuentros en cada una de las variables continuas, exceptuando el número de diagnósticos y los procedimientos de laboratorio, repitiéndose el caso de estos últimos para todos los algoritmos. El grupo 0 corresponde a los encuentros con los más altos índices de medicamentos administrados, procedimientos realizados (no de laboratorio), y con mayor duración en el hospital, siendo en su mayoría diagnosticados con "Neoplasia Respiratoria" y con un número significativo de ellos teniendo un resultado del test de hemoglobina A1c superior al 8 % y con una prescripción de insulina. Esto, concuerda con el criterio médico debido a la naturaleza del diagnóstico, el cual requiere de medicamentos que inmunosuprimen además de altas dosis de corticoides, lo que puede producir hiperglicemia relacionando el diagnóstico con la diabetes, presentando una fuerte posibilidad de obesidad dentro de la población del grupo. Así mismo, un valor superior al 30 % de los encuentros en el conjunto de datos pertenecen a este grupo. Un caso similar ocurre con el grupo 1, concordando tanto en diagnóstico primario y secundario, pero presentando mayores valores de hospitalizaciones previas y menores en cuanto a los medicamentos y tiempo en hospital. Sin embargo, no es posible atribuir el historial de hospitalizaciones a la reducción de medicamentos, procedimientos, y tiempo en el hospital, pues por su diagnóstico es probable que las hospitalizaciones sean debido a distintas patologías. Adicionalmente, la mitad de los encuentros pertenecen a este grupo. El grupo 2 presenta los niveles más altos en cuanto al historial de los pacientes (hospitalizaciones previas, emergencias, diagnósticos, y consultas externas), y en contraste, niveles bajos de procedimientos y tiempo en el hospital durante el encuentro. Este comportamiento, puede deberse a la particularidad del diagnóstico a la des-compensación, explicando el alto valor de hospitalizaciones, y así mismo la tendencia del paciente a adoptar comportamientos de mayor control de su enfermedad. Cerca del 10 % de los encuentros pertenecen a este grupo. Para el grupo 3, se consideran pacientes en las mejores condiciones del agrupamiento, pues acorde al criterio médico sus medicamentos sugieren diabetes sin

requerimiento de insulina, presentan el menor tiempo en el hospital, procedimientos, y medicamentos, además de una particular mayoría del género masculino en el grupo. Esta sub-población es bastante reducida, con cerca del 1 % de los encuentros. En contraste, el grupo 4 muestra algunas similitudes con el grupo 0, y podría representar a los pacientes con peor pronóstico. Representa a los pacientes con cáncer añadiendo la enfermedad hepática sobre el diagnóstico de diabetes, manteniendo un manejo con insulina y haciendo principal énfasis en la readmisión del paciente al poco tiempo (mayor a 30 días). Cerca del 7 % de los encuentros hacen parte de esta sub-población.

- Para el caso del agrupamiento DBSCAN, los grupos resultantes muestran una identificación más rigurosa de sub-poblaciones dentro de los diagnósticos comunes de pacientes con diabetes mellitus. En primer lugar, existe un bloque de grupos relacionado a los diagnósticos de "Isquemia" y "Enfermedad Hepática" (diagnóstico principal y auxiliar respectivamente), donde los grupos 4, 8, y 11 (aproximadamente el 12 % del conjunto de datos), muestran encuentros con pacientes sin readmisión, un historial de consulta externa, hospitalizaciones, y emergencias intermedio, y un alto número de diagnósticos. Es una población que particularmente no muestra cambio de medicamentos. Específicamente, el grupo 8 sugiere pacientes con baja insulinodependencia y podría representar a los pacientes en mejor estado de esta sub-población. Por otra parte, los grupo 4 y 11 separan pacientes insulinodependientes, donde prevalecen pacientes hombres en el grupo 11, además de una cantidad representativa de encuentros donde el manejo se hace con medicamentos distintos a la insulina. Los grupos -1, 3, y 7 (cercano al 7 % del conjunto de datos), muestran un historial con tasas de hospitalización, emergencia, y consultas más altas, además de mayor tiempo en el hospital y medicaciones. Así mismo, este bloque de grupos se caracteriza por presentar cambio en los medicamentos y una tasa de readmisión mayor a los 30 días, denotando un estado de deterioro mayor del paciente, entre otras razones. Se hace una particular separación de la insulinodependencia del paciente, mostrando pacientes sin insulinodependencia en el grupo -1, pero con las duraciones en el hospital más altas, pudiéndose tratar de otras patologías relacionadas con el diagnóstico. El grupo 3 indica pacientes con un manejo mediante insulina. Así, el grupo 7, incluye tanto pacientes con requerimientos de insulina, como pacientes cuyo manejo se hace con medicamentos distintos. Aunque la dosis de insulina se reduce en este grupo, lo que

sugeriría un estado controlado del paciente, es frecuente la readmisión, concluyendo un mal manejo de diabetes haciendo hincapié en la falta de los test de hemoglobina y glucosa plasmática. Por último, el grupo 10 (>1 % del conjunto de datos) referencia a los pacientes en peor estado, cuya readmisión es menor a los 30 días y muestran insulinodependencia. Esta sub-población muestra grandes similitudes con el grupo 2 generado por el agrupamiento jerárquico a pesar de su diferencia en tamaño. Existe otra sub-población derivada del diagnóstico de isquemia, que reemplaza el diagnóstico secundario por obstrucción de las vías respiratorias y está presente en el grupo 9 (cercano al 4 % del total de encuentros). En este grupo, también muestra readmisión en un periodo mayor a los 30 días, sin embargo, podría deberse a las características propias del diagnóstico pues los pacientes no muestran insulinodependencia o algún otro manejo de diabetes en su mayoría.

- Otra de las poblaciones identificadas a partir del algoritmo DBSCAN, muestran pacientes en un estado controlado, los cuales en su mayoría han sido admitidos por remisión médica y no por una emergencia. Si bien muestran tasas de historial previo, como hospitalizaciones, emergencias, y consultas externas, en ningún caso muestran re admisión. Este bloque de grupos consiste en cerca del 40 % del total de encuentros con los grupos 0, 1, y 6, y su diagnósticos consisten en neoplasia respiratoria y diabetes mellitus. Los grupos 0 y 6, muestran características muy similares en todas las variables, y no presentan medicación relacionada con diabetes. Además, se diferencian en la información disponible sobre las admisiones y descargas. En este caso, son sub-poblaciones que presentan algunas diferencias pequeñas, pero bien podrían ser una sola. En el caso del grupo 1, se diferencia mayormente en la presencia de medicamentos para la diabetes, mayormente por una alta cantidad de pacientes cuyo examen de hemoglobina fue superior al 8 %. Otra sub-población similar a la expuesta se encuentra en el grupo 2, con cantidad de diagnósticos, historial de consultas, emergencias, y hospitalizaciones altas, número de medicamentos, y un tiempo en el hospital bastante altos. Aquí, el diagnóstico secundario cambia a ser “Enfermedad Hepática”, sin embargo, se consideran encuentros diabéticos por la presencia del test de hemoglobina (con resultados también superiores al 8 %) y algunos medicamentos para la diabetes. El grupo 2 contempla cerca del 35 % de los encuentros totales por si mismo.

- Por último, el grupo 5 hace énfasis en pacientes con "Neoplasia de Riñón" como diagnóstico principal, siendo el único algoritmo que identificó esta sub-población. También, el diagnóstico auxiliar contempla "Diabetes Mellitus", sin embargo el test de hemoglobina A1c dió como resultados valores normales en su mayoría. Esta sub-población se puede considerar controlada en términos de diabetes mellitus y se refiere a un poco más del 3 % de los encuentros.

Si bien todos los agrupamientos se realizaron sobre el mismo conjunto de datos, es visible que la determinación de pertenencia a un grupo u otro depende de distintas categorías según el tipo de agrupamiento utilizado. Así, es válido mencionar la sensibilidad de un agrupamiento específico con alguna categoría sobre los demás. No obstante, este comportamiento puede provocar la creación de sub-poblaciones muy similares entre sí. Adicionalmente, y pese al mal desempeño con relación a las métricas establecidas, el algoritmo DBSCAN presenta resultados prometedores en la identificación de sub-poblaciones de diabetes mellitus, especialmente por la utilización de UMAP como reductor de dimensionalidad para todos los tipos de variables identificados en el conjunto de datos.
El agrupamiento jerárquico ha realizado una clara identificación de sub-poblaciones de diabetes mellitus acorde a distintas variables (como diagnósticos, tiempo en hospital, etc.) con una clara separación entre ellas. Sin embargo, los altos costes de procesamiento hacen difícil el uso de esta herramienta en conjuntos de datos médicos (que usualmente son extensos).

Capítulo 7

Conclusiones y Trabajo Futuro

7.1. Conclusiones

Si bien el análisis de métricas proporciona una noción del desempeño de un algoritmo de agrupamiento, se hace necesario el análisis de los grupos creados, sobretodo cuando se realiza agrupamiento sobre conjuntos de datos con alta cantidad de variables categóricas. Así, también se evidencia el beneficio de los índices seleccionados sobre técnicas de agrupamiento cuyos grupos formados sean convexos (diferentes a los generados DBSCAN o agrupamiento jerárquico), donde el agrupamiento K-Means obtuvo excelentes puntajes, pero carecía de sensibilidad respecto a muchas variables tanto categóricas como numéricas. Aunque este resultado podría deberse a la transformación ortogonal por PCA, sigue existiendo un agrupamiento pobre en las variables continuas, denotando poca sensibilidad. El agrupamiento jerárquico muestra grupos diferenciables y congruentes con criterios médicos expertos referentes a la diabetes mellitus y las demás enfermedades relacionadas. Si bien, estos resultados se obtuvieron mediante una muestra representativa del conjunto de datos, se puede considerar un agrupamiento idóneo para la identificación y caracterización de sub-poblaciones, pues presenta sensibilidad a variables y características tanto categóricas como continuas. Ahora, su gran costo computacional puede ser un factor de descarte del algoritmo, pero de igual manera puede servir como un identificador preliminar de sub-poblaciones presentes en un conjunto de datos.

El algoritmo DBSCAN presentó una detección de sub-poblaciones sensible a distintos factores como la readmisión, denotando los distintos estados de los pacientes con el mismo diagnóstico, el grado de insulinodependencia, y el grado de control que tiene un paciente mediante su manejo e historial de admisión y descarga. Aunque creó una gran cantidad

de grupos, pocos podían ser considerados pertenecientes a una misma sub-población. Así mismo, no es complejo computacionalmente y permite el procesamiento de grandes conjuntos de datos. Vale la pena aclarar que estos resultados fueron en gran parte gracias al uso de UMAP como reductor de dimensionalidad, así mismo, se le atribuye la sensibilidad a las variables categóricas. En conjunto, son herramientas con gran aplicación sobre datos epidemiológicos.

Aunque los grupos generados por el algoritmo K-Means tienen un buen rendimiento en términos de métricas, A pesar de gran presencia de pacientes inmunodependientes o con diversos manejos de diabetes, se observan encuentros (y en algunos casos poblaciones específicas), donde el paciente es re admitido después de disminuir su dosis, denotando un mal manejo de diabetes y la falta de exámenes de glucosa plasmática y hemoglobina A1c.

7.2. Trabajo Futuro

En este trabajo se exploraron técnicas de aprendizaje no supervisado con distintos principios de funcionamiento con el fin de identificar su desempeño en conjuntos de datos diversos además de la identificación de poblaciones. También, son técnicas que han sido utilizadas previamente en conjuntos de datos de diabetes. Se anima a continuar la exploración de distintas técnicas no supervisadas para el hallazgo de nuevas sub-poblaciones con conjuntos de datos de mayor envergadura y diversidad en sus variables, tomando como referencia las características y poblaciones identificadas en el presente estudio. Preferiblemente, utilizando conjuntos de datos con mayor diversidad étnica, o distintos conjuntos de datos provenientes de diferentes etnias, pues beneficiaría la identificación de diversidad étnica en algunas características de la diabetes mellitus. Así mismo, las técnicas ya expuestas pueden ser complementadas y potenciadas para incrementar la detección de patrones dentro de distintos conjuntos de datos de diabetes mellitus.

Bibliografía

Adame, L. G. M., Pérez, F. J. G., and Rodrigo, J. A. R. (2002). Diagnóstico y clasificación de la diabetes mellitus, conceptos actuales. *Revista de Endocrinología y Nutrición*, 10(2):62–68.

Ahlqvist, E., Storm, P., Käräjämäki, A., Martinell, M., Dorkhan, M., Carlsson, A., Vikman, P., Prasad, R. B., Aly, D. M., Almgren, P., et al. (2017). Clustering of adult-onset diabetes into novel subgroups guides therapy and improves prediction of outcome. *BioRxiv*, page 186387.

Ahlqvist, E., Storm, P., Käräjämäki, A., Martinell, M., Dorkhan, M., Carlsson, A., Vikman, P., Prasad, R. B., Aly, D. M., Almgren, P., et al. (2018). Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. *The lancet Diabetes & endocrinology*, 6(5):361–369.

Alpaydin, E. (2016). *Machine Learning: The New AI*. The MIT Press Essential Knowledge series. MIT Press.

Atlas, D. (2019). International diabetes federation. idf diabetes atlas, 9th edn. brussels, belgium: International diabetes federation, 2019.

Batool, F. and Hennig, C. (2021). Clustering with the average silhouette width. *Computational Statistics & Data Analysis*, 158:107190.

Bej, S., Sarkar, J., Biswas, S., Mitra, P., Chakrabarti, P., and Wolkenhauer, O. (2022). Identification and epidemiological characterization of type-2 diabetes sub-population using an unsupervised machine learning approach. *Nutrition & Diabetes*, 12(1):27.

Bhatia, K. and Syal, R. (2017). Predictive analysis using hybrid clustering in diabetes diagnosis. In *2017 Recent Developments in Control, Automation & Power Engineering (RDCAPE)*, pages 447–452. IEEE.

Bro, R. and Smilde, A. K. (2014). Principal component analysis. *Analytical methods*, 6(9):2812–2831.

Cai, D., Zhang, C., and He, X. (2010). Unsupervised feature selection for multi-cluster data. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 333–342.

Caliński, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27.

Cam, L. and Neyman, J. (1967). *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability: Weather modification*. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability: Held at the Statistical Laboratory, University of California, June 21-July 18, 1965 and December 27, 1965-January 7, 1966. University of California Press.

Care, D. (2022). Care in diabetes 2022. *Diabetes Care*, 45:S17.

Carrillo-Larco, R. M., Castillo-Cara, M., Anza-Ramirez, C., and Bernabé-Ortiz, A. (2021). Clusters of people with type 2 diabetes in the general population: unsupervised machine learning approach using national surveys in latin america and the caribbean. *BMJ Open Diabetes Research and Care*, 9(1):e001889.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R., et al. (2000). Crisp-dm 1.0: Step-by-step data mining guide. *SPSS inc*, 9(13):1–73.

Chauhan, T., Rawat, S., Malik, S., and Singh, P. (2021). Supervised and unsupervised machine learning based review on diabetes care. In *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*, volume 1, pages 581–585. IEEE.

Chen, W., Chen, S., Zhang, H., and Wu, T. (2017). A hybrid prediction model for type 2 diabetes using k-means and decision tree. In *2017 8th IEEE International conference on software engineering and service science (ICSESS)*, pages 386–390. IEEE.

Cho, S. B., Kim, S. C., and Chung, M. G. (2019). Identification of novel population clusters with different susceptibilities to type 2 diabetes and their impact on the prediction of diabetes. *Scientific reports*, 9(1):1–9.

Davies, D. and Bouldin, D. (1979). A cluster separation measure: Ieee transactions on pattern analysis and machine intelligence. itpidj 0162-8828, pami-1, 2 224–227. *Crossref Web of Science*.

Dennis, J. M., Shields, B. M., Henley, W. E., Jones, A. G., and Hattersley, A. T. (2019). Disease progression and treatment response in data-driven subgroups of type 2 diabetes compared with models based on simple clinical features: an analysis using clinical trial data. *The lancet Diabetes & endocrinology*, 7(6):442–451.

D'Orazio, M. (2021). Distances with mixed type variables some modified gower's coefficients. *arXiv preprint arXiv:2101.02481*.

Dua, D. and Graff, C. (2017). UCI machine learning repository.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231.

Group, N. D. D. (1979). Classification and diagnosis of diabetes mellitus and other categories of glucose intolerance. *diabetes*, 28(12):1039–1057.

Humaira, H. and Rasyidah, R. (2020). Determining the appropiate cluster number using elbow method for k-means algorithm. In *Proceedings of the 2nd Workshop on Multidisciplinary and Applications (WMA)*.

Jameson, J. L., Kasper, D. L., Longo, D. l., Fauci, A. S., Hauser, S. L., and Loscalzo, J., editors (2018). *Harrison, principios de medicina interna, 20e*. McGraw-Hill Interamericana.

Kahkoska, A. R., Geybels, M. S., Klein, K. R., Kreiner, F. F., Marx, N., Nauck, M. A., Pratley, R. E., Wolthers, B. O., and Buse, J. B. (2020). Validation of distinct type 2 diabetes clusters and their association with diabetes complications in the devote, leader and sustain-6 cardiovascular outcomes trials. *Diabetes, Obesity and Metabolism*, 22(9):1537–1547.

Khashei, M., Eftekhari, S., and Parvizian, J. (2012). Diagnosing diabetes type ii using a soft intelligent binary classification model. *Review of Bioinformatics and Biometrics*, 1(1):9–23.

Lance, G. N. and Williams, W. T. (1967). Mixed-data classificatory programs i - agglomerative systems. *Australian Computer Journal*, 1(1):15–20.

Likas, A., Vlassis, N., and Verbeek, J. J. (2003). The global k-means clustering algorithm. *Pattern recognition*, 36(2):451–461.

McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Metzger, B. E., Coustan, D. R., Committee, O., et al. (1998). Summary and recommendations of the fourth international workshop-conference on gestational diabetes mellitus. *Diabetes care*, 21:B161.

Ministerio de Sanidad y Consumo, S. G. T. *CIE-9-MC: Clasificación Internacional de Enfermedades*.

Mujumdar, A. and Vaidehi, V. (2019). Diabetes prediction using machine learning algorithms. *Procedia Computer Science*, 165:292–299.

Murphy, K. (2012). *Machine Learning: A Probabilistic Perspective*. Adaptive Computation and Machine Learning series. MIT Press.

Murtagh, F. and Legendre, P. (2014). Wards hierarchical agglomerative clustering method: which algorithms implement wards criterion? *Journal of classification*, 31(3):274–295.

Nainggolan, R., Perangin-angin, R., Simarmata, E., and Tarigan, A. F. (2019). Improved the performance of the k-means cluster using the sum of squared error (sse) optimized by using the elbow method. In *Journal of Physics: Conference Series*, volume 1361, page 012015. IOP Publishing.

Nallaperumal, S., Bhavadharini, B., Mahalakshmi, M. M., Maheswari, K., Jalaja, R., Moses, A., Anjana, R. M., Deepa, M., Ranjani, H., and Mohan, V. (2013). Comparison of the world health organization and the international association of diabetes and pregnancy study groups criteria in diagnosing gestational diabetes mellitus in south indians. *Indian Journal of Endocrinology and Metabolism*, 17(5):906–909.

Nedyalkova, M., Madurga, S., and Simeonov, V. (2021). Combinatorial k-means clustering as a machine learning tool applied to diabetes mellitus type 2. *International Journal of Environmental Research and Public Health*, 18(4):1919.

Ojugo, A. and Otakore, D. (2018). Improved early detection of gestational diabetes via intelligent classification models: A case of the niger delta region in nigeria. *Journal of Computer Sciences and Applications*, 6(2):82–90.

on the Diagnosis, E. C. and of Diabetes Mellitus, C. (1998). Report of the expert committee on the diagnosis and classification of diabetes mellitus. *Diabetes care*, 21(Supplement_1):S5–S19.

Pandit, S., Gupta, S., et al. (2011). A comparative study on distance measuring approaches for clustering. *International journal of research in computer science*, 2(1):29–31.

Pes, G. M., Delitala, A. P., Errigo, A., Delitala, G., and Dore, M. P. (2016). Clustering of immunological, metabolic and genetic features in latent autoimmune diabetes in adults: evidence from principal component analysis. *Internal and emergency medicine*, 11(4):561–567.

Qi, J., Yu, Y., Wang, L., and Liu, J. (2016). K-means: An effective and efficient k-means clustering algorithm. In *2016 IEEE international conferences on big data and cloud computing (BDCloud), social computing and networking (SocialCom), sustainable computing and communications (SustainCom)(BDCloud-SocialCom-SustainCom)*, pages 242–249. IEEE.

Raihan, M., Islam, M. T., Farzana, F., Raju, M. G. M., and Mondal, H. S. (2019). An empirical study to predict diabetes mellitus using k-means and hierarchical clustering techniques. In *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–6. IEEE.

Rodriguez León, C. (2016). *Adecuación a un procedimiento de minería de datos para guiar la categorización no supervisada*. PhD thesis, Universidad Central Marta Abreu de Las Villas.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Roux, M. (2018). A comparative study of divisive and agglomerative hierarchical clustering algorithms. *Journal of Classification*, 35:345–366.

Russell, S., Norvig, P., and Rodríguez, J. (2004). *Inteligencia artificial: un enfoque moderno*. Colección de Inteligencia Artificial de Prentice Hall. Pearson Educación.

Safai, N., Ali, A., Rossing, P., and Ridderstråle, M. (2018). Stratification of type 2 diabetes based on routine clinical markers. *Diabetes research and clinical practice*, 141:275–283.

Sarría-Santamera, A., Orazumbekova, B., Maulenkul, T., Gaipov, A., and Atageldiyeva, K. (2020). The identification of diabetes mellitus subtypes applying cluster analysis techniques: A systematic review. *International Journal of Environmental Research and Public Health*, 17(24):9523.

Solorio-Fernández, S., Carrasco-Ochoa, J. A., and Martínez-Trinidad, J. F. (2020). A review of unsupervised feature selection methods. *Artificial Intelligence Review*, 53(2):907–948.

Sosenko, J. M., Palmer, J. P., Greenbaum, C. J., Mahon, J., Cowie, C., Krischer, J. P., Chase, H. P., White, N. H., Buckingham, B., Herold, K. C., et al. (2006). Patterns of metabolic progression to type 1 diabetes in the diabetes prevention trial–type 1. *Diabetes care*, 29(3):643–649.

Strack, B., DeShazo, J. P., Gennings, C., Olmo, J. L., Ventura, S., Cios, K. J., and Clore, J. N. (2014). Impact of hba1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. *BioMed research international*, 2014.

Tebar Masso, F. J. and Ferrer Gomez, M. (2009). *La Diabetes Mellitus En La Practica Clinica*. Editorial Medica Panamericana Sa de.

Temurtas, H., Yumusak, N., and Temurtas, F. (2009). A comparative study on diabetes disease diagnosis using neural networks. *Expert Systems with applications*, 36(4):8610–8615.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244.

Xu, R. and Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Transactions on neural networks*, 16(3):645–678.

Zaharia, O. P., Strassburger, K., Strom, A., Bönhof, G. J., Karusheva, Y., Antoniou, S., Bódis, K., Markgraf, D. F., Burkart, V., Müssig, K., et al. (2019). Risk of diabetes-associated diseases in subgroups of patients with recent-onset diabetes: a 5-year follow-up study. *The lancet Diabetes & endocrinology*, 7(9):684–694.

Zou, X., Zhou, X., Zhu, Z., and Ji, L. (2019). Novel subgroups of patients with adult-onset diabetes in chinese and us populations. *The Lancet Diabetes & Endocrinology*, 7(1):9–11.

Apéndice A

Apendices

Comparativa de técnicas de aprendizaje no supervisado para la identificación y caracterización de sub-poblaciones de diabetes mellitus tipo II

Giorgio Enrique Hernández Pineda

Universidad Internacional de la Rioja, Logroño (España)

1 de febrero de 2024

RESUMEN

Se propone la evaluación de tres técnicas de aprendizaje no supervisado en un conjunto de datos de diabetes mellitus, que recopila datos de aproximadamente 100.000 pacientes en más de 130 hospitales en Estados Unidos, con el objetivo de identificar y caracterizar distintas sub-poblaciones. La metodología implementada fue una versión de CRISP-DM acondicionada a los problemas de clasificación no supervisada. Fueron evaluados algoritmos con principios de funcionamiento diferentes, siendo el K-Means, DBSCAN, y agrupamiento jerárquico los seleccionados. Para evaluarlos, se seleccionaron 3 índices diferentes, sin embargo, el análisis de sub-poblaciones resultantes fue el más determinante en la evaluación. Finalmente, el agrupamiento jerárquico es una excelente opción, siempre y cuando la capacidad de cómputo permita utilizarle. Seguido, los algoritmos como K-Means Y DBSCAN requieren de representaciones adecuadas, siendo UMAP la preferida en cuanto a conjuntos de datos con alta cantidad de variables epidemiológicas y gran cantidad de datos.

PALABRAS CLAVE

Aprendizaje no Supervisado, DBSCAN, Agrupamiento Jerárquico, K-Means, Diabetes Mellitus.

I. INTRODUCCIÓN

La diabetes mellitus es un problema de salud pública que afecta a casi 500 millones de personas en todo el mundo, y se proyecta que para el año 2030, este número llegará hasta los 578 millones. Añadiendo, también se considera que al rededor de 193 millones de personas tienen la enfermedad sin estar enterados debido a su naturaleza asintomática [1]. La diabetes, es una enfermedad metabólica que se caracteriza por niveles altos de glucosa en la sangre debido a que el cuerpo no produce la suficiente insulina, o que éste es resistente a los efectos de la misma. También, es asociada con varias complicaciones, como el riesgo a la ceguera, presión arterial, y enfermedades cardiovasculares. Su detección temprana es de extrema dificultad, por lo que se han intentado distintas soluciones utilizando tanto aprendizaje supervisado como no supervisado [2].

Los algoritmos de aprendizaje no supervisado poseen la ventaja de detectar patrones complejos en los datos disponibles y dan la posibilidad de encontrar relaciones entre ellos que no se podrían de otra forma. Por lo tanto, se planeó la comparación de diversas técnicas de aprendizaje no supervisado sobre un conjunto de datos recopilado por más de 130 hospitales en Estados Unidos durante 10 años (1999-

2008). Este conjunto de datos, está disponible en el sitio web de la universidad de California, Irvine, escuela de información y ciencias de la computación *UCI Machine Learning Repository* y fue compartido por el Centro de investigación clínica y traslacional (*Center for Clinical and Translational Research*) en conjunto con la universidad Commonwealth en Virginia [3].

Específicamente, se seleccionaron 3 algoritmos distintos de aprendizaje no supervisado con principios de funcionamiento distinto, los cuales son K-Means, DBSCAN, y agrupamiento jerárquico. Se siguió la metodología CRISP-DM, adecuada a problemas de clasificación no supervisada [4] y se planteó una evaluación comparativa dividida en dos etapas, analizando métricas de rendimiento y un análisis de la distribución e información del contenido de cada grupo. Como resultado, se obtuvo un buen rendimiento por parte del agrupamiento jerárquico y DBSCAN, donde el flujo de procesamiento y reducción de dimensionalidad con UMAP jugó un papel importante.

II. ESTADO DEL ARTE

Numerosos estudios se han realizado al rededor de la temática, siendo al rededor del 65 % implementaciones de algoritmos de aprendizaje supervisado, 25 % de una combinación de aprendizaje supervisado con no supervisado, y por último, un 10 % de aprendizaje no supervisado [2]. De tal manera, se explorarán las más recientes soluciones existentes referentes al aprendizaje no supervisado.

En el estudio [5], se separó una población de pacientes de diabetes mellitus mediante algoritmos de agrupamiento. Entre los datos disponibles, utilizaron la edad, sexo, índice de masa corporal, historial de diabetes mellitus en la familia, entre otros para el agrupamiento. Al haber analizado conjuntos de datos de más de trece países distintos, fue necesario el análisis del componente principal como reductor de dimensionalidad. El algoritmo utilizado fue el K-Means. De tal forma, se realizó primero el agrupamiento en cada conjunto de datos de cada país, para así encontrar diferencias con el agrupamiento global. la distancia euclidiana, el método de elbow y silhouette, fueron utilizados para la selección del número de grupos. Luego, mediante el coeficiente de jaccard, se evaluó la estabilidad de cada grupo. Como resultado, se obtuvieron cuatro grupos donde predominan algunos valores de las características usadas. Así, se concluyó que los grupos pueden separar los pacientes en 4 perfiles que pueden ayudar a identificar una etapa temprana de la enfermedad o factores de riesgo no encontrados en el pasado. Igualmente, pacientes en distintos grupos pueden necesitar un tratamiento o prevención a medida.

En [6], se presenta la división en 6 distintos subgrupos basándose en los principales factores de riesgo conocidos (Edad, sexo, índice de masa corporal, hipertensión, historial de diabetes en la familia) de la diabetes. Para esto, en un conjunto de datos inicial (*Discovery data*), se utiliza agrupamiento jerárquico junto con la distancia de Gower como medida de similitud. El número de grupos a crear fue obtenido mediante el análisis de dendograma. La reproducibilidad de los grupos fue probada con otros 3 cohortes distintos *(HEXA, CAVAS, KNHANES)*. Por último, fue aplicado un modelo basado en máquinas de vectores de soporte para predecir la pertenencia a cualquiera de los 6 grupos basándose en los 5 factores de riesgo. El entrenamiento del modelo se realizó con *Discovery data*, para luego validar la clasificación con los 3 cohortes restantes. Las diferencias entre la prevalencia de la enfermedad a través de los distintos grupos fue altamente reproducible en todos los conjuntos de datos utilizados. A pesar de pequeñas discrepancias con la distribución de los factores de riesgo dentro de los grupos en *Discovery data* y el conjunto de validación, la tendencia en general fue consistente.

Es posible utilizar técnicas de reducción de dimensionalidad para realizar agrupamientos, como en el caso de [7], quienes proponen el uso de UMAP (*Uniform Manifold Approximation and Projection*) para separar grupos distintos de pacientes de diabetes mellitus tipo 2 basándose en información epidemiológica (información dietaria, historial de adicciones, patrones

socio-económicos y de estilo de vida, etc.). Para conseguir grupos relevantes, se agruparon variables continuas, ordinales, y nominales de forma distribuida. Integrando las dimensiones reducidas de cada tipo de característica, se obtuvieron 4 grupos diferentes mediante el uso del algoritmo DBSCAN. Dos de los grupos, representan pacientes no obesos de diabetes mellitus tipo 2.

III. OBJETIVOS Y METODOLOGÍA

El objetivo general para la investigación planteada es: Extracción de características y comparar y evaluar distintas técnicas de aprendizaje no supervisado para el agrupamiento de un conjunto de datos de pacientes de diabetes mellitus tipo II con prevalencia en características nominales. Así mismo, los objetivos específicos son:

- Realizar un análisis exploratorio del conjunto de datos para comprender la distribución y naturaleza de las variables nominales.
- Generar un flujo de datos que permita el pre-procesamiento necesario para el ingreso de los datos a los modelos a comparar.
- Implementación y evaluación de las diferentes técnicas de clustering con los datos pre-procesados.
- Interpretar y evaluar los agrupamientos resultantes de cada algoritmo seleccionado.

Se seleccionó una adecuación de la metodología CRISP-DM para problemas de clasificación no supervisada, presentada en [4].

IV. CONTRIBUCIÓN

A. Conjunto de Datos

El datos, fue resultado de una recopilación y pre-procesamiento realizado en [8], donde inicialmente se contaba con una base de datos de 41 tablas con 117 características, correspondientes a 74'036.643 de visitas y 17'880,231 pacientes únicos. En primera instancia, encuentros de interés fueron seleccionados de la base de datos con 55 atributos. Seguido a esto, se aplicó análisis y pre-procesamiento buscando que los datos contengan suficiente información. Como resultado, se obtuvo un conjunto de datos de 101.766 encuentros y 50 características que cumplen con los criterios definidos previamente. Así mismo, expertos clínicos realizaron la selección de características.

B. Análisis de Datos

Se puede percibir cierta relación entre algunas variables numéricas, como la relación que tienen la cantidad de medicamentos suministrados, tiempo en el hospital, procedimientos y laboratorios. Así mismo, la distribución del número de diagnósticos sugiere un promedio alto para todos los encuentros relacionados con al diabetes, lo que corrobora la teoría médica que indica la variedad de enfermedades subyacentes que los distintos pacientes de diabetes mellitus tipo 2 pueden presentar. En las variables categóricas, se evidenció la cantidad de valores faltantes en las variables *Weight, Payer Code, y Medical Specialty* (97 %, 40 %, y 49 % respectivamente). Además, las variables *A1Cresult y Glucose serum test* presentan una gran mayoría de datos "none". La ausencia del test en los encuentros es bastante cercana al 100 %, indicando la fuerte relación entre la posibilidad de readmisión y los test mencionados.
De igual forma, la exploración del balance de clases en cada variable categórica corrobora el criterio médico, como el caso de la edad y género, donde una mayoría de casos se concentran en pacientes con edades superiores a los 40 años, y así mismo, una proporción mayor del género femenino. Cabe resaltar que el diagnóstico primario no presenta comúnmente la diabetes mellitus (esta se ve más como un diagnóstico secundario, o secundario adicional), razón por la cual es poco frecuente que se efectúe el test de hemoglobina en los encuentros. Claramente, se encuentran algunas de las enfermedades subyacentes como la falla cardiaca,

hipertensión e isquémia.

C. Verificación de Calidad de Datos

Las variables *Payer Code, Weight, y Medical Specialty* se descartaron. Complementando, variables como *Payer Code y Medical Specialty* no tienen mayor relevancia en el agrupamiento, además de que su valor no varía demasiado entre las clases disponibles, siendo pocas las categorías las que se reparten el porcentaje de observaciones restantes. A pesar de que el peso del paciente es una variable considerada relevante en el estudio, se decidió su descarte debido a la alta proporción de valores faltantes (cerca al 97 %).

D. Transformación de Datos

En primer lugar, las variables fueron definidas según su simetría. Seguido, se debió realizar un codificado a las variables categóricas (similar al ya existente en las variables *Admission Type, Admission Source, y Discharge Disposition*) de acuerdo a su simetría.

E. Pre-Procesamiento de los Datos

El agrupamiento jerárquico es una técnica costosa computacionalmente, razón por la cual se realizó con una muestra representativa del conjunto de datos (30.000 muestras, el 30 % del conjunto). Luego, se calculó la distancia de Gower y posteriormente se transformó su salida en una matriz de similaridad. Por último se transformó a una matriz condensada. Para el agrupamiento K-Means, se propuso una transformación ortogonal mediante el algoritmo PCA, donde se activó el parámetro *whiten* que multiplica los vectores por la raíz cuadrada de las muestras, lo que garantiza salidas no correlacionadas con variaciones unitarias de los componentes. Así, se extrajeron los 2 primeros componentes principales, que explicaron el 91,18 % y el 8,82 % de la varianza total del conjunto de datos respectivamente. En el caso del agrupamiento DBSCAN, se utilizó UMAP como una herramienta de reducción de dimensionalidad, con la que se realizó la extracción de embeddings de forma individual para cada tipo de variable. Así, para las variables continuas se utilizó la distancia Euclidiana, en las variables nominales se utilizó la distancia de Hamming, y por último para las variables ordinales se utilizó la distancia de Canberra. Se escogieron las dos primeras dimensiones reducidas de UMAP para las variables continuas y ordinales, mientras que se escogió la primera dimensión para las variables nominales, por lo que con una concatenación de todas las respuestas deja una representación de 5 dimensiones. Los parámetros de *n neighbours* y *min distance* fueron de 30 y 0.1 respectivamente. Se realizó una última reducción de dimensionalidad con UMAP sobre los datos, lo produjo una representación bidimensional.

F. Métricas de rendimiento

F.1. Average Silhouette Width

El índice de Silhouette, es utilizado para medir la calidad de un agrupamiento. Si el valor de ASW es cercano a cero, indica que hay grupos que se están superponiendo entre sí. Si este valor es más cercano a -1, indica que la observación pertenece más a otro grupo que al actual. Por último, un valor cercano a 1 indica una buena clasificación [9].

F.2. Índice de Elbow

El método de Elbow se efectúa de manera visual graficando la suma de distancias cuadradas contra distintos valores de k. Luego, se ubica la cantidad de grupos que posee el clasificador actual para encontrar la suma de distancias cuadradas correspondiente. Se utilizó la gráfica completa del método y de cual es el valor óptimo de k para el clasificador. Además, se utilizaron otras métricas de rendimiento para la selección de k, como el índice de Calinski-Harabasz y el índice de Silhouette [10].

F.3. Índice de Davies-Bouldin

El índice de Davies-Bouldin es una medida de similaridad que compara la distancia entre grupos con el tamaño de los mismos. Se define como la similaridad promedio entre cada

cluster $C_i = 1, ..., k$ y su cluster más similar C_j. Así, el mejor valor posible del índice de Davies-Bouldin es cero, y valores cercanos a cero indican una buena partición [11].

F.4. Índice de Calinski-Harabasz

El índice de Calinski-Harabasz, conocido como criterio de relación de varianza, indica grupos mejor definidos entre más alto sea su valor. Para un conjunto de datos E de tamaño n_E que ha sido partido en k clusters, el índice de Calinski-Harabasz se define como la relación entre la dispersión promedio entre los clustres y la dispersión dentro de los clusters [12].

G. Desarrollo de modelos

G.1. Agrupamiento Jerárquico

En el caso del agrupamiento jerárquico, se realizó el agrupamiento con el método de enlace completo y se dibujó el dendograma. Sin embargo, al ser una muestra de datos grande, se dificulta encontrar un punto para el corte del árbol generado de manera visual, por lo que se decidió seguir el criterio de Elbow para la selección de grupos. Acorde al criterio de Elbow, el número óptimo de clusters fué de $k = 5$.

G.2. Agrupamiento K-Means

Para el algoritmo K-Means, todos los parámetros del clasificador se dejaron en su valor por defecto y se escogió la cantidad de grupos siguiendo tanto el método de Elbow como el de Silhouette iterando el algoritmo con $k = 2, ..., 6$. No obstante, ambos métodos presentaban discrepancias, recomendando un $k = 6$ y un $k = 3$ respectivamente. En la figura 1, se puede observar el resultado del índice de Silhouette junto con la gráfica de los grupos creados por el algoritmo en el mejor de los casos.

G.3. Agrupamiento DBSCAN

El parámetro más importante del agrupamiento DBSCAN, *eps*, fue establecido en 1.1, y

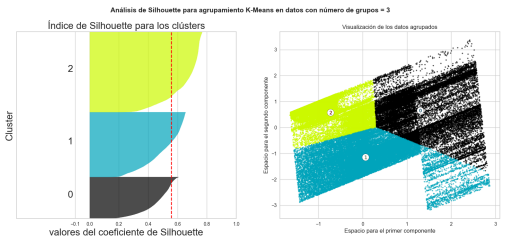


Figura 1: Índice de Silhouette para el mejor caso de k.

representa la distancia máxima entre dos unidades para considerar una en el vecindario de la otra. Otro parámetro representativo fue el *min samples*, configurado en 200, y representa el número de muestras que debe tener un vecindario para considerarlo un núcleo. Estos parámetros fueron definidos a prueba y error, corroborando los resultados del agrupamiento. Una vez configurado el clasificador, detectó 13 distintos grupos en los datos de entrada, los cuales se pueden observar en la figura 2.

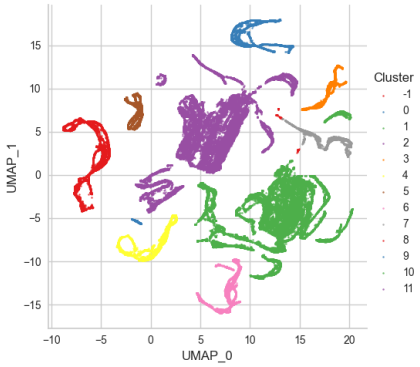


Figura 2: Grupos detectados por DBSCAN en la reducción de dimensionalidad UMAP del conjunto de datos.

V. RESULTADOS

En la tabla 1, se encuentran los resultados de las métricas establecidas para todos los algoritmos. El agrupamiento jerárquico creó un valor de $k = 5$ grupos, los cuales muestran las siguientes características: El historial de emergencias es distinto en el grupo 3, con encuentros sin visitas. Para las hospitalizaciones, los valores de menor a mayor son representados en

los grupos 3, 4, 0, 1, y 2. Caso similar en los medicamentos administrados, con los grupos 1, 3, 2, 0, y 4. Por último, los procedimientos y tiempo en el hospital con los grupos 2, 3, 1, 4, y 0. Para los diagnósticos principales en los grupos, se separan entre "Isquémia" y "Neoplasia Respiratoria". Los diagnósticos auxiliares, con "Diabetes Mellitus", "Enfermedad Hepática", y "Obstrucción de las vías respiratorias". Existe división de grupos dominados por género femenino (0, 1, y 2) y masculino (3 y 4). En su mayoría, los grupos no muestran readmisión, excepto por el grupo 4. En los grupos 0 y 1, se encontró alta presencia de prescripción de insulina cuya dosis se mantuvo. En adición, el grupo 3 muestra alta prescripción de Metmorfina y Gliburida.

En el caso del agrupamiento K-Means, se generaron $k = 3$ grupos, donde: Para la variable diagnósticos, el grupo 2 muestra los valores más bajos, siendo los grupos 1 y 0, los que poseen valores intermedios y altos respectivamente. El resto de variables numéricas tienen comportamientos similares con ligeras variaciones. En términos de cambios de medicación, los grupos 1 y 2 muestran casos donde no hubo cambio, caso contrario al del grupo 1. Para las variables de diagnóstico, los diagnósticos principales son en su mayoría "Isquémia" y "Neoplasia Respiratoria". En cuanto al diagnóstico secundario, predomina mayormente "Enfermedad Hepática" y "Obstrucción de las vías respiratorias". En los grupos 0 y 2, dominan resultados normales para el test de glucosa plasmática, para el grupo 1, se encuentran pacientes con un resultado mayor a 300 mg/dL.

El agrupamiento DBSCAN generó $k = 13$ grupos, donde: Para la variable de consultas externas, los grupos 4, 0, 1, 8, 11, 5, 6, y -1, tienen un comportamiento similar y valores intermedios. El resto de grupos, concentra valores por encima de esto. El caso de variables de diagnóstico, diagnósticos, y procedimientos, muestra comportamientos similares con algunas diferencias. En cuanto a emergencias, hospitalizaciones, medicamentos, y tiempo en hospital, se dividen en valores altos, intermedios, y bajos. También, existe distinción para las variables de género, tipo de admisión, y fuente de admisión entre grupos. Los diagnósticos principales incluyen "Isquémia", "Neoplasia Respiratoria", y "Neoplasia de Riñón". En cuanto al diagnóstico secundario, es más visible la "Diabetes Mellitus", "Enfermedad Hepática", y "Obstrucción de las vías Respiratorias" para el caso del grupo 9. Existen grupos donde predomina tanto la presencia de medicación de diabetes y cambios en la misma como donde no la hay. También, existe clara distinción del manejo de cada encuentro con distintos medicamentos, test, y periodos de readmisión entre grupos.

Cuadro 1: Métricas de rendimiento de todos los agrupamientos

Agrupamiento	**Resultado**
Índice promedio de Silhouette	
Jerárquico	-0.1889
K-Means	0.4690
DBSCAN	-0.2526
Índice de Davies-Bouldin	
Jerárquico	10.5643
K-Means	0.6830
DBSCAN	35.6171
Índice de Calinski-Harabasz	
Jerárquico	221.8803
K-Means	139025.0573
DBSCAN	156.2110

VI. DISCUSIÓN

En la tabla 1, se evidencia un mejor desempeño del algoritmo K-Means sobre los otros métodos de agrupamiento. El agrupamiento jerárquico, de igual manera, supera en métricas al algoritmo DBSCAN. No obstante, el agrupamiento jerárquico es una técnica bastante exigente computacionalmente, por lo que el algoritmo DBSCAN le supera en este ámbito.

En el caso del algoritmo K-Means, los grupos se comportan de manera muy similar en la mayoría de características, diferenciándose principalmente en las variables relacionadas con los diagnósticos. Así mismo, el número de diagnósticos demarca la pertenencia a cada gru-

po. Fuera de esto, variables como el rango de edad, género, raza, relacionadas con admisiones y descargas, medicamentos, test de hemoglobina A1c, re-admisiones, medicamentos de diabetes, procedimientos, tiempo en hospital y hospitalizaciones previas, se comportan de forma muy similar entre cada grupo y no permiten la identificación de sub-poblaciones más allá de la clasificación por diagnósticos.

Hablando del agrupamiento jerárquico, el grupo 0 corresponde a los más altos índices de medicamentos administrados, procedimientos (no de laboratorio), y con mayor duración en el hospital. También, se refiere a diagnósticos como neoplasia respiratoria, con un resultado del test de hemoglobina A1c superior al 8 %, y con una prescripción de insulina. Este diagnóstico, requiere de medicamentos que inmunosuprimen además de altas dosis de corticoides, lo que puede producir hiperglicemia relacionándolo con diabetes y presentando una fuerte posibilidad de obesidad dentro de la población del grupo. En el grupo 1, se presentan mayores valores de hospitalizaciones previas y menores en cuanto a los medicamentos y tiempo en hospital. Sin embargo, no es posible atribuir relación entre estas variables, pues por su diagnóstico es probable que sea debido a distintas patologías. El grupo 2 presenta los niveles más altos en cuanto al historial de los pacientes, y en contraste, niveles bajos de procedimientos y tiempo en el hospital. Este comportamiento, puede deberse a la particularidad del diagnóstico a la des-compensación, y así mismo la tendencia del paciente a adoptar comportamientos de mayor control de su enfermedad. Para el grupo 3, acorde al criterio médico, sus medicamentos sugieren diabetes sin requerimiento de insulina, presentan el menor tiempo en el hospital, procedimientos, y medicamentos, además de una particular mayoría del género masculino. En contraste, el grupo 4 muestra algunas similitudes con el grupo 0, y podría representar a los pacientes con peor pronóstico, siendo los pacientes con cáncer añadiendo la enfermedad hepática sobre el diagnóstico de diabetes, manteniendo un manejo con insulina y haciendo principal énfasis en la readmisión del paciente al poco tiempo (mayor a 30 días).

Para el caso del agrupamiento DBSCAN, los grupos resultantes muestran una identificación más rigurosa de sub-poblaciones dentro de los diagnósticos comunes de pacientes con diabetes mellitus. En primer lugar, existe un bloque de grupos relacionado a los diagnósticos de isquemia y enfermedad hepática, separando encuentros con pacientes sin readmisión ni cambio de medicamentos. Específicamente, el grupo 8 sugiere pacientes con baja insulinodependencia y podría representar a los pacientes en mejor estado de esta sub-población. Por otra parte, los grupo 4 y 11 separan pacientes insulinodependientes, además de una cantidad representativa de encuentros donde el manejo se hace con medicamentos distintos a la insulina. otra sub-población, que se caracteriza por presentar cambio en los medicamentos y una tasa de readmisión mayor a los 30 días, denotando un estado de deterioro mayor del paciente, entre otras razones. Aunque la dosis de insulina se reduce en este grupo, lo que sugiere un estado controlado del paciente, es frecuente la readmisión, concluyendo un mal manejo de diabetes. Por último, otra sub-población referencia a los pacientes en peor estado, cuya readmisión es menor a los 30 días y muestran insulinodependencia. Esta sub-población muestra grandes similitudes con el grupo 2 generado por el agrupamiento jerárquico. Otra de las poblaciones identificadas a partir del algoritmo DBSCAN, muestran pacientes en un estado controlado, los cuales en su mayoría han sido admitidos por remisión médica y no por una emergencia. Por último, el grupo 5 hace énfasis en pacientes con neoplasia de riñón como diagnóstico principal, siendo el único algoritmo que identificó esta sub-población.

VII. CONCLUSIONES

Si bien el análisis de métricas proporciona una noción del desempeño de un algoritmo de agrupamiento, se hace necesario el análisis de los grupos creados, sobretodo cuando se realiza agrupamiento sobre conjuntos de datos con alta cantidad de variables categóricas. También, el agrupamiento jerárquico muestra grupos

diferenciables y congruentes con criterios médicos expertos referentes a la diabetes mellitus y las demás enfermedades relacionadas. Por lo tanto, se puede considerar un agrupamiento idóneo para la identificación y caracterización de sub-poblaciones. Por último, el algoritmo DBSCAN presentó una detección de sub-poblaciones sensible a distintos factores como la readmisión, denotando los distintos estados de los pacientes con el mismo diagnóstico, el grado de insulinodependencia, y el grado de control que tiene un paciente mediante su manejo e historial de admisión y descarga. Vale la pena aclarar que estos resultados fueron en gran parte gracias al uso de UMAP como reductor de dimensionalidad, que en conjunto con DBSCAN, son herramientas con gran aplicación sobre datos epidemiológicos.

Referencias

[1] Diabetes Atlas. International diabetes federation. idf diabetes atlas, 9th edn. brussels, belgium: International diabetes federation, 2019, 2019.

[2] Tannu Chauhan, Surbhi Rawat, Samrath Malik, and Pushpa Singh. Supervised and unsupervised machine learning based review on diabetes care. In *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*, volume 1, pages 581–585. IEEE, 2021.

[3] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[4] Ciro Rodriguez León. *Adecuación a un procedimiento de minería de datos para guiar la categorización no supervisada*. PhD thesis, Universidad Central Marta Abreu de Las Villas, 2016.

[5] Rodrigo M Carrillo-Larco, Manuel Castillo-Cara, Cecilia Anza-Ramirez, and Antonio Bernabé-Ortiz. Clusters of people with type 2 diabetes in the general population: unsupervised machine learning approach using national surveys in latin america and the caribbean. *BMJ Open Diabetes Research and Care*, 9(1):e001889, 2021.

[6] Seong Beom Cho, Sang Cheol Kim, and Myung Guen Chung. Identification of novel population clusters with different susceptibilities to type 2 diabetes and their impact on the prediction of diabetes. *Scientific reports*, 9(1):1–9, 2019.

[7] Saptarshi Bej, Jit Sarkar, Saikat Biswas, Pabitra Mitra, Partha Chakrabarti, and Olaf Wolkenhauer. Identification and epidemiological characterization of type-2 diabetes sub-population using an unsupervised machine learning approach. *Nutrition & Diabetes*, 12(1):27, 2022.

[8] Beata Strack, Jonathan P DeShazo, Chris Gennings, Juan L Olmo, Sebastian Ventura, Krzysztof J Cios, and John N Clore. Impact of hba1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. *BioMed research international*, 2014, 2014.

[9] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[10] H Humaira and R Rasyidah. Determining the appropiate cluster number using elbow method for k-means algorithm. In *Proceedings of the 2nd Workshop on Multidisciplinary and Applications (WMA)*, 2020.

[11] D Davies and D Bouldin. A cluster separation measure: Ieee transactions on pattern analysis and machine intelligence. itpidj 0162-8828, pami-1, 2 224–227. *Crossref Web of Science*, 1979.

[12] Tadeusz Caliński and Jerzy Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27, 1974.