Universidad Nacional Autónoma de México

Facultad de Ciencias

Aplicación de algoritmos de conglomeración para la identificación de patrones en las reflexiones de los profesores sobre sus experiencias en el proyecto *El Aula del Futuro*

T E S I S

que para obtener el título de:

actuaria

P R E S E N T A :

Daniela Andaluz Ramírez

TUTOR

Dr. Gustavo De la Cruz Martínez

Ciudad Universitaria, CDMX, 2021

A mi madre ...

Agradecimientos

Una vida de agradecimiento al Dr. Gustavo de la Cruz Martínez, por el tiempo que dedicó a revisar y guiar este trabajo.

Resumen

La conglomeración es el arte de encontrar grupos en los datos; la conglomeración de textos es una técnica ampliamente utilizada para analizar y extraer información importante de grandes colecciones de documentos, así como para agrupar documentos a partir de características compartidas o patrones. En este trabajo se utilizarán algoritmos de conglomeración para encontrar patrones en las reflexiones de los profesores sobre sus experiencias en el proyecto *El Aula del Futuro*, un proyecto que busca ayudar a lograr un cambio profundo en las estrategias didácticas a partir de las TICS.

Índice general

Índice de figuras

Capítulo 1

Introducción

Últimamente la automatización de tareas se ha vuelto de especial interés en el campo del análisis de datos. Una herramienta no solo apropiada sino que también moderna para automatizar tareas es la Inteligencia Artificial.

Existen dos corrientes que definen a la Inteligencia Artificial, la primera la define como la capacidad de imitar acciones humanas de forma artificial y la segunda ve a la Inteligencia Artificial como el proceso de lograr un objetivo o resolver una tarea de la mejor forma posible. Este trabajo se guiará con la segunda corriente.

El aprendizaje automatizado es una rama de la Inteligencia Artificial basada en la idea de que los sistemas pueden aprender de los datos con mínima intervención humana.

Dentro del aprendizaje automatizado se encuentra el aprendizaje no supervisado, que nos brinda un instrumento para poder automatizar la tarea de encontrar patrones, categorizar y extraer datos importantes: la conglomeración.

La conglomeración es el arte de encontrar grupos en los datos; la conglomeración de textos puede proporcionar información sobre la composición de una colección de documentos y, a menudo, se utiliza como el paso inicial en el análisis de datos.

En este trabajo se intentará encontrar patrones en las reflexiones de los profesores sobre sus experiencias, éxitos y dificultades en el proyecto *El Aula del Futuro*, utilizando algoritmos de conglomeración.

En el capítulo 2 se iniciará con una introducción a la Inteligencia Artificial, al

aprendizaje automatizado y a los distintos tipos de aprendizaje que existen.

En el capítulo 3 se profundizará en algoritmos de conglomeración, en donde se revisarán cinco de los algoritmos más conocidos, además de los conceptos de disimilitudes, similitudes y distancias. También se explicarán formas de encontrar el número óptimo de grupos y la conglomeración de texto, pasando superficialmente por el procesamiento de lenguaje natural.

En el capítulo 4 se presentará el proyecto *El aula del Futuro*, un proyecto que busca ayudar a lograr un cambio profundo en las estrategias didácticas, donde el uso de las TIC sea la piedra angular en la innovación educativa. Dentro de las actividades de este proyecto se compilarán las reflexiones de los profesores sobre sus experiencias en *El Aula del Futuro*. En este mismo capítulo se presentará el planteamiento del problema, hipótesis y metodología de este trabajo. Se procederá a agrupar el conjunto de textos de las reflexiones de los profesores sobre sus experiencias en el proyecto *El Aula del Futuro*. Los resultados se presentarán en este capítulo, así como un análisis de ellos.

En el capítulo 5 se presentarán las conclusiones de este trabajo, en donde se discutirán las razones por las que la hipótesis de este trabajo es cierta o no.

Capítulo 2

Aprendizaje

En los últimos años, temas como Inteligencia Artificial y aprendizaje automatizado se han vuelto muy populares. La razón principal es que los datos han tomado mayor relevancia debido a que cada día se genera una cantidad inimaginable de ellos. La pregunta que surge inmediatamente es: ¿cómo aprovechar todos estos datos? Una de las herramientas que hace posible esta tarea es el aprendizaje automatizado.

Como señala Ian Goodfellow [1], cuando las computadoras recién se inventaron, una de las principales preguntas que nos hacíamos era ¿las máquinas algún día se volverán inteligentes? Hoy, la Inteligencia Artificial se está desarrollando en un campo con muchas aplicaciones como: automatizar tareas rutinarias, entender el lenguaje, reconocer imágenes y hacer diagnósticos en medicina, entre muchos otros.

Goodfellow [1] resalta que en los principios de la Inteligencia Artificial, su principal tarea era resolver de forma rápida problemas que eran difíciles para los humanos, pero relativamente fáciles para las computadoras. Sin embargo, el verdadero reto para la Inteligencia Artificial es resolver tareas que son fáciles para la gente pero difíciles de describir formalmente, problemas que resolvemos instintiva o automáticamente, como reconocer rostros o manejar un automóvil.

2.1. Inteligencia Artificial

Russell y Norvig en su libro *Artificial Intelligence: A Modern Approach* [2] proponen la existencia de cuatro enfoques para definir a la Inteligencia Artificial:

1. Los que se basan en el pensamiento humano: "[La automatización de] actividades que asociamos con el razonamiento humano, actividades como tomar decisiones, resolver problemas, aprender.." [3]

2. Los que se basan en el pensamiento racional: "El estudio de cálculos que hacen posible percibir, razonar y actuar" [4]

3. Los que miden el éxito en términos de fidelidad al comportamiento humano: "El arte de crear máquinas que realicen funciones que requieren inteligencia cuando son realizadas por personas" [5]

4. Los que miden el éxito en términos de un comportamiento ideal, llamado racionalidad: "La inteligencia computacional es el estudio del diseño de agentes inteligentes" [6]

A partir de esto se pueden distinguir dos corrientes. Una corriente abarca la primera y tercera definición, y propone a la Inteligencia Artificial como un proceso o sistema que pretende imitar las acciones de los humanos de forma artificial. La otra corriente abarca la segunda y cuarta definición, y ve a la Inteligencia Artificial como el proceso de lograr un objetivo o resolver una tarea de la mejor forma posible. Este trabajo se guiará por la segunda corriente.

Para hablar de Inteligencia Artificial desde el punto de vista de la racionalidad, primero se deben definir un par de cosas. Siguiendo las ideas de Russell y Norvig [2], un agente es simplemente algo que actúa. Un agente racional, por otro lado, es aquel que actúa para alcanzar el mejor resultado posible.

Una de las diferencias entre un agente y un agente racional, radica en que un agente racional hace las inferencias correctas. Inferir se refiere a extraer un juicio o conclusión a partir de hechos, y una forma de actuar racionalmente es justamente

inferir cuál de todas las posibles formas de actuar va a alcanzar el objetivo deseado, y ejecutar sobre esa inferencia. A pesar de esto, las inferencias no abarcan toda la racionalidad, en algunas ocasiones no existe solución correcta o incorrecta pero aún así se requiere actuar.

Russell y Norvig [2] definen a un agente racional de la siguiente manera: "Para cada posible secuencia de percepciones, un agente racional debe seleccionar la acción que se espera maximice su medida de rendimiento, dada la evidencia provista por la secuencia de percepciones y el conocimiento previamente incorporado que el agente tenga".

El aprendizaje automatizado (el cual se definirá más adelante), se basa en la idea de agentes racionales.

2.2. Aprendizaje

Russell y Norvig [2] proponen al aprendizaje como la capacidad que tiene un ente, computadora o sistema, de mejorar su desempeño en la resolución de una tarea, a partir de su experiencia y conocimiento previo del entorno. Se dice que un agente aprende cuando es capaz de mejorar su desempeño a partir de sus experiencias previas.

A partir de esta definición, es natural que se plantee la siguiente pregunta: ¿Para qué se necesita que un agente aprenda? Si existe una manera óptima de actuar, ¿por qué no simplemente se programa al agente para que ejecute de esa forma? La respuesta es que muchas veces el programador no tiene ni idea de cuál podría ser la solución, tampoco se puede programar de una forma en que se anticipen todas las posibles situaciones a las que el agente se puede enfrentar, y mucho menos se pueden prever todos los cambios que podrían existir a través del tiempo a la hora de programar la solución. [2]

Tom M. Mitchell en su libro *Machine Learning* [7] define al aprendizaje automatizado de la siguiente manera: *"Se dice que un programa de computadora aprende de experiencia E con respecto a algún tipo de tarea T y medida de desempeño P, si su desempeño en la tarea T, medido por P, mejora con la experiencia E"*.

La tarea se refiere al proceso que deseamos lleve a cabo el agente, procesos como clasificación o predicción. Las medidas de desempeño pueden ser, por ejemplo, la **precisión** en el caso de clasificaciones. La **precisión** es la proporción de ejemplos que son clasificados de manera correcta por el agente.

En algunos casos es difícil decidir cómo se medirá el desempeño del agente, en el caso de problemas de regresión, ¿se debería penalizar más por raramente cometer errores grandes o por frecuentemente cometer errores no tan grandes? La respuesta depende del criterio del programador.

Otra definición bastante adecuada es la que da Ian Goodfellow [1]: *"Los sistemas de inteligencia artificial necesitan la habilidad de adquirir su propio conocimiento extrayendo patrones de los datos brutos. Esta capacidad es conocida como aprendizaje automatizado"*. Este trabajo se guiará por esta última definición.

Así, el **aprendizaje automatizado** abarca a agentes racionales que adquieren su propio conocimiento aprendiendo de la experiencia que extraen de los datos.

Según Russell y Norvig [2], dentro del aprendizaje, existen cuatro tipos principales y se distinguen entre ellos por la retroalimentación que reciben: En el **no supervisado** el agente aprende patrones en los datos aunque no haya alguna especie de retroalimentación, por ejemplo, un agente taxista podría desarrollar gradualmente el concepto de *días de tráfico buenos* y *días de tráfico malos* sin que en algún momento se le hayan proporcionado ejemplos etiquetados de cada uno.

En el aprendizaje **por refuerzo** el agente aprende por una serie de recompensas y castigos. Por ejemplo, la falta de propina al final de un viaje le da la señal al agente taxista de que hizo algo mal.

En el aprendizaje **supervisado** el agente tiene acceso a datos etiquetados: observa ejemplos de pares de datos de entrada y salida y aprende una función que asocia los ejemplos de entrada con los ejemplos de salida, por ejemplo, un dato de entrada sería una imagen y un dato de salida diría *eso es un camión*, indicándole al agente qué contiene la imagen.

En la práctica la distinción entre aprendizajes no siempre es clara, por esta razón se definió el aprendizaje **semi-supervisado**. En este aprendizaje se proporcionan

pocos ejemplos etiquetados y se debe hacer lo que se pueda con una colección grande de datos no etiquetados.

2.2.1. Aprendizaje supervisado

Siguiendo a Russell y Norvig [2], el aprendizaje supervisado se refiere al tipo de aprendizaje donde el objetivo del agente es aprender el mapeo entre los datos de entrada y los datos de salida (también conocidos como etiquetas). En este caso, el agente tiene acceso a los datos de entrada y de salida en el momento del aprendizaje, facilitando el desarrollo del mapeo adecuado. Dado un conjunto con N observaciones $(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)$ donde cada y_i es generado por una función $f(x) = y$, el objetivo es aprender f (suponiendo siempre que existe f), o, en su defecto, aproximarse lo mejor posible a f. Aquí, el aprendizaje se refiere a buscar dentro del espacio de posibles soluciones, a la función h que mejor se aproxime a f. Para medir qué tan bien se aproxima h a f se utiliza un conjunto de prueba, distinto al que utiliza el agente para aprender, y se dice que h generaliza bien si asigna correctamente el valor $y = h(x)$ para cada x del conjunto de prueba.

Se tiene interés en saber qué tan bien generaliza el modelo, es decir, qué tan bueno es su desempeño en un conjunto de datos que no ha visto antes porque esto da una pista del desempeño que tendrá cuando se emplee en el mundo real.

Cuando el valor de y está en un conjunto de valores finitos, como {*soleado, nublado, cálido, frío*} se dice que se trata de un problema de **clasificación**, si y es una variable continua, por ejemplo, la estatura de una persona, se dice que es un problema de **regresión**.

2.2.2. Aprendizaje por refuerzo

Este aprendizaje está inspirado en la forma de aprendizaje de los animales. La forma en que un perro aprende a sentarse es un excelente ejemplo: lo premian si se sienta y lo castigan si no se sienta.

Aterrizando esta idea en la inteligencia artificial, el agente debe aprender a llevar a

cabo una tarea mediante prueba y error, donde el único indicador de su desempeño es un castigo si lo hace mal, o una recompensa si lo hace bien, sin ningún tipo de guía de un operador humano. Este tipo de aprendizaje no tiene por experiencia un conjunto de datos, contrario al aprendizaje supervisado. Los algoritmos de aprendizaje por refuerzo interactúan con un ambiente, así que hay un ciclo de retroalimentación entre el ente y sus experiencias. [1]

En otras palabras, la técnica de aprendizaje por refuerzo se refiere al problema de encontrar acciones factibles para ejecutar en una situación con el objetivo de maximizar la recompensa [8].

Regresando al ejemplo de un agente aprendiendo a convertirse en un conductor de taxi. La falta de propina al final del viaje sería un castigo, y da al agente la señal de que hizo algo mal, tal vez condujo demasiado rápido y no se fijó en los topes y baches, o que no respetó los semáforos ni las señales de alto. Recibir propina sería una recompensa, y el agente aprendería que hizo un buen trabajo.

2.2.3. Aprendizaje no supervisado

Haciendo referencia a lo que Christopher Bishop [9] señala, en el aprendizaje no supervisado se cuenta con un conjunto de datos sin su correspondiente etiqueta, contrario al aprendizaje supervisado. Cuando el objetivo es encontrar patrones en los datos y agruparlos respecto a estos, se habla de **conglomeración**. Por otro lado, si se quiere determinar la distribución de los datos, se habla entonces de **estimación de la densidad**, o bien, **visualización** si el objetivo es proyectar datos de una dimensión alta a una más baja con el propósito de visualizarlos.

El tema de algoritmos de conglomeración se retomará más adelante, ya que es parte del tema principal de este trabajo.

2.2.4. Aprendizaje semi-supervisado

En este aprendizaje se cuenta con unos cuantos ejemplos etiquetados y se debe hacer lo que se pueda con una colección grande de ejemplos no etiquetados. Por

ejemplo: intentar construir un sistema que adivine la edad de una persona en una foto. Se toman unos cuantos ejemplos etiquetados preguntándole a las personas su edad. Eso es aprendizaje supervisado. Pero en la realidad, algunas personas mintieron sobre su edad. No es solo que haya un ruido aleatorio en los datos; más bien las inexactitudes son sistemáticas, y descubrirlas es un problema de aprendizaje no supervisado que involucra imágenes, edades autoinformadas y edades reales (desconocidas). Por lo tanto, tanto el ruido como la falta de etiquetas crean una mezcla entre aprendizaje supervisado y no supervisado. [2]

2.3. Resumen

En este capítulo se revisaron conceptos básicos de Inteligencia Artificial para contextualizar al aprendizaje automatizado, mismo que también se revisó de forma básica. Existen cuatro tipos de aprendizaje, pero este trabajo se centrará solamente en uno.

En los capítulos posteriores a este, se explorará con más detalle el aprendizaje no supervisado, en particular, los algoritmos de conglomeración, el tema central de este trabajo.

Una de las ventajas de trabajar con aprendizaje no supervisado, es que no hay resultados correctos o incorrectos, simplemente el mejor resultado es aquel que tiene más sentido de acuerdo al contexto en el que se esté trabajando.

Respecto a los algoritmos de conglomeración, a pesar de que existen medidas para evaluar la calidad de agrupamiento, en la práctica un buen agrupamiento se mide más por la coherencia que tienen los grupos en la vida real que por estas medidas.

Capítulo 3

Algoritmos de conglomeración

Antes de profundizar en el tema de algoritmos de conglomeración, es pertinente repasar las similitudes, disimilitudes y distancias, debido a que los algoritmos de conglomeración se basan completamente en estos conceptos. En este mismo capítulo se repasarán los algoritmos de conglomeración más populares, seguido de una explicación de cómo encontrar el número óptimo de grupos, finalizando con un acercamiento meramente teórico a la conglomeración de texto, con el objetivo de aplicar todo esto a datos reales en el capítulo siguiente.

3.1. Similitudes, disimilitudes y distancias

Esta sección se basa en lo propuesto por Kaufman y Rousseeuw [10]. Las **disimilitudes** son coeficientes no negativos que son cercanos a cero cuando los objetos x y y son bastante parecidos y grandes cuando x y y son poco parecidos, en general la desigualdad del triángulo no se cumple con las disimilitudes.

Una forma muy común de calcular disimilitudes entre los objetos x y y es primero calculando el coeficiente de correlación de Pearson:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}} \tag{3.1}$$

En el numerador de 3.1 las variables se centran alrededor de cero al restarles la media

$\bar{y}$ y $\bar{x}$ respectivamente, y luego los productos cruzados de las variables centradas se acumula. El denominador ajusta la escala de las variables para que ambas estén en las mismas unidades. Así, r es la suma estandarizada y centrada del producto cruzado de dos objetos.

El numerador en la ecuación 3.1 es la covarianza entre los objetos x y y. La covarianza mide la relación lineal entre dos objetos, pero no siempre es útil porque su valor depende de la escala de los objetos. El coeficiente de correlación de Pearson corrige este problema estandarizando ambos para que queden en la misma escala. [11]

Por otro lado, el coeficiente de correlación de Pearson mide la fuerza y dirección de la relación lineal entre dos objetos. Si $r > 0$, quiere decir que si uno crece, el otro crece también. Si $r < 0$ se tiene que si uno crece, el otro decrece y si la correlación es cero, no existe relación lineal entre ellos.

Ahora bien, para calcular qué tan linealmente relacionados están los objetos x y y, se calcula lo siguiente:

$$r(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{3.2}$$

donde

$$x = (x_1, x_2, ..., x_n), y = (y_1, y_2, ..., y_n) \in \mathbb{R}^n,$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

y

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Posteriormente, para calcular la disimilitud $d(x, y)$:

$$d(x,y) = \frac{1 - r(x,y)}{2}$$

Así, objetos con coeficiente de correlación de Pearson positivo, reciben un coeficiente de disimilitud cercano a cero, y objetos con un coeficiente de correlación negativo reciben un coeficiente de disimilitud alto.

Por el contrario, las **similitudes** muestran qué tan similares el objeto x y y son, entre más parecidos son x y y, la similitud $s(x, y)$ se vuelve más grande. Una forma común de calcular similitudes entre dos objetos es también utilizando la correlación de Pearson, pero haciendo algunos ajustes, ya que las correlaciones por sí mismas pueden tomar valores negativos, por lo que hay que eliminar ese efecto [10]:

$$s(x, y) = \frac{1 + r(x, y)}{2}$$

Así, cuando dos objetos x y y tienen correlación $r(x, y)$ grande (cercana a 1), la similitud es también cercana a 1. Si pasa al revés, que la correlación es cercana a -1, la similitud es cercana a 0.

Existe una relación entre similitudes y disimilitudes [10]:

$$d(x, y) = 1 - s(x, y)$$

La **distancia** $dis(x, y)$ también mide qué tan parecido es el objeto x a y, pero además cumple las siguientes condiciones:

1. Nunca es negativa: $dis(x, y) \leq 0$
2. La distancia entre un objeto y él mismo es cero: $dis(x, x) = 0$
3. Es simétrica: $dis(x, y) = dis(y, x)$
4. Cumple la desigualdad del triángulo:

   Siendo x, y, z objetos, $dis(x, y) \leq dis(x, z) + dis(z, y)$

La distancia más utilizada es la distancia euclidiana:

$$dis(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (x_p - y_p)^2}$$

Aunque la distancia Manhattan es ampliamente utilizada también:

$$dis(x, y) = |x_1 - y_1| + |x_2 - y_2| + ... + |x_p - y_p|$$

Es importante mencionar que existe una gran variedad de similitudes, disimilitudes y distancias, pero en este trabajo solo se repasaron las más conocidas.

3.2. Algoritmos de conglomeración

La conglomeración es el arte de encontrar grupos en los datos, y puede ser utilizada no solo para identificar una estructura ya presente en los datos, sino también para imponer una estructura en un conjunto de datos más o menos homogéneo [10] (como se muestra en la figura 3.1).

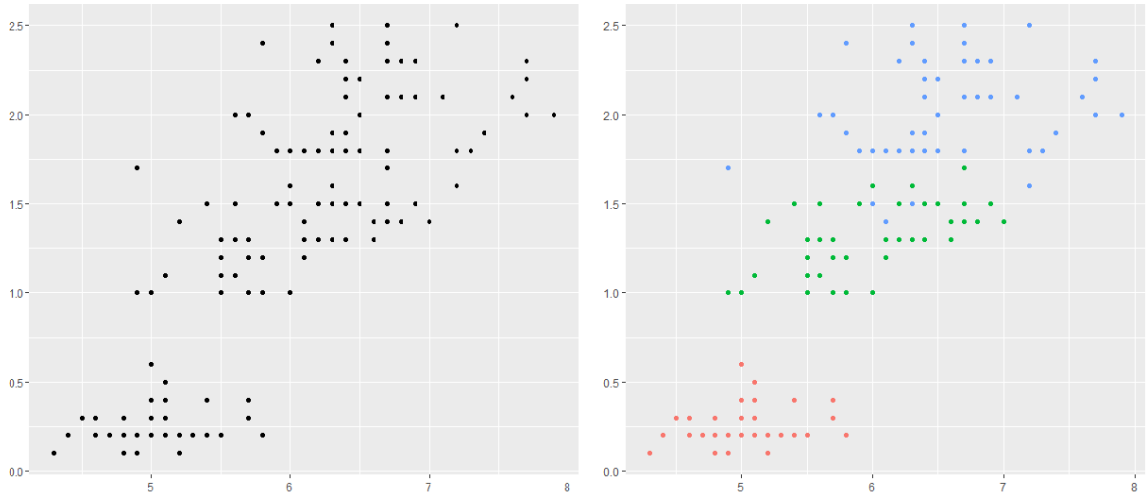

Figura 3.1: Estructura impuesta a datos más o menos homogéneos.

La forma de clasificar a los algoritmos de conglomeración depende completamente del autor. Kaufman y Rousseeuw [10] sugieren que existen dos tipos de técnicas de agrupamiento: **Jerárquicos** y **basados en particiones**. Alboukadel Kassambara en su libro *Practical Guide To Cluster Analysis in R. Unsupervised Machine Learning* [12] propone una tercera: **basados en densidad**.

3.2.1. Algoritmos basados en particiones

De acuerdo con Kaufman y Rousseeuw [10], estos algoritmos construyen k grupos a partir de n objetos, y con las siguientes características:

1. Cada grupo debe contener al menos un objeto.

2. Cada objeto debe pertenecer a exactamente un grupo.

Estas condiciones implican que debe haber, a lo más, un grupo por objeto: $k \leq n$.

Uno de los parámetros que requiere este método de conglomeración es la especificación de cuántos grupos (k) se quiere generar a partir de los datos. Existen métodos para encontrar la k óptima, y se explicarán más adelante.

El objetivo de los algoritmos basados en particiones es agrupar datos de forma que dentro de cada grupo formado, los datos sean lo más parecidos entre sí como sea posible. Dentro de esta técnica, los algoritmos más utilizados son: *k-medias* y PAM (Partitioning Around Medoids).

k-medias

Esta sección se desarrolló usando como guía lo descrito por Hartigan y Wong [13]. El objetivo de *k-medias* es dividir el conjunto de M datos en k grupos de tal forma que la suma de distancias cuadradas entre objetos de cada grupo se minimice, es decir, *k-medias* forma k grupos de tal forma que los elementos entre grupos sean lo más cercanos entre ellos posible.

Paso 0: Se inicializan los centroides (centros de los grupos) aleatoriamente. En la figura 3.2 se muestra un conjunto de datos en color gris, y tres centroides en color rojo inicializados aleatoriamente.

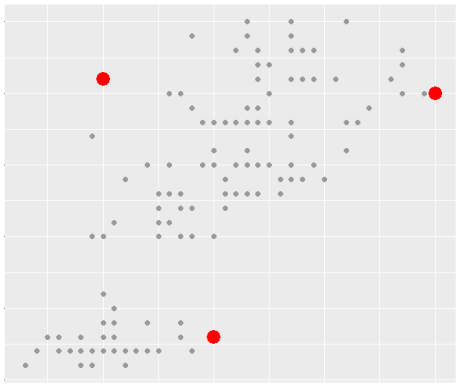

Figura 3.2: Paso 0. Centroides inicializados aleatoriamente.

Paso 1: Para cada elemento i en el conjunto de datos, $i = 1, 2, .., M$ se calcula la distancia euclidiana a todos los centroides, se encuentra el centroide más cercano, y se asigna i al grupo de ese centroide. En la figura 3.3 se observan los datos agrupados, el color de los datos representa el grupo al que fueron asignados, en rojo los centroides.

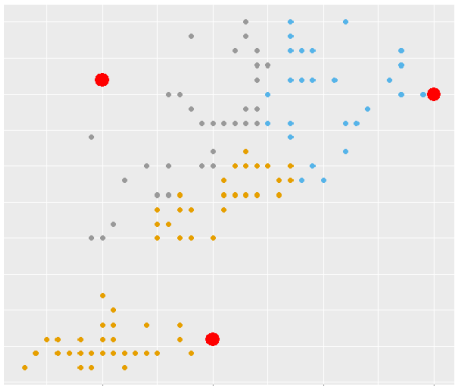

Figura 3.3: Paso 1. Datos asignados al grupo del centroide más cercano.

Paso 2: Una vez que todos los puntos han sido asignados a algún grupo, se actualiza el centroide: el nuevo centroide C_j es la media de todos los objetos contenidos en dicho grupo j (figura 3.4).

$$C_j = \sum_{X \in j} \frac{X}{|j|}, j = 1, 2, ..., k$$

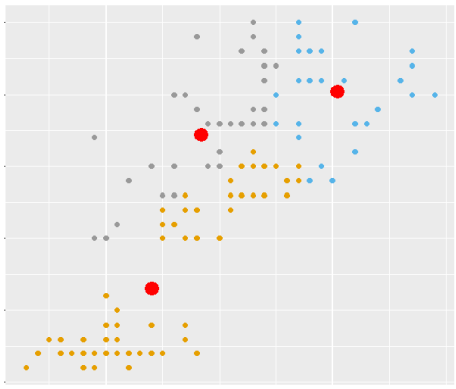

Figura 3.4: Paso 2. En rojo: Centroides actualizados.

Tomando la media como el centroide se minimiza la suma total de distancias cuadradas de cada punto del grupo a su centro [14].

Paso 3: Se repite el paso 1 (figura 3.5).

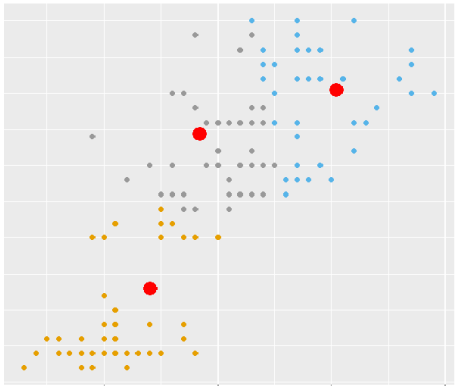

Figura 3.5: Paso 3. Reasignación de grupos después de actualización de centroides.

Los pasos 1 y 2 se repiten hasta que el centroide de cada grupo no cambie respecto a la iteración anterior. El resultado se muestra en la figura 3.1 de lado derecho.

Una ventaja de usar este algoritmo es que es simple y efectivo, y por lo general lleva a buenos agrupamientos. Entre sus desventajas está el hecho de que a pesar de que al iterar los pasos 1 y 2 se llega a un mínimo en la suma de distancias cuadradas entre los puntos y sus centroides, se trata de un mínimo **local**, y no hay garantía de que se trate de un mínimo **global**, por lo que los grupos finales son sensibles a sus centroides iniciales [14].

PAM (Partitioning Around Medoids)

El método de *k-medoides* es parecido a *k-medias* en dividir el conjunto de datos en k grupos. Como señalan Kaufman y Rousseew [10], en *k-medoides*, cada grupo es representado por un elemento del mismo grupo, este elemento es llamado **medoide**, y es considerado como el ejemplo más representativo del grupo. Los demás elementos son asignados al grupo del medoide más cercano.

La forma más común de aplicar este método es mediante el algoritmo PAM. En lo referente a PAM en esta sección, se utilizará lo planteado por Kaufman y Rousseew [10].

El algoritmo se divide en dos fases: **construcción** e **intercambio**. PAM empieza con una selección de k objetos representativos o medoides, y en cada iteración hace un intercambio entre un objeto no seleccionado como medoide y un medoide, solamente si este cambio mejora la calidad de los grupos, la cual se mide con la función de costos [15].

Se definen los siguientes conjuntos: El conjunto de los medoides S, el conjunto de todos los datos O y el conjunto con los datos no seleccionados como medoides $U = O$ - S. El objetivo del algoritmo es minimizar la disimilitud promedio entre los elementos del grupo y el medoide, esto quiere decir, que se intenta minimizar la distancia promedio que hay entre el medoide y los objetos pertenecientes a su grupo. Equivalentemente, y como se hace en la práctica, se puede minimizar la **suma** de las disimilitudes entre los objetos y sus medoides más cercano.

En la primera fase, **construcción**, una colección de k objetos es seleccionada para construir S. En la segunda fase, **intercambio**, se intenta mejorar la calidad del agrupamiento intercambiando medoides por objetos no seleccionados como medoides, es decir, intercambiando elementos de S por elementos de U.

En la fase **construcción** se hace lo siguiente:

1. Para cada elemento i en el conjunto de datos, $i = 1, 2, ..., M$, se calcula la disimilitud con los demás puntos, se encuentra el punto l cuya disimilitud con los demás puntos sea mínima y se inicializa S asignándole l. En la figura 3.6 se muestra un conjunto de datos en negro y en rojo el medoide inicial l.

2. Se considera un candidato $i \in U$ para incluir en S.

3. Se toma un objeto $j \in U | \ j \neq i$ y se calcula la disimilitud D_j entre j y el medoide más cercano.

4. Se calcula

$$C_{i_j} = max\left\{D_j - d(i,j), 0\right\}$$

siendo $d(i,j)$ la disimilitud entre i y j y se calcula la ganancia total g_i que se

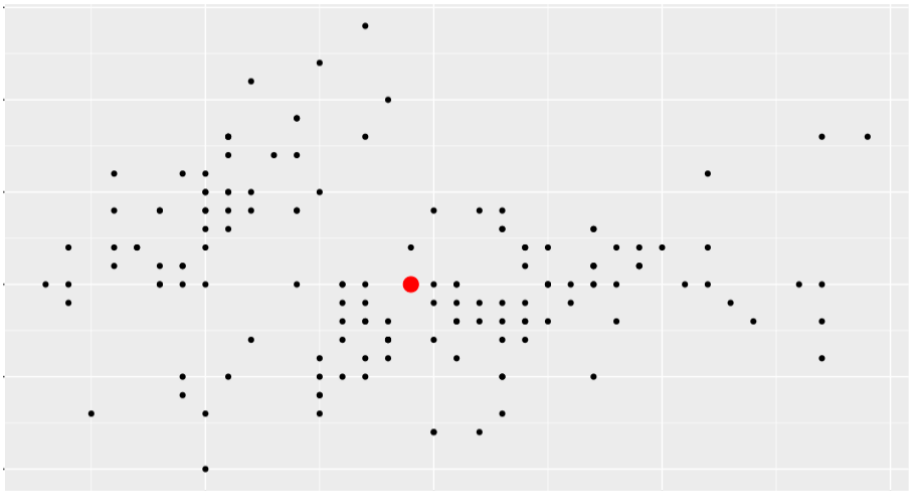

Figura 3.6: Paso 1. Primer medoide seleccionado.

obtiene si se añade a i en S

$$g_i = \sum_{j \in U} C_{i_j}$$

5. Se escoge el objeto $i \in U$ que maximice g_i. En la figura 3.7 se muestra en azul el segundo objeto seleccionado como medoide, en negro el conjunto de datos y en rojo el primer punto seleccionado como medoide.

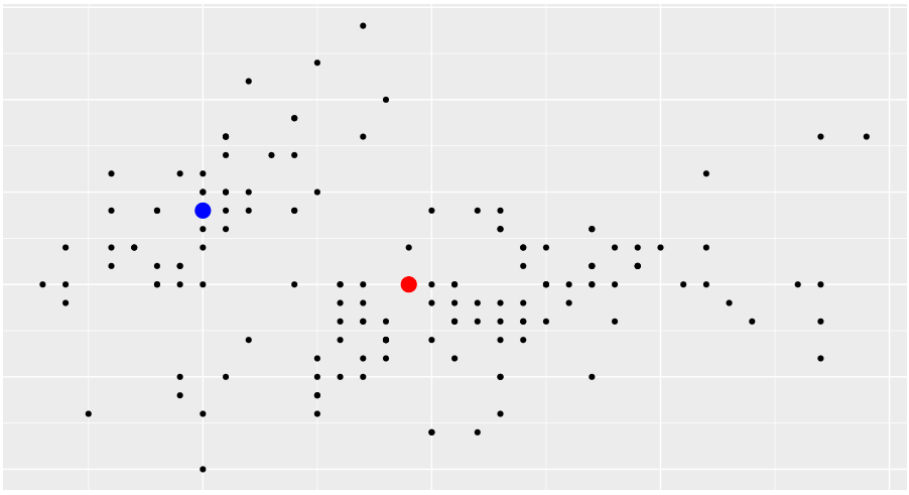

Figura 3.7: Paso 5. Segundo punto seleccionado como medoide.

Y se redefinen S y U de la siguiente forma:

$$S = S \cup \{i\}$$

$$U = U - \{i\}$$

Los pasos dos al cinco se siguen hasta que S contenga k elementos. En la figura 3.8 se muestran tres medoides seleccionados en colores azul, rojo y verde. En negro se muestra el conjunto de datos.

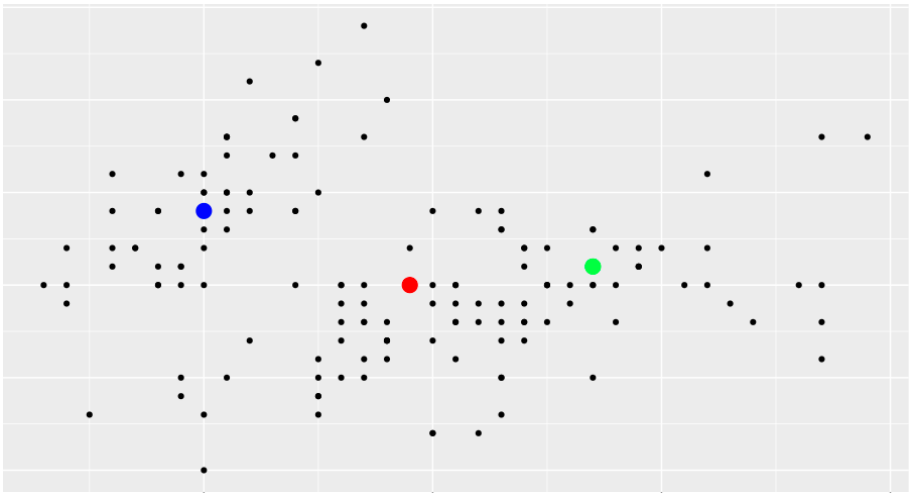

Figura 3.8: Con $k = 3$. Medoides seleccionados en la etapa CONSTRUCCIÓN.

Una vez encontrados los k elementos de S, es momento de empezar la segunda etapa del algoritmo: **intercambio**, cuyo objetivo es mejorar la calidad de los grupos. Hay que tener en mente que el **valor de agrupamiento** determinado por k elementos representativos es **la suma de disimilitudes entre cada objeto del conjunto de datos y el objeto representativo más similar a él**. En esta segunda etapa, se intenta ver el efecto que tiene en el valor de agrupamiento intercambiar un objeto seleccionado como representativo por un objeto no seleccionado. Para calcular el efecto del intercambio de i con h se consideran todos los puntos $(i, h) \in S \times U$, y se realiza lo siguiente:

1. Se toma $j \in U - \{h\}$ y se calcula K_{jih} dependiendo el caso:

   a) Si $d(j, i) > D_j$:

   $$K_{jih} = min\left\{d(j, h) - D_j, 0\right\}.$$

b) Si $d(j,i) = D_j$:

$$K_{jih} = min\{d(j,h), E_j\} - D_j$$

siendo E_j la disimilitud o distancia entre j y el **segundo** medoide más cercano.

2. Se calcula $T_{i,h}$:

$$T_{ih} = \sum_{j \in U} K_{jih}.$$

3. Se selecciona el par $(i,h) \in S \times U$ que minimice T_{ih}.

4. Si el mínimo T_{ih} es negativo, se realiza el intercambio y se regresa al paso 1, si es positivo o cero, quiere decir que el valor objetivo no puede decrecer más intercambiando los puntos y el algoritmo se detiene. En la figura 3.9 se muestran en rojo, verde y azul los medoides seleccionados en la etapa anterior **construcción**. En morado se muestra el objeto h que fue seleccionado en la primera iteración de la etapa **intercambio** y que va a sustituir al medoide rojo. En negro se muestra el conjunto de datos.

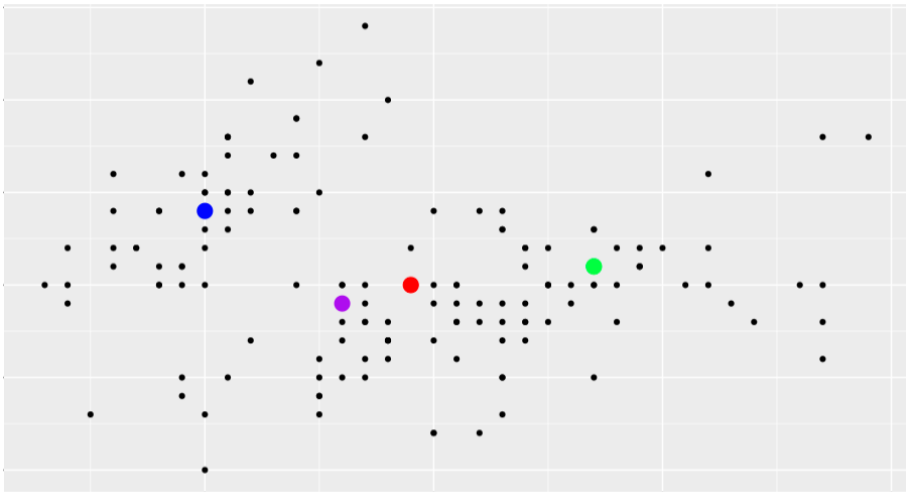

Figura 3.9: Con $k = 3$. Primera iteración de la etapa INTERCAMBIO.

A partir de que el algoritmo se detiene, se asigna cada punto al grupo del medoide más cercano. En la figura 3.10 se muestra un conjunto de datos agrupados con PAM, tomando $k = 3$. En negro se muestran los medoides.

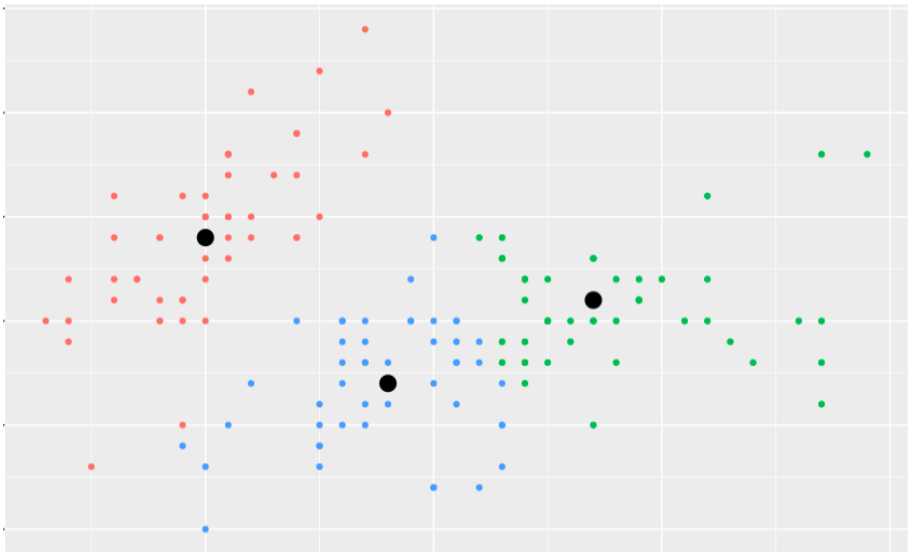

Figura 3.10: Datos agrupados utilizando PAM.

Entre las ventajas de este algoritmo está el hecho de que es más robusto que *k-medias*, esto tiene que ver con la selección del centro del grupo, *k-medias* selecciona el centro del grupo como la media, la cual se ve fuertemente influenciado por valores atípicos. PAM selecciona como centro del grupo el medoide, el objeto con menor disimilitud con los demás objetos del grupo [16].

Entre sus desventajas está que puede llegar a ser computacionalmente costoso si se trabaja con muchos datos, por lo que PAM no es práctico en conjuntos grandes debido a que su principal esfuerzo computacional es buscar entre un conjunto grande de subconjuntos de k objetos, al subconjunto que produzca un agrupamiento localmente óptimo satisfactorio. Conforme el conjunto de datos crece, el número de subconjuntos crece dramáticamente [10].

3.2.2. Métodos jerárquicos

Esta sección se desarrolló tomando como punto de partida lo expuesto por Kaufman y Rousseeuw [10]. Los métodos jerárquicos crean una descomposición jerárquica del conjunto de datos. Estos algoritmos no necesariamente compiten con los métodos basados en particiones, ya que cada uno intenta describir los datos de forma distinta.

Los métodos jerárquicos diseñan una jerarquía en forma de árbol para cada valor

de k posible, de esta forma lidian con todas las posibles particiones en una sola iteración. Esto es, el resultado que arroja este método construye todos los números de particiones posibles, desde $k = 1$ (todos los datos en un solo grupo) hasta $k = n$ (cada uno de los n datos formando su propio grupo).

En consecuencia, los métodos jerárquicos se han encontrado muy útiles en campos como la biología, para clasificar animales y plantas.

Dentro de este método existen dos tipos de algoritmos, los **divisivos** y los **aglomerativos**, y justamente el término **jerárquico** hace referencia al orden con que estos algoritmos tratan a todas las posibles k.

Los algoritmos aglomerativos construyen su jerarquía empezando en $k = n$ (cada objeto forma su propio grupo) y en cada paso van juntando el par de grupos menos disimilar entre sí, hasta terminar en un solo grupo formado por todos los objetos.

La jerarquía de los algoritmos divisivos se construye de forma opuesta a los aglomerativos, ya que empiezan con todos los datos formando un solo grupo y en cada paso dividen un grupo en dos hasta que en el último paso terminan con cada uno de los n objetos formando su propio grupo.

La descomposición generada por estos algoritmos es representada por un **dendograma** (figura 3.11), un árbol que iterativamente divide el conjunto de datos en subconjuntos más pequeños hasta que cada subconjunto consista en un solo objeto. En esta jerarquía, cada nodo del árbol representa un grupo. Un dendograma puede ser creado ya sea de las hojas hacia arriba la raíz (enfoque aglomerativo) o desde la raíz hacia abajo (enfoque divisivo) [17].

Además de su definición de jerarquía, lo que distingue a un algoritmo jerárquico de otro es la forma en que define la **disimilitud entre grupos**, la cual se definirá en la siguiente sección.

Una ventaja de los métodos jerárquicos es que se obtienen todas las posibles particiones en una sola corrida, pero, una vez que este método separó (o unió, según sea el caso) dos grupos, en ninguno de los pasos posteriores los puede volver a juntar (o separar), por lo que no puede corregir sus errores una vez hechos.

Esta rigidez es la clave de su éxito y también de su fracaso, ya que permite generar

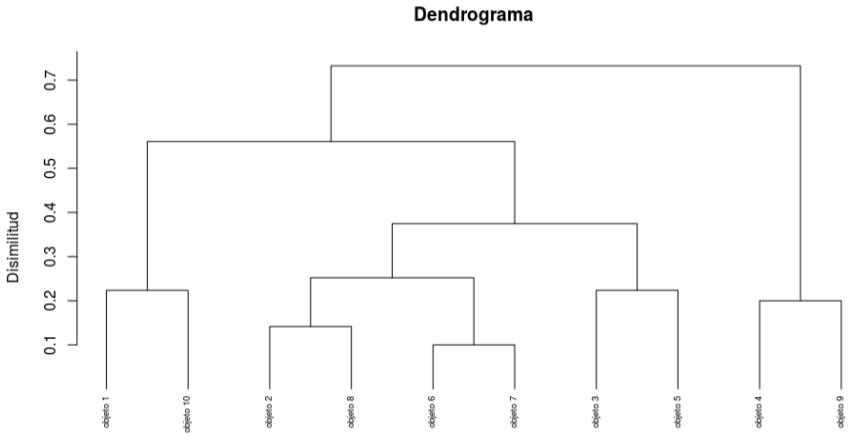

Figura 3.11: Dendograma.

grupos en un tiempo computacional muy corto pero sin la capacidad de corregir decisiones erróneas.

Los algoritmos más populares son:

1. AGNES (AGlomerative NESting)
2. DIANA (DIvisive ANAlysis)

AGNES es un tipo de algoritmo jerárquico aglomerativo y DIANA es un tipo de algoritmo jerárquico divisivo, ambos serán explicados a fondo más adelante.

3.2.3. Disimilitud entre grupos

Según Kaufman y Rousseeuw [10], las tres formas más comunes de definir la disimilitud entre grupos a partir de disimilitud entre objetos son:

1. Media grupal

   La disimilitud entre los grupos Q y P se toma como la **media** de todas las disimilitudes $d(i, j)$ donde $i \in Q$ y $j \in P$

2. Vecino más cercano

   La disimilitud entre los grupos Q y P se toma como el **mínimo** de todas las disimilitudes $d(i, j)$ donde $i \in Q$ y $j \in P$

3. Vecino más lejano

   La disimilitud entre los grupos Q y P se toma como el **máximo** de todas las disimilitudes $d(i, j)$ donde $i \in Q$ y $j \in P$

AGNES (AGlomerative NESting)

Siguiendo lo propuesto por Kaufman y Rouseeuw [10] se tiene lo siguiente.

El Anidado Aglomerativo o AGNES (AGlomerative NESting) es un método jerárquico aglomerativo que se compone de fusiones sucesivas de grupos.

Lo que distingue a AGNES de otros algoritmos de métodos jerárquicos aglomerativos es que AGNES utiliza el método de media grupal para calcular la disimilitud entre grupos.

En el paso 0 cada uno de los n objetos del conjunto de datos está formando un grupo por sí mismo, por lo que en el paso 0 se tienen n grupos.

En pasos posteriores, AGNES básicamente encuentra los dos grupos menos disimilares entre sí y los une en uno solo.

Ahora bien, el sentido común sugiere que en cada paso se recalcule la matriz de disimilitudes entre grupos, ya que tanto el número de grupos como su composición cambia. Existen formas eficientes de hacer estos cálculos sin comprometer el tiempo computacional.

Kaufman y Roussew [10] proponen lo siguiente: si en un paso se unen los grupos A y B para formar el nuevo grupo R, en el siguiente paso la disimilitud del nuevo grupo R con algún grupo Q será:

$$d(Q, R) = \frac{|A|}{|R|} d(A, Q) + \frac{|B|}{|R|} d(B, Q)$$

donde $d(Q, R)$ es la disimilitud entre Q y R, $d(Q, A)$ es la disimilitud entre el Q y A, y $d(Q, B)$ es la disimilitud entre el Q y B.

Lo anterior quiere decir que basta utilizar las disimilitudes que ya se tenían de A y B para calcular las del nuevo grupo R. Esto se puede hacer desde el inicio donde A y B están compuestos por un solo objeto, teniendo que solo es necesario que la

matriz de disimilitudes se calcule una vez, disminuyendo así el tiempo computacional radicalmente.

DIANA (DIvisive ANAlysis)

El Análisis Divisivo o DIANA (DIvisive ANAlysis) es un método jerárquico divisivo que calcula divisiones sucesivas de grupos. DIANA inicia con todos los datos formando un solo grupo y en cada paso siguiente divide un grupo en dos.

Teóricamente, DIANA debe considerar todas las posibles formas de dividir un grupo en dos, lo cual en la práctica es inalcanzable debido al gran número de combinaciones posibles y el gran tiempo computacional que conlleva.

Para mitigar este efecto, Kaufman y Rousseew [10] propusieron para DIANA que en vez de considerar todas las posibles formas en que se pueden generar dos grupos (lo más disimilares entre sí posibles) a partir de uno, primero se debe localizar el objeto del grupo R que en promedio sea más disimilar a los demás, de la siguiente forma:

Tomando el grupo R con intención de dividirlo en dos (grupo A y grupo B), se inicializan los grupos como $A = R$ y $B = \emptyset$. Para cada objeto $i \in A$ se calcula lo siguiente:

$$d(i, A - \{i\}) = \frac{1}{|A| - 1} \sum_{j \in A, i \neq j} d(i, j) \tag{3.3}$$

Si se tiene que el objeto h fue el que maximizó la ecuación 3.3, significa que h es el objeto más disimilar en promedio de A, lo cual lo vuelve el candidato ideal para iniciar un nuevo grupo lo más disimilar a A posible, así que se actualizan los grupos:

$$A = A - \{k\}$$

$$B = B \cup \{k\}.$$

En las siguientes etapas de este paso se intenta mover objetos de A a B buscando que el objeto que se mueva de A a B sea **en promedio más disimilar** a los objetos de A que a los objetos de B.

Así, mientras $|A| \geq 1$, se calcula para cada objeto de A:

$$d(i, A - \{i\}) - d(i, B) = \frac{1}{|A| - 1} \sum_{j \in A, i \neq j} d(i, j) - \frac{1}{|B|} \sum_{h \in B} d(h, i). \tag{3.4}$$

Sea x el objeto que maximizó la ecuación 3.4. Si la ecuación 3.4 evaluada en x es mayor que cero, quiere decir que x es más disimilar en promedio con los objetos de A que con los de B, por lo que tiene sentido mover a x de A a B.

Una vez que se actualiza A y B con el movimiento de x, se busca otro objeto en el nuevo A que maximice la ecuación 3.4.

Esto se repite hasta que el valor máximo de la ecuación 3.4 sea cero o negativo, ya que no tiene sentido mover un objeto de A a B si es más disimilar a los objetos de B que a los de A. Ahí se detiene el proceso de división del grupo R en A y B.

Es importante recordar que en cada paso se divide un grupo. Para elegir qué grupo se va a dividir en ese paso, se calcula para cada grupo R existente en el paso anterior:

$$diam(R) = \max_{i,j \in R} d(i, j) \tag{3.5}$$

y se divide el grupo que maximice la ecuación 3.5, es decir, el grupo más ancho o alargado.

El algoritmo termina cuando cada uno de los objetos forma un grupo por sí mismo. En la figura 3.12 Se muestra el dendograma de un agrupamiento hecho por DIANA. Cada línea de corte en el dendograma muestra cómo se ve el conjunto de cinco datos a esa altura del proceso.

3.2.4. Métodos basados en densidad

Cuando es posible observar un conjunto de datos gráficamente, es fácil detectar grupos y puntos ruidosos, debido a que la densidad de objetos en los grupos es considerablemente más alta que la densidad de objetos fuera de estos, como se muestra en la figura 3.13. En el primer recuadro de la izquierda se muestran grupos convencionales: en forma de esfera. Los algoritmos basados en particiones y jerárquicos encuentran

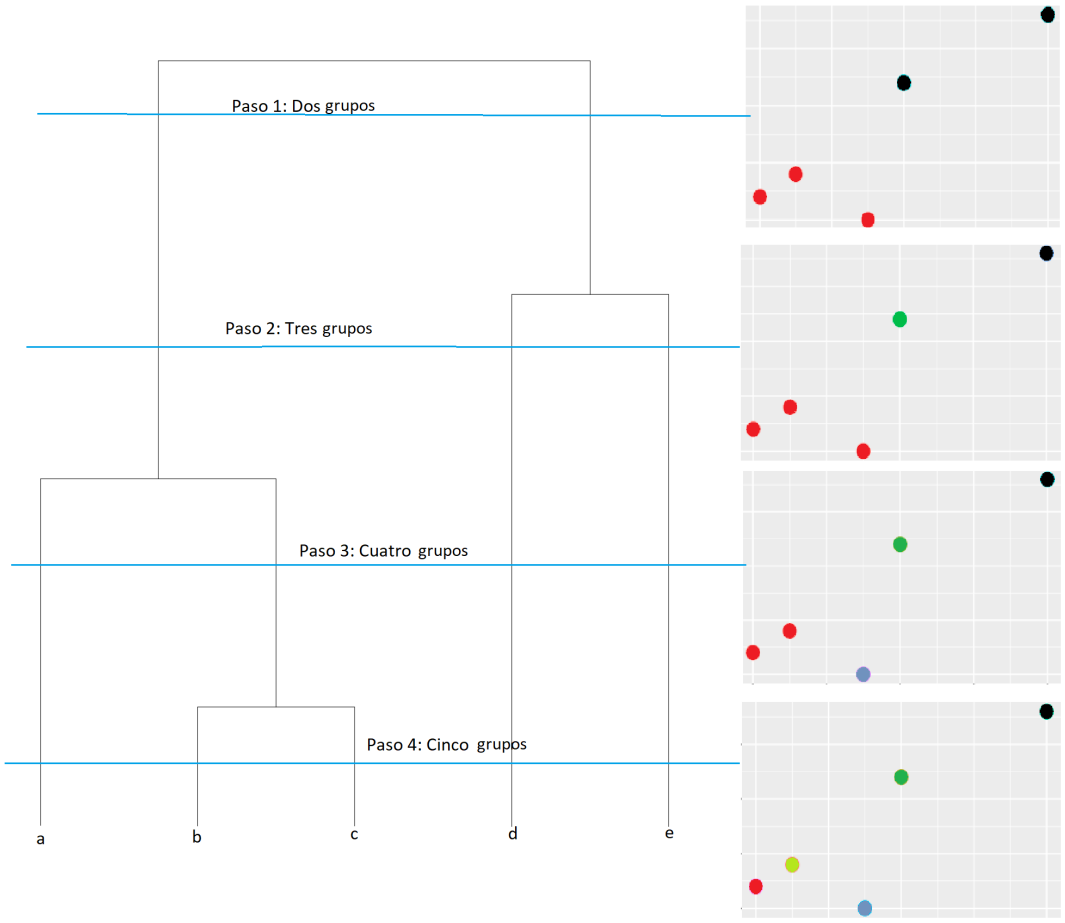

Figura 3.12: Dendograma ilustrando un agrupamiento de DIANA.

este tipo de grupos con gran facilidad [17]. En el recuadro de la derecha se muestran grupos de formas arbitrarias o menos uniformes.

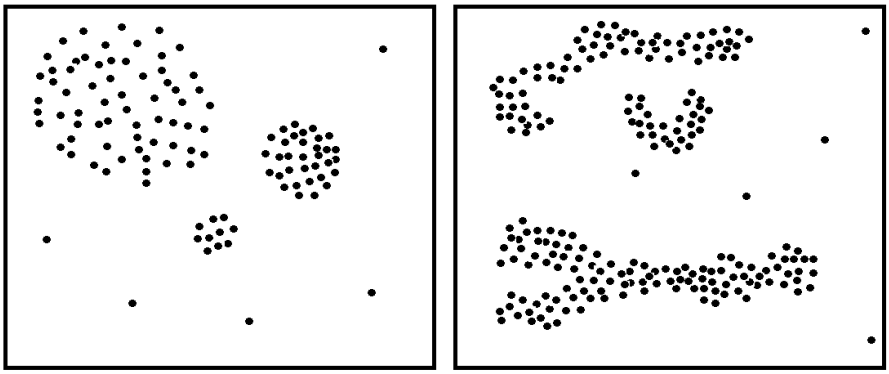

Figura 3.13: Grupos de distintas formas.

Los algoritmos basados en densidad proponen que los **grupos** son regiones con alta densidad de datos separadas por áreas de poca densidad en el espacio de los datos [18].

Este enfoque permite encontrar grupos de formas arbitrarias, además de dar la ventaja de no necesitar indicar previamente el número de grupos, ya que no se asume alguna distribución paramétrica o se minimiza algún tipo de medida para encontrar buenos grupos [17], contrario a los algoritmos revisados anteriormente.

Los algoritmos basados en densidad también manejan adecuadamente el ruido, donde los objetos que viven en áreas de baja densidad no son asignados a un grupo, sino que se tratan como datos atípicos [18]. En los algoritmos basados en particiones y jerárquicos estos datos sí son asignados a algún grupo.

3.2.5. Noción de grupos basados en densidad

Antes de profundizar en la teoría de grupos basados en densidad, hay que definir algunos conceptos importantes, basados en lo propuesto en: *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise* [17].

Dado D un conjunto de datos, $x \in D$ y $r > 0$, se define la **vecindad de radio** r (r-vecindad) con centro en x como:

$$B_r(x) = \{y \in D | dis(x, y) < r\}$$

donde $dis(x, y)$ es una distancia entre x y y.

Se define **punto denso**, **punto frontera** y **punto ruidoso**:

1. Punto denso:

   x es un punto denso si $|B_r(x)| \geq minPts$ donde $minPts$ es especificado por el usuario y es un número estrictamente positivo [18].

2. Punto Frontera:

   x es un punto frontera si no es un punto denso pero está en la vecindad de uno [18].

3. Punto Aislado

   x es un punto aislado o ruidoso si no es un punto denso ni un punto frontera [18].

En la figura 3.14 se muestra un grupo. El punto rojo representa un punto denso, el punto azul representa un punto frontera y el punto verde representa un punto ruidoso.

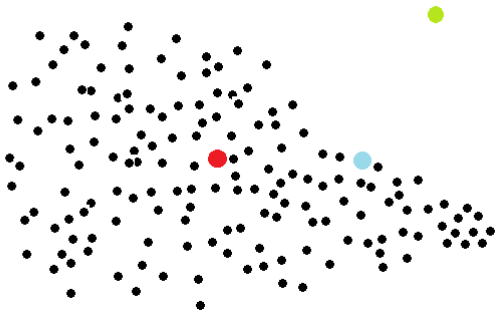

Figura 3.14: Tipos de puntos en un grupo.

Se definen ahora algunas nociones de cercanía entre puntos:

Un punto $y \in A$ es **directamente densidad-alcanzable** de $x \in A$ respecto a r y $minPts$ sii:

1. $|B_r(x)| \geq minPts$

2. $y \in B_r(x)$

Es decir, si x es un punto denso y y está en su r-vecindad [17].

En la figura 3.15 se muestra un conjunto de puntos. El punto rojo representa un punto denso y el punto azul un punto frontera. Aunque el punto rojo está en la vecindad del punto azul, y el punto azul en la vecindad del rojo, solo el punto azul es directamente densidad-alcanzable del punto rojo, al revés no se cumple porque la densidad-alcanzabilidad directa solo es simétrica si se trata de dos puntos densos.

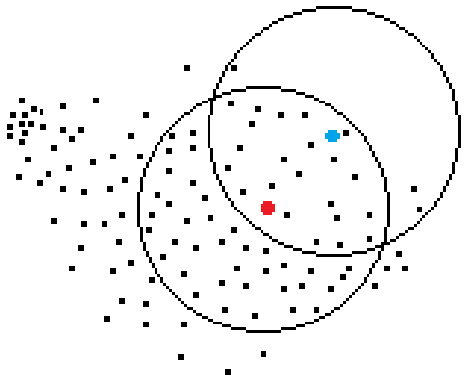

Figura 3.15: El punto azul es directamente densidad-alcanzable del punto rojo.

Un punto y es **densidad-alcanzable** de x si existe en D una secuencia ordenada de puntos $(p_1, ..., p_n)$ con $x = p_1$ y $y = p_n$ tal que $\forall\ i \in \{1, 2, ..., n-1\}$ p_{i+1} es directamente densidad-alcanzable de p_i [17].

En la figura 3.16 se muestra un ejemplo de esta definición, donde el punto verde es densidad-alcanzable del punto rojo, si se toma p_1 = punto rojo, p_2 = punto azul y p_3 = punto verde. La condición se cumple ya que p_3 es directamente densidad-alcanzable de p_2 y p_2 es directamente densidad-alcanzable de p_1.

Si un punto es directamente densidad-alcanzable de otro, también es densidad-alcanzable. La densidad-alcanzabilidad es una extensión de la densidad-alcanzabilidad directa [17].

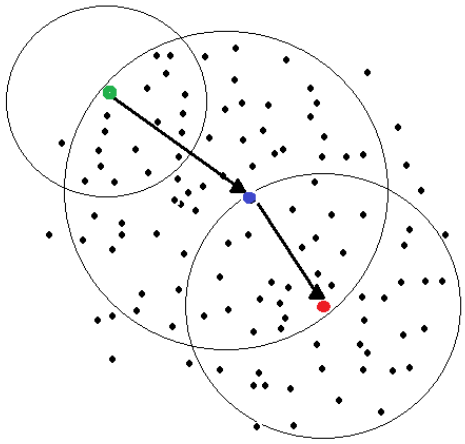

Figura 3.16: Ejemplo de puntos densidad-alcanzables.

Dos puntos frontera del mismo grupo pueden no ser densidad-alcanzables uno del otro debido a que la condición de ser punto denso no se cumple para alguno de ellos (de hecho para ambos). Sin embargo, debe haber un punto denso en el grupo para el cual ambos puntos frontera sean densidad-alcanzables de él, por lo que se define el siguiente concepto:

Un punto $x \in A$ es **densidad-conectado** a un punto $y \in A$ si existe un punto $z \in A$ tal que tanto x como y sean densidad-alcanzables de z. En la figura 3.17 los dos puntos negros son densidad-conectados [17].

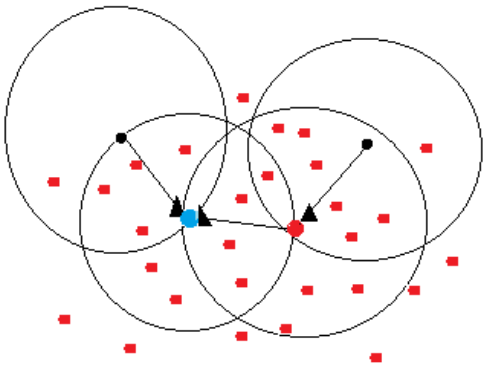

Figura 3.17: Ejemplo de dos puntos densidad-conectados.

La densidad-conectividad es una relación simétrica [17].

Definidos estos conceptos se puede intuir que un grupo es un conjunto de pun-

tos densidad-alcanzables y el ruido es simplemente un conjunto de puntos que no pertenecen a ningún grupo. Formalmente se definen de la siguiente forma:

1. Sea D un conjunto de datos. Un **grupo** C con respecto a r y $MinPts$ es un subconjunto no vacío de D que cumple las siguientes condiciones [17]:

   a) $\forall x, y$: si $x \in C$ y y es densidad-alcanzable de x con respecto a r y $MinPts$, entonces $y \in C$ (Maximalidad).

   b) $\forall x, y \in C$: x es densidad-conectado a y con respecto a r y $MinPts$ (Conectividad).

2. Sean $C_1, ..., C_k$ los grupos de D. Se define el **ruido** como el subconjunto de puntos de D que no pertenecen a ningún grupo [17]:

$$ruido = \{x \in D | \forall i, x \notin C_i\}.$$

Una vez definido lo anterior, se puede introducir el algoritmo basado en densidad más conocido: DBSCAN.

DBSCAN (Density Based Spatial Clustering of Applications with Noise)

DBSCAN es un algoritmo de conglomeración basado en densidad que encuentra grupos y ruido de acuerdo a las definiciones anteriores. DBSCAN necesita dos parámetros: r y $MinPts$. Para entender cómo funciona DBSCAN, hay que enunciar los siguientes dos lemas:

Lema 1: Sea $p \in D$ y $|B_r(p)| \geq MinPts$. Entonces el conjunto $O = \{x \in D \mid x$ es densidad-alcanzable de p con respecto a r y $MinPts\}$ es un **grupo** con respecto a r y $MinPts$ [17].

Es decir que si se toma un punto denso del conjunto de datos y se recuperan todos los puntos que son densidad-alcanzables de este punto, el conjunto de puntos densidad-alcanzables que se forme cumple las condiciones para ser un grupo.

Lema 2: Sea C un grupo respecto a r y $MinPts$ y sea $x \in C$, x un punto denso. Entonces C es igual a $O = \{y \in D \mid y$ es densidad-alcanzable de x con respecto a r y $MinPts\}$ [17].

Es decir, si se toma un grupo C y un punto denso x dentro de él, el grupo O que forme al recuperar todos los puntos densidad-alcanzables de x (O es grupo por lema 1) va a ser igual a C sin importar qué punto denso x ($x \in C$) seleccione.

Así, no hay manera única de determinar un grupo a partir de alguno de sus puntos densos, cada punto en el grupo es densidad alcanzable de cualquier punto denso del grupo y por esto un grupo contiene exactamente los puntos que son densidad-alcanzables de cualquier punto denso arbitrario [17].

El algoritmo empieza con un punto arbitrario p y recupera toda su vecindad. Si p es un punto denso, DBSCAN va a empezar un grupo recuperando los puntos del r-vecindario de p, como se muestra en la figura 3.18. Aquí, p es el punto rojo y los puntos verdes son los puntos recuperados de la vecindad de p.

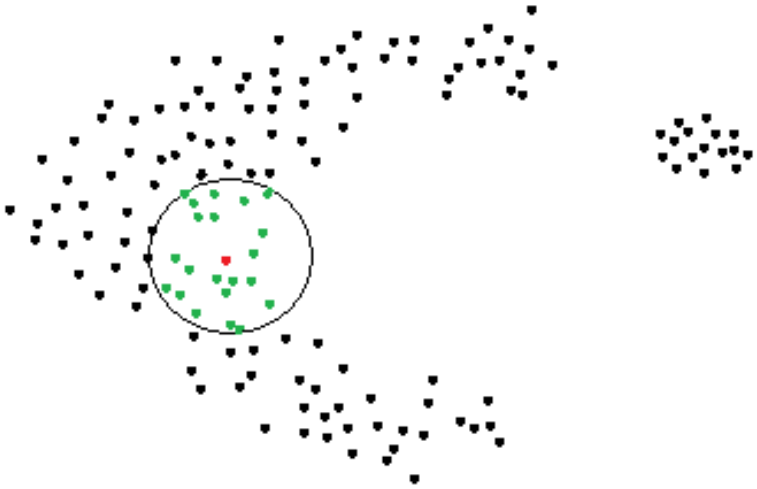

Figura 3.18: Vecindad recuperada de un punto denso.

Si encuentra un punto denso q dentro de la vecindad de p, se expande la vecindad de p incluyendo los puntos en la vecindad de q, como se muestra en la figura 3.19, donde el punto rojo es q y los puntos verdes corresponden a los puntos que están en

la vecindad expandida.

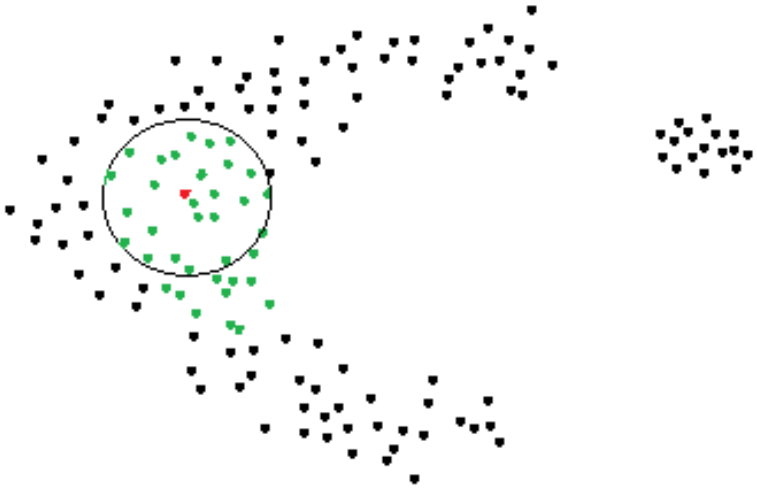

Figura 3.19: Vecindad expandida.

Se repite lo anterior hasta que no se encuentren más puntos densos en la vecindad expandida, entonces el grupo está completo y se busca en los puntos restantes en el conjunto de datos para ver si otro punto denso puede ser encontrado para iniciar un nuevo grupo [18].

Después de procesar todos los puntos en el conjunto de datos, los puntos que no fueron asignados a algún grupo se consideran puntos ruidosos.

Las ventajas de usar DBSCAN para construir grupos son: no se requiere ningún tipo de conocimiento sobre los datos para determinar los parámetros del algoritmo. DBSCAN incluso propone parámetros que considera adecuados, por lo que el usuario puede o no introducir parámetros. DBSCAN es capaz de descubrir grupos con formas arbitrarias, además de ser eficiente en bases de datos grandes [17].

Entre sus desventajas está la forma en que asigna a los puntos frontera a un grupo, ya que un punto frontera puede ser densidad-alcanzable de puntos densos de distintos grupos y DBSCAN lo asigna al primero de estos grupos que es procesado, que depende en el orden de los datos [18]. Además, DBSCAN es incapaz de encontrar grupos de datos de densidad variable [18], ya que usa parámetros fijos globales para todos los

grupos.

3.3. Número ideal de grupos

Theodoridis y Koutroubas [19] propusieron tres enfoques para investigar la validación de grupos.

El primero es basado en criterios externos, que consiste en comparar los resultados obtenidos por los algoritmos de conglomeración con resultados externos ya conocidos, como proveer las etiquetas reales de cada dato.

El segundo enfoque se basa en criterios internos, que utilizan la información obtenida desde el proceso de conglomeración para evaluar qué tan bien los resultados del análisis se ajustan a los datos sin referencia a información externa.

En el tercer enfoque se evalúan los resultados del algoritmo con otros resultados de la aplicación del mismo algoritmo o de otro, con diferentes parámetros.

En la práctica son más utilizados los métodos del tercer enfoque, y la mayoría consiste en tomar alguna medida de calidad de conglomeración y calcularla para cada número de grupos posible, posteriormente graficar los resultados contra el número de grupos, y elegir visualmente el número de grupos óptimo dependiendo de qué medida de calidad se eligió. Dentro de estos métodos se tiene a la *Silueta Promedio* y al *Método del codo*, mismos que se utilizarán en este trabajo.

3.3.1. Silueta promedio

Para calcular las siluetas [10] se toma un objeto i y se denota como A el grupo al que i pertenece.

Se calcula la disimilitud promedio $a(i)$ del objeto i con los demás objetos de su grupo:

$$a(i) = \frac{\sum d(i,j)}{|A|}, j \in A$$

Luego se calcula la disimilitud promedio $b(i)$ de i con los objetos del grupo más

cercano C:

$$b(i) = \min_{C \neq A} \frac{\sum d(i,j)}{|C|} j \in C$$

Una vez obtenidos $a(i)$ y $b(i)$ se puede calcular la **silueta** de i:

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

Si $s(i)$ está cercana a 1, quiere decir que la *disimilitud interna* $a(i)$ es mucho más pequeña que la *disimilitud externa* $b(i)$. Entonces i está bien clasificado: la segunda mejor opción de grupo no está tan cerca como la elección real. Si $s(i)$ está cercana a 0 entonces $a(i)$ y $b(i)$ son casi iguales, y no está claro si i debió haber sido asignado al grupo en el que está o a su segunda mejor opción. La peor situación sería si $s(i)$ es cercana a -1, entonces $a(i)$ es mucho más grande que $b(i)$, significando que i en promedio es más distinto a los objetos de su grupo que a los objetos de su segundo grupo más cercano.

La **silueta promedio** $\hat{s}$ se calcula entonces como el promedio de las siluetas de los n objetos que fueron agrupados.

Con lo anterior, el **número óptimo de grupos** es el que **maximiza** la silueta promedio en un rango de valores posibles para k, como se muestra en la gráfica 3.20:

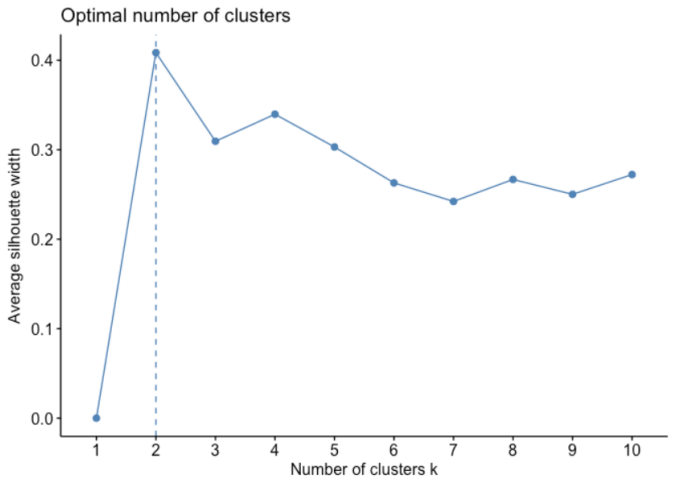

Figura 3.20: Silueta Promedio: El número ideal de grupos es 2.

3.3.2. Método del codo

El método del codo es un método que analiza el porcentaje de varianza explicada (suma de distancias cuadradas de cada punto al centro de su grupo) en función del número de grupos [20].

Se calcula la suma de distancias cuadradas SSE de cada punto x al centro de su grupo C_k para cada k [21]:

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} ||x - C_i||_2^2$$

El SSE se calcula para todas las k y se grafica contra el número de grupos.

Al principio, las primeras k agregarán mucha información pero en algún momento, la ganancia marginal caerá drásticamente y dará un ángulo en la gráfica. El número óptimo de grupos será este punto [20].

En la gráfica 3.21 se muestra un ejemplo de este método, en donde la k óptima es $k = 4$:

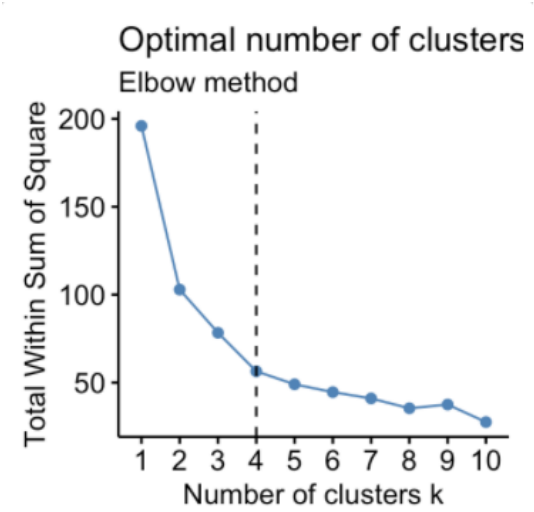

Figura 3.21: Método del codo: El número ideal de grupos es 4.

3.4. Conglomeración de texto

El Procesamiento de Lenguaje Natural (PLN) es una rama de la Inteligencia Artificial que brinda a las máquinas la capacidad de leer, comprender y derivar el significado de los lenguajes humanos [22].

El objetivo de conglomerar texto es el mismo que se tiene al conglomerar datos numéricos: agruparlos de acuerdo con la similitud que existe entre ellos. Para conglomerar texto, este se debe transformar a vectores, ya que los algoritmos de conglomeración solo funcionan con datos numéricos.

Para poder representar texto de forma numérica, se necesita hacer uso de algunas herramientas de PLN, como tokenización, remoción de palabras vacías (*stop words*), cortar las palabras para recuperar su *Stem* y TF-IDF, mismas que se explicarán más adelante.

A pesar de que no existe una forma estándar para procesar textos, este trabajo será guiado por lo que hicieron Marutho y otros [21].

Antes de procesar el texto, primero se procederá a limpiarlo: transformar mayúsculas a minúsculas, quitar caracteres especiales y números.

Después se hará lo siguiente:

1. Tokenizar

   Separar un texto o una oración en unidades más pequeñas, como frases o palabras. En este trabajo se separará por palabras.

2. Quitar palabras vacías (*stop words*)

   Las palabras vacías son aquellas que son muy comunes en el lenguaje y cuyo significado no agrega valor al texto, como conectores de oraciones, preposiciones y pronombres.

3. Cortar las palabras para recuperar su *Stem* (*Stemming*)

   Se refiere al proceso de cortar el final de la palabra con la intención de considerar a todas las derivaciones de una palabra como una misma, por ejemplo: tanto *dirección* como *direccionar* se transforman a *direcc*.

4. Utilizar el método TF-IDF para mapear el texto a vectores

   Se genera un vocabulario con las palabras únicas que aparecen en el texto. Posteriormente, con este vocabulario se crea una matriz, en donde cada columna representa una palabra del vocabulario, y los renglones representan a los textos u oraciones. El valor de cada columna va a ser TF-IDF calculado para cada palabra de cada texto.

   Para calcular TF-IDF para una palabra se hace lo siguiente [21]:

   - Frecuencia del término (tf)

     Se calcula tomando el número de ocurrencias l de la palabra en un texto y se divide entre el número total de palabras t en el texto:

     $$tf = -\frac{l}{t}$$

   - Frecuencia inversa del documento (idf)

     Se calcula tomando el número de textos D en el conjunto de textos y dividiéndolo entre el número total DF de documentos que contienen la palabra de interés:

     $$idf = \frac{D}{DF}$$

   Con esto finalmente se puede calcular:

   $$\text{TF-IDF}=tf * idf$$

Conglomerar textos ha sido estudiado por décadas, y está lejos de ser una tarea trivial, ya que involucra varios retos como: escoger la forma apropiada de tokenizar el texto, escoger la medida de similitud o distancia adecuada, y escoger el algoritmo de conglomeración que mejor capture la estructura en los datos [23].

Capítulo 4

Agrupamiento de las reflexiones de los profesores en el proyecto *El Aula del Futuro*

En el capítulo anterior se repasaron algunos de los algoritmos de conglomeración más famosos para dar una noción general de cómo funcionan. A pesar de que todos los algoritmos de conglomeración sirven para segmentar datos, no todos lo hacen de la misma manera y es por eso que por lo general cada algoritmo de conglomeración entrega resultados distintos.

Se repasaron varios algoritmos propuestos por Kauffman y Rousseew [10] que fueron implementados por ellos en Fortran en 1990. Como ellos mismos señalan, estos algoritmos tenían limitantes en la cantidad de datos que podían ser procesados, debido a la capacidad computacional de ese entonces.

Actualmente, dado que las implementaciones de Kauffman y Rousseew no son muy eficientes cuando se enfrentan a conjuntos de datos muy grandes y dado que las computadoras tienen mucha más capacidad de procesamiento, las implementaciones actuales siguen conservando la esencia de lo propuesto por sus creadores pero con ligeras modificaciones que los hacen más eficientes.

En vez de intentar construir en código estas implementaciones originales (ya que no es el objetivo de este trabajo), se utilizarán las implementaciones actuales que

existen en Python de estos algoritmos para aplicarlas a datos reales en este capítulo.

A pesar de que en el capítulo anterior se repasaron cinco algoritmos en total, se decidió no aplicar PAM y DIANA a los datos para evitar redundancia y sintetizar en este trabajo solamente resultados interesantes; ya que debido a la poca cantidad de datos, PAM y DIANA entregan resultados sumamente similares a los de sus semejantes (*k-medias* y AGNES respectivamente).

A partir de lo anterior, en este capítulo se aplicarán algoritmos de conglomeración a fragmentos de textos obtenidos en el proyecto *El Aula del Futuro*.

4.1. Introducción al proyecto Aula del Futuro

El Aula del Futuro es un proyecto por parte del Instituto de Ciencias Aplicadas y Tecnología de la Universidad Nacional Autónoma de México para apoyar a la innovación y mejoramiento de la enseñanza.

La introducción de Tecnologías de la Información y la Comunicación (TIC) en las aulas no es suficiente para promover cambios significativos en el aprendizaje, así como un dominio adecuado de las TIC por parte de los profesores no implica la modificación de su práctica docente.

Por ello, el principal objetivo de El Aula del Futuro es coadyuvar a lograr un cambio profundo en las estrategias didácticas, en las que el uso de las TIC trascienda la simple sustitución (ej. video en vez de cátedra) y pase a esquemas de transformación educativa e innovación.

La metodología incluye la instalación de las tecnologías desarrolladas en el Aula, un programa de acompañamiento y la evaluación y sistematización colegiada de las experiencias obtenidas.

Ello incluye preparar al docente para un cambio radical en su función, la apropiación de una visión integral y una forma diferente de abordar su disciplina, acentuar el énfasis en la cultura de lo digital, capacitar para la producción de materiales en línea con calidad y aplicar el nuevo paradigma de aprendizaje en la elaboración del contenido pedagógico de los programas.

La metodología propuesta en El Aula del Futuro consta de cinco etapas:

1. Adaptación del espacio educativo. A partir del análisis del tipo de dinámicas educativas que se desea llevar a cabo en el espacio (seminario, taller, laboratorio, desarrollo de proyectos, lectura de curso, combinaciones de las anteriores), los objetivos educativos de los profesores, y las características del espacio, se generan las propuestas que faciliten dichos objetivos. Éstas incluyen, de manera razonada, las tecnologías que se consideran necesarias, buscando reutilizar todos los elementos con que ya cuenta la institución.

2. Instalación y calibración de tecnologías colaborativas adecuadas a los objetivos del proyecto. Una vez asignado y adquirido el equipo de cómputo requerido, se supervisa su instalación, se instala el software necesario y se calibra, dejando el espacio listo para su uso.

3. Trabajo con profesores. A través de convocatorias abiertas e invitaciones personalizadas, se formarán grupos de profesores para el rediseño de sus propuestas educativas. Este rediseño no tiene como objetivo principal introducir TIC, sino generar situaciones en las que sus estudiantes asuman una participación activa y crítica dentro del curso. Para ello, se ofrece un diplomado de 120 horas, que permite a los participantes revisar estrategias de aprendizaje activo apoyadas con TIC, para luego pasar al rediseño de sus clases, aplicar la nueva propuesta con sus estudiantes y evaluarla.

4. Ejecución y evaluación de secuencias didácticas modificadas. Una parte fundamental de la propuesta radica en que todas las secuencias diseñadas por los profesores, deben ser ejecutadas y evaluadas con sus alumnos. Este ejercicio es fundamental para hacer una medición objetiva de los logros alcanzados, así como de las áreas que deben seguirse mejorando. Asimismo, introduce al profesor en una dinámica de práctica reflexiva, en la que las propuestas educativas siempre son objeto de observación y mejora.

5. Socialización y sistematización de resultados. La última etapa está dirigida a

fortalecer el sentido de comunidad de aprendizaje, y busca acercar a los profesores al trabajo colaborativo y el aprendizaje entre pares, a partir de compartir sus experiencias, sus éxitos y sus dificultades. En ese sentido, el espacio renovado puede fungir como un punto de anclaje físico, donde los profesores sepan que pueden encontrarse con otros colegas que comparten estas inquietudes.

4.2. Planteamiento del problema

Dentro de las actividades del proyecto Aula del Futuro, se compilarán y editarán, en versión electrónica, todas las actividades que los profesores hayan trabajado durante el diplomado. Estas experiencias serán presentadas bajo el siguiente formato: Estrategia original, Estrategia rediseñada y aplicada con estudiantes, Resultados obtenidos, Reflexiones del profesor sobre la experiencia. Esta compilación reunirá todos los trabajos de los profesores participantes en el proyecto.

La última etapa de la metodología de El Aula del Futuro está dirigida a fortalecer el sentido de comunidad de aprendizaje, y busca acercar a los profesores al trabajo colaborativo y el aprendizaje entre pares, a partir de compartir sus experiencias, sus éxitos y sus dificultades.

La tarea de leer y analizar las reflexiones para **encontrar** las inquietudes que comparten los profesores se puede tornar sumamente larga y pesada. En este trabajo se propone el uso de algoritmos de conglomeración para automatizar esta tarea.

4.3. Hipótesis

La hipótesis bajo la que se desarrollará este trabajo es la siguiente:

Existen patrones en las reflexiones de los profesores sobre sus experiencias, éxitos y dificultades en el proyecto *El Aula del Futuro* que pueden ser identificados con algoritmos de conglomeración.

4.4. Metodología

Los 105 testimonios existentes derivados del proyecto *El Aula del futuro* fueron recolectados. Posteriormente se procedió a procesarlos:

1. Se transformó cada texto a minúsculas, se eliminaron signos de puntuación y números.

2. Se tokenizó cada texto utilizando la función *word_tokenize* de la biblioteca *nltk* en Python que funciona para texto en español. Se tokenizó por palabras. Se extrajeron 4,485 palabras únicas en todo el conjunto de textos.

3. Se eliminaron palabras vacías a partir de la lista de palabras vacías que la biblioteca *nltk* provee para palabras en español, pero se agregaron palabras a esta lista que son particulares para este contexto. La lista de palabras vacías que se utilizó para este trabajo se incluirá en el apéndice A al final de este trabajo. Después de la limpieza de palabras vacías, quedaron 4,303 palabras únicas en el conjunto de textos.

4. Se cortó cada palabra para recuperar su *Stem* utilizando la función *SnowballStemmer* de la biblioteca *nltk* en Python que funciona para texto en español. A partir de este proceso quedaron 2,327 *Stem* únicos en todo el conjunto de textos.

5. Se eliminaron también *Stem* que solo ocurrieron una vez en el conjunto de textos, esto para reducir dimensionalidad y evitar agrupar en espacios tan dispersos. Esta lista de *Stem* con frecuencia 1 se incluye en el apéndice B. Después de este proceso, quedaron solo 1,437 *Stem* únicas en todo el conjunto de textos.

6. Se mapeó el conjunto de textos a matriz utilizando el método TF-IDF a través de la función $TfidfVectorizer$ de la librería *sklearn* en Python, resultando una matriz de dimensión 105 x 1,437 (en la figura 4.1 se muestra un extracto de esta matriz).

	abarc	abiert	abord	abp	absolut	abstraccion	abstract	academ	academi	acced	...	visual	visualiz	vital	viv	vivencial	vocabulari	volv
0	0.0	0.0	0.025112	0.0	0.0	0.0	0.0	0.049356	0.0	0.000000	...	0.027629	0.000000	0.0	0.0	0.0	0.000000	0.0
1	0.0	0.0	0.048973	0.0	0.0	0.0	0.0	0.000000	0.0	0.040714	...	0.000000	0.000000	0.0	0.0	0.0	0.000000	0.0
2	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	...	0.000000	0.000000	0.0	0.0	0.0	0.027165	0.0
3	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	...	0.000000	0.000000	0.0	0.0	0.0	0.000000	0.0
4	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	...	0.000000	0.240531	0.0	0.0	0.0	0.000000	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
100	0.0	0.0	0.107649	0.0	0.0	0.0	0.0	0.000000	0.0	0.357978	...	0.000000	0.000000	0.0	0.0	0.0	0.000000	0.0
101	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	...	0.000000	0.000000	0.0	0.0	0.0	0.000000	0.0
102	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	...	0.000000	0.000000	0.0	0.0	0.0	0.000000	0.0
103	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.105211	0.0	0.000000	...	0.000000	0.000000	0.0	0.0	0.0	0.000000	0.0
104	0.0	0.0	0.109358	0.0	0.0	0.0	0.0	0.053734	0.0	0.000000	...	0.000000	0.000000	0.0	0.0	0.0	0.000000	0.0

Figura 4.1: Extracto de matriz obtenida con método TF-IDF.

Una vez obtenida esta matriz, se procedió a utilizar algoritmos de conglomeración.

Con *k-medias* y AGNES, para encontrar el número óptimo de grupos se utilizó tanto el método del codo como el método de la silueta promedio.

En el caso de AGNES (figura 4.2) la gráfica de lado izquierdo representa el método del codo y la de lado derecho el método de silueta promedio. En la gráfica de la izquierda, la línea en azul representa la suma de distancias cuadradas para cada k y la línea en verde representa el tiempo que tomó calcular esta suma para cada k. Por otro lado, en la gráfica de la derecha, la línea azul representa la silueta promedio para cada k, y la línea en verde representa el tiempo que tomó calcular esta silueta promedio para cada k.

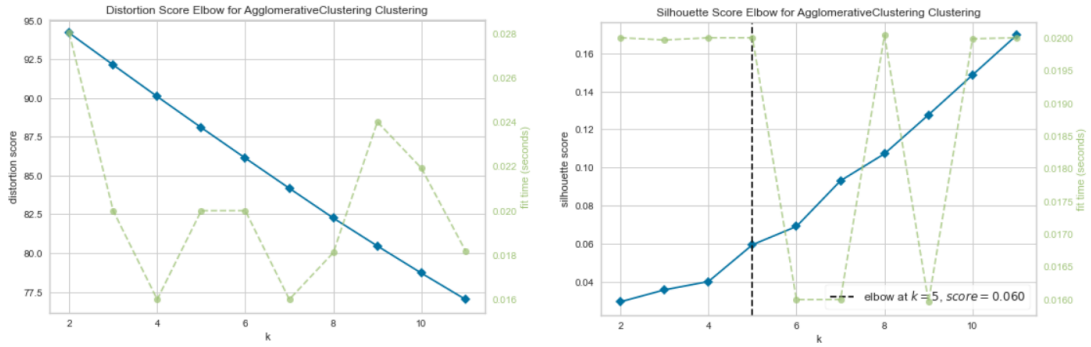

Figura 4.2: AGNES: Métodos para obtener el número óptimo de grupos.

Como se puede observar, el método del codo no encontró k óptima y el método de silueta promedio encontró que la k óptima es 5.

En el caso de *k-medias* (figura 4.3), también la gráfica de lado izquierdo representa el método del codo y la de lado derecho el método de silueta promedio. En la gráfica

de la izquierda, la línea en azul representa la suma de distancias cuadradas para cada k y la línea en verde representa el tiempo que tomó calcular esta suma para cada k. Por otro lado, en la gráfica de la derecha, la línea azul representa la silueta promedio para cada k, y la línea en verde representa el tiempo que tomó calcular esta silueta promedio para cada k.

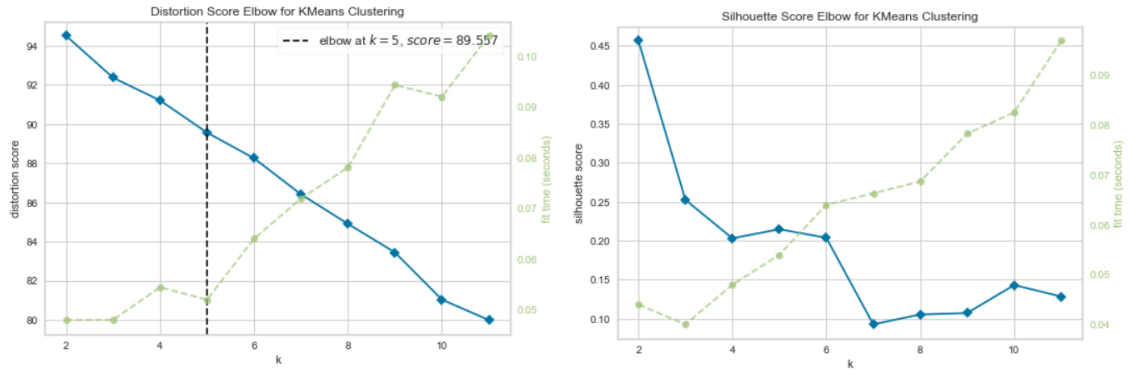

Figura 4.3: *k-medias*: Métodos para obtener el número óptimo de grupos

Como se puede observar, en *k-medias* el método del codo encontró que el número óptimo de grupos era 5, mientras que el método de silueta promedio no encontró número óptimo.

Para encontrar los mejores parámetros de DBSCAN no se utilizaron los métodos antes expuestos, debido a que no se encontró una librería que tenga esta función para DBSCAN en Python (e implementarlos en código no es el objetivo de este trabajo), así que se realizó una búsqueda entre diferentes configuraciones de parámetros, generadas aleatoriamente.

A partir de esto, se agrupó con DBSCAN utilizando los parámetros $r = .05$ y $MinPts = 2$, resultando en 10 grupos. Estos parámetros se eligieron debido a que los grupos resultantes presentaron mayor coherencia cuando se analizaron vs. grupos obtenidos con otros parámetros. Estos resultados se presentan y explican a continuación.

4.5. Resultados

Para visualizar los resultados se utilizó el método de Análisis de Componentes Principales (PCA) para poder graficar los datos en dos y tres dimensiones.

La figura 4.4 muestra los datos agrupados por *k-medias* con $k = 5$. La figura 4.5 muestra los datos agrupados por AGNES con $k = 5$ y la figura 4.6 muestra los datos agrupados por DBSCAN con $p = 0.5$ y $MinPts = 2$.

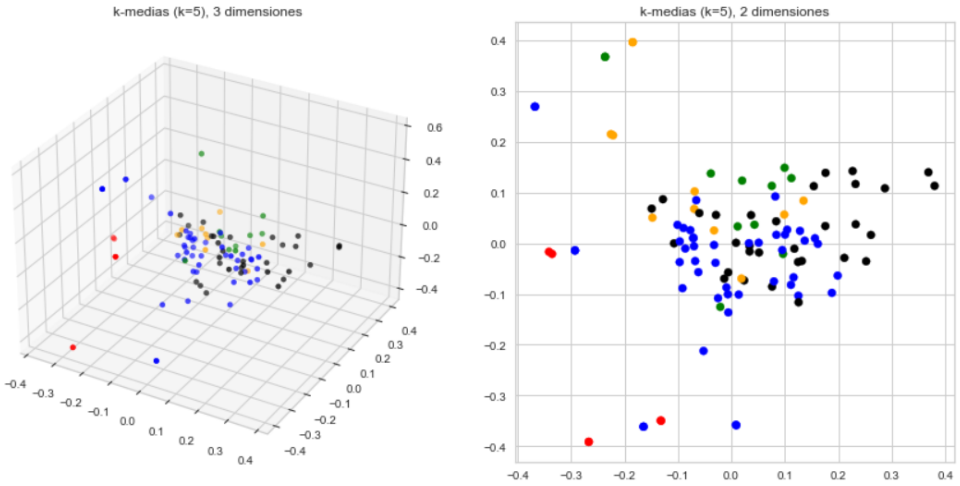

Figura 4.4: Visualización datos agrupados con *k-medias*, $k = 5$

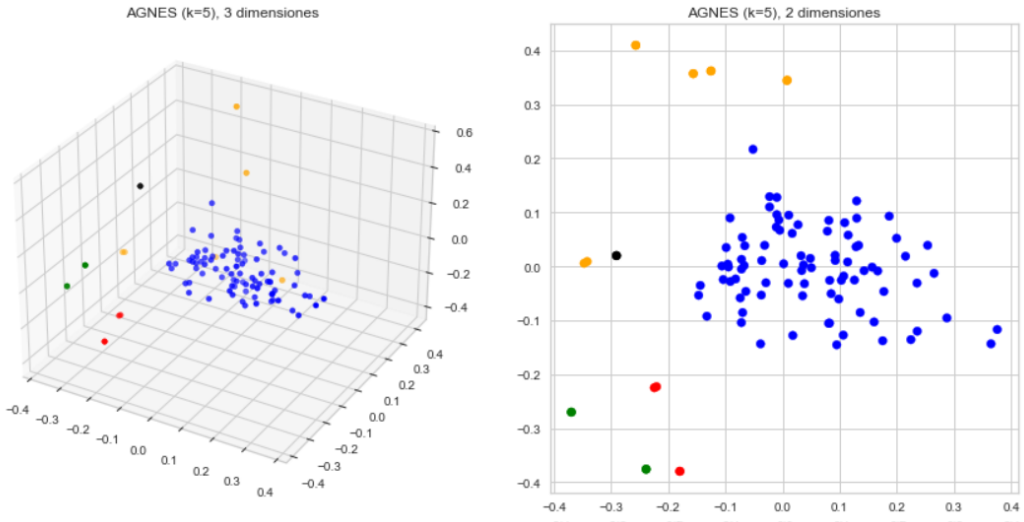

Figura 4.5: Visualización datos agrupados con AGNES, $k = 5$

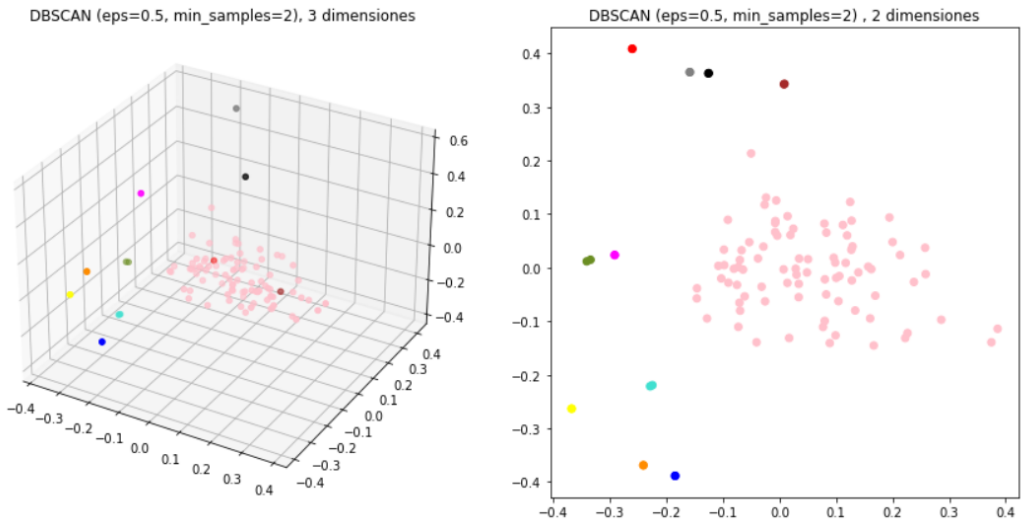

Figura 4.6: Visualización datos agrupados con DBSCAN, con parámetros: $p = 0.5$ y $MinPts = 2$

Como se puede observar, AGNES y DBSCAN entregaron resultados bastante parecidos, a pesar de que no tienen el mismo número de grupos. No así AGNES y *k-medias* que sí tienen el mismo número de grupos pero entregaron resultados muy distintos.

En las tablas 4.1, 4.2 y 4.3 se muestra la cantidad de textos de reflexiones de los profesores asignadas a cada grupo por *k-medias*, AGNES y DBSCAN respectivamente.

Grupo	1	2	3	4	5
cantidad	11	11	6	46	31

Tabla 4.1: Número de textos asignados por grupo, *k-medias*

Grupo	1	2	3	4	5
cantidad	10	4	4	85	2

Tabla 4.2: Número de textos asignados por grupo, AGNES

Grupo	-1	1	2	3	4	5	6	7	8	9	10
cantidad	85	2	2	2	2	2	2	2	2	2	2

Tabla 4.3: Número de textos asignados por grupo, DBSCAN

El grupo -1 de DBSCAN es el conjunto de datos ruidosos.

Al observar la distribución de los grupos que encontró cada algoritmo, se confirma que AGNES y DBSCAN encontraron cosas parecidas, ya que ambos algoritmos encontraron un grupo muy grande que contiene 85 textos en el caso de DBSCAN y 84 textos en el caso de AGNES. Hay que resaltar algo muy importante que es que este grupo grande que encontró DBSCAN es en realidad el conjunto de puntos ruidosos, es decir, este grupo grande no es un grupo como tal. AGNES sí trata a este grupo como un grupo de puntos que comparten un patrón.

Desde un punto de vista matemático, los algoritmos funcionan bien agrupando objetos cercanos, pero también es de especial interés en este trabajo comprobar que esta forma de agrupar funciona bien cuando se le da contexto, es decir, que existe una coherencia en el agrupamiento de las reflexiones de los profesores respecto a la temática que hablan.

Entonces, una vez que los algoritmos de conglomeración dividieron los datos en grupos hay que verificar que existe cierta coherencia en estas agrupaciones, para lo cual se utilizarán los bigramas y trigramas de *Stem* más frecuentes en cada grupo.

La tabla 4.4 muestra los bigramas más frecuentes del primer grupo encontrado por *k-medias*. La tabla 4.5 muestra los trigramas más frecuentes. Las columna *Palabra 1*, *Palabra 2* y *Palabra 3* de la tabla indican las palabras asociadas a los *Stem*.

Stem	frecuencia	Palabra 1	Palabra 2
(trabaj, social)	11	trabajan,trabajadores,trabajador,trabajos,trab...	social,sociales
(vid, social)	5	vida,vidas	social,sociales
(sustent, fuent)	4	sustentados,sustentado,sustento,sustenten,sust...	fuentes
(fuent, acerc)	4	fuentes	acercar,acercado,acerca,acercamiento
(histor, trabaj)	4	histórica,histórico,históricos	trabajan,trabajadores,trabajador,trabajos,trab...
(revolu, industrial)	4	revolución	industrial
(ident, profesional)	3	identidad	profesionales,profesional
(ruptur, cambi)	3	rupturas,ruptura	cambio,cambian,cambios,cambiar,cambiando
(deven, histor)	3	devenir	histórica,histórico,históricos
(desarroll, histor)	3	desarrollando,desarrollaban,desarrolladas,desa...	histórica,histórico,históricos

Tabla 4.4: Bigramas frecuentes, primer grupo encontrado por *k-medias*, con k=5.

Stem	frecuencia	Palabra 1	Palabra 2	Palabra 3
(histor, trabaj, social)	4	histórica,histórico,históricos	trabajan,trabajadores,trabajador,trabajos,trab...	social,sociales
(desarroll, histor, trabaj)	3	desarrollando,desarrollaban,desarrolladas,desa...	histórica,histórico,históricos	trabajan,trabajadores,trabajador,trabajos,trab...

Tabla 4.5: Trigramas frecuentes, primer grupo encontrado por *k-medias*, k=5.

A partir de estas tablas, se concluye que los temas que se tratan en este grupo son

históricos y sociales.

Las tablas 4.6 y 4.7 muestran los bigramas y trigramas de *Stem* más frecuentes en el segundo grupo encontrado por *k-medias*.

Stem	frecuencia	Palabra 1	Palabra 2
(oral, escrit)	7	oral,orales,oralmente	escrita,escrito,escritos
(verb, regular)	6	verbo,verbos	regulares,regular,regularmente
(regular, irregular)	6	regulares,regular,regularmente	irregulares
(sesion, virtual)	6	sesiones,sesión	virtual,virtuales
(italian, aleman)	5	italianas,italiana,italiano	alemán,alemanes,alemana,alemanas
(trabaj, colabor)	5	trabajan,trabajadores,trabajador,trabajos,trab...	colaborativas,colaborativa,colaboración,colabo...
(pas, simpl)	5	pasa,pasamos,pasando,paso,pasadas,pasado,pasar...	simple,simples
(maner, oral)	5	manera	oral,orales,oralmente
(pais, habl)	5	países,país	habla,hablar,habladas
(aul, virtual)	4	aula,aulas	virtual,virtuales

Tabla 4.6: Bigramas frecuentes, segundo grupo encontrado por *k-medias* con k=5.

Stem	frecuencia	Palabra 1	Palabra 2	Palabra 3
(verb, regular, irregular)	6	verbo,verbos	regulares,regular,regularmente	irregulares
(maner, oral, escrit)	5	manera	oral,orales,oralmente	escrita,escrito,escritos
(pronunci, verb, regular)	4	pronunciación	verbo,verbos	regulares,regular,regularmente
(expres, maner, oral)	3	expresarse,exprese,expresa,expresar,expresarla...	manera	oral,orales,oralmente

Tabla 4.7: Trigramas frecuentes, segundo grupo encontrado por *k-medias*, con k=5.

A partir de estas tablas, se concluye que este segundo grupo se centra en la enseñanza o aprendizaje de idiomas.

Las tablas 4.8 y 4.9 muestran los bigramas y trigramas de *Stem* más frecuentes.

Stem	frecuencia	Palabra 1	Palabra 2
(trabaj, social)	12	trabajan,trabajadores,trabajador,trabajos,trab...	social,sociales
(landing, pag)	8	landing	page
(plan, marketing)	7	planeadas,planes,plan,planos,planeado	marketing
(fuent, inform)	6	fuentes	información,informará
(pensamient, critic)	6	pensamiento	crítica,críticos,crítico,críticas
(comun, visual)	5	común,comunicaciones,comunica,comunicación,com...	visuales,visual
(social, grup)	4	social,sociales	grupos,grupo
(prim, problem)	4	primer	problemas,problema
(licenciatur, pedagog)	4	licenciaturas,licenciatura	pedagógicos,pedagógico,pedagógicas,pedagogía,p...
(falt, interes)	4	falta	interesado,interesantes,interesarse,interesar,...

Tabla 4.8: Bigramas frecuentes, tercer grupo encontrado por *k-medias*, k=5.

Stem	frecuencia	Palabra 1	Palabra 2	Palabra 3
(prototip, landing, pag)	4	prototipo	landing	page
(trabaj, social, grup)	4	trabajan,trabajadores,trabajador,trabajos,trab...	social,sociales	grupos,grupo
(licenciatur, trabaj, social)	3	licenciaturas,licenciatura	trabajan,trabajadores,trabajador,trabajos,trab...	social,sociales
(diseñ, comun, visual)	3	diseñadas,diseñen,diseñe,diseñar,diseñarán,dis...	común,comunicaciones,comunica,comunicación,com...	visuales,visual
(teor, grup, trabaj)	3	teoría,teorías	grupos,grupo	trabajan,trabajadores,trabajador,trabajos,trab...
(target, client, potencial)	3	target	cliente	potencial
(grup, trabaj, social)	3	grupos,grupo	trabajan,trabajadores,trabajador,trabajos,trab...	social,sociales

Tabla 4.9: Trigramas frecuentes, tercer grupo encontrado por *k-medias*, k=5.

A partir de estas tablas se concluye que la temática de este tercer grupo gira

alrededor del trabajo social y se relacionan las estrategias de pensamiento crítico y búsqueda de información.

Las tablas 4.10 y 4.11 muestran los bigramas y trigramas de *Stem* más frecuentes en el cuarto grupo encontrado por *k-medias*.

Stem	frecuencia	Palabra 1	Palabra 2
(aul, invert)	14	aula,aulas	invertida
(secuenci, didact)	14	secuencias,secuencia	didácticos,didácticas,didáctico,didáctica
(solucion, problem)	9	solución,solucionan,solucionar,soluciones,solu...	problemas,problema
(trabaj, colabor)	9	trabajan,trabajadores,trabajador,trabajos,trab...	colaborativas,colaborativa,colaboración,colabo...
(artes, diseñ)	8	artes	diseñadas,diseñen,diseñe,diseñar,diseñarán,dis...
(imag, corporal)	8	imagen	corporal
(resolv, problem)	8	resolverla,resolver,resolverán,resolverlo,reso...	problemas,problema
(enseñ, aprendizaj)	8	enseñar,enseñan,enseñó,enseñanza,enseña,enseño	aprendizaje,aprendizajes
(aprendizaj, tem)	8	aprendizaje,aprendizajes	tema,temas
(comun, visual)	8	común,comunicaciones,comunica,comunicación,com...	visuales,visual

Tabla 4.10: Bigramas frecuentes, cuarto grupo encontrado por *k-medias*, k=5.

Stem	frecuencia	Palabra 1	Palabra 2	Palabra 3
(diseñ, comun, visual)	6	diseñadas,diseñen,diseñe,diseñar,diseñarán,dis...	común,comunicaciones,comunica,comunicación,com...	visuales,visual
(escuel, nacional, preparatori)	6	escuela,escuelas	nacional	preparatoria
(realiz, activ, fisic)	4	realizarán,realizadas,realizan,realizados,real...	activas,activación,activo,actividades,activa,a...	física,físico,físicas,físicos
(bachillerat, escuel, nacional)	4	bachilleratos,bachillerato	escuela,escuelas	nacional
(are, cienci, experimental)	4	área,áreas	ciencias,ciencia	experimentales,experimental
(clas, frent, grup)	4	clase,clases	frente	grupos,grupo
(inici, cicl, escol)	3	inicia,inicien,inicio,iniciaba,iniciado,inicia...	ciclos,ciclo	escolar
(aul, invert, abp)	3	aula,aulas	invertida	abp
(lineament, legal, administr)	3	lineamiento,lineamientos	legales	administración,administrativos,administrativa
(produccion, text, academ)	3	producción	texto,textos	académico,académica,académicas,académicos

Tabla 4.11: Trigramas frecuentes, cuarto grupo encontrado por *k-medias*, k=5.

A partir de estas tablas, se distinguen varios enfoques de enseñanzas, como trabajo colaborativo, secuencias didácticas, actividades físicas y resolución de problemas.

En la tabla 4.12 se muestran los bigramas de *Stem* más frecuentes en el quinto grupo encontrado por *k-medias*.

Stem	frecuencia	Palabra 1	Palabra 2
(text, literari)	5	texto,textos	literario,literarios

Tabla 4.12: Bigramas frecuentes, quinto grupo encontrado por *k-medias*, con k=5.

A pesar de que este grupo fue el más poblado, solamente se encontró un bigrama frecuente. En el caso de trigramas, no se encontraron con frecuencia significativa (mayor a dos veces). Por lo que se caracterizará este grupo a partir de *textos literarios.*

Por otro lado, AGNES encontró 5 grupos. Los bigramas *Stem* más frecuentes del primer grupo se muestran en la tabla 4.13.

Stem	frecuencia	Palabra 1	Palabra 2
(tip, roc)	6	tipo,tipos	rocas
(represent, espacial)	6	representativas,representar,representación,rep...	espacial,espaciales
(tip, represent)	4	tipo,tipos	representativas,representar,representación,rep...
(represent, espaci)	4	representativas,representar,representación,rep...	espacio,espacios
(sustent, fuent)	4	sustentados,sustentado,sustento,sustenten,sust...	fuentes
(fuent, acerc)	4	fuentes	acercar,acercado,acerca,acercamiento
(ventaj, desventaj)	4	ventajas	desventajas,desventaja
(revolu, industrial)	4	revolución	industrial
(tem, fundamental)	4	tema,temas	fundamentales,fundamental
(vid, social)	4	vida,vidas	social,sociales

Tabla 4.13: Bigramas frecuentes, primer grupo encontrado por AGNES, k=5.

Para este grupo tampoco se encontraron trigramas con frecuencia importante, por lo que no se incluyeron en este trabajo. Al intentar caracterizar este grupo a partir de los bigramas más frecuentes, no se pudo inferir un tema central y coherente en este grupo, ya que los bigramas más frecuentes tienen que ver con tipos de rocas y representaciones espaciales, y en conjunto no parecen tener coherencia, por lo que se concluye que el tema es indefinido.

Para el segundo grupo encontrado por AGNES de nueva cuenta no se encontraron ni bigramas ni trigramas con frecuencia relevante, cosa comprensible ya que el grupo solamente tiene 4 integrantes, por lo que se dirá también que el tema es indefinido.

Las tablas 4.14 y 4.15 muestran los bigramas y trigramas de *Stem* más frecuentes en el tercer grupo encontrado por AGNES.

Stem	frecuencia	Palabra 1	Palabra 2
(imag, corporal)	8	imagen	corporal
(propi, imag)	4	propios,propio,propias,propia	imagen
(educ, cuid)	4	educando,educativas,educativo,educación,educat...	cuidarse,cuidaran,cuidado,cuidados
(activ, fisic)	4	activas,activación,activo,actividades,activa,a...	física,físico,físicas,físicos
(realiz, activ)	4	realizarán,realizadas,realizan,realizados,real...	activas,activación,activo,actividades,activa,a...
(propi, cuerp)	4	propios,propio,propias,propia	cuerpo
(deb, sab)	3	debería,debieron,debida,deba,deberían,deban,de...	sabían,sabe,sábados,saben,saber

Tabla 4.14: Bigramas frecuentes, tercer grupo encontrado por AGNES, k=5.

Stem	frecuencia	Palabra 1	Palabra 2	Palabra 3
(realiz, activ, fisic)	4	realizarán,realizadas,realizan,realizados,real...	activas,activación,activo,actividades,activa,a...	física,físico,físicas,físicos

Tabla 4.15: Trigramas frecuentes, tercer grupo encontrado por AGNES, k=5.

A partir de esto, se concluye que en el tercer grupo encontrado por AGNES se habla sobre imagen corporal y realización de actividades físicas.

Las tablas 4.16 y 4.17 incluyen los bigramas y trigramas de *Stem* más frecuentes del cuarto grupo encontrado por AGNES.

Stem	frecuencia	Palabra 1	Palabra 2
(trabaj, social)	23	trabajan,trabajadores,trabajador,trabajos,trab...	social,sociales
(aul, invert)	17	aula,aulas	invertida
(trabaj, colabor)	15	trabajan,trabajadores,trabajador,trabajos,trab...	colaborativas,colaborativa,colaboración,colabo...
(secuenci, didact)	13	secuencias,secuencia	didácticos,didácticas,didáctico,didáctica
(comun, visual)	13	común,comunicaciones,comunica,comunicación,com...	visuales,visual
(enseñ, aprendizaj)	11	enseñar,enseñan,enseñó,enseñanza,enseña,enseño	aprendizaje,aprendizajes
(aprendizaj, bas)	10	aprendizaje,aprendizajes	basados,basado,base,basada,bases
(solucion, problem)	9	solución,solucionan,solucionar,soluciones,solu...	problemas,problema
(pensamient, critic)	9	pensamiento	crítica,críticos,crítico,críticas
(diseñ, comun)	9	diseñadas,diseñen,diseñe,diseñar,diseñarán,dis...	común,comunicaciones,comunica,comunicación,com...

Tabla 4.16: Bigramas frecuentes, cuarto grupo encontrado por AGNES, k=5.

Stem	frecuencia	Palabra 1	Palabra 2	Palabra 3
(diseñ, comun, visual)	9	diseñadas,diseñen,diseñe,diseñar,diseñarán,dis...	común,comunicaciones,comunica,comunicación,com...	visuales,visual
(aprendizaj, bas, proyect)	6	aprendizaje,aprendizajes	basados,basado,base,basada,bases	proyecto,proyectos
(verb, regular, irregular)	6	verbo,verbos	regulares,regular,regularmente	irregulares
(maner, oral, escrit)	5	manera	oral,orales,oralmente	escrita,escrito,escritos
(desarroll, pensamient, critic)	4	desarrollando,desarrollaban,desarrolladas,desa...	pensamiento	crítica,críticos,crítico,críticas
(trabaj, social, grup)	4	trabajan,trabajadores,trabajador,trabajos,trab...	social,sociales	grupos,grupo
(aprendizaj, bas, problem)	4	aprendizaje,aprendizajes	basados,basado,base,basada,bases	problemas,problema
(histor, trabaj, social)	4	histórica,histórico,históricos	trabajan,trabajadores,trabajador,trabajos,trab...	social,sociales
(proces, enseñ, aprendizaj)	4	procesamiento,proceso,procesar,procesarlos,pro...	enseñar,enseñan,enseñó,enseñanza,enseña,enseño	aprendizaje,aprendizajes
(pronunci, verb, regular)	4	pronunciación	verbo,verbos	regulares,regular,regularmente

Tabla 4.17: Trigramas frecuentes, cuarto grupo encontrado por AGNES, k=5.

En este grupo en particular es en donde mejor se puede apreciar la temática del mismo, siendo esta temática los distintos enfoques de enseñanza: trabajo colaborativo, solución de problemas, secuencias didácticas, comunicación visual, aprendizaje basado en problemas, aprendizaje basado en proyectos, entre otros.

La tabla 4.18 muestra los bigramas de *Stem* más frecuentes que se encontraron en el quinto grupo de AGNES.

Stem	frecuencia	Palabra 1	Palabra 2
(plac, tecton)	4	placas,placer	tectónicas

Tabla 4.18: Bigramas frecuentes, quinto grupo encontrado por AGNES, k=5.

En este grupo no se encontraron trigramas con frecuencia relevante, por lo que no se incluyeron, así que se concluye que la temática de este quinto grupo encontrado por AGNES es placas tectónicas.

Para el caso de DBSCAN, debido a que los grupos que encontró contienen todos exactamente dos integrantes, una frecuencia de 2 en sus bigramas y trigramas de *Stem* es bastante relevante.

En la tabla 4.19 y 4.20 se muestran los bigramas y trigramas de *Stem* más frecuentes en el primer grupo encontrado por DBSCAN.

Stem	frecuencia	Palabra 1	Palabra 2
(realiz, lectur)	2	realizarán,realizadas,realizan,realizados,real...	lecturas,lectura
(lectur, minim)	2	lecturas,lectura	mínimo,mínimas,mínima
(minim, entreg)	2	mínimo,mínimas,mínima	entregan,entregar,entrega,entregaban
(entreg, material)	2	entregan,entregar,entrega,entregaban	materiales,material
(material, basic)	2	materiales,material	basicos,básicos,básicas,básica,básico
(basic, plataform)	2	basicos,básicos,básicas,básica,básico	plataforma
(plataform, suay)	2	plataforma	suay,suayed
(suay, surg)	2	suay,suayed	surge,surgen
(surg, divers)	2	surge,surgen	diversa,diversas,diversos,diversidad
(divers, problemat)	2	diversa,diversas,diversos,diversidad	problemático,problemáticas,problemática
(problemat, atras)	2	problemático,problemáticas,problemática	atraso
(atras, tar)	2	atraso	tareas
(tar, desercion)	2	tareas	deserción

Tabla 4.19: Bigramas frecuentes, primer grupo encontrado por DBSCAN.

Stem	frecuencia	Palabra 1	Palabra 2	Palabra 3
(realiz, lectur, minim)	2	realizarán,realizadas,realizan,realizados,real...	lecturas,lectura	mínimo,mínimas,mínima
(lectur, minim, entreg)	2	lecturas,lectura	mínimo,mínimas,mínima	entregan,entregar,entrega,entregaban
(minim, entreg, material)	2	mínimo,mínimas,mínima	entregan,entregar,entrega,entregaban	materiales,material
(entreg, material, basic)	2	entregan,entregar,entrega,entregaban	materiales,material	basicos,básicos,básicas,básica,básico
(material, basic, plataform)	2	materiales,material	basicos,básicos,básicas,básica,básico	plataforma
(basic, plataform, suay)	2	basicos,básicos,básicas,básica,básico	plataforma	suay,suayed
(plataform, suay, surg)	2	plataforma	suay,suayed	surge,surgen
(suay, surg, divers)	2	suay,suayed	surge,surgen	diversa,diversas,diversos,diversidad
(surg, divers, problemat)	2	surge,surgen	diversa,diversas,diversos,diversidad	problemático,problemáticas,problemática
(divers, problemat, atras)	2	diversa,diversas,diversos,diversidad	problemático,problemáticas,problemática	atraso
(problemat, atras, tar)	2	problemático,problemáticas,problemática	atraso	tareas
(atras, tar, desercion)	2	atraso	tareas	deserción

Tabla 4.20: Trigramas frecuentes, primer grupo encontrado por DBSCAN

A partir de estas tablas se concluye que la temática de este grupo es la problemática que existen cuando los alumnos trabajan en la plataforma: atraso de tareas y deserción. Y las actividades que se publican en la plataforma: lecturas y otros materiales.

En la tabla 4.21 y 4.22 se muestran los bigramas y trigramas de *Stem* más frecuentes en el segundo grupo encontrado por DBSCAN.

Stem	frecuencia	Palabra 1	Palabra 2
(imag, corporal)	8	imagen	corporal
(propi, imag)	4	propios,propio,propias,propia	imagen
(activ, fisic)	4	activas,activación,activo,actividades,activa,a...	física,físico,físicas,físicos
(educ, cuid)	4	educando,educativas,educativo,educación,educat...	cuidarse,cuidaran,cuidado,cuidados
(propi, cuerp)	4	propios,propio,propias,propia	cuerpo
(realiz, activ)	4	realizarán,realizadas,realizan,realizados,real...	activas,activación,activo,actividades,activa,a...
(deb, sab)	3	debería,debieron,debida,deba,deberían,deban,de...	sabían,sabe,sábados,saben,saber

Tabla 4.21: Bigramas frecuentes, segundo grupo encontrado por DBSCAN, k=5.

Stem	frecuencia	Palabra 1	Palabra 2	Palabra 3
(realiz, activ, fisic)	4	realizarán,realizadas,realizan,realizados,real...	activas,activación,activo,actividades,activa,a...	física,físico,físicas,físicos

Tabla 4.22: Trigramas frecuentes, segundo grupo encontrado por AGNES, k=5.

A partir de esto, se infiere que el tema de este grupo tiene que ver con educación física e imagen corporal.

Para el tercer grupo encontrado por DBSCAN, las tablas 4.23 y 4.24 muestran los bigramas y trigramas de *Stem* más frecuentes.

Stem	frecuencia	Palabra 1	Palabra 2
(cambi, climat)	2	cambio,cambian,cambios,cambiar,cambiando	climático
(inform, disposicion)	2	información,informará	disposición,disposiciones
(pensamient, critic)	2	pensamiento	crítica,críticos,crítico,críticas
(critic, tom)	2	crítica,críticos,crítico,críticas	tomen,tome,tomará,tomado,toman,tomar,tomando,toma
(tom, postur)	2	tomen,tome,tomará,tomado,toman,tomar,tomando,toma	postura,posturas
(postur, bas)	2	postura,posturas	basados,basado,base,basada,bases
(bas, inform)	2	basados,basado,base,basada,bases	información,informará
(climat, problem)	2	climático	problemas,problema
(reconoc, inform)	2	reconocen,reconocer,reconoce	información,informará
(inform, encuentr)	2	información,informará	encuentros,encuentren,encuentra,encuentro,encu...
(encuentr, util)	2	encuentros,encuentren,encuentra,encuentro,encu...	utilidad,útil,útiles
(util, confiabl)	2	utilidad,útil,útiles	confiable,confiables
(confiabl, buen)	2	confiable,confiables	buenas,buena,buen
(buen, calid)	2	buenas,buena,buen	calidad
(seleccion, inform)	2	seleccionado,seleccionados,selección,seleccion...	información,informará
(oportun, desarroll)	2	oportunidades,oportunas,oportunidad	desarrollando,desarrollaban,desarrolladas,desa...
(desarroll, competent)	2	desarrollando,desarrollaban,desarrolladas,desa...	competencias,competencia
(competent, busqued)	2	competencias,competencia	búsquedas,búsqueda,busqueda

Tabla 4.23: Bigramas frecuentes, tercer grupo encontrado por DBSCAN, k=5.

Stem	frecuencia	Palabra 1	Palabra 2	Palabra 3
(cambi, climat, problem)	2	cambio,cambian,cambios,cambiar,cambiando	climático	problemas,problema
(climat, problem, complej)	2	climático	problemas,problema	complejos,complejo,compleja,complejidad
(pensamient, critic, tom)	2	pensamiento	crítica,críticos,crítico,críticas	tomen,tome,tomará,tomado,toman,tomar,tomando,toma
(critic, tom, postur)	2	crítica,críticos,crítico,críticas	tomen,tome,tomará,tomado,toman,tomar,tomando,toma	postura,posturas
(tom, postur, bas)	2	tomen,tome,tomará,tomado,toman,tomar,tomando,toma	postura,posturas	basados,basado,base,basada,bases
(postur, bas, inform)	2	postura,posturas	basados,basado,base,basada,bases	información,informará
(inform, encuentr, util)	2	información,informará	encuentros,encuentren,encuentra,encuentro,encu...	utilidad,útil,útiles
(encuentr, util, confiabl)	2	encuentros,encuentren,encuentra,encuentro,encu...	utilidad,útil,útiles	confiable,confiables
(util, confiabl, buen)	2	utilidad,útil,útiles	confiable,confiables	buenas,buena,buen
(confiabl, buen, calid)	2	confiable,confiables	buenas,buena,buen	calidad
(acced, pensamient, critic)	2	acceder,accediendo	pensamiento	crítica,críticos,crítico,críticas
(busqued, seleccion, inform)	2	búsquedas,búsqueda,busqueda	seleccionado,seleccionados,selección,seleccion...	información,informará
(problem, complej, diferent)	2	problemas,problema	complejos,complejo,compleja,complejidad	diferente,diferentes,diferencias,diferencia
(complej, diferent, punt)	2	complejos,complejo,compleja,complejidad	diferente,diferentes,diferencias,diferencia	puntos,punto
(diferent, punt, vist)	2	diferente,diferentes,diferencias,diferencia	puntos,punto	vistas,vistos,vista,visto
(investig, eje, problemat)	2	investigó,investigativas,investigaciones,inves...	eje	problemático,problemáticas,problemática
(problemat, necesit, ser)	2	problemático,problemáticas,problemática	necesita,necesitarán	ser
(necesit, ser, reflex)	2	necesita,necesitarán	ser	reflexiva,reflexivo
(ser, reflex, represent)	2	ser	reflexiva,reflexivo	representativas,representar,representación,rep...
(reflex, represent, oportun)	2	reflexiva,reflexivo	representativas,representar,representación,rep...	oportunidades,oportunas,oportunidad
(represent, oportun, desarroll)	2	representativas,representar,representación,rep...	oportunidades,oportunas,oportunidad	desarrollando,desarrollaban,desarrolladas,desa...
(oportun, desarroll, competent)	2	oportunidades,oportunas,oportunidad	desarrollando,desarrollaban,desarrolladas,desa...	competencias,competencia
(desarroll, competent, busqued)	2	desarrollando,desarrollaban,desarrolladas,desa...	competencias,competencia	búsquedas,búsqueda,busqueda

Tabla 4.24: Trigramas frecuentes, tercer grupo encontrado por DBSCAN.

En este tercer grupo se logra intuir que la temática es sobre el cambio climático y el desarrollo de un pensamiento crítico.

Los resultados del cuarto grupo se muestran en las tablas 4.25 y 4.26.

Por lo que se concluye que el cuarto grupo habla de las problemáticas que enfrentan los estudiantes de biología.

Stem	frecuencia	Palabra 1	Palabra 2
(tip, roc)	6	tipo,tipos	rocas

Tabla 4.25: Bigramas frecuentes, cuarto grupo encontrado por DBSCAN.

Stem	frecuencia	Palabra 1	Palabra 2	Palabra 3
(carrer, biolog, confund)	2	carrera,carreras	biólogos,biológicos,biología,biológicas	confundida,confunden,confundidos,confundirse,c...
(abord, problemat, promov)	2	aborden,abordar,aborda,aborde,abordan,abordado	problemático,problemáticas,problemática	promoviendo,promover
(terminolog, roc, igne)	2	terminologías	rocas	ígneas
(aprend, divers, terminolog)	2	aprende,aprender,aprendidos,aprendido,aprendie...	diversa,diversas,diversos,diversidad	terminologías
(biolog, confund, caracterist)	2	biólogos,biológicos,biología,biológicas	confundida,confunden,confundidos,confundirse,c...	característica,características
(plante, abord, problemat)	2	planteo,planteó,plantea	aborden,abordar,aborda,aborde,abordan,abordado	problemático,problemáticas,problemática
(problemat, promov, realic)	2	problemático,problemáticas,problemática	promoviendo,promover	realice,realicé,realicen
(realic, activ, busqued)	2	realice,realicé,realicen	activas,activación,activo,actividades,activa,a...	búsquedas,búsqueda,busqueda
(identif, similitud, diferent)	2	identificación,identifica	similitudes	diferente,diferentes,diferencias,diferencia
(similitud, diferent, apoy)	2	similitudes	diferente,diferentes,diferencias,diferencia	apoyados,apoyan,apoyarlo,apoyo,apoyar,apoyen,a...
(diferent, apoy, visual)	2	diferente,diferentes,diferencias,diferencia	apoyados,apoyan,apoyarlo,apoyo,apoyar,apoyen,a...	visuales,visual
(comprend, aprend, divers)	2	comprender,comprendan,comprenden,comprendido,c...	aprende,aprender,aprendidos,aprendido,aprendie...	diversa,diversas,diversos,diversidad
(distint, diferent, clasif)	2	distintas,distintos,distintivas	diferente,diferentes,diferencias,diferencia	clasificación,clasificaciones
(confund, caracterist, descript)	2	confundida,confunden,confundidos,confundirse,c...	característica,características	descriptivas

Tabla 4.26: Trigramas frecuentes, cuarto grupo encontrado por DBSCAN.

Para el quinto grupo encontrado por DBSCAN, las tablas 4.27 y 4.28 muestran los resultados.

Stem	frecuencia	Palabra 1	Palabra 2
(diferent, context)	2	diferente,diferentes,diferencias,diferencia	contextos,contexto
(expres, mism)	2	expresarse,exprese,expresa,expresar,expresarla...	misma,mismos,mismas,mismo
(mism, maner)	2	misma,mismos,mismas,mismo	manera
(maner, context)	2	manera	contextos,contexto
(context, result)	2	contextos,contexto	resultaba,resultados,resultado,resulta,resulten
(result, convenient)	2	resultaba,resultados,resultado,resulta,resulten	conveniente
(convenient, mostr)	2	conveniente	mostrando,mostraban,mostrarán,mostrados,mostra...
(mostr, ejempl)	2	mostrando,mostraban,mostrarán,mostrados,mostra...	ejemplo,ejemplos
(ejempl, divers)	2	ejemplo,ejemplos	diversa,diversas,diversos,diversidad
(logr, identific)	2	logra,lograran,lograrlo,logrará,logros,lograr,...	identificamos,identificar,identificado,identif...
(identific, moment)	2	identificamos,identificar,identificado,identif...	momentos,momento
(moment, adecu)	2	momentos,momento	adecuados,adecuadamente,adecuaciones,adecuado,...
(adecu, usar)	2	adecuados,adecuadamente,adecuaciones,adecuado,...	usar,usará
(cuid, lectur)	2	cuidarse,cuidaran,cuidado,cuidados	lecturas,lectura
(lectur, elabor)	2	lecturas,lectura	elaboración,elaborará,elaborar,elaborando,elab...
(elabor, text)	2	elaboración,elaborará,elaborar,elaborando,elab...	texto,textos
(text, academ)	2	texto,textos	académico,académica,académicas,académicos
(academ, uso)	2	académico,académica,académicas,académicos	uso
(uso, registr)	2	uso	registro,registros
(registr, adecu)	2	registro,registros	adecuados,adecuadamente,adecuaciones,adecuado,...
(registr, formal)	2	registro,registros	formales,formal
(formal, informal)	2	formales,formal	informal

Tabla 4.27: Bigramas frecuentes, quinto grupo encontrado por DBSCAN.

Stem	frecuencia	Palabra 1	Palabra 2	Palabra 3
(expres, mism, maner)	2	expresarse,exprese,expresa,expresar,expresarla...	misma,mismos,mismas,mismo	manera
(mism, maner, context)	2	misma,mismos,mismas,mismo	manera	contextos,contexto
(mostr, ejempl, divers)	2	mostrando,mostraban,mostrarán,mostrados,mostra...	ejemplo,ejemplos	diversa,diversas,diversos,diversidad
(uso, form, expresion)	2	uso	forma,formando,formar,forme,forman,formas	expresiones,expresión
(form, expresion, ambos)	2	forma,formando,formar,forme,forman,formas	expresiones,expresión	ambos
(logr, identific, moment)	2	logra,lograran,lograrlo,logrará,logros,lograr,...	identificamos,identificar,identificado,identif...	momentos,momento
(identific, moment, adecu)	2	identificamos,identificar,identificado,identif...	momentos,momento	adecuados,adecuadamente,adecuaciones,adecuado,...
(moment, adecu, usar)	2	momentos,momento	adecuados,adecuadamente,adecuaciones,adecuado,...	usar,usará
(tiend, expres, mism)	2	tiende,tienden	expresarse,exprese,expresa,expresar,expresarla...	misma,mismos,mismas,mismo
(ser, cuid, lectur)	2	ser	cuidarse,cuidaran,cuidado,cuidados	lecturas,lectura
(cuid, lectur, elabor)	2	cuidarse,cuidaran,cuidado,cuidados	lecturas,lectura	elaboración,elaborará,elaborar,elaborando,elab...
(elabor, text, academ)	2	elaboración,elaborará,elaborar,elaborando,elab...	texto,textos	académico,académica,académicas,académicos
(nocion, diferent, registr)	2	noción	diferente,diferentes,diferencias,diferencia	registro,registros
(diferent, registr, formal)	2	diferente,diferentes,diferencias,diferencia	registro,registros	formales,formal
(registr, formal, informal)	2	registro,registros	formales,formal	informal

Tabla 4.28: Trigramas frecuentes, quinto grupo encontrado por DBSCAN.

En este quinto grupo se infiere que la temática se centra en la importancia de saber diferenciar entre el lenguaje formal e informal y su uso adecuado.

El resultado de analizar al sexto grupo de DBSCAN se muestra en las tablas 4.29 y 4.30.

Stem	frecuencia	Palabra 1	Palabra 2
(fuent, acerc)	4	fuentes	acercar,acercado,acerca,acercamiento
(revolu, industrial)	4	revolución	industrial
(sustent, fuent)	4	sustentados,sustentado,sustento,sustenten,sust...	fuentes
(vid, social)	4	vida,vidas	social,sociales

Tabla 4.29: Bigramas frecuentes, sexto grupo encontrado por DBSCAN.

Stem	frecuencia	Palabra 1	Palabra 2	Palabra 3
(estudi, tem, central)	2	estudio,estudiado,estudian,estudios,estudiará,...	tema,temas	centrales,central
(industrializ, distint, etap)	2	industrialización	distintas,distintos,distintivas	etapa,etapas
(plan, vid, social)	2	planeadas,planes,plan,planos,planeado	vida,vidas	social,sociales
(vid, social, estudi)	2	vida,vidas	social,sociales	estudio,estudiado,estudian,estudios,estudiará,...
(estudi, repercusion, industrializ)	2	estudio,estudiado,estudian,estudios,estudiará,...	repercusiones	industrialización
(tercer, revolu, industrial)	2	tercer,tercera	revolución	industrial
(segund, tercer, revolu)	2	segunda,segundo	tercer,tercera	revolución
(nuev, proyect, industrializ)	2	nuevo,nuevos,nuevas,nueva	proyecto,proyectos	industrialización
(red, comun, transport)	2	redes	común,comunicaciones,comunica,comunicación,com...	transporte
(desarroll, vid, social)	2	desarrollando,desarrollaban,desarrolladas,desa...	vida,vidas	social,sociales
(vid, social, apropi)	2	vida,vidas	social,sociales	apropiarse,apropia,apropiación,apropiada
(permit, asum, postur)	2	permita,permitirá,permiten,permitan,permite,pe...	asumir,asume,asumiendo	postura,posturas
(asum, postur, critic)	2	asumir,asume,asumiendo	postura,posturas	crítica,críticos,crítico,críticas
(punt, vist, sustent)	2	puntos,punto	vistas,vistos,vista,visto	sustentados,sustentado,sustento,sustenten,sust...
(especif, aprendizaj, proces)	2	específicamente,específicas,específica,especif...	aprendizaje,aprendizajes	procesamiento,proceso,procesar,procesarlos,pro...
(siempr, prov, fuent)	2	siempre	provean,proveerse,proveer	fuentes
(prov, fuent, trabaj)	2	provean,proveerse,proveer	fuentes	trabajan,trabajadores,trabajador,trabajos,trab...
(present, sustent, fuent)	2	presentaciones,presentar,presentan,presentes,p...	sustentados,sustentado,sustento,sustenten,sust...	fuentes
(sustent, fuent, estudi)	2	sustentados,sustentado,sustento,sustenten,sust...	fuentes	estudio,estudiado,estudian,estudios,estudiará,...
(fuent, estudi, deven)	2	fuentes	estudio,estudiado,estudian,estudios,estudiará,...	devenir
(hech, present, sustent)	2	hecho,hechos	presentaciones,presentar,presentan,presentes,p...	sustentados,sustentado,sustento,sustenten,sust...
(posicion, critic, acerc)	2	posiciones,posición,posicionaba	crítica,críticos,crítico,críticas	acercar,acercado,acerca,acercamiento

Tabla 4.30: Trigramas frecuentes, sexto grupo encontrado por DBSCAN.

A partir de estas tablas se concluye que el tema central del grupo no se distingue fácilmente, ya que por un lado de habla sobre revolución industrial y por otro lado de habla sobre la capacidad de buscar fuentes confiables de información, por lo que concluiremos que la temática de este grupo es indefinida.

Los resultados del séptimo grupo encontrado por DBSCAN se muestran en las tablas 4.31 y 4.32.

Stem	frecuencia	Palabra 1	Palabra 2
(represent, espacial)	6	representativas,representar,representación,rep...	espacial,espaciales
(represent, espaci)	4	representativas,representar,representación,rep...	espacio,espacios
(ventaj, desventaj)	4	ventajas	desventajas,desventaja
(tip, represent)	4	tipo,tipos	representativas,representar,representación,rep...
(tem, fundamental)	4	tema,temas	fundamentales,fundamental

Tabla 4.31: Bigramas frecuentes, séptimo grupo encontrado por DBSCAN.

En este grupo se distingue que la temática tiene que ver con los problemas que se

Stem	frecuencia	Palabra 1	Palabra 2	Palabra 3
(divers, ventaj, desventaj)	2	diversa,diversas,diversos,diversidad	ventajas	desventajas,desventaja
(ofrec, divers, ventaj)	2	ofrecen,ofrecer,ofrece	diversa,diversas,diversos,diversidad	ventajas
(represent, espacial, ofrec)	2	representativas,representar,representación,rep...	espacial,espaciales	ofrecen,ofrecer,ofrece
(utiliz, ejempl, represent)	2	utilizará,utilizadas,utilizan,utilizarlos,util...	ejemplo,ejemplos	representativas,representar,representación,rep...
(acontec, present, espaci)	2	acontecimientos,acontecidos	presentaciones,presentar,presentan,presentes,p...	espacio,espacios
(tip, represent, espacial)	2	tipo,tipos	representativas,representar,representación,rep...	espacial,espaciales
(cuant, tip, represent)	2	cuántos,cuanto,cuánto	tipo,tipos	representativas,representar,representación,rep...
(identifiqu, cuant, tip)	2	identifiquen,identifique	cuántos,cuanto,cuánto	tipo,tipos
(import, identifiqu, cuant)	2	importante,importar,importancia,importantes	identifiquen,identifique	cuántos,cuanto,cuánto
(map, import, identifiqu)	2	mapas,mapa	importante,importar,importancia,importantes	identifiquen,identifique
(tod, map, import)	2	todas,toda	mapas,mapa	importante,importar,importancia,importantes
(form, represent, tierr)	2	forma,formando,formar,forme,forman,formas	representativas,representar,representación,rep...	tierra
(temat, exist, dificult)	2	temáticos,temáticas,temática	existe,existen,existido	dificulta,dificultad
(dich, temat, exist)	2	dicha,dichas,dichos,dicho	temáticos,temáticas,temática	existe,existen,existido
(abord, dich, temat)	2	aborden,abordar,aborda,aborde,abordan,abordado	dicha,dichas,dichos,dicho	temáticos,temáticas,temática
(problemat, represent, aprendizaj)	2	problemático,problemáticas,problemática	representativas,representar,representación,rep...	aprendizaje,aprendizajes
(geograf, mencion, realiz)	2	geográfico,geografía	mencionado,mencionar,mencionan,mencionados	realizarán,realizadas,realizan,realizados,real...
(identific, tem, fundamental)	2	identificamos,identificar,identificado,identif...	tema,temas	fundamentales,fundamental
(geograf, año, bachillerat)	2	geográfico,geografía	año	bachilleratos,bachillerato

Tabla 4.32: Trigramas frecuentes, séptimo grupo encontrado por DBSCAN.

enfrentan profesores de geografía al impartir su materia: que los alumnos identifiquen las ventajas y desventajas de las representaciones espaciales.

En las tablas 4.33 y 4.34 se muestran los bigramas y trigramas de *Stem* más frecuentes en el octavo grupo encontrado por DBSCAN.

Stem	frecuencia	Palabra 1	Palabra 2
(plac, tecton)	4	placas,placer	tectónicas

Tabla 4.33: Bigramas frecuentes, octavo grupo encontrado por DBSCAN.

Stem	frecuencia	Palabra 1	Palabra 2	Palabra 3
(estudi, punt, calient)	2	estudio,estudiado,estudian,estudios,estudiará,...	puntos,punto	calientes
(punt, calient, hotspots)	2	puntos,punto	calientes	hotspots
(veloc, plac, tecton)	2	velocidad	placas,placer	tectónicas
(direccion, veloc, plac)	2	dirección	velocidad	placas,placer
(desplaz, direccion, veloc)	2	desplazamiento	dirección	velocidad
(identific, desplaz, direccion)	2	identificamos,identificar,identificado,identif...	desplazamiento	dirección
(mant, import, geolog)	2	manto	importante,importar,importancia,importantes	geológico,geológica,geólogos,geológicos

Tabla 4.34: Trigramas frecuentes, octavo grupo encontrado por DBSCAN.

Se concluye que en el octavo grupo encontrado por DBSCAN se habla del estudio de movimiento y velocidad de las placas tectónicas, además de su dirección y desplazamiento.

En las tablas 4.35 y 4.36 se muestran los bigramas y trigramas de *Stem* más frecuentes en el noveno grupo encontrado por DBSCAN. Con estas tablas se concluye que la temática del décimo grupo encontrado por DBSCAN tiene que ver con el valor de una buena lectura: generación de opinión propia, razonamiento y reflexión.

En las tablas 4.37 y 4.38 se muestran los bigramas y trigramas de *Stem* más frecuentes en el décimo grupo encontrado por DBSCAN.

Stem	frecuencia	Palabra 1	Palabra 2
(verdader, leer)	2	verdadero	leer
(buen, lectur)	2	buenas,buena,buen	lecturas,lectura
(implic, buen)	2	implicaciones,implica,implicaba,implicado	buenas,buena,buen
(reflexion, implic)	2	reflexiona,reflexión,reflexionar,reflexione,re...	implicaciones,implica,implicaba,implicado
(razon, reflexion)	2	razones,razonamiento,razón	reflexiona,reflexión,reflexionar,reflexione,re...
(opinion, razon)	2	opiniones	razones,razonamiento,razón
(ide, opinion)	2	idea,ideas	opiniones
(propi, ide)	2	propios,propio,propias,propia	idea,ideas
(desarroll, propi)	2	desarrollando,desarrollaban,desarrolladas,desa...	propios,propio,propias,propia

Tabla 4.35: Bigramas frecuentes, noveno grupo encontrado por DBSCAN.

Stem	frecuencia	Palabra 1	Palabra 2	Palabra 3
(lectur, ademas, poc)	2	lecturas,lectura	además	pocos,poca,poco,pocas
(buen, lectur, ademas)	2	buenas,buena,buen	lecturas,lectura	además
(implic, buen, lectur)	2	implicaciones,implica,implicaba,implicado	buenas,buena,buen	lecturas,lectura
(reflexion, implic, buen)	2	reflexiona,reflexión,reflexionar,reflexione,re...	implicaciones,implica,implicaba,implicado	buenas,buena,buen
(razon, reflexion, implic)	2	razones,razonamiento,razón	reflexiona,reflexión,reflexionar,reflexione,re...	implicaciones,implica,implicaba,implicado
(opinion, razon, reflexion)	2	opiniones	razones,razonamiento,razón	reflexiona,reflexión,reflexionar,reflexione,re...
(ide, opinion, razon)	2	idea,ideas	opiniones	razones,razonamiento,razón
(propi, ide, opinion)	2	propios,propio,propias,propia	idea,ideas	opiniones
(desarroll, propi, ide)	2	desarrollando,desarrollaban,desarrolladas,desa...	propios,propio,propias,propia	idea,ideas
(incapacit, desarroll, propi)	2	incapacitados	desarrollando,desarrollaban,desarrolladas,desa...	propios,propio,propias,propia

Tabla 4.36: Trigramas frecuentes, noveno grupo encontrado por DBSCAN.

Stem	frecuencia	Palabra 1	Palabra 2
(criteri, semej)	2	criterios	semejanza,semejanzas
(semej, triangul)	2	semejanza,semejanzas	triángulo,triángulos
(curs, matemat)	2	cursado,cursará,cursar,cursaron,cursos,cursan,...	matemático,matemática,matemáticas,matemáticos
(resolv, problem)	2	resolverla,resolver,resolverán,resolverlo,reso...	problemas,problema
(familiariz, figur)	2	familiarizando,familiarizados,familiarizarnos	figurativa,figuración,figurativo,figura
(figur, geometr)	2	figurativa,figuración,figurativo,figura	geométrica,geometría
(matemat, dificil)	2	matemático,matemática,matemáticas,matemáticos	difícilmente,difícil,difíciles
(dificil, entend)	2	difícilmente,difícil,difíciles	entender,entendamos,entendimiento,entendido
(entend, ven)	2	entender,entendamos,entendimiento,entendido	ven,venido
(ven, util)	2	ven,venido	utilidad,útil,útiles
(teorem, tal)	2	teoremas,teorema	tal,tales
(tal, semej)	2	tal,tales	semejanza,semejanzas
(dificult, comprend)	2	dificulta,dificultad	comprender,comprendan,comprenden,comprendido,c...
(demostr, teorem)	2	demostración,demostrar,demostraciones	teoremas,teorema
(geometr, general)	1	geométrica,geometría	general,generalmente,generales

Tabla 4.37: Bigramas frecuentes, décimo grupo encontrado por DBSCAN.

Stem	frecuencia	Palabra 1	Palabra 2	Palabra 3
(criteri, semej, triangul)	2	criterios	semejanza,semejanzas	triángulo,triángulos
(familiariz, figur, geometr)	2	familiarizando,familiarizados,familiarizarnos	figurativa,figuración,figurativo,figura	geométrica,geometría
(analiz, cumpl, algun)	2	analizará,analizar,analizan,analizada,analiza	cumplirse,cumplen,cumple,cumpliendo,cumplir	algún,alguna,alguno
(dificult, analiz, cumpl)	2	dificulta,dificultad	analizará,analizar,analizan,analizada,analiza	cumplirse,cumplen,cumple,cumpliendo,cumplir
(matemat, dificil, entend)	2	matemático,matemática,matemáticas,matemáticos	difícilmente,difícil,difíciles	entender,entendamos,entendimiento,entendido
(dificil, entend, ven)	2	difícilmente,difícil,difíciles	entender,entendamos,entendimiento,entendido	ven,venido
(tem, teorem, tal)	2	tema,temas	teoremas,teorema	tal,tales
(teorem, tal, semej)	2	teoremas,teorema	tal,tales	semejanza,semejanzas
(semej, dificult, comprend)	2	semejanza,semejanzas	dificulta,dificultad	comprender,comprendan,comprenden,comprendido,c...
(difficult, comprend, demostr)	2	dificulta,dificultad	comprender,comprendan,comprenden,comprendido,c...	demostración,demostrar,demostraciones
(comprend, demostr, teorem)	2	comprender,comprendan,comprenden,comprendido,c...	demostración,demostrar,demostraciones	teoremas,teorema

Tabla 4.38: Trigramas frecuentes, décimo grupo encontrado por DBSCAN.

A partir de estas tablas se distingue que la temática central de este grupo es la dificultad que enfrentan los alumnos para comprender temas de matemáticas como semejanzas entre triángulos y el teorema de tales.

En la tabla 4.39 se muestra un resumen de las temáticas de cada grupo encontrado por cada algoritmo de conglomeración.

A partir de la tabla 4.39 se pueden notar algunas cosas. La primera es que los grupos encontrados por DBSCAN tratan temáticas mucho más específicas que los grupos encontrados por los demás algoritmos, esto se debe a que DBSCAN no agrupa puntos ruidosos, por lo que los resultados que entrega son siempre los más puros. En compensación DBSCAN encontró que gran parte de los datos eran puntos ruidosos, por lo que la gran parte de los datos no fueron agrupados.

Temáticas por grupo			
grupo	*k-medias*	AGNES	DBSCAN
1	Histórico sociales.	Indefinido.	Problemas derivados de trabajar en plataformas: atraso de tareas y deserción.
2	Enseñanza o aprendizaje de idiomas.	Indefinido.	Educación física e imagen corporal.
3	Trabajo social, estrategias de pensamiento crítico y búsqueda de información.	Imagen corporal.	Cambio climático y desarrollo de un pensamiento crítico.

4	Enfoques de enseñanza: trabajo colaborativo, secuencias didácticas, actividades físicas y resolución de problemas.	Enfoques de enseñanza: trabajo colaborativo, solución de problemas, secuencias didácticas, comunicación visual, apendizaje basado en problemas y aprendizaje basado en proyectos.	Problemáticas que enfrentan los estudiantes de biología.
5	Textos literarios.	Placas tectónicas.	Importancia de saber diferenciar entre lenguaje formal e informal, así como su uso adecuado.
6			Indeterminado.
7			Problemáticas que enfrentan los profesores de geografía: ventajas y desventajas de representaciones espaciales.
8			Estudio de placas tectónicas: movimiento, velocidad, dirección y desplazamiento.

9			El valor de la buena lectura: generación de opinión propia, razonamiento y reflexión.
10			Dificultad que enfrentan los alumnos para comprender temas de matemáticas como semejanza de triángulos y el teorema de Tales.

Tabla 4.39: Temáticas de cada grupo encontrado por *k-medias*, AGNES y DBSCAN.

En el caso de AGNES ocurrieron cosas muy interesantes. Por un lado, la mayoría de sus grupos no tienen una temática bien determinada, ya que o agruparon muy pocos datos, o los datos que agruparon no cumplen con una temática coherente como en el caso del primer grupo. Por otro lado, AGNES encontró un grupo (el cuarto) con la información que se prevía encontrar antes de iniciar este trabajo: enfoques de enseñanza.

Como se mencionó en el planteamiento del problema de este trabajo, una de las razones que motivó la elaboración de este fue facilitar tanto las problemáticas, como las formas de atacar estas problemáticas a los profesores que experimentan o experimentaron situaciones similares. En el cuarto grupo encontrado por AGNES podemos identificar las formas que utilizaron los profesores para atacar sus problemáticas: trabajo colaborativo, secuencias didácticas, comunicación visual, aprendizaje basado en problemas, entre otros. Por esta razón se concluyó que el cuarto grupo de AGNES entregó resultados muy valiosos.

Los grupos encontrados por *k-medias* fueron también bastante prometedores. Se puede observar que las temáticas de sus grupos están bien determinadas (no tanto como las de DBSCAN, pero mucho mejor que la mayoría de AGNES), y se puede observar que *k-medias* también encontró el grupo que reúne los enfoques de enseñanza.

Capítulo 5

Conclusiones

En este trabajo se inició dando una introducción a la Inteligencia Artificial, al aprendizaje, al aprendizaje automatizado y a los diferentes tipos de aprendizaje. Posteriormente se profundizó en disimilitudes, similitudes y distancias, para poder contextualizar a los algoritmos de conglomeración.

Se repasaron cinco de los algoritmos de conglomeración más utilizados. Este repaso estuvo basado en las publicaciones originales de cada autor de cada algoritmo. Se revisó también el tema de conglomeración de textos de forma teórica, pasando superficialmente por el procesamiento de lenguaje natural.

Se presentó el proyecto *El Aula del Futuro* y se presentó la problemática de este trabajo: automatizar la tarea de encontrar las inquietudes que comparten los profesores participantes en el proyecto. Se propuso la hipótesis de este trabajo: identificar, mediante algoritmos de conglomeración, los patrones (inquietudes compartidas) en las reflexiones de los profesores. Se presentó la metodología utilizada en el análisis: se utilizaron métodos de procesamiento de lenguaje natural para mapear el conjunto de reflexiones de los profesores a una matriz numérica.

Se agrupó con las implementaciones actuales en Python de *k-medias*, AGNES y DBSCAN. Se utilizó el método del codo para encontrar el número óptimo de grupos para *k-medias* y AGNES. Para DBSCAN se buscó entre diferentes configuraciones de parámetros generadas aleatoriamente. Al final, *k-medias* y AGNES encontraron cinco grupos y DBSCAN encontró 10.

Se analizó cada grupo obtenido por cada algoritmo de conglomeración para comprobar que había coherencia en la forma de agrupamiento. Este análisis consistió en revisar los bi-gramas y tri-gramas de *Stem* más frecuentes.

Los resultados fueron bastante prometedores, ya que en su mayoría se logró caracterizar a los grupos de acuerdo a una temática en común que era coherente con lo observado en las reflexiones elaboradas por los profesores. En particular, el grupo más grande que encontró AGNES fue el grupo más prometedor porque en él se logró encontrar lo que se buscaba originalmente: problemáticas que podrían compartir los profesores y las estrategias que utilizaron para enriquecer sus clases.

Por otro lado, los grupos encontrados por *k-medias* fueron los que por sí solos tuvieron un patrón más visible en sus grupos. AGNES no tuvo un patrón tan visible en sus grupos pero sí encontró el grupo más relevante. DBSCAN entregó grupos muy pequeños y por la forma en que agrupa, entregó los grupos más limpios, por lo que era de esperarse que la temática de sus grupos fuera también bastante perceptible.

Cuando se habla de resultados prometedores, se hace referencia a que utilizando esta misma metodología, pero trabajando adicionalmente con los nuevos datos que se vayan generando en los siguientes años del proyecto, este enfoque podría ser una muy buena técnica para identificar las problemáticas que comparten los profesores y con esto las estrategias que usan para resolverlos, lo cual será de gran ayuda para el proyecto del Aula del Futuro.

Hay que mencionar que este enfoque es atractivo porque ayuda a relacionar las reflexiones de los profesores que han tenido las mismas problemáticas educativas en distintos niveles (bachillerato, universidad, etc.), y esto permitiría que los profesores puedan comparar cómo se resolvieron y así mejorar sus estrategias didácticas.

Por lo que es buena idea para trabajos futuros complementar este enfoque con técnicas de visualización de información que faciliten la exploración de los grupos encontrados, para una interpretación más amigable.

Desde el punto de vista personal, considero que los conocimientos adquiridos en la carrera fueron cruciales para desarrollar este trabajo, más puntualmente, sin los conocimientos de estadística, manejo de bases de datos, análisis numérico, programa-

ción, y la capacidad de análisis profundo que se desarrolla en la carrera, no hubiera podido llevar a cabo este trabajo.

Así que, por lo anterior se concluye este trabajo afirmando que existen patrones en las reflexiones de los profesores sobre sus experiencias, éxitos y dificultades en el proyecto *El Aula del Futuro* que pueden ser identificados gracias a los algoritmos de conglomeración.

Apéndice A

Lista de palabras vacías

a uea al algo algunas algunos ante antes así asi aunque cada como cómo con contra cual cuál cuáles cuales cuando cuándo cuyo cuya cuyos cuyas de dé del desde donde dónde durante e él el ella ellas ello ellos en entre era erais eramos éramos eran eras eres es esa esas ese eso esos esta está están estaba estabais estábamos estaban estabas estad estada estadas estado estados estáis estamos estando estar estaremos estará estarán estarás estaré estaréis estaría estaríais estaríamos estarían estarías estas estás este esté estéis estemos estén estés esto estos estoy estuve estuviera estuvierais estuvieran estuvieras estuvieron estuviese estuvieseis estuviesen estuvieses estuvimos estuviste estuvisteis estuviéramos estuviésemos estuvo fue fuera fuerais fueran fueras fueron fuese fueseis fuesen fueses fui fuimos fuiste fuisteis fuéramos fuésemos ha habida habidas habido habidos habiendo habremos habrá habrán habrás habré habréis habría habríais habríamos habrían habrías habéis había habíais habíamos habían habías han has hasta hay haya hayamos hayan cada hayas hayáis he hemos hube hubiera hubierais hubieran hubieras hubieron hubiese hubieseis hubiesen hubieses hubimos hubiste hubisteis hubiéramos hubiésemos hubo hace hacia la las le les lo los mas me mi mis mucho muchos muy más mí mía mías mío míos nada ni no nos nosotras nosotros nuestra nuestras nuestro nuestros o os otra otras otro otros para pero por porque que quien quienes quién quiénes qué quizá quizás quiza quizás se sé sea seamos sean seas sentid sentida sentidas sentido sentidos seremos será serán serás seré seréis sería seríais seríamos serían serías seáis si sí siente sin sintiendo sobre sois somos son soy su sus

suya suyas suyo suyos tambien también tan tanto tantos te tendremos tendrá tendrán tendrás tendré tendréis tendría tendríais tendríamos tendrían tendrías tened tenemos tenga tengamos tengan tengas tengo tengáis tenida tenidas tenido tenidos teniendo tenéis tenia teniamos tenía teníais teníamos tenían tenian tenias tenías ti tiene tienen tienes todo todos tu tú tus tuve tuviera tuvierais tuvieran tuvieras tuvieron tuviese tuvieseis tuviesen tuvieses tuvimos tuviste tuvisteis tuviéramos tuviésemos tuvo tuya tuyas tuyo tuyos tú traves través un u e una uno unos varias varios vosotras vosotros vuestra vuestras vuestro vuestros y ya yo i iii ii v iv vi vii viii ix uama x llevar cabo estudiante estudiantes alumno alumnas alumna alumnos

Apéndice B

Lista de *Stem* con frecuencia 1

['anim', 'penultim', 'horari', 'profesionaliz', 'mancomun', 'impos', 'acomod', 'guionism', 'productor', 'director', 'escenograf', 'vestuari', 'sonid', 'postproduccion', 'edicion', 'credit', 'escalafon', 'lider', 'subalt', 'satisfag', 'narrat', 'terci', 'balanc', 'tonal', 'confusion', 'escen', 'dig', 'siqu', 'urgenci', 'veloz', 'exces', 'descu', 'generaliz', 'estrict', 'exig', 'sugerent', 'cuantit', 'cognit', 'conversion', 'recuerd', 'supon', 'acab', 'extracurricul', 'expliqu', 'manipul', 'personaliz', 'flu', 'net', 'emergent', 'oid', 'audicion', 'amplitud', 'sonor', 'antepas', 'supervis', 'simultan', 'sencill', 'especializ', 'cuerd', 'resort', 'cub', 'grabador', 'musical', 'guitarr', 'flaut', 'tecl', 'voc', 'cor', 'daran', 'equivalent', 'identità', 'lavor', 'messic', 'ofici', 'jorn', 'suger', 'repart', 'fastid', 'music', 'favorit', 'desperdici', 'certific', 'familiaric', 'austriac', 'imagin', 'aprehend', 'lexical', 'serv', 'rode', 'menud', 'alemani', 'austri', 'sensibiliz', 'desemple', 'optimiz', 'persig', 'que', 'convert', 'inclin', 'propus', 'tecnopedagog', 'abproyect', 'realist', 'residual', 'urban', 'anunci', 'rendimient', 'regulariz', 'prevalec', 'educativoscompañer', 'forj', 'soci', 'refuerc', 'reeduqu', 'perici', 'induc', 'plagi', 'escasez', 'demerit', 'obvied', 'reaprendizaj', 'escaz', 'propedeut', 'zon', 'proxim', 'afianz', 'tronc', 'rezag', 'fertil', 'frut', 'congres', 'metacognit', 'autorregul', 'peldañ', 'apa', 'sintet', 'moviliz', 'mimesis', 'estiliz', 'acert', 'bondad', 'for', 'septiembr', 'transdisciplin', 'actu', 'extraccion', 'opuest', 'vs', 'capilar', 'por', 'humed', 'transitori', 'contadur', 'negoci', 'fidedign', 'flojer', 'imprescind', 'porqu', 'afirm', 'infier', 'cataliz', 'acid', 'proust', 'volumen', 'replic', 'reafirm', 'romp', 'covalent', 'diatom', 'celd', 'electrolit', 'recipient', 'reactor', 'racional',

'deduct', 'convergent', 'divergent', 'proactiv', 'conserv', 'magic', 'distingu', 'pseudocient', 'decision', 'asunt', 'tecnocientif', 'multifactorial', 'generacional', 'aburr', 'irrelev', 'añad', 'mayoritari', 'emocion', 'curi', 'noven', 'penetr', 'nominal', 'dict', 'youtub', 'consistent', 'tesiunam', 'formulari', 'temporal', 'sucint', 'resist', 'hoj', 'prob', 'predec', 'reconfigur', 'leyer', 'quer', 'pow', 'point', 'recod', 'val', 'sinopt', 'diagram', 'apell', 'dav', 'jos', 'zamudi', 'angel', 'leydy', 'ale', 'eraz', 'ñañez', 'extension', 'envuelv', 'cualidad', 'predomin', 'evidenci', 'inconvenient', 'reprsent', 'holist', 'anatom', 'pilar', 'despert', 'urgent', 'tedios', 'progres', 'moderniz', 'disciplinar', 'epidemiolog', 'rang', 'termodinam', 'phreeqc', 'ansied', 'diaposit', 'pantall', 'ir', 'instal', 'pestañ', 'interf', 'comp', 'captur', 'avis', 'catalog', 'matern', 'milenari', 'tradicion', 'xxi', 'clerig', 'erudit', 'xvii', 'decrec', 'mercantil', 'colonial', 'franc', 'anquil', 'xix', 'suprim', 'españ', 'desinent', 'rus', 'bat', 'desdeñ', 'defiend', 'traz', 'abriend', 'zanj', 'cimient', 'descart', 'franj', 'neutr', 'cans', 'extrañ', 'cancion', 'sobrecarg', 'literario...', 'omit', 'extint', 'dialectal', 'renac', 'neurolingüist', 'factibl', 'corrig', 'error', 'calc', 'renacent', 'neolog', 'consid', 'argentin', 'inglaterr', 'ludus', 'schol', 'collegium', 'institutum', 'antigu', 'sinergi', 'enriquezc', 'organ', 'incident', 'disposit', 'invident', 'solidaric', 'preven', 'obstacul', 'gent', 'manifest', 'huell', 'intensific', 'region', 'migratori', 'caravan', 'migrant', 'centroamer', 'estadounidens', 'cruz', 'fronter', 'trayect', 'emigr', 'inmigr', 'reorden', 'rotur', 'reactiv', 'nanoscop', 'simbol', 'octav', 'parcializ', 'inmers', 'alons', 'conexion', 'charl', 'coincident', 'museograf', 'gal', 'luis', 'nishizaw', 'interrump', 'coincid', 'simil', 'logist', 'especific', 'exhibicion', 'planteamient', 'situacional', 'eventual', 'festiv', 'toler', 'division', 'privilegi', 'priorid', 'sort', 'interdependent', 'contrarrest', 'vici', 'menor', 'incurr', 'copy', 'past', 'transcripcion', 'induzc', 'inspir', 'tics', 'aparezc', 'preferent', 'intercomun', 'saqu', 'mediocr', 'confianz', 'inmediat', 'leyeron', 'recodific', 'exponent', 'coadyuv', 'realidad', 'intrafamiliar', 'preliminar', 'multicausal', 'gestacion', 'zapat', 'luch', 'libertad', 'hombr', 'iconograf', 'reconstruccion', 'rectangul', 'rig', 'gobiern', 'aquel', 'rect', 'aal', 'faltant', 'errone', 'sexagesimal', 'radian', 'habitu', 'diferenci', 'imaginari', 'parafrasis', 'citacion', 'pers', 'reiter', 'definicion', 'contrapon', 'mision', 'jerarquiz', 'clasic', 'grand', 'recr', 'prejuici', 'matadit', 'influenci', 'leeran', 'anticip', 'contraposicion', 'presum', 'ignor', 'grues', 'encomend', 'resumen', 'platiqu',

'breved', 'compr', 'vertient', 'mient', 'engañ', 'acas', 'dij', 'ayer', 'asidu', 'salg', 'biograf', 'quijot', 'contest', 'signif', 'coevalu', 'infogram', 'historiet', 'guion', 'sustitu', 'silenci', 'inculc', 'busqu', 'transf', 'postul', 'estereoscop', 'usaron', 'optic', 'agudiz', 'tej', 'dandol', 'organel', 'membran', 'nucle', 'cloroplast', 'mitocondri', 'citoplasm', 'transversal', 'virus', 'sars', 'cumplimient', 'motor', 'natural', 'alusion', 'constitu', 'justifiqu', 'subjet', 'dogm', 'corpuscul', 'mov', 'unirs', 'olor', 'textur', 'bell', 'difusion', 'pie', 'tac', 'sociedad', 'omision', 'ocho', 'mont', 'latinoamerican', 'auxili', 'organizacion', 'recodificacion', 'modelis', 'sistematizacion', 'falsacion', 'contrast', '«', '»', 'deductiv', 'inductiv', '•', 'ido', 'parezc', 'pul', 'diari', 'debil', 'intervin', 'hardwar', 'hered', 'aparat', 'firm', 'er', 'fich', 'word', 'chicag', 'instruct', 'ocasional', 'estratigraf', 'pincipal', 'entrevist', 'campus', 'ciud', 'impresion', 'sic', 'lagun', 'reclut', 'percepcion', 'simplific', 'subterrane', 'hidrocarbur', 'solut', 'ecuacion', 'num', 'com', 'codig', 'algoritm', 'imparticion', 'rediseñ', 'voy', 'usad', 'cubiert', 'hav', 'should', 'must', 'categor', 'usam', 'conjug', 'drill', 'demasi', 'precari', 'geobiolog', 'postgr', 'planet', 'licenci', 'ingenier', 'prepar', 'haran', 'examin', 'europe', 'ampliacion', 'raic', 'loabl', 'herenci', 'grieg', 'visibl', 'diccionari', 'list', 'dispers', 'fraccion', 'cincuent', 'increment', 'vean', 'forz', 'min', 'narracion', 'crisis', 'gubernamental', 'videoconferent', 'grab', 'conector', 'lugar', 'driv', 'dand', 'finaliz', 'marcador', 'be', 'pragmat', 'vacacion', 'oracion', 'first', 'second', 'third', 'lat', 'then', 'finally', 'and', 'but', 'or', 'also', 'becaus', 'sociolingüist', 'automonitor', 'curricul', 'psicoanal', 'membres', 'liderazg', 'origen', 'secuencial', 'incomplet', 'disponibil', 'iniciar', 'arid', 'remembr', 'benef', 'optad', 'desinteres', 'admision', 'imprecision', 'inconsistent', 'reconstru', 'repit', 'memoric', 'discrimin', 'popul', 'saber', 'incis', 'imper', 'carid', 'filantrop', 'cap', 'impuls', 'receptor', 'entusiasm', 'apasion', 'exposit', 'fot', 'gir', 'refer', 'progresion', 'acent', 'literatur', 'novel', 'abordaj', 'primari', 'pobr', 'estet', 'ficcional', 'arquitectur', 'repeticion', 'curricular', 'agudic', 'echar', 'convencional', 'ric', 'conect', 'sensibil', 'rigidez', 'cumul', 'diseccion', 'autopsi', 'tridimensional', 'vea', 'fals', 'heterogene', 'jov', 'entrelaz', 'demand', 'recab', 'cmi', 'planific', 'exit', 'subdivid', 'record', 'inmens', 'dobl', 'efectu', 'comprens', 'compet', 'gast', 'dañ', 'riesgos', 'sintomatolog', 'alergi', 'dolor', 'naus', 'antibiot', 'cep', 'vent', 'instanci', 'neuron', 'recodif', 'afin', 'insert', 'suscept', 'deduc', 'fren', 'praxis', 'men-

saj', 'ed', 'sustant', 'ca', 'ce', 'aprendient', 'ejecut', 'contrari', 'prove', 'crucial', 'asegur', 'subhabil', 'pre', 'kahoot', 'mindmast', 'h', 'hit', 'terminal', 'organizacional', 'inaplic', 'decl', 'psicosocial', 'diferencial', 'libert', 'habitual', 'examen', 'prediseñ', 'demuestr', 'imposibil', 'cer', 'deteccion', 'procur', 'tir', 'parabol', 'bidimensional', 'desaparec', 'vuel', 'confus', 'acostumbr', 'empez', 'problematiz', 'sosten', 'pandem', 'sac', 'flot', 'comorbil', 'supermerc', 'refresc', 'enfermedad', 'sindrom', 'cuell', 'descrit', 'soport', 'telefon', 'entreten', 'oci', 'conceb', 'subordin', 'encicloped', 'vigil', 'castig', 'suprem', 'vuelc', 'responsabiliz', 'contrapart', 'concib', 'omar', 'mand', 'extra', 'esfer', 'dra', 'rosari', 'freix', 'agent', 'heteroevalu', 'cualit', 'infinit', 'dtransparent', 'democratiz', 'aclar', 'cierr', 'condens', 'plenari', 'tik', 'tok', 'peligr', 'drenaj', 'concientiz', 'sobrant', 'ambiental', 'farmaceut', 'farmac', 'inadvert', 'excrecion', 'inadecu', 'estigmatiz', 'predisposicion', 'local', 'preparatorian', 'dim', 'entrad', 'subproces', 'venc', 'almacen', 'string', 'precedent', 'rubr', 'esencial', 'inminent', 'codif', 'semiot', 'introductori', 'organic', 'terapi', 'capitul', 'treatment', 'levi', 'et', 'inferent', 'jueg', 'panel', 'fung', 'audienci', 'moder', 'voz', 'bastant', 'confirm', 'descripcion', 'expong', 'premur', 'jug', 'land', 'traid', 'encuadr', 'fusion', 'dr', 'guillerm', 'davil', 'supervisor', 'recib', 'conmut', 'consej', 'brev', 'audiograb', 'estatut', 'pinterest', 'aterrizaj', 'tactic', 'finalic', 'pyme', 'exigent', 'tendenci', 'eficaz', 'boton', 'tentat', 'leal', 'links', 'requ', 'wi', 'fi', 'neurocient', 'psicoanalisis', 'mas', 'gestalt', 'ganch', 'persuas', 'masiv', 'wan', 'na', 'click', 'wix', 'back', 'instagram', 'emot', 'globaliz', 'imposicion', 'geopolit', 'transnacional', 'reform', 'anteproyect', 'anuanci', 'estan', 'ocacion', 'dem', 'pobrez', 'tant', 'rep', 'cusion']

Bibliografía

[1] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT Press, 2016.

[2] Stuart J. Russell and Peter Norvig. *Artificial intelligence: a modern approach*. Prentice-Hall, 2010.

[3] Richard Bellman. *An introduction to artificial intelligence: can computers think?* Boyd & Fraser Pub. Co., 1978.

[4] Patrick Henry. Winston. *Artificial intelligence*. Addison-Wesley, 1992.

[5] Ray Kurzweil. *The age of intelligent machines*. MIT Press, 1990.

[6] David Poole, Alan K. Mackworth, and Randy G. Goebel. *Computational intelligence: a logical approach*. Oxford University Press, 1998.

[7] Tom M. Mitchell. *Machine learning*. McGraw Hill, 2017.

[8] Richard S Sutton and Andrew Barto. *Reinforcement learning: an introduction*. The MIT Press, 2018.

[9] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag New York, 2016.

[10] Leonard Kaufman and Peter J. Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. Wiley-Interscience, 2005.

[11] Joseph Lee Rodgers and W. Alan Nicewander. Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1):59, Feb 1988.

[12] Alboukadel Kassambara and Alboukadel Kassambara. *Practical guide to cluster analysis in R: unsupervised machine learning*. STHDA, 2017.

[13] J. A. Hartigan and M. A. Wong. Algorithm as 136: A k-means clustering algorithm. *Applied Statistics*, 28(1):100, 1979.

[14] Ian H. Witten, Eibe Frank, and Mark A. Hall. *Data mining: practical machine learning tools and techniques*. Elsevier, 2011.

[15] SUSANA A. LEIVA-VALDEBENITO and FRANCISCO J. TORRES-AVILÉS. Una revisión de los algoritmos de partición más comunes en el análisis de conglomerados: un estudio comparativo. *Revista Colombiana de Estadística*, 33:321 – 339, 12 2010.

[16] Erich Schubert and Peter J. Rousseeuw. Faster k-medoids clustering: Improving the pam, clara, and clarans algorithms. *Similarity Search and Applications Lecture Notes in Computer Science*, page 171–187, 2019.

[17] Jörg Sander Xiaowei Xu Martin Ester, Hans-Peter Kriegel. A density-based algorithm for discovering clusters in large spatial databases with noise. *Institute for Computer Science. University of Munich*, 1996.

[18] Michael Hahsler, Matthew Piekenbrock, and Derek Doran. dbscan: Fast density-based clustering with r. *Journal of Statistical Software*, 91(1), 2019.

[19] Sergios Theodoridis and Konstantinos Koutroumbas. *Pattern recognition*. Syngress, 2008.

[20] Purnima Bholowalia and Arvind Kumar. Article: Ebk-means: A clustering technique based on elbow method and k-means in wsn. *International Journal of Computer Applications*, 105(9):17–24, November 2014. Full text available.

[21] Dhendra Marutho, Sunarna Hendra Handaka, Ekaprana Wijaya, and Muljono. The determination of cluster number at k-mean using elbow method and purity

evaluation on headline news. *2018 International Seminar on Application for Technology of Information and Communication*, 2018.

[22] Diego Lopez Yse. Your guide to natural language processing (nlp), Apr 2019.

[23] Sudha-Nadchal. Text clustering, Apr 2020.