



UNIVERSIDAD TÉCNICA DE BABAHOYO

**FACULTAD DE ADMINISTRACIÓN, FINANZAS E
INFORMÁTICA**

PROCESO DE TITULACIÓN

NOVIEMBRE 2021 – ABRIL 2022

**EXAMEN COMPLEXIVO DE GRADO O DE FIN DE CARRERA
PRUEBA PRACTICA**

**INGENIERÍA EN SISTEMAS
PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERO (A)
EN SISTEMAS**

TEMA

**ANÁLISIS DE LAS CARACTERÍSTICAS DE LOS TIPOS DE
ALGORITMOS DE CLUSTERING EN EL APRENDIZAJE NO
SUPERVISADO**

EGRESADO

ANGEL STEVEN CHOEZ FRANCO

TUTOR:

ING. JOSÉ TEODORO MEJÍA VITERI, MSC

AÑO

2022

Resumen

En los últimos años, los avances tecnológicos han llevado a la generación de grandes volúmenes de datos, y uno de los problemas es la hora de clasificar y extraer datos, por ello el aprendizaje no supervisado juega un papel fundamental en el proceso utilizando los tipos de algoritmos de clustering.

Por ello el presente estudio de caso, se basa en realizar un “análisis de las características de los tipos de algoritmos de clustering en el aprendizaje no supervisado”, cuyo objetivo es analizar las características de los tipos de algoritmos de clustering ya que estos algoritmos se basan en la suposición de que los patrones se pueden agrupar en función de su similitud. Es decir que realiza un proceso para explorar y analizar los datos donde se desconoce la estructura que tienen, cuya finalidad es encontrar patrones en los datos que formen grupos con características similares.

Palabras clave

Clustering, análisis, patrones de datos.

Abstract

In recent years, technological advances have led to the generation of large volumes of data, and one of the problems is the time to classify and extract data, so unsupervised learning plays a fundamental role in the process using the types of clustering algorithms.

Therefore, the present case study is based on performing an "analysis of the characteristics of the types of clustering algorithms in unsupervised learning", whose objective is to analyze the characteristics of the types of clustering algorithms since these algorithms are based on the assumption that patterns can be grouped according to their similarity. That is, it performs a process to explore and analyze the data where the structure they have is unknown, whose purpose of finding patterns in the data that form groups with similar characteristics.

Keywords

Clustering, analysis, data patterns.

ÍNDICE

Resumen.....	i.
Abstract.....	ii.
1. INTRODUCCION.....	1-2
2. DESARROLLO.....	3
2.1. Planteamiento del problema.....	3
2.2. Justificación.....	3
2.3. Metodología.....	4
2.4. Marco teórico.....	4
.....	
2.4.1 Aprendizaje no supervisado.....	4
2.4.2 Análisis de clustering.....	5
2.4.3 Tipos de clustering.....	5
2.4.3.1 Agrupamiento por particiones.....	6
2.4.4.2 Basados en densidad.....	6
2.4.4.3 DBSCAN.....	6
2.4.4.4 Basados en grafos.....	7
2.4.4.5 Algoritmo Kruskal.....	7
2.4.4.6 Mínimo error cuadrático.....	8
2.4.4.7 K-Medias.....	8
2.4.4.8 Característica de los algoritmos agrupamiento por particiones.....	9
2.4.4.9 Ventajas y desventajas de los algoritmos agrupamiento por particiones.....	10-13
2.4.4.10 Jerárquico.....	11
2.4.4.11 Probabilístico.....	
2.5 Propuesta de la solución.....	14-15
3. CONCLUSIONES.....	16
4. BIBLIOGRAFICA.....	17-19
ANEXOS	

Introducción

En los últimos años, los avances tecnológicos han llevado a la generación de grandes volúmenes de datos, en su mayoría datos digitales, destacando el valor de procesarlos para extraer conocimiento e información.

En el campo del reconocimiento de patrones de datos, es común encontrar problemas en extraer datos debido a los grandes volúmenes de datos, lo que puede representar un problema para tareas de clasificación posteriores. Por ello el aprendizaje no supervisado juega un papel fundamental en el proceso de modelado que lleva a cabo sobre un conjunto de ejemplos formado tan solo por entradas al sistema, que tiene ser capaz de reconocer patrones para poder etiquetar las nuevas entradas es decir que trabaja con datos que no han sido etiquetado. No se tiene una etiqueta que predecir.

Uno de los principales tipos de algoritmos del aprendizaje no supervisado son los de “clustering”, que tienen el propósito de agrupar datos por similitud, sin previa imposición de restricciones por parte del experto o analista, buscando la extracción de características donde cada objeto queda representado por una colección de descriptores, permitiendo generar información valiosa que será convertida en conocimiento.

El presente estudio de caso, se basa en realizar un “análisis de las características de los tipos de algoritmos de clustering en el aprendizaje no supervisado”, que tiene como objetivo analizar las características de los tipos de algoritmos clustering, basados en técnicas de agrupación. La línea de investigación para el desarrollo de la presente se vincula con los tipos de algoritmos de clustering en el aprendizaje no supervisado que son utilizados en muchas aplicaciones que requieran clasificación de datos, como la minería de datos, inteligencia artificial, aprendizaje de máquina, estadísticas, entre otros.

La metodología aplicada para recopilar información fue la cualitativa e interpretativa, de tipo documental. Se utilizo artículos publicados en revista de divulgación científica, libros y documentos, entre otros, como herramienta que, a través de ello se recaudó información

necesaria para analizar e interpretar temas relacionados de la temática. Este estudio de caso está organizado de la siguiente forma: en la primera parte se plantea el problema encontrado en los tipos de algoritmos de clustering, en la segunda parte se justifica el análisis de las características de los tipos de algoritmos de clustering, en la tercera parte se realiza la metodología aplicada en este estudio investigativo, en la cuarta parte se muestran información de los tipos de algoritmos de clustering, por último, se presentan las conclusiones.

Desarrollo

El aprendizaje no supervisado busca identificar patrones existentes, por ello el análisis de datos de los diferentes algoritmos clustering en el aprendizaje no supervisado es útil en aplicaciones que requieren clasificación de datos, como en la minería de datos, agrupar palabras con definiciones similares para una mejor precisión del motor de búsqueda, etc (Llaque, 2018).

Este estudio de caso pretende ser una respuesta a la formulación del problema desarrollado a partir de las características de los tipos de algoritmos clustering:

¿Cuál es el análisis de las características de los tipos de algoritmos clustering en el aprendizaje no supervisado?

El análisis de las características de los tipos de algoritmos clustering es una técnica exploratoria de análisis de datos para resolver problemas de clasificación. Consiste en identificar qué relaciones existen entre las variables del estudio como, por ejemplo, entre esos están variables, plantas, personas, animales, cosas, marketing, biología así entre otros (Ceron, 2018).

Con el fin de clasificar y analizar los datos, descubrir conocimiento sobre ellos se emerge como un campo de investigación interdisciplinario de áreas como bases de datos, aprendizaje de máquina, inteligencia artificial, estadística, entre otros. Entre esos están los tipos de clustering o métodos de clustering, tales como los de agrupamiento por particiones y los jerárquico que son algoritmos para la precisión en la tarea de clasificar datos que contribuyen en la definición formal de un sistema de clasificación como un clasificador para un conjunto de objetos (Alvaro, 2017).

La metodología aplicada en el siguiente estudio de caso fue la investigación documental, que a través de un proceso sistemático de indagación, análisis e interpretación de información se define con el objetivo de analizar las características de los tipos de algoritmos de clustering en el aprendizaje no supervisado, como primera parte de este estudio analítico tenemos el aprendizaje no supervisado que trabaja con datos no etiquetados. Es decir que no tiene etiquetas para predecir, la cual a su vez menciona su clasificación o tipos de clustering que se dividen en: agrupamiento por particiones, agrupamiento jerárquico, agrupamiento probabilístico, a partir de los tipos de clustering se analizara la funcionalidad de los algoritmos con el fin de indagar los diferentes algoritmos en técnicas de agrupamiento.

En el diseño del estudio de caso, como segunda parte se realizará un estudio comparativo de las características, ventajas y desventajas de los diferentes tipos de algoritmos de clustering, la cual permite determinar la forma en la que se agrupa los datos ya que cada uno de los diferentes tipos de algoritmos de clustering tienen técnicas diferentes de interpretar en la manera de agrupar los datos.

Finalmente, en este estudio se determina el resultado del análisis de los algoritmos, a través de la herramienta Jupyter Notebook con el lenguaje de programación python la cual, se realizó la funcionalidad de los algoritmos como resultado se observó, unos de los primeros pasos a realizar es la normalización de los datos, como segundo paso se indica el número de cluster, como tercer paso nos muestra un resultado en representación gráfica la cantidad de grupos formados, pero dependiendo del tipo de algoritmo su grafica puede ser tablas, dendrogamas, gráficos entre otros para luego ser interpretados con facilidad.

Aprendizaje no supervisado

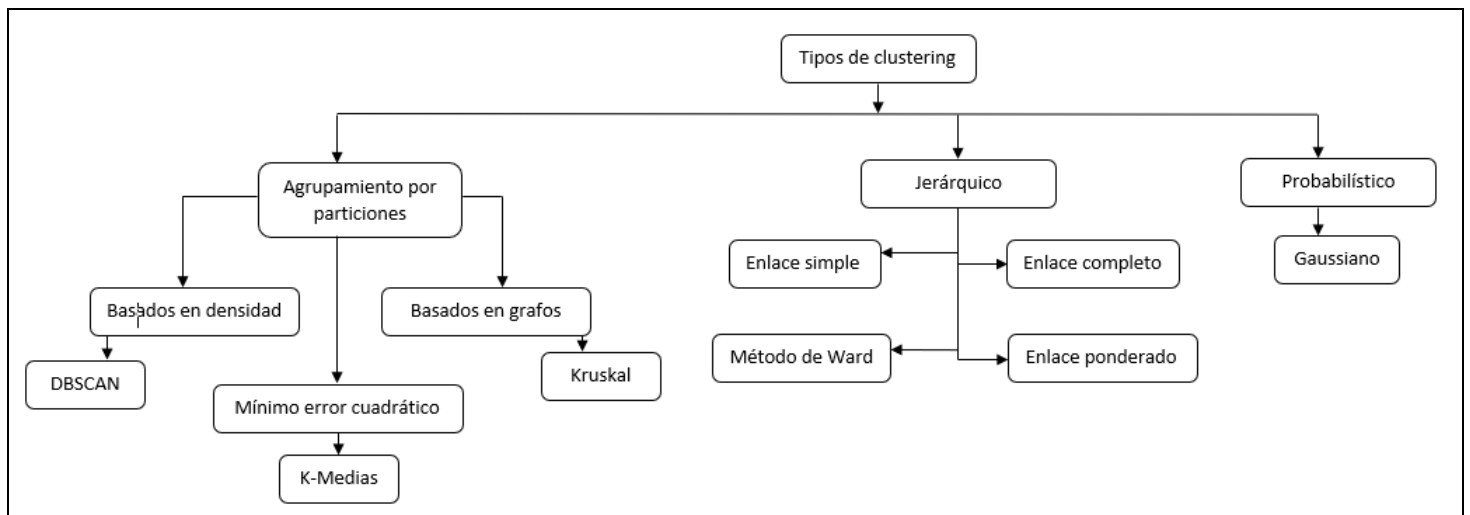
El aprendizaje no supervisado, funciona con datos no etiquetados. No tienes etiquetas para predecir. Estos algoritmos se utilizan principalmente en tareas en las que se

deben analizar datos para obtener nuevos conocimientos o agrupar entidades por relación (Vallalta Rueda, 2019).

Análisis de clustering

Análisis de clúster (también conocido como análisis de clustering o simplemente agrupamiento). Es la tarea de agrupar un conjunto de objetos (no etiquetados) en subconjuntos de objetos llamados clusters. Cada cluster consta de una colección de objetos que son similares o se considerados similares entre sí, pero diferentes de los objetos de otros clusters (Moya, 2016).

Es decir que los clustering es una técnica para explorar y analizar los datos donde se desconoce la estructura que tienen, un modelo de clustering tendrá la finalidad de encontrar patrones en los datos que formen grupos de datos con características similares.



Tipos de clustering Figura 1 *Tipos de Clustering*

Agrupamiento por particiones

Los algoritmos de agrupamiento de particiones generan múltiples particiones y luego las puntúan o evalúan según ciertos criterios. También se conocen como no jerárquicos, ya que cada instancia se sitúa exactamente no de k-clusters mutuamente excluyentes. Dado que solo un conjunto de clústeres es el resultado de un algoritmo de clúster de partición atípico,

Una división simple de un conjunto de datos en subconjuntos discretos que no se superponen, de modo que cada punto del conjunto pertenezca a uno de dichos subconjuntos o clusters (Alcalde, 2018).

Basados en densidad

La herramienta de Clustering basado en densidad se basa en detectar qué áreas tienen una concentración de puntos y dónde están separados por áreas vacías o de puntos bajos. Los puntos que no forman parte de un grupo o clúster se etiquetan como ruido.

En este tipo de clustering, un cluster o grupo es una región densa de objetos rodeada por un área de baja densidad. Suele usarse cuando hay ruido y outliers presentes en los datos (Alcalde, 2018). El principal algoritmo utilizada en esta técnica se puede mencionar: DBSCAN.

Density Based Scan Clustering (DBSCAN)

El algoritmo DBSCAN identificar clústeres y filtrar valores atípicos sin saber el número de clústeres reales lo que se basa en el concepto de áreas densas para formar clústeres de datos (Zhang, 2018). Es decir que Consiste en medir la densidad como el número de puntos que caen dentro de un radio especificado.

Funcionamiento del algoritmo Density Based Scan Clustering (DBSCAN)

- **Paso 1:** inicia con un punto de datos de inicio arbitrario. El vecindario de este punto se extrae usando la distancia ϵ , todos los puntos que están dentro de la distancia de la distancia de ϵ son puntos de vecindario.
- **Paso 2:** si hay un suficiente de puntos dentro de este vecindario, entonces el proceso de agrupación comenzará y el punto de datos actual se convierte en el primer punto del nuevo grupo. Por ende, el punto se etiquetará como ruido en su lugar.
- **Paso 3:** para este punto en el nuevo cluster, los puntos dentro de su vecindario distante ϵ también pasan a formar parte del mismo cluster.
- **Paso 4:** en este proceso de los pasos 2 y 3 se repite hasta que se hayan identificado todos los puntos en el cluster, es decir, que se hayan visitado y etiquetado todos los puntos dentro del vecindario ϵ del cluster.
- **Paso 5:** una vez que se haya terminado con el cluster actual, se buscare y procesara un nuevo punto no alcanzado o visitado, lo que resultara en el descubrimiento de otro cluster o ruido. Este procedimiento se repite hasta que todos los puntos se marcan como visitados.

Basados en grafos

Según Sandra (2018), un grafo es una estructura de datos que consiste en un conjunto de objetos llamados vértices o nodos conectados por enlaces llamados aristas o arcos, que ayudan a representar las relaciones binarias entre elementos de un conjunto. El principal algoritmo utilizada en esta técnica se puede mencionar: Kruskal.

Algoritmo Kruskal

El algoritmo de Kruskal es un proceso que permite unir todos los nodos de un grafo formando un árbol, tomando en cuenta el peso de las aristas y cuyo coste total es el mínimo posible (Benavides, 2017).

Funcionamiento del algoritmo Kruskal

- **Paso 1:** Ordenar las aristas de menor a mayor peso.
- **Paso 2:** Unir las aristas con sus vértices siempre y cuando estos no formen ciclos (Martínez, 2018).

Mínimo error cuadrático

Según (Alcalde, 2018), afirma: “en este algoritmo, utiliza la minimización error cuadrático para determinar a qué cluster o grupo pertenece el punto. Esta técnica es utilizada por el algoritmo K-Medias”.

K-Medias

K-Means es un algoritmo de agrupamiento de partición. Tiene un parámetro de entrada, k , que indica la cantidad de clústeres a generar, por lo que la cantidad de clústeres a buscar debe conocer de antemano.

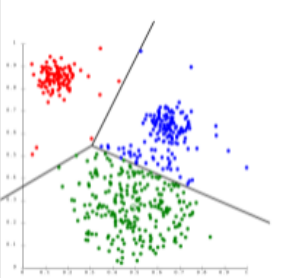
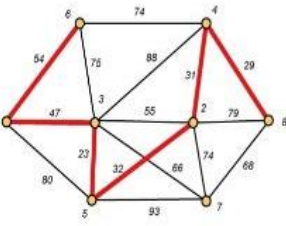
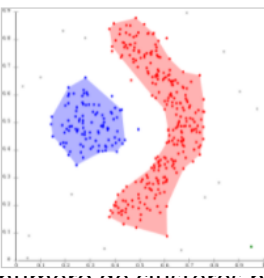
En este algoritmo, el usuario especifica el número de agrupaciones K . Cada agrupación tiene un vector medio (centroide). La distancia Euclídea se utiliza para encontrar el clúster más cercano al objeto (Vintimilla, 2017).

Funcionamiento del algoritmo K-Medias

- **Paso 1:** Seleccionar el número de cluster (K) que deseas identificar en los datos. En este caso $K = 3$.
- **Paso 2:** Seleccionar aleatoriamente K puntos. No tienen que ser puntos de nuestros datos, pueden ser puntos nuevos.
- **Paso 3:** Medimos la distancia entre cada uno de los datos y los puntos seleccionados, asignándole el punto que se encuentre más cerca.
- **Paso 4:** Colocamos nuevos K puntos y repetimos el procedimiento.

Particional			
	Mínimo error cuadrático	Modelo de grafos	Modelo de densidades
	K-Medias	Kruskal	DBSCAN
Características	Agrupar observaciones con características similares (Capretto, 2018).	Selecciona y analiza las aristas más pequeñas que no se han agregado a la solución (Concatto, 2017)	Establece un radio para buscar vecinos cercanos (Salazar, 2017)

Fuente: Angel Steven Choez

<p>Representación grafica</p> <p>Ventajas</p>	 <p>Es bajo lo que calcula</p> <p>Figura 2 Algoritmo K-Medias</p> <p>que hay muy pocos cálculos</p> <p>Fuente: https://www.cs.us.es/~fsancho/?e=230</p>	 <p>Figura 3 Algoritmo Kruskal</p> <p>Complejidad menor con otros algoritmos (Sanchez, 2017)</p> <p>Fuente: https://www.slideshare.net/erickestradamancilla/algorithmodekruskal-72069816</p>	 <p>numero de cluisteres para</p> <p>Figura 4 Algoritmo DBSCAN</p> <p>Identica a los valores atipicos</p> <p>Fuente: https://www.cs.us.es/~fsancho/?e=230</p>
<p>Tabla 1 Comparación de las Características del Método Particional</p> <p>del Método Particional, incluso si el punto es muy diferente.</p>			

A continuación, se expresa la información de la presente tabla:

Comparación de las Características del Método Particional tabla 1, muestra que el algoritmo K-Medias es para soluciones iterativas es decir que es simple, rápido, adecuado para conjuntos de datos regulares, el algoritmo Kruskal trata de encontrar el camino o la ruta más corta hacia un nodo y el algoritmo DBSCAN encuentra valores atípicos es decir que es bueno para tareas de detección.

Desventaja	<p>-> Tiene que seleccionar cuantos clústeres hay. Esto no siempre es trivial e idealmente con el algoritmo de clustering que queremos que los resolviera por nosotros porque el objetivo es obtener información de los datos.</p> <p>-> Inicia con una selección aleatoria de centros de conglomerados o clústeres y, por lo tanto, puede producir distintos resultados de clústeres en diferentes ejecuciones del algoritmo.</p>	Las únicas limitaciones que se presentan con el problema de la implementación del algoritmo de Kruskal es la creación de un algoritmo adicional que nos compruebe que al adicionar una arista al grafo no nos haga un ciclo (Sanchez, 2017).	Si la base de datos contiene puntos de datos que forman clústeres de diferentes densidades, DBSCAN no puede agrupar bien los puntos de datos, porque el agrupamiento depende del parámetro ϵ y no se pueden seleccionar los puntos mínimos por separado para todos los grupos.
------------	--	--	---

A continuación, se expresa la información de la presente tabla:

Comparación de Ventajas y Desventajas del Método Particional tabla 2, muestran que los algoritmos mencionados en la presente tabla son buenos trabajando dependiendo del tipo de problema o problemática a resolver de acuerdo a su funcionalidad del algoritmo.

Jerárquico

Tabla 2 Comparación de Ventajas y Desventajas del Método Particional

El agrupamiento jerárquico permite que cada clúster tenga sub-clusters se obtiene

Fuente: Ángel Steven Choez

un clustering jerárquico. Incluye permitir que los clusters o grupos puedan anidarse, organizado en forma de árbol. Cada nodo del árbol, un clúster en este caso a excepción de los nodos de hoja, forman la unión de sus hijos los sub-clusters. La raíz del árbol es el clúster que contiene todos los datos. Los nodos de hoja generalmente corresponden a una sola pieza de dato, pero esto no es obligatorio (Alcalde, 2018).

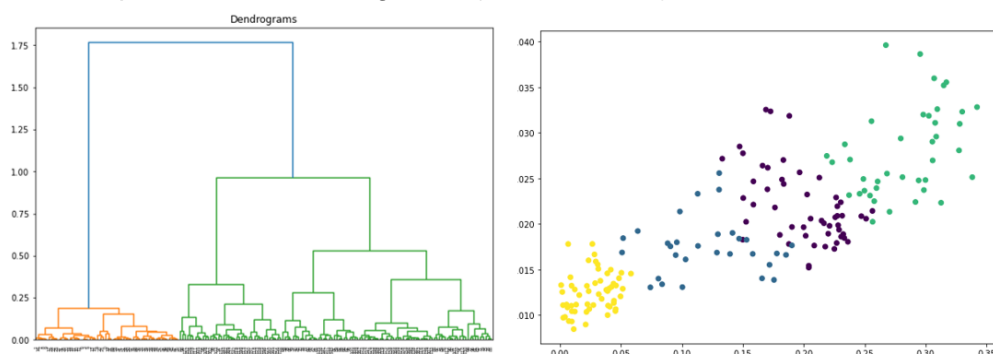


Figura 5 Agrupamiento Jerárquico
Fuente: Angel Steven Choez

Enlace simple: El enlace simple es la distancia entre dos grupos o la distancia más corta entre dos puntos de cada grupo. Este enlace se puede usar para detectar valores altos en su conjunto de datos, posiblemente valores atípicos, ya que se fusionarán al final. Según Alcalde (2018), “la cercanía entre dos grupos o clusters se da como la distancia entre los dos puntos más cercanos de cada clusters”.

Enlace completo: El enlace completo se calcula la distancia máxima entre los grupos o clústeres antes de la fusión, es decir, la distancia de los elementos más lejanos. Según Alcalde (2018), concluye que: “utiliza la distancia de los dos puntos más lejanos en cada cluster”.

Enlace ponderado: Usa las distancias pares a pares de todos los puntos en cada cluster, es decir que utiliza la distancia media entre los pares de clústeres.

Método de Ward: Mide la proximidad entre dos clusters usando el incremento del error cuadrático medio producido al unir dos clusters (Alcalde, 2018).

Jerárquico				
	Enlace Simple	Enlace ponderado	Enlace completo	Método Ward
Característica	Se identifica por la distancia mínima entre grupos	Se identifica por la distancia entre centros de grupos	Se identifica por la distancia máxima entre grupos	Se caracteriza por calcular la expresión anterior cada par de grupo

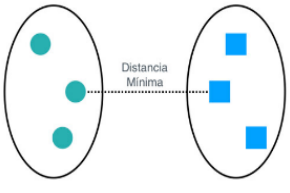
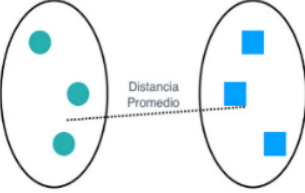
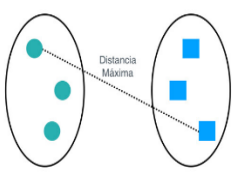
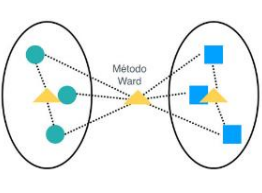
Representación grafica	 <p>Figura 6 Enlace Simple</p> <p>Fuente: https://aprendeia.com/algorithm-agrupamiento-jerarquico-teoria/</p>	 <p>Figura 7 Enlace promedio</p> <p>Fuente: https://aprendeia.com/algorithm-agrupamiento-jerarquico-teoria/</p>	 <p>Figura 8 Enlace Completo</p> <p>Fuente: https://aprendeia.com/algorithm-agrupamiento-jerarquico-teoria/</p>	 <p>Figura 9 Método Ward</p> <p>Fuente: https://aprendeia.com/algorithm-agrupamiento-jerarquico-teoria/</p>
------------------------	---	---	--	---

Tabla 3 Comparación de las Características del Agrupamiento Jerárquico

Fuente: Angel Steven Choez

A continuación, se expresa la información de la presente tabla:

Comparación de las Características del Agrupamiento Jerárquico tabla 3, muestra que estos algoritmos realizan un proceso de agrupación jerárquica que permite disminuir la dimensión del problema desde el punto de vista descriptivo.

Probabilístico

Esta técnica utiliza la distribución de probabilidad para crear los clústeres. El algoritmo utilizado en esta técnica se puede mencionar: Modelo de agrupamiento Gaussiana.

Modelo de agrupamiento gaussiano

El Modelo de agrupamiento Gaussiana es un modelo probabilístico que supone que todas las muestras se generan a partir de una combinación de un número finito de distribuciones gaussianas, con parámetros desconocidos. Pertenece al grupo de algoritmos de agrupamiento blando donde cada punto de datos pertenecerá a cada grupo existente en el conjunto de datos, pero con diferentes niveles de pertenencia a cada grupo (Roman, 2019).

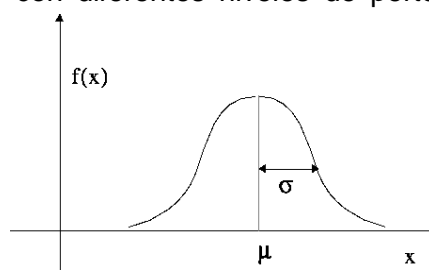


Figura 10 *Algoritmo Gaussiano*

Fuente: <https://lamfo-unb.github.io/2017/05/11/Aprendizado-Semi-Supervisionado-para-Deteccion-de-Fraudes-Parte-2/>

Ventajas

- Son mucho más flexibles en términos de varianza de clústeres que los K-Means, porque debido al parámetro de desviación estándar, los clústeres pueden tener cualquier forma de elipse, en lugar de estar limitados a círculos.
- Los modelos de agrupamiento gaussianos usan probabilidad, pueden tener múltiples grupos por punto de datos.

Desventajas

- Sin suficientes puntos para cada mezcla, el algoritmo diverge y encuentra soluciones con una probabilidad infinitas, a menos que ajustemos artificialmente las covarianzas entre los puntos de datos.

Característica

- Es un modelo probabilístico que asume todas las mediciones provienen de una distribución N-dimensional de Gaussianas con parámetro desconocidos (Muñoz, 2019).

Funcionamiento del algoritmo Modelo de agrupamiento gaussiano

- **Paso 1:** Iniciamos eligiendo el número de cluster, como se hace en K-Means, e inicializamos aleatoriamente los parámetros de distribución gaussiana para cada cluster.
- **Paso 2:** Dadas estas reparticiones gaussianas para cada cluster, se calcula la probabilidad de que cada punto de datos pertenezca a un cluster en particular. Cuando más cerca esté un dato del centro de los gaussianos, más probable es que esté en este grupo.
- **Paso 3:** Calculamos un nuevo parámetro para las distribuciones gaussianas de manera que maximicemos las probabilidades de los puntos dentro de los clústeres.
- **Paso 4:** Se repiten muchas veces los pasos 2 y 3 hasta la convergencia, donde las distribuciones no cambian mucho de una iteración en iteración.

Propuesta de la solución

Al final se cumplió el objetivo de analizar las características de los tipos de algoritmos de clustering con el fin de demostrar el funcionamiento de los algoritmos, tales como el algoritmo K-Media perteneciente al agrupamiento por particiones se basa en agrupar y encontrar los puntos de datos que tienen una similitud entre ellos, es decir que cada grupo está representado por su centro, además se caracteriza porque la asignación de los clusters es más robusta.

Con respecto al algoritmo Gaussiana, perteneciente al agrupamiento probabilístico, se basa en la probabilidad que supone que todas las muestras se generan a partir de una combinación de un número finito de distribuciones gaussianas, con parámetros desconocidos, cabe mencionar que el algoritmo de gaussiana tiene funciones similares al del algoritmo K-Media cuya diferencia que el algoritmo gaussiano al momento de agrupar los datos, si un punto de dato se encuentra muy lejos su agrupación no tiene peso, mientras

que el algoritmo K-Media al momento de agrupar los datos todos los puntos de datos del mismo centro tienen el mismo peso.

Con respecto al algoritmo método de Ward, perteneciente al agrupamiento jerárquico, se basa que a partir de un dendrograma con la técnica de la observación se determina el número de clusters ya que este algoritmo calcula la suma de las distancias cuadradas en los clústeres y los fusiona para minimizarlas, la única dificultad que presenta el agrupamiento jerárquico es que si trabaja con grandes volúmenes de datos es muy difícil interpretar el gráfico.

Como lo anterior mencionado de los tipos de algoritmos de clustering en el aprendizaje no supervisado, es importante mencionar que los algoritmos más utilizados son los K-Media, DBSCAN, Gaussiana, método de Ward, todos estos algoritmos son buenos resolviendo problemas en clasificar los datos solo hay que tener en cuenta que cada algoritmo trabaja de acuerdo al planteamiento del problema o la problemática a resolver, esto se debe porque cada uno de los algoritmos identifican diferentes patrones dentro de un grupo de datos.

Conclusión

Al finalizar el estudio de caso análisis de las características de los tipos de algoritmos de clustering en el aprendizaje no supervisado se concluye:

Los tipos de algoritmos de clustering tiene como finalidad encontrar patrones con características similares dentro de un conjunto de datos y mediante un proceso no supervisado lo particionan en cierto número de clusters.

Además, se estudió los de agrupamiento por particiones que generan varias particiones y luego las evalúan según algún criterio y el agrupamiento jerárquico produce una jerarquía de grupos llamados dendogramas que consiste en permitir que los clusters puedan anidar, y mostrar en forma de árbol.

También se estudió el algoritmo gaussiano perteneciente al agrupamiento probabilístico, se basa en las probabilidades para formar los clusters cuya dificultad que si los puntos de datos están muy dispersos encuentra un conjunto de probabilidades infinitas es decir con soluciones irregulares.

Bibliografía

Alcalde, A. (3 de Abril de 2018). *El Baul del Programador*. Obtenido de El Baul del Programador: <https://elbauldelprogramador.com/aprendizaje-nosupervisado-clustering/>

Alvaro, J. (29 de Diciembre de 2017). Organización de datos multidimensionales en un sistema de recomendaciones basado en data clustering e inteligencia de enjambres.

Barragan, S. (Julio de 2018). *Modelacion con teoria de grafos para la unidimensionalidad de un instrumento de evaluacion*. Obtenido de scielo: http://www.scielo.org.ar/scielo.php?script=sci_arttext&pid=S1668-70272018000100001&lang=es

Benavides, F. (Mayo de 2017). *Expansión óptima del sistema de transmisión mediante el algoritmo de PRIM*. Obtenido de (Bachelor's thesis, Universidad Politécnica Salesiana. Carrera de Ingeniería Eléctrica. Sede Quito): <https://dspace.ups.edu.ec/bitstream/123456789/14358/4/UPS-KT01401.pdf>

Capretto, Tomas, Mari, & Gonzalo. (2018). Método de agrupamiento geoespacial para la segmentación de una población de viviendas. *Instituto de Investigaciones Teóricas y Aplicadas, Escuela de Estadística*. Obtenido de rehip.unr.edu.ar.

Ceron, J. (5 de Diciembre de 2018). *Análisis de algoritmos de clustering para datos categóricos*. Obtenido de *Análisis de algoritmos de clustering para datos categóricos*: <https://repositorio.uniandes.edu.co/bitstream/handle/1992/39125/u820979.pdf?sequence=1>

Concatto, F., Raimundo, C., & Rese, A. (2017). Análise de Algoritmos da Árvore Geradora Mínima para o Problema de Reconfiguración de Redes de Distribuição. *Revista de Informática Aplicada*. Obtenido de https://seer.uscs.edu.br/index.php/revista_informatica_aplicada/article/view/6922/3013

Estrada Mancilla, E. (13 de Febrero de 2017). *algoritmo Kruskal*. Obtenido de algoritmo Kruskal: <https://www.slideshare.net/erickestradamancilla/algoritmo-de-kruskal-72069816>

- Facure, M. (11 de Mayo de 2017). *Aprendizaje semisupervisado para la detección de fraudes [Parte 2]*. Obtenido de Lamfo: <https://lamfo-unb.github.io/2017/05/11/Aprendizado-Semi-Supervisionado-para-Deteccion-de-Fraudes-Parte-2/>
- Ligdi, G. (08 de Septiembre de 2020). *Algoritmo Agrupamiento Jerárquico – Teoría*. Obtenido de aprendeia: <https://aprendeia.com/algoritmo-agrupamiento-jerarquico-teoria/>
- Llaque, V., & Jennifer, Y. (2018). *Analisis de trayectorias vehiculares GPS para evaluar su calidad de agrupamiento utilizando algoritmos clustering de minería de datos*. Obtenido de (Doctoral dissertation, Universidad de Guayaquil. Facultad de Ciencias Matemáticas y Físicas. Carrera de Ingeniería En Sistemas Computacionales): <http://repositorio.ug.edu.ec/handle/redug/32669>
- Martínez, L. (2018). *Aplicaciones en la industria del problema de Steiner y su resolución mediante algoritmos genéticos*. Obtenido de idus.us.es: <https://idus.us.es/bitstream/handle/11441/82990/TFG-2106-MARTINEZ.pdf?sequence=1&isAllowed=y>
- Moya, R. (25 de Marzo de 2016). *jarroba.com*. Obtenido de jarroba.com: <https://jarroba.com/que-es-el-clustering/#jumpToEntryData>
- Muñoz, U. A. (2019). *Deteccion temprana de desviaciones del comportamiento nominal de sistemas utilizando algoritmos de Machine learnin (Doctoral dissertation Universidad Nacional de Cuyo)*. Obtenido de Deteccion temprana de desviaciones del comportamiento nominal de sistemas utilizando algoritmos de Machine learnin (Doctoral dissertation Universidad Nacional de Cuyo): http://ricabib.cab.cnea.gov.ar/843/1/Mu%C3%B1oz_U.pdf
- Roman, V. (12 de Junio de 2019). *Aprendizaje no supervisado en machine learning: agrupacion*. Obtenido de Aprendizaje no supervisado en machine learning:

agrupacion: <https://medium.com/datos-y-ciencia/aprendizaje-no-supervisado-en-machine-learning-agrupaci%C3%B3n-bb8f25813edc>

Salazar, M. (30 de Enero de 2017). *Modelamiento de confiabilidad y análisis para flotas: Un enfoque basado en clustering para manejo de datos no homogéneos*. Obtenido de Repositorio Academico de la Universidad de Chile: <https://repositorio.uchile.cl/bitstream/handle/2250/144639/Modelamiento-de-confiabilidad-y-an%C3%A1lisis-para-flotas-Un-enfoque-basado-en-clustering-para-manejo-de-datos-no-homog%C3%A9neos.pdf?sequence=1&isAllowed=y>

Sanchez, B. (19 de Junio de 2017). *Algoritmo Kruscal*. Obtenido de Algoritmo Kruscal: https://prezi.com/nrhlsgh2w_py/algoritmo-kruscal/

Sancho Caparrini, F. (20 de Diciembre de 2020). *Algoritmos de Clustering*. Obtenido de Algoritmos de Clustering: Algoritmos de Clustering: <https://www.cs.us.es/~fsancho/?e=230>

Vallalta Rueda, F. (04 de Agosto de 2019). *Aprendizaje supervisado y no supervisado*. Obtenido de healthdataminer: <https://healthdataminer.com/data-mining/aprendizaje-supervisado-y-no-supervisado/>

Vintimilla, C., Astudillo, F., Severeyn, E., & Encalada, L. (2017). Agrupamiento de K-medias para estimación de insulino-resistencia en adultos mayores de Cuenca. *Maskana*.

Zhang, R., Du, T., Qu, S., & Sun, H. (20 de Enero de 2018). *Algoritmo de agrupación en clústeres basado en la densidad adaptativa con juego de conflicto KNN compartido*. Obtenido de Ciencias de la Información: <https://www.sciencedirect.com/science/article/abs/pii/S0020025521001596>

Anexo



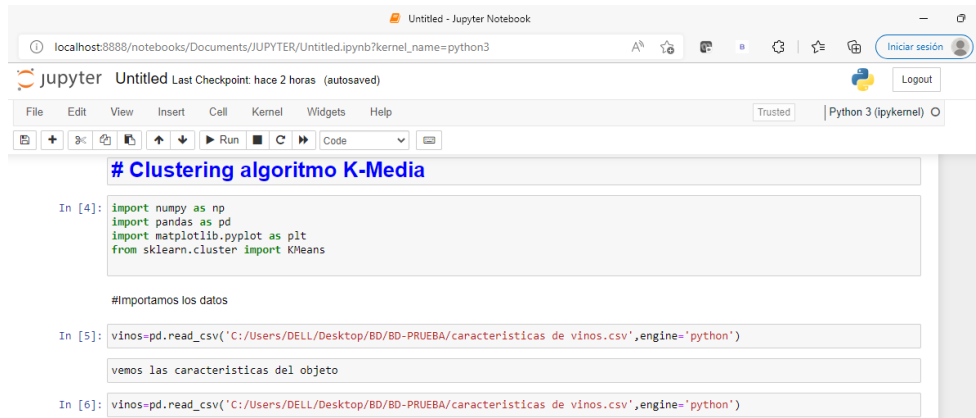
Document Information

Analyzed document	Angel_Choez.docx (D131744206)
Submitted	2022-03-28T03:52:00.0000000
Submitted by	Jorge Mejia
Submitter email	jmejia@utb.edu.ec
Similarity	4%
Analysis address	jmejia.utb@analysis.orkund.com

Sources included in the report

W	URL: https://elbauldelprogramador.com/aprendizaje-nosupervisado-clustering/ Fetched: 2021-04-10T06:43:49.5200000	 10
W	URL: https://jarroba.com/que-es-el-clustering/#jumpToEntryDataMu Fetched: 2022-03-28T03:52:00.0000000	 1

Para demostrar el análisis de las características de los tipos de algoritmos clustering en el aprendizaje no supervisado se utilizó el método de agrupamiento particional utilizando el **algoritmo K-medias** y para el manejo de los datos se usó una base de datos llamada “características de vinos”, la cual vamos agrupar o clasificar los datos.



```
# Clustering algoritmo K-Media

In [4]: import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans

#Importamos los datos

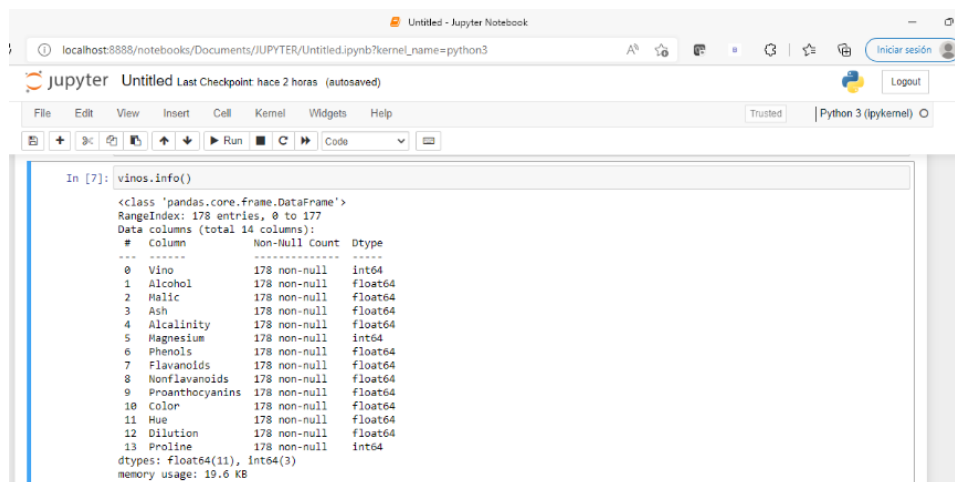
In [5]: vinos=pd.read_csv('C:/Users/DELL/Desktop/BD/BD-PRUEBA/características de vinos.csv',engine='python')

veamos las características del objeto

In [6]: vinos=pd.read_csv('C:/Users/DELL/Desktop/BD/BD-PRUEBA/características de vinos.csv',engine='python')
```

Figura 11 *Importación de Librerías*
Fuente: Angel Steven Choez

Importación de Librerías figura 11, aquí en esta parte importamos las librerías necesarias para analizar los datos, y a su vez importamos el origen de dato con su respectiva ruta de archivo.



```
In [7]: vinos.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 178 entries, 0 to 177
Data columns (total 14 columns):
#   Column             Non-Null Count  Dtype
---  -
0   Vino                178 non-null   int64
1   Alcohol             178 non-null   float64
2   Malic               178 non-null   float64
3   Ash                 178 non-null   float64
4   Alkalinity          178 non-null   float64
5   Magnesium           178 non-null   int64
6   Phenols             178 non-null   float64
7   Flavonoids          178 non-null   float64
8   Nonflavonoids       178 non-null   float64
9   Proanthocyanins     178 non-null   float64
10  Color               178 non-null   float64
11  Hue                 178 non-null   float64
12  Dilution            178 non-null   float64
13  Proline              178 non-null   int64
dtypes: float64(11), int64(3)
memory usage: 19.6 KB
```

Figura 12 *Información Origen de Datos*
Fuente: Angel Steven Choez

Información origen de datos figura 12, aquí vemos los datos que contiene la base de datos.

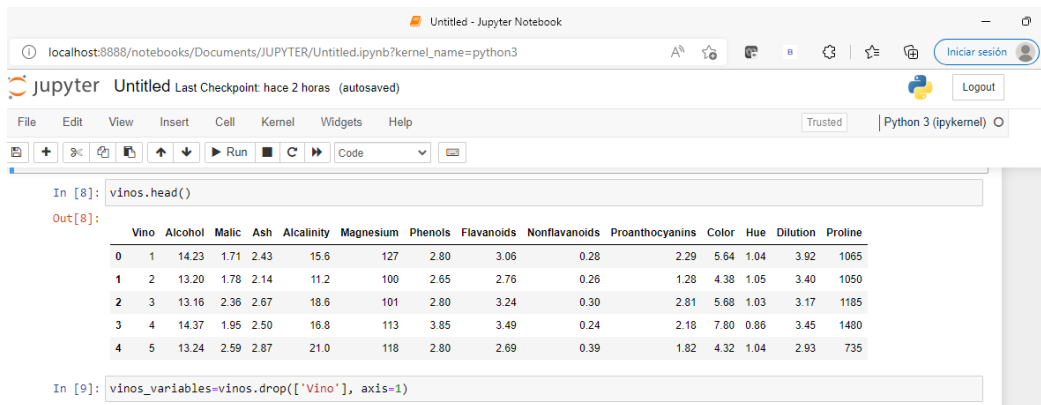


Figura 13 *Primeras*
Fuente: Angel Steven Choez

Primeras Filas figura 13, aquí en esta parte ya tenemos desplegada la primera fila, correspondiente a la primera columna de la figura 12.

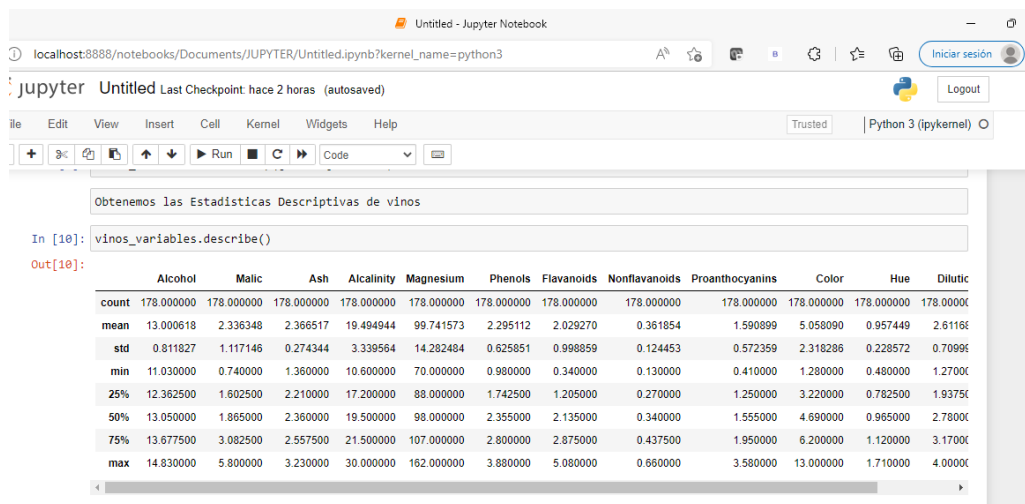


Figura 14 *Estadísticas de los*
Fuente: Angel Steven Choez

Estadísticas de los Datos figura 14, aquí en esta nos muestra los valores estadísticos tales como: la desviación estándar, el promedio, valores mínimos, valores máximos, y los cuartiles de cada columna.

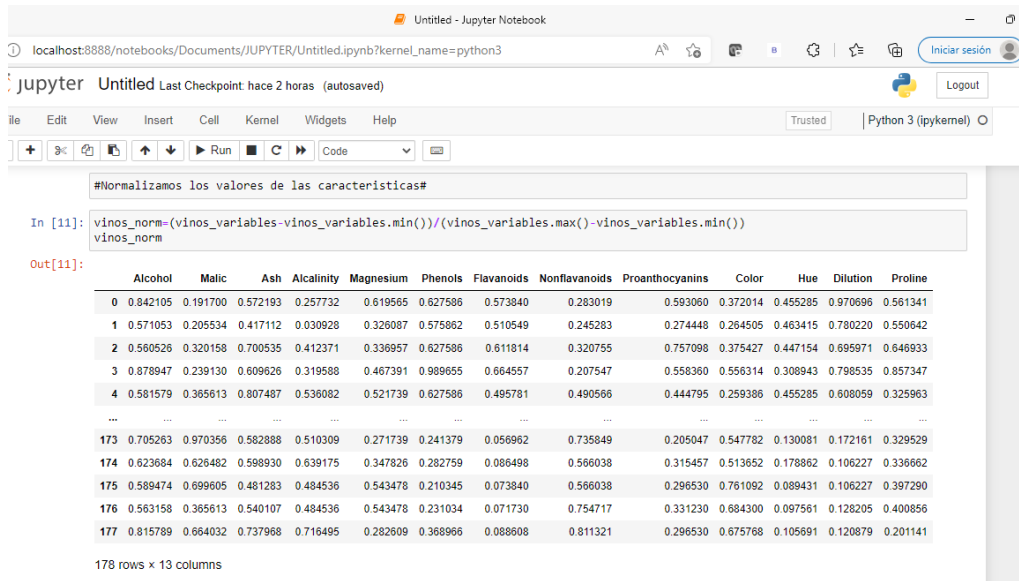


Figura 15 Normalizar
Fuente: Angel Steven Choez

Normalizar Valores figura 15, aquí en esta parte normalizamos los valores para después encontrarlo dentro del mismo rango.

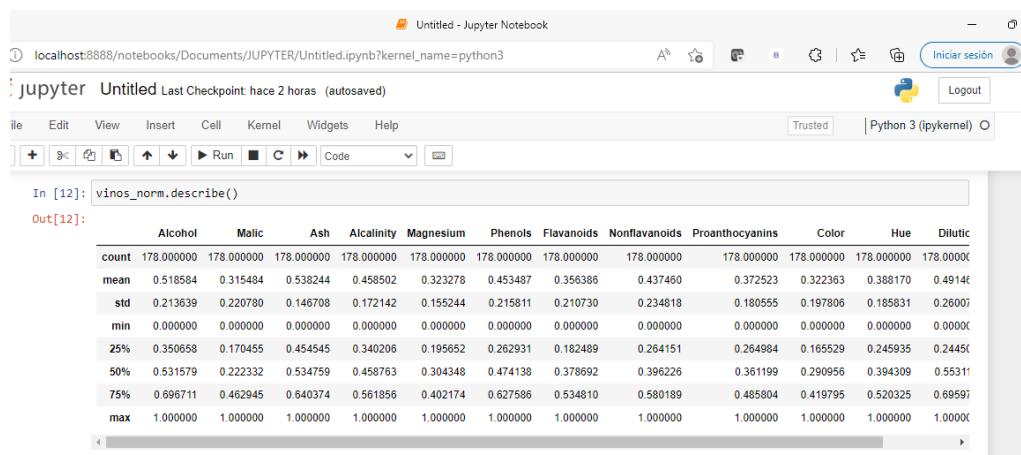


Figura 16 Valores Normalizados Descriptos
Fuente: Angel Steven Choez

Valores Normalizados Descriptos figura 16, aquí en esta parte nos muestra los valores estadísticos ya normalizados

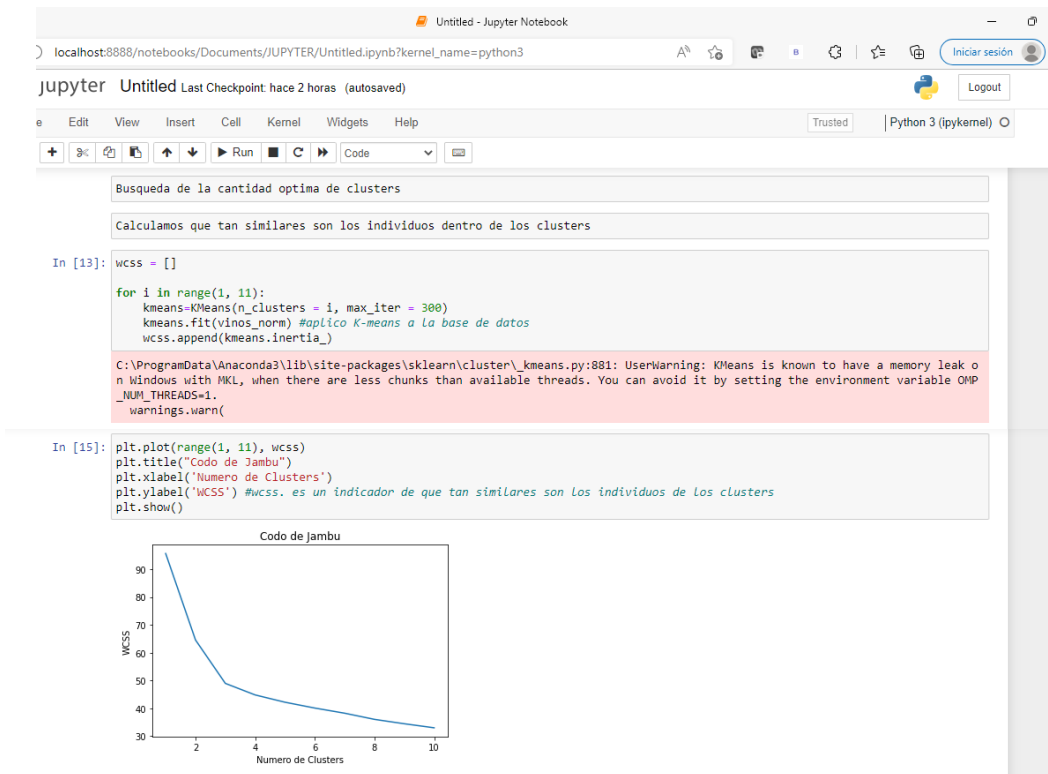


Figura 17 *Búsqueda de Clústers Optima*

Fuente: Angel Steven Choez

Búsqueda de Clústers Optima figura 17, aquí en esta parte aplicamos el método de búsqueda de clusters, pero antes de eso aplicamos un método llamado “Codo de Jambu” que es para crear una cantidad de clusters similares dentro de los mismo, luego de eso nos muestra la gráfica del Codo de Jambu donde elegimos en la gráfica el número 3 porque a partir de ese número deja de disminuir los clusters y es la cantidad optima de clusters a formar.

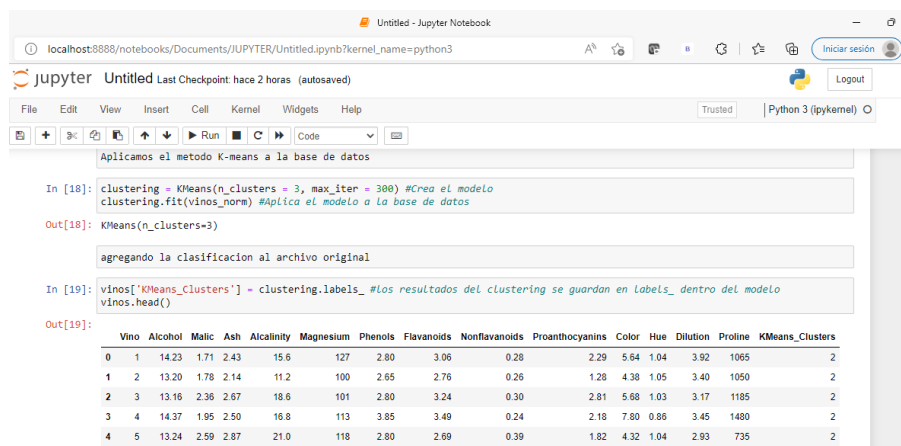
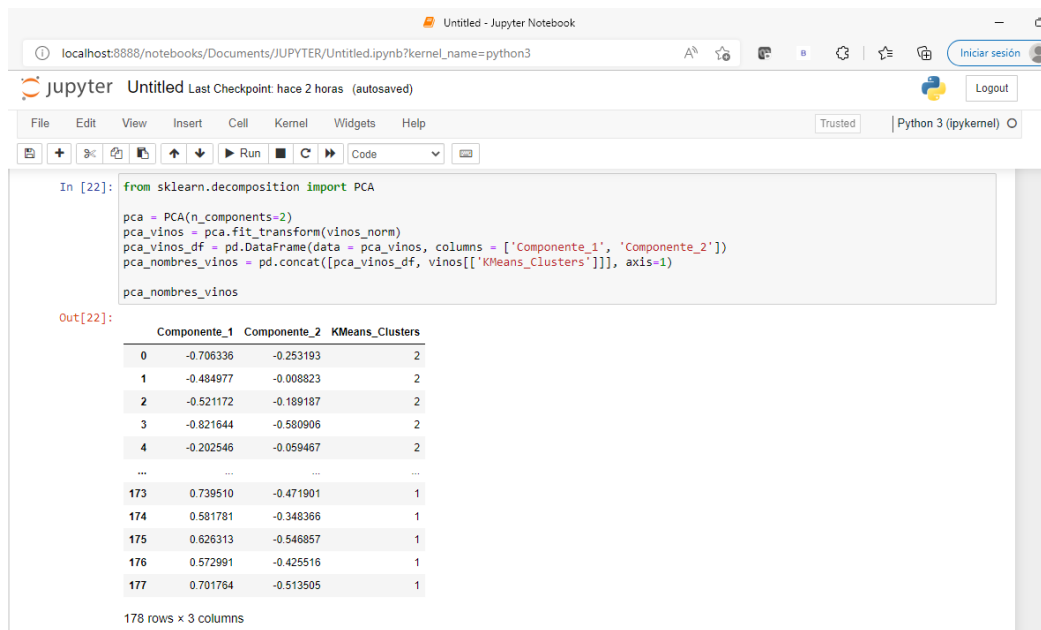


Figura 18 *Aplicamos el método K-Medias*

Fuente: Angel Steven Choez

Aplicamos el método K-Medias figura 18, aquí en esta parte aplicamos el método de K-medias a la base de datos con el objetivo de agregar la clasificación al archivo original.



```
In [22]: from sklearn.decomposition import PCA
pca = PCA(n_components=2)
pca_vinos = pca.fit_transform(vinos_norm)
pca_vinos_df = pd.DataFrame(data = pca_vinos, columns = ['Componente_1', 'Componente_2'])
pca_nombres_vinos = pd.concat([pca_vinos_df, vinos[['KMeans_Clusters']], axis=1)
pca_nombres_vinos
```

Out[22]:

	Componente_1	Componente_2	KMeans_Clusters
0	-0.706336	-0.253193	2
1	-0.484977	-0.008823	2
2	-0.521172	-0.189187	2
3	-0.821644	-0.580906	2
4	-0.202546	-0.059467	2
...
173	0.739510	-0.471901	1
174	0.581781	-0.348366	1
175	0.626313	-0.546857	1
176	0.572991	-0.425516	1
177	0.701764	-0.513505	1

178 rows x 3 columns

Figura 19 *Análisis de Componentes Principales*

Fuente: Angel Steven Choez

Análisis de Componentes Principales figura 19, aquí en esta nos muestra un cuadro como se formaron los clusters.

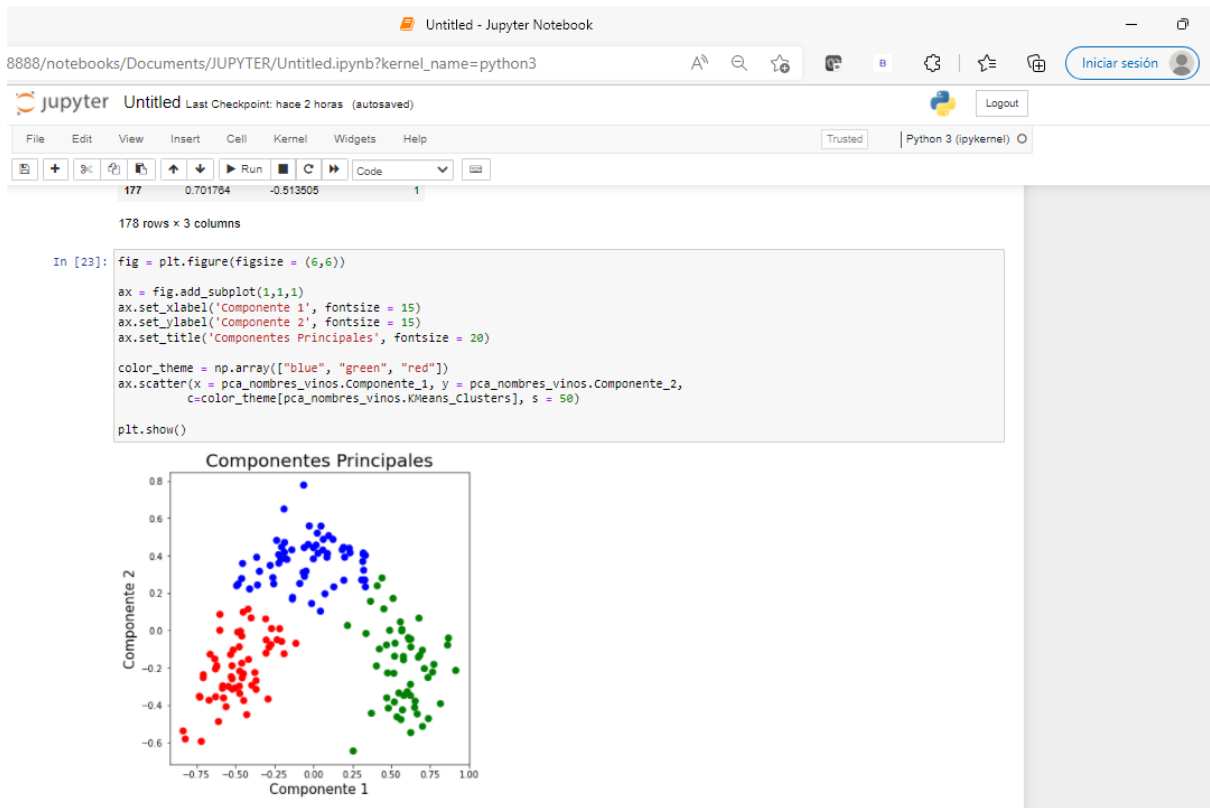


Figura 20 Visualización de los Clusters

Fuente: Angel Steven Choez

Visualización de los clusters figura 20, aquí en esta parte nos muestra la visualización grafica de los cluster formados con sus respectivos grupos, luego de la agrupación guardamos los datos nuevos datos ya clasificado.

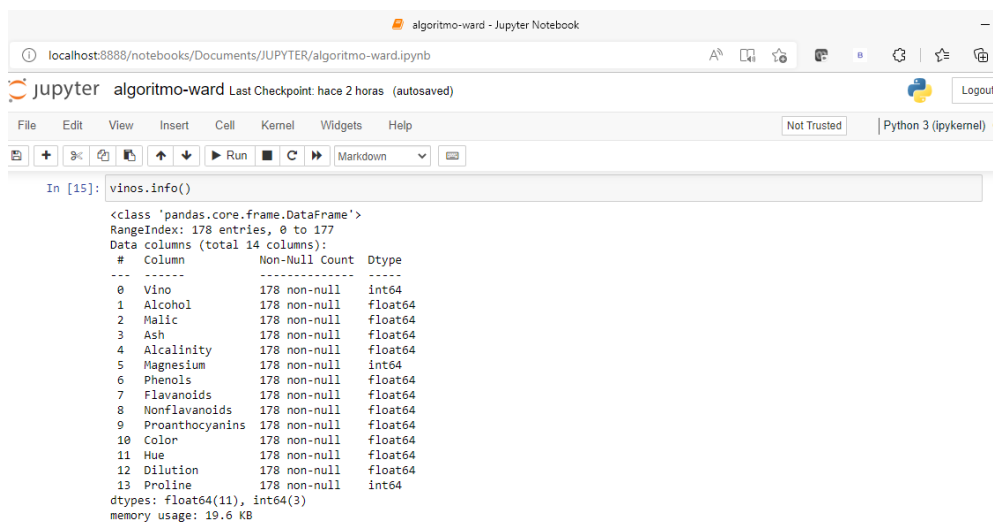
Para esta demostración en analizar de las características de los tipos de algoritmos clustering en el aprendizaje no supervisado se usó el método de agrupamiento jerárquico utilizando el **algoritmo método Ward** y para el manejo de los datos se usó la misma base de datos llamada “características de vinos”, la cual vamos agrupar o clasificar los datos.



```
algorithmo-ward - Jupyter Notebook
localhost:8888/notebooks/Documents/JUPYTER/algorithmo-ward.ipynb
jupyter algorithmo-ward Last Checkpoint: hace 2 horas (autosaved)
Python 3 (ipykernel)
File Edit View Insert Cell Kernel Widgets Help
+ - Undo Redo Copy Paste Run Stop Refresh Markdown
In [13]: import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
In [14]: vinos=pd.read_csv('C:/Users/DELL/Desktop/BD/BD-PRUEBA/características de vinos.csv',engine='python')
```

Figura 21 *Importación de Librerías*
Fuente: Angel Steven Choez

Importación de Librerías figura 21, aquí en esta parte importamos las librerías necesarias para analizar los datos, y a su vez importamos el origen de dato con su respectiva ruta de archivo.



```
algorithmo-ward - Jupyter Notebook
localhost:8888/notebooks/Documents/JUPYTER/algorithmo-ward.ipynb
jupyter algorithmo-ward Last Checkpoint: hace 2 horas (autosaved)
Python 3 (ipykernel)
File Edit View Insert Cell Kernel Widgets Help
+ - Undo Redo Copy Paste Run Stop Refresh Markdown
In [15]: vinos.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 178 entries, 0 to 177
Data columns (total 14 columns):
#   Column              Non-Null Count  Dtype
---  -
0   Vino                 178 non-null   int64
1   Alcohol             178 non-null   float64
2   Malic               178 non-null   float64
3   Ash                 178 non-null   float64
4   Alcalinity          178 non-null   float64
5   Magnesium           178 non-null   int64
6   Phenols             178 non-null   float64
7   Flavanoids          178 non-null   float64
8   Nonflavanoids       178 non-null   float64
9   Proanthocyanins     178 non-null   float64
10  Color               178 non-null   float64
11  Hue                 178 non-null   float64
12  Dilution            178 non-null   float64
13  Proline              178 non-null   int64
dtypes: float64(11), int64(3)
memory usage: 19.6 KB
```

Figura 22 *Información Origen de Datos*
Fuente: Angel Steven Choez

Información origen de datos figura 22, aquí vemos los datos que contiene la base de datos.

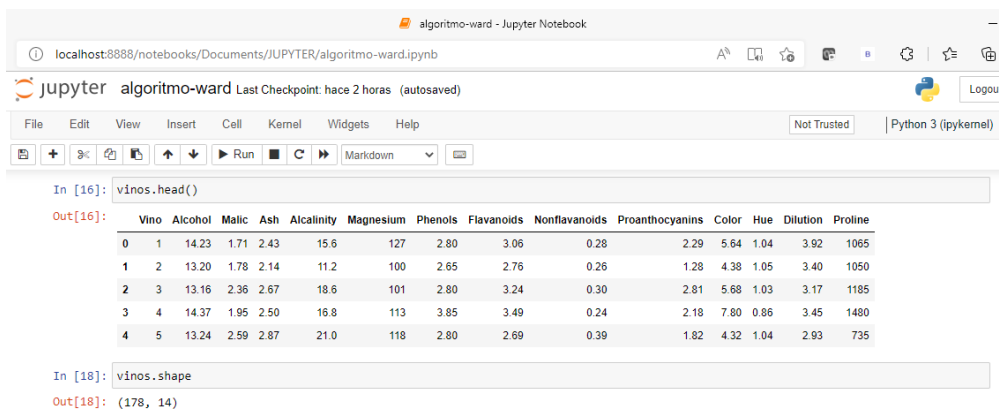


Figura 23 *Primeras*
Fuente: Angel Steven Choez

Primeras Filas figura 23, aquí en esta parte ya tenemos desplegada la primera fila, correspondiente a la primera columna de la figura 22.

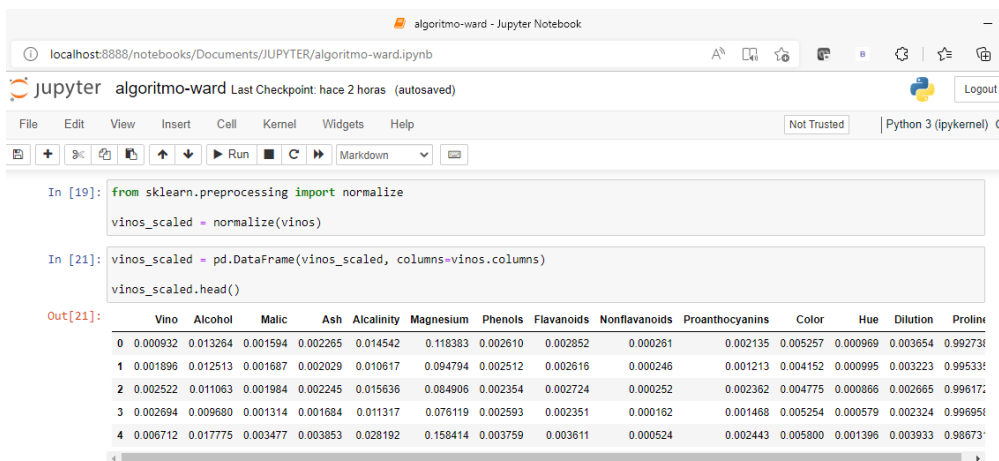


Figura 24 *Normalizar*
Fuente: Angel Steven Choez

Normalizar Valores figura 24, aquí en esta parte vemos valores numéricos es decir con valores normalizados.

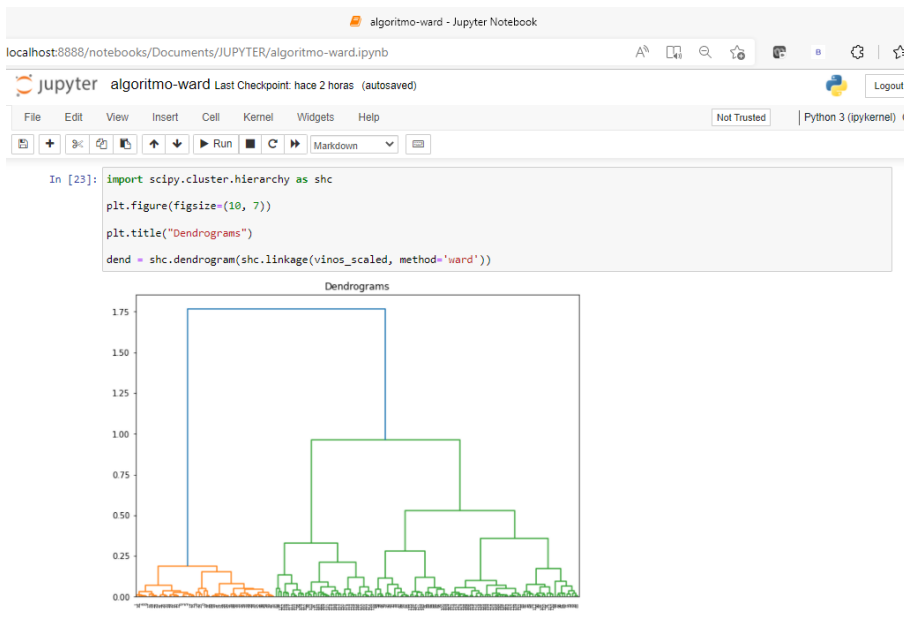


Figura 25 Dendrograma de los

Fuente: Angel Steven Choez

Dendrograma de los Datos figura 25, aquí en esta parte vemos una representación gráfica de dendrograma la cual nos permitirá calcular la distancia de cada uno de ello y con la técnica de la observación vemos como se forman los clusters.

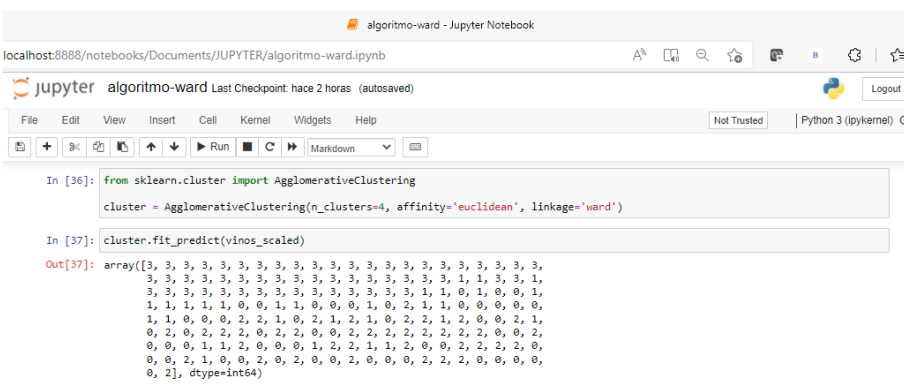


Figura 26 Grupos de Clusters

Fuente: Angel Steven Choez

Grupos de Clusters figura 26, aquí en esta parte nos muestra los grupos de clusters formados.

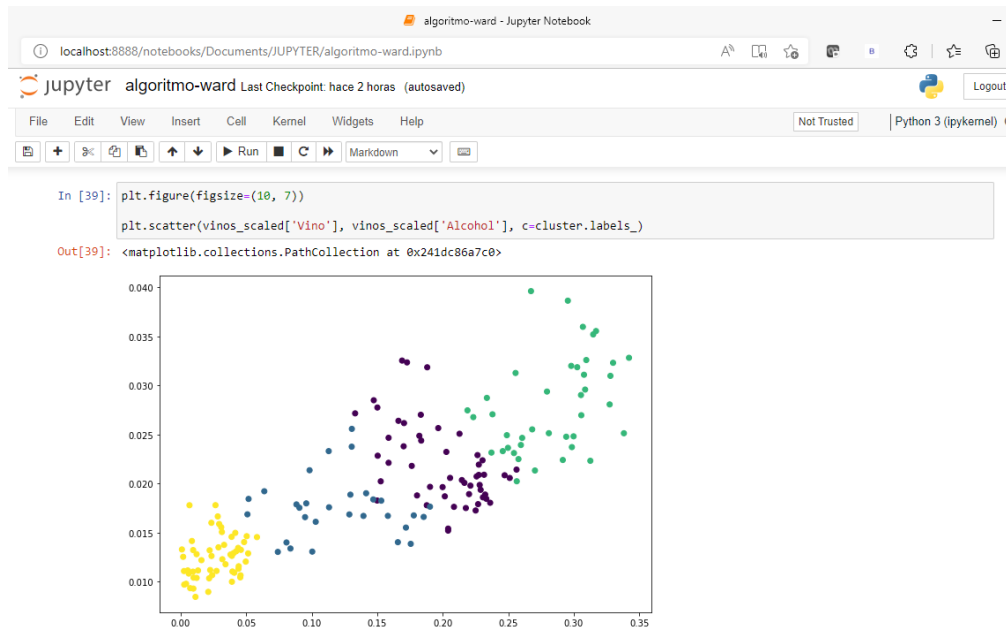


Figura 27 Visualización de los Clusters

Fuente: Angel Steven Choez

Visualización de los clusters figura 27, aquí en esta parte nos muestra la visualización grafica de los cluster formados con sus respectivos grupos, luego de la agrupación guardamos los datos nuevos datos ya clasificado.

Nota: con respecto a la gráfica final los puntos de datos parecen estar mezclados con otros grupos, pero no es así debido porque la gráfica está en 2D, se apreciaría mejor los puntos de datos si la gráfica fuera en 3D y se vería que los puntos están más distante.