TRABAJO OBLIGATORIO:

Uso y aplicación de los métodos de agrupamiento

Aprendizaje no supervisado

Alexandro López González
Leonardo Pacheco Garduño
Manuel Garcia Sanes

7 de Abril de 2019

Máster en
Inteligencia Artificial

DESCRIPCIÓN DEL PROBLEMA

En Machine Learning, el área de estudio de aprendizaje no supervisado se basa en el paradigma de descubrir en los datos relaciones ocultas que nos permitan describir los datos del mismo conjunto de entrenamiento. A diferencia del aprendizaje supervisado, los datos no se consideran completos en el sentido de que existe una salida que queremos encontrar.

La principal herramienta del aprendizaje no supervisado es la de realizar grupos o "clusters" en subconjuntos de características similares. Diferentes suposiciones sobre el origen de los datos llevan a las diferentes técnicas para descubrir diferentes agrupaciones en los datos. La finalidad de estas técnicas es básicamente descriptiva, aunque, una vez identificado un agrupamiento, es posible incorporar nuevos ejemplos de manera sencilla al subconjunto que corresponda.

El problema de clustering se define de la siguiente manera, para un set de datos D=(x1,x2,...,xn), y dadas las V características (X1,...,Xv) el objetivo es aprender la mejor división en subconjuntos homogéneos del set D basándose en las características V de los datos, de acuerdo a criterios establecidos.

En este trabajo se presentan diversas metodologías de agrupamiento aplicadas a dos sets de datos, el primero Seeds, en el cual el agrupamiento de los datos es conocido, y el set Online shoppers purchasing intention donde el agrupamiento real no es conocido.

Algoritmos

En ambos casos se intentarán resolver aplicando, 5 métodos diferentes:

- Particiones
- Jerárquico
- Espectral
- Densidad
- Mediante modelos probabilísticos

A continuación, los algoritmos elegidos serán utilizados para la resolución de los dos problemas seleccionados. Cada algoritmo y los parámetros asociados se ajustarán de manera apropiada (mediante una validación justa y razonable) para cada uno de los problemas por separado.

Evaluación

Se usarán las métricas de evaluación que se consideren oportunas para ello. Nótese que en el primer problema (el agrupamiento real de los datos es conocido) se podrán usar métricas de evaluación extrínseca, mientras que para el segundo dataset (el agrupamiento real es desconocido) se podrán usar solamente métricas de evaluación intrínseca.

Las medidas de evaluación extrínsecas son; Información Mutua, Valor de Error, Pureza del Agrupamiento, Precision, Recall, F1 y la Entropía.

Información mutua
Mide la dependencia mutua entre dos variables. En el contexto de aprendizaje no supervisado, mide la dependencia entre clusters. Entonces, es deseable cuando la Información mutua es más cercana a uno.

Valor del error
En cuanto al error cometido, un valor próximo al cero es deseable.

Pureza del agrupamiento
Es la medida en la cual se evalúa que un cluster únicamente tenga elementos de una clase. El valor más alto de pureza es 1.

Precisión
Se refiere a la cantidad de resultados que son relevantes.Tiene un rango de 0 a 1 y más próximo al uno es mejor.

Recall
Se refiere al porcentaje total de resultados relevantes correctamente clasificados por el algoritmo. Tiene un rango de 0 a 1 y más próximo al uno es mejor.

F1
Usa ambas métricas precisión y recall obteniendo la media armónica de ambas. Tiene un rango de 0 a 1 y más próximo al uno es mejor.

Entropía
Es la medida de la incertidumbre. La mejor entropía posible es de cero.

Las medidas de evaluación intrínsecas son; RMSSTD, R cuadrado, Calinski-Harabasz, Medida I y Davies-Bouldin

RMSSTD
Mide que tan homogéneos son los clusters.

R cuadrado
Es la relación de distancia entre los clusters con respecto a la distancia hipotética o ideal. Más cercana a cero es mejor

Calinski-Harabasz
Mide la bondad de un cluster basado en la suma promedio de distancias entre los grupos.

Medida I
Mide la separación entre los clusters con la distancia máxima entre los centros de los diferentes clusters.

Davies-Bouldin
Mide la distancia promedio entre los centroides

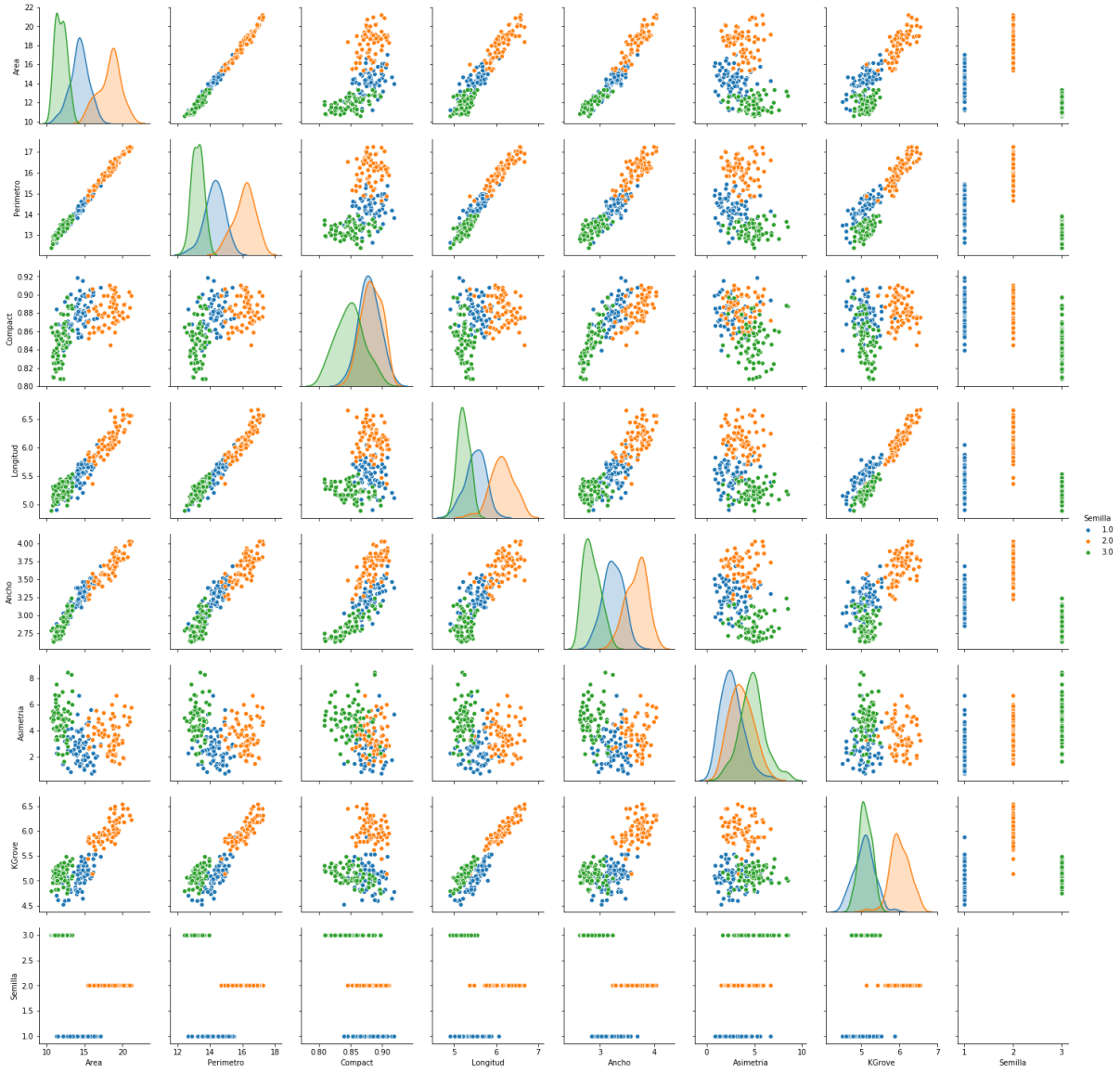
BASES DEL TRABAJO

Dataset 1:

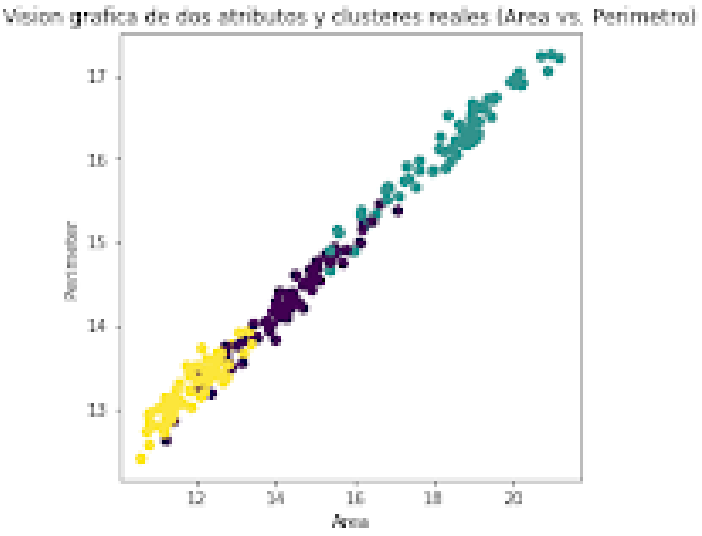
El primer grupo de datos proviene del repositorio de datos para el aprendizaje automático de la Universidad de California en Irvine (UCI). El set se denomina “Seeds” se puede encontrar en la página:
https://archive.ics.uci.edu/ml/datasets/seeds#
Los datos se refieren a las mediciones de propiedades geométricas de granos pertenecientes a tres variedades diferentes de trigo. El set es multivariable y tiene la principal finalidad de clasificación y clustering. Contiene 7 atributos y 210 instancias.

Haciendo el análisis de los datos, hemos elegido para la representación visual, dos atributos que muestran una separación clara de los agrupamientos. En todos los notebooks, la representación gráfica será de los atributos Area y Perimetro.

Los atributos son, área, perímetro, densidad, largo, ancho, asimetría y largo del surco central.



La ventaja de este set es que se puede notar de forma evidente los agrupamientos. La gráfica mostrada compara dos atributos, área y perímetro. Es fácil notar los clusters mostrados en colores.



Dataset 2:

Para el segundo dataset donde se esperan no tener conocimiento de los clusters se usará el mismo dataset “seeds”, pero se obvia la clase para cumplir con la especificación. Por lo cual se llevará a cabo la regla del codo y la métrica R cuadrado para estimar la cantidad de agrupaciones a encontrar.

Algoritmos

Los algoritmos escogidos representan cinco métodos de agrupamiento, por particiones, jerárquico, espectral, densidad y modelos probabilísticos. Específicamente son:

- K-means ++ (Particiones)
- Aglomerativo (Jerárquico)
- Matriz Laplaciana (Espectral)
- DBSCAN (Densidad)
- EM (Mediante modelos probabilísticos)

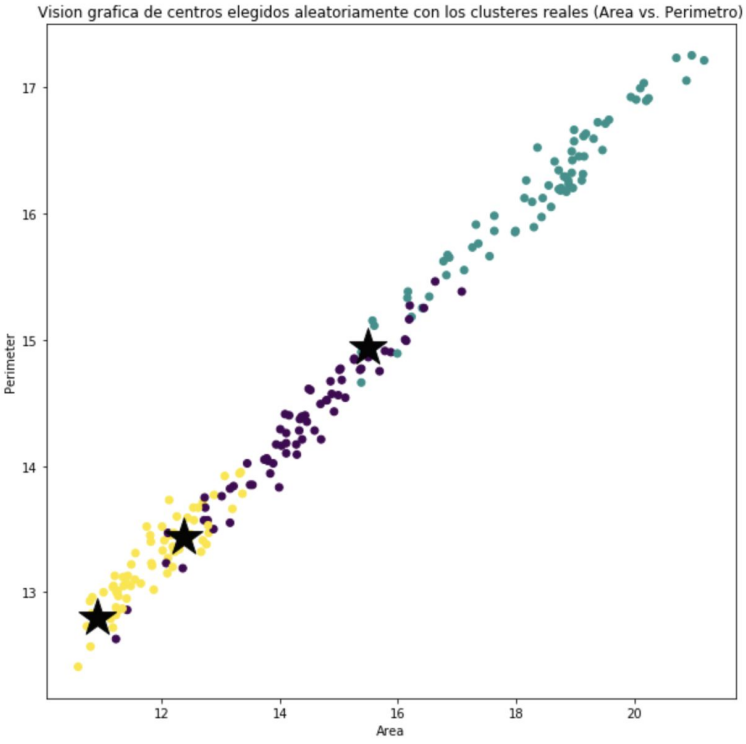
APLICACIÓN DE LOS MÉTODOS DE AGRUPAMIENTO

Agrupación por particiones: K-means++

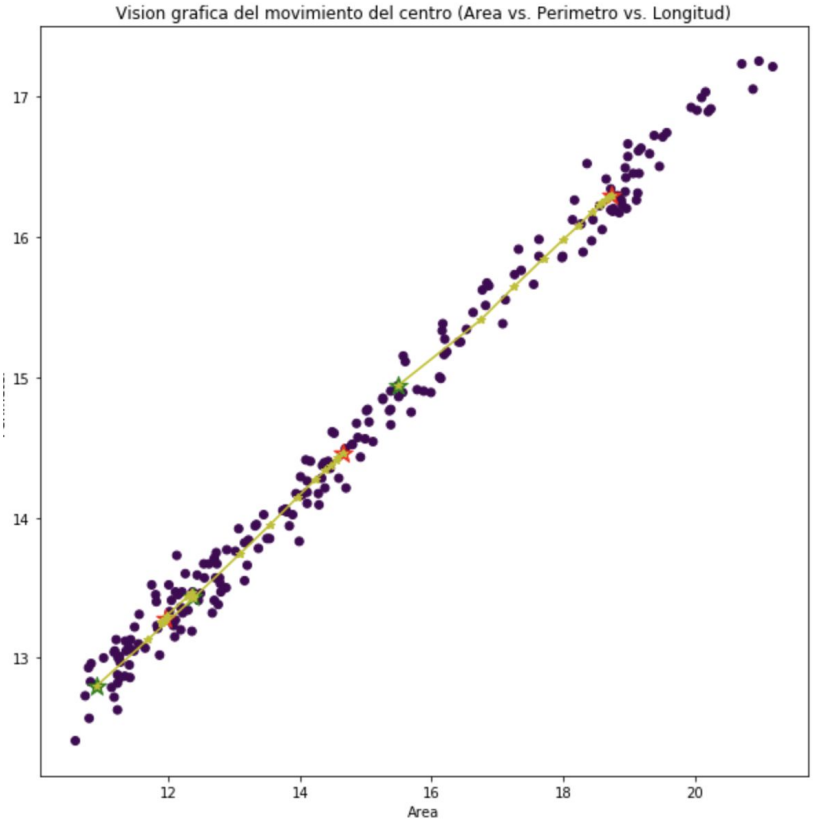
El algoritmo K-means es un método iterativo que inicia con K clústeres y revisa con cada iteración la asignación de los clústeres hasta que ninguno de los ejemplos cambia de clúster. Se eligió la mejora del algoritmo llamada K-means++ debido a que es más robusto en cuanto a la selección inicial de centros.

Dataset 1:

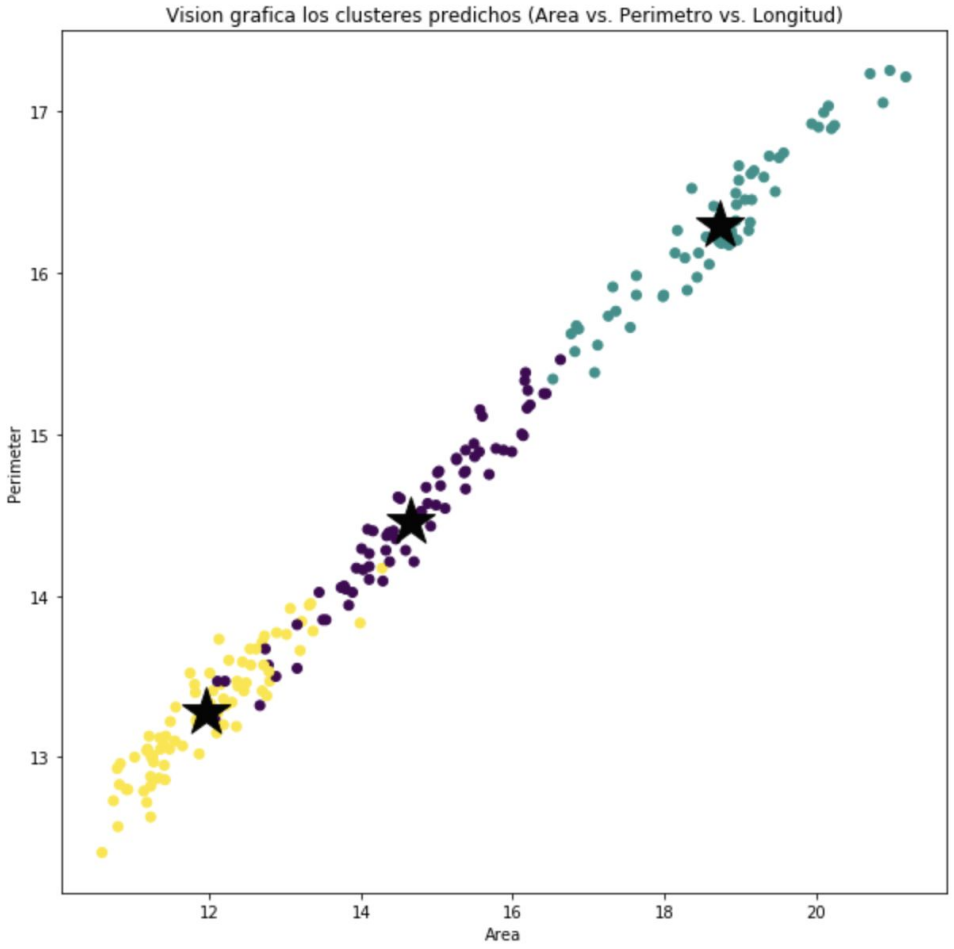
Para el primer dataset se usan los atributos área y perímetro del núcleo. Se eligió K=3 dado que conocemos las agrupaciones reales. En la siguiente figura se muestra la elección aleatoria de los centros:



Los centros se desplazan a lo largo de las iteraciones como se muestra a continuación:



La posición final de los centros y por lo tanto la separación en clústeres se queda de la siguiente forma:



Métricas de evaluación y análisis

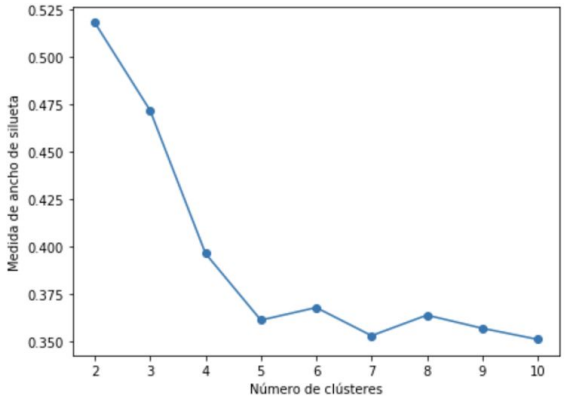
Se usaron medidas de evaluación extrínsecas, específicamente, Información Mutua, Valor de Error, Pureza del Agrupamiento, Precision, Recall, F1 y la Entropía. Los resultados fueron:

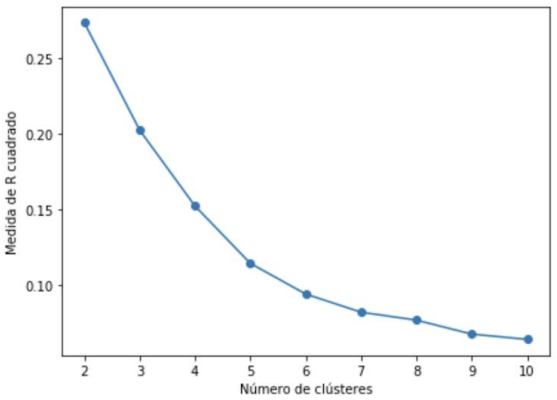
La información mutua es = 0.7785404720566337
El valor del error cometido es = 0.22857142857142854
La pureza del agrupamiento obtenido es = 0.8761904761904762
Precisión de l=1,k=1 es : 1.0
Recall de l=1,k=1 es : 0.6857142857142857
El valor F1 es = 0.8352152176563834
La entropía es = 0.3200718166114761

Conclusión

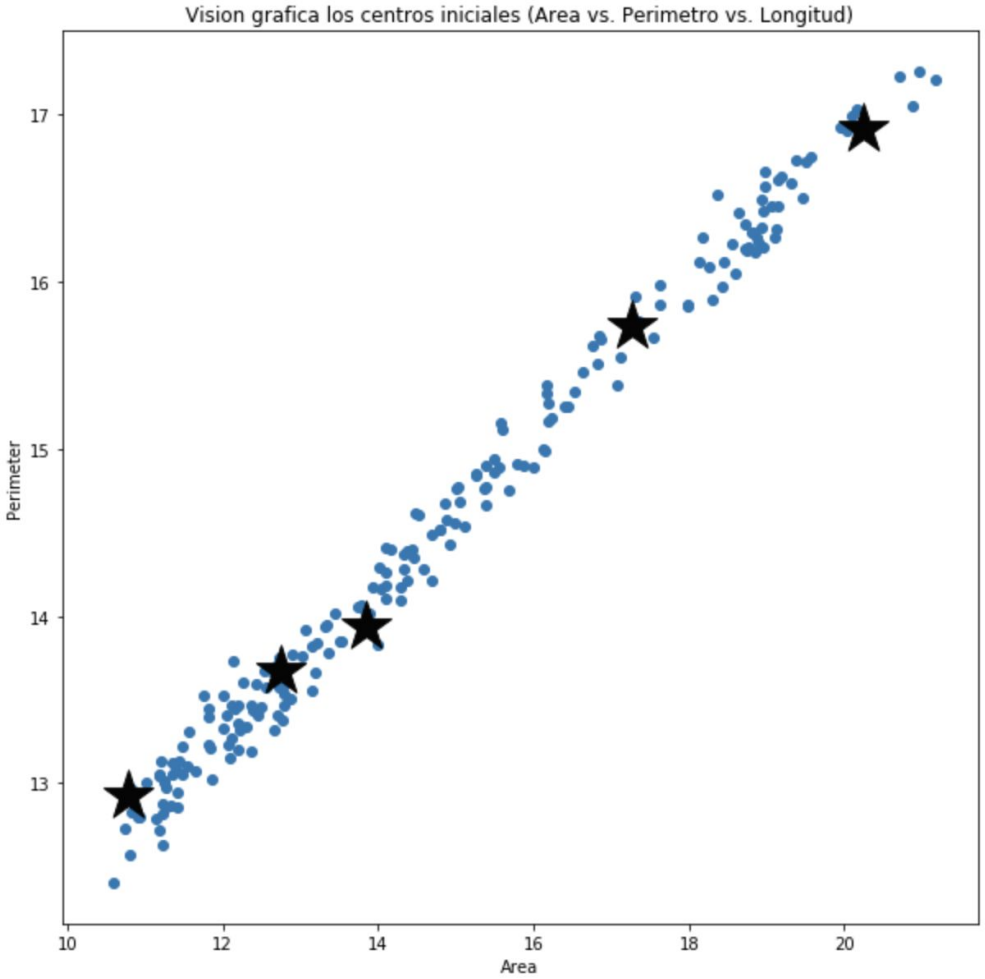
Evaluando las medidas de evaluación se puede notar una tendencia positiva en cuanto a la elección de los clústeres realizados por el algoritmo. El error y la entropía se aproximan al cero, mientras que las demás medidas se aproximan al uno. Aunque la pureza es alta, el error no es tan bajo como quisieramos.

Dataset 2:
Ahora se aplica la metodología del codo. La gráfica del ancho de silueta parece indicar un k=5, mientras que el R cuadrado no es muy convincente. Aunque sabemos que la agrupacion real es k=3, con el propósito académico continuamos con k=5.

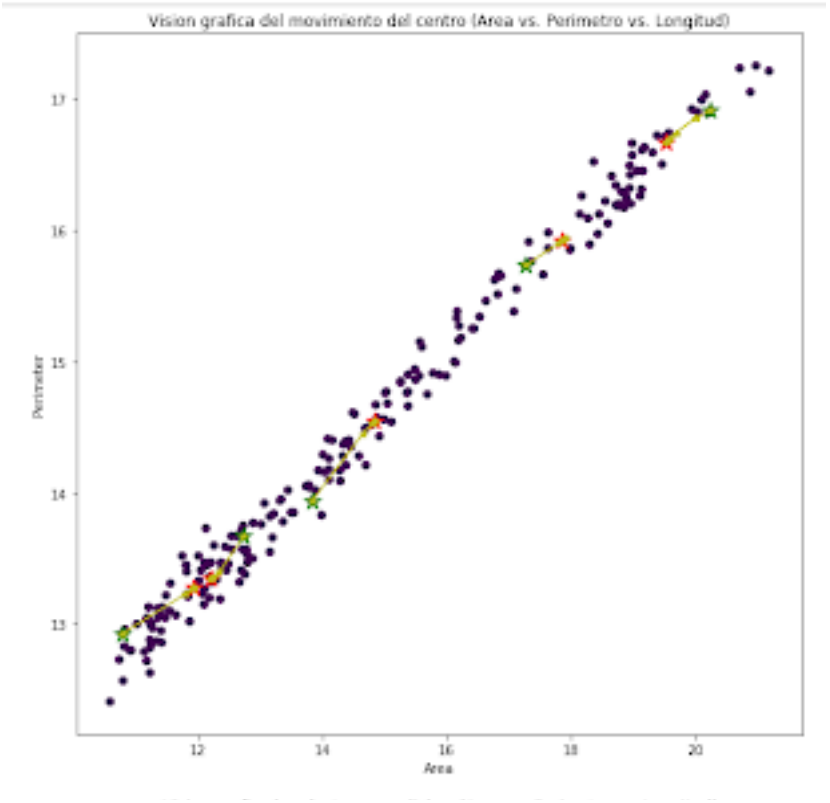




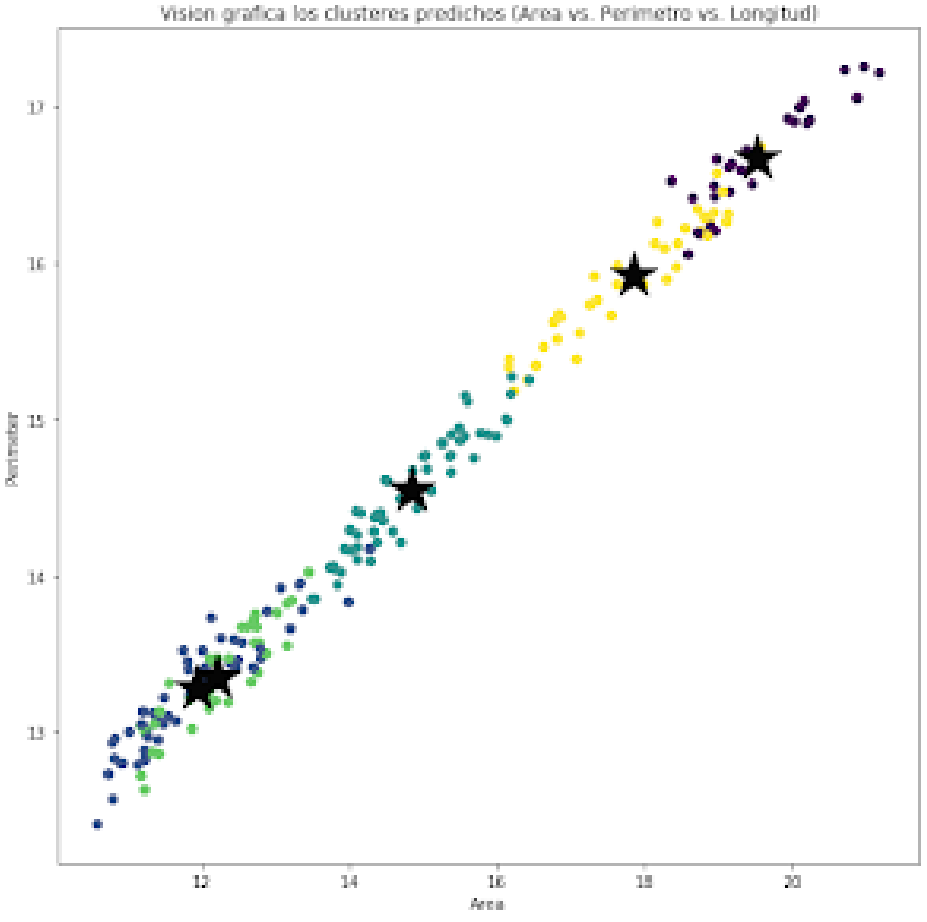
Se inicia el análisis de K-means++ con un K=5



Se puede notar la evolución del algoritmo en la siguiente gráfica



Mientras que las posiciones finales encontradas fueron las mostradas a continuación



Métricas de evaluación y análisis
Las métricas intrínsecas obtenidas son:

El error RMSSTD es = 0.663941977723
El valor de la medida R cuadrado es = 0.106138043266
El valor de la medida Calinski-Harabasz es = 283.772408468
El valor de la medida I es = 19.3009758684
El valor de la medida Davies-Bouldin es = 1.07432943241

Conclusión

Se puede notar por inspección visual que los resultados concuerdan con la evaluación anterior. Más aún, las medidas intrínsecas nos permiten identificar que la formación de los clusters es adecuada considerando índices adecuados para las cinco medidas.

Agrupación jerárquica: Aglomerativo

La agrupación jerárquica aglomerativa se basa en las medidas de disimilitud, iniciando con K=n clústeres y reduciendo el tamaño con cada iteración dependiendo de las características de disimilitud. Se eligió este algoritmo por curiosidad en cuanto a su desempeño con respecto a los datasets seleccionados.

Dataset 1:

Métricas de evaluación y análisis:

- Matriz Confusión
- Información Mutua
- Valor de Error
- Pureza del Agrupamiento
- F1
- Entropía

Con el siguiente resultado:

Matriz de Confusión con K=3

[[66 3 1]
 [6 0 64]
 [9 61 0]]

La información mutua con K=3 es = 0.783452191021
El valor del error cometido con K=3 es = 0.0904761904762
La pureza del agrupamiento con K=3 obtenido es = 0.909523809524
El valor F1 con K=3 es = 0.910922698258
La entropía con K=3 es = 0.315160097647

Conclusión:
El error es mínimo, el menor que hemos obtenido **siendo el mejor método para nuestro dataset**. La pureza es también muy alta

Dataset 2:

Métricas de evaluación y análisis:

- Silueta
- Calinski-Harabasz
- Davies-Bouldin

Con los siguientes resultados:
La medida de Silueta con K = 3 es 0.45811237501
La medida de Calinski Harabaz con K = 3 es 0.45811237501
La medida de Davie Bouldin con K = 3 es 0.760425242988

Las métricas aquí, también con k=3, muestran concordancia con los obtenidos con las métricas extrínsecas

Agrupación espectral: Matriz Laplaciana

El método de la matriz laplaciana consiste en llevar a cabo el análisis de autovalores y autovectores de la matriz laplaciana. La matriz laplaciana es una descripción del grafo que representa los datos del dataset a clusterizar. El análisis de la matriz laplaciana se basa en la teoría de grafos. Se eligió el método por las características de la relación con la teoría de grafos.

Dataset 1:
Métricas de evaluación y análisis:

1. Matriz Confusion
2. Informacion Mutua
3. Valor de Error
4. Pureza del Agrupamiento
5. Precision
6. Recall
7. F1
8. Entropia

Matriz de Confusión

[[61 8 1]

[10 0 60]

[3 67 0]]

La información mutua es = 0.755849783666

El valor del error cometido es = 0.104761904762

La pureza del agrupamiento obtenido es = 0.895238095238

El valor F1 es = 0.895796895869

La entropía es = 0.342762505002

El error es bajo, pero no mejor que el método aglomerativo. La pureza es también menor.
Dataset 2:

Métricas de evaluación y análisis:

1. Silueta
2. Calinski-Harabasz
3. Davies-Bouldin

Obteniendo los valores siguientes:

La medida de Silueta con K = 5 es 0.366263982158

La medida de Calinski Harabaz con K = 5 es 0.366263982158

La medida de Davie Bouldin con K = 5 es 0.927095745999

Las metricas son menos deseables que con el método aglomerativo.

Agrupación por densidad: DBSCAN

El algoritmo DBSCAN se basa en la identificación de agrupamientos por densidad, seleccionando varios núcleos adecuados a partir de los cuales se pueden determinar los elementos restantes a la densidad del núcleo al que pertenecen. Es un algoritmo robusto el cual solo tiene problemas cuando existen clústeres con densidades muy dispares. Se eligió por ser uno de los algoritmos más usados y tener variaciones que aumentan su robustez.

Dataset 1:

Se utilizaron métricas de bondad del agrupamiento :

1. Matriz confusión
2. Medida error
3. Medida pureza
4. Medida precisión
5. Medida recall
6. Medida f1 específica
7. Medida f1

Conclusión:

A pesar de aplicar el algoritmo y haberlo intentado con varias combinaciones no encontramos valores de EPS (epsilon) y M que nos generen agrupaciones con el algoritmo DBSCAN.

Por lo tanto se concluye que la distribución de los datos en nuestro dataset no es la mejor para el algoritmo DBSCAN.

Dataset 2:

Conclusión:

Se considera la misma conclusión que en el primer dataset.

Agrupación basado en métodos probabilísticos: EM

El algoritmo de esperanza maximización o EM es un proceso iterativo que ajusta desde las condiciones iniciales los clústeres de forma probabilística. Es parecido al algoritmo K-means en el procedimiento, pero se diferencia en que EM es probabilista, mientras que K-means es determinista. Aunque depende de las condiciones iniciales para poder lograr una convergencia adecuada y evitar mínimos locales, la simplicidad de implementación es por lo que lo elegimos.

Dataset 1:

Métricas de evaluación y análisis:
Las métricas extrínsecas son:

1. Matriz Confusión
2. Información Mutua
3. Valor de Error
4. Pureza del Agrupamiento
5. Precisión
6. Recall
7. F1
8. Entropía

Obteniendo los siguientes resultados:
Matriz de Confusión
[[45 23 2]
[0 1 69]
[70 0 0]]
La información mutua es = 0.6688967228
El valor del error cometido es = 0.12380952381
La pureza del agrupamiento obtenido es = 0.771428571429
Precisión de l=1,k=1 es : 0.0416666666667
Recall de l=1,k=1 es : 0.0142857142857
El valor F1 es = 0.74161395438
La entropía es = 0.429715565868

Bueno error, pero no tan bueno como el aglomerativo.

Dataset 2:

1. RMSSTD
2. R cuadrado
3. Silueta
4. Calinski-Harabasz
5. Davies-Bouldin

Obteniendo los siguientes resultados

El error RMSSTD es = 0.920336133806
El valor de la medida R cuadrado es = 0.205930300166
La medida de Silueta con K = 3 es 0.313449977956
La medida de Calinski Harabaz con K = 3 es 0.313449977956
La medida de Davie Bouldin con K = 3 es 0.822860536092
Estas métricas fueron calculadas con K=3, y aunque son buenas, el método aglomerativo es mejor.

COMPARACIÓN DE RESULTADOS DE LOS DIFERENTES ALGORITMOS

	kemeans++	**Aglomerativo**	Laplace	DBSCAN	EM
Error	0.22	**0.09**	0.1047	NA	0.123
Pureza	0.87	**0.909**	0.895	NA	0.771
Inf. Mutua	0.77	0.78	0.755	NA	0.668
R Cuadrado	0.106				0.205
Calinkski H	283.77	0.458	0.366	NA	0.3134
DaviesBouldin	1.07	0.760	0.92709	NA	0.822
Silueta		0.458	0.366	NA	0.313

El método aglomerativo ha mostrado los mejor resultados para nuestro dataset.

Bibliografía

Repositorio UCI
https://archive.ics.uci.edu/ml/datasets/seeds#

Github usados

https://github.com/mgarciasanes/metodosdeagrupamiento_nosupervisado
https://github.com/pachecoleonardo/metodosdeagrupamiento_nosupervisado
https://github.com/pachecoleonardo/aprendizaje_no_supervisado/tree/master/Grupo_10

Documento VIU
Manual de la asignatura 06MAIR_JHernandez.pdf, Jerónimo Hernández-González
https://campus.viu.es/bbcswebdav/pid-1936825-dt-content-rid-22714459_1/courses/2018_10_A_10379/06MAIR_JHernandez.pdf

https://es.wikipedia.org/wiki/Informaci%C3%B3n_mutua

https://towardsdatascience.com/precision-vs-recall-386cf9f89488

https://scikit-learn.org/stable/auto_examples/model_selection/plot_precision_recall.html

https://en.wikipedia.org/wiki/Cluster_analysis#Evaluation_and_assessment