Comparación de metodologías de trabajo para la identificación de variedades de arroz mediante algoritmos de Machine Learning

Clas Giulia
Maestría en Explotación de Datos y Descubrimiento del Conocimiento
Universidad de Buenos Aires
clas.giulia.s@gmail.com

Massara Lautaro
Maestría en Explotación de Datos y Descubrimiento del Conocimiento
Universidad de Buenos Aires
malautaro@gmail.com

Uribe Alejandro
Maestría en Explotación de Datos y Descubrimiento del Conocimiento
jauriberamirez@gmail.com

Abstract

El arroz es el alimento más consumido en todo el mundo, su proceso de producción, separación y distribución fue evolucionando con el tiempo. En este trabajo se tomaron cinco variedades diferentes de arroz trabajando con un total de 25000 imágenes, 5000 imágenes por variedad. De estas imágenes se obtuvieron las características principales mediante dos métodos, PCA y UMAP. Luego de la extracción de los componentes principales se utilizaron diferentes algoritmos de clustering para comprar la eficiencia de separación. Se observa un desempeño superior en UMAP respecto a PCA, esta etapa demostró ser la esencial. Los diferentes algoritmos utilizados para la clusterización son notablemente más eficientes cuando las componentes principales forman los clústeres necesarios.

1 Introducción

El avance en la automatización de procesos es una constante en el desarrollo de la industria del Machine Learning. A medida que avanza la tecnología y, con las constantes mejoras en los algoritmos, entender cuales son las herramientas efectivas y como deben utilizarse es imperioso. En el presente trabajo se evaluará un dataset de 25000 imágenes[1], 5000 por cada especie. El curado de las imágenes es de muy buena calidad, el tamaño es constante para todas ellas (250x250 pixeles) y poseen un preprocesamiento. Esto pone el foco del trabajo en como utilizar las herramientas analíticas para poder obtener una mejor separación de las especies. En el trabajo original[1] se escogieron parámetros físicos que representan las características de las diferentes variedades de arroz y una serie de algoritmos para clasificar los granos en sus diferentes tipos. Nos proponemos evaluar imágenes de granos de arroz y, utilizando métodos automatizados para la obtención de características principales, compara los resultados obtenidos mediante la aplicación de diferentes técnicas de clusterizado y evaluar la performance obtenidas.

Data Mining en Ciencia y Tecnología 2023.

2 Métodos

El objetivo de este trabajo es evaluar tres métodos de clusterización mediante dos técnicas de obtención de componentes principales.

2.1 Imágenes

Las imágenes de este dataset[1] fueron procesadas por Matlab. Posee imágenes de 250x250, el curado de las imágenes es de alta calidad y la no se encontraron anomalías en las mismas. El desafío relacionado con este dataset es el número de imágenes total (75000, 15000 por cada variedad de arroz) que requiere un poder de procesamiento elevado. Para este trabajo se optó utilizar 5000 imágenes por variedad (25000 imágenes totales).

2.2 Modelado y componentes principales

A estas 25000 imágenes se les aplico el modelo VGG16[2], modelo que posee 16 capas y que utilizaremos las primeras 14 capas para el desarrollo de este trabajo. Se utilizó PCA [3] y UMAP [4]. El único parámetro a optimizar con PCA es la cantidad de componentes a conservar. En el caso de UMAP existen dos hiperparámetros importantes a considerar: la cantidad de vecinos (n_neighbors) y la distancia mínima (min_dist). Se realizó una exploración de los hiperparámetros.

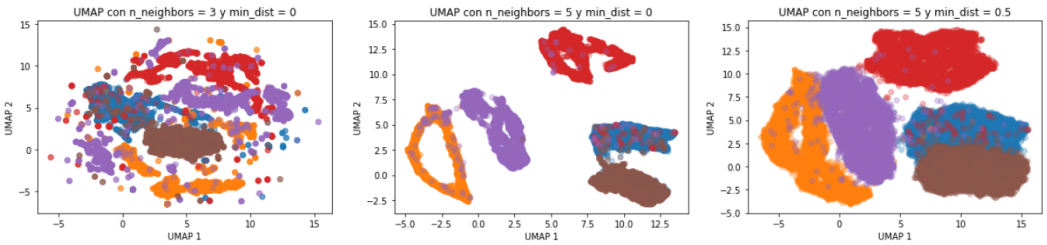

Figure 1: Diferente combinación para hiperparámetros de UMAP

En la figura 1 se encuentran en los ejes los componentes principales de UMAP y el color representa la etiqueta verdadera de ese punto. Se puede observar que frente a un bajo número de vecinos, no logran formarse los clusters de forma clara, mientras que si la aumentamos, podemos diferenciar los clusters con claridad. La min_dist queda fijada en el valor mínimo, si la aumentamos los clusters comienzan a solaparse.

2.3 Algoritmos de clusterización

Se trabajó con 3 algoritmos de clusterización KMean[5], AgglomerativeClustering y HDBSCAN. Cada uno de ellos bajo los parámetros obtenidos en PCA y UMAP. Para estos algoritmos se evaluaron diferentes hiperparámetros y se trabajó con los mejores valores obtenidos. En el caso del KMeans se modificó el número de clusters, para el AgglomerativeClustering se modificó el tipo de vinculación y en HDBSCAN el tamaño mínimo de cluster.

2.4 Parámetros de evaluación

Para evaluar la performance de los modelos se utilizaron las siguientes métricas:

Silhouette score[6]: Interpreta y valida la coherencia dentro del análisis de grupo:

Van Donguen[7]: Mide la pureza de la separación.

$$VD_n = \frac{2n - \sum_i max_j n_i j - \sum_j max_i n_i j}{2n - max_i n_i - max_j n_j}$$

Donde:

n: es el número total de elementos

i y j: las columnas y filas

Rand ajustado[8]: El índice de Rand compara la similitud entre dos métodos de agrupación diferentes

$$R = \frac{a+b}{a+b+c+d}$$

Donde:

a: número de pares de elementos que aparecen juntos en un clúster y además pertenecen a la misma clase.

b: número de pares de elementos que pertenecen a clases diferentes y además están en clústeres diferentes.

c: número de pares de elementos que comparten la clase, pero se ubican en diferentes clústeres.

d: número de elementos que pertenecen a clases diferentes, sin embargo se agrupan en el mismo clúster.

Además de estos parametros se utilzará una matriz de confución para ver el agrupamiento de los resultados.

3 Resultados y discusión

3.1 Componentes principales mediante PCA

3.1.1 Modelado

Para este dataset en particular debido a que las caracterisiticas de los granos de arroz son similares, estrategias como el PCA para obtener los parametros principales no demostraron ser eficientes para distinguir las diferentes variedades. Esto genera que los algoritmos de clusterización no aumentan significativamente la eficiencia de la separación.

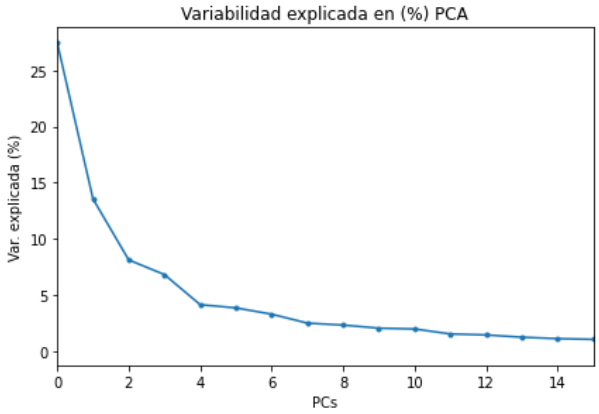

Figure 2: % de la variabilidad explicadad por variable

Con 30 variables se explica el 90% de la variabilidad, la calidad de separación por especie de arroz no es muy buena, en particular hay un solapamiento de dos especies (Basmati y Jasmine) lo que dificulta notablemente la capacidad de los modelos de clusterización para separar las especies.

Analizando los resultados obtenidos mediante los diferentes modelos tenemos:

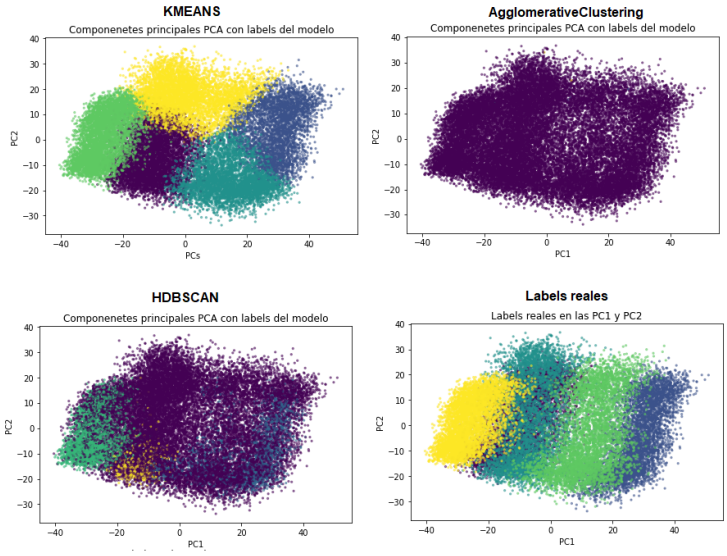

Figure 3: Resultados obtenidos mediante diferentes algoritmos de clusterización sobre PCA (De izquierda a derecha y de arriba a abajo: KMeans, AgglomerativeClustering, HDBSCAN, etiquetas reales)

En la figura 3 se comparan el resultado de los 3 modelos utilizados (KMeans, AgglomerativeClustering y HDBSCAN) con las etiquetas reales de las especies de arroz. Se puede apreciar la dificultad para realizar la separación de clusters mediante la técnica de PCA. A pesar de que contamos con etiquetas de cada imagen y que podemos distinguir tres grupos relativamente bien separados (Amarillo, Verde y Azul) El turquesa se solapa con el violeta y es imposible distinguir un grupo del otro mediante esta combinación de componentes principales y algoritmo de separación.

Si procedemos a analizar los resultados obtenidos mediante el Silhouette score y los residuos en función de la cantidad de clústeres elegidos para el modelado en KMeans, podemos observar que no hay un punto claro de quiebre para ninguno de los dos graficos. El Slhouette siempre es decreciente y a medida que aumentamos el k, el SSE disminuyendo paulatinamente.

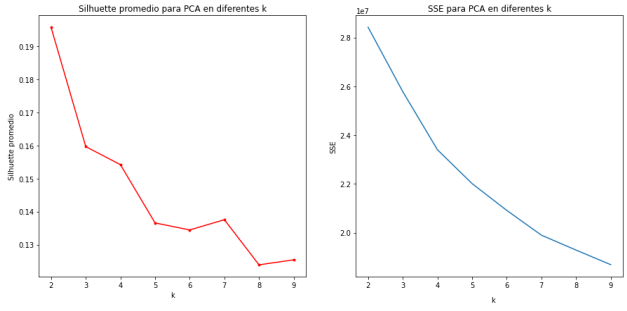

Figure 4: Silhouette score y SSE para KMeans

Resulta complicado poder elegir la cantidad de clusters a ser modelados con la información obtenida a partir de la figura 4, lo que es coherente con lo observado en la figura 3

Los puntajes obtenidos con el índice de Rand ajustado 0,4334 que indica una baja correlación entre las coincidencias entre las etiquetas asignadas por el modelo y las etiquetas reales. Mientras que la pureza de vanDogen de 0,4292 es relativamente alta, indicando que la pureza en la clasificación por elemento es baja.

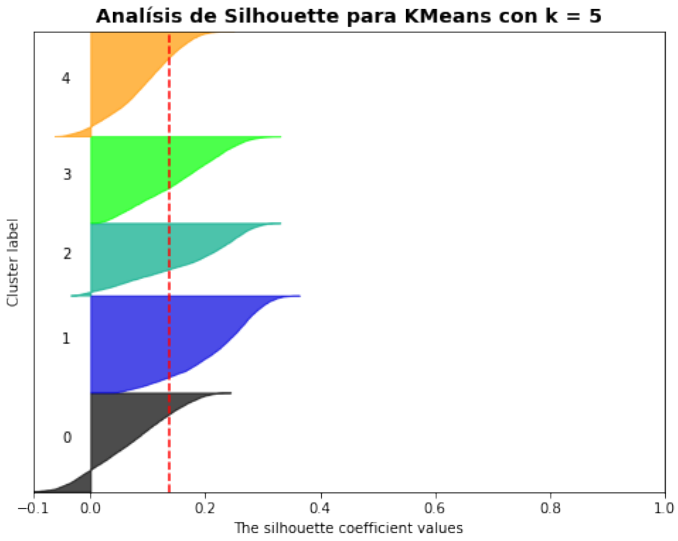

Figure 5: Silhouette score por cluster para KMeans con 5 clusters con PCA

El gráfico de Silhouette por grupo no aporta información significativa, el valor medio es bajo (menor a 0,25) lo que es un indicador de que no tenemos ningún agrupamiento bien definido.

3.1.2 Hiperparámetros

La exploración de hiperparámetros con PCA no genero beneficios significativos. En Kmeans se modificaron la cantidad de clústeres, obteniendo un mejor Silhouette pero que no representa una mejor separación de las etiquetas conocidas. Si modificó la métrica de distancia utilizando KMedoids y tampoco se obtuvo una mejora significativa. En AgglomerativeClustering se modificó el tipo de **linkage** y número de clústeres, pero no hubo grandes cambios en los resultados Finalmente, en HDBSCAN se cambió el tamaño mínimo de clústeres, pero debido al agrupamiento de los puntos, tampoco se obtuvo una mejora significativa. El elemento limitante en estas modificación de hiperparámetros fue el PCA inicial que no genera clústeres lo suficientemente claros como para poder obtener un buen ajuste con los algoritmos.

3.2 Componentes principales sobre UMAP

3.2.1 Modelado

Cuando utilizamos UMAP para la selección de las componentes principales, se pueden apreciar los diferentes agrupamientos. En la exploración de los hiperparámetros en la sección de métodos en la figura 1, se observa una mejor separación, a pesar de que existe el solapamiento entre algunas de las variedades de arroz. La separación y formación de clusters es superior a la obtenida mediante PCA. Con esta información podemos pensar en **tres** o **cinco** clusters como opciones viables para correr los algoritmos.

Utilizando cinco clústeres la separación realizada por el AgglomerativeClustering y el HDBSCAN es muy clara. En el caso del KMeans tiene mayor dificultad en separar el clúster celeste del amarillo. Como demuestra el grafico con las etiquetas correctas, la separación del UMAP no es completa y los clústeres obtenidos por los algoritmos se verán afectados por esa separación. Para optimizar la separación se deberá trabajar más en detalle en la separación de componentes principales.

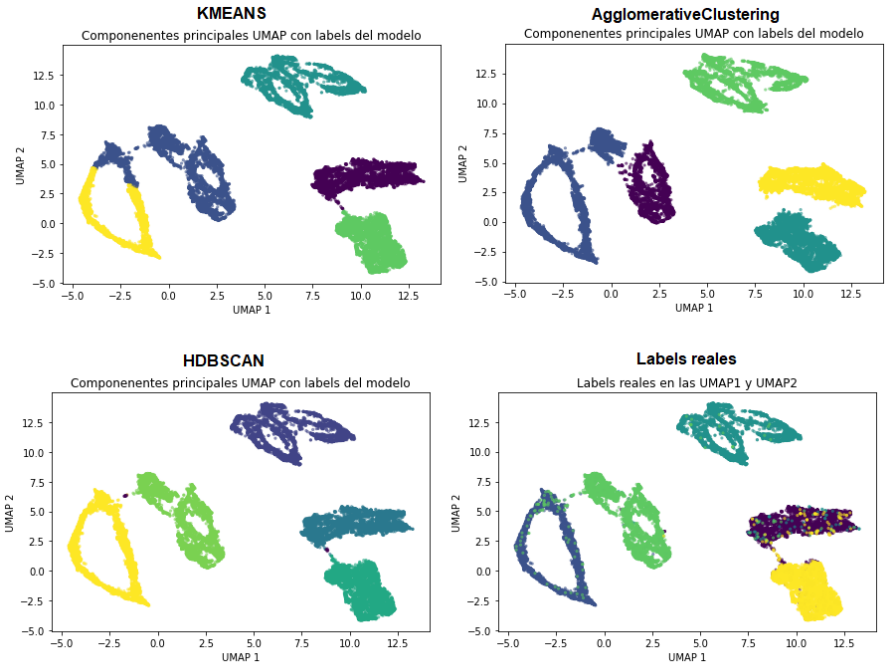

Figure 6: Resultados obtenidos mediante diferentes algoritmos de clusterización sobre UMAP (De izquierda a derecha y de arriba a abajo: KMeans, AgglomerativeClustering, HDBSCAN, etiquetas reales)

El promedio del Silhouette score es superior al obtenido mediante PCA en todos los modelos, la caída es más abrupta en la figura 7, lo que indica un clusteres más homogeneos. Cabe destacar la existencia de un pico en el Silhouette score para tres clusters. Esto se puede apreciar en KMeans y AgglomerativeClustering. Este comportamiento es coherente con lo observado en la figura 6, Se pueden observar 5 clústeres, pero 4 de los 5 se encuentran agrupados de a pares, en este tipo de situaciones el Silhouette score da mejor para menor cantidad de grupos.

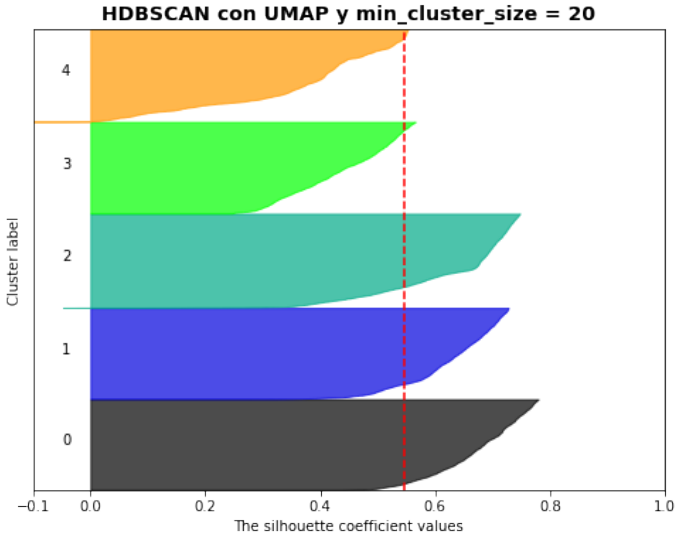

Figure 7: Silhouette Score para HDBSCAN utilizando UMAP

Table 1: Rand Score y vanDongen para modelos con UMAP

Modelo	Rand Score	vanDongen
KMeans	0.8420	0.0892
AgglomerativeClustering	0.8093	0.1126
HDBSCAN	0.9412	0.0312

Los resultados de los puntajes de Rand y vanDongen van en la misma dirección que las inspecciones visuales. El algoritmo de HDBSCAN realizó el mejor trabajo separando los clusters y tiene el Rand score y vanDongen más altos.

En la tabla 2 se encuentra la matriz de confusión que nos permite aprecias tres comportamientos particularmente descriptivos de este dataset y la forma en la que fue trabajado.

Table 2: Matriz de confusión para modelos aplicados sobre UMAP

		Clase predicha				
Especie	Algoritmo	Arborio	Basmati	Ipsala	Jasmine	Karacadag
Arborio	KMeans	4793	0	3	20	184
	AC	4793	1	3	19	184
	HDBSCAN	4813	0	2	18	167
Basmati	KMeans	1	3673	0	1326	0
	AC	1	3824	0	1175	0
	HDBSCAN	1	4926	0	49	0
Ipsala	KMeans	60	0	4901	38	1
	AC	60	4	4901	34	1
	HDBSCAN	66	0	4890	39	3
Jasmine	KMeans	28	34	17	4920	1
	AC	28	1	17	4301	653
	HDBSCAN	22	83	18	4825	1
Karacadag	KMeans	78	1	0	0	4921
	AC	78	0	0	1	4921
	HDBSCAN	91	3	0	0	4908

- En primera instancia tenemos los limites sobre los cuales pueden trabajar los algoritmos de clusterizado en función de la separación en los componentes principales. Las especies de Arborio, Ipsala y Karacadag tienen cotas superiores bien definidas, estas operan como los límites máximos con los parámetros utilizados del UMAP.
- También podemos detectar cuales son las especies conflictivas, con las componentes principales obtenidas es posible la separación eficiente de Jasmine, pero no podemos distinguir de forma clara un Basmati de un Jasmine.
- La eficiencia de los algoritmos para poder tratar de la mejor forma estos resultados. Vimos resultados parciales de esta performance en la figura (ref figura modelo), ahora tenemos la confirmación mediante la matriz de confusión. Utilizando HDBSCAN es posible la correcta separación del Basmati y Jasmine.

3.2.2 Hiperparámetros

Al tener una separación más eficiente mediante UMAP, la optimización de hiperparámetros tiene un impacto mayor que en PCA. UMAP es donde enfocaremos el análisis de los hiperparámetros

KMeans Para todos los algoritmos que nos hacen elegir un grupo de clústeres a priori tenemos el mismo problema. El Silhouette tiene su pico en tres clústeres, como fue mencionado en la discusión del modelado. Esto se debe a como fueron separadas las componentes con el UMAP y la similitud que tienen algunas variedades de arroz.

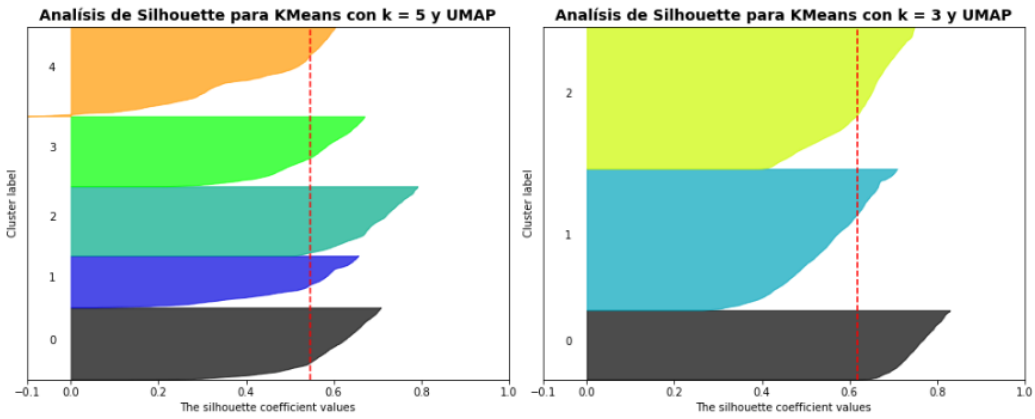

Figure 8: KMeans con 3 clústeres y 5 clústeres

En la figura 9 se realizó la comparación de los dos modelos contra las etiquetas reales. En esta figura se puede aprecia claramente el efector de los diferentes clústeres. Cuando el número de clústeres es tres, las diferencias con los otros clusteres son significativas, en cambio cuando son 5 clústeres, hay un clúster separado y dos pares de clústeres muy cercanos, lo que disminuye el puntaje obtenido.

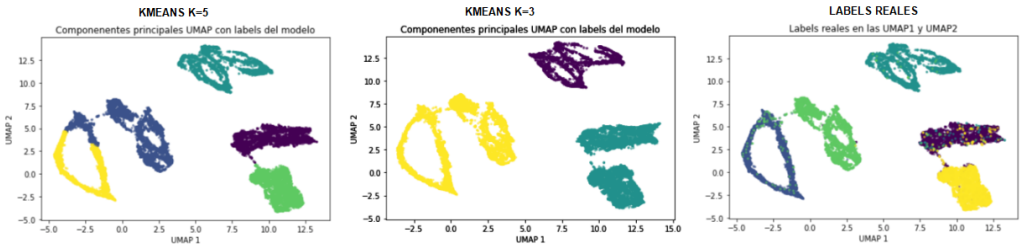

Figure 9: De izquierda a derecha, KMeans con 5 clústeres, KMeans 3 clústeres y etiquetas reales

AgglomerativeClustering El tipo de linkage elegido puede modificar significativamente el resultado de la agrupación de valores. En la optimización de hiperparámetros del AgglomerativeClustering utilizamos un número constante de clústeres y modificamos el tipo de linkage. Obtuvimos resultados diferentes. Desde una agrupación similar a la de KMeans con linkage promedio, hasta un grupo de un solo punto con linkage single y una separación más acorde a lo que estábamos buscando utilizando linkage "Ward".

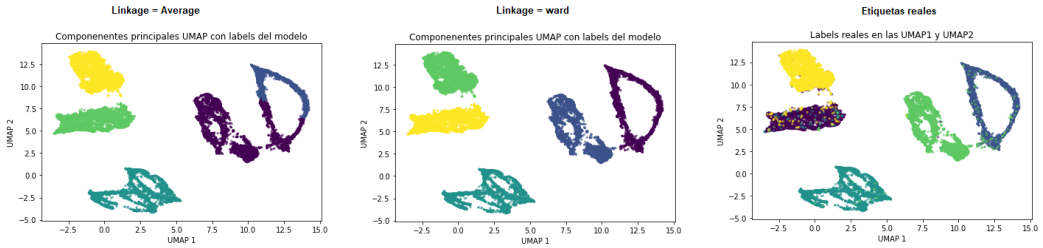

Figure 10: De izquierda a derecha, AgglomerativeClustering con linkage promedio, ward y etiquetas reales

HDBSCAN En el caso de HDBSCAN trabajos particularmente sobre el tamaño mínimo de clúster, El tamaño mínimo para que un punto sea considerado un centro está limitado por el tamaño mínimo de clúster, por defecto al definir el tamaño mínimo de clúster fijamos el tamaño mínimo de muestra.

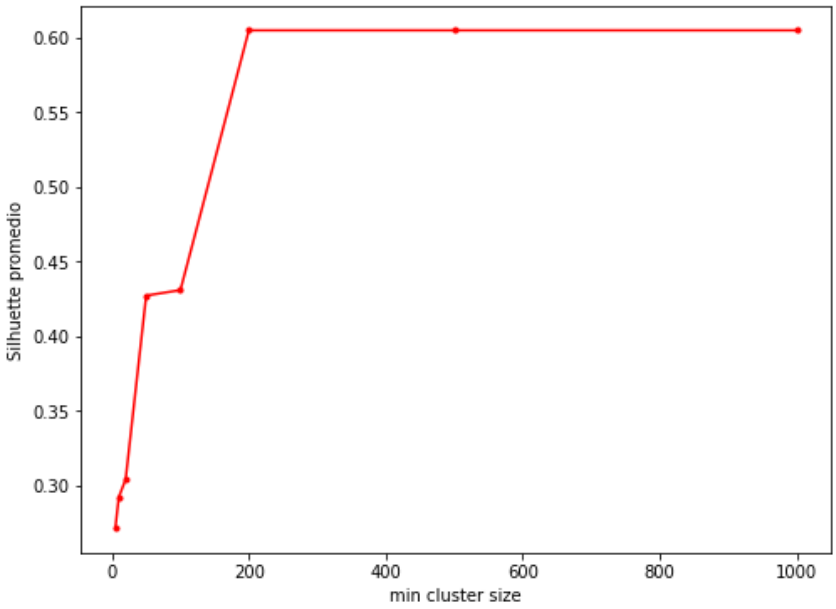

Figure 11: Silhouette score vs tamaño mínimo del clúster

Encontramos que el Silhouette score obtiene un valor máximo a partir del tamaño mínimo de cluster de 200. Esta información hay que compararla con los resultados visuales obtenidos para ver si existe coherencia entre lo que estamos buscando y lo que obtuvimos.

3.3 Detección de objetos dentro de una imagen

Teniendo en cuenta el pre procesamiento que poseen las imágenes, se decidió combinar múltiples imágenes con el fin de observar mejor la performance de los diferentes algoritmos de detección de objetos. Para ello construimos una nueva imagen que está compuesta de múltiples imágenes y dos objetos pegados para apreciar la diferencia entre los dos algoritmos utilizados.

3.3.1 Connected-component labelling y clustering espectral

La clasificación de objetos obtenida por el Connected-component labeling es muy buena para objetos que se encuentran aislados. Otra ventaja de este algoritmo es que no hay que indicarle un número inicial de objetos a detectar. Para el caso del spectral clustering no obtuvimos tan buenos resultados, se intentó con múltiples imágenes y modificando algunos de los hiperparámetros, los resultados para separar objetos sin intersección funcionaron adecuadamente, pero cuando aparece alguna intersección comienza a dividir los objetos de manera cuestionable.

Figure 12: De izquierda a derecha, Component labeling y Spectral clustering

4 Concluciones

Mediante la separación de componentes principales, este trabajo cubrió dos aspectos: Como el preprocesamiento de los datos impacta en el resultado obtenido y cual es el impacto de los algoritmos de clusterización en esos datos. En una primera instancia el trabajo estaba orientado en verificar la performance de los modelos de clusterizado y optimización de hiperparametros, pero en el proceso, y con los resultados obtenidos mediante PCA, el foco se corrió en la búsqueda de una mejor separación de dichos componentes. PCA no resultó eficiente para obtener una separación de componentes principales en este dataset. Los granos de arroz poseen características muy similares y a pesar de que se puede explicar una gran parte de la variabilidad mediante pocas variables con el PCA no es suficiente para distinguir las variedades. UMAP genero clústeres claros visualmente y los resultados de las variables elegidas para los puntajes acompañan en el mismo sentido. Los índices de Rand y van Dongen son muy superiores a los obtenidos mediante PCA y las mismas técnicas de clusterización. Todavía existe un problema de solapamiento en la separación de componentes principales, que limita los resultados que podemos obtener. El resultado obtenido es muy bueno para un método automático, incluso siendo aplicado sin un conocimiento profundo en el área de trabajo. Utilizando UMAP y HDBSCAN se obtuvo una tasa de clasificación correcta superior al 95%.

La separación de objetos mediante Connected-component labelling y clustering espectral no fue del todo satisfactoria. Connected-component labelling funcionó de la forma esperada, pero el clustering espectral puede ser mejorado.

References

[1] lkay Cinar and Murat Koklu. Classification of rice varieties using artificial intelligence methods. International Journal of Intelligent Systems and Applications in Engineering, 7(3):188–194, 2019.

[2] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014

[3] Andrzej Maćkiewicz, Waldemar Ratajczak, Principal components analysis (PCA). Computers & Geosciences, Volume 19, Issue 3, 1993.

[4] Leland McInnes, John Healy, James Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv:1802.03426v3 2018.

[5] Alsabti, Khaled & Ranka, Sanjay & Singh, Vineet. An Efficient K-Means Clustering Algorithm. Proc First Workshop High Performance Data Mining. 2000.

[6] Peter J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, Volume 20, 1987, Pages 53-65,

[7] Stichting, Copyright & Centrum, Mathematisch & Dongen, Stijn. Performance Criteria for Graph Clustering and Markov Cluster Experiments. (2000).

[8] W. M. Rand (1971). Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association. American Statistical Association. 66 (336): 846–850