

Agrupamientos basados en modelos: Semilla de trigo

Pedro Pablo Silva Antilef¹[0000–1111–2222–3333]

Universidad de Santiago de Chile, Facultad de ingeniería, Departamento de ingeniería informática, Estación Central, Santiago de Chile `pedro.silva@usach.cl`

Abstract. En el presente trabajo se presenta un análisis de una base de datos con información geométricas de semillas de trigo a través de clusterización basada en modelos, específicamente con el modelo *Gaussian Mixture*. Los mejores modelos se eligen con los criterios de selección *Bayesian Information Criterion* (BIC) y *Integrated Completed Likelihood* (ICL). Se encontró que la eliminación de variables con información redundante podría ayudar a mejorar la precisión del agrupamiento.

Keywords: Model-based Clustering · Wheat Seeds · Data Mining.

1 Introducción

El reconocimiento y autenticación son tareas esenciales para garantizar la calidad de las semillas en la cadena productiva industrial, con énfasis en el proceso de certificación. Estas tareas se siguen haciendo de manera manual en varias industrias, por lo que la aplicación de métodos computacionales y automatizados podría significar una gran ayuda para diferentes empresas [1].

El reconocimiento de variedades también es una tarea crítica para limitar cualitativa y cuantitativamente las posibles pérdidas en el campo de cultivo, así como también para predecir la cosecha productiva asincrónicas [2].

2 Estado del arte

En [3] se utiliza este el algoritmo de cluster gradiente completo o CGCA por sus siglas en ingles (*Complete Gradient Clustering Algorithm*), con el que busca definir grupos que permitan representar las características de los granos de trigo, en base a la distribución de los datos, ya que el CGCA necesita calcular la estimación de densidad de kernel [5]. Cada cluster es caracterizado por un máximo local de la estimación de densidad de kernel. Con esto, las regiones con una mayor densidad de objetos son definidos como un grupo o cluster. Cada dato es asignado a cada cluster utilizando el método de gradiente ascendente.

Por otro lado existen varios acercamiento en el último tiempo que se han desarrollado con técnicas de redes neuronales profundas, mejorando los resultados, utilizando una gran cantidad de datos, pudiendo entrenar estos modelos de manera mucho más robusta [1][4].

3 Datos

El *dataset* utilizado es denominado *seeds*, fue creado por el Instituto de Agrofísica de la Academia Polaca de Ciencias en Lublin, pero extraídos desde el repositorio *UCI Machine Learning Repository* de la Universidad de California Irvine [6]. Contiene 210 registros donde se describen 7 características geométricas del núcleo de 3 variedades de semillas de trigo: Kama, Rosa y Canadian, 70 elementos para cada variedad. Para esto se utilizaron técnicas de rayos X débil. Las variables contenidas en el *dataset* son: Área A , perímetro P , compactibilidad $C = 4\pi A/P^2$, largo del núcleo l_kernel , ancho del núcleo w_kernel , coeficiente de asimetría *assimetry* y largo de la ranura del núcleo l_groove . Todas las variables son numéricas continuas y no existen datos con valores nulos o faltantes.

4 Metodología

Clustering basado en modelos considera que los datos vienen de una distribución de dos o mas clusters. A diferencia de otros métodos como k-means, la clusterización basada en modelos utiliza una asignación del tipo "soft", donde cada punto de datos tiene una probabilidad de pertenecer a cada cluster. Uno de los modelos más utilizados en este tipo de agrupamiento es la función *Gaussian Mixture*, el cual es una función compuesta por k distribuciones gaussianas, donde la asignación a cada cluster se hace de manera probabilística, en ese sentido la matriz de covarianza describe los cluster en función del volumen, forma y orientación del cluster.

Para la selección de modelos, se utilizan los siguientes criterios: *Bayesian Information Criterion* (BIC), el cual utiliza máxima verosimilitud con una combinatoria de los parámetros de clustering y el *Integrated Completed Likelihood* (ICL), el cual maximiza la verosimilitud de los datos del dataset.

Se realizaron 3 experimentos para el análisis de este tipo de clusterización en el *dataset* de semillas de trigo:

- Experimento 1: Se utilizan todas las variables del *dataset*.
- Experimento 2: Se elimina la variable "area" al estar esta muy correlacionada con la variable "perimeter".
- Experimento 3: Se elimina la variable "area" y "perimeter", ya que tienen una alta correlación respecto a las métricas de largo y ancho del núcleo de la semilla, lo que probablemente pueda estar agregando información que ya se encuentra contenida en otra variable.

5 Resultados

En esta sección se presentan los resultados obtenidos de los diferentes experimentos planteados y explicados en la sección 4, así como también la matriz de correlación de las variables numéricas continuas del *dataset*. Estos resultados serán discutidos en la siguiente sección.

A continuación se muestran los resultados obtenidos para el experimento 1, considerando los métodos BIC y ICL. También se pueden apreciar la matriz de clusterización generada por el algoritmo en el anexo, específicamente en la Figura 8.

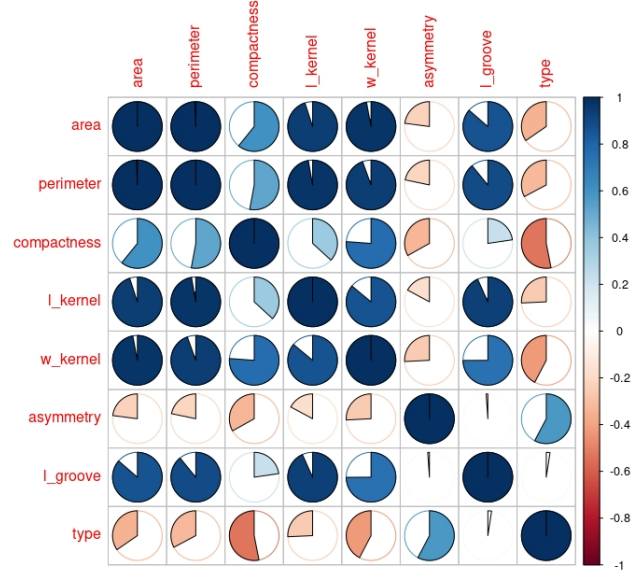


Fig. 1: Matriz de correlación de variables

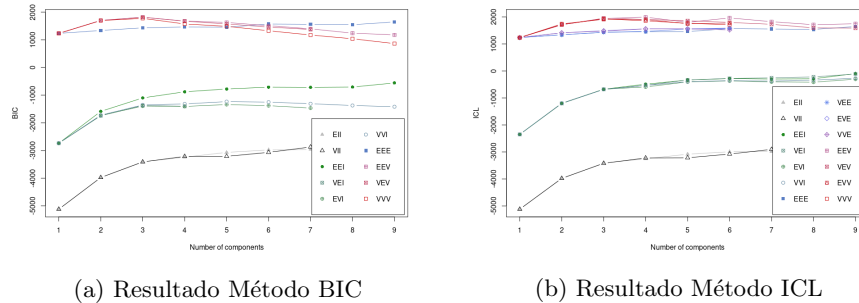


Fig. 2: Resultado para experimento 1

	EEV,3	VEV,3	VVV,3
BIC	1817.581	1813.006	1773.101
BIC diff	0.000	-4.575	-44.481

Table 1: Mejores resultados para BIC en experimento 1

Variedad l	Grupo	1	2	3
1		58	2	10
2		1	69	0
3		0	0	70

Table 3: Matriz de distribución de clases por grupo usando BIC

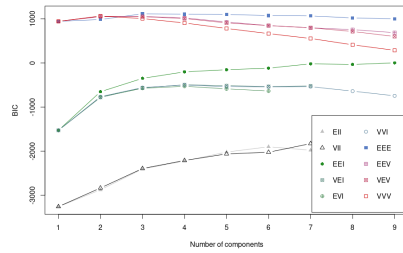
	EEV,3	VEV,3	VVV,3
ICL	1812.852	1807.849	1766.297
BIC diff	0.000	-5.003	-46.556

Table 2: Mejores resultados para ICL en experimento 1

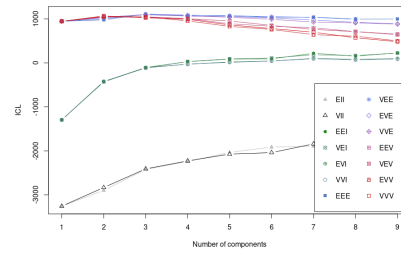
Variedad l	Grupo	1	2	3	4
1		59	1	10	4
2		1	69	0	0
3		0	0	65	5

Table 4: Matriz de distribución de clases por grupo usando ICL

A continuación se muestran los resultados obtenidos para el experimento 2, considerando los métodos BIC y ICL. También se pueden apreciar la matriz de clusterización generada por el algoritmo en el anexo, específicamente en la Figura 9.



(a) Resultado Método BIC



(b) Resultado Método ICL

Fig. 3: Resultado para experimento 2

	EEE,3	EEE,4	EEE,5
BIC	1113.879	1103.652	1099.189
BIC diff	0.000	-10.227	-14.690

Table 5: Mejores resultados para BIC en experimento 2

	VEE,3	EEE,3	EVE,3
ICL	1108.053	1105.996	1099.201
BIC diff	0.000	-2.057	-8.851

Table 6: Mejores resultados para ICL en experimento 2

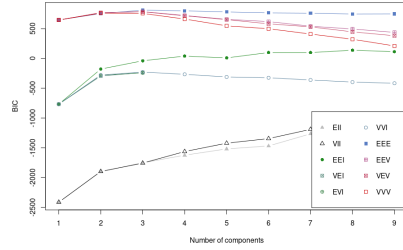
Variedad \ Grupo	1	2	3
1	67	2	1
2	0	70	0
3	5	0	65

Table 7: Matriz de distribución de clases por grupo usando BIC

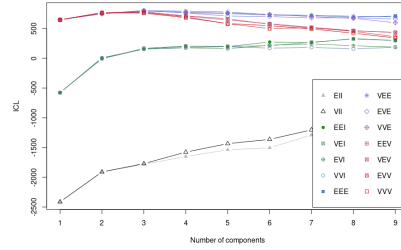
Variedad \ Grupo	1	2	3
1	62	3	5
2	5	65	0
3	6	0	64

Table 8: Matriz de distribución de clases por grupo usando ICL

A continuación se muestran los resultados obtenidos para el experimento 3, considerando los métodos BIC y ICL. También se pueden apreciar la matriz de clusterización generada por el algoritmo en el anexo, específicamente en la Figura 10.



(a) Resultado Método BIC



(b) Resultado Método ICL

Fig. 4: Resultado para experimento 3

	EEV,3	VEV,3	VVV,3
BIC	811.066	799.262	786.728
BIC diff	0.000	-11.804	-24.337

Table 9: Mejores resultados para BIC en experimento 3

	EVE,3	VEE,3	EEE,3
ICL	804.937	800.570	798.417
BIC diff	0.000	-4.367	-6.521

Table 10: Mejores resultados para ICL en experimento 3

Variedad \ Grupo	1	2	3
1	67	2	1
2	0	70	0
3	6	0	64

Table 11: Matriz de distribución de clases por grupo usando BIC

Variedad \ Grupo	1	2	3
1	63	2	5
2	0	70	0
3	4	0	66

Table 12: Matriz de distribución de clases por grupo usando ICL

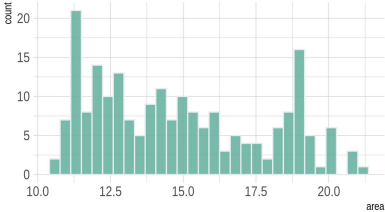
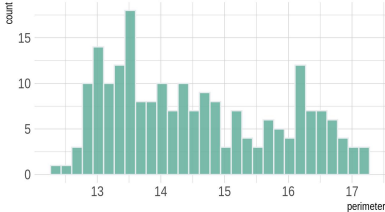
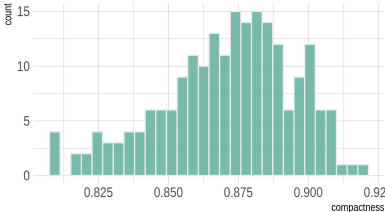
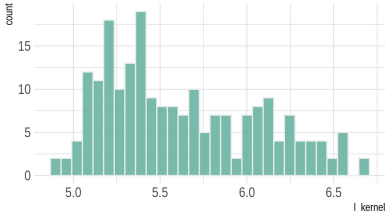
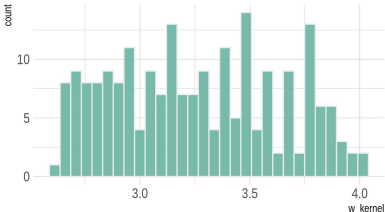
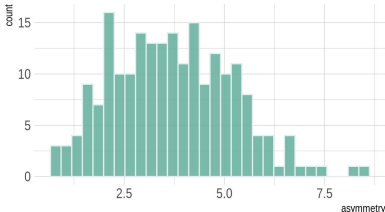
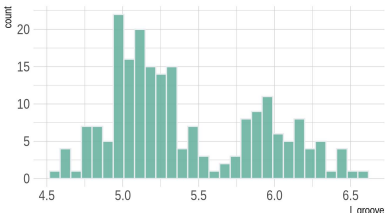
6 Discusión y conclusiones

La matriz de correlación muestra que existe una correlación muy grande entre las variables de área y perimeter, así como estas a su vez están fuertemente correlacionadas con las métricas de largo y ancho del núcleo de la semilla de trigo. En este sentido, se buscó eliminar las variables área y perimeter, con la idea de eliminar información redundante, lo que no mejoró de manera considerable. Esto permitió mejorar la asignación de grupos en los experimentos 2 y 3, donde se obtuvieron los mejores resultados para el experimento 2 utilizando el método BIC, levemente mejor utilizando el mismo método en el experimento 3. En general para el *dataset* analizado, se obtuvieron los mejores resultados con el método BIC. Por otro lado la variedad de semilla Rosa (valor 2), fue la que se agrupó de mejor manera en el cluster correcto, obteniendo un 100% de acierto en el experimento 2 utilizando BIC y en el experimento 3 utilizando ambos métodos.

Una razón que podría generar problemas con la clusterización es el hecho de que las variables no tenían una distribución normal, solo la variable *asymmetry* y *compactness* tienen distribuciones que se asemejan a una del tipo normal.

Se puede concluir que al remover una variable redundante (casi 1 de correlación entre las variables *area* y *perimeter*), se mejoraron los resultados, teniendo un error de 3 asignaciones para la variedad kama (tipo 1), 0 error en la variedad rosa (tipo 2) y finalmente 6 asignaciones erróneas para la variedad de trigo canadian (tipo 3). Se propone realizar otros experimentos para identificar las variables que entregan el mejor resultados en términos de una correcta agrupación de los datos. Así como también probar con otros modelos, así como también con otros métodos de selección de modelos (criterios).

7 Anexos

<div><p>Histograma de variable: area</p><p>count</p><p>area</p></div>	<div><p>Histograma de variable: perimeter</p><p>count</p><p>perimeter</p></div>
Histograma variable area	Histograma variable perimeter
<div><p>Histograma de variable: compactness</p><p>count</p><p>compactness</p></div>	<div><p>Histograma de variable: l_kernel</p><p>count</p><p>l_kernel</p></div>
Histograma variable compactness	Histograma variable large kernel
<div><p>Histograma de variable: w_kernel</p><p>count</p><p>w_kernel</p></div>	<div><p>Histograma de variable: asymmetry</p><p>count</p><p>asymmetry</p></div>
Histograma variable width	Histograma variable assymetry
<div><p>Histograma de variable: l_groove</p><p>count</p><p>l_groove</p></div>	
Histograma variable large kernel groove	

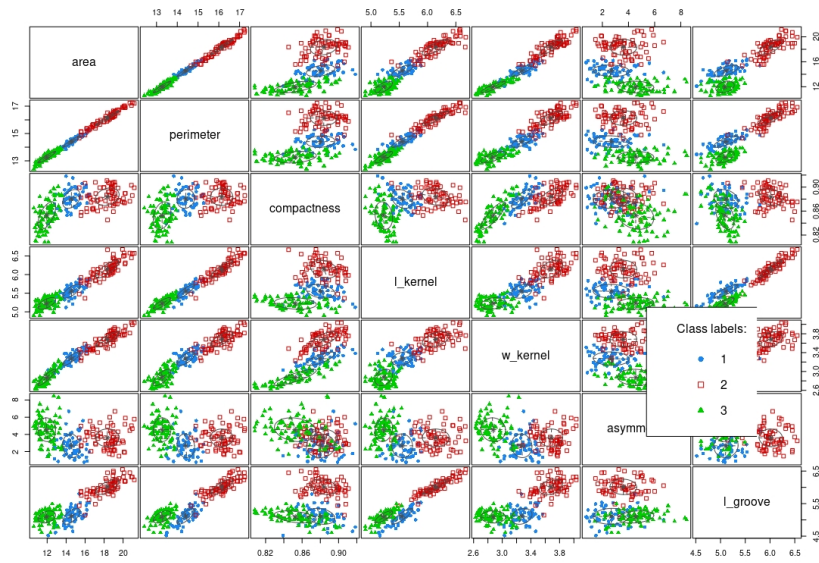


Fig. 5: Gráfico método BIC para experimento 1

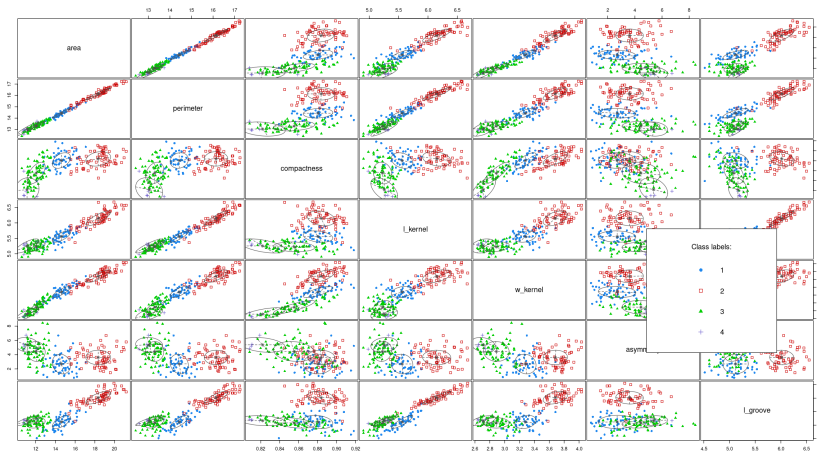


Fig. 6: Gráfico método ICL para experimento 1

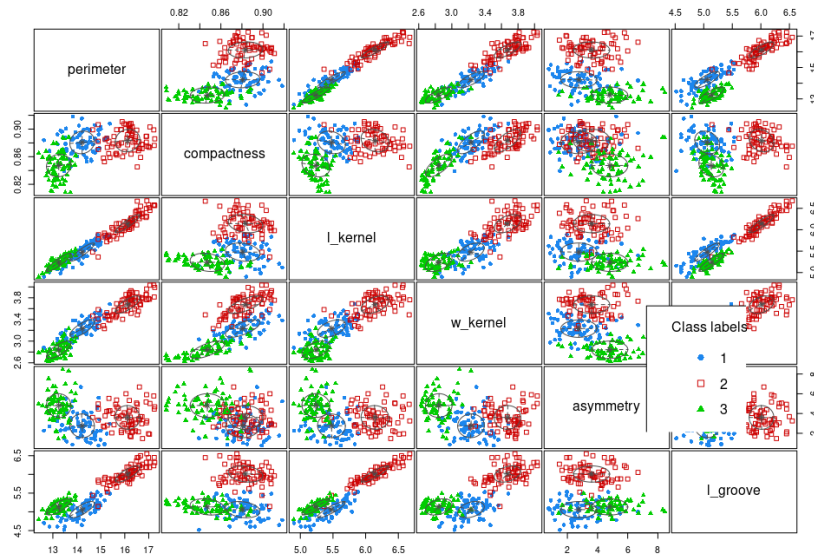


Fig. 7: Gráfico método BIC para experimento 2

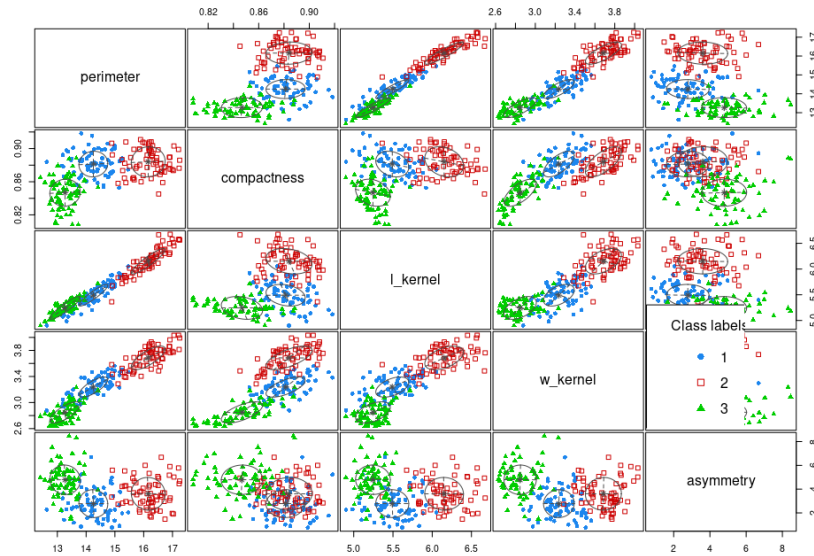


Fig. 8: Gráfico método ICL para experimento 2

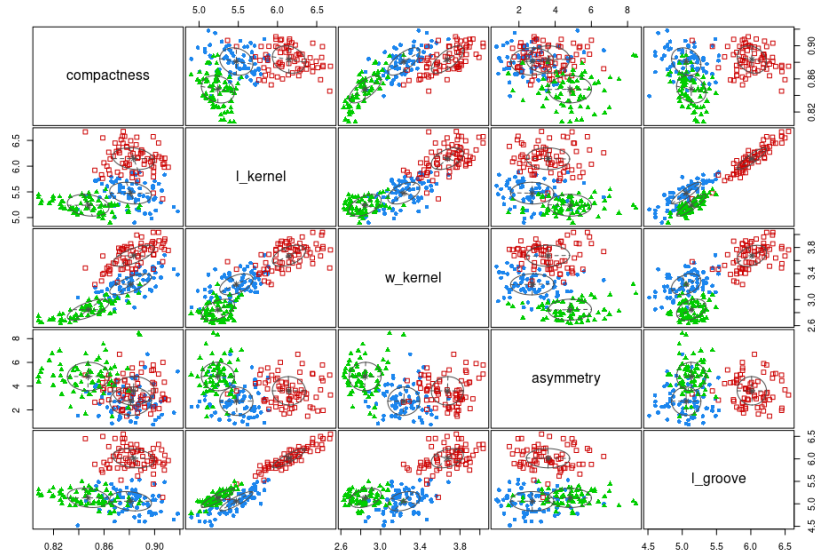


Fig. 9: Gráfico método BIC para experimento 3

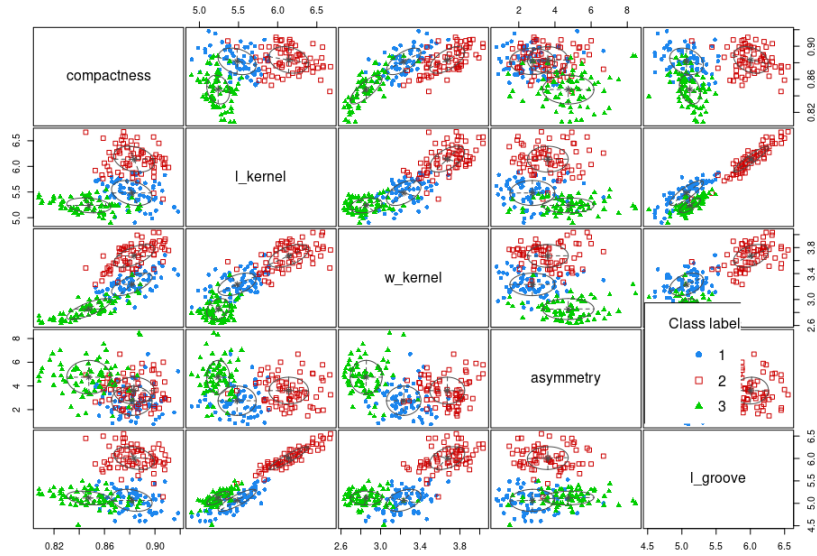


Fig. 10: Gráfico método ICL para experimento 3

References

1. Karim Laabassi: Wheat varieties identification based on a deep learning approach. *Journal of the Saudi Society of Agricultural Sciences* (2021)
2. Taheri-Garavand, A., Nasiri, A., Fanourakis, D., Fatahi, S., Omid, M., & Nikoloudakis, N.: Automated In Situ Seed Variety Identification via Deep Learning: A Case Study in Chickpea. *Journal of the Saudi Society of Agricultural Sciences Plants* (Basel, Switzerland), 10(7), 1406 (2021). <https://doi.org/10.3390/plants10071406>
3. Charytanowicz, M., Niewczas, J., Kulczycki, P., Kowalski, P. A., Łukasik, S., & Żak, S. : Complete gradient clustering algorithm for features analysis of x-ray images. In *Information technologies in biomedicine*. (pp. 15-24). Springer, Berlin, Heidelberg. (2010)
4. Zhou Lei, Zhang Chu, Taha Mohamed Farag, Wei Xinhua, He Yong, Qiu Zhengjun, Liu Yufei: CWheat Kernel Variety Identification Based on a Large Near-Infrared Spectral Dataset and a Novel Deep Learning-Based Feature Selection Method. *Frontiers in Plant Science*. (2020).
5. Medium Estimación de densidad de kernel, <https://medium.com/@garzonsergio/m%C3%A9todos-de-estimaci%C3%B3n-de-densidad-de-kernel-de-odf-a-ebsd-b4a143dc9eee>: :text=Los%C3%A9todos%20de%20estimaci%C3%B3n%20de. Last accessed 12 Abr 2022
6. UCI Repository, <https://archive.ics.uci.edu/>. Last accessed 12 Abr 2022