

Selección de variables informativas en imágenes hiperespectrales

Informative Variable Selection in Hyperspectral Images

Doctoranda:

Maria Laura Vranic

<https://orcid.org/0000-0002-9627-6575>

Universidad Tecnológica Nacional-Facultad Regional Buenos Aires, Argentina.

mvranic@frba.utn.gov.ar

Director:

Claudio DELRIEUX

Codirector:

Juan VOROBIOFF

Resumen

Este trabajo analiza métodos de selección de variables para reducir la alta dimensionalidad, en imágenes hiperespectrales de maní contaminado con aflatoxinas. Se evaluaron 194 bandas en el rango 500–700 nm mediante técnicas estadísticas, teoría de la información, métodos penalizados en grupos y algoritmos Relief. Los resultados muestran que las bandas cercanas a 640–700 nm son las más discriminantes, mientras que el espectro completo presenta redundancia elevada. La validación cruzada con un clasificador CART confirma que la selección de características mejora la estabilidad del modelo y disminuye el sobreajuste, permitiendo identificar subconjuntos de bandas útiles para aplicaciones de control de calidad en la industria alimentaria.

Palabras clave: Imágenes hiperespectrales; Selección de variables; Información mutua; grLASSO / grSCAD; Relief.

Abstract

This work analyzes variable selection methods to reduce high dimensionality in hyperspectral images of peanuts contaminated with aflatoxins. A total of 194 spectral bands in the 500–700 nm range were evaluated using statistical techniques, information theoretic measures, group penalized methods, and Relief based algorithms. The results show that bands near 640–700 nm are the most discriminative, while the full spectrum exhibits substantial redundancy. Cross validation with a CART classifier confirms that feature selection improves model stability and reduces overfitting, enabling the identification of informative band subsets suitable for quality control applications in the food industry.

Keywords: Hyperspectral imaging; Feature selection; Mutual information; grLASSO / grSCAD; Relief.

Introducción

En los últimos años, la tecnología de Imágenes Hiperespectrales (HSI) ha surgido como una herramienta prometedora al combinar las ventajas de la espectroscopía infrarroja y la visión por computadora. Esta técnica permite obtener simultáneamente información espacial y química de cada muestra a través de un "cubo de datos" que abarca cientos de bandas espectrales contiguas, consolidándose como una metodología no destructiva de referencia para la evaluación de la calidad, autenticidad y seguridad microbiológica o tóxica en la industria agroalimentaria (Jiang et al., 2023). Sin embargo, el procesamiento e interpretación de estos datos enfrenta un obstáculo metodológico crítico: la maldición de la dimensionalidad y la presencia inherente de multicolinealidad.

Trabajos previos como el de Vranić et al. (2024a) han demostrado un desempeño analítico sobresaliente en la discriminación o cuantificación de aflatoxinas en maní mediante HSI; sin embargo, sus estimaciones y modelos óptimos dependen del cómputo exhaustivo sobre la totalidad del perfil espectral (operando con 194 variables o bandas independientes). Aunque estos enfoques con espectro completo alcanzan métricas de ajuste elevadas en entornos controlados, la altísima redundancia entre esas 194 dimensiones no solo satura el costo computacional, sino que mantiene latente el riesgo de inestabilidad frente a nuevas muestras y dificulta el desarrollo de sensores industriales en línea. Para superar la inestabilidad derivada de la multicolinealidad y reducir dicha carga dimensional, la selección de características o selección de bandas se posiciona como una etapa indispensable en el flujo de procesamiento de datos. En este contexto, el objetivo de este trabajo es evaluar y comparar el desempeño de diferentes técnicas de selección de variables.

Desarrollo

Descripción de los datos: Se analizó un conjunto de datos obtenidos de imágenes hiperespectrales de muestras de maní (500–700 nm, 194 bandas espectrales) dividido en tres clases: control (SC), contaminado con aflatoxina B₁ (CB1) y con una mezcla de aflatoxinas (CMIX). Tras el procesamiento, la matriz de datos quedó estructurada en 21145 observaciones de entrenamiento y 5287 observaciones de prueba. Los detalles sobre la preparación de las muestras, adquisición de imágenes y su procesamiento se pueden consultar en Vranic et al. (2024b).

Métodos de selección de variables: En la Tabla 1 se resumen los métodos de selección de características evaluados en este trabajo, clasificados según su enfoque analítico (univariado o multivariado), la metodología empleada, que abarca desde estadística tradicional y teoría de la información hasta modelos penalizados en grupos (grLASSO y grSCAD) y algoritmos basados en vecindad.

	Método	Enfoque
Métodos estadísticos tradicionales	ANOVA no paramétrico (Kruskal–Wallis) + Test de Dunn	Univariado (Xs vs. Y)
	Tamaño del efecto (ϵ_{KW}^2)	Univariado (Xs vs. Y)
	Asimetría y Curtosis (AED)	Univariado (Xs)
	Matriz de Correlación de Pearson	Multivariado (Xs vs. Xs)
	Análisis de Clustering	Multivariado (Xs vs. Xs)
Regresión Penalizada en Grupos	grLASSO (<i>Group LASSO</i>)	Selecciona/elimina grupos completos de predictores correlacionados, evaluando la importancia individual y promedio por grupo.
	grSCAD (<i>Group SCAD</i>)	Penalización no convexa por grupos. Reduce el sesgo en la estimación de coeficientes grandes respecto a grLASSO, seleccionando grupos de variables con mayor significancia estructural.
Teoría de la Información	Información Mutua	Univariado /Supervisado (Xs vs. Y). Medir la dependencia probabilística general (lineal y no lineal) con la variable objetivo.
	Entropía de Shannon	Univariado (Xs). Cuantificar la incertidumbre o contenido de información intrínseco.
	mRMR	Filtro Multivariado. Maximizar la relevancia hacia Y penalizando la redundancia entre atributos.
Relief Family	ReliefF, Multi-SURF	Filtro Multivariado basado en vecindad. Estimar la importancia considerando interacciones no lineales y la estructura local del espacio de características.

Tabla 1: Resumen de las técnicas de selección de variables

Métodos estadísticos tradicionales: Como que las variables espectrales no cumplieron con los supuestos de normalidad (prueba de Shapiro–Wilks, $p < 0,05$) ni de homogeneidad de varianzas (test de Levene), se empleó la prueba no paramétrica de Kruskal–Wallis para evaluar la existencia de diferencias significativas entre las medias de las categorías de contaminación

(SC, CB1 y CMIX) en cada longitud de onda analizada. Se encontraron diferencias estadísticamente significativas en todas las variables espectrales, excepto a 519,06 nm. Cuando el contraste global resultó significativo, se realizaron comparaciones múltiples por pares de medias mediante el test de Dunn, aplicando el ajuste de Bonferroni para controlar la tasa de error asociada a comparaciones múltiples (Dunn, 1964). las bandas pueden clasificarse en cuatro categorías según su comportamiento discriminante: aquellas que diferencian las tres clases, las que distinguen muestras contaminadas de no contaminadas y las que son sensibles específicamente a aflatoxina B1 o a contaminación mixta. Los resultados detallados pueden consultarse en Vranić et al. (2024b).

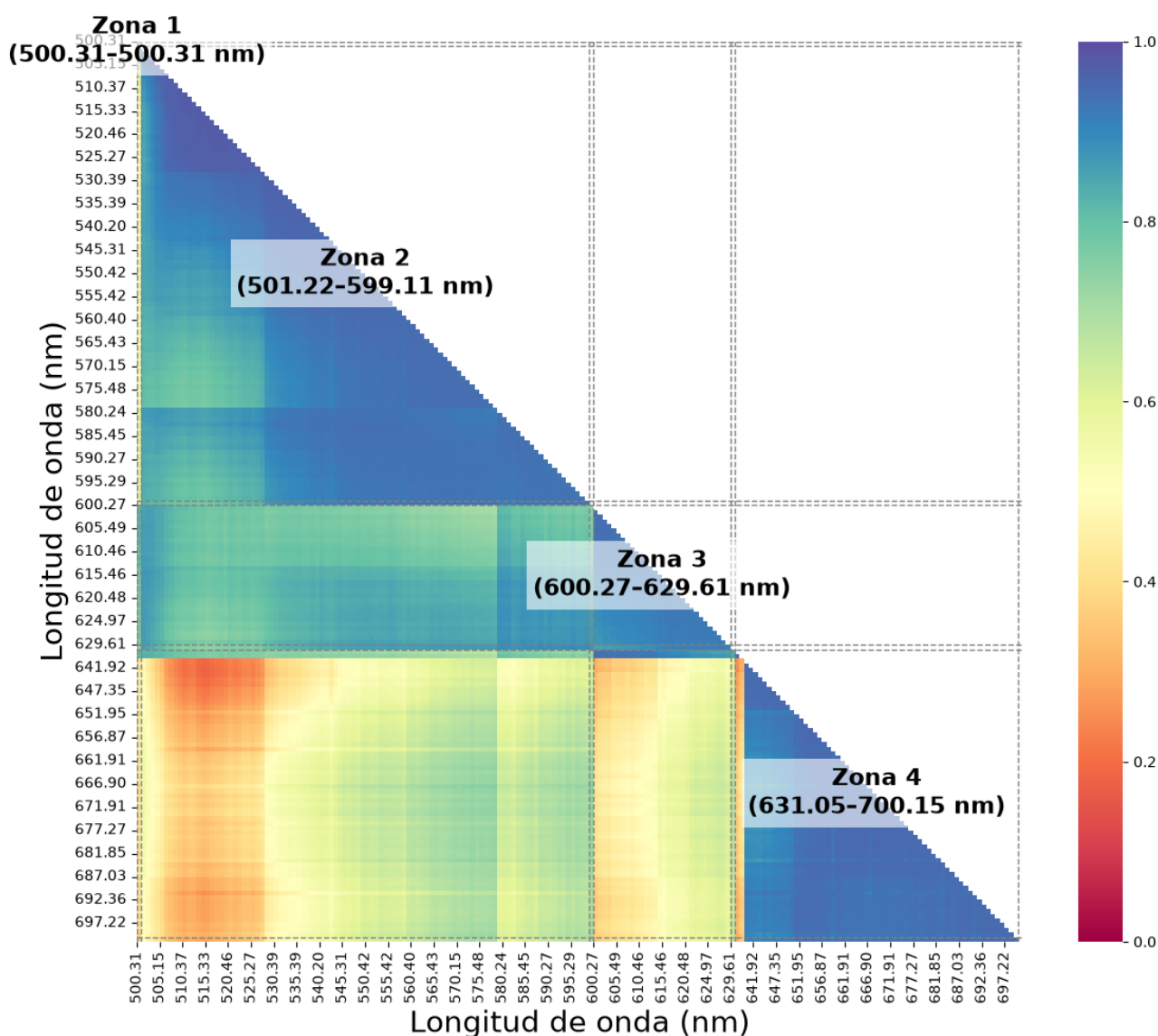


Figura 1: Matriz de correlación de Pearson de los datos de entrenamiento.

Dado el elevado tamaño muestral del conjunto de datos ($n > 5000$), se siguió la recomendación de complementar los valores p con una medida de tamaño del efecto. La magnitud de las diferencias se cuantificó mediante el índice ε_{KW}^2 , basado en el estadístico H de Kruskal–Wallis (King et al., 2018, pp. 384–387). Se observaron valores grandes ($\varepsilon_{KW}^2 \geq 0,14$) en el rango 640–700 nm, lo que indica una marcada separación entre las categorías de contaminación en esta región espectral. Asimismo, los índices de asimetría y curtosis resultaron elevados en este mismo rango.

El análisis de la matriz de correlación de Pearson (Figura 2) evidenció una elevada redundancia espectral, caracterizada por bloques contiguos de bandas con coeficientes de correlación cercanos a la unidad. El análisis de agrupamiento jerárquico basado en la distancia espectral ($1 - r$) confirmó estos patrones, identificando cuatro regiones homogéneas: una banda inicial aislada (≈ 500.31 nm), una región visible amplia (501.22–599.11 nm), una zona intermedia de transición (600.27–631.51 nm)

y un bloque final (640.14–700.15 nm) compuesto por bandas altamente correlacionadas entre sí. La coincidencia entre ambos enfoques refuerza la segmentación espectral observada y sustenta la reducción dimensional basada en la selección de bandas informativas. los hallazgos del tamaño del efecto (ε_{KW}^2) y las métricas de asimetría, justificando la selección de bandas representativas en lugar de la totalidad del espectro.

Regresión Penalizada en Grupos: Para evaluar la relevancia de las regiones espectrales en la discriminación de las tres categorías de la variable respuesta, las 194 bandas se organizaron en los cuatro grupos definidos previamente mediante el análisis de clustering. Esta estructura grupal se incorporó a los métodos de selección mediante regresión penalizada en grupos (Group LASSO y Group SCAD), implementados mediante algoritmos de descenso por grupos aplicados a regresión logística penalizada (Breheny & Huang, 2015). La importancia de cada banda se cuantificó mediante el valor absoluto de los coeficientes estimados. Debido a la alta colinealidad interna, los métodos no generaron selección de grupos completa, por lo que se emplearon percentiles (95 y 99) para identificar las bandas más relevantes tanto a nivel global como dentro de cada grupo. La Tabla 2 resume las importancias totales y promedio por grupo. Dado el tamaño desigual de los grupos, las mayores importancias individuales se concentran en los Grupos 2 y 4. Sin embargo, al considerar la importancia promedio, la mayor contribución corresponde a los Grupos 2 y 4 en el caso de LASSO en grupos, y a los Grupos 1 y 4 para SCAD en grupos.

		Grupo 1	Grupo 2	Grupo 3	Grupo 4
	Nº de variables	1	99	33	61
Lasso en grupos (grLasso)	Importancia	0.0661	5.3273	1.4841	5.3226
	Importancia Promedio	0.0661	0.5380	0.0450	0.0873
SCAD en grupos (grSCAD)	Importancia	0.1294	6.8339	1.8133	5.2551
	Importancia promedio	0.1290	0.0690	0.0549	0.0891

Tabla 2: Importancia de los grupos para los métodos de regresión penalizados

Teoría de la Información: La Información Mutua (MI) cuantifica cuánta dependencia existe entre una banda espectral y la clase. A diferencia de la correlación, captura relaciones no lineales, por lo que resulta especialmente útil para identificar qué bandas aportan información para discriminar entre SC, CB1 y CMIX. Permite obtener un ranking de relevancia de cada banda. La Entropía de Shannon mide la incertidumbre asociada a una variable aleatoria y, en este contexto, se utiliza como referencia para evaluar cuánta información aporta cada banda respecto de la variabilidad total del sistema. Las definiciones formales y fundamentos teóricos de ambas medidas pueden consultarse en Cover & Thomas (2006).

Método mRMR (Minimum Redundancy – Maximum Relevance): es una técnica de selección de características que busca identificar bandas espectrales que sean simultáneamente informativas para la clasificación y no redundantes entre sí. Evalúa la relevancia de cada banda mediante información mutua y penaliza aquellas que aportan información repetida. De este modo, selecciona bandas que contribuyen con información nueva y evita elegir bandas altamente correlacionadas. Se incluye en este trabajo porque permite reducir la dimensionalidad del espectro de manera coherente con su estructura interna, priorizando bandas representativas y evitando redundancia espectral. Es especialmente adecuado para datos hiperespectrales, donde múltiples bandas vecinas presentan alta colinealidad. El método fue introducido por Peng, Long y Ding (2005).

Métodos de Selección de Características Basados en la Familia Relief: son algoritmos de selección de características que estiman la importancia de cada banda espectral evaluando cómo cambian sus valores entre muestras cercanas de la misma clase y de clases distintas. Su fundamento es identificar características que contribuyen a separar correctamente instancias similares, capturando relaciones no lineales y patrones locales sin necesidad de entrenar un modelo predictivo. ReliefF extiende el algoritmo original para problemas multiclase y es robusto al ruido, mientras que MultiSURF elimina la necesidad de fijar un número de vecinos y define automáticamente un vecindario adaptado a la estructura del conjunto de datos, lo que mejora la estabilidad en contextos con alta colinealidad espectral. Se incluyen en este trabajo porque permiten obtener un ranking complementario de relevancia espectral basado en diferencias locales entre instancias, reforzando la identificación de bandas informativas más allá de los criterios globales como la información mutua o las penalizaciones (Urbanowicz et al., 2018).

Resultados

Variables seleccionadas: en la Tabla 3 se presentan las variables consideradas relevantes, por método. Cada uno de los modelos reducidos se evaluó mediante el algoritmo CART, optimizando sus hiperparámetros por validación cruzada. La métrica utilizada para comparar y seleccionar el mejor modelo fue la balanced accuracy.

Método	Variables Seleccionadas (nm)	Cantidad
grLasso95Global	503.3, 506.49, 649.04, 659.15, 664.38, 669.41, 673.26, 683.17, 686.46, 690.07	10
grLasso99Global	659.15, 669.41	2
grLasso95_porgrupo	500.31, 503.3, 505.15, 506.49, 523.28, 609.23, 629.61, 659.15, 669.41, 686.46	10
grLasso99_global	501.22, 506.49, 609.23, 659.15	4
grSCAD95Global	501.22, 503.3, 505.15, 506.49, 649.04, 659.15, 664.38, 669.41, 686.46, 690.07	10
grSCAD99Global	503.3, 659.15	2
grSCAD95_porgrupos	500.31, 501.22, 503.3, 505.15, 506.49, 531.39, 609.23, 612.31, 659.15, 669.41, 686.46	11
grSCAD99_porgrupo	500.31, 503.3, 609.23, 659.15	4
MI_95	640.14, 641.33, 641.92, 643.11, 644.15, 646.23, 688.06, 691.44, 694.35, 695.35	10
MI_99	640.14, 644.15	2
mMRM_95	561.19, 673.94, 675.29, 683.17, 689.09, 692.36, 694.35, 696.28, 697.22, 698.15	10
Relief_95	500.31, 529.28, 537.22, 552.34, 565.43, 573.18, 574.33, 575.48, 577.41, 578.18	10
Multisurf_95	641.33, 641.92, 643.11, 644.15, 688.06, 690.07, 691.44, 692.36, 693.36, 695.35	10
Multisurf_99	641.33, 688.06	2

Tabla 3: Variables seleccionadas (La notación utilizada indica el método, el percentil del umbral y si la selección se realizó a nivel global o por grupo.)

Análisis de los modelos evaluados: en la Tabla 4 se ordenaron los modelos según el valor de Balanced Accuracy del conjunto de test (BA_T) priorizando aquellos con mejor desempeño en datos no vistos.

Método	BA_T	BA_E	$\overline{BA}_{ECV}$	BA_{ECVsd}	$\overline{BA}_{TCV}$	BA_{TCVsd}	$BA_E - BA_T$	$\overline{BA}_{TCV} - BA_T$	$BA_{ECVsd} - BA_{TCVsd}$
grLasso95_porgrupo	0,9129	0,9856	0,9847	0,0004	0,9141	0,0047	0,073	0,001	-0,004
grSCAD95_Grupos	0,8967	0,9476	0,9507	0,0008	0,9135	0,0060	0,051	0,017	-0,005
Relief_95	0,8650	0,9164	0,9233	0,0028	0,8566	0,0033	0,051	-0,008	-0,001
grSCAD99_Grupo	0,8645	0,9119	0,9162	0,0030	0,8741	0,0045	0,047	0,010	-0,001
Multisurf_95	0,8458	0,9077	0,9131	0,0031	0,8470	0,0043	0,062	0,001	-0,001
MI_95	0,8425	0,9097	0,9131	0,0013	0,8517	0,0025	0,067	0,009	-0,001
grLasso95Global	0,8405	0,9018	0,9089	0,0021	0,8456	0,0055	0,061	0,005	-0,003
mMRM_95	0,8137	0,8728	0,8818	0,0039	0,8058	0,0048	0,052	-0,004	-0,002
grSCAD95Global	0,7994	0,8582	0,8637	0,0073	0,8049	0,0066	0,059	-0,008	-0,001
grLasso99_global	0,7921	0,8427	0,8472	0,0027	0,7946	0,0122			
MI_99	0,7638	0,9335	0,9298	0,0010	0,7356	0,0074			
Multisurf_99	0,7345	0,7824	0,7853	0,0026	0,7478	0,0078			
grLasso99Global	0,7167	0,7275	0,7268	0,0029	0,7178	0,0063			
grSCAD99Global	0,6500	0,6542	0,6549	0,0022	0,6476	0,0083			
Entropia_95	0,4587	0,5411	0,5523	0,0151	0,4496	0,0050			
Entropia_99	0,3999	0,4028	0,4042	0,0025	0,3989	0,0067			

Table 4: Evaluación modelos seleccionados

Se descartaron los modelos con $BA_T < 0.80$, ya que en este caso se consideran modelos con subajuste o con capacidad insuficiente de generalización. En todos los casos se observó una diferencia ≥ 0.05 entre el BA del entrenamiento directo (BA_E) y el promedio del BA obtenido en validación cruzada ($\overline{BA}_{ECV}$), lo que indica un sobreajuste moderado, esperable en modelos CART. Las desviaciones estándar del BA en validación cruzada fueron bajas para entrenamiento (BA_{ECVsd}) y para test (BA_{TCVsd}), reflejando buena estabilidad; los modelos más estables fueron **grLasso95_porgrupo** y **grSCAD95_Grupos**. La generalización de los modelos es adecuada: la diferencia entre el promedio del Balanced Accuracy del test en validación cruzada ($\overline{BA}_{TCV}$) y BA_T es menores a 0.02 en todos los casos. Asimismo, las desviaciones estándar del test (BA_{TCVsd}) son bajas (menores a 0.006), lo que indica que el desempeño en test es consistente entre particiones. Las

desviaciones estándar del entrenamiento en validación cruzada (BA_{ECVsd}) son bajas en general, lo que indica buena estabilidad del proceso de entrenamiento. Los modelos más estables son **grLasso95_porggrupo** y **grSCAD95_Grupos**, cuyos valores de desviación estándar son menores a 0.005. La diferencia entre las desviaciones estándar de entrenamiento y test ($BA_{ECVsd} - BA_{TCVsd}$) es muy pequeña, lo cual sugiere que todos los modelos presentan buena estabilidad global. El **grLasso95_porggrupo** es el modelo con mejor desempeño global, mayor estabilidad y mejor capacidad de generalización, por lo que se selecciona como el modelo óptimo para este conjunto de datos.

Conclusiones

A pesar de la presencia de multicolinealidad, los métodos basados en agrupamiento previo y selección dentro de grupos, en particular Lasso en grupos, logra reducir la dimensionalidad sin pérdida de precisión, demostrando que es posible reemplazar el espectro completo por un conjunto reducido de longitudes de onda informativas. Estos resultados evidencian que la selección de características no solo mejora la estabilidad y la capacidad de generalización de los modelos, sino que también constituye un paso clave hacia el desarrollo de sensores hiperspectrales más eficientes, interpretables y aplicables en entornos industriales de control de calidad alimentaria. Como línea de trabajo futura, se propone evaluar en profundidad el subconjunto de 10 bandas seleccionadas para determinar si es posible reducir aún más la dimensionalidad sin comprometer la capacidad discriminante del modelo. Este análisis permitirá avanzar hacia la identificación de un conjunto mínimo de bandas óptimas, potencialmente útil para el diseño de sensores espectrales compactos y de bajo costo.

Referencias

- Breheny, P., & Huang, J. (2015). Group descent algorithms for nonconvex penalized linear and logistic regression models. *Computational Statistics & Data Analysis*, 92, 160–171. <https://doi.org/10.1016/j.csda.2015.07.003>
- Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory* (2nd ed.). Wiley. <https://doi.org/10.1002/047174882X>
- Dunn, O. J. (1964). Multiple comparisons using rank sums. *Technometrics*, 6(3), 241–252. <https://doi.org/10.1080/00401706.1964.10490181>
- Jiang, H., Jiang, X., Ru, Y., Chen, Q., & Li, X. (2023). Hyperspectral imaging technology combined with feature selection algorithms for non-destructive detection of food quality: A review. *Trends in Food Science & Technology*, 133, 89–102. <https://doi.org/10.1016/j.tifs.2023.01.018>
- King, B. M., Rosopa, P. J., & Minium, E. W. (2018). *Statistical reasoning in the behavioral sciences* (7th ed.). John Wiley & Sons. ISBN: 978-1-119-37973-7.
- Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8), 1226–1238. <https://doi.org/10.1109/TPAMI.2005.159>
- Urbanowicz RJ, Meeker M, La Cava W, Olson RS, Moore JH. Relief-based feature selection: Introduction and review. *J Biomed Inform*. 2018 Sep; 85:189-203. doi: 10.1016/j.jbi.2018.07.014. Epub 2018 Jul 18. PMID: 30031057; PMCID: PMC6299836.
- Vranic, M. L., Cristos, D. S., Vorobioff, J., Martínez Vargas, S., Pesquero, N. V., & Delrieux, C. (2024). Análisis de imágenes hiperspectrales de maní contaminado con aflatoxinas. XXX Congreso Argentino de Ciencias de la Computación (CACIC), La Plata, Argentina. ISBN: 978-950-34-2428-5. Pages: 1098-1107
- Vranic, M. L., Delrieux, C., & Vorobioff, J. (2024). Identificación de Bandas Espectrales claves en Imágenes Hiperspectrales para la Detección de Aflatoxinas en Maní. AJEA (Actas De Jornadas Y Eventos Académicos De UTN), (AJEA 37). <https://doi.org/10.33414/ajea.1730.2024>