

Fecha de recepción: 06 de marzo de 2026, fecha de publicación en línea: junio de 2026.

Método Híbrido Kendall–Kolmogorov–Smirnov para la Selección de Características Orientada a la Evaluación Automática de la Calidad del Durazno

Humberto Huerta Flores, Yesenia Nohemí González Meneses, Blanca Estela Pedroza Méndez, y María Guadalupe Medina Barrera

Tecnológico Nacional de México – Instituto Tecnológico de Apizaco. San Andrés Ahuashuatepec, Municipio de Tzompantepec, Tlaxcala, C.P. 90491, México.
Autor de correspondencia: Humberto Huerta Flores (m24370005@apizaco.tecnm.mx)

Abstract- The present work presents a hybrid methodology for the selection of traits in peach samples from the municipality of Atltzayanca, (Tlaxcala). The methodology consists of the application of Kendall's Tau coefficient, which measures the correlation by ranges between a variable and the class. This allows you to detect if, when the value of a feature increases, the class also changes in an orderly fashion (high values belong to one class and low values to another). This method is useful for identifying monotonic relationships that are not necessarily linear. In a complementary way, the Kolmogorov–Smirnov test is used, which compares the distributions of the values of the same variable between two classes; If these distributions are different with a significance level $p = 0.05$, the attribute is considered to clearly distinguish between the two classes. Feature extraction is done using trained YOLO models, accurately trimming the fruits to measure their attributes. The hybrid method is applied to the entire set of images as a subset stabilized by bubble sorting; in both cases, results close to 100% are obtained, which demonstrates the high capacity of each characteristic to separate peaches of acceptable quality (CA) from those of unacceptable quality (NA). By using the hybrid method, not only correlation and distributions are considered, but the characteristic is also required to be consistent in both directions, reducing the risk of selecting misleading variables and allowing the selected characteristics to be analytically justified.

Keywords: feature selection, Kendall tau, Kolmogorov–Smirnov test, peach quality, hybrid method.

I. INTRODUCCIÓN

La selección de características es una etapa clave en los sistemas de análisis automático, especialmente cuando se dispone de pocas muestras y muchas variables. Utilizar todas las características disponibles puede aumentar el ruido, el costo de cómputo y el riesgo de sobreajuste. Por ello, se requieren criterios estadísticos que permitan identificar qué variables aportan información relevante para la clasificación. Este trabajo se basa en un enfoque de filtrado estadístico que combina dos criterios complementarios a diferencia de trabajos previos reportados en la literatura que emplean el coeficiente Tau de Kendall o la prueba de Kolmogorov–Smirnov de manera independiente como criterios de filtrado, el coeficiente tau de Kendall mide la relación monotónica entre cada característica y la clase sin suponer normalidad ni linealidad, y es robusto ante valores atípicos. Por otro lado, la prueba de Kolmogorov–Smirnov (K-S) compara la distribución de una característica entre dos clases y detecta diferencias significativas en su comportamiento. La combinación de ambos criterios proporciona una visión más completa: Kendall captura la tendencia ordenada atributo–

clase y K-S evalúa si las distribuciones por clase son realmente distintas.

La novedad del método radica en exigir significancia estadística simultánea tanto en la asociación monotónica como en la separabilidad de las distribuciones, integrando además un análisis estructural basado en el ordenamiento tipo burbuja para cuantificar la coherencia NA (calidad no aceptable) – CA (calidad aceptable). Esta combinación proporciona un mecanismo de selección de características más robusto, interpretable y particularmente adecuado para conjuntos de datos a analizar. El método propuesto se aplica a un conjunto de datos de duraznos que integra variables de color, textura, tamaño, circularidad y una lectura de sensor de gas. La idea es conservar solo aquellas características que, de forma simultánea, muestran una asociación monotónica significativa con la clase y una diferencia clara de distribución entre los niveles de calidad del fruto.

II. ESTADO DEL ARTE

A continuación, se presenta un estado del arte de artículos que emplean diversos métodos de selección de características

como Kendall, Kolmogorov–Smirnov, Chi2, etc., así como un artículo de visión por computadora donde se sustenta la toma de características y uno de sensores de gas.

Chi2-MI: hybrid feature selection for chronic kidney disease diagnosis

En este estudio propusieron un enfoque híbrido de selección de atributos denominado Chi2-MI, que integra la prueba chi-cuadrada y la información mutua para preseleccionar variables antes de entrenar distintos clasificadores. En el conjunto de datos de enfermedad renal crónica, el algoritmo Extra Trees alcanzó 98 % de precisión, 99 % de precisión (precision), 98 % de sensibilidad y 98 % de F1; otros modelos como AdaBoost, Random Forest y SVM obtuvieron entre 97-96 % de exactitud. Estos resultados demuestran que un filtro basado en χ^2 combinado con información mutua mejora la selección y el rendimiento de los clasificadores [1].

Benchmarking feature selection and feature extraction methods in biomedical data analysis

Este benchmark comparo distintos métodos de selección (incluyendo el test de Kolmogorov–Smirnov y Boruta) combinados con técnicas de extracción (PLS-DA, LDA, XGBoost). Para un conjunto de datos de cerebro, la combinación Boruta + PLS-DA alcanzó una exactitud balanceada de 89,5 % con 92,0 % de sensibilidad, 89,0 % de especificidad y 96,9 % de F1. Para el conjunto de mama, Boruta + LDA logró 85,9 % de exactitud balanceada y 92,7 % de sensibilidad. Estos resultados respaldan el uso de KS y Boruta como filtros supervisados previos a la extracción [2].

Variable selection of Kolmogorov–Smirnov maximization with a penalized surrogate loss

En este trabajo extendieron el método DMKS (Direct Maximization of KS) introduciendo un surrogate loss que permite optimizar la estadística KS e incorporar una penalización no cóncava (por ejemplo, SCAD) para selección de variables. Las simulaciones y un caso aplicado a credit scoring demostraron que el nuevo método (SKS) logra mejoras en la estadística KS y el área bajo la curva (AUC) frente a DMKS sin selección, reduciendo a la vez el número de variables. El artículo no reporta porcentajes de exactitud, pero enfatiza que la selección de variables mejora la interpretabilidad y no sacrifica la capacidad predictiva [3].

An optimized hybrid approach for feature selectin based on Chi-Square and PSOGWO

En este artículo desarrollaron un método híbrido donde la prueba chi-cuadrada filtra atributos irrelevantes y, a continuación, una combinación de Grey Wolf Optimizer y Particle Swarm Optimization (PSOGWO) refina el subconjunto final. La variante híbrida chi2-PSOGWO mejoró la exactitud hasta un 95,961 % y redujo el tiempo de ejecución respecto a PSOGWO solo. El trabajo demuestra que incorporar un filtro χ^2 antes de la metaheurística permite mejorar la precisión y reducir el costo computacional [4].

Interaction screening by Kendall's partial correlation for ultrahigh-dimensional survival data

En este artículo implementaron un estadístico basado en la correlación parcial de Kendall ponderada por probabilidad inversa de censura (PC-IPCW- τ) para detectar interacciones significativas en datos de supervivencia con miles de variables. En el estudio de validación, el método PC-IPCW- τ con penalización MCP obtuvo valores de prueba de verosimilitud significativamente mejores ($p=0,0027$) y mayor índice de concordancia (c-index) que otros métodos, y además identificó interacciones génicas relevantes. No reporta una tasa de exactitud porque se centra en métricas de supervivencia [5].

Network-adjusted Kendall's tau measure for feature screening

En este trabajo adoptaron el coeficiente de Kendall incorporando información de redes de coexpresión génica. Al ponderar las correlaciones de Kendall mediante PageRank, se obtiene un filtro que prioriza genes centrales. En un conjunto de datos de cáncer, esta variante alcanzó c-index de hasta 0,6178, frente a valores de 0,5915–0,5938 obtenidos por métodos sin ajuste. La mejora demuestra que añadir la estructura de red a Kendall mejora la predicción en datos de supervivencia [6].

Kendall transformation brings a robust categorical representation of ordinal data

En este estudio propusieron la transformación de Kendall para convertir variables ordinales en representaciones categóricas basadas en comparaciones par a par. La técnica ofrece una forma robusta de calcular medidas de información (como la información mutua) y aplicar filtros en datos ordinales o discretizados. Los autores destacan que la transformación reduce el efecto de interacciones complicadas y que fue la solución más versátil y robusta entre las evaluadas, nunca inferior a los mejores métodos de comparación. El artículo no muestra métricas de exactitud, sino que argumenta la utilidad teórica de la transformación [7].

Fruits and vegetables quality evaluation using computer vision

En este artículo revisaron exhaustivamente los sistemas de visión por computadora para la inspección de frutas y hortalizas. Describe las etapas de preprocesamiento, segmentación, extracción de características y clasificación y señala que los descriptores más utilizados para evaluar la calidad son el color, la textura, el tamaño, la forma y la detección de defectos superficiales. La revisión destaca que la combinación de diferentes espacios de color y perspectivas puede mejorar el rendimiento de los clasificadores, pero no reporta porcentajes de exactitud específicos porque sintetiza muchos estudios [8].

Fruits and vegetables quality evaluation using computer vision: A review

En este artículo analizaron el uso de una nariz electrónica para detectar infecciones tempranas en durazno. Mediante análisis de discriminante lineal sobre señales de sensores de gas, el sistema clasificó los frutos en sanos, con decaimiento inicial y en descomposición media, obteniendo tasas de reconocimiento superiores al 97 % y 100 % de exactitud para infecciones en etapas tempranas. Este resultado demuestra la relevancia de incorporar lecturas de sensores de gas como una variable más en sistemas de clasificación [9].

A partir del análisis de los trabajos revisados, se observa que la mayoría de los enfoques utilizan métodos de selección de características basados en un único criterio estadístico o combinaciones con técnicas metaheurísticas y de regularización. Sin embargo, no se identifican propuestas que integren simultáneamente un criterio de asociación monótona, un criterio de divergencia de distribuciones y un análisis estructural de ordenamiento en un mismo marco metodológico. En este sentido, el método híbrido Kendall–K-S propuesto busca cubrir este vacío mediante un esquema de doble evaluación estadística.

III. METODOLOGÍA

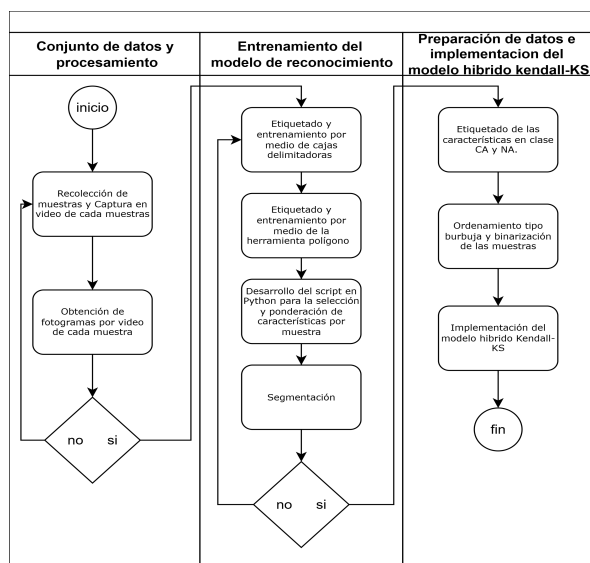


Figura 1. Diagrama de flujo de la metodología propuesta.

IV. EXPERIMENTACIÓN

Conjunto de datos y preprocesamiento

Recolección de muestras y captura en video de las muestras. El conjunto de datos utilizado se compone de duraznos recolectados en Atltzayanca, municipio del estado de Tlaxcala, México, bajo condiciones controladas. Obteniendo un conjunto de datos de 100 muestras.

En la figura 2 se observan las muestras recolectadas y el área controlada, donde se extrajeron los videos representativos de los duraznos.

Obtención de fotogramas por video de cada muestra

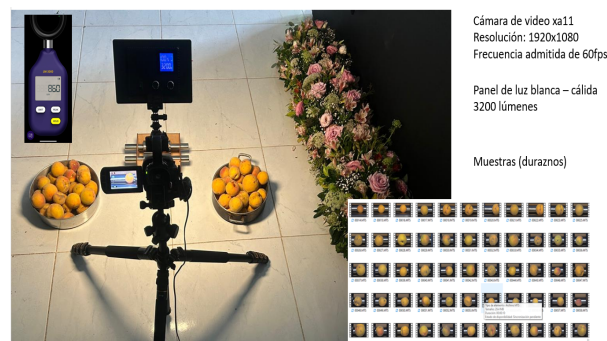


Figura 2. Recolección de datos.

Se realizó un script en Python donde se ingresa el video de la muestra, se procesa y se obtienen como resultado 10 fotogramas por video, obteniendo un total de 1000 fotogramas de los 100 videos procesados.

Entrenamiento del modelo de reconocimiento.

Etiquetado y entrenamiento por medio de cajas delimitadoras Para el reconocimiento del fruto se entrenó un modelo YOLO (You Only Look Once) etiquetando manualmente las cajas delimitadoras con la aplicación Make Sense. El conjunto de datos estuvo compuesto por 500 fotogramas las cuales fueron etiquetadas y divididas en tres subconjuntos 70% muestras para el entrenamiento del modelo 20% para la validación y 10% para la prueba, dando una precisión: 0.994 ($\approx 99.4\%$) y un recall: 1.000 ($\approx 100\%$). Una vez entrenado el modelo de detección por medio de cajas delimitadoras fue posible centrar las imágenes de durazno mediante un script en Python ya que la toma del video es captada en 2 rodillos giratorios y el durazno tiende a cambiar de posición.

En la figura 3 podemos observar el proceso de etiquetado de cada muestra utilizando la herramienta bounding box (caja delimitadora) en Make Sense.

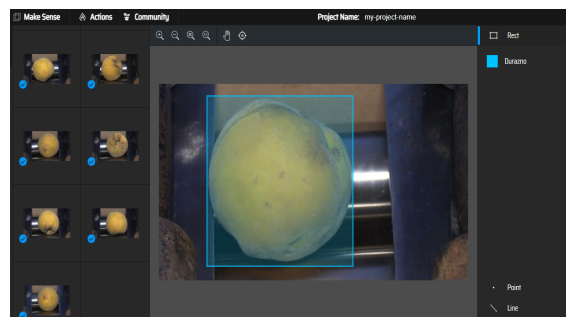


Figura 3. Entrenamiento mediante cajas delimitadoras.

Etiquetado y entrenamiento por medio de la herramienta polígono

En la figura 4 se aplica un enmascarado por polígonos para segmentar con mayor precisión el durazno y eliminar la mayor parte del fondo. Para esta etapa se utilizó el conjunto de 500 imágenes restante, las cuales fueron etiquetadas y divididas en tres subconjuntos 70% para el entrenamiento del modelo 20% para la validación y 10% para la prueba, bajo este esquema se tuvo una precisión: 0.999 ($\approx 99.9\%$) y un Recall: 1.000 ($\approx 100\%$).

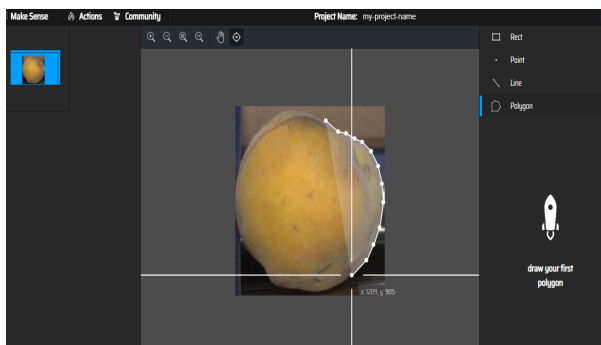


Figura 4. Entrenamiento por medio de polígonos.

Segmentación

Una vez detectando la forma del durazno, es posible quitar el fondo y la mayor parte del ruido mediante el reconocimiento y la segmentación de la máscara polígono. Se desarrolló un código en Python que implementa un proceso automatizado de segmentación de objetos basado en el modelo YOLO previamente entrenado con capacidad de segmentación. Este procedimiento tiene como objetivo aislar el objeto de interés en cada imagen en este caso, el fruto eliminando completamente el fondo y generando una representación con transparencia que facilite el análisis posterior de características visuales. En la figura 5 se observa el resultado de la segmentación de la muestra de durazno una vez procesado por el script.



Figura 5. Segmentación de un fotograma de la muestra de durazno.

Desarrollo del script en Python para la selección y ponderación de características por muestra.

En un Script en Python utilizando el sistema de segmentación previamente desarrollado se ingresaron los videos de las 100 muestras de durazno obteniendo 3 fotogramas representativos y se seleccionaron diferentes características por muestra, la selección de estos fotogramas se basó en estos criterios: adecuada calidad de imagen (no borrosos o pixelados), visibilidad completa del durazno y que no se repitan.

Se extrajeron características como el color, el tamaño, la circularidad (forma) y la textura, adicionalmente se incorporó una medida asociada al grado de madurez de la fruta obtenida mediante un sensor (TGS2602) que mide los compuestos volátiles que puede desprender el fruto y con ello poder predecir parcialmente el grado de madurez del fruto. Para la forma: el algoritmo no mide si es un círculo perfecto, sino qué tan redondeada es la forma del durazno en la imagen ponderando valores entre 0 y 1.

Para el color: Analiza los colores dominantes en varios fotogramas del durazno usando K-Means, promedia los porcentajes entre fotogramas, agrupa esos colores en categorías (café claro/oscuro, amarillo, verde, negro) y con base en umbrales el durazno es ponderado entre 0 y 1.

Tamaño: Este algoritmo no mide el tamaño físico en centímetros directamente, sino el: tamaño proyectado del durazno en la imagen, medido en píxeles, usando el área del contorno segmentado, ponderando de 0 a 1.

La textura del durazno: se evalúa mediante el uso del operador Laplaciano aplicado sobre la región segmentada del fruto. Este operador resalta variaciones locales de intensidad, permitiendo identificar zonas con mayor presencia de manchas, irregularidades o cambios abruptos de tono ponderando cada durazno de 0 a 1.

Para la madurez, el durazno es encapsulado en un recipiente de cristal junto con el sensor (TGS2602) pronosticando valores en consola, de estos valores es ponderado cada durazno entre 0 y 1. Cabe resaltar que las decisiones de ponderación asignadas a cada muestra fueron supervisadas por un experto agricultor.

Preparación de datos e implementación del modelo híbrido Kendall-KS

Etiquetado de las características en clases CA Y NA

Con base a estas características cada muestra fue etiquetada en dos clases, CA (calidad aceptable) mayores a 0.5 y NA (calidad no aceptable) menores a 0.5. Se eliminó información vacía y se verificó la correspondencia entre los registros de clase y los valores numéricos.

En la tabla 1 se muestra una matriz con los pesos pronosticados de las 100 muestras y sus clases por cada característica.

Tabla 1. Matriz con los pesos de cada característica y su respectiva clase.

Muestra	tamaño	forma	textura	color	sensor(madurez)
1	NA	CA	NA	CA	CA
	0.49	0.86	0.47	0.72	0.8
	CA	CA	CA	NA	CA
2	0.53	0.79	0.76	0.46	0.79
	CA	CA	NA	CA	CA
3	0.9	0.8	0.48	0.88	0.79
	CA	NA	NA	NA	CA
4	0.68	0.5	0.46	0.49	0.78
	CA	CA	NA	CA	CA
5	0.88	0.55	0.47	0.7	0.78
	CA	NA	CA	CA	CA
6	0.79	0.45	0.63	0.58	0.77
	CA	CA	CA	CA	CA
7	0.77	0.88	0.58	0.61	0.77
	CA	NA	NA	NA	CA
8	0.52	0.46	0.49	0.49	0.76
	NA	CA	CA	CA	CA
9	0.5	0.8	0.75	0.59	0.76
	CA	CA	CA	CA	CA
10	0.91	0.61	0.69	0.85	0.75
	CA	NA	CA	CA	CA
11	0.54	0.49	0.6	0.76	0.75
	CA	CA	CA	CA	CA
12	0.79	0.85	0.91	0.52	0.74
	NA	CA	CA	NA	CA
13	0.48	0.88	0.89	0.5	0.74
	NA	NA	CA	CA	CA

Ordenamiento tipo burbuja y binarización de las muestras

Antes de aplicar el filtro híbrido Kendall con K-S, se ordenaron los elementos por clase y se contó el número de intercambios necesarios para separar ambas clases (similar a un ordenamiento burbuja). Este número de movimientos

sirvió como un índice de separabilidad: a menor cantidad de movimientos, mayor capacidad discriminante del atributo. La tabla 2 presenta las características ordenadas y los saltos que se ejecutan por muestra para el ordenamiento.

Tabla 2. Matriz ordenada.

Muestra	tamaño	forma	textura	color	sensor[madurez]	
1	0.49	NA	0.47	0.86	0.72	0.8
	NA	CA	CA	CA	CA	CA
2	0.46	0.53	0.79	0.76	0.79	
	NA	CA	CA	CA	CA	CA
3	0.48	0.9	0.8	0.88	0.79	
	NA	NA	NA	CA	NA	CA
4	0.5	0.46	0.49	0.68	0.78	
	NA	CA	CA	CA	CA	CA
5	0.47	0.88	0.55	0.7	0.78	
	NA	CA	CA	CA	CA	CA
6	0.45	0.79	0.63	0.58	0.77	
	CA	CA	CA	CA	CA	CA
7	0.77	0.88	0.58	0.61	0.77	
	NA	NA	NA	CA	CA	CA
8	0.46	0.49	0.49	0.52	0.76	
	NA	CA	CA	CA	CA	CA
9	0.5	0.8	0.75	0.59	0.76	
	CA	CA	CA	CA	CA	CA
10	0.91	0.61	0.69	0.85	0.75	
	NA	CA	CA	CA	CA	CA
11	0.49	0.54	0.6	0.76	0.75	
	CA	CA	CA	CA	CA	CA
12	0.79	0.85	0.91	0.52	0.74	

Muestra	Salts_NA_CA
1	1
2	3
3	2
4	3
5	2
6	1
7	0
8	3
9	0
10	0
11	1
12	0

En la tabla 3 se observa una binarización de cada característica por cada muestra.

Tabla 3. Matriz con la implementación de la columna binaria.

Muestra	Característica	Valor	Clase_texto	Clase_bin	Salto	NA_CA
1	tamaño	0.49	NA	0	1	1
1	forma	0.86	CA	1	1	1
1	textura	0.47	NA	0	1	1
1	color	0.72	CA	1	1	1
1	sensor(madure	0.8	CA	1	1	1
2	tamaño	0.53	CA	1	3	3
2	forma	0.79	CA	1	3	3
2	textura	0.76	CA	1	3	3
2	color	0.46	NA	0	3	3
2	sensor(madure	0.79	CA	1	3	3
3	tamaño	0.9	CA	1	2	2
3	forma	0.8	CA	1	2	2
3	textura	0.48	NA	0	2	2
3	color	0.88	CA	1	2	2
3	sensor(madure	0.79	CA	1	2	2
4	tamaño	0.68	CA	1	3	3
4	forma	0.5	NA	0	3	3
4	textura	0.46	NA	0	3	3
4	color	0.49	NA	0	3	3
4	sensor(madure	0.78	CA	1	3	3
5	tamaño	0.88	CA	1	2	2
5	forma	0.55	CA	1	2	2
5	textura	0.47	NA	0	2	2
5	color	0.7	CA	1	2	2
5	sensor(madure	0.78	CA	1	2	2

Implementación del modelo hí-brido Kendall-KS

La metodología propuesta no se limita a aplicar pruebas estadísticas de manera independiente, sino que formula un criterio híbrido que exige coherencia estructural y consistencia estadística simultánea para validar la relevancia de cada característica. En la tabla 4 se aplica el modelo híbrido Kendall Y K-S sin ordenamiento en todo el conjunto de datos (columnas_full) obteniendo los siguientes valores:

Modelo híbrido aplicado sin saltos

Tabla 4. Resultados utilizando las 100 muestras.

Característica	Kendall_tau_full	p_Kendall_full	KS_full	p_KS_full
tamaño	0.644376161	4.95424E-28	1	1.0008E-50
forma	0.687230423	1.90418E-31	1	1.9644E-55
textura	0.677917262	1.29633E-30	1	2.7511E-54
color	0.651883634	1.34169E-28	1	1.6067E-51
sensor(madurez)	0.584634569	2.36293E-23	1	6.6033E-44

Donde; Kendall_tau_full: es la fuerza de la relación entre el valor de la característica y la clase (0 = NA, 1 = CA). Valor

0.68 aproximadamente es alto (máximo posible 1), Cuando el valor de esta característica aumenta, casi siempre pasa de NA a CA de manera ordenada.

[illegible]

KS_full. Es el estadístico de Kolmogorov–Smirnov entre la distribución de valores en NA vs CA, los resultados se encuentran en 1 (y uno ~ 0.99). El valor de KS está entre 0 y 1 donde: 0 distribuciones idénticas y 1 distintas al máximo (no se traslapan). Así que KS = 1 significa: los valores de esta característica para NA y CA están prácticamente separados; casi no se traslapan.

p_KS_full. Probabilidad de que esa diferencia tan grande (KS) sea por azar. Valores tipo 1.0E-50, 6.6E-44 de nuevo casi cero. La diferencia entre NA y CA para esta característica es fuertemente significativa

Podemos observar que para todas las características de las muestras por separado existe una gran discriminación entre CA y NA.

En la tabla 5 se aplica el híbrido Kendall y K-S sobre el subconjunto de cinco muestras con baja cantidad de movimientos en la fase estructural (columnas `_filtrado`), verificando que las características mantuvieran su capacidad discriminante incluso en las muestras con mejor organización NA-CA.

Tabla 5. Resultados con las muestras de mejor organización.

Kendall_tau_filtrado	p_Kendall_filtrado	KS_filtrado	p_KS_filtrado	Relevancia_hibrida
0.698338935	1.89757E-20	1	1.31544E-34	✓ Fuerte (full + pocos saltos)
0.683927963	1.07649E-19	1	1.34115E-33	✓ Fuerte (full + pocos saltos)
0.660400928	1.61519E-18	1	5.49065E-32	✓ Fuerte (full + pocos saltos)
0.455466465	1.21221E-09	1	2.07262E-18	✓ Fuerte (full + pocos saltos)
0.22791745	0.003144423	1	6.60893F-06	✓ Fuerte (full + pocos saltos)

Aplicado a muestras con pocos saltos.

Donde:

Kendall_tau_filtrado y p_Kendall_filtrado, Tau es aproximado a 0.22–0.69, significando que sigue siendo alto, el p-values en el orden de E-20, E-19, E-18 y E-09 son casi cero. Si me quedo solo con las muestras que tienen pocos saltos, la relación valor–clase sigue siendo muy fuerte y significativa.

KS_filtrado y p_KS_filtrado. KS_filtrado vuelve a salir 1, máxima diferencia entre NA y CA en ese subconjunto.

p_KS_filtrado reportan rangos muy pequeños (E-34, E-06, etc.) lo que da más significancia a los resultados obtenidos.

Relevancia_hibrida. La regla propuesta fue: Si Kendall y K-S son significativos en el conjunto de datos completo y siguen siendo significativos en el subconjunto con pocos saltos, se marca con una etiqueta y con el texto Fuerte (full + pocos saltos). Como podemos observar la tabla reporta en todas las características este tipo de etiqueta interpretando que pasan el filtro híbrido discriminando muy bien NA de CA. Esa capacidad discriminante se mantiene incluso si solo usamos las muestras con mejor estructura (pocos saltos).

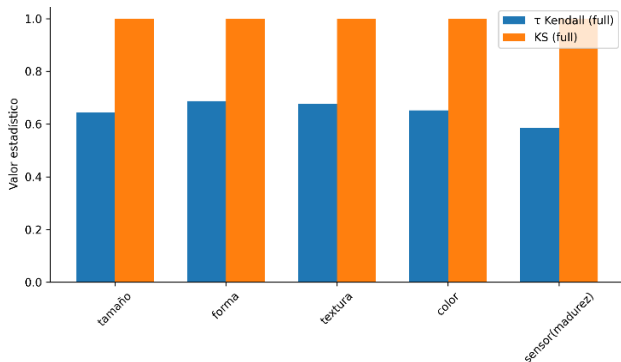


Figura 6. Resumen estadístico por característica.

V. RESULTADOS

Los cinco atributos evaluados (tamaño, forma, textura, color y señal del sensor) mostraron una alta asociación monotónica con la clase (CA/NA) y diferencias máximas en la distribución entre clases:

$Kendall_ \tau_ full \approx 0.68$: indica una asociación monotónica fuerte entre el valor de la característica y la pertenencia a la clase

$p_ Kendall_ full \approx 10^{-28}$: p -valores prácticamente cero demuestran que la relación observada no se debe al azar. $KS_ full \approx 1$: las distribuciones de NA y CA para cada atributo están casi completamente separadas, evidenciando una capacidad discriminante extrema.

$p_ KS_ full \approx 10^{-44}$ refuerza la significancia de la diferencia entre distribuciones.

Al analizar únicamente las muestras con pocos saltos se observaron patrones similares: τ permaneció entre 0.22 y 0.69 y los valores de KS siguieron siendo prácticamente 1, con p -valores en el orden de 10^{-18} a 10^{-06} . Esto confirma que los atributos mantienen su relevancia incluso en las muestras con mejor organización NA–CA.

En la figura 6 se presenta el resumen estadístico obtenido con el método híbrido Kendall–K-S para las cinco características analizadas (tamaño, forma, textura, color y señal del sensor de madurez). En todos los casos se observó un coeficiente τ de Kendall cercano a 0.7, lo que indica una asociación monotónica fuerte entre el valor del atributo y la clase asignada (NA/CA). De forma complementaria, el estadístico K-S fue prácticamente igual a 1 para todas las variables, lo que evidencia diferencias máximas entre las distribuciones de valores correspondientes a frutos no aceptables (NA) y aceptables (CA).

En la figura 7, muestra las medias de cada característica separadas por clase. En las cinco variables se aprecia el mismo patrón: los duraznos clasificados como NA presentan valores medios bajos (~ 0.48), mientras que los duraznos CA presentan valores medios elevados (~ 0.72). Esta separación consistente entre medias confirma visualmente la capacidad discriminante de los atributos y es coherente con los resultados de Kendall y K-S. En conjunto, estos resultados indican que tamaño, forma, textura, color y la señal del sensor de madurez son características relevantes y robustas para diferenciar entre frutos aceptables y no aceptables dentro del método propuesto.

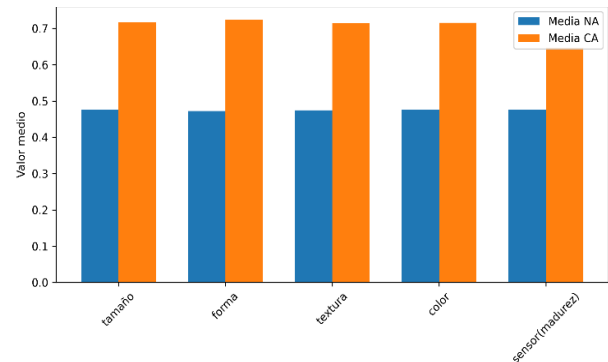


Figura 7. Media de cada característica por clase (NA vs CA).

VI. CONCLUSIONES

El presente trabajo propuso y evaluó un método híbrido para analizar la capacidad discriminante de cinco atributos (tamaño, forma, textura, color y señal del sensor de madurez) en la clasificación de duraznos en dos categorías: calidad aceptable (CA) y calidad no aceptable (NA). El enfoque combina una etapa estructural, basada en el número de movimientos necesarios para ordenar las etiquetas NA–CA dentro de cada muestra, con una etapa estadística, que integra el coeficiente de correlación por rangos de Kendall (τ) y la prueba de Kolmogorov–Smirnov (K-S) aplicada a las distribuciones de valores por clase.

En la fase estructural, el conteo de “saltos” mediante un algoritmo tipo burbuja permitió cuantificar qué tan coherente era el patrón de clases dentro de cada muestra. Un número reducido de movimientos indica que las características de esa muestra ya se encontraban alineadas con la separación NA–CA, mientras que un número alto de saltos sugiere mayor desorden o ruido estructural. Esta información se utilizó como filtro de calidad: las muestras con menos saltos se consideraron más confiables desde el punto de vista de la organización NA–CA.

En la fase estadística, se construyó una tabla con todas las observaciones posibles (muestra–característica–valor–clase) y se evaluó, para cada atributo, su capacidad de distinguir entre NA y CA. Los resultados fueron consistentes y muy robustos: para las cinco características se obtuvo un τ de Kendall cercano a 0.68, con valores de p extremadamente pequeños (del orden de 10^{-28}), lo que evidencia una asociación monotónica fuerte entre el aumento del valor del atributo y la probabilidad de pertenecer a la clase CA. De manera complementaria, el estadístico K-S fue prácticamente igual a 1 en todos los casos (con p también del orden de 10^{-06} a 10^{-18}), indicando que las distribuciones de valores para NA y CA son máximamente diferentes y apenas presentan traslape.

Al repetir el análisis únicamente en el subconjunto de muestras con baja cantidad de saltos, los resultados se mantuvieron: los valores de τ y de K-S siguieron siendo altos y estadísticamente significativos. Esto se refleja en la columna de “Relevancia híbrida”, donde los cinco atributos se clasifican como “Fuerte (full + pocos saltos)”, es decir, no solo discriminan bien en el conjunto completo de datos, sino que también conservan su poder discriminante cuando se consideran únicamente las muestras estructuralmente más

ordenadas. Las gráficas de barras de medias NA vs CA confirman visualmente este comportamiento: en todos los atributos los duraznos NA se concentran en valores bajos (~ 0.48), mientras que los CA se agrupan en valores altos (~ 0.72).

En conjunto, el método híbrido propuesto muestra que tamaño, forma, textura, color y sensor de madurez son atributos altamente relevantes y estables para diferenciar duraznos aceptables y no aceptables. La combinación de un filtro estructural basado en saltos con la validación estadística Kendall–K-S permite no solo identificar características discriminantes, sino también verificar que su desempeño se mantiene en las muestras con mejor organización NA–CA. Estos resultados no solo confirman la alta capacidad discriminante de los atributos evaluados, sino que validan la efectividad del criterio híbrido de doble consistencia, demostrando que la integración de asociación monotónica y separabilidad de distribuciones reduce el riesgo de seleccionar variables engañosas. Este enfoque puede servir como base metodológica para el diseño de sistemas automáticos de evaluación de calidad, y es extensible a otros productos agrícolas y a conjuntos de datos donde sea importante integrar información estructural y estadística en la selección de atributos.

REFERENCIAS

- [1] S. K. Dey, K. M. M. Uddin, H. M. H. Babu, M. M. Rahman, A. Howlader, and K. A. Uddin, “Chi2-MI: A hybrid feature selection based machine learning approach in diagnosis of chronic kidney disease,” *Intelligent Systems with Applications*, vol. 16, art. 200144, 2022.
- [2] J. Labory, E. Njomgue-Fotso, and S. Bottini, “Benchmarking feature selection and feature extraction methods to improve the performances of machine-learning algorithms for patient classification using metabolomics biomedical data,” *Computational and Structural Biotechnology Journal*, vol. 23, pp. 1274–1287, Dec. 2024.
- [3] X. Lin and F. Fang, “Variable selection of Kolmogorov–Smirnov maximization with a penalized surrogate loss,” *Computational Statistics & Data Analysis*, vol. 195, art. 107944, 2024.
- [4] A. Abdo, R. Mostafa, and L. Abdel-Hamid, “An optimized hybrid approach for feature selection based on chi-square and particle swarm optimization algorithms,” *Data*, vol. 9, no. 2, art. 20, 2024.
- [5] J.-H. Wang and Y.-H. Chen, “Interaction screening by Kendall’s partial correlation for ultrahigh-dimensional data with survival trait,” *Bioinformatics*, vol. 36, no. 9, pp. 2763–2769, May 2020.
- [6] J.-H. Wang and Y.-H. Chen, “Network-adjusted Kendall’s tau measure for feature screening with application to high-dimensional survival genomic data,” *Bioinformatics*, vol. 37, no. 15, pp. 2150–2156, Aug. 2021.
- [7] M. B. Kursa, “Kendall transformation brings a robust categorical representation of ordinal data,” *Scientific Reports*, vol. 12, art. 8341, May 2022.
- [8] A. Bhargava and A. Bansal, “Fruits and vegetables quality evaluation using computer vision: A review,” *Journal of King Saud University – Computer and Information Sciences*, vol. 33, no. 3, pp. 243–257, 2021.
- [9] A. Martínez, A. Hernández, P. Arroyo, J. Lozano, M. de G. Córdoba, and A. Martín, “Early detection of *Monilinia laxa* in yellow-fleshed peach using a non-destructive E-nose approach,” *Foods*, vol. 14, no. 18, art. 3155, Sept. 2025.



Humberto Huerta Flores obtuvo la licenciatura en ingeniería en Procesos y Operaciones Industriales por la Universidad Tecnológica de Tlaxcala, México, en 2021. Actualmente está cursando la maestría en Sistemas Computacionales en el Tecnológico Nacional de México Campus Apizaco.



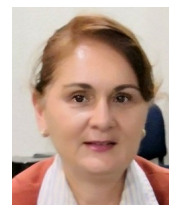
computadora.

Yesenia Nohemí González Meneses obtuvo en el año 2021 el grado de Doctor en Ingeniería del Lenguaje y del Conocimiento por parte de la Universidad Autónoma de Puebla. Actualmente trabaja en la División de Estudios de Posgrado del Tecnológico Nacional de México Campus Apizaco. Sus áreas de interés incluyen reconocimiento automático de emociones, aplicaciones de IA, ingeniería de software e interacción humano



hasta diciembre del 2029. Sus áreas de investigación incluyen modelos de lógica difusa, tutoriales inteligentes y aplicaciones de internet de las cosas.

Blanca Estela Pedroza Méndez obtuvo el grado de doctor en la Facultad de Ciencias de la Electrónica de la Benemérita Universidad Autónoma de Puebla (BUAP) en 2018. Actualmente es docente investigador del área de posgrado y de licenciatura en el Tecnológico Nacional de México, campus Apizaco. Tiene el reconocimiento como Perfil PRODEP, hasta el 2026 y el reconocimiento en el SNII, en el nivel 1,



Actualmente, cuenta con el reconocimiento como candidata en el Sistema Nacional de Investigadores del CONAHCYT, además del reconocimiento al Perfil Deseable y de ser miembro del cuerpo académico “Sistemas de Información”, ambos reconocidos por PRODEP. Sus áreas de trabajo son: Gestión y Desarrollo de Proyectos de Software, Automatización de Procesos, Interfaces Humano Computadora y Reconocimiento de Patrones.

María Guadalupe Medina Barrera es Doctora en Planeación Estratégica y Dirección de Tecnología por la Universidad Popular Autónoma del Estado de Puebla (UPAEP). Realizó estudios de Maestría en Ciencias en Ciencias Computacionales en el Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET) y estudios de Licenciatura en Informática en el Instituto Tecnológico de Tepic, Nayarit, México.