**Métodos de selección de variables y características para la construcción de modelos predictivos de la calidad de alimentos usando imágenes THz.**

**Quiroz Burga, Luis Augusto**

**Universidad Privada del Norte**

**Resumen**

Resumen.

*Palabras claves: Imágenes THz, selección de variables, selección de características, Machine Learning, modelos predictivos, clasificación, calidad de alimentos.*

**Methods for variable and feature selection through the analysis in THz images for the construction of predictive models of food quality.**

**Quiroz Burga, Luis Augusto**

**Universidad Privada del Norte**

**Abstract (en inglés)**

Abstract.

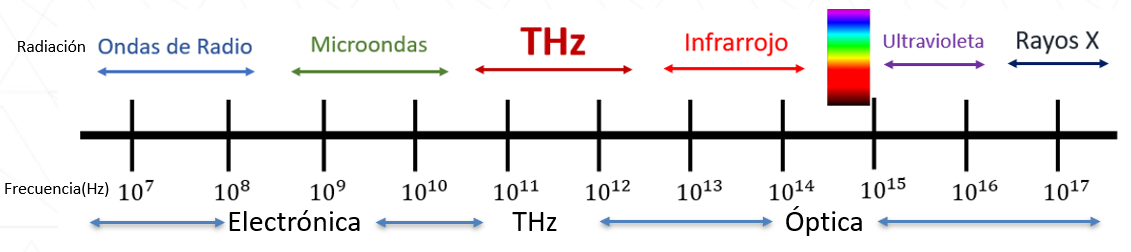
*Keywords: THz images, variable selection, feature selection, Machine Learning, predictive models, classification, food quality.*

**Introducción**

Se conoce que a partir de los inicios del presente milenio, existen investigaciones en selección de variables que exploran dominios con cientos de miles de variables o características (Guyon & Elisseeff, 2003) y además, estudios en analítica de datos abarcan cantidades gigantescas de datos donde se llega a tener millones de datos que existen en miles de dimensiones que representan miles de grupos (McCallum, Nigam, & Ungar, 2000). Estas inmensas cantidades de datos requieren, en muchos estudios y aplicaciones, ser pre procesada extrayendo subconjuntos de variables y características relevantes que aporten a la resolución de problemas específicos.

El análisis de imágenes hiperespectrales consiste en analizar imágenes compuestas por varias bandas espectrales de información a través de todo el espectro electromagnético (Roman-Gonzalez & Vargas-Cuentas, 2013, p.). Esta técnica de análisis se a utilizado ampliamente en aplicaciones de diferentes segmentos tecnológicos, principalmente en el control de calidad de alimentos (Condezo-Hoyos & Castro, 2018; Gowen, O’Donnell, Cullen, Downey, & Frias, 2007; D. Liu, Sun, & Zeng, 2014); específicamente en la determinación de la calidad de frutas y verduras (Nicolaï et al., 2007), la determinación de enfermedades, infecciones e índices importantes en productos agrícolas (Castro, Oblitas, Maicelo, & Avila-George, 2011; Fu, Meacham-Hensold, Guan, & Bernacchi, 2019; Hue et al., 2014), predicción de la calidad de productos agropecuarios (Li, Shan, & Peng, 2011), además de otras aplicaciones como el diagnóstico de enfermedades, cirugía y medicina (G. Lu & Fei, 2014), telemetría (Crowley et al., 2006), etc. Sin embargo, la utilización de imágenes hiperespectrales que integran espectroscopía e imágenes convencionales para obtener información espacial y espectral de un objeto se reduce al análisis de la superficie el alimento (Gowen et al., 2007), impidiendo analizar su interior e imposibilitando la detección de adulteraciones, la presencia de contaminantes, la distribución de compuestos internos, entre otras características adicionales. Se conoce el uso de instrumentos que usan la tecnología de pulso en el rango de los terahercios en diversas aplicaciones. Entre las principales aplicaciones se encuentra el análisis interno de objetos (Fukunaga, 2016) utilizando espectroscopía e imágenes, logrando detectar exitosamente distintas propiedades como residuos de antibióticos (Redo-Sanchez et al., 2011), melamina (Baek, Lim, & Chun, 2014) y la inspección de alimentos (Yan, Ying, Zhang, & Yu, 2006) para otros tipos de diagnósticos.

Con la necesidad de métodos de inspección y predicción de la calidad de alimentos con alta precisión para ambientes de líneas de producción, con poca demanda de tiempo que además eviten el uso de reactivos químicos, equipamientos de laboratorio y la destrucción de las muestras a analizar, surgen técnicas de análisis de la calidad de alimentos mediante el uso de espectroscopía de terahercios (Qin, Ying, & Xie, 2013). La banda del rango de los Terahercios (THz) es la región del espectro electromagnético desde los 100 GHz hasta los 30 THz. Esta banda situada entre el microondas y el infrarojo fue ignorada hasta mediados de los 90’s debido a la baja eficiencia de detectores y fuentes de energía THz (Gowen, O’Sullivan, & O’Donnell, 2012). La radiación de 1 THz tiene un periodo de , con una longitud de onda de y el número de ondas de , la energía del fotón 4.1 y la temperatura equivalente de o - (Zhang & Xu, 2009). Estas características dotan de ventajas frente a otros métodos de inspección interna por su onda no ionizante, dada la baja energía de fotón por cada THz.



1 THz = 1ps = 300 μm = 4.1meV = 33cm-1

Figura 1: Rango del espectro electromagnético y la conversión de Terahertz (THz) a distintas medidas. Elaboración propia.

El rango de los THz tienen diversas aplicaciones, entre ellos: sistemas de física de materia condensada, ciencia de materiales, bioquímica, medicina, diagnóstico, transferencia de datos, comunicación segura de banda ancha a gran distancia, seguridad aeroportuaria, aplicación de la ley y defensa militar (Prabhu, 2017). Desde hace algunos años, han surgido estudios enfocados en la exploración de la tecnología THz para explotar sus ventajas y ser aplicada en diversas industrias, especialmente en la industria agroalimentaria (Wang, Sun, & Pu, 2017), por ejemplo, para la discriminación de semillas de arroz transgénicas (W. Liu, Liu, Hu, Yang, & Zheng, 2016), inspección de cuerpos extraños en alimentos (Ok, Park, Lim, Jang, & Choi, 2018; Shin, Choi, & Ok, 2018), la cuantificación de aminoácidos (S. H. Lu, Li, Zhai, Zhang, & Zhang, 2018) e inclusive la discriminación del origen geográfico aplicado a los aceites de oliva extra virgen (Wei Liu et al., 2018) en combinación con técnicas quimiometricas u otras relacionadas.

Las imágenes hiperespectrales THz contienen desde cientos de miles, hasta millones de datos espectrales por cada imagen. Para poder construir herramientas con diferentes aplicaciones utilizando el análisis de imágenes hiperespectrales, es importante reducir las dimensiones del conjunto de datos a utilizarse en los modelos para elevar la precisión y efectividad en los objetivos de dichas herramientas. La selección de variables tiene tres objetivos (Jiménez, 2012): mejorar el rendimiento de predicción de los predictores, proporcionar predictores más rápidos y eficientes y proporcionar un mejor entendimiento del proceso subyacente que generó el conjunto de datos.

La selección de características aumenta la probabilidad de predicción de algoritmos al disminuir la dimensionalidad, eliminando características irrelevantes (Tarakçi Kalayci, Alkaya, & Algin, 2020). Para alimentar las entradas de modelos predictivos de la calidad de alimentos y obtener una alta precisión de predicción, se requiere pre procesar la información presente en imágenes hiperespectrales en orden de la relevancia de los atributos de la calidad de los alimentos (D. Liu et al., 2014). mediante métodos de selección de variables y características.

El objetivo de la presente investigación es realizar una revisión sistemática de los métodos de selección de variables y características más utilizados presentes en la literatura, de manera que al analizar diferentes imágenes THz y extraer conjuntos de datos relevantes para el análisis predictivo de la calidad de alimentos lograr extraer subconjuntos de datos mediante diferentes algoritmos y técnicas de selección de variables y características y transformar el conjunto de datos inicial en variables y características específicas y así, determinar cuáles son los métodos de selección de variables y características más adecuados para ser aplicados en dominios relacionados a la predicción y clasificación de la calidad de alimentos usando los datos extraídos del análisis de imágenes THz.

**Materiales y Métodos**

**Espectroscopia de Imágenes THz**

La espectroscopía con Imágenes THz es una técnica de obtención de datos moderna e inexplorada. Las formas más utilizadas para la generación y detección de la radiación pulsada de THz son antenas fotoconductoras y cristales ópticos electrónicos; además, plasmas inducidos por aire y láser, aceleradores de electrones y emisores de foto-Dember (Jepsen, Cooke, & Koch, 2011). Para la espectroscopía THz comercial, la especificación del dominio de tiempo THz Troscopy (THz-TDS) es una herramienta de medición muy efectiva ampliamente utilizada (Wang et al., 2017).

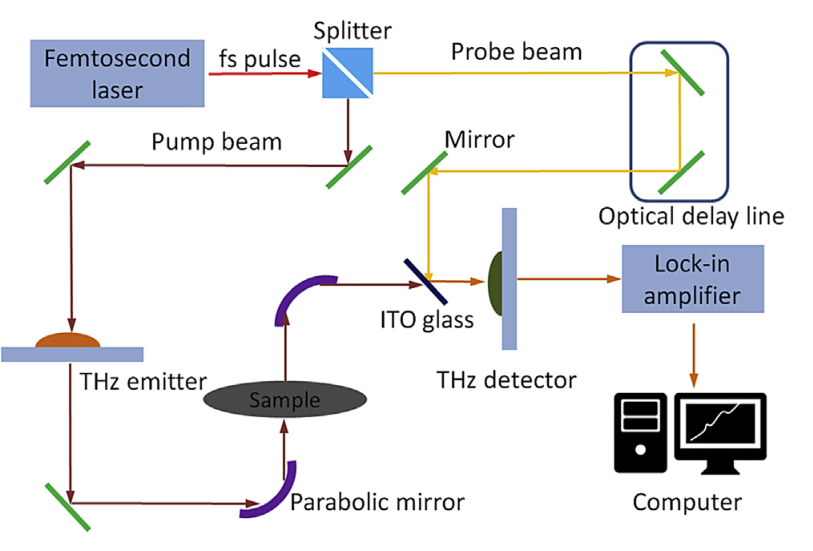


Figura 2: Esquema de configuración típica de espectroscopía THz-TDS (Wang et al., 2017). Esta configuración consiste principalmente de un láser de femtosegundos (fs), una línea de retardo óptico, un emisor y un detector de THz, y un amplificador de bloqueo. Los pulsos THz son enfocados por los espejos parabólicos, donde interactúan con la muestra para combinarse con la sonda del detector y posteriormente ser amplificada por el amplificador de bloqueo, y así, grabar en función del tiempo el campo eléctrico THz aplicando la transformada de Fourier para convertir la señal del dominio del tiempo en el espectro de frecuencias (Leahy-Hoppa, Fitch, & Osiander, 2009).

El espectómetro de pulso THz TeraPulse 4000 (TeraView Ltd, 2019) cuenta con una fuente emisor y receptor semiconductor fotoconductor de activación por láser, con láser de fibra de pulso ultracorto en el rango espectral desde los 0.06 THz hasta los 4 THz (2 cm-1 – 133 cm-1) con un rango dinámico pico de 90dB y compartimiento de muestreo a base de nitrógeno. Este espectómetro se utilizará como fuente de obtención de datos de las muestras de cacao para la presente investigación. TeraPulse 4000 obtiene imágenes hiperespectrales multivariadas con millones de datos que se almacenan en archivos en el rango de los Gigabytes.

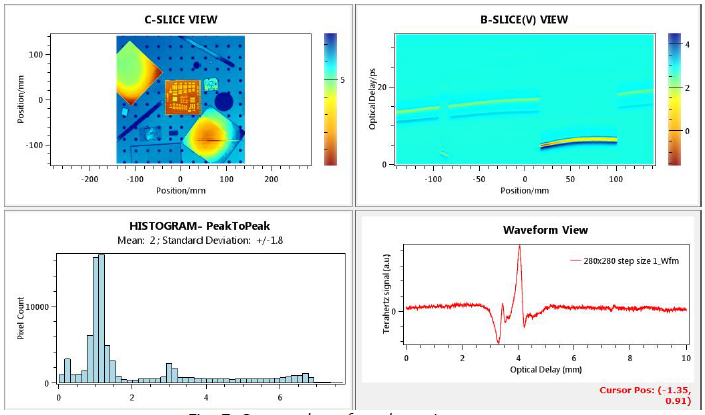


Figura 3: Captura de pantalla de una imagen THz obtenida por el espectrómetro TeraView4000. (TeraView Ltd, 2019)

**Análisis de Imágenes THz**

Las imágenes THz hiperespectrales son almacenadas en archivos, los cuales deberán ser pre procesados para extraer las variables relevantes necesarias en el orden de prioridad de los objetivos de este estudio. Se implementaron un conjunto de funcionalidades reunidas en una sola interfaz gráfica de usuario (GUI) diseñada y programada con el lenguaje de programación Python 3.6.8. Las funcionalidades de la GUI desarrollada ayudarán en la selección primaria de datos, para ello se implementó la lectura de archivos hiperespectrales, visualización de la información de cabecera, visualización combinada de tres bandas para efectos RGB (HIT por NIR), funcionalidad de ajuste y configuración de imágenes, espectrograma por selección de píxel, visualización y renderizado de la imagen hiperespectral en hipercubos 3D, selección de ***Regiones de Interés*** (ROIs), almacenamiento de datos de RIOs seleccionadas , gráfica de perfiles espectrales de ROIs seleccionados, selección de bandas y la selección de variables y características relevantes. Esta herramienta integra, además, los métodos de selección de variables y características para la obtención de subconjuntos de datos más pequeños y específicos que serán utilizados inicialmente para entrenar un modelo predictivo de la calidad de alimentos.

*Selección de Regiones de Interés*

Para obtener un conjunto inicial de datos de alta relevancia para diferentes propósitos, de acuerdo con la presente investigación, se es necesaria la selección manual de ***Regiones de Interés*** para obtener subconjuntos de datos más pequeño, aislando los datos de las regiones donde un analista experto pueda seleccionar los píxeles que contiene la información de las regiones con características de especial interés.

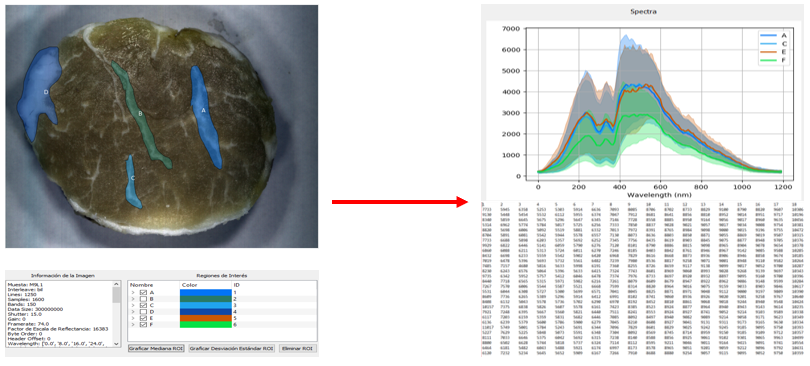


Figura 4: Extracción de ROIs. Estas regiones generan un subconjunto de datos que pueden ser almacenados en formato txt o utilizados para otros procedimientos dentro de la misma herramienta. Elaboración propia.

Este subconjunto de datos específicos es el subconjunto inicial de datos y a pesar de que se reduce la cantidad de datos a un conjunto de píxeles con sus respectivas firmas espectrales, la cantidad de dimensiones es aún bastante grande, por lo que este subconjunto de datos es el que se someterá a los métodos de selección de variables con el objetivo de reducir la dimensionalidad y obtener subconjunto de características para alimentar las entradas de un modelo predictivo de la calidad de alimentos.

**Selección de Variables**

Existen tres conjuntos destacados en las que podemos agrupar las técnicas de selección de variables:

**Wrappers** (envoltorios): Utilizada como métodos universales para puntuar subconjuntos de variables acorde a su potencial predictivo.

**Filters** (filtros): Selecciona subconjuntos de variables como un paso de preprocesado independientemente del predictor utilizado.

**Embedded** (integrados): Específico para cada modelo de aprendizaje, realiza la selección de variables de manera integrada mientras ocurre el proceso de entrenamiento en el modelo predictivo.

*Wrappers*

Estas metodologías son simples de implementar. Independientemente del modelo elegido para abordar determinado problema, las metodologías wrappers son potentes abordando la selección de variables basándose en el potencial predictivo del modelo. Este rendimiento de predicción es utilizado para evaluar la utilidad relativa de cada subconjunto de variables. Para implementar un método wrapper es esencial tener en cuenta el espacio de todos los posibles subconjuntos de variables, el cómo evaluar el rendimiento de predicción de la RNA para guiar la búsqueda y detenerla en el momento oportuno y definir el modelo de predictor a usar.

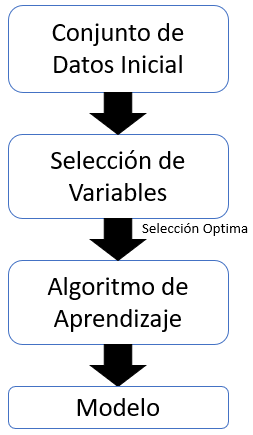
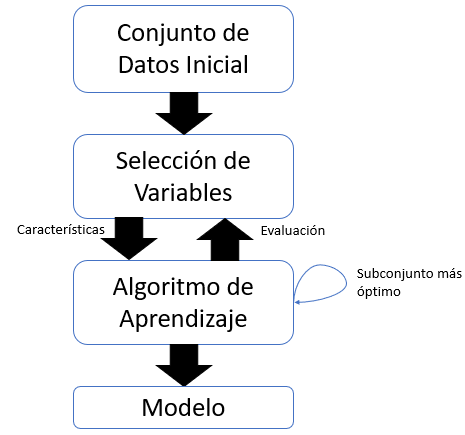
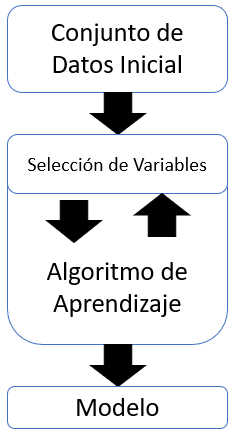
Los métodos wrappers son particularmente eficaces al trabajar con un npumero de variables no tan largo, ya que se puede utilizar un método de búsqueda exhaustiva. La búsqueda se vuelve intratable con problemas con alta dimensionalidad de variables. Alguna de las estrategias de búsqueda que pueden utilizarse son la técnica best-first, branch and boud, simulated annealing, algoritmos genéticos, etc. (Jiménez, 2012). Las validaciones de los métodos wrapper se realizan con un conjunto de datos de validación o por validación cruzada (*cross-validation*).

*Filters*

Son más rápidos en comparación con los métodos wrappers. Son esencialmente una propuesta de solución genérica a la selección de variables, sin tener en cuenta las características del modelo que aborda el problema global y se pueden usar como métodos de pre procesado para la reducción del espacio y la dimensionalidad de las variables de entrada del modelo y superar así el sobreajuste de aprendizaje en el entrenamiento.

*Embedded*

Estos métodos generalmente son característicos de ser dependientes en incorporados en cada modelo específico de aprendizaje. Estos métodos se ejecutan al mismo tiempo en que el proceso de aprendizaje se realiza y esta característica puede resultar beneficiosa con respecto al tratamiento de los datos: hacen un mejor uso de los datos disponibles al no tener que dividir los datos de entrenamiento en un conjunto de entrenamiento y validación; alcanzan una solución más rápida al evitar el reentrenamiento de los predictores desde cero para cada subconjunto de variables investigadas (Guyon & Elisseeff, 2003). Estos métodos, en combinación con algoritmos Greedy para la selección de variables, obtienen un subconjunto de variables anidadas.

(A) (B) (C)

Figura 5: Esquematización de las técnicas de selección de variables: Wrappers (A), Filters (B) y Embedded (C). Elaboración propia.

**Algoritmos de búsqueda local de variables**

Al abordar problemas de selección de variables, existen tres clasificaciones de los algoritmos o métodos para la búsqueda local de variables con principios fundamentales diferenciados en cada uno de ellos. Estos algoritmos son del tipo Greedy y obtienen como resultado el subconjunto de variables. Estos algoritmos los podemos agrupar en tres categorías:

**Algoritmos Foward** (Selección hacia adelante): Se inicia con un modelo sencillo y se van agregando términos bajo algún criterio hasta que no procede añadir algún termino. En resumen, se añade la variable más significativa hasta el cumplimiento de una regla que detenga el proceso.

**Algoritmos Backward** (Eliminación hacia atrás): Se parte de un complejo modelo que contiene todos los efectos que pueden influir en la respuesta y en cada fase se suprime la variable menos influyente hasta que no proceda eliminar ningún término más.

**Algoritmos Stepwise** (Foward-Backward): Es una combinación de métodos foward y backward. Comienza como el algoritmo foward, pero en cada etapa se evalúa si todas las variables introducidas deben permanecer en el modelo.

Las evaluaciones del rendimiento generalmente son realizadas con un conjunto de datos de validación o mediante la validación cruzada (cross-validation). Utilizada para estimar el riesgo de un estimador o para realizar la selección del modelo, en este de la característica seleccionada, la validación cruzada es una estrategia generalizada debido a su simplicidad y su aparente universalidad (Arlot & Celisse, 2010).

**Algoritmos de búsqueda global de variables**

**Búsqueda Exhaustiva**: Es un algoritmo simple que encuentra el valor óptimo de la función objetivo buscando entre todas las combinaciones posibles de entradas y salidas. Esta búsqueda es computacionalmente inviable, debido a que es un algoritmo de fuerza bruta y testea todas las combinaciones posibles y el tiempo en encontrar la solución podría ser demasiado largo.

**Búsqueda Tabu**: La búsqueda TS es un algoritmo metaheurístico que puede usarse para guiar los métodos de búsqueda local para explorar el espacio de soluciones más allá de la optimización local. Unidos a diferentes componentes, TS es poderoso y capaz de obtener excelentes resultados en diferentes problemas. Usa métodos *neighborhood* para moverse iterativamente de una solución *x* a una solución *x’* hasta que un criterio de parada que detenga la búsqueda.

**Algoritmos Genéticos**: Es una técnica de búsqueda utilizada para encontrar soluciones exactas o aproximadas a problemas de optimización de búsqueda, pertenecientes a la clase de algoritmos evolutivos que usan técnicas inspiradas en la biología evolutiva (herencia, mutuación y selección crossover).

**Constructores de Características**

Se puede asociar la selección de Variables y la transformación de dimensionalidad al concepto de constructores de características. Es posible encontrar una gran variedad de constructores de características. Entre las más comunes están las funciones de reducción de variables, la transformación lineal de variables de entrada, transformación espectral (e.g. Transformada de Fourier), transformación por *wavelets*, clustering, entre otros. A continuación, haremos una revisión de los constructores de características más utilizados:

***Remover características con baja varianza***

Es un método de eliminación de variables básico que elimina todas las características cuya variación no alcanza algún umbral en específico.

donde, hace referencia a cada dato, la media de los datos y el número total de datos.

***Selección de los K-ésimos mejores puntuados***

Es posible seleccionar el conjunto de variables que contengan las mejores puntuaciones según alguna función diseñada para puntuar las variables en base a diferentes criterios, como por ejemplo, la función chi-cuadrado de características no negativas para tareas de clasificación, el valor F entre etiqueta/característica para tareas de regresión, información sobre continuidad de objetos, tazas de falsos positivos, tasas de falsos descubrimientos, tasas de error family-wise o selectores de funciones univariadas configurables (Scikit Learn Documentation, 2019).

***Análisis de Componentes Principales***

También conocido como método de construcción lineal de variables, PCA (por sus siglas en inglés) reduce la dimensionalidad al encontrar la mejor proyección de los datos sobre un vector, utilizando los valores propios o ***eigenvalues*** y los vectores propios o ***eigenvectors***. Dado un conjunto de datos centrados por su centro de masa, , PCA diagonaliza la matriz de covarianza :

Hallando el valor propio o *eigenvalue* de cada dimensión mediante la ecuación:

y puede ser reescrito de la siguiente manera:

Los valores de los eigenvalues y los eigenvectors pueden determinarse por:

donde es la matriz identidad.

Una vez obtenido los valores y vectores propios, podemos crear un subconjunto de características utilizando los eigenvertors transformando de la siguiente manera:

donde cada media está dada por y cada dato transformado por .

Este método, por ejemplo, reduce la dimensionalidad de un dominio de datos considerado intratable de 5 dimensiones a un subconjunto de datos de dos dimensiones, los cuales sí podrían ser cubiertos por modelos de aprendizaje.

***Análisis de Componentes Principales basado en Núcleo***

k-PCA es una generalización no lineal del PCA en el sentido de que si usamos el kernel recuperamos el PCA original (Jiménez, 2012) usando un núcleo, las operaciones lineales del PCA se realizan el espacio Hilbert con núcleo de reproducción (Moisés, s. f.) para realizar la selección de características en dominios de datos de altas dimensiones.

kPCA es útil en agrupación de datos. Si bien los puntos N no se pueden separar linealmente en dimensiones, pero casi siempre se pueden separar linealmente en dimensiones. Si tenemos *N* datos, si podemos mapearlos en un espacio N-dimensional con

fácilmente se pueden construir hiperplanos que divida los datos en agrupaciones arbitrarias. crea vectores con independencia lineal, por lo que no contienen covarianza. Una función no trivial se escoge para que los datos no sean independientes en . En lugar de elegir explícitamente , escoge*mos*

Donde K es la matriz de Gram (Shawe-Taylor, Williams, Cristianini, & Kandola, 2005) en el espacio de alta dimensionalidad.

K-PCA nos permite operar en muchos espacios sin mapear explícitamente los datos en el espacio de alta dimensionalidad. K-PCA puede ser descrito como un problema en términos de productos internos con la matriz de datos transpuesta, así podemos concluir

k-PCA podría ser desventajoso con un enorme número de observaciones que resulte en una extensa matriz *K*.

***Regresión Parcial de Mínimos Cuadrados***

La regresión PLS es un método estadístico ciertamente relacionado con la regresión de componentes principales, a diferencia de encontrar hiperplanos de máxima varianza entre las salidas y las variables independientes, encuentra un modelo lineal proyectando las variables predichas y las variables observables en un nuevo espacio. Adecuado para predictores que tienen más variables que observaciones, PLS intentará encontrar la dirección multidimensional en el espacio X que explica la dirección máxima de la varianza multidimensional en el espacio Y.

Fundamentalmente, encuentra la relación entre dos matrices *X* e *Y*. Se conoce como una manera de modelar ecuaciones estructurales basada en variaciones o componentes (Jiménez, 2012).

Para una sola salida *y* y *p* predictores, el modelo PLS con *h (h ≤ p)* variables latentes, se puede expresar como:

Donde , ,, y usados por predictores, *X* puntajes, *X* Cargas, una salida *y*, y los coeficientes de regresión *T*. El k-esimo elemento del vector de columna b explica la relación entre *y* y , la k-esima columna del vector *T*. y representan los errores de *X* e *y* respectivamente. Generalmente usando el algoritmo de mínimos cuadrados parciales iterativos no lineales (NIPALS), se obtiene una matriz de peso para hacer lo más pequeña posible y al mismo tiempo encontrar la relación útil entre *X* e *y* (Jiménez, 2012).

***k-Vecinos más cercanos***

KNN (por sus siglas en inglés) es un método utilizado para predicción y clasificación. KNN se basa en la conjetura de que los conjuntos de datos de entrada similares tienen valores de salida similares. Basa en la distancia euclidiana y los valores de salida correspondientes son utilizados para obtener la máxima aproximación de salida deseada. La salida estimada se calcula promediando las salidas del vecino más cercano:

donde es el estimado o la aproximación de la salida, es la salida de la j-ésimo vecino más cercano de la muestra y es el número de vecinos utilizados . Las distancias entre las muestras están influenciadas por la selección de entradas. Para el método KNN es necesario determinar el número de , la cual puede ser resultante de aplicar diferentes ténicas como k-fold cross-validation, LOO, Bootstrap (con mejor rendimiento junto con LOO) y Bootstrap 632 (Jiménez, 2012). Esta selección es la que minimiza el error generalizado.

**k-Means**

La agrupación de K-means es un método comúnmente utilizado para particionar automáticamente un conjunto de datos en k grupos (Wagstaﬀ, Cardie, Rogers, & Schroedl, 2001). Este método de agrupación (clustering) es un método de clasificación no supervisado basado en la agrupación de elementos con similares características, buscando patrones en los datos sin tener una predicción específica como objetivo, es decir, sin variable dependiente.

La manera en que el algoritmo k-means clustering opera es la siguiente:

1. Cada instancia se asigna a su grupo más cercano centrar.

2. Cada centro de clúster se actualiza

El algoritmo converge cuando no hay más cambio en la asignación de instancias a clústeres (Wagstaﬀ et al., 2001).

***Nonparametric noise estimation***

Los métodos NNE selecciona las mejores características basadas solamente en el conjunto de dato, por lo que no es necesario construir un modelo de regresión para encontrar el mejor conjunto de variables. Estos métodos se basan en calcular un criterio que determina las dependencias entre cada combinación de variables de entrada y la salida correspondiente utilizando previsibilidad, correlación u otra información estadística relacionada. La referencia a la estimación del “ruido” es la estimación de la varianza de alguna función desconocida que relaciona las variables de entrada y salida de un modelo. Una NNE estima simultáneamente el rendimiento a priori del mejor modelo de regresión no lineal que se puede construir con las variables seleccionadas.

***La prueba Gamma***

El test Gamma (GT) es una técnica NNE basada en una generalización de la desigualdad de Chybechov y la propiedad de las estructuras vecinas k-más cercanas (Lendasse, Ji, Reyhani, & Verleysen, 2005). Esta prueba es mayormente utilizada para determinar la correlación no lineal entre dos variables aleatorias. Las condiciones para realizar una prueba Gamma son las siguientes:

* existen las derivadas parciales primera y segunda de la función subyacente;
* existen los momentos primero a cuarto de la distribución del ruido;
* el ruido es independiente de la entrada.

Con estas tres condiciones, la variación del ruido se da por la intersección de la línea de regresión entre y con la línea vertical donde y

donde es el índice del k-ésimo vecino más cercano de . Esta es una técnica no paramétrica en la cual podemos aplicar sus resultados independientemente de los métodos particulares que se utilizan posteriormente para construir el modelo.

***La prueba Delta***

Este enfoque estima la varianza del ruido en la salida (Jones, 2004) basada en la similitud del comportamiento del ruido entre dos puntos de datos cerrados. La prueba Delta puede interpretarse como una particularización de la prueba Gamma considerando solo el primer vecino más cercano, lo que produce un valor no paramétrico, ya que elimina el único hiperparámetro que se tenía que elegir para la prueba Gamma (el número de vecinos).

Donde es la salida de . La prueba Delta presenta mayor rendimiento de cómputo que la prueba Gamma, con una pequeña fracción de su tiempo de cómputo y requisitos de memoria, ya que solo es necesario almacenar la distancia desde cada punto hasta su vecino más cercano (Jiménez, 2012).

Estos estimadores de varianza de ruido presentan problemas al momento de realizar la selección de variables. Uno de los principales problemas relacionados es la de fallar al momento de filtrar entradas irrelevantes que para ciertos contextos pueden contener información adicional que conduzca a una estimación desvariada. Pese a esto, la prueba Delta tiene la propiedad interesante de que, al agregar entradas no relacionadas, aumentan la estimación y así se separarlas de la mayoría de otros estimadores, haciéndola efectiva de otros selectores de entradas (Eirola, Liitiäinen, Lendasse, Corona, & Verleysen, 2008).

**Métodos de Validación**

***Raíz del error cuadrático medio RMSE***

*Root mean squared error* o RMSE es un estimador que mide el promedio de los errores al cuadrado, que vendría a ser la diferencia entre el estimador y lo que se estima. RMSE nos puede ayudar a determinar la información que podría producir una estimación más precisa en nuestro modelo de predicción del tiempo de fermentación del cacao.

Donde es la esperanza, es el estimador y es el parámetro.

***Coeficiente de Correlación***

El coeficiente de correlación mide la fuerza y el sentido de la relación lineal entre dos variables cuantitativas. El coeficiente viene dado por:

Donde, de un conjunto de ejemplos {}, y las variables de entrada con una variable de salida para cada .

***Coeficiente de Determinación***

El coeficiente de determinación para una regresión lineal utilizado en modelos estadísticos cuyo propósito es predecir futuros resultados o probar una hipótesis. Este coeficiente viene dado por:

**Resultados**

Con el propósito de cumplir con los objetivos del presente trabajo de investigación, se realizaron un total de tres experimentos estadísticos, donde se utilizaron diferentes técnicas de selección de variables y características, específicamente se realizó la selección de variables mediante un algoritmo *Filter* no paramétrico llamado SelectKBest o Variable Ranking en base a la selección de bandas hiperespectrales de un total de 6 regiones de interés extraídas manualmente usando un Test Gamma Chi-2 como función de puntuación, el segundo experimento de extracción de variables realizado es la extracción de características mediante un método *Wrapper* con el Análisis de Componentes Principales y por último el tercer experimento realizando la extracción de variables utilizando el algoritmo k-means para clasificar de manera no supervisada las características del total de píxeles espectrales.

El hipercubo contiene datos de reflectancia en el espectro NIR, que por cuestiones de carencia de datos en el espectro THz se decidió utilizar debido a la necesidad de completar con los objetivos de la presente investigación. Estos datos se distribuyen en 150 bandas que equivalen al tamaño de 1192nm de muestra. Además, la imagen hiperespectral se divide en 1250 líneas (que equivalen al eje X en una imagen bidimensional) y 1600 muestras (que equivalen al eje Y de una imagen bidimensional), dando así el total de 300 millones de datos.

La figura 6 muestra una imagen donde se combinan los datos de reflectancia de tres bandas para formar una sola imagen visible en el color original, como si se hubiese tomado fotografía común y corriente.



Figura 6: Imagen Hiperespectral RGB en las bandas 80, 69 y 57.Elaboración Propia

**Selección de Variables mediante Variable Ranking**

Para realizar la selección de variables por el método de puntuación, utilizaremos el criterio de puntuación calculando la estadística chi-cuadrado, la cual mide la dependencia entre variables aleatorias, por lo que el uso de esta función “elimina” las características que tienen más probabilidad de ser independientes de la clase y, por lo tanto, irrelevantes para la clasificación (Scikit Learn Documentation, 2019). Iniciamos con la selección manual de regiones de interés utilizando las herramientas desarrolladas para la presente investigación:

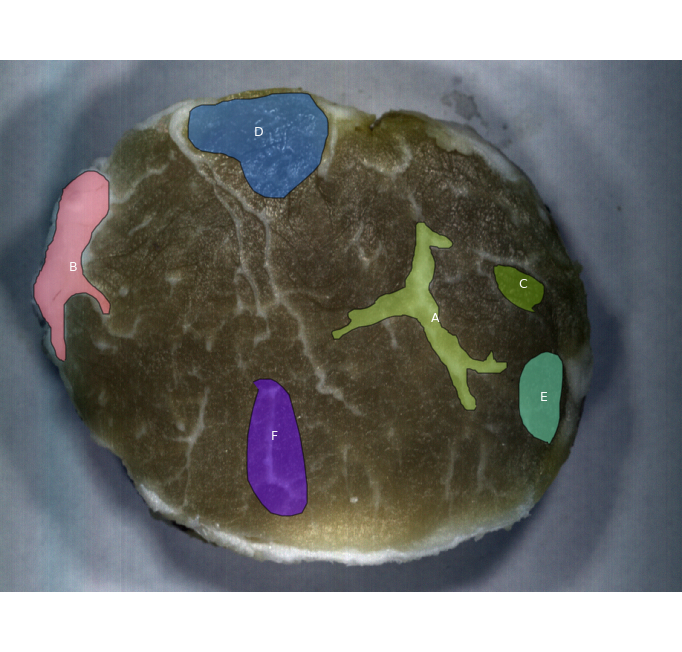


Figura 7: Selección manual de 6 Regiones de Interés. Elaboración Propia

Estas selecciones de regiones de interés dan por resultados 6 perfiles espectrales, los cuales se muestran en la figura 8.

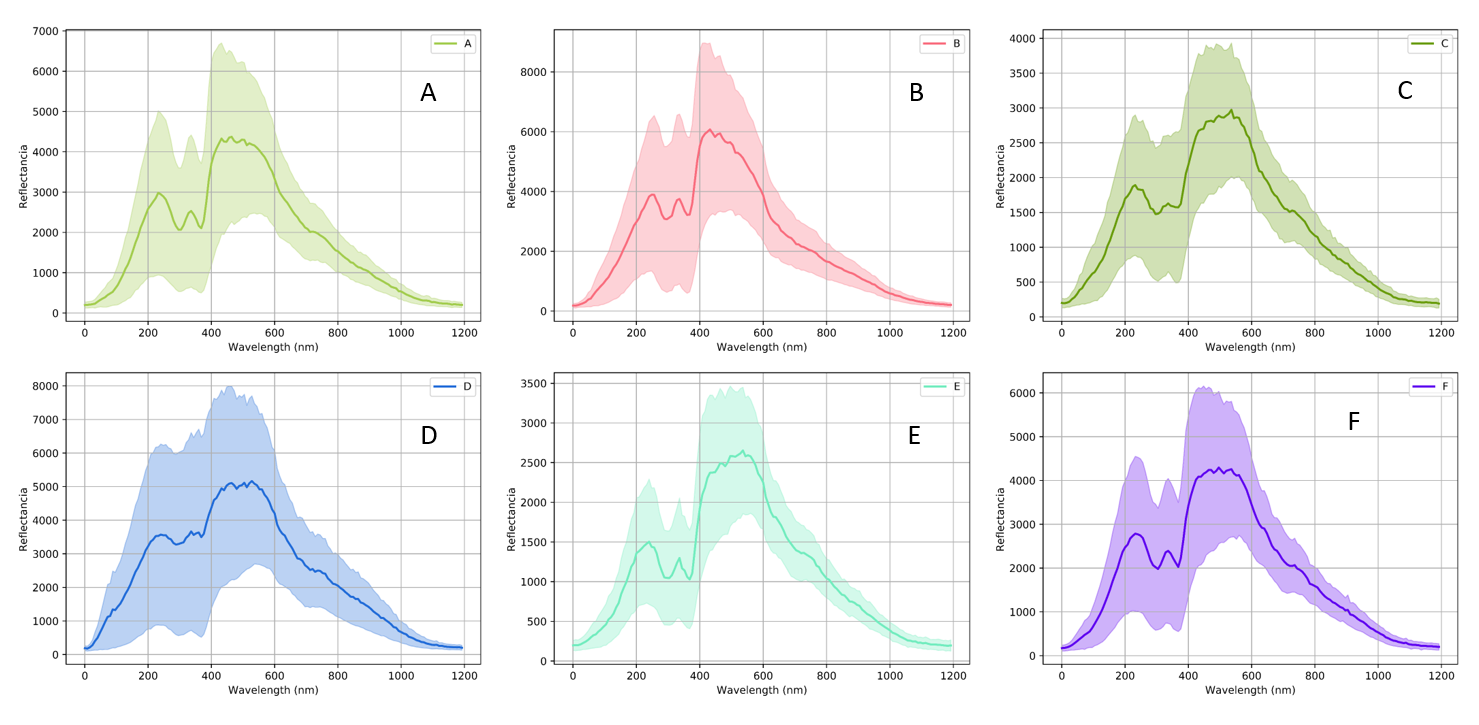


Figura 8: Perfiles espectrales de las regiones de interés seleccionadas. Elaboración Propia

Una vez obtenidas las regiones de interés se almacenan en archivos de texto plano utilizando la función Exportar. De estas regiones de interés podemos asignar un rango de calidad según el color de la zona seleccionada en 5 categorías, en un intervalo de 0 a 100, asemejando un rango de probabilidad del 0% a 100%:

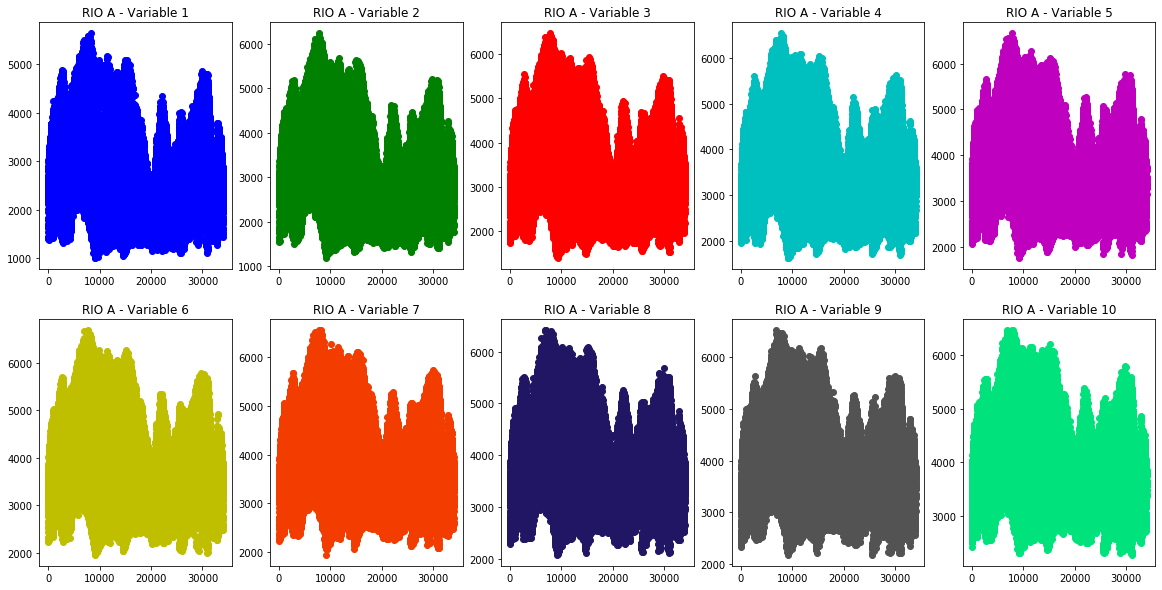
* Calidad Alta: de 0% al 12% (Totalmente Blanca)
* Calidad Media-Alta: de 12% al 37% (Zona parcialmente blancas)
* Calidad Media: de 37% al 62% (Zona con surcos)
* Calidad Media-Baja: de 62% al 87% (Zona parcialmente Obscura)
* Calidad Baja: de 87% al 100% (Zona Obscura)

Para realizar dicha clasificación, usamos la función *randint* disponible en la librería *Numpy* para Python. Una vez realizada la carga de datos y la definición de los rangos de clasificación con variables aleatorias para cada banda, seleccionamos la cantidad de bandas que vamos a filtrar para obtener nuestros subconjuntos de datos finales. Utilizando la función SelectKBest, la función chi2 disponibles en la librería *scikit-learn* y el parámetro de 10 dimensiones, obtenemos los siguientes subconjuntos de datos.

Las dispersiones de las regiones de interés, en una vista inicial de los diagramas de dispersión generados, hace referencia a una selección uniforme de los datos.

Región de Interés A:

Tabla 1:



**Discusión**

Discusión

**Referencias bibliográficas**

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, *4*, 40-79. https://doi.org/10.1214/09-SS054

Baek, S. H., Lim, H. B., & Chun, H. S. (2014). Detection of Melamine in Foods Using Terahertz Time-Domain Spectroscopy. *Journal of Agricultural and Food Chemistry*, *62*(24), 5403-5407. https://doi.org/10.1021/jf501170z

Castro, W., Oblitas, J., Maicelo, J., & Avila-George, H. (2011). Evaluation of Expert Systems Techniques for Classifying Different Stages of Coffee Rust Infection in Hyperspectral Images. *International Journal of Computational Intelligence Systems*, *11*(1), 86-100. https://doi.org/10.2991/ijcis.11.1.8

Condezo-Hoyos, L., & Castro, W. (2018). *Artificial Neural Networks and Hyperspectral Images for Quality Control in Foods*. https://doi.org/10.1201/9781315209203-11

Crowley, M. D., Chen, W., Sukalac, E. J., Sunt, X., Coronado, P. L., & Zhang, G.-Q. (2006). *Visualization of remote hyperspectral image data using google earth*. 907-910. https://doi.org/10.1109/IGARSS.2006.233

Edwards, H. M. (1987). An Appreciation of Kronecker. *The Mathematical Intelligencer*, *9*(1), 28-35. https://doi.org/10.1007/BF03023570

Eirola, E., Liitiäinen, E., Lendasse, A., Corona, F., & Verleysen, M. (2008). *Using the Delta Test for Variable Selection.*

Fu, P., Meacham-Hensold, K., Guan, K., & Bernacchi, C. J. (2019). Hyperspectral Leaf Reflectance as Proxy for Photosynthetic Capacities: An Ensemble Approach Based on Multiple Machine Learning Algorithms. *Frontiers in Plant Science*, *10*, 730. https://doi.org/10.3389/fpls.2019.00730

Fukunaga, K. (2016). THz Instruments. En K. Fukunaga (Ed.), *THz Technology Applied to Cultural Heritage in Practice* (pp. 11-22). https://doi.org/10.1007/978-4-431-55885-9\_2

Gowen, A. A., O’Donnell, C. P., Cullen, P. J., Downey, G., & Frias, J. M. (2007). Hyperspectral imaging – an emerging process analytical tool for food quality and safety control. *Trends in Food Science & Technology*, *18*(12), 590-598. https://doi.org/10.1016/j.tifs.2007.06.001

Gowen, A. A., O’Sullivan, C., & O’Donnell, C. P. (2012). Terahertz time domain spectroscopy and imaging: Emerging techniques for food process monitoring and quality control. *Trends in Food Science & Technology*, *25*(1), 40-46. https://doi.org/10.1016/j.tifs.2011.12.006

Guyon, I., & Elisseeff, A. (2003). *An Introduction to Variable and Feature Selection*. 26.

Hue, C., Gunata, Z., Bergounhou, A., Assemat, S., Boulanger, R., Sauvage, F. X., & Davrieux, F. (2014). Near infrared spectroscopy as a new tool to determine cocoa fermentation levels through ammonia nitrogen quantification. *Food Chemistry*, *148*, 240-245. https://doi.org/10.1016/j.foodchem.2013.10.005

Jepsen, P. U., Cooke, D. G., & Koch, M. (2011). Terahertz spectroscopy and imaging – Modern techniques and applications. *Laser & Photonics Reviews*, *5*(1), 124-166. https://doi.org/10.1002/lpor.201000011

Jiménez, F. M. (2012). *Redes neuronales y preprocesado de variables para modelos y sensores en bioingeniería*. Universidad Politécnica de Valencia, Valencia, España.

Jones, A. J. (2004). New tools in non-linear modelling and prediction. *Computational Management Science*, *1*(2), 109-149. https://doi.org/10.1007/s10287-003-0006-1

Leahy-Hoppa, M. R., Fitch, M. J., & Osiander, R. (2009). Terahertz spectroscopy techniques for explosives detection. *Analytical and Bioanalytical Chemistry*, *395*(2), 247-257. https://doi.org/10.1007/s00216-009-2803-z

Lendasse, A., Ji, Y., Reyhani, N., & Verleysen, M. (2005). *LS-SVM Hyperparameter Selection with a Nonparametric Noise Estimator*. 6.

Li, Y., Shan, J., & Peng, Y. (2011). *A hyperspectral imaging system for real-time detection of beef quality*. *3*, 2196-2204. Recuperado de Scopus.

Liu, D., Sun, D.-W., & Zeng, X.-A. (2014). Recent Advances in Wavelength Selection Techniques for Hyperspectral Image Processing in the Food Industry. *Food and Bioprocess Technology*, *7*(2), 307-323. https://doi.org/10.1007/s11947-013-1193-6

Liu, W., Liu, C., Hu, X., Yang, J., & Zheng, L. (2016). Application of terahertz spectroscopy imaging for discrimination of transgenic rice seeds with chemometrics. *Food Chemistry*, *210*, 415-421. https://doi.org/10.1016/j.foodchem.2016.04.117

Liu, Wei, Liu, C., Yu, J., Zhang, Y., Li, J., Chen, Y., & Zheng, L. (2018). Discrimination of geographical origin of extra virgin olive oils using terahertz spectroscopy combined with chemometrics. *Food Chemistry*, *251*, 86-92. https://doi.org/10.1016/j.foodchem.2018.01.081

Lu, G., & Fei, B. (2014). Medical hyperspectral imaging: A review. *Journal of Biomedical Optics*, *19*(1). https://doi.org/10.1117/1.JBO.19.1.010901

Lu, S. H., Li, B. Q., Zhai, H. L., Zhang, X., & Zhang, Z. Y. (2018). An effective approach to quantitative analysis of ternary amino acids in foxtail millet substrate based on terahertz spectroscopy. *Food Chemistry*, *246*, 220-227. https://doi.org/10.1016/j.foodchem.2017.11.016

McCallum, A., Nigam, K., & Ungar, L. H. (2000). *Efficient clustering of high-dimensional data sets with application to reference matching*. 169-178. Recuperado de Scopus.

Moisés, B. C. R. (s. f.). *Espacios de Hilbert con núcleo reproductor*. 13.

Nicolaï, B. M., Beullens, K., Bobelyn, E., Peirs, A., Saeys, W., Theron, K. I., & Lammertyn, J. (2007). Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: A review. *Postharvest Biology and Technology*, *46*(2), 99-118. https://doi.org/10.1016/j.postharvbio.2007.06.024

Ok, G., Park, K., Lim, M.-C., Jang, H.-J., & Choi, S.-W. (2018). 140-GHz subwavelength transmission imaging for foreign body inspection in food products. *Journal of Food Engineering*, *221*, 124-131. https://doi.org/10.1016/j.jfoodeng.2017.10.011

Prabhu, S. S. (2017). Terahertz Spectroscopy: Advances and Applications. En *Mol. And Laser Spectrosc.: Adv. And Appl.* (pp. 65-85). https://doi.org/10.1016/B978-0-12-849883-5.00004-8

Qin, J., Ying, Y., & Xie, L. (2013). The Detection of Agricultural Products and Food Using Terahertz Spectroscopy: A Review. *Applied Spectroscopy Reviews*, *48*(6), 439-457. https://doi.org/10.1080/05704928.2012.745418

Redo-Sanchez, A., Salvatella, G., Galceran, R., Roldós, E., García-Reguero, J.-A., Castellari, M., & Tejada, J. (2011). Assessment of terahertz spectroscopy to detect antibiotic residues in food and feed matrices. *Analyst*, *136*(8), 1733-1738. https://doi.org/10.1039/C0AN01016B

Roman-Gonzalez, A., & Vargas-Cuentas, N. I. (2013). Análisis de imágenes hiperespectrales. *Revista Ingenieria & Desarrollo*, *Año 9*(N° 35), 14-17.

Scikit Learn Documentation. (2019). Scikit Learn, Feature Selection Module [Documentación]. Recuperado 10 de octubre de 2019, de sklearn.feature\_selection.SelectKBest—Scikit-learn 0.21.3 documentation website: https://scikit-learn.org/stable/modules/generated/sklearn.feature\_selection.SelectKBest.html#sklearn.feature\_selection.SelectKBest

Shawe-Taylor, J., Williams, C. K. I., Cristianini, N., & Kandola, J. (2005). On the eigenspectrum of the gram matrix and the generalization error of kernel-PCA. *IEEE Transactions on Information Theory*, *51*(7), 2510-2522. https://doi.org/10.1109/TIT.2005.850052

Shin, H. J., Choi, S.-W., & Ok, G. (2018). Qualitative identification of food materials by complex refractive index mapping in the terahertz range. *Food Chemistry*, *245*, 282-288. https://doi.org/10.1016/j.foodchem.2017.10.056

Tarakçi Kalayci, G., Alkaya, A. F., & Algin, R. (2020). Exploitation and comparison of computational intelligence techniques on the feature selection problem. *Advances in Intelligent Systems and Computing*, *1029*, 1243-1249. https://doi.org/10.1007/978-3-030-23756-1\_146

TeraView Ltd. (2019, septiembre 17). TeraView | TeraPulse, Providing Pulsed Imaging and Spectroscopy. Recuperado 17 de septiembre de 2019, de TeraView website: https://teraview.com/terapulse/

Wagstaﬀ, K., Cardie, C., Rogers, S., & Schroedl, S. (2001). *Constrained K-means Clustering with Background Knowledge*. 8.

Wang, K., Sun, D.-W., & Pu, H. (2017). Emerging non-destructive terahertz spectroscopic imaging technique: Principle and applications in the agri-food industry. *Trends in Food Science & Technology*, *67*, 93-105. https://doi.org/10.1016/j.tifs.2017.06.001

Yan, Z., Ying, Y., Zhang, H., & Yu, H. (2006). Research progress of terahertz wave technology in food inspection. *Terahertz Physics, Devices, and Systems*, *6373*, 63730R. https://doi.org/10.1117/12.686840

Zhang, X.-C., & Xu, J. (2009). *Introduction to THz Wave Photonics* (2010 edition). Springer.