UNIVERSIDAD DE PIURA

FACULTAD DE INGENIERÍA

Determinación del contenido de cadmio en granos de cacao mediante la aplicación de redes neuronales e imágenes hiperespectrales

Tesis para optar el Título de
Ingeniera Mecánico - Eléctrica

Keyla Virginia Checa Roman

Asesor:
Dr. Ing. William Ipanaqué Alama

Piura, diciembre de 2022

NOMBRE DEL TRABAJO

09. Tesis_04092022.docx

RECUENTO DE PALABRAS

23654 Words

RECUENTO DE CARACTERES

124819 Characters

RECUENTO DE PÁGINAS

90 Pages

TAMAÑO DEL ARCHIVO

5.7MB

FECHA DE ENTREGA

Sep 7, 2023 8:24 AM GMT-5

FECHA DEL INFORME

Sep 7, 2023 8:26 AM GMT-5

13% de similitud general

El total combinado de todas las coincidencias, incluidas las fuentes superpuestas, para cada base d

- 12% Base de datos de Internet
- 3% Base de datos de publicaciones
- Base de datos de Crossref
- Base de datos de contenido publicado de Crossr
- 7% Base de datos de trabajos entregados

Excluir del Reporte de Similitud

- Material bibliográfico
- Material citado
- Coincidencia baja (menos de 10 palabras)
- Fuentes excluidas manualmente

A Dios y a la Virgen María por darme fortaleza y cuidarme siempre.

A mis seres queridos y en especial a mi madre por su apoyo e inspiración durante esta aventura llena de grandes aprendizajes.

UNIVERSITAS STUDIORUM PIURENSIS

Agradecimientos

A Dios y la Virgen María por todas estas oportunidades que me han formado y enseñado lecciones de vida que atesoro y guardo de manera muy especial en lo más profundo de mi corazón.

A mi madre, mi fuente de inspiración y el motivo que me impulsa a seguir mis sueños y a mis seres queridos por confiar en mí y darme su apoyo en todo momento.

Al Programa Nacional de Becas y Crédito Educativo del Perú (PRONABEC) por creer en mí y haberme dado la oportunidad de estudiar con una beca completamente subvencionada en la Universidad de Piura.

Al Fondo Nacional de Desarrollo Científico, Tecnológico y de Innovación Tecnológica (FONDECYT), que permitió el desarrollo de la investigación titulada "Determinación del contenido de cadmio en granos de cacao mediante la aplicación de redes neuronales e imágenes hiperespectrales" en el marco del esquema de financiamiento "Becas de Mentorías María Reiche 2021-01" [Contrato N°E053-2021-PROCIENCIA]. Además, permitió el desarrollo de una pasantía en el Laboratorio de Robótica y Machine Learning (AIRLab) del Politécnico de Milán, Italia e hizo posible contar con la mentoría de la Dra. Bióloga Hanna Cáceres Yparraguirre, quien sumo esfuerzos en el desenvolvimiento de los objetivos transversales del plan de mentoría como el desarrollo de habilidades blandas, de investigación y de gestión de proyectos.

Al Dr. Ing. William Ipanaque Alama, por su apoyo constante a lo largo del planteamiento, desarrollo y conclusión de esta investigación.

A los miembros del Laboratorio de Sistemas Automáticos de Control de la Universidad de Piura por su apoyo en la elaboración de esta tesis.

A los miembros del Laboratorio de Robótica y Machine Learning (AIRLab) del Politécnico de Milán y al Dr. Ing. Matteo Matteucci por su contribución en esta investigación.

Finalmente, a la Dra. Rachel Atkinson de Alliance Biodiversity- CIAT quién facilitó al Laboratorio de Sistemas Automáticos de Control de la Universidad de Piura las muestras para este estudio.

Resumen

El cacao es uno de los cultivos peruanos con mayor reconocimiento a nivel internacional, declarado como patrimonio natural de la nación (FAO, 2012) debido a sus características organolépticas únicas y de importante contribución al sector agrícola nacional. Este producto es altamente cotizado en el mercado extranjero especialmente en Europa y Asia (Observatorio de Commodities, 2021). El Perú posee el 60% de las variedades de cacao del mundo y el 75% de su producción es cacao fino y de aroma según la Organización Internacional del Cacao (López Cuadra et al., 2020) por lo que se encuentra en el octavo lugar a nivel mundial de producción de cacao en grano. Además, las exportaciones e importaciones de cacao son una fuente importante de ingresos en 16 departamentos del Perú. De tal modo que las plantaciones de este producto abarcan una superficie aproximada de 143 mil hectáreas y tienen una producción de 122 mil toneladas (Ministerio de Agricultura y Riego, 2019).

En el 2019, el Perú enfrentó a una problemática que afectó a miles de productores de cacao orgánico debido a la entrada en vigencia del Reglamento (UE) No 488/2014 de la Unión Europea, en el que se especifican los contenidos máximos de cadmio en los derivados del cacao. En esta normativa, se estipula que el chocolate con un contenido de materia seca total de cacao mayor e igual al 50% puede contener como máximo niveles de cadmio menores a 0.80 µg/g (Reglamento UE N° 488/2014, 2014). Sin embargo, los comercializadores de cacao, quienes juegan un papel intermedio entre los productores y las grandes empresas confiteras, utilizaban esta norma para establecer niveles máximos de contenido de cadmio a los granos de cacao, los cuales no están cubiertos por el reglamento (Banco de Desarrollo de América Latina, 2020). Esto trajo como consecuencia un impacto negativo en las exportaciones de cacao de los pequeños y medianos agricultores peruanos debido a que no se tenía claridad con respecto a la aplicabilidad de la norma a los granos de cacao.

Los efectos de la normativa europea de las concentraciones de cadmio en el cacao peruano representan un desafío nacional, aún más si se considera que en el Perú los más de 90 mil productores de cacao utilizan principalmente métodos tradicionales en el proceso productivo y tienen un acceso limitado a la asistencia técnica y a las nuevas tecnologías (Banco de Desarrollo de América Latina, 2018). Es por esto, que se requiere el desarrollo de iniciativas tanto de identificación adecuada de los niveles de cadmio en el cacao, así como también técnicas de mitigación de ese metal pesado en el desarrollo del cultivo. En este sentido es de

suma importancia desarrollar metodologías de vanguardia para la medición del contenido de cadmio que sean no destructivas y empleen menor tiempo que las técnicas tradicionales.

Esta investigación se concentra en implementar metodologías de Machine Learning, Deep Learning combinadas con imágenes hiperespectrales y análisis químico para predecir el contenido de cadmio en muestras de granos de cacao. Los objetivos de esta investigación se centraron en investigar la aplicabilidad de las imágenes hiperespectrales y de las metodologías inteligencia artificial para determinar el contenido de cadmio en granos de cacao seco.

Se aplicaron las metodologías de Machine Learning: Regresión de mínimos cuadrados parciales (PLSR) y Regresión de vectores de soporte (SVR) y de Deep Learning, Perceptrón multicapa (MLP) con retropropagración. Además, se usaron los algoritmos de selección de bandas debido a la gran multicolinealidad de las variables de entrada: Algoritmo de proyecciones sucesivas (SPA) y Muestreo Reponderado Adaptativo Competitivo (CARS). Finalmente, se destacó la superioridad del modelo CARS-MLP para encontrar las bandas espectrales que contienen la información crucial para la predicción del contenido de cadmio.

Se evidenció la aplicabilidad de esta metodología confiable, rápida y basada en métodos no destructivos. Además, de la gran capacidad de las imágenes hiperespectrales para la determinación del contenido de cadmio en granos de cacao secos aplicando modelos de Machine Learning y Deep Learning. Finalmente, se demostró la importancia del procesamiento de los datos en el desarrollo de modelos predictivos, ya que, al tener una mejor calidad de datos como variables de entrada, mejores serán los resultados del modelo.

Tabla de contenido

UNIVERSITAS STUDIORUM PIURENSIS

Lista de tablas

UNIVERSITAS STUDIORUM PIURENSIS

Lista de figuras

Introducción

La presente investigación tiene como propósito el desarrollo de metodologías de Machine Learning y Deep Learning a través de la aplicación de imágenes hiperespectrales y métodos de análisis químico para la determinación del contenido de cadmio en muestras de granos de cacao secos. Como objetivos específicos se tienen tres principalmente: Implementar metodologías basadas en Machine Learning, el desarrollo de modelos de Deep Learning y la aplicación de metodologías de selección de bandas espectrales caracterizadas por poseer la información más relevante para la predicción del contenido de cadmio.

Se divide principalmente en tres capítulos:

En el primer capítulo, se presenta el marco teórico de la investigación, iniciando con los conceptos básicos del cacao orgánico, la descripción botánica, el proceso productivo, la bioacumulación de cadmio y la revisión de estrategias de mitigación de este metal pesado en los cultivos cacaoteros. Además, se hace una recopilación de los fundamentos de la visión hiperespectral, en donde se describe a detalle en qué consisten las imágenes hiperespectrales y sus métodos de obtención con distintos sensores espectrales y a través de los diferentes tipos de técnicas de escaneo.

En el segundo capítulo, se establece la metodología utilizada para la determinación del contenido de cadmio aplicando imágenes hiperespectrales y redes neuronales. Se realiza una revisión de la técnica de adquisición, procesamiento de las imágenes, análisis espectral y análisis químico para realizar en base a estas dos últimas componentes: variables de entrada y variable de respuesta, respectivamente, un modelo capaz de determinar el contenido de cadmio en base a información espectral de una muestra de granos de cacao.

En el tercer capítulo se desarrollan las metodologías de Machine Learning: Regresión de mínimos cuadrados parciales (PLSR) y Regresión de vectores de soporte (SVR) y de Deep

Learning, Perceptrón multicapa (MLP) con retropropagación. Además, se aplican los algoritmos de selección de bandas debido a la gran multicolinealidad de las variables de entrada: Algoritmo de proyecciones sucesivas (SPA) y Muestreo Reponderado Adaptativo Competitivo (CARS). Finalmente se exponen los mejores modelos alcanzados con sus métricas respectivas y se hace una revisión de los resultados experimentales obtenidos de este estudio.

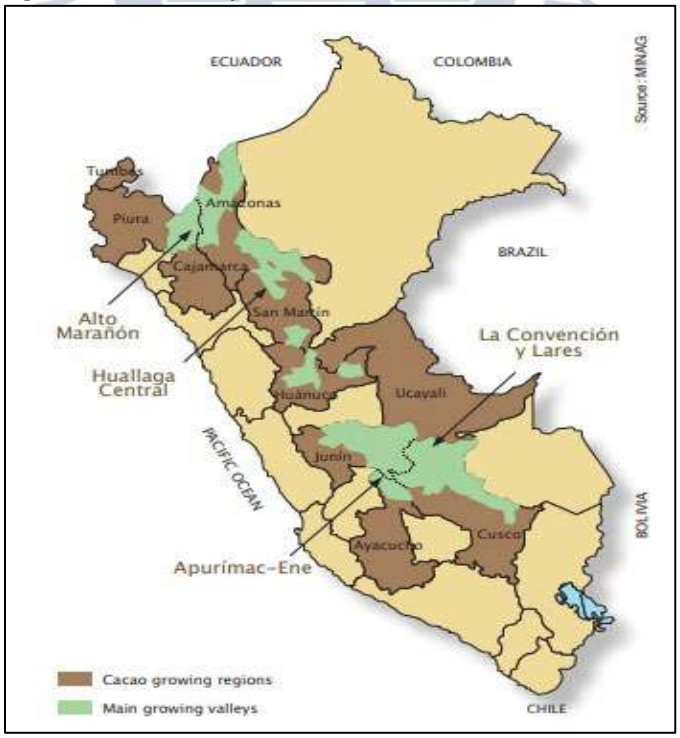
Capítulo 1

Marco teórico

El cacao de nombre científico *Theobroma cacao Linneo,* es una especie nativa de América originaria de la Amazonía Occidental cuyo origen proviene de la región comprendida entre las cuencas de los ríos tributarios del río Amazonas: Caquetá, que recorre Colombia y Brasil, Putumayo que nace en Colombia, desemboca en Brasil y es límite natural entre Perú y Colombia; y Napo que recorre territorios Amazónicos de Ecuador y Perú.

El cacao en el Perú se cultiva en la selva de Loreto, San Martín, Ucayali, Huánuco, Junín, Pasco, Madre de Dios, Cuzco y Ayacucho y en la costa en Tumbes, Piura entre los 0 a los 500 msnm**.** Además, se producen tres variedades de cacao que se componen de un 53.3% del total de producción nacional de la variedad Trinitario en Junín, un 37.3% de Forastero en Cuzco y Ayacucho y un 9.4% de Criollo en la zona norte abarcando San Martín, Amazonas y Cajamarca (Observatorio de Commodities, 2021).

Figura 1. Ciudades productoras de cacao



Nota. **Adaptado de "*Cacao : Propiedades y beneficios del cacao peruano*" PromPerú (2018)**

El cacao peruano es considerado como cacao cien por ciento fino y de aroma, es decir de calidad superior de aromas y sabores únicos cuyas propiedades organolépticas, químicas y físicas se pueden percibir e identificar claramente (Banco de Desarrollo de América Latina, 2018) debido a esto ha logrado el 36% de la producción mundial según cifras de la Organización Internacional del Cacao (ICCO). Además, ha recibido numerosos premios internacionales que lo catalogan como uno de los mejores del mundo. Recientemente, el chocolate Cuzco 80 - Cocoasuyo ganó el galardón de oro en la categoría "Dark Origin" en el evento "Premio Internacional del Chocolate 2021 - Internacional Chocolate Awards 2021" (PromPerú, 2021).

La producción de cacao en grano en el Perú ha venido en aumento desde hace más de diez años a una tasa de 12.6% anual. Durante el 2020, esta se incrementó en un 6.9% comparación del 2019, ya que se dio un aumento de la producción principalmente en Ucayali, San Martin, Junín, Huánuco. Por otro lado, se exportaron US$273.4 millones en productos de cacao y sus derivados, lo cual es una disminución del 7.1% en comparación con el 2019. Esto debido al contexto mundial por la pandemia del COVID-19 lo cual restringió la salida de algunos derivados del cacao, como la manteca de cacao, cacao en grano y chocolate a nivel mundial (Ministerio de Desarrollo Agrario y Riego, 2020).

Según el Ministerio de Desarrollo Agrario y Riego, en enero del 2021 se exportaron 3.9 miles de toneladas de cacao en sus diferentes derivados con un crecimiento del 54% comparado con enero del año anterior, generando 10.6 millones de dólares en sus exportaciones (Ministerio de Desarrollo Agrario y Riego, 2021b).

Por otro lado, en el Observatorio de Commodities de abril y setiembre de 2021, se afirma que entre las regiones que durante los últimos cinco años han tomado un lugar destacable en la producción de cacao se encuentran, San Martín como el mayor productor nacional con 48.4 mil toneladas (35,6%); Junín con 25.5 mil toneladas (18.8%); Ucayali con 17 mil toneladas, es decir un 12.5% de la producción total. Entre las regiones con menores cantidades de producción que han ido en aumento han sido Pasco con un 37% y Piura con un 25% de producción mayor que en años anteriores (Ministerio de Desarrollo Agrario y Riego, 2021a).

En el primer capítulo se hace una revisión de las generalidades del cacao en el Perú, abordando sus características botánicas y los procesos productivos de cosecha y postcosecha que se siguen a nivel nacional. Luego se estudia el cadmio en el ambiente y cómo es que llega a acumularse en el cacao. Además, se analiza la problemática que se genera debido a esto y las estrategias y proyectos peruanos para la mitigación del cadmio en el cacao. Por otro lado, se hace una recopilación de los fundamentos de la visión hiperespectral, en donde se describe a detalle en qué consisten las imágenes hiperespectrales y como se logran obtener con distintos sensores espectrales.

1.1 Antecedentes

A continuación, se presenta la revisión bibliográfica de las investigaciones y libros publicados acerca del tema de investigación, el cual gira en torno a tres puntos principales: el uso de metodologías de inteligencia artificial, el procesamiento de imágenes hiperespectrales y la determinación de metales pesados en productos agrícolas. Por cual, se profundizó en las investigaciones más relevantes que constituyen una base fundamental en el planteamiento y desarrollo de esta investigación.

En el 2017, Arévalo-Gardini et al. analizan los niveles de diferentes metales pesados, entre ellos el cadmio, en muestras de granos y hojas de cacao de 70 plantaciones de diferentes regiones cacaoteras del Perú, entre ellas Tumbes, Piura, Cajamarca, Amazonas, San Martín, Huánuco, Junín y Cuzco. Luego, determinan las concentraciones de metales pesados de las muestras y caracterizan cada uno de los genotipos utilizados. Por consiguiente, concluyen que los lugares de donde se obtuvieron los genotipos con mayores contenidos de cadmio fueron en las regiones de Amazonas, Piura y Tumbes (Arévalo-Gardini et al., 2017).

Por su parte, (Jun et al., 2019) evalúan la capacidad de las imágenes hiperespectrales VIS-NIR, es decir del rango visible y del infrarrojo cercano (400-1000 nm), para la determinación en menor tiempo y no destructiva de los niveles de metales pesados en hojas de tomate sometidos a diferentes tipos de estrés por cadmio. Para su estudio, aplican la técnica de Transformación de la longitud de onda y Regresión por máquina de vectores de apoyo de mínimos cuadrados (WT-LSSVR) con el objetivo de encontrar la longitud de onda óptima para construir el modelo. Se obtuvieron las imágenes hiperespectrales de 405 muestras de hojas de tomate de un invernadero de laboratorio en Jiangsu, China. Se procesaron y en definitiva la técnica aplicada combinada con las imágenes hiperespectrales tienen un gran potencial para la detección del contenido de metales pesados en las hojas de tomate.

En, (Checa et al., 2019) se analiza la relación entre el contenido de cadmio y la firma hiperespectral de granos de cacao orgánico y aplican metodologías de regresión de Machine Learning como Regresión de vectores de soporte (SVR) y Regresión parcial por mínimos cuadrados (PLSR) para la predicción de contenido de cadmio en base a la firma espectral de las muestras de cacao de diferentes zonas cacaoteras de Piura, Perú. Como resultado se encuentra una relación entre el contenido de cadmio y la firma hiperespectral, es decir se puede determinar el contenido de cadmio a través de la firma espectral de una muestra de granos de cacao. Además, las bandas óptimas o más influyentes, dentro del rango de trabajo, para la predicción de niveles de cadmio en granos de cacao.

En el 2020, (Pandey et al., 2020) hacen un análisis del uso de la imagen hiperespectral (HSI) la cual integra la información de imágenes bidimensionales en la amplia gama del espectro electromagnético en los diferentes campos como la arqueología, el arte, control de vegetación, recursos hídricos, seguridad alimentaria, medicina forense, biomedicina, entre

otras. Además, resaltan sus principales características en cuanto a su potencial uso no invasivo y de aplicabilidad en línea tanto en sistemas de laboratorio como de teledetección para agricultura de precisión. Asimismo, se hace una revisión de las metodologías de Machine Learning y Deep Learning con las que en conjunto conforman una herramienta poderosa para la detección de enfermedades, metales pesados, componentes químicos presentes en un objeto de estudio determinado.

En el 2020, (Xin et al., 2020) aplican técnicas de Inteligencia Artificial, auto-encoders y Regresión por máquina de vectores de soporte de mínimos cuadrados parciales (LSSVR) para desarrollar un modelo de detección de contenido de cadmio haciendo uso de espectroscopía VIS-NIR. Se tienen como objeto de estudio 1120 muestras de hojas de lechuga de un invernadero ubicado en Jiangsu, China y se concluyó que los métodos de aprendizaje profundo junto con las imágenes hiperespectrales tienen un gran potencial para predecir el contenido de metales pesado en las hojas de lechuga.

En el 2020, en un estudio realizado en España por (Zea, 2020) plantea un estudio aplicando HSI para predecir concentraciones de cadmio en dos plantas de hoja verde: la col rizada y la albahaca, usando muestras con y sin el uso de biochar o carbón activado, sometidas a distintas concentraciones de cadmio en el suelo. Para los modelos de predicción se hizo uso de metodologías de Machine Learning como Análisis de Componentes Principales (PCA) y Regresión de mínimos cuadrados (PLS); y de Deep Learning aplicando Redes neuronales artificiales (ANN), obteniendo con esta última metodología los mejores resultados para predecir los niveles de contenido de cadmio en las verduras de hoja verde.

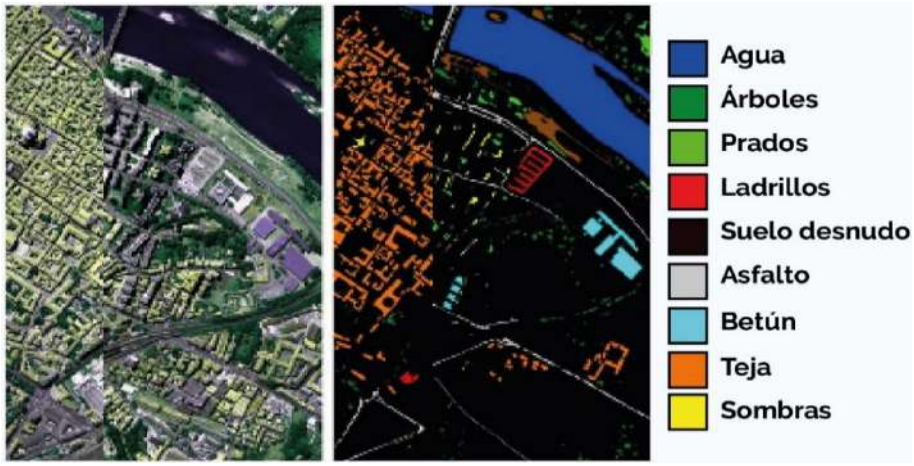
El aporte más reciente alrededor de este tema, es el desarrollado por (Saeidan et al., 2021) en Irán. El cual es un estudio de discriminación de materias extrañas como madera, plástico, piedra y órganos de plantas en 250 muestras de grano de cacao. En base a la información de las imágenes hiperespectrales tomadas de las muestras efectúan el estudio aplicando las técnicas de Discriminante Lineal de Máquina de Vector de Soporte (LDA-SVM) y K-Vecinos más Próximos (KNN). Además, utilizan PCA para encontrar las bandas óptimas en donde se tiene la información más relevante para la clasificación. Concluyen que el mejor clasificador es LDA-SVM cuando se usan las bandas óptimas como variables de entrada para la clasificación de los mariales extraños presentes en los granos de cacao.

1.2 Estado del arte

Recientes investigaciones afirman que los algoritmos de aprendizaje profundo tienen buen rendimiento al combinarse con imágenes hiperespectrales (Lu et al., 2020). Por lo que esta combinación ha sido muy utilizada en los últimos años en aplicaciones en distintos campos, entre ellos la evaluación de la calidad y bioseguridad de productos agrícolas y alimentarios como frutas (Jun et al., 2019), hortalizas (Bao et al., 2019), cereales, entre otros.

Entre las aplicaciones se pueden diferenciar dos tipos: clasificación y regresión y cada una de ellas utiliza técnicas de Machine Learning y Deep Learning específicas de acuerdo a la naturaleza del problema planteado. Para la clasificación, debido a que se tiene un conjunto de datos que necesitan agruparse en una clase determinada se utilizan técnicas de Machine Learning SVM, KNN, Árboles de decisión, Bosques Aleatorios, entre otras. Por otro lado, cuando se tienen imágenes de grandes áreas de cultivo a través de aplicaciones de teledetección y agricultura de precisión se utilizan técnicas de Deep Learning en donde se busca diferenciar áreas con enfermedades en el cultivo y tipos de áreas de componentes específicos (Figura 2). Entre las más usadas están 2D y 3D Red neuronal convolucional (CNN), Auto-Encoders (AE), Redes Residuales, Red de creencia profunda (DBN), Red neuronal recurrente (RNN), entre otras (Paoletti et al., 2019).

Figura 2. Aplicaciones en teledetección en campos abiertos



***Nota*. Adaptado de "Deep Learning Applications for Hyperspectral Imaging: A Systematic Review" Ozdemir & Polat (2020)**

En el caso de aplicaciones de regresión, en donde se desea determinar en base diversas variables independientes relacionadas entre sí el valor de la(s) variable(s) respuesta o dependiente. Se tienen aplicaciones de Machine Learning como Regresión Linear, SVM, Árboles de decisión, Bosques Aleatorios y de Deep Learning como Red neuronal artificial (ANN), Perceptrón multicapa (MLP).

En esta investigación se presenta un problema de regresión debido a que se desea determinar el contenido de cadmio a través de información espectral de una muestra de granos de cacao por lo que se desarrollan diversas metodologías de Deep Learning para obtener el mejor modelo predictor.

1.3 Taxonomía y descripción botánica

La taxonomía del objeto de estudio pertenece a la clase *Magnoliopsida (= Dicotyledoneae),* de orden *Malvales*, de familia *Sterculiaceae*, de género *Theobroma* y de especie *cacao* (Fundación Charles Darwin, 2022). A continuación se presenta la descripción botánica del cacao se describe según (Dostert et al., 2012).

1.4 Morfología

1.4.1 Planta

Es un árbol semicaducifolio de hasta 12 a 20 m de altura en general y en cultivo se mantiene normalmente a 4 a 8 m. El tallo es lampiño, la corteza del tronco es café oscuro y las ramas son cafés y vellosas.

1.4.2 Hojas

Son coriáceas angostamente ovadas y elípticas, levemente asimétricas, alternas y glabras o laxamente pubescentes en ambas caras.

1.4.3 Flores

Son pentámeras, hermafroditas, actinomorfas, y 5 a 20 mm de diámetro, el pedúnculo floral es de 1 a 3 cm de largo. Los sépalos son verdosos, blancos o rosa claro, 5 a 8 mm de largo, 1.5 a 2 mm de ancho, angostamente lanceoladas, persistentes y fusionados en la base.

1.4.4 Fruto

Es una mazorca, polimorfa, esférica a fusiforme, púrpura o amarilla en la madurez, glabro, 10 a 35 cm de largo y 7 cm ancho. Llega a tener un peso de 200 a 1000 gr. y se caracteriza por poseer de 5 a 10 surcos longitudinales en la cascara. El endocarpio es de 4 a 8 mm de grosor y es de aspecto duro, carnoso, y leñoso.

1.4.5 Semilla

Son café-rojizas, ovadas, ligeramente comprimidas de 10 a 30 mm de largo, 12 a 16 mm de ancho y 7 a 12 mm de grosor aproximadamente.

1.5 Descriptores morfológicos de las semillas

1.5.1 Forma

1.5.1.1 Forma de la sección longitudinal. Puede tener forma oblonga, elíptica, ovada o irregular (Figura 3).

Figura 3. Forma de la sección longitudinal

Nota. **Adaptado de "*Cultivares de cacao en el Perú*" (García Carrión, 2010) y Adaptado de "Protocolo para la caracterización morfológica de árboles élite de cacao (Theobroma cacao L.)" Compañía Nacional de Chocolates (2018)**

1.5.1.2 Forma de sección transversal. Puede tener forma aplanada, intermedia y redondeada (Figura 4).

Figura 4. Forma de la sección transversal

***Nota*. Adaptado de "*Cultivares de cacao en el Perú*"(García Carrión, 2010) y Adaptado de "Protocolo para la caracterización morfológica de árboles élite de cacao (Theobroma cacao L.)" Compañía Nacional de Chocolates (2018)**

1.5.2 Color

Los cotiledones pueden ser de color blanco, rosado, violeta, morado y moteado o manchado (Figura 5).

Figura 5. Color de los cotiledones

***Nota*. Adaptado de "Protocolo para la caracterización morfológica de árboles élite de cacao (Theobroma cacao L.)" Compañía Nacional de Chocolates (2018)**

1.5.3 Textura

La textura del fruto de las diferentes variedades del cacao se divide en cuarto principalmente: ausente, ligera, intermedia e intensa.

1.6 Variabilidad

1.6.1 Criollo

Crecen bajo condiciones semi - silvestres y se distribuyen desde México hasta Colombia, Venezuela, Nicaragua y Guatemala. Son árboles poco vigorosos y delgados, de lento crecimiento y más susceptibles a las enfermedades e insectos que los Forasteros. Además, son conocidos por su alta diversidad morfológica. Entre sus características más destacables se tiene:

- Bajo rendimiento y mayor susceptibilidad a enfermedades.
- Menos amargo y más aromático que los demás.

- Mazorca de cascara fina en comparación a las demás y de color amarillo o rojo.
- Contiene de 20 a 30 granos aproximadamente.

En la Figura 6 y Figura 7 se muestra el fruto y la semilla respectivamente del tipo de grano criollo de nombre varietal “Guasare”.

Figura 6. Fruto del cacao criollo Guasare

***Nota*. Adaptado de “*Cultivares de cacao en el Perú*” García Carrión (2010)**

Figura 7. Semilla del cacao criollo Guasare

***Nota.* Adaptado de “*Cultivares de cacao en el Perú*” García Carrión (2010)**

1.6.2 Forastero

Es la variedad más abundante y representa el 80% de la producción mundial. Mayormente se desarrollan en estado silvestre y domesticado en la Amazonía alta del Perú, Ecuador, Colombia, Brasil, África Central y el Caribe. Se caracterizan por ser árboles robustos que poseen frutos verdes y de forma variable. Se puede encontrar variedad en el color de los cotiledones, como por ejemplo cotiledones blancos como en el Porcelana de Piura. En el caso de la producción de la variedad Forastero del Alto Amazonas es de calidad corriente o básica (García Carrión, 2010). Entre sus características más destacables se tiene:

- Cotiledones amargos y marrones o púrpuras.
- Mazorcas amarillas y duras.
- Mejor rendimiento y menos susceptible a enfermedades.
- Contiene 30 a 60 granos aproximadamente.

En la Figura 8 y Figura 9 se muestra el fruto y la semilla respectivamente del tipo de grano forastero de nombre varietal “PA -150”.

Figura 8. Fruto del cacao forastero PA - 150

***Nota*. Adaptado de "*Cultivares de cacao en el Perú*" García Carrión (2010)**

Figura 9. Semilla del cacao forastero PA - 150

***Nota*. Adaptado de "*Cultivares de cacao en el Perú*" García Carrión (2010)**

1.6.3 Trinitario

Originario de la Isla Trinidad, son árboles rara vez se han encontrado en estado silvestre y en su mayoría se caracterizan por poseer propiedades similares a los Criollos y Forasteros debido a que son de origen híbrido de ambas formas. Es una forma muy irregular genética y morfológicamente, no siendo posible de determinar a simple vista. Aproximadamente el 15% de la producción mundial es de cacao Trinitario (Dostert et al., 2012). Entre sus características más destacables se tiene:

- Mayor rendimiento y menor sensibilidad a las enfermedades que el forastero.2
- Contiene 30 a 45 granos aproximadamente.

En la Figura 10 y Figura 11 se muestra el fruto y la semilla respectivamente del tipo de grano trinitario de nombre varietal "ICS -1".

Figura 10. Fruto del cacao trinitario ICS - 1

***Nota*. Adaptado de "*Cultivares de cacao en el Perú*" García Carrión (2010)**

Figura 11. Semilla del cacao trinitario ICS - 1

***Nota*. Adaptado de "*Cultivares de cacao en el Perú*" García Carrión (2010)**

1.7 Proceso productivo del cacao

Para obtener los granos de cacao como producto final, se realizan una serie de productivos característicos de este sembrío. La cosecha y la postcosecha son procesos cruciales en la calidad del cacao y deben ser efectuados bajo ciertos parámetros de control. En este apartado se profundiza en el proceso productivo considerando la cosecha y la apertura como punto de partida para desarrollar cada uno de los procesos siguientes hasta conseguir el objeto de estudio de esta investigación que es el grano de cacao seco.

1.7.1 Cosecha y apertura

La maduración completa se reconoce de manera empírica a través de la variación de color en el fruto. El proceso de cosecha cosiste en retirar manualmente los frutos de los árboles, donde el pedúnculo se corta cuidadosamente con un utensilio afilado como se muestra en la Figura 12. Una práctica habitual de los agricultores consiste en almacenar las mazorcas de cacao que se han cosechado sin abrir aproximadamente 12 días, dependiendo de cada cultivar, ya que esperan acumular una cantidad de mazorcas adecuada para la fermentación. En el Perú generalmente las épocas del año de cosecha de mayor producción son de enero a julio y se hace con una frecuencia de 11 días. En cuanto a la cosecha de menor producción se realiza de agosto a noviembre y se hace con una frecuencia de medio mes (Viera, 2018).

Figura 12. Cosecha del cacao

***Nota*. Adaptado de *"Caja de herramientas para cacao: Aprendiendo e Innovando sobre el Manejo Sostenible del Cultivo de Cacao en Sistemas Agroforestales. Cosecha, fermentación y secado del cacao"* Lutheran World Relief (2013)**

El proceso de apertura de las mazorcas que han llegado a la maduración adecuada se realiza con un mazo y seguido a esto se realiza la extracción de los granos. Además, parte del proceso consiste en separar los granos que estén dañados, germinados, inmaduros, o infectados por hongos e insectos (Figura 13).

Figura 13. Proceso de apertura de mazorcas de cacao en ASPROBO - Buenos Aires-Morropón, Piura

***Nota*. Adaptado de Laboratorio de Sistemas Automáticos de Control. Universidad de Piura (2019)**

1.7.2 Postcosecha

El proceso de postcosecha tiene una repercusión fundamental en el aroma y el contenido de polifenoles y una mala práctica de estos pueden generar granos no germinados o mohosos que degradan la calidad del grano del cacao.

1.7.2.1 Fermentación. Durante este proceso se produce la muerte del embrión, la eliminación de la pulpa mucilaginosa, así como la formación de precursores de las cualidades organolépticas de aroma, sabor y color, haciendo de la fermentación una etapa imprescindible y crucial para la calidad del grano de cacao (Augstburger et al., 2000; Schwan & Wheals, 2004).

La duración de la apertura y el comienzo del proceso de fermentación debe ser como máximo un día. Los granos extraídos se colocan en cajones de madera (Figura 14), cestas tejidas u otros recipientes adecuados que estén hechos de material natural, sin recurrir al metal, los cuales estarán aislados y protegidos de las variaciones bruscas del clima (Augstburger et al., 2000). Existe otro método más rudimentario que el anterior utilizado en África Occidental que consiste en apilar los granos sobre trozos de madera dispuestos radialmente cubiertos en la parte superior por hojas de plátano (Nigam & Singh, 2014).

Figura 14. Fermentación en cajas de madera

El proceso de fermentación dura de 7 a 8 días, variando entre cada cultivar. Consiste en una transformación bioquímica que tiene dos etapas: Anaeróbica, ausencia de oxígeno y aeróbica con presencia de oxígeno.

En la primera etapa se genera la transformación de los azúcares del mucílago, en alcohol etílico y luego en ácido láctico, por la intervención, en las primeras 24 horas, de levaduras y, de las 24 a 36 horas, de las bacterias lácticas. En carencia de oxígeno se incrementan las diferentes especies de levaduras, lo que genera un incremento de temperatura en la masa de fermentación entre 30°C y 40°C, un pH menor a 4 con una acidificación del medio, y esto a su vez, la descomposición de la pulpa y la salida del jugo (Augstburger et al., 2000; Cardona Velásquez, 2016; Lagunes Gálvez et al., 2007). La proliferación de levaduras conlleva a la producción de etanol y secreción de enzimas pectinolíticas. Luego de las 24 horas se da una disminución de la población de levaduras, lo que favorece el crecimiento de las bacterias ácido láctico (BAL), que alcanzan su pico después de alrededor de 36 horas desde el inicio de la fermentación, cuya actividad principal es degradar la glucosa en ácido láctico (Saltini et al., 2013).

En la segunda etapa se da luego de las 48 horas con la incorporación de oxígeno. La población de las bacterias ácido lácticas (BAL) decrece y se da paso al crecimiento de las bacterias de ácido acético (BAA). Las reacciones exotérmicas de las BAA consisten en la oxidación del etanol en ácido acético, seguido de su oxidación en dióxido de carbono y agua, lo que produce un incremento la temperatura hasta los 50°C o más. Por lo que se dan

diferentes reacciones químicas que producen la formación de los precursores de sabor y aroma. En este punto, se tiene como resultado debido a la acidez, un sabor amargo y un olor a amoníaco. La agitación regular es necesaria para promover la aireación, a fin de lograr una fermentación rápida y uniforme (Augstburger et al., 2000; Lagunes Gálvez et al., 2007; Saltini et al., 2013; Schwan & Wheals, 2004).

Para conocer el estado de la fermentación, los productores hacen la prueba de corte que consiste en el fraccionamiento o corte de cierta cantidad de granos para determinar por su coloración el grado de fermentación de los granos.

1.7.2.2 Secado. Partiendo de una humedad de los granos de un 55% luego de la fermentación, en el secado los granos deben llegar a tener una humedad de entre 6 a 7% lo que indica su calidad (Augstburger et al., 2000; Dostert et al., 2012; Nigam & Singh, 2014; Saltini et al., 2013). En este proceso continúan desarrollándose las reacciones iniciadas en la fermentación, disminuye el amargor, se completan las reacciones de oxidación responsables del aroma y del sabor y se produce un cambio del color de los granos a un tono café oscuro (Augstburger et al., 2000; Cardona Velásquez, 2016; Dostert et al., 2012; Nigam & Singh, 2014).

El secado al sol, que toma alrededor de 7 días, consiste en el aprovechamiento del sol con la exposición de los granos de cacao sobre tejidos de yute, superficies de madera como parihuelas de madera (Figura 15), tarimas entre otros. En este método, los granos de cacao se deberán remover constantemente y es reconocido como el método que da mejor resultado en cuanto a la calidad del grano (Isla Ramírez & Andrade Adaniya, 2009; Nigam & Singh, 2014; Saltini et al., 2013). Sin embargo, no es muy eficiente ante climas húmedos y variables, homogeneidad del secado, tiempo y mano de obra.

Figura 15. Secado en cajas de madera

Por su parte, en el secado artificial se hace uso de secadores adiabáticos, donde se expone el sólido a un gas caliente, que principalmente es aire. Además, se hace uso de secadores no adiabáticos donde la transferencia de calor se puede realizar transportando los sólidos sobre una superficie horizontal que aumenta su temperatura usando vapor de agua. También se puede realizar este procedimiento, al reposar los sólidos sobre una superficie

caliente de forma cilíndrica haciendo uso de agitadores (Parra Rosero, 2017). No obstante, según (Saltini et al., 2013) el secado artificial puede mejorar el proceso de secado, haciendo que se pueda hacer una estandarización y los lotes de granos de cacao después de este proceso sean más homogéneos, pero se requiere más investigación.

Para dar por concluido el proceso de secado se hace la prueba del corte, en donde se secciona transversalmente el grano y se verifica la existencia de espaciamientos o grietas. (Isla Ramírez & Andrade Adaniya, 2009)

1.7.2.3 Selección. Consiste en seleccionar los granos de cacao que se encuentran en condiciones óptimas para su comercialización. Principalmente se pueden diferenciar dos métodos de selección: manual y mecánico. En el primer método se procede a retirar toda la materia extraña que se encuentre junto con los granos y separar todos los granos pizarrosos, aglomerados, planos, rotos y mohosos. Mientras que en el segundo método se hace uso de una máquina tamizadora que puede hacer el paso anterior o no, (en ese caso se hace manualmente) y selecciona los granos por tamaños (End & Dand, 2015; Ruiz, 2016).

1.7.2.4 Almacenamiento. Después del secado, los granos de cacao se reservan en sacos de yute o cestas que permitan la ventilación de los mismos. Se debe tener en cuenta que, los granos de cacao son altamente sensibles a los aromas del ambiente debido a su alto contenido de grasas. Por lo tanto, lo óptimo es que sean almacenados en ambientes pulcros, bien ventilados, evitando el contacto con el suelo o las paredes, productos químicos y separados de otros cultivos. (Augstburger et al., 2000; Dostert et al., 2012; Isla Ramírez & Andrade Adaniya, 2009).

1.8 Bioacumulación de cadmio en cacao

El cadmio es un metal de fácil absorción que se presenta en diversas formas en los suelos, sin embargo necesita estar disponible para su absorción lo que depende de la especie y genotipos de la planta debido a la variación morfológica, las etapas de crecimiento de la planta, edad, tipo de cultivo y las condiciones fisicoquímicas del suelo (El Rasafi et al., 2022).

Las raíces ejercen un rol clave en la absorción de metales del suelo. Las características de las raíces como la estructura y el tamaño son determinantes para la absorción de cadmio. Recientes estudios señalan que las raíces peludas y delgadas muestran una alta absorción y acumulación de metales (Gupta et al., 2019). Además, la absorción de metales por las raíces está controlada por diversos factores del suelo como el contenido soluble de oligoelementos en el suelo, el pH del suelo, la materia orgánica, la capacidad de intercambio catiónico, fertilizantes, tipo de suelo entre otros (Gupta et al., 2019).

1.9 Generalidades del cadmio

El cadmio es un metal pesado no esencial, considerado uno de lo más tóxicos; para las plantas, los animales y los seres humanos; y más móviles del medio ambiente (Kubier & Pichler, 2019; Lewis et al., 2018). Es considerado un oligoelemento o metal traza debido a que

tiene una densidad relativamente alta y tóxica incluso a baja concentración (Haider et al., 2021).

1.10 Principales fuentes

El principal sumidero de contaminantes es el suelo. Además, La mayoría de los contaminantes inorgánicos no sufren degradación química ni microbiana por lo que su concentración persiste en el suelo durante largos periodos de tiempo (Kubier & Pichler, 2019).

Según (Haider et al., 2021), existen diferentes factores de contaminación de suelos por cadmio que se deben a numerosas actividades antropogénicas y emisiones al medio ambiente. Esto se da principalmente por la alta concentración de metaloides y metales traza a través de las emisiones del sector industrial en expansión, la eliminación descontrolada de residuos con alto contenido de metales, gases de escape de vehículos, los plaguicidas, fertilizantes sintéticos, abonos, riego de aguas residuales, vertidos petroquímicos, residuos mineros, deposición atmosférica, lodos de depuradora, entre otros. Entre las fuentes naturales se encuentran la erupción volcánica, la meteorización de las rocas y la erosión (Gupta et al., 2019).

1.10.1 Las aguas residuales y de alcantarillado

Utilizadas para el riego de los cultivos son las principales fuentes de metales traza en el medio ambiente. Estudios realizados a través de años muestran que se ha producido un considerable riego a lo largo del tiempo con aguas residuales, lo cual ha traído como consecuencia la acumulación de metales pesados en los cultivos. Además, se demuestra que el riego a largo plazo con aguas residuales aumenta de manera considerable el contenido de metales pesados como el cadmio de 2.0 a 3.4 mg/kg, níquel de 9 a 19 mg/kg, cromo de 33 a 225 mg/kg y plomo de 22 a 41 mg/kg en los cultivos (Gupta et al., 2019).

1.10.2 Pesticidas y fertilizantes

Las prácticas agrícolas que usan indiscriminadamente los pesticidas y fertilizantes para aumentar el rendimiento de los cultivos elevan la concentración de metales pesados en los cultivos. Entre los fertilizantes mayormente usados se encuentran los basados en fosfatos como el Superfosfato Triple (TSP) y el Fosfato Di-amónico (DAP) debido a que el fósforo se considera como un nutriente esencial para el desarrollo de los cultivos. Estos fertilizantes contienen cadmio, plomo, arsénico, cromo y zinc con una concentración que depende de su tipo y marca, lo cual causa la aumenta el contenido de estos metales pesados en los cultivos (Gupta et al., 2019).

En el caso de los pesticidas, un estudio realizado en la Nigeria por (Bawa et al., 2021) en el que se analizaron las concentraciones de metales pesados como el cadmio, cromo, cobre y zinc en 20 plaguicidas usados mayormente en las zonas de estudio reveló los siguientes resultados: concentraciones de 0.10 a 6.05 para el cadmio, 2.33 a 46.44 para el plomo, 3.98 a 18.10 para el cromo, 0.62 a 5.69 para el cobre y 0.34 a 34.11para el zinc.

1.10.3 Actividad industrial

Las múltiples actividades industriales contribuyen directamente a través de la liberación de efluentes industriales, residuos sólidos o indirectamente con emisiones de gases a la atmósfera que después se depositan suelo, agua y aire. Los suelos cercanos a zonas industriales tienen mayor vulnerabilidad a la contaminación por metales pesados debido al vertido de efluentes no tratados adecuadamente y la eliminación de residuos sólidos en el área (Gupta et al., 2019).

Dependiendo de la actividad industrial, se asocian metales específicos dependiendo de sus productos y procesos de manufactura. Por lo cual las industrias de cemento están asociadas a la generación de un alto nivel cadmio, cromo, cobre, plomo y zinc a la atmósfera debido a que se usa el níquel, cobalto, plomo y cobre como catalizadores y secadores. Por su parte, las curtidurías están asociadas a la generación de cromo y el zinc en la producción de agroquímicos como los fertilizantes. Las refinerías de petróleo están asociadas al plomo, y el níquel se asocia con emisiones petroquímicas. Por otro lado, cerca de una industria de fundición se encontró acumulación foliar de cadmio, antimonio, zinc y plomo en espinacas y coles. La gasolina y sus fuentes relacionadas son las que tienen una mayor contribución de plomo, cadmio, cobre, zinc y níquel. El consumo de aceite de motor es responsable de la mayor emisión de cadmio, el desgaste de neumáticos contribuye a las emisiones de zinc y el desgaste de los frenos a las emisiones de cobre y plomo (Gupta et al., 2019).

1.10.4 Minería

Es una de mayores fuentes de contaminación principalmente en países en desarrollo. Durante el proceso de explotación minera, algunos metales pesados se generan y se esparcen en fosas abiertas y parcialmente cubiertas.

Tabla 1. Principales países mineros del mundo de algunos ETs con su dosis oral de referencia

ETs	Principales países mineros del mundo	Dosis oral de referencia (mg/kg/día)
As	China, Chile, Marruecos, Rusia	3.0 × 10−4
Cd	China, Corea, Japón, México, Canadá	1.0 × 10−3
Cr	Sudáfrica, Kazakstán, India, Turquía, Rusia	1.5 × 10−1
Cu	Chile, China, Perú, Australia, Estados Unidos	4.0 × 10−2
Hg	China, Kyrgyzstan, Chile, Rusia	3.0 × 10−4
Ni	Filipinas, Rusia, Brasil, Indonesia, Canadá	2.0 × 10−2
Pb	China, Australia, Estados Unidos, Perú, México, India	4.0 × 10−3
Zn	China, Australia, Perú, India, Estados Unidos	3.0 × 10−1
Mn	Sudáfrica, Australia, China	1.4 × 10−1

***Nota*. Adaptado de *"Trace elements in soil-vegetables interface: Translocation, bioaccumulation, toxicity and amelioration - A review"* Gupta et al. (2019).**

En la Tabla 1, se muestra una lista de los principales países mineros del mundo con algunos de sus ETs (oligoelementos o elementos traza) y su dosis oral de referencia según la Agencia de Protección Ambiental de Estados Unidos (USEPA) (Gupta et al., 2019).

En el Perú, (Castro-Bedriñana et al., 2021) realizaron un estudio en zonas cercanas a las industrias minero-metalúrgica de los Andes centrales del Perú y se encontraron altas concentraciones de plomo de 577 ± 18.2 µg/kg y de cadmio de 18.35 ± 5.4 µg/kg en la leche cruda de vaca.

1.11 Toxicidad

La toxicidad del cadmio afecta directamente a varios órganos del cuerpo humano. Recientes estudios señalan que se acumula principalmente en los riñones y es la causa de graves daños pulmonares y renales como enfisema pulmonar, reducción de la función de reabsorción tubular renal y cálculos renales (Mahajan & Kaushal, 2018). Así mismo, provoca daños en el hígado y a los huesos generando osteoporosis ya que el cadmio puede llegar a reducir la absorción de calcio. Esto debido a que tienen una carga idéntica, similar comportamiento y radio iónico (Kubier & Pichler, 2019; Mahajan & Kaushal, 2018). Por lo que, según la Organización de las Naciones Unidas para la Agricultura y la Alimentación (FAO) y la Organización Mundial de la Salud (OMS) la exposición tolerable de cadmio en una persona es de 25 µg/kg del peso corporal por mes, es decir 62 µg/kg para una persona de 70 kg (Satarug et al., 2017).

La toxicidad del cadmio afecta a las plantas a nivel morfológico , bioquímico, fisiológico y molecular (El Rasafi et al., 2022). La presencia de cadmio en las plantas afecta en su crecimiento, debido a que reduce las concentraciones de nitratos, inhibe la fijación de carbono y disminuye el contenido de clorofila. Esto genera desequilibrios en el metabolismo del cloroplasto lo que impacta directamente en el desarrollo de la fotosíntesis, reduce la absorción de hierro y zinc provocando clorosis (Llatance et al., 2018; Xu et al., 2017), genera un crecimiento desviado de la planta y conduce a su necrosis (Haider et al., 2021). El cadmio puede dar lugar a desordenes fisiológicos en las plantas, debido al aumento de estrés oxidativo es decir se da una sobreproducción de especies reactivas de oxígeno (ROS) (El Rasafi et al., 2022).

En cuanto a las consecuencias en el ADN que produce la presencia del cadmio en las plantas principalmente genera la destrucción de las membranas celulares y ácidos nucleicos, daños y disminución en las proteínas fotosintéticas, lo que influye en el crecimiento de todo el organismo (Haider et al., 2021).

1.12 Problemática y contexto actual

Según el último reporte del Ministerio de Agricultura y Riego de abril a setiembre del 2021, en el transcurso de los últimos siete años las exportaciones de cacao y sus derivados tuvieron un ascenso con una taza de aumento de 0.5% en promedio al año. Como se observa

en la Tabla 2 , entre los principales derivados de cacao mayormente exportados se tiene el cacao en grano, entero o partido, crudo o tostado siendo un 77% de las exportaciones; la manteca de grasa y aceite de cacao y el cacao el polvo (Observatorio de Commodities, 2021).

Tabla 2. Perú: Exportaciones agregadas de cacao y sus derivados (En miles de US$)

AÑO	2015	2016	2017	2018	2019	2020	2020*	2021*
TOTAL	**266972**	**293681**	**235341**	**257232**	**294263**	**273437**	**199797**	**207624**
Cacao en grano, entero o partido, crudo o tostado	192274	201669	148357	152772	153463	145747	112652	104560
Manteca, grasa y aceite de cacao	42940	54455	50274	64488	88 997	65 913	44 964	55 528
Chocolate y demás preparaciones alimenticias que contengan cacao	15960	14974	18220	21890	26 581	25 396	18 662	21 007
Cacao en polvo sin adición de azúcar ni de otro edulcorante	10642	12161	13418	12303	15 591	21 077	13 122	15 440
Pasta de cacao, incluso desgrasado	5128	10313	4976	5724	9 493	15 222	10 348	10 944
Cáscara, películas y demás residuos de cacao	28	210	97	55	137	82	49	146

***enero-septiembre**

Nota*. Adaptado de *"Observatorio de Commodities: Cacao" Observatorio de Commodities (2021).

En el periodo de enero a setiembre del 2021, se enviaron al extranjero aproximadamente US$208 millones en cacao y sus derivados, con un incremento de 3.9% respecto al mismo periodo del año anterior, esto a causa de una mayor demanda de manteca de cacao, chocolate, cacao en polvo y grano.

En las exportaciones durante los tres últimos años del cacao en grano, entero o partido, crudo o tostado ha ido en disminución debido a varios factores. En el año 2019, las exportaciones se redujeron en 59.7 miles de toneladas y de 53.7 miles de toneladas en el 2020, con una tasa de reducción del 9.9% debido al impacto de la entrada en vigencia del reglamento de la Unión Europea y la pandemia del COVID-19.

En el año 2021, de enero a setiembre se exportaron 39,2 mil toneladas, lo que supone una caída de 6% respecto al mismo periodo en el año anterior. Los países en donde se han disminuido las exportaciones principalmente son en Bélgica que cayeron en 3,8 mil toneladas y Estados Unidos en 1,6 mil toneladas. Estos países son los principales mercados del cacao peruano, donde también se suman Indonesia, Bélgica, Malasia, Italia, México y España, los cuales son en su mayoría países europeos.

En cuanto a las exportaciones de manteca, grasa y aceite de cacao, el segundo producto más exportado de los derivados del cacao peruano se han tenido resultados más alentadores en los tres últimos años. Se registró una exportación de US$89 miles de dólares siendo la más alta en los últimos 10 años. Sin embargo, en el 2020, debido a los acontecimientos anteriormente mencionados se dio una caída del 22.2 % del volumen de producción, debido a una menor demanda de los países europeos en un 28%, de Holanda en un -50.0%, Inglaterra de un -24.9% y Alemania de un -1.8%.

En base a lo presentado anteriormente se conoce el contexto de la problemática de la que parte la presente investigación. A continuación, se expone una de las razones de fuerza por las cuales se ha dado una disminución de la demanda y exportaciones de cacao peruano en su principal mercado, la Unión Europea.

En el año 2014 la Unión Europea publicó el *"Reglamento (UE) No 488/2014 de la Comisión de 12 de mayo del 2014 que modifica el Reglamento (CE) Nº1881/2006 por lo que respecta al contenido máximo de cadmio en los productos alimenticios"* en el que se especifican los niveles máximos de cadmio en los derivados del cacao que entraban en vigencia el 1 de enero del 2019 (Tabla 3) (Reglamento UE N° 488/2014, 2014) y que su vez puso en peligro las exportaciones peruanas del cacao y sus derivados debido a que se ha encontrado un contenido de cadmio mayor al reglamentado en ellas.

Tabla 3. Niveles máximos para el cacao y sus derivados por la *Unión Europea*

Productos específicos de cacao y chocolate	Contenido en µg/g
- Chocolate con leche con un contenido de materia seca total de cacao < 30 %	0.10 a partir del 1 de enero de 2019
- Chocolate con un contenido de materia seca total de cacao < 50 %; chocolate con leche con un contenido de materia seca total de cacao ≥ 30 %	0.30 a partir del 1 de enero de 2019
- Chocolate con un contenido de materia seca total de cacao ≥ 50	0.80 a partir del 1 de enero de 2019
- Cacao en polvo vendido al consumidor final o como ingrediente en cacao en polvo edulcorado vendido al consumidor final (chocolate para beber)	0.60 a partir del 1 de enero de 2019

***Nota*. Adaptado de *"Reglamento (UE) No 488/2014 de la Comisión de 12 de mayo del 2014 que modifica el Reglamento (CE) Nº1881/2006 por lo que respecta al contenido máximo de cadmio en los productos alimenticios"* Reglamento UE N° 488/2014 (2014)**

Frente a esta normativa el Perú desde el 2017 hasta la actualidad a presentado diversas solicitudes a la Comisión de la Unión Europea, donde plantea su preocupación sobre este tema, debido a que no se proporciona un nivel máximo correspondiente al insumo que se comercializa que son los granos de cacao, sino que está referida a los derivados del mismo. Esto hace que no sea aplicable en la práctica debido a que no se tiene una medida referencial con la cual fijar el contenido máximo de cadmio a los granos de cacao. Sin embargo, no se ha obtenido hasta ahora una solución a esta problemática que sigue afectando a la producción de cacao peruano por parte de la Unión Europea.

En febrero del 2018, el Perú envió sus observaciones en el informe *G/SPS/GEN/1602* (Ministerio de Agricultura y Riego del Perú, 2018a) en el cuales expuso que la exposición al cadmio por el consumo de los derivados del cacao no está clasificada como un riesgo según las directrices JEFCA (Comité Mixto FAO/OMS de Expertos en Aditivos Alimentarios). Además, solicitó a la Unión Europea que excluya los derivados del cacao del ámbito de la aplicación del Reglamento (UE) N° 488/2014 hasta que el Comité del Codex sobre Contaminantes en los Alimentos (CCCF) haya podido emitir una norma de referencia internacional, y se haya proporcionado evidencia científica adecuada, evitando así restricciones comerciales

innecesarias (Sistema de gestión de información sanitaria y fitosanitaria, 2019). En junio del 2018, en *G/SPS/GEN/1624* (Ministerio de Agricultura y Riego del Perú, 2018b) solicita extender el periodo de inicio de la ejecución de resolución hasta inicios del 2022.

En contra parte la Unión Europea expresó que las nuevas medidas en el Reglamento (UE).N° 488/2014, en cuestión, fue emitida en el 2014 y se había dado un plazo de cinco años para su aplicación en enero de 2019. Además, refirió que los niveles máximos que figuran en la normativa están dados por EFSA (*European Food Safety Authority*).

Recientemente el Perú presentó una solicitud a la Organización Mundial de Comercio (OMC) a mediados del 2020 en donde se solicita un sustento con una norma del *Codex Alimentarius* sobre el tema (Observatorio de Commodities, 2021). Esto debido a que actualmente no existe una norma de este ente internacional al respecto de los niveles máximos de contenido de cadmio en el cacao.

Por su parte, el Perú durante este proceso ha desarrollado una serie de proyectos junto a entidades nacionales e internacionales para mitigar el contenido de cadmio en el cacao orgánico. El Ministerio de Desarrollo Agrario y Riego (MIDAGRI) ha realizado diversas acciones para nutrir sus lazos con los pequeños y medianos productores centros cacaoteros del país. A través del Instituto Nacional de Innovación Agraria (INIA), organizaron cursos de entrenamiento en San Martín a un total de 50 productores a través de talleres de manejo integral de plagas y enfermedades, sistema de riego, fertilización entre otros temas cruciales en el manejo agronómico de este cultivo. Además, a través de Agroideas se han aprobado 85 planes de negocio en el proceso productivo de cacao beneficiando a seis mil productores de Amazonas, Ayacucho, Huánuco, entre otras.

1.13 Lugares con mayor contenido de cadmio en el Perú

En el 2017, Arévalo-Gardini et al. analizaron diferentes niveles de metales pesados, como el cadmio, cromo, cobre, hierro, magnesio, níquel, plomo y zinc en muestras de hojas y granos de cacao en 70 plantaciones cacaoteras del Perú, entre ellas Tumbes, Piura, Cajamarca, Amazonas, San Martín, Huánuco, Junín y Cuzco. Los resultados del estudio muestran que el 57% de las muestras recogidas superaron el límite crítico de 0.8 µg/g para el cadmio impuesto por la Unión Europea (Reglamento UE N° 488/2014, 2014). Las concentraciones medias de contenido de cadmio en las ciudades de estudio estuvieron en un rango de 0.17±0.41 µg/g donde Cuzco ocupa el primer lugar, con menor contenido de cadmio, hasta 1.78±0.35 µg/g en donde Tumbes es el lugar en donde se encontró el mayor contenido de cadmio. Además, Piura, uno de los lugares de estudio de la presente investigación, ocupó el segundo de lugar de mayor contenido de cadmio. En cuanto a las variedades, las concentraciones medias de contenido de cadmio van desde la variedad Nativo de Satipo con 0.15 ± 0.8 µg/g hasta 1.8 ± 0.4 µg/g de híbridos espontáneos. Por lo que los mayores valores medios de concentraciones de metales pesados en granos de cacao en los genotipos muestreados se registraron en Blanco

Piurano (Zn); CCN51/ICS95/ICS39 (Cr, Cu); Híbridos Espontáneos (Cd); Nativo de Marañón (Fe) y Nativo de Satipo (Mn, Ni, Pb) (Arévalo-Gardini et al., 2017).

En 2018, (Llatance et al., 2018), analizaron el potencial fitorremediador de siete especies vegetales: *Pouteria caimito, Matisia cordata Bonpl., Malvaviscus sp., Vochysia sp., Carludovica palmata, Attalea sp.* y *Theobroma cacao L.,* las cuales fueron expuestas al cadmio de manera natural en la Comunidad Nativa de Pakun, Amazonas. Como resultado, la especie *Theobroma cacao L.,* presentó mayores niveles de cadmio que las otras especies. Además, se concluyó que una planta de cacao acumula mayor contenido de cadmio como sigue: Raíz > Tallo > Hoja > Fruto, según los experimentos realizados, siendo el grano de cacao en donde se acumula menor contenido de cadmio.

1.14 Mecanismos para mitigar el contenido de cadmio en las plantas

Las plantas en suelos contaminados desarrollan múltiples mecanismos para hacer frente al estrés por cadmio. Según (El Rasafi et al., 2022; Haider et al., 2021), se pueden distinguir dos mecanismos principales para mitigar los efectos tóxicos del cadmio en las plantas: la prevención y la tolerancia. El primero, incluye la limitación de absorción de cadmio en la planta. Por otra parte, la tolerancia considera la acumulación del cadmio mediante su unión de péptidos, aminoácidos y proteínas y la inmovilización del mismo en algunos tejidos de la planta. La tolerancia de las plantas al cadmio se potencia a través de la producción de fitoquelatinas, la síntesis de antioxidantes y la restauración de los pigmentos de las plantas. Además, hacen frente al estrés por cadmio mediante la inmovilización de la pared celular, la inducción de enzimas y proteínas antioxidantes y la compactación vacuolar (El Rasafi et al., 2022; Haider et al., 2021).

La hiperacumulación es una estrategia de mitigación que implica la redistribución y la captación del cadmio para reducir su contenido en la planta. Las plantas hiperacumuladoras retienen los iones metálicos que se toman del suelo en las células de la raíz, los desintoxica en el citoplasma almacenándolos en las vacuolas y los traslada rápidamente a los tallos de la planta con la ayuda del xilema (El Rasafi et al., 2022; Haider et al., 2021).

Según (Haider et al., 2021), aproximadamente 450 especies de angiospermas fueron clasificadas como hiperacumuladoras de metales hasta 2011 y se continúan encontrando nuevas plantas con flores con este potencial. Un estudio desarrollado por (Liu et al., 2022) en China muestra el potencial hiperacumulador de las siete plantas con flores para remediar de forma sostenible los suelos urbanos contaminados por cadmio. Luego de 60 días de exposición a 100 mg/kg de cadmio, *Calendula officinalis L.* fue la que mostró una gran capacidad de tolerancia, en términos de un aumento significativo de biomasa y sin cambios evidentes en su altura. A su vez acumuló 279.51 ± 13.67 μg/g, lo cual está por encima del valor crítico definido para un hiperacumulador (100 μg/g). Al final del experimento, pudo acumular hasta 926.68 ± 29.1 μg/g en la raíz y 1206.19 ± 23.06 μg/g en la planta por lo que puede convertirse en una potencial hiperacumuladora de cadmio para la fitorremediación.

Por su parte, el uso de biochar o biocarbón en los suelos agrícolas contaminados con cadmio ha logrado relevante importancia en los últimos años. Se trata de un material carbonoso natural y poroso formado por ausencia de oxígeno por pirólisis de estiércol orgánico y residuos de cultivos. La incorporación del biocarbón minimiza la disponibilidad de cadmio y su consiguiente acumulación y toxicidad en las plantas debido a sus propiedades fisicoquímicas (Haider et al., 2021).

1.15 Imágenes hiperespectrales

La imagen hiperespectral (HSI) combina la espectroscopía, ciencia relacionada con la emisión o reflexión de luz que se produce en los materiales (Ozdemir & Polat, 2020) y la imagenología convencional para obtener información espectral e información espacial de una muestra. Es así como se genera una pila de imágenes bidimensionales que conforman un cubo de datos (Jia et al., 2020).

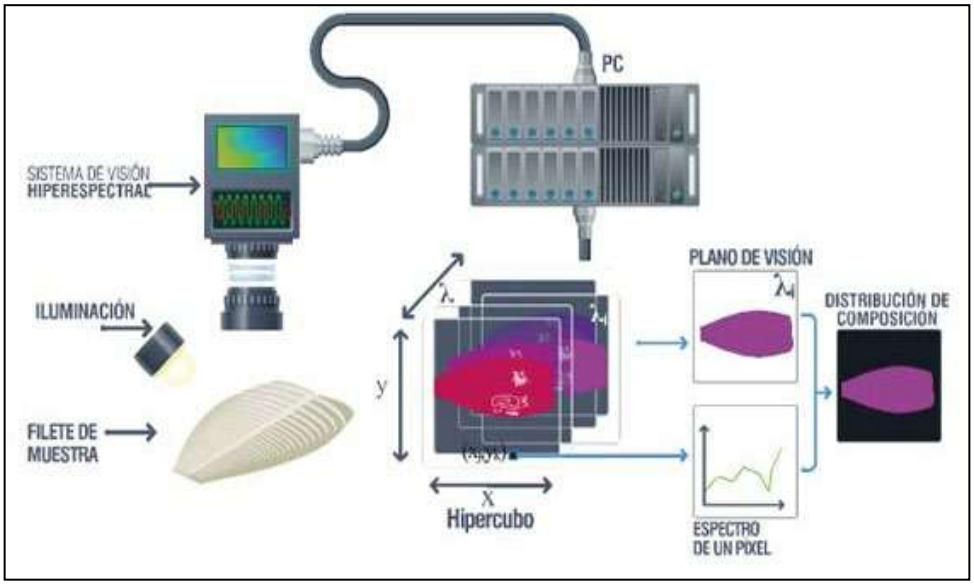
Las cámaras hiperespectrales adquieren información de un objeto a lo largo del espectro electromagnético, lo cual permite identificar con mayor precisión su composición físico-química. Además, se caracterizan por su potencial uso no invasivo y de aplicabilidad tanto en sistemas de laboratorio, sistemas robóticos terrestres y aéreos en teledetección para su utilización en la agricultura de precisión (Pandey et al., 2020).

El acceso rápido y la confiabilidad de la concentración de metales pesados en productos agrícolas, como la determinación del cadmio en los granos de cacao que se propone en esta investigación, son cruciales para el monitoreo de los cultivos y la remediación de suelos. Por lo tanto, se desarrollaron metodologías para la medición del contenido de cadmio con menos tiempo y complejidad que las técnicas tradicionales.

La imagen hiperespectral está formada por bandas espectrales contiguas, una después de la otra dependiendo del rango de trabajo de la cámara y de su resolución espectral a lo largo del espectro electromagnético. Es decir, a diferencia de las imágenes ordinarias de RGB en donde se tiene información de tres bandas: rojo, verde y azul; las imágenes hiperespectrales tienen una alta resolución espectral y abundante información de numerosas bandas espectrales. Al adquirir una imagen hiperespectral, esta actúa como una huella espectral única del objeto de estudio. Por lo que esta información puede reflejar la estructura física y la composición química del objeto de estudio (Lv & Wang, 2020).

La imagen hiperespectral está conformada por un hipercubo que contiene datos tridimensionales (x, y, λ) que comprenden dos dimensiones espaciales (x, y) y una dimensión de la longitud de onda (λ) (Figura 16). Dependiendo del sensor con el que se realice la adquisición de las imágenes hiperespectrales se tendrán distintos rangos de trabajo dentro del espectro electromagnético. Se pueden diferenciar tres tipos de bases de datos dependiendo de los sensores utilizados para aplicaciones específicas.

Figura 16. Representación esquemática de un sistema de visión hiperespectral y de un hipercubo



Nota*. Adaptado de *"Visión hiperespectral para inspección de calidad" TECHPRESS (2014)

1.15.1 Técnicas de monitoreo usando imágenes hiperespectrales

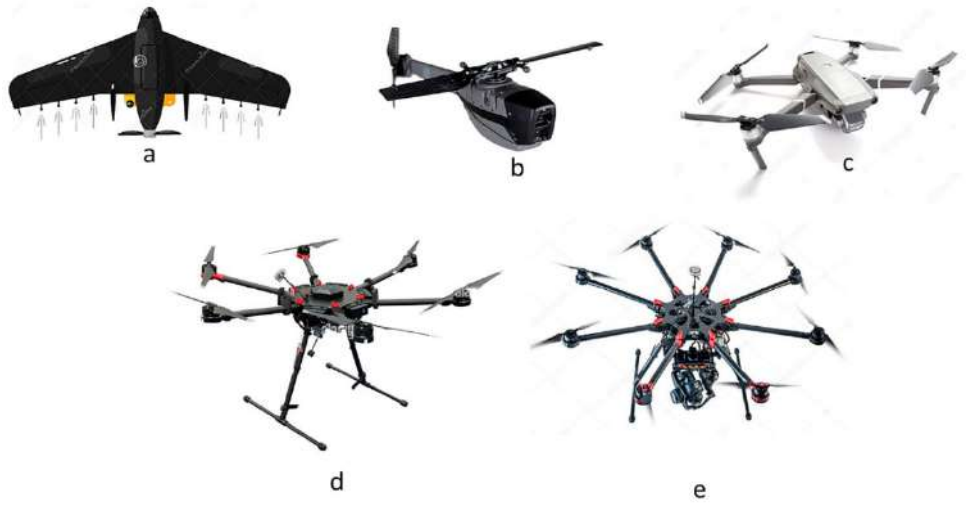
1.15.1.1 Imágenes de UAV y sensores. Existe un gran número de aplicaciones en donde se utilizan este tipo de sensores, principalmente en la agricultura de precisión.

Un UAV (Vehículo aéreo no tripulado) es un avión sin piloto humano controlado por un canal de radio. Se caracterizan por sus diversas aplicaciones en la adquisición de imágenes aplicando metodologías de topología y geodésica para el procesamiento de mosaicos georreferenciados de grandes extensiones de terreno. Esta tecnología es aplicada mayormente en el área de la agricultura, debido a se puede realizar el monitoreo de cultivos de manera más automatizada, lo que reduce costos y aumenta el rendimiento de la cosecha. Además, esta tecnología ha llevado a desarrollar mapas de estrés, infestación de plagas, estudios de estado del suelo, control de enfermedades, entre otros (Awais et al., 2022). Existen en general cinco modelos de UAV: de ala fija, de rotor simple con un solo motor, cuadricóptero, hexacóptero y octocóptero (Figura 17). Dependiendo de cada modelo, los UAVs soportan un peso máximo y sobrevuelan a una velocidad determinada (Saddik et al., 2022).

En cuanto a los sensores, se pueden identificar dos tipos principalmente: cámaras multiespectrales e hiperespectrales. Ambas se caracterizan por ser un dispositivo que capta varias longitudes de onda en un solo plano que permite un análisis con mayor precisión en menor escala (Saddik et al., 2022). En cuanto a las diferencias entre estas cámaras se resumen principalmente en que las cámaras multiespectrales poseen un número reducido de bandas específicas no contiguas dentro del espectro electromagnético. Mientras que las cámaras

hiperespectrales suelen abarcar rangos mayores de bandas espectrales contiguas dentro del espectro electromagnético.

Figura 17. Modelos de vehículos aéreos no tripulados



***Nota*. Adaptado de "Computer development based embedded systems in precision agriculture: tools and application" Saddik et al. (2022)**

Las cámaras multiespectrales más utilizadas son Parrot Sequoia+ (Figura 18) y RedEdge-MX (Figura 19). La primera que posee cuatro cámaras monocromáticas con diferentes filtros de banda estrecha en el dominio visible e infrarrojo cercano (NIR) (Cubero-Castan et al., 2018). Además, cuenta con cuatro bandas espectrales: Verde (550 nm), Roja (660 nm), Borde Rojo (735 nm) y NIR (790 nm). Mientras que la cámara RedEdge-MX ofrece cinco bandas espectrales: Borde Rojo (717 nm), Azul (475 nm), Rojo (668 nm), Verde (560 nm) y NIR (840 nm). Ambas se caracterizan porque pueden calcular mediante su software integrado, diferentes índices de vegetación y se usan principalmente para aplicaciones relacionadas con fenotipado y mapeo de salud de cultivos, manejo de fertilizantes, identificación de enfermedades, detección de malezas, modelados de terreo, entre otras (Geotop, 2022; Saddik et al., 2022).

Figura 18. Imagen de la cámara Parrot Sequoia+

***Nota*. Adaptado de PIX4D (2022)**

Figura 19. Imagen de la cámara RedEdge-MX

***Nota*. Adaptado de Geotop (2022)**

En cuanto a las cámaras hiperespectrales más utilizadas son de las marcas: Hyspex, Specim (Saddik et al., 2022). Dentro del catálogo de Hyspex, ofrecen cámaras para sistemas aerotransportados de alta resolución y velocidad con bajo peso, para sistemas de laboratorio que cubren rangos espectrales de 400 a 1000 nm (VNIR: visible e infrarrojo cercano), 930 a 2500 nm (SWIR: infrarrojo de longitud de onda corta) con diferentes resoluciones y para aplicaciones en campo ofrecen cámaras montables con trípode. En el caso de aplicaciones para UAV ofrece un cuadricóptero (Figura 20) con opciones para dos cámaras hiperespectrales: VNIR (400 a 1000 nm) y SWIR (1000 a 2500 nm) y para aplicaciones industriales ofrecen cámaras que cubren el rango visible e infrarrojo (Neo, 2022).

Figura 20. Hyspex UAV

***Nota*. Adaptado de Neo (2022)**

Specim también ofrece una gran gama de cámaras hiperespectrales tanto para aplicaciones industriales como para la investigación, como sistemas aerotransportados, sistemas de geología, cámaras espectrales y espectrógrafos. Además, sus cámaras cubren gran

parte del espectro electromagnético, como el rango de luz visible (380 - 800 nm), luz visible e infrarrojo cercano (400 - 1 000 nm) (Figura 21), infrarrojo cercano (900 - 1 700 nm), infrarrojo cercano de onda corta (1 000 - 2 500 nm), infrarrojo cercano de onda media (2,7 - 5,3 µm) e infrarrojo cercano de onda larga (8 – 12,4 µm) (Specim, 2022).

Figura 21. Cámara hiperespectral Specim FX10

***Nota*. Adaptado de Specim (2022)**

Estudios realizados a lo largo del tiempo han demostrado un rendimiento superior de las imágenes hiperespectrales sobre las imágenes multiespectrales en el análisis de propiedades en la vegetación, la evaluación del contenido de metales pesados en diferentes cultivos y la discriminación de diferentes tipos de sembríos. Esto debido a que son capaces de captar mayor información y por ende de detectar variaciones sutiles a lo largo del tiempo (Lv & Wang, 2020).

1.15.1.2 Imágenes de satélite. Principalmente utilizadas para la vigilancia o monitoreo de grandes campos agrícolas a gran escala. Por su parte las imágenes multiespectrales como las generadas por los satélites Landsat 9 (NASA) y Sentinel 3 (European Space Agency) poseen un número más reducido de bandas en comparación con las cámaras hiperespectrales convencionales debido a las aplicaciones por las que fueron lanzados. Esto en algunas aplicaciones suele tener limitaciones en la resolución espectral y en la precisión de las variables recuperadas(Lv & Wang, 2020).

Landsat 9 fue lanzada el setiembre del 2021 por la NASA en asociación con el Servicio Geológico de Estados Unidos y está compuesto de dos instrumentos: el Operational Land (OLI-2) y el sensor infrarrojo térmico 2 (TIRS-2). OLI-2 es una cámara multiespectral posee bandas 11 espectrales de resolución espacial de 30 metros y cubre un área de 185 km. TIRS-2, detecta la radiación emitida por la Tierra utilizando la tecnología Quantum Well Infrared Photoreceptors (QWIPS). Es usado para múltiples aplicaciones, entre ellas: determinar índices vegetales, identificaciones de ecosistemas, zonas oscuras como bosques y aguas costeras (USGS & NASA, 2019).

En el caso de Sentinel-3 OLCI lanzado en el 2016, posee una cámara hiperespectral de 21 bandas que abarca el rango de 400 a 1020 nm con una resolución de 300 m y posee un campo de visión de 290 km. Esta cámara es usada principalmente para determinar índices vegetales relacionados con la clorofila y el contenido de agua para aplicaciones agrícolas, forestales y seguridad alimentaria (ESA, 2015; Saddik et al., 2022).

1.15.1.3 Robots de suelo. Tienen una gran ventaja en el monitoreo de invernaderos a menor escala. Se trata de robots móviles diseñados de manera personalizada para las aplicaciones por las que fue desarrollado. Estos poseen en su mayoría de sensores de GPS, odómetros, guías lineales, planes de trayectoria, cámaras multiespectrales, entre otros sensores. Además, suelen tener sistemas adicionales montados para su aplicación específica como brazos o tubos de riego.

El robot GRAPE (Ground Robot for VineyArd Monitoring and ProtEction) es un ejemplo de robot de suelo desarrollado en colaboración conjunta entre Agile X Robotics y el Politécnico de Milán. Está constituido por un chasis compacto de tipo Scout (Agile X, 2022), compuesto por cuatro servomotores sin escobillas con un sistema de rotación diferencial y de suspensión independiente de doble eje que permite giros suaves y estables. En cuanto a los sensores que se encuentran sobre el chasis, posee sensores ZED 2 para la detección de la profundidad, sensores IMU (Unidad de Medición Inercial) y un acelerómetro para la medición de las velocidades angulares y aceleraciones, una estación GPS para identificar la ubicación dada una referencia. Además, dependiendo de la aplicación se pueden montar cámaras portables RGB, multiespectrales e hiperespectrales que cubren el rango visible y el infrarrojo cercano (Figura 22). Este robot se usa en aplicaciones de agricultura para la navegación de campo y mapeo de cultivos (METRICS, 2022).

Figura 22. Robot de suelo GRAPE (Ground Robot for VineyArd Monitoring and ProtEction) Politécnico de Milán - 2022 ACRE 1ra campaña de campo – Montoldre, Francia

1.15.2 Metodologías para la obtención de imágenes hiperespectrales

Existen tres técnicas para la captura de imágenes hiperespectrales: reflectancia, transmitancia e interactancia. La distribución de la fuente de luz, que puede ser los focos

halógenos y el detector óptico, el lente, varían según el modo de adquisición (Wu & Sun, 2013) como se muestra en la Figura 23.

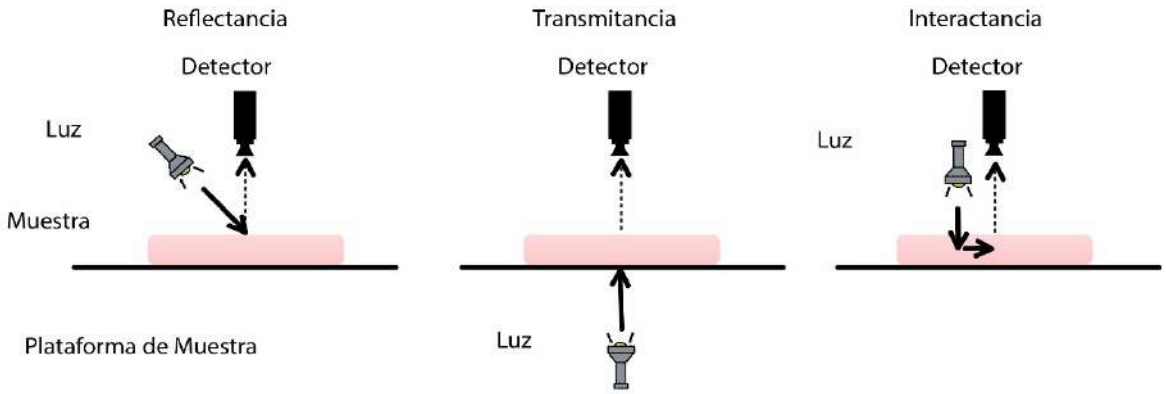
1.15.2.1 Reflectancia. El detector captura la luz reflejada del objeto de estudio iluminado en una conformación específica para evitar la reflexión especular. Las características externas de calidad son típicamente detectadas usando un modelo de reflectancia, como tamaño, forma, color, textura de la superficie y defectos externos.

1.15.2.2 Transmitancia. El detector se encuentra en la posición lado opuesto de la fuente de luz y captura los datos transmitidos que a menudo es muy débil. Se utiliza normalmente para determinar la concentración de componentes internos, así como también para detectar rasgos internos de diferentes materiales. Sin embargo, el modo de transmisión tiene un nivel de señal bajo debido a la atenuación de la luz y se ve afectado por el grosor de la muestra.

1.15.2.3 Interactancia. En este modo, tanto la fuente de luz como el detector están localizados paralelos a la muestra, dando como resultado una detección de información más profunda reduciendo la influencia del espesor y tiene menos defectos superficiales en comparación con el modo de reflectancia.

En esta investigación se ha utilizado el método de detección de la reflectancia, la cual es la información de la muestra que proporciona la imagen hiperespectral tomada por la cámara Resonon Pika II.

Figura 23. Modos de detección de imágenes



Nota*. Adaptado de *"Advanced applications of hyperspectral imaging technology for food quality and safety analysis and assessment: A review — Part I: Fundamentals" Wu & Sun (2013)
***Las direcciones de escaneo se muestran con flechas**

1.15.3 Método de escaneo

Existen diversas técnicas de obtención de imágenes hiperespectrales a través de sensores, las cuales tienen sus propias características y requisitos de complejidad del plano focal y la estabilidad de la plataforma (Chang, 2006):

- Escáneres de línea (Line Scanners)
- Escáneres Whiskbroom (Whiskbroom Scanners)
- Escáneres Pushbroom (Pushbroom Scanners)
- Cámaras de encuadre (Framing Cameras)

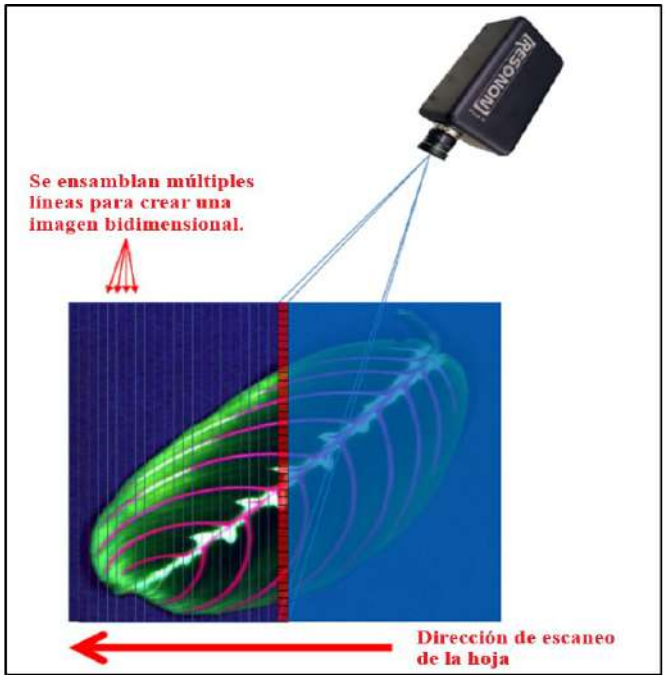
En el siguiente apartado se profundizará en el método de escaneo pushbroom debido a que es el utilizado para la presente investigación.

1.15.3.1 Método de escaneo Pushbroom.

La configuración usada por la cámara de esta investigación, Resonon Pika II, conocida como *"pushbroom"* implica la adquisición de mediciones espectrales simultáneas desde una serie de posiciones espaciales adyacentes, lo que requiere un movimiento relativo entre el objeto y el detector, donde la muestra se mueve mientras la cámara está estacionaria.

Los espectrómetros de imágenes por resonancia magnética son escáneres de línea, lo que significa que recogen datos de una línea, es decir se generan varias imágenes línea por línea a la vez para ensamblar una imagen bidimensional completa (Figura 24) (Resonon, 2017).

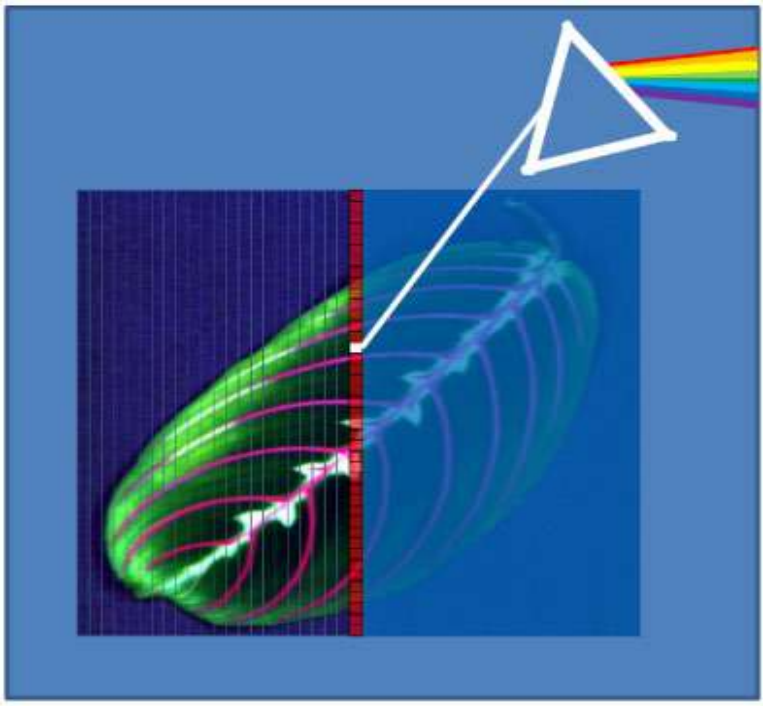
Figura 24. Construcción de imagen hiperespectral



Nota*. Adaptado de *"SpectrononPro Manual" Resonon (2017)

Para obtener los datos hiperespectrales, la señal de cada píxel se difracta en sus componentes espectrales de manera similar al pasar la luz de cada píxel a través de un prisma. Este proceso ocurre para cada píxel de la línea, (cuadros rojos de la Figura 25) recogiendo toda la información espectral como el color para cada uno. El resultado es una curva espectral detallada para cada píxel de la imagen (Resonon, 2017).

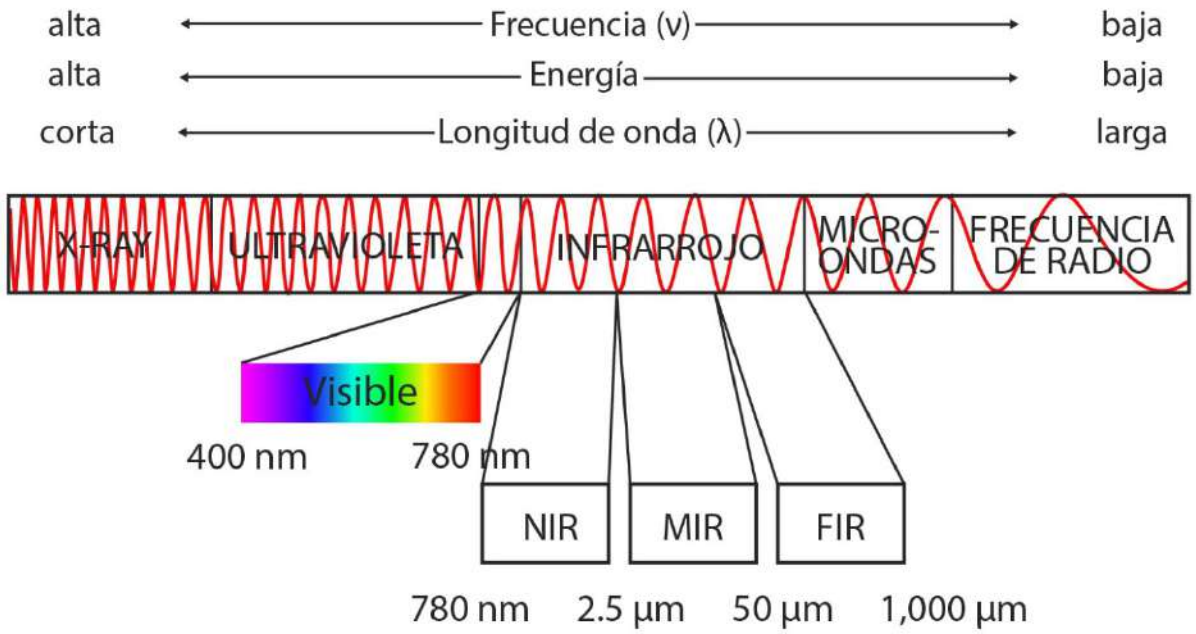
Figura 25. Proceso de difracción de los componentes espectrales



Nota*. Adaptado de *"SpectrononPro Manual" Resonon (2017)

Los hipercubos obtenidos utilizando esa configuración se almacenan en el formato *Band Interleaved by Line* (BIL), muy utilizado en aplicaciones de la industria alimentaria (Gowen et al., 2007).

Figura 26. Espectro electromagnético: Longitud de onda (nm y µm)



***Nota*. Adaptado de: *"Automatización del análisis de imágenes hiperespectrales para identificación de aptitud de patatas"* Ayala Martini (2018)**

Dependiendo del tipo de espectrómetro se pueden adquirir imágenes hiperespectrales en distintos rangos del espectro electromagnético (Figura 26). El sistema más utilizado en análisis de alimentos es el que trabaja en el rango VIS, VIS-NIR y NIR. En (Morin et al., 2017) se hace un estudio de la respuesta de la vegetación frente al estrés ambiental y los cambios que se dan en los distintos rangos del espectro electromagnético. Tomando como referencia que el estrés causa una reducción de los pigmentos en las hojas, en los que se destaca la clorofila, así como también los carotenoides, las xantofilas y las antocianinas, caracterizados por ser los principales absorbentes de la luz en las plantas en la parte visible del espectro de 400 a 700 nm; se analiza el cambio en la firma hiperespectral del objeto de estudio.

UNIVERSITAS STUDIORUM PIURENSIS

Capítulo 2

Materiales y métodos

En este capítulo se realiza una revisión de la metodología experimental empleada en esta investigación haciendo énfasis en tres componentes principales: la adquisición, el procesamiento de las imágenes hiperespectrales y el análisis químico de las muestras. Además, se hace una caracterización del perfil de la muestra y los rangos de trabajo definidos los cuales son clave para construcción y aplicación del modelo predictor.

Por otro lado, se exponen las dos técnicas propuestas para definir las variables de entrada del modelo. En una de ellas se considera toda la firma espectral obtenida de la información espectral al capturar la imagen y la segunda considera la utilización de algoritmos de selección de bandas óptimas para encontrar las bandas espectrales que explican con mayor precisión la variable a determinar que es el contenido de cadmio.

2.1 Área de estudio y perfil de la muestra

El estudio consistió en 99 plantaciones, 14 de ellas situadas en las ciudades de Piura y 85 en Huánuco del Perú como se observa en Figura 27 y Figura 28. De todas las plantaciones de las zonas cacaoteras, se recolectaron 300 muestras de cacao seco, las cuales fueron seleccionadas aplicando la normativa de muestreo estandarizada propuesta por el Ministerio de Agricultura y Riego en la Resolución Ministerial N°0451-2018-MINAGRI (Ministerio de Agricultura y Riego, 2018). Esta normativa contiene las pautas de muestreo para la determinación de los contenidos de cadmio en suelos, hojas, granos y productos derivados de cacao. Además, tiene como objetivo principal que exista un instrumento de referencia para que se realice una adecuada toma de muestras bajo estándares nacionales e internacionales.

El objeto de estudio fueron granos de cacao seco al seis por ciento de humedad. La muestra de estudio consistió en 500 gr. de granos de cacao seco y se dividió en dos submuestras. La primera submuestra se envió a un laboratorio externo para determinar su contenido de cadmio y con la segunda submuestra se realizó la captura de las imágenes hiperespectrales. A esta segunda submuestra se le aplicó un submuestreo aleatorio, donde se tomaron ciertos granos de la submuestra, para capturar la imagen de esos granos aislados y

así asignar una firma promedio a la submuestra total. En la Figura 30, se muestra la imagen de un grano de cacao, el cual constituye el objeto de estudio de esta investigación.

En la Figura 27 y Figura 28 se muestran los gráficos de barras con la frecuencia de las muestras de la presente investigación en cada uno de los lugares de estudio, Piura y Huánuco respectivamente.

Figura 27. Gráfico del contenido de cadmio en µg/g de las muestras de Piura

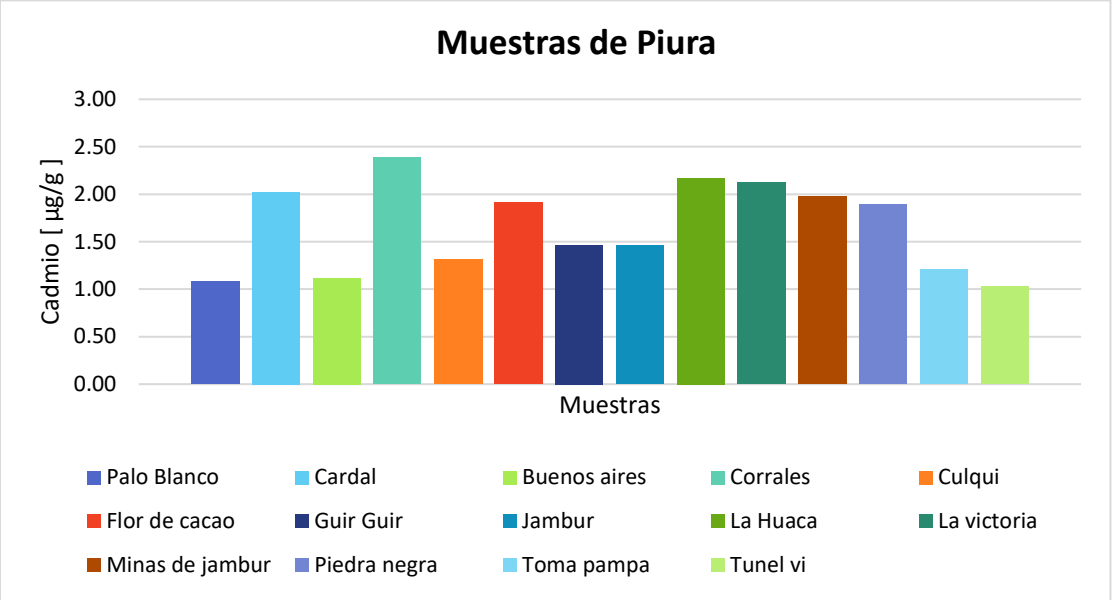
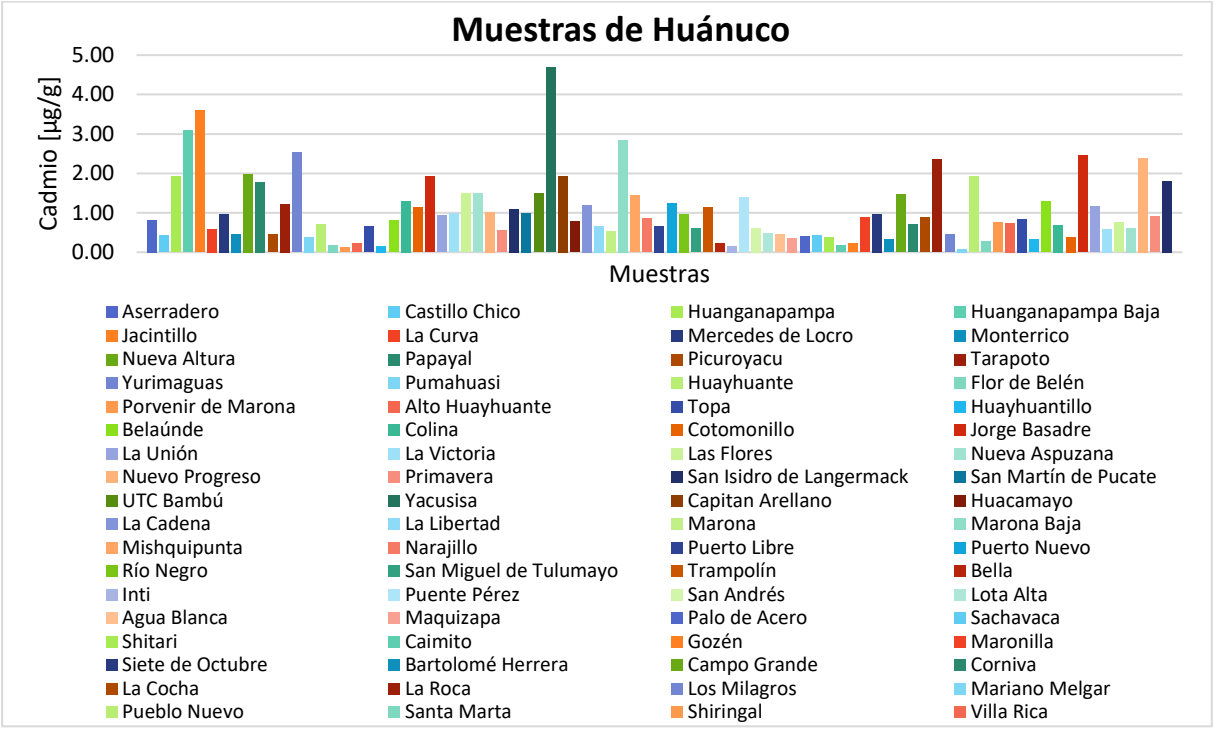


Figura 28. Gráfico del contenido de cadmio en µg/g de las muestras de Huánuco

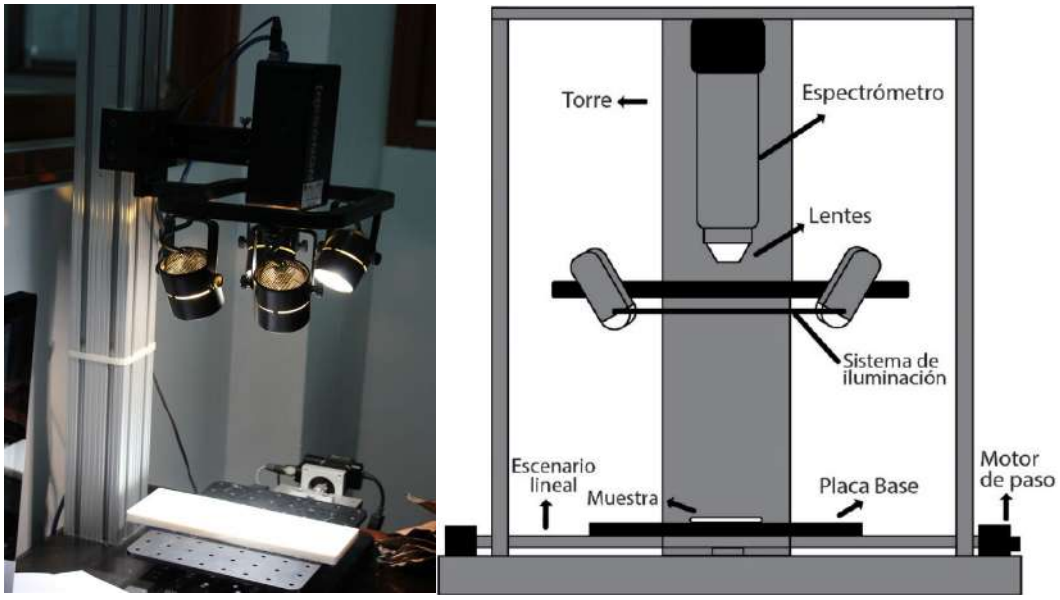


2.2 Adquisición de imágenes hiperespectrales

Para la toma de las imágenes hiperespectrales se utilizó la cámara hiperespectral Resonon Pika II G, ubicada en el Laboratorio Sistemas Automáticos de Control de la Facultad de Ingeniería de la Universidad de Piura. Tiene un rango de trabajo de los 400 a los 900 nm abarcando el espectro visible y parte del infrarrojo cercano (VIS-NIR) y posee una resolución espectral de 2.1 nm. Es uno de los componentes de todo el sistema de mesa RESONON que está compuesto de un lente objetivo de longitud focal de 16 mm, 30.8 grados de FOC y 1.47 mmrad de IFOV, 4 luminarias halógenas y etapa de exploración lineal los cuales unidos componen el sistema de captura de imágenes hiperespectrales de escaneo lineal (Resonon, 2017).

La distancia entre el lente de la cámara y la muestra fue de 30 cm, el tiempo de exposición fue de 0.02 segundos y la velocidad del motor tornillo fue de 1.6 cm/s, controlado por un sistema informático encargado de procesar las imágenes. Las muestras se colocaron en una placa sobre el motor de tornillo para escanearlas mediante el método de escaneo de barrido espacial línea por línea. En la Figura 29. se muestra un gráfico del sistema de la cámara hiperespectral con sus componentes.

Figura 29. Cámara Pika II G RESONON

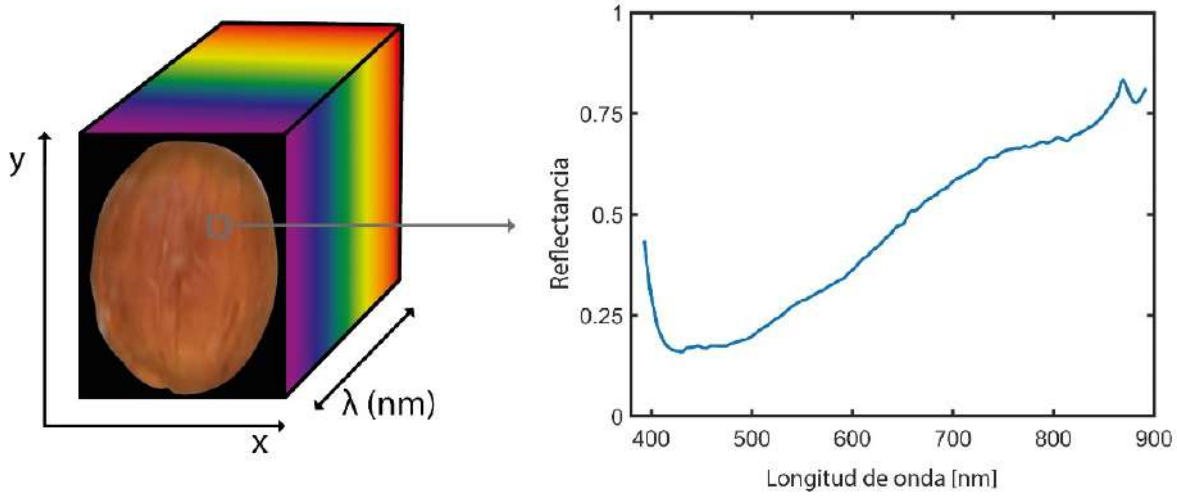


2.3 Procesamiento de imágenes

A través de la captura de las imágenes hiperespectrales se obtiene la firma espectral de cada uno de los pixeles de la imagen final (Figura 30). La firma espectral es un identificador único de la muestra ya que contiene información acerca de su composición físico-química (Pandey et al., 2020).

Para el procesamiento de las HSI tomadas se definió una región de interés (ROI) mediante la cual se excluyen los valores de reflectancia de las zonas irrelevantes del fondo de la imagen tomada y se obtiene la información relevante para el análisis que se concentra en el área que abarca desde el centro hacia el contorno del grano de cacao para cada una de las muestras individuales (Feng et al., 2019). Por lo cual se emplearon programas en MATLAB R2020a para el preprocesamiento de las firmas espectrales de las muestras.

Figura 30. Firma espectral de la muestra de grano de cacao



La firma espectral promedio de la ROI de la muestra se asigna como la firma espectral de la muestra del grano de cacao y esta utiliza como información de entrada para la determinación del contenido de cadmio a través del algoritmo de Deep Learning entrenado.

2.4 Análisis espectral

Se tienen 300 muestras en total de granos de cacao seco, constituidas por 257 muestras de Huánuco y 43 de Piura. Se procedió a hacer un análisis de las firmas espectrales y los rangos de trabajo que abarcan, tanto de su contenido de cadmio en µg/g como de reflectancia.

En la Figura 31, se observa un gráfico 3D, en donde en el eje x, se tiene el intervalo de longitud de onda en el espectro electromagnético de 400 a 900 nm correspondiente al rango de trabajo de la cámara hiperespectral. En el eje y, se tiene el porcentaje de reflectancia de la muestra y en eje z, se tiene el contenido de cadmio de las muestras pertenecientes a Huánuco. Como se puede observar, se tiene la mayor cantidad de los datos en el rango de 0 a 2.5 µg/g aproximadamente. Además, se tienen datos con niveles de cadmio mayores de 3 µg/g, llegando a tener como máximo de 7.4 µg/g.

Por otro lado, en la Figura 32 se tiene análogamente el gráfico 3D de los datos de Piura. Se puede observar que, a diferencia de los datos de Huánuco, todos los datos se concentran en un rango de 0 a 3.7 µg/g, siendo el máximo nivel de cadmio alcanzado de 3.66 µg/g.

Figura 31. Gráfico 3D: x-Longitud de onda, y-Reflectancia e z-Cadmio [µg/g] de los datos de Huánuco

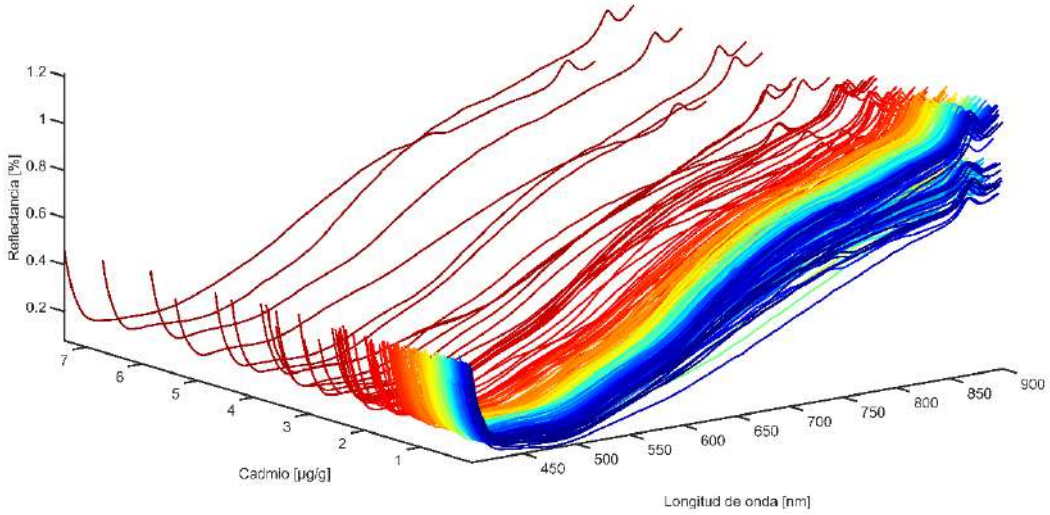
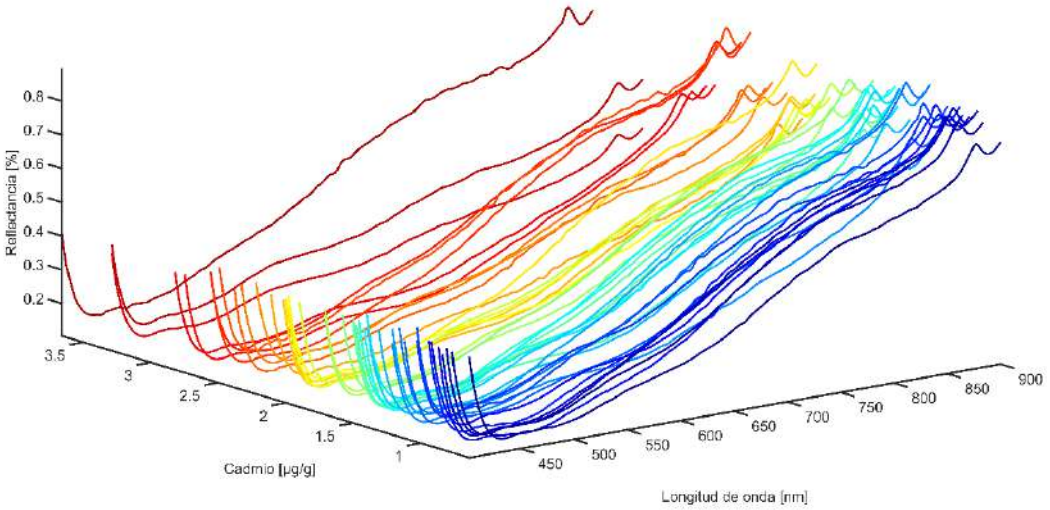


Figura 32. Gráfico 3D: x-Longitud de onda, y-Reflectancia e z-Cadmio [µg/g] de los datos de Piura

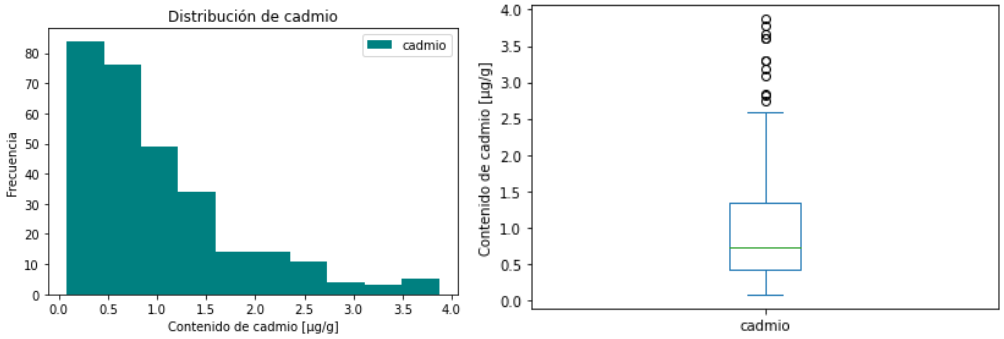


Según lo observado, el conjunto de datos se concentra principalmente entre 0 y 4 µg/g principalmente, así que se hizo un análisis de la distribución de los datos excluyendo los que estaban por encima de 4 µg/g y se obtuvo lo que se muestra en la Figura 33.

Según la gráfica a) de la Figura 33, al haberse hecho una primera acotación de los datos totales de 0 a 4 µg/g, aún se sigue observando a través del histograma de distribución de contenido de cadmio, que los datos se encuentran mayormente concentrados en el rango de 0 a 3 µg/g. Por su parte, para analizar la variabilidad de los datos se realizó el diagrama de caja y bigotes de la variable cadmio. En la gráfica b) de la Figura 33, se observa que el primer cuartil, es decir el 25% de los datos tiene valor en 0.42 µg/g, el 50% o la mediana en 0.73 µg/g y el tercer cuartil o el 75% de los datos en 1.35 µg/g. Por lo que se puede corroborar una mayor dispersión de los datos en la segunda mitad de los datos. Además, se observa que los valores mayores a 2.8 µg/g se representan como datos atípicos debido a que exceden en 1.5 veces el

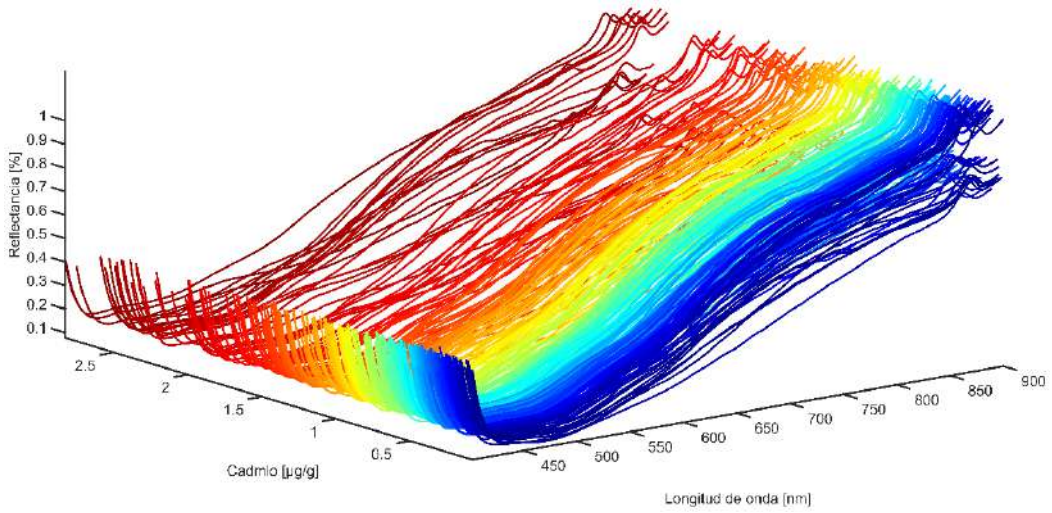
rango intercuartílico es decir superan la varianza esperada, los cuales merecen un análisis más a profundidad debido a que de considerarse en un modelo predictor, podrían tener efecto negativo. Esto debido a que no describen la variabilidad de la mayoría del conjunto de datos.

Figura 33. a) Histograma de distribución de contenido de cadmio y b) Diagrama de caja de contenido de cadmio



a) b)

Figura 34. Gráfico 3D: x-Longitud de onda, y-Reflectancia e z-Cadmio [µg/g] de los datos finales



Por lo que, en base al análisis previo se decide tomar como rango de trabajo en primera instancia todos los datos que tengan niveles de cadmio de 0 a 3 µg/g. Los cuales son 285 muestras que se representan en el gráfico 3D en la Figura 34. Así mismo, esto quiere decir que los modelos generados a partir del conjunto de datos acotados del rango elegido, serán aplicables a muestras de granos de cacao seco con niveles de cadmio del mismo rango, es decir entre 0 a 3 µg/g.

2.5 Análisis químico

Como parte de la metodología y la validación de los modelos de estudio las muestras fueron enviadas a un laboratorio externo que utilizó la metodología NTP N° 208.030: 2015 (INACAL, 2015). En esta se estipula que se debe emplear el método estandarizado de

espectrometría de masas de plasma acoplado inductivamente (ICP-MS) para la determinación de los niveles de cadmio de una muestra de cacao.

2.6 Metodologías de Inteligencia Artificial

Las primeras contribuciones a la Inteligencia Artificial (IA) fueron desarrolladas por Alan Turing en 1950. Mediante el desarrollo de la prueba de Turing propuso una pregunta llena de genialidad en medio de la Segunda Guerra Mundial: "Propongo considerar la pregunta: ¿Pueden las máquinas pensar?". Luego en 1956, John McCarthy propuso la primera definición de la IA: "La rama de la informática que se ocupa de hacer que los ordenadores se comporten como los humanos" (Huawei, 2020a).

De manera general, la Inteligencia Artificial comprende todas las técnicas que permiten a las computadoras imitar el comportamiento humano y reproducir o superar la toma de decisiones humana para resolver tareas de alta complejidad independientemente o con la mínima intervención humana (Janiesch et al., 2021). Dentro de este marco, se pueden distinguir diferentes metodologías, entre las cuales se encuentran las técnicas de Machine Learning y Deep Learning utilizadas en este estudio.

2.6.1 Metodologías de Machine Learning

Machine Learning (ML) es un campo de investigación de la IA que se centra en el estudio de cómo las computadoras pueden obtener nuevos conocimientos mediante la simulación o la realización de comportamientos de aprendizaje de los seres humanos. Además comprende la reorganización de la arquitectura del conocimiento existente para mejorar su rendimiento (Huawei, 2020a). ML se puede clasificar en cuatro subramas: Aprendizaje supervisado, aprendizaje no supervisado, aprendizaje semi supervisado y aprendizaje de refuerzo.

El aprendizaje supervisado se trata de la tarea de aprendizaje automático en donde se desea aprender una función basándose en variables de entrada que se componen de observaciones etiquetadas para inferir una respuesta o salida. En este tipo de aprendizaje se conocen desde el principio los fines que deben cumplirse (Sarker, 2021). Una vez que el modelo ha sido entrenado con éxito, puede utilizarse para predecir la variable objetivo a partir de observaciones nuevos o no visto en las características de entrada (Janiesch et al., 2021).

Las tareas más comunes de aprendizaje supervisado son de clasificación y regresión. Esta última es la tarea desarrollada en profundidad en esta investigación. En la clasificación, se asignan a las muestras una categoría determinada mediante un modelo de clasificación. Se mapea una función f desde variables de entrada "x" y variables de salida "y" como objetivos, las cuales son etiquetas o categorías. En la regresión se buscar predecir una variable de resultado continua "y" a partir del valor de una o más variables predictoras "x". La diferencia más importante entre clasificación y regresión es que la primera predice distintas etiquetas de clase mientras que la segunda busca la predicción de una cantidad continua.

El aprendizaje no supervisado analiza conjuntos de datos u observaciones no etiquetados sin necesidad de la intervención humana o especificaciones preexistentes, y genera un proceso impulsado enteramente por los datos. Entre sus aplicaciones principales se encuentran la extracción características generativas, identificar tendencias y estructuras significativas, agrupaciones de resultados, entre otras (Janiesch et al., 2021; Sarker, 2021). El clustering es una forma habitual de este tipo de aprendizaje, en donde se agrupan muestras similares, se calcula la similitud entre las nuevas muestras y las existentes y se clasifica dicha similitud (Huawei, 2020a).

El aprendizaje semi supervisado puede definirse como una combinación entre el aprendizaje supervisado y el no supervisado, ya que utiliza tanto datos etiquetados como no etiquetados. Esta es una rama muy interesante dentro de ML debido a que simula lo que sucede en un proceso real la mayoría de ocasiones, en donde para determinado contexto se tienen datos etiquetados y muchos otros sin etiquetar. Su objetivo final es proporcionar un mejor resultado de predicción con ambos conjuntos de datos, etiquetados y no etiquetados, que el producido si solo se hubieran usado datos etiquetados (Sarker, 2021).

El aprendizaje de refuerzo permite a los softwares y a las máquinas que evalúen de manera automática el comportamiento óptimo en un contexto o entorno particular para mejorar su eficacia. Es decir, en lugar de proporcionar pares de entrada y salida, se requiere descubrir el estado actual del sistema, por lo que se trata de un enfoque basado en el entorno. Además, se basa en la recompensa o penalidad, y su objetivo final es utilizar la información obtenida de las restricciones del entorno para tomar medidas que aumenten la recompensa y minimicen el riesgo. Es muy utilizada en aplicaciones en donde se necesita aumentar la automatización y la eficacia operativa de sistemas de robótica, tareas de conducción autónoma, fabricación, logística y cadena de suministro, videojuegos, mercados electrónicos, entre otros. No obstante, no es recomendable utilizar este tipo de aprendizaje para problemas sencillos (Janiesch et al., 2021; Sarker, 2021).

Entre los algoritmos regresión de ML se encuentra la regresión linear, las máquinas de vectores de soporte (SVM), árboles de decisiones, bosque aleatorio, la regresión por mínimos cuadrados parciales (PLSR), entre otras. Las metodologías de SVR y PLSR son las que se han usado en este estudio y se verán más en detalle a continuación.

2.6.1.1 Regresión de vectores de soporte (SVR). El algoritmo SV se basa en el algoritmo “Generalized Portrait” creado en Rusia en los años sesenta por (Vapnik & Lerner, 1963) y está basado en la teoría de aprendizaje estadístico VC (Vapnik- Chervonenkis) desarrollada en 1974. Las máquinas de vectores de soporte (SVM) se caracterizan por su capacidad de generalizar sobre datos que aún no ha visto el algoritmo, por lo que se ha generalizado para los problemas de regresión (SVR) y también para clasificación (SVM).

La regresión de vectores de soporte (SVR) es conocida por el uso de kernels, control del margen y la cantidad de vectores soporte. SVR entrena utilizando una función de pérdida

simétrica que penaliza por igual cálculos errados altos y bajos. Además, forma simétricamente un margen de radio mínimo alrededor de la función estimada, por lo que los valores absolutos de los errores inferiores a cierto umbral se excluyen tanto por encima como por debajo de la estimación. Por lo tanto, lo puntos fuera del margen son penalizados y lo que están dentro por encima o por debajo de la función no son penalizados. Ha demostrado un gran rendimiento en la estimación de funciones de valor real, tiene un gran capacidad de generalización y alta precisión (Awad & Khanna, 2015).

A continuación se presenta la generalización de la función de regresión SVR desarrollada por (Rui et al., 2019). El algoritmo de SVR requiere un conjunto de muestras $x_i \in X = R^n$ junto con las siguientes propiedades donde $y_i \in Y = R, i = 1,2, \dots N$. Por lo que se puede establecer la siguiente función linear para resolver el problema de SVR basado en los datos de entrenamiento:

$$f(x) = w.x + b \text{ , donde } w \in R \text{ y } b \in R \tag{1}$$

Donde $w.x$ representa el producto punto de los dos vectores $w\ y\ x$. Si la función de regresión $f(x)$ puede estimar todos los puntos de entrenamiento dentro de la precisión ε, entonces esta simplificación podría transformarse en los siguiente:

$$min\frac{1}{2}w^2 \begin{cases} y - w.x_i - b \leq \varepsilon \\ w.x_i + b - y \leq \varepsilon \end{cases} \text{ donde } i = 1,2, \dots N \tag{2}$$

Debido a que, en algunos casos, la función anterior no consigue procesar todos los datos con precisión, se introducen las variables de holgura ξ y ξ^* para penalizar las funciones de estimación. Por lo que esto puede convertirse en:

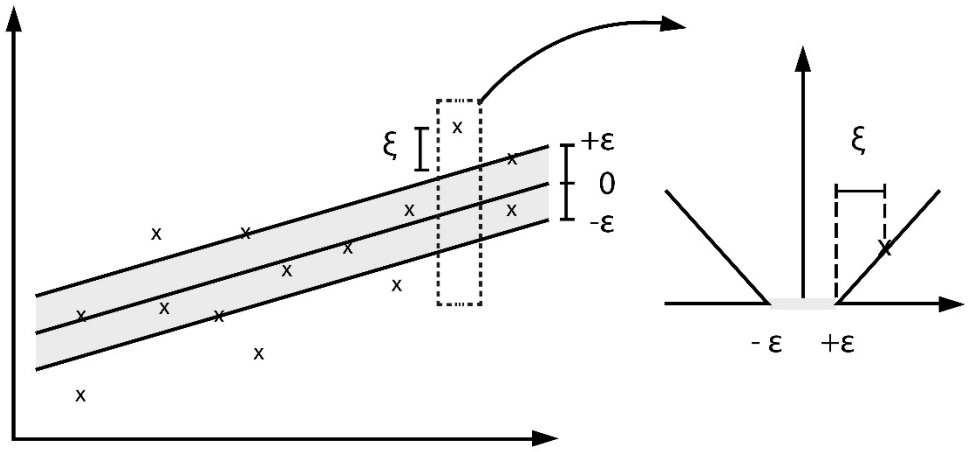
$$min\frac{1}{2}w^2 + C\sum_{i=0}^{N}(\xi_i + \xi^*{}_i) \begin{cases} y - w.x_i - b \leq \varepsilon + \xi_i \\ w.x_i + b - y \leq \varepsilon + \xi^*{}_i \\ \xi_i, \xi^*{}_i \geq 0, \quad i = 1, \dots, N \end{cases} \tag{3}$$

Además, la función de pérdida épsilon puede definirse como:

$$|y - f(x)|_\varepsilon = \begin{cases} 0, & |\xi| \leq \varepsilon \\ |\xi| - \varepsilon & \end{cases} \tag{4}$$

La función $f(x)$ calculada por SVR es la estimación del valor medido y, y los puntos de entrenamiento que caen dentro de la zona positiva y negativa de ε que no se tomarán en cuenta en la función de pérdida. Se puede ver en la Figura 35, que los puntos que se encuentran fuera de la región sombreada son penalizados de forma lineal. Asimismo, la constante C es un coeficiente de penalización cuando los resultados de los puntos de entrenamiento están fuera del canal y esta puede penalizar el error determinando el compromiso (trade-off) entre el error de entrenamiento y la complejidad del modelo.

Figura 35. Definición de la pérdida de margen suave para un SVR lineal



***Nota*. Adaptado de "Total organic carbon content prediction based on support-vector-regression machine with particle swarm optimization" Rui et al. (2019)**

Cuando se tiene un valor de C muy grande, significa que se tiene un gran margen de la función de regresión, mientras que un C menor, representa que la función de estimación permite desviarse de ε con un coste menor.

Dado que la ecuación (3) es un problema de optimización convexo, se introducen los multiplicadores Lagrangianos para obtener la función Lagrangiana:

$$\begin{aligned} L(w, b, \alpha, \alpha^*, \xi, \xi^*, \mu, \mu^*) \\ = \frac{1}{2} w^2 \\ + C \sum_{i=1}^{m} (\xi_i + {\xi^*}_i) - \sum_{i=1}^{m} \mu_i \xi_i - \sum_{i=1}^{m} {\mu_i}^* {\xi_i}^* \\ + \sum_{i}^{m} \alpha_i (f(x_i) - y_i - \varepsilon - \xi_i) \\ + \sum_{i}^{m} {\alpha_i}^* (y_i - f(x_i) - \varepsilon - {\xi_i}^*) \end{aligned} \tag{5}$$

Por lo que problema original puede transformarse en el correspondiente problema dual y se puede obtener la función de regresión:

$$f(x) = \sum_{i=1}^{N} (\alpha_i - {\alpha_i}^*) x_i . x + b \tag{6}$$

2.6.1.2 Regresión por mínimos cuadrados parciales (PLSR). PLS ha demostrado una gran capacidad cuando se usa para conjunto de datos multivariados y con alta colinealidad. Propone una selección de variables para una mejor interpretación y compresión de los

fenómenos que se necesitan modelar y una relación interpretable entre las variables explicativas y la variable respuesta. PLS propone una amplia gama de métodos que abordan la selección de variables que se pueden clasificar según su construcción en tres categorías principales: métodos de filtro, envolventes e integrados (Mehmood et al., 2020). En el caso de PLSR, extrae las componentes latentes, que explican la covarianza entre las variables de entrada y las de respuesta y realiza una estimación de los parámetros de regresión (IBM, 2018).

Introduce el término de componentes principales (CP) las cuales son las direcciones que maximizan la covarianza entre la matriz X e Y, y son combinaciones lineales de las variables originales. PLSR es especialmente útil cuando las variables predictoras son colineales, están altamente correlacionadas entre sí y se tiene mayor número de predictores que observaciones. Además, se puede aplicar para problemas con una variable de respuesta, como también para el cálculo multivariante.

A continuación se presenta la generalización del modelo PSLR lineal desarrollado por (Ramoelo et al., 2013). En este modelo, las matrices de datos centrados X e Y se proyectan en las matrices de menor puntuación y dimensión T y U, respectivamente.

$$X = TP' + E \tag{7}$$

$$Y = UC' + G \tag{8}$$

Donde P y C son los coeficientes o cargas de regresión. Por su parte, T y U en PLS son desarrollados del mismo modo como en PCA, pero la diferencia es que en PLSR se utilizan tanto las variables dependientes como las independientes para descomponer los datos de entrada en variables latentes.

Cuando los pesos no están normalizados, la relación lineal entre las matrices de puntuación T y U puede representarse como:

$$U = T + H \tag{9}$$

Y luego también,

$$Y = TC' + F \tag{10}$$

Donde las matrices E, G, F contienen residuos. Lo que resulta en las matrices:

$$Y = XW * C' + F \tag{11}$$

$$B = W * C' \tag{12}$$

$$Y = XB + F \tag{13}$$

Donde:

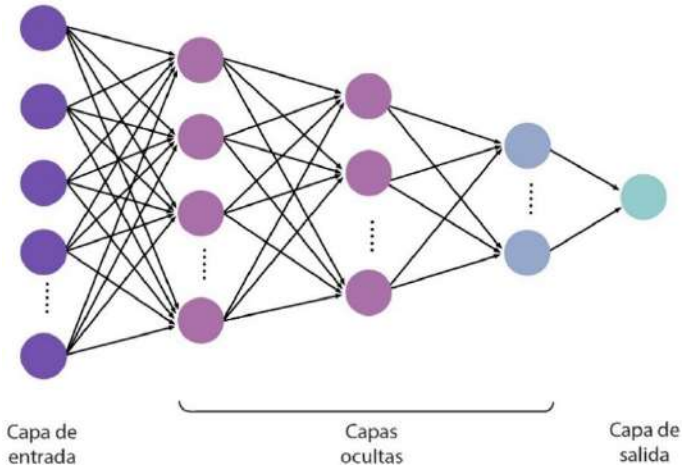
$$T = XW \tag{14}$$

2.6.2 Metodologías de Deep Learning

El Deep Learning (DL) forma parte de una familia más amplia de enfoques de ML basados en redes neuronales artificiales (Sarker, 2021). La arquitectura de DL es una red neuronal, en donde profundo se refiere al número de capas de la red neuronal que combinan las capas de entrada, las capas ocultas y la capa de salida para aprender de los datos. Como se puede observar en la Figura 36, se tiene una arquitectura de red compuesta por una capa de entrada, tres capas ocultas y una capa de salida con una neurona.

Las redes neuronales artificiales (ANN) son un sistema informático inspirado en las redes neuronales biológicas debido a que se componen de neuronas interconectadas por capas que procesan información mediante respuestas dinámicas a entradas externas. Esta concepción fue formulada por primera vez por (Hecht-Nielsen, 1992) en su libro "Theory of the Backpropagation Neural Network". Por lo que puede expresarse como un sistema de información diseñado para imitar la estructura y el funcionamiento del cerebro humano a partir de sus fuentes, características y explicaciones (Huawei, 2020b).

Figura 36. Estructura de una red neuronal



***Nota*. Adaptado de Deep learning classifiers for hyperspectral imaging: A review Paoletti et al. (2019)**

Una red neuronal, según (Burkov, 2020) es una función matemática:

$$y = f_{NN}(x) \tag{15}$$

Donde la función f_{NN} es una función anidada. En donde, para una red neuronal de 3 capas que devuelve un escalar, la función f_{NN} resulta en lo siguiente:

$$y = f_{NN}(x) = f_3(f_2(f_1(x))) \tag{16}$$

Donde f_1 y f_2 son funciones vectoriales de la siguiente forma:

$$f_l(z) \stackrel{\text{def}}{=} g_l(W_l z + b_l) \tag{17}$$

Para esta función, l es el índice de la capa y puede ser desde 1 a cualquier número de capas. La función g_l es la función de activación, la cual es una función fija, normalmente no lineal debido a que su función principal es dotar de no linealidad a la estructura de la red neuronal y es elegida antes de iniciar el aprendizaje dependiendo del problema al resolver. Los parámetros W_l y b_l para cada capa se aprenden mediante el descenso del gradiente optimizado y dependiendo del problema se elige una función de costo particular.

En este caso se usa una matriz W_l y no un vector w_l, debido a que g_l es una función vectorial. Donde cada fila de $w_{l,u}$ (u por unidad) de la matriz W_l es una vector de la misma dimensionalidad que z, sea $a_{l,u} = w_{l,u} z + b_{l,u}$. La salida de $f_l(z)$ es un vector $[g_l(a_{l,1}), g_l(a_{l,2}), \dots, g_l(a_{l,m})]$, donde g_l es una función escalar, m es el número de unidades en la capa l. Para analizarlo de manera concreta, se considera una arquitectura de red neuronal llamada perceptrón multicapa en el siguiente apartado (Burkov, 2020).

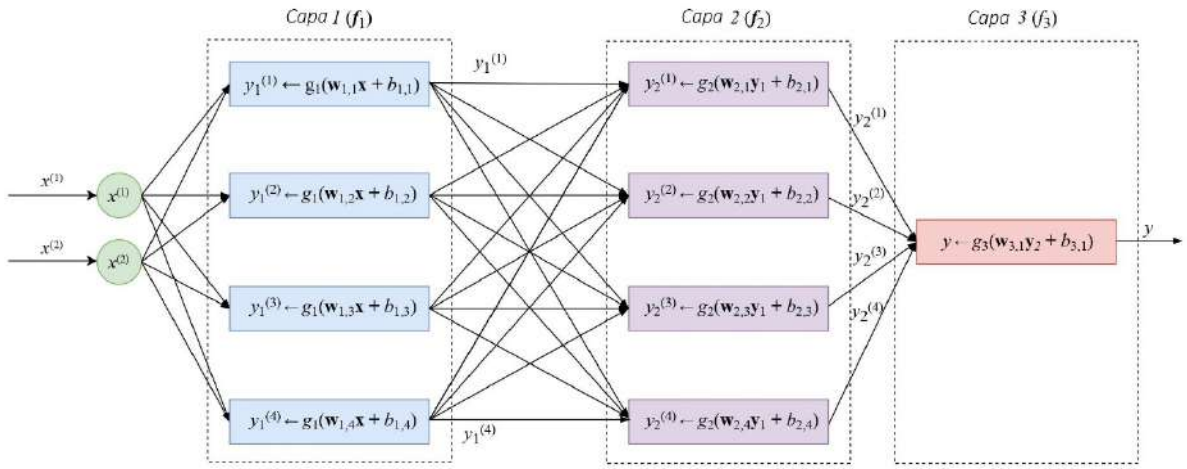
2.6.2.1 Perceptrón Multicapa (MLP). Analizamos una configuración particular de redes neuronales llamada redes neuronales con retroalimentación (FFNN), y más específicamente la arquitectura llamada perceptrón multicapa (MLP). Para reducir el problema, se considera una MLP con 3 capas. La red toma un vector de características bidimensionales como entrada y produce como salida un número. Esta FFNN puede ser un modelo de regresión o clasificación, dependiendo de la función de activación usada en la tercera o última capa de salida (Burkov, 2020).

Se puede observar en la Figura 37, la representación gráfica de la arquitectura de MLP que se va a analizar. Cada unidad se representa por un círculo o un rectángulo, donde las flechas indican la entrada y la salida cada una de las unidades. Además, la salida de cada unidad es el resultado de una transformación matemática que se ha aplicado al vector de entrada. Este resultado a su vez se vuelve una entrada para las unidades de la capa siguiente. Es decir, todas las capas están conectadas a la capa siguiente, siendo una conexión completa, lo que se conoce como capas totalmente conectadas (Burkov, 2020).

La función de activación g_l tiene el índice l que representa el índice de la capa a la que pertenece la unidad. Dependiendo de la aplicación la arquitectura de red puede tener la misma función de activación en todas las unidades o una diferente en cada unidad.

Cada unidad tiene sus parámetros $w_{l,u}$ y $b_{l,u}$, donde u es el índice de la unidad. El vector y_{l-1} es definido en cada unidad como $\left[{y_{l-1}}^{(1)}, {y_{l-1}}^{(2)}, {y_{l-1}}^{(3)}, {y_{l-1}}^{(4)}\right]$. Asimismo, el vector x en la primera capa es definido como $\left[x^{(1)}, \dots, x^{D}\right]$ (Burkov, 2020).

Figura 37. Arquitectura de perceptrón multicapa con entrada bidimensional, dos capas con cuatro neuronas y una capa de salida con una neurona



***Nota*. Adaptado de "The Hundred-Page Machine Learning Book" Burkov (2020)**

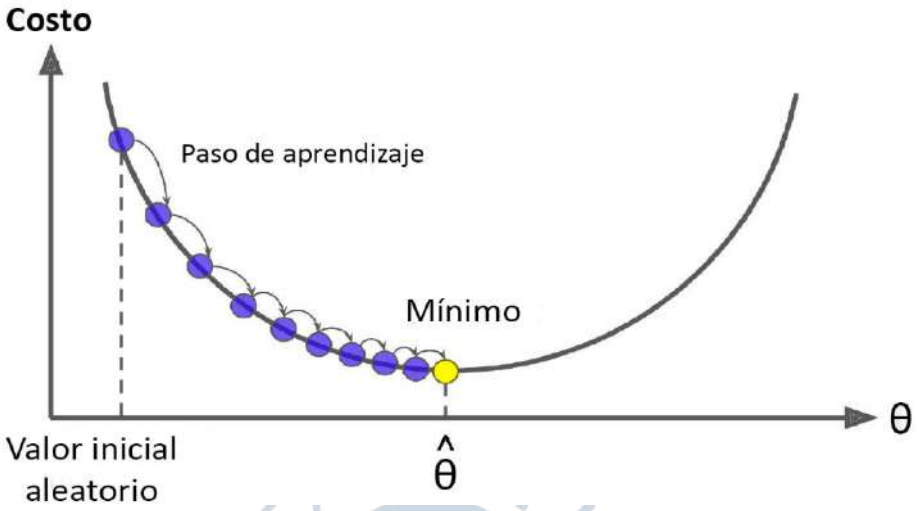
Si la aplicación implica un problema de regresión o de clasificación, la última capa (en rojo) suele contener una sola unidad. Si la función de activación de la última unidad es lineal, entonces se refiere a una arquitectura de red de un modelo de regresión. En cambio, si la función de activación es una función logística, se refiere a un modelo de clasificación binaria.

2.6.2.2 Gradiente de descenso y función de costo. El gradiente de descenso es un algoritmo de optimización que tiene como objetivo ajustar los parámetros de forma iterativa para minimizar la función de costo. Para poder explicar esto, se suele usar el ejemplo una zona de montañas en donde se requiere llegar al fondo del valle lo más rápido posible. Entonces se necesita encontrar una forma de poder llegar en la dirección cuesta abajo usando la pendiente más pronunciada. Por lo que, el gradiente de descenso funciona de esta manera, mide el gradiente local de la función de error con respecto al vector de parámetros θ y va en la dirección del gradiente descendente. Cuando el gradiente es 0, significa que se ha alcanzado un mínimo. Como se muestra en la Figura 38, este proceso se empieza con una inicialización aleatoria de θ y luego se va actualizando este valor, tratando de dar pasos de magnitud adecuada para evitar demorar en el proceso de convergencia al mínimo o alejarse del mínimo debido a que se han dado pasos muy grandes (Géron, 2019).

Para poder encontrar la dimensión de un paso adecuado usando por ejemplo la función MSE como función de costo, se requiere encontrar primero la dirección del vector gradiente, el cual apunta hacia arriba. Entonces se debe seguir la dirección contraria para ir cuesta baja en busca de un mínimo local por lo que se resta $\nabla_\theta MSE(\theta)$ de θ. Además, dependiendo de la tasa de aprendizaje η se determina que tan grande se realiza el paso (Géron, 2019).

$$\theta^{(siguiente\ paso)} = \theta - \eta \nabla_\theta MSE(\theta) \tag{18}$$

Figura 38. Representación del gradiente de descenso y el paso de aprendizaje para encontrar el mínimo local



***Nota*. Adaptado de "Hands-on Machine Learning with Scikit-Learn, Keras, and Tensorflow" Géron (2019)**

Asimismo, con objetivo de poder realizar el entrenamiento del algoritmo y penalizar la clasificación errónea de un ejemplo i, se requiere minimizar la función de costo J encontrando valores de w y b óptimos, donde N es el tamaño de la matriz de observaciones.

A continuación se muestra la función de costo MSE según (Burkov, 2020).

$$J(w,b) = \frac{1}{N}\sum_{i=1,\ldots,N}(f_{w,b}(x_i) - y_i)^2 \tag{19}$$

El algoritmo de retropropagación, consiste en un gradiente de descenso que encuentra, haciendo uso de la regla de la cadena, los gradientes de manera automática y optimizada. En solo dos pasadas por la red, una hacia adelante y otra hacia atrás, el algoritmo de retropropagación es capaz de calcular el gradiente del error de la red para cada uno de los parámetros del modelo (Burkov, 2020). Es decir, va ajustando los valores de los pesos de cada parámetro de la red, como si de una perilla se tratara, para que, al analizar la contribución de cada conexión, se obtenga un menor error al calcular el parámetro de salida. Una vez se tienen estos pesos, se repite el procedimiento varias veces, llamadas épocas, hasta que la red converge en una solución.

En el presente estudio, se utilizó la red neuronal multicapa (MLP) con retropropagación para generar modelos con el objetivo de determinar con precisión el contenido de cadmio a través de la información espectral de las muestras.

2.7 Determinación de las longitudes de onda de mayor importancia

Debido a la alta colinealidad de las variables de entrada a la red neuronal, se aplicaron dos técnicas de selección de longitudes de onda óptima con el objetivo de encontrar las bandas espectrales que describen mejor la variable que se busca predecir a partir de estas, la cual es el contenido de cadmio. Para la aplicación de estos dos algoritmos se desarrollaron programas en MATLAB R2020a.

2.7.1 Algoritmo de proyecciones sucesivas (SPA). Desarrollado por (Araújo et al., 2001), es un método de selección hacia adelante que analiza la información en cada longitud de onda para minimizar la información espectral redundante y minimizar la colinealidad existente. Es un método muy utilizado cuando se trabaja con HSI debido a están conformadas por un gran número de variables con un alto grado de colinealidad. Consta de tres fases principales (Paiva et al., 2012). En la fase 1, los datos respuesta se disponen en una matriz centrada en la media X_{cal} de dimensiones $(N_{cal} x\ K)$ donde la variable x_k se asocia al kth columna vector $x_k \in \Re^{N_{cal}}$. Estos vectores columna se someten a operaciones de proyección para generar K cadenas de M de variables, donde $M = \min(N_{cal} - 1, K)$ es el número máximo de variables que se pueden incluir en un modelo multivariable. Al final de la primera fase, las cadenas se almacenan en una matriz $SEL(MxK)$ tal que $SEL(1,k), SEL(2,k), \dots, SEL(M,k)$ corresponden a los índices de las M variables de la k-ésima cadena, y se construye según los siguientes pasos descritos por (Paiva et al., 2012):

1. Paso 1 (Inicialización):

Se expresa el vector que define las operaciones de proyecciones iniciales $z^1 = x_k$.

$x_j^1 = x_j$, $j = 1, \dots, K$ y $SEL(1xk) = k$, donde i es el contador de iteraciones.

2. Paso 2: Se calcula la matriz P^i de la proyección en un subespacio ortogonal de z^1 como:

$$P^i = I - \frac{z^i (z^i)^T}{(z^i)^T z^i} \tag{20}$$

Donde I es una matriz de identidad $(N_{cal} x N_{cal})$.

3. Paso 3: Se calculan los vectores proyectados x_j^{i+1} como:

$$x_j^{i+1} = P^i x_j^1 \tag{21}$$

Para todo $j = 1, \dots, K$.

4. Paso 4: Se determina el índice j^* del mayor vector proyectado y se almacena este dato en la posición $(i+1, k)$ de la matriz SEL:

$$j^* = arg_{j=1,\dots,K} max \left\| x_j^{i+1} \right\| \tag{22}$$

$$SEL(i+1,k) = j^* \tag{23}$$

5. Paso 5: Sea $z^{i+1} = x_{j^*}^{i+1}$ el vector que define las operaciones de proyección de la siguiente iteración.
6. Paso 6: Sea $i = i + 1$. Si $i < M$, regresar al paso 2.

La segunda fase consiste en evaluar los subconjuntos de datos que conforman las cadenas de variables almacenadas en la matriz SEL. La tercera fase consiste en eliminar las variables que no aportan información relevante para la variable respuesta.

2.7.2 Muestreo Reponderado Adaptativo Competitivo (CARS). Propuesto por (Li et al., 2009), es un método de selección óptima de longitudes de onda clave de datos multiespectrales. Se definen como longitudes de onda clave como aquellas que poseen los mayores coeficientes absolutos obtenidos a través de la aplicación de PLS según las ecuaciones desde la (9) hasta la (14) presentadas anteriormente. CARS consiste en cuatro pasos iterativos, que incluyen la selección de variables con alta adaptabilidad aplicando Monte Carlo, la reducción forzada de longitud de onda mediante la función exponencialmente decreciente (EDF) y el muestreo reponderado (ARS). Finalmente se elige el subconjunto de datos con el valor más bajo de RMSECV como el más óptimo, que contiene los valores de las posiciones de las longitudes de onda más importantes según CARS.

En la reducción forzada de longitud de onda mediante EDF, se supone que el especto completo contiene p longitudes de onda y se realizan N muestreos de ejecución realizados por CARS. El proceso de CARS consiste en dos pasos, en el primer paso EDF es utilizado para eliminar las longitudes de onda que tienen un coeficiente con un valor absoluto relativamente pequeño. En la i-ésimo ejecución de muestreo, las relaciones de longitudes de onda que se deben mantener se calculan usando una EDF definida como:

$$r_i = ae^{-ki} \tag{24}$$

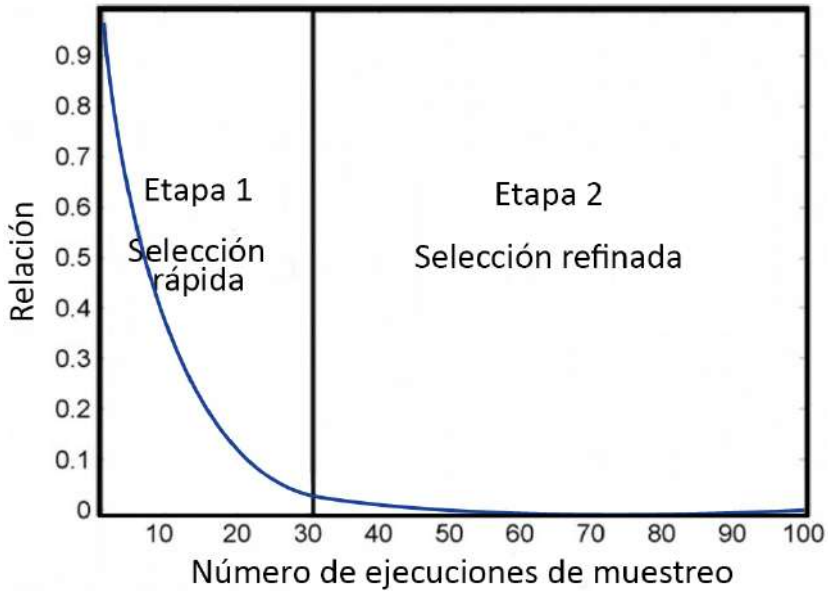
Donde a y k son constantes determinadas siguiendo dos condiciones. La primera condición se considera que, en la primera ejecución de muestreo, todas las longitudes de onda p se toman para el modelado, lo que supone que $r_1 = 1$. La segunda condición, toma en cuenta que en el N-ésimo muestreo, solo dos longitudes de onda se reservan, lo que significa que $r_n = 2/p$. Con estas dos condiciones, a y k pueden ser calculadas como:

$$a = (\frac{p}{2})^{1/(N-1)} \tag{25}$$

$$k = \frac{\ln(p/2)}{N-1} \tag{26}$$

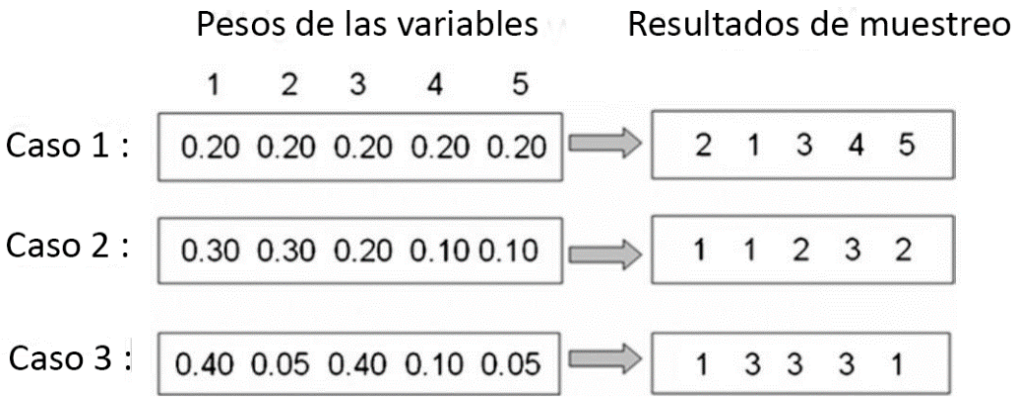
Por lo que en la primera fase EDF elimina rápidamente las longitudes de onda mediante una selección rápida, mientras que, en la segunda fase, las longitudes de onda se reducen de manera suavizada mediante una selección refinada como se muestra en el gráfico de la Figura 39.

Figura 39. Ilustración gráfica de la función exponencialmente decreciente, la primera y la segunda etapa



***Nota*. Adaptado de "Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration" Li et al. (2009)**

Figura 40. Ilustración de la técnica de muestreo reponderado adaptativo con cinco variables



***Nota*. Adaptado de "Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration" Li et al. (2009)**

Luego de realizar los pasos de EDF vistos anteriormente, se aplica el muestreo reponderado (ARS) para eliminar más longitudes de onda de forma competitiva. En este paso se imita el principio de supervivencia del más fuerte de Darwin. En donde, asumiendo que se tienen cinco variables se someten a un muestreo ponderado aleatorio con reemplazo (Figura

40). En el caso 1, cada variable tiene una ponderación igual a 0.20, lo cual indica que pueden ser muestreadas con igual probabilidad. Por lo que el resultado ideal sería que cada variable sea muestreada una vez. En el caso 2, se muestra que las variables 1 y 2 tienen el mayor peso con un valor de 0.3, mientras que las variables 3, 4 y 5 tienen ponderaciones menores. Por lo que, según ARS, las primeras dos variables se muestrean dos veces, mientras que la variable 3 una vez y las dos últimas variables no se eliminan. En el caso 3, demuestra que las variables 1 y 3 debido a sus pesos mayores son más competitivas, mientras que las variables 2, 4 y 5 son mucho menos competitivas debido a sus pesos bastantes bajos. Por lo tanto, ARS elimina estas últimas 3 variables.

UNIVERSITAS STUDIORUM PIURENSIS

Capítulo 3

Análisis de resultados

En este capítulo se hace una revisión de metodología para la construcción de las arquitecturas de redes neuronales y el análisis para la definición de los parámetros e hiperparámetros de cada uno de los modelos propuestos para la predicción del contenido de cadmio.

Además, se exponen los modelos con los mejores desempeños encontrados, se analizan según las métricas consideradas para este estudio y se hace una comparativa entre cada uno de ellos para así encontrar el mejor modelo con la capacidad de determinar el contenido de cadmio en una muestra de cacao a través de su información espectral con mayor precisión.

3.1 Metodología para la construcción de los modelos predictores

Para construir los modelos, se tienen como parámetros de entrada la información espectral de la muestra, a la cual se le asigna como parámetro de respuesta su contenido de cadmio en µg/g. Se procedió a programar usando la versión de Python 3.9.7, haciendo uso de Tensorflow y Keras para el desarrollo de los modelos. Los datos se estandarizaron, con media en cero y con varianza uno. Luego se desarrollaron las metodologías de Machine Learning y Deep Learning detalladas en la Tabla 4.

Tabla 4. Metodologías de Machine Learning y Deep Learning usadas

Machine Learning	Deep Learning
Regresión por vectores de soporte (SVR)	Modelo 1 MLP: 240 variables
Regresión por mínimos cuadrados parciales (PLSR)	Modelo 2: SPA-MLP
	Modelo 3: CARS-MLP

Para los modelos de Machine Learning se usaron dos metodologías anteriormente estudiadas: SVR y PLSR, de las cuales se construyeron 2 modelos. Análogamente, para los modelos de Deep Learning, se construyeron 4 modelos usando como arquitectura base MLP con retropropagación variando el número de variables de entrada. Asimismo, se ajustaron los hiperparámetros de la red neuronal para encontrar el mejor modelo con la capacidad de determinar con mayor precisión el contenido de cadmio a partir de información espectral de una muestra de granos de cacao.

Cabe resaltar que para cada uno de los modelos de Machine Learning y Deep Learning se hizo uso de tres distintos conjuntos de datos, según lo descrito a continuación:

1. Datos acotados de 0 a 3 µg/g

En la Figura 41, se observa el gráfico 3D de los datos acotados de 0 a 3 µg/g, en adelante conjunto de datos 1, según lo descrito anteriormente en el apartado de análisis espectral del anterior capítulo. Este conjunto de datos está conformado por 285 muestras, donde 40 muestras son de Piura y 245 son de Huánuco. Para cada uno de los conjuntos de datos, debido a que se tiene una mayor cantidad de muestras de Huánuco, en cada grupo de la validación cruzada se tiene la misma proporción de datos de Piura y de Huánuco.

Figura 41. Gráfico 3D: x-Longitud de onda, y-Reflectancia e z-Cadmio [µg/g] del conjunto de datos 1

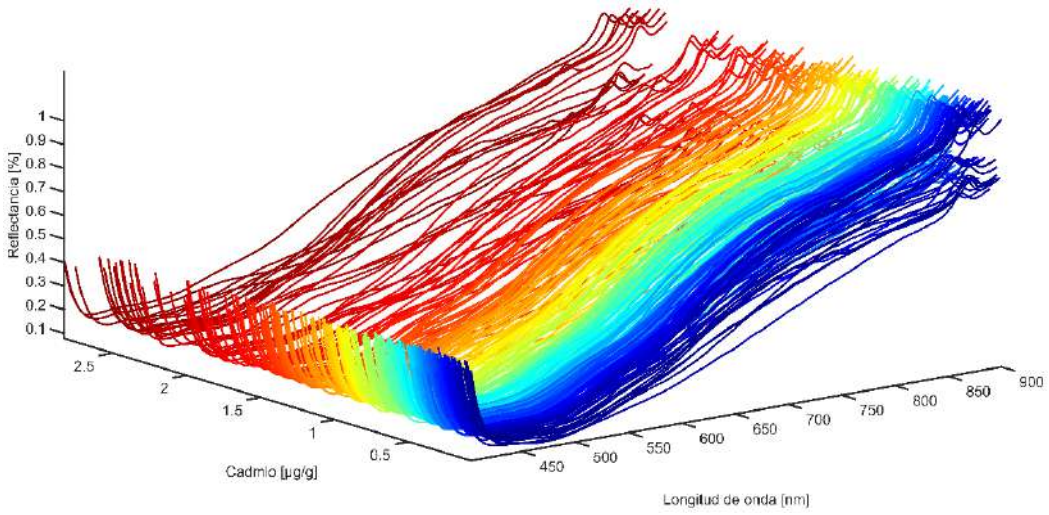
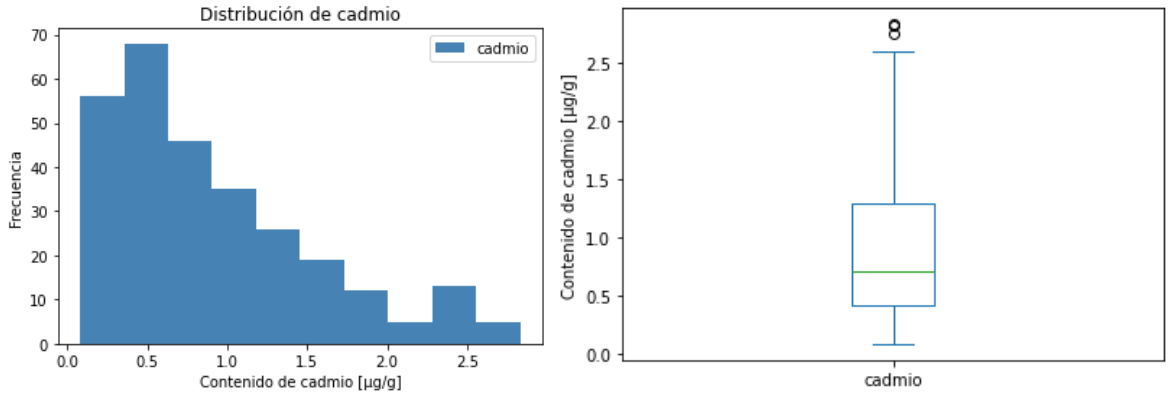


Figura 42. a) Histograma de distribución del conjunto de datos 1 y b) Diagrama de caja del de conjunto de datos 1



a) b)

En la Figura 42, se muestra la distribución del conjunto de datos 1, la cual tiene un valor de 0.41 µg/g en el primer cuartil, 0.71 µg/g en el segundo cuartil, 1.29 µg/g en el tercer cuartil y 2.83 µg/g como valor máximo.

2. Datos acotados de 0 a 2 µg/g

En la Figura 43, se observa el gráfico 3D de los datos acotados de 0 a 2 µg/g, en adelante conjunto de datos 2. Este conjunto de datos está conformado por 262 muestras, donde 29 muestras son de Piura y 233 son de Huánuco.

Figura 43.Gráfico 3D: x-Longitud de onda, y-Reflectancia e z-Cadmio [µg/g] del conjunto de datos 2

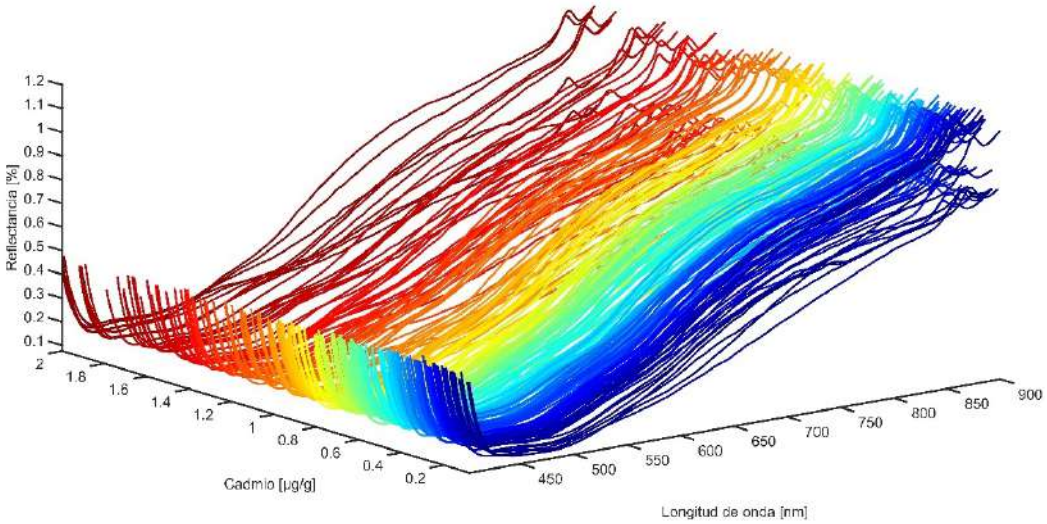
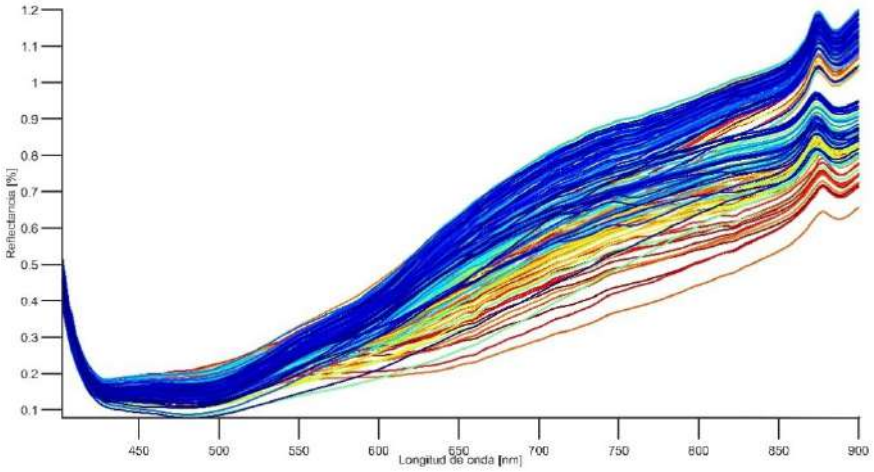
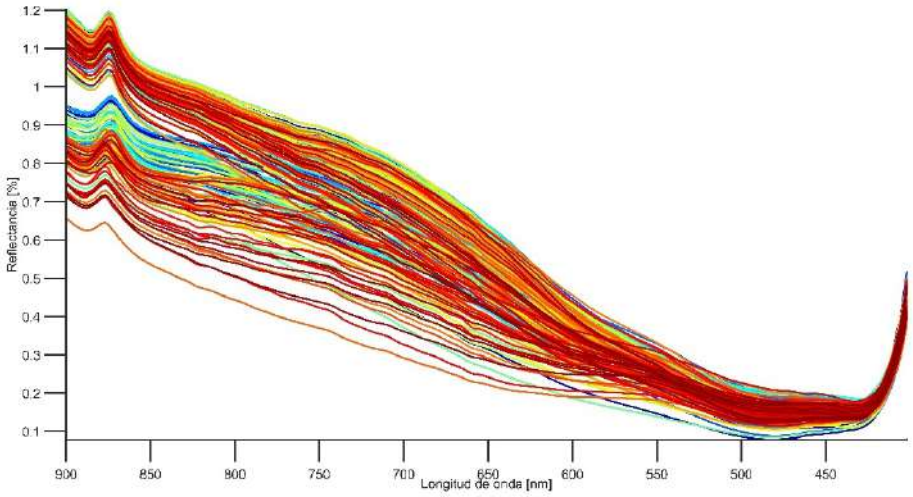


Figura 44. Gráfico 2D x-Longitud de onda, y-Reflectancia [µg/g] del conjunto de datos 2



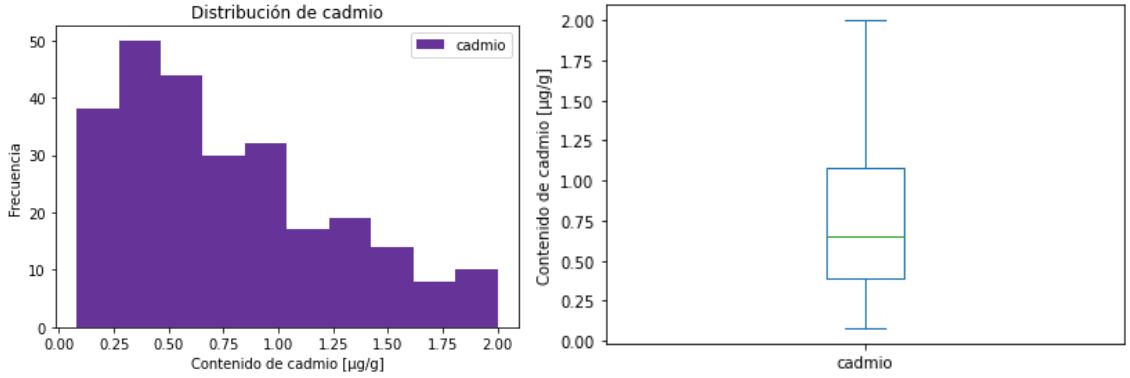
a)



b)

En la Figura 44, se muestran las vistas 2D del conjunto de datos 2. Se puede apreciar que se tienen dos comportamientos diferenciados entre sí. Un conjunto de datos en donde la firma espectral tiene un máximo de 0 a 1 de porcentaje de reflectancia y el otro donde la firma espectral tiene un máximo de 1 a 1.2 de porcentaje de reflectancia. Lo cual no debería ocurrir debido a que el porcentaje de reflectancia solo puede fluctuar entre 0 a 1. Este comportamiento atípico de los datos puede deberse a diversos factores como la una variación en los parámetros y las calibraciones en la cámara hiperespectral cuando se dio la captura de las imágenes de esas muestras en específico.

Figura 45. a) Histograma de distribución del conjunto de datos 2 y b) Diagrama de caja del de conjunto de datos 2



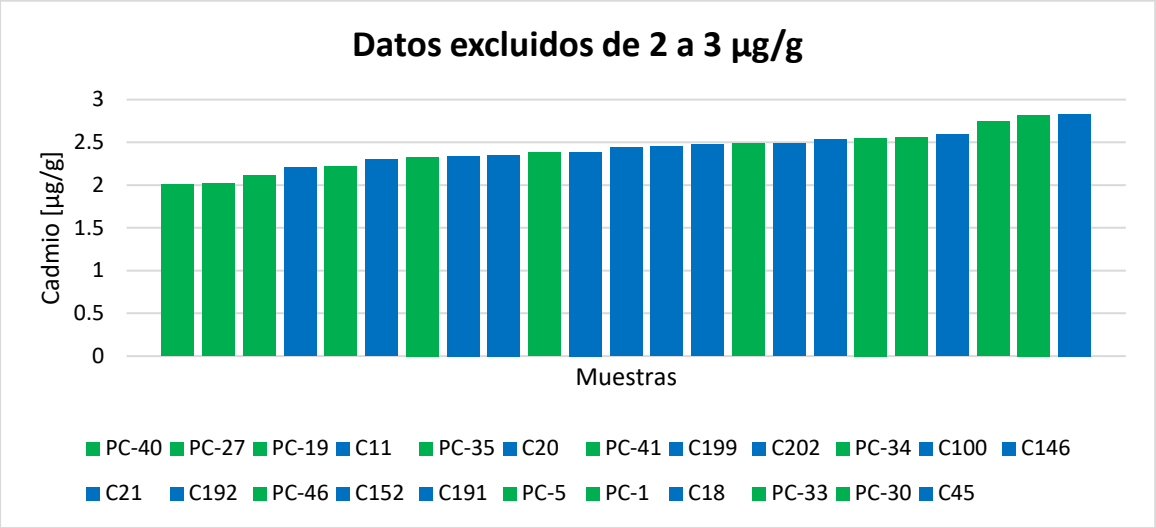
a) b)

En la Figura 45, se muestra la distribución del conjunto de datos 2, la cual tiene un valor de 0.39 µg/g en el primer cuartil, 0.65 µg/g en el segundo cuartil, 1.08 µg/g en el tercer cuartil y 2.0 µg/g como valor máximo.

Debido a que la frecuencia de datos entre 2 a 3 µg/g es de 23 datos, 2 a 1 µg/g es 75 y entre 0 a 1 µg/g es 187. Se decidió excluir los datos de 2 a 3 µg/g debido a que representa la tercera parte aproximadamente de la cantidad de datos de 2 a 1 µg/g y la octava parte de la cantidad de datos de 0 a 1 µg/g. Además, como se observa en la Figura 46, los 23 datos

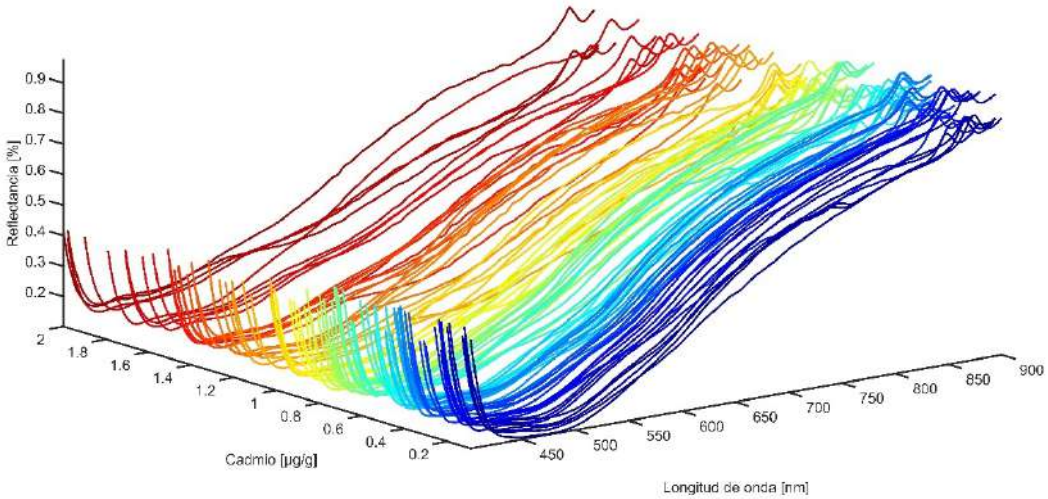
excluidos se conforman de 11 son de Piura y 12 de Huánuco, por lo que se puede decir que se excluyeron similar cantidad de muestras de ambas zonas de análisis.

Figura 46. Datos excluidos de 2 a 3 µg/g del conjunto de datos 2



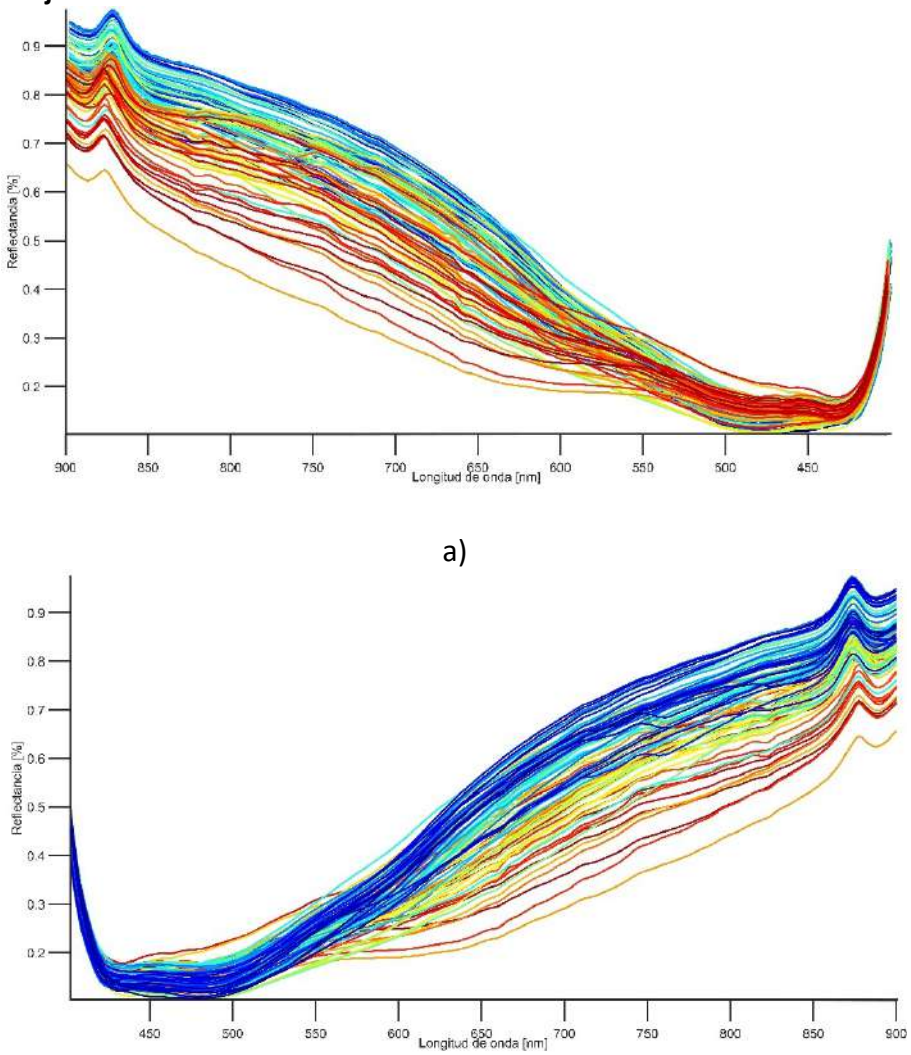
3. Datos acotados de 0 a 2 µg/g excluyendo datos atípicos

Figura 47. Gráfico 3D: x-Longitud de onda, y-Reflectancia e z-Cadmio [µg/g] del conjunto de datos 3



En la Figura 47, se observa el gráfico 3D de los datos acotados de 0 a 2 µg/g, en adelante conjunto de datos 3, excluyendo datos atípicos definidos en la sección anterior, los cuales se caracterizaban por superar el porcentaje de reflectancia de 1. Este conjunto de datos está conformado por 89 muestras, donde 29 muestras son de Piura y 60 son de Huánuco.

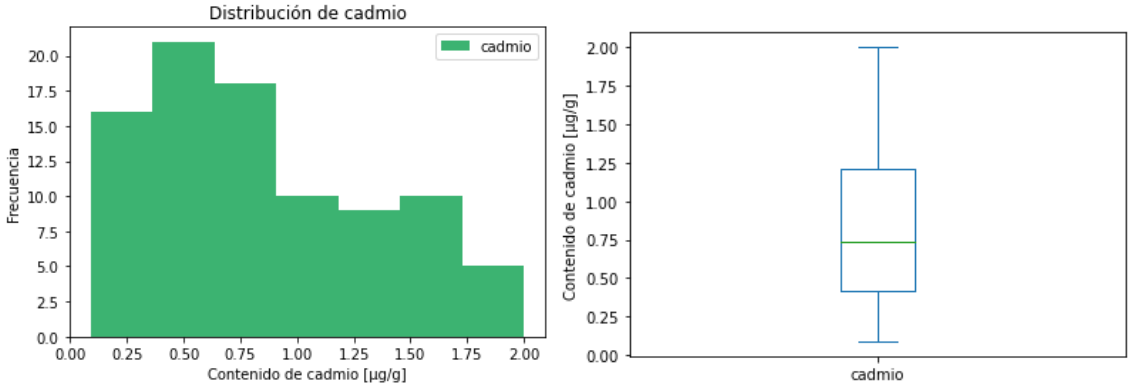
Figura 48. Gráfico 2D: x-Longitud de onda, y-Reflectancia [µg/g] del conjunto de datos 3



a)

b)

Figura 49. a) Histograma de distribución del conjunto de datos 2 y b) Diagrama de caja del de conjunto de datos 3



a) b)

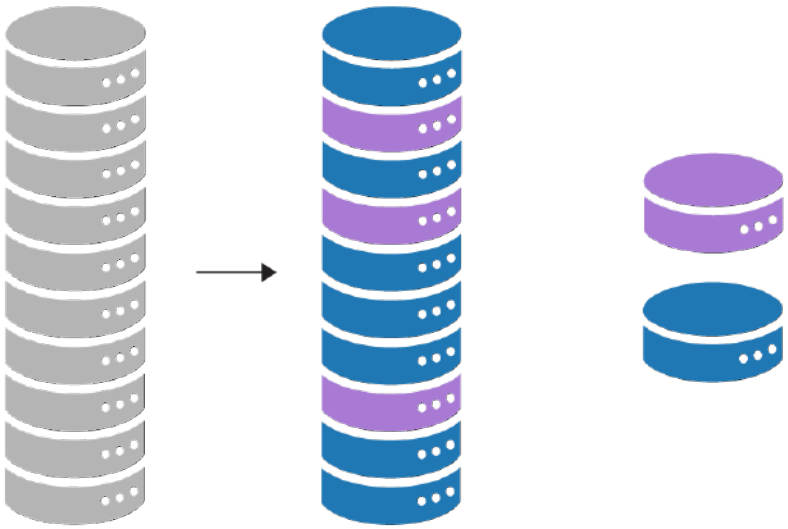
Como se observa en la Figura 48, todas las firmas espectrales de las muestras de este conjunto de datos 3 tienen un comportamiento similar y su valor máximo es 1 de porcentaje de reflectancia. Lo cual es coherente debido a que esta es la medida en la que se escalan todas

las firmas espectrales. Por lo que un valor mayor a 1, como se ha comentado anteriormente se atribuye a una variación en los parámetros de calibración con los que se tomaron las imágenes hiperespectrales de esas muestras.

Cabe resaltar que se realizaron modelos con un cuarto conjunto de datos en donde se tenía un rango de trabajo de 0 a 3 µg/g sin los datos atípicos descritos anteriormente, caracterizados por tener una reflectancia mayor de 1. Sin embargo, estos resultados fueron superados por el conjunto de datos de 0 a 2 µg/g sin los datos atípicos, lo que demuestra según la muestra tomada para esta investigación, que el rango en donde se encuentran los datos más representativos de contenido de cadmio y son capaces de describir mejor la variabilidad de los datos es de 0 a 2 µg/g. Por lo que, para efectos de análisis, los resultados mostrados son los mejores que se obtuvieron con los distintos rangos de trabajo con los que se realizaron experimentos.

En la Figura 49, se muestra la distribución del conjunto de datos 3, la cual tiene un valor de 0.42 µg/g en el primer cuartil, 0.74 µg/g en el segundo cuartil, 1.21 µg/g en el tercer cuartil y 2.0 µg/g como valor máximo.

Figura 50. División de los datos



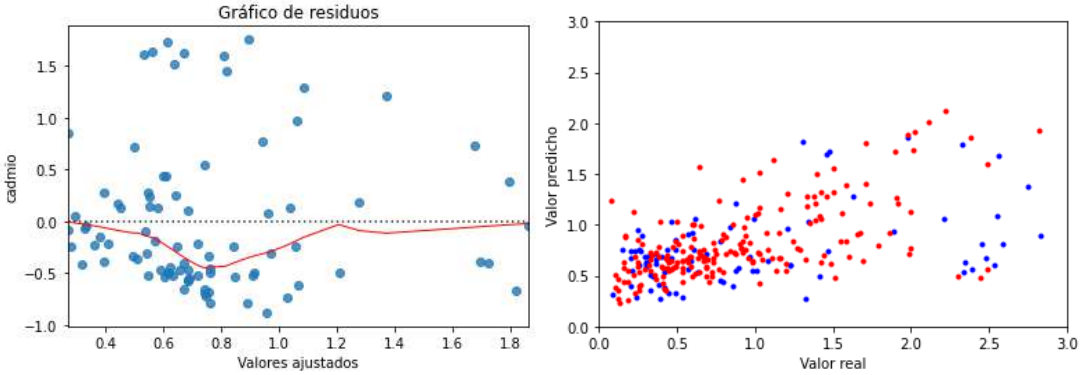
En la Figura 50, se muestra una representación de la división de los datos de cada uno de los conjuntos anteriormente descritos. Se tomó el 70% de los datos para el entrenamiento y 30% de los datos para la validación.

3.2 Modelos de Machine Learning

3.2.1 SVR

Se aplicó este algoritmo a los 3 conjuntos de datos y se obtuvieron los resultados que se muestran en Tabla 5.

Figura 51. Resultados de SVR con conjunto de datos 1

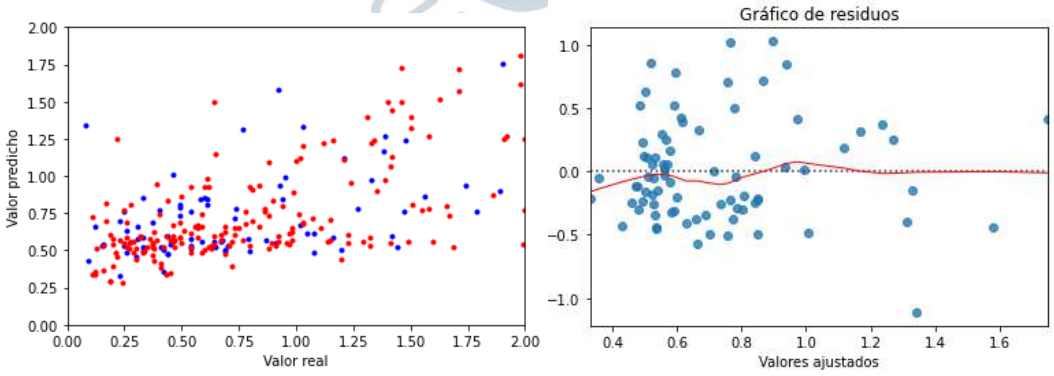


a) b)

En la Figura 51, se muestran los resultados de SVR aplicado al conjunto de datos 1. En la parte a) se grafican los valores reales frente a los valores predichos de contenido de cadmio en µg/g obtenidos aplicando SVR. Se obtuvo un coeficiente de determinación, R^2, con un valor de 0.341. Lo cual refleja que este modelo no puede explicar de manera óptima la variabilidad de la variable respuesta, debido a que el valor de R^2 se encuentra por debajo de 0.50. Mientras mayor sea la variabilidad que puede explicar un modelo de la variable respuesta, el valor de R^2 es cercano a 1.

En la parte b) de la Figura 51 se muestra el gráfico de residuos versus valores ajustados. Esta gráfica se caracteriza por detectar alguna deficiencia en el modelo. Si los residuos están distribuidos entorno a cero y no se da ninguna tendencia o patrón en los datos entonces el modelo es adecuado. Cuando se presenta alguna tendencia, esto quiero decir que el modelo lineal no es capaz de explicar el comportamiento de los datos debido a que existen otras variables explicativas que no han sido consideradas o que las variables predictoras explican la variable de respuesta en un modo más complejo (Aparicio et al., 2004).

Figura 52. Resultados de SVR con conjunto de datos 2



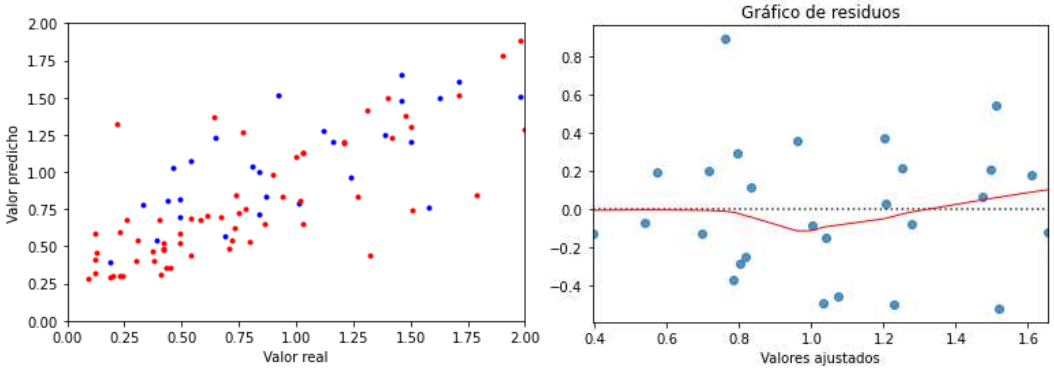
a) b)

En este caso, se aprecia una gran dispersión en la gráfica de los valores ajustados con valores de errores de ± 1.0, es decir este modelo predice la variable respuesta con máximo 1 µg/g. Lo cual es un valor elevado debido a que representa un error de 33% como máximo.

Además, se tiene un MSE-V, es decir un error cuadrático medio del modelo en los datos de validación de 0.516, lo cual es relativamente alto.

En la Figura 52, se muestran los resultados de SVR aplicado al conjunto de datos 2. En la parte a) se grafican los valores reales frente a los valores predichos de contenido de cadmio en µg/g obtenidos aplicando SVR. Se obtuvo un coeficiente de determinación, R^2, con un valor de 0.362. Este valor sigue siendo similar al anterior, lo cual refleja que este modelo no puede explicar de manera óptima la variabilidad de la variable respuesta. En la parte b) de la Figura 52, se muestra el gráfico de residuos versus valores ajustados. Se aprecia una menor dispersión en la gráfica de los valores ajustados que el caso anterior. Además, se tienen valores de errores de ± 1.0 µg/g y un MSE-V de 0.174. Estos resultados han tenido una mejora en comparación con al primer modelo, lo cual se debe principalmente a las variables de entrada.

Figura 53. Resultados de SVR con conjunto de datos 3



a) b)

En la Figura 53, se muestran los resultados de SVR aplicado al conjunto de datos 3. En la parte a) se grafican los valores reales frente a los valores predichos de contenido de cadmio en µg/g obtenidos aplicando SVR. Se obtuvo un coeficiente de determinación, R^2, con un valor de 0.601. En comparación a los resultados anteriores, con este conjunto de datos se obtuvieron mejores resultados. Esto refleja que este modelo puede explicar mejor la variabilidad de la variable respuesta, que los dos anteriores modelos de SVR. En la parte b) de la Figura 53, se muestra el gráfico de residuos versus valores ajustados. Se aprecia una menor dispersión en los valores ajustados, se tienen valores de errores en su gran mayoría de ± 0.4 µg/g y un valor de MSE-V de 0.121.

Estos resultados demuestran que este modelo de SVR es capaz de explicar de manera más óptima el comportamiento de la variable respuesta. Esto se explica debido a que el conjunto de datos 3, es decir las variables de entrada tienen menor variabilidad y no se tienen datos atípicos. Lo que resalta la importancia del conjunto de datos en el desarrollo de modelos predictivos, ya que, al tener una mejor calidad de datos como variables de entrada, mejores serán los resultados del modelo.

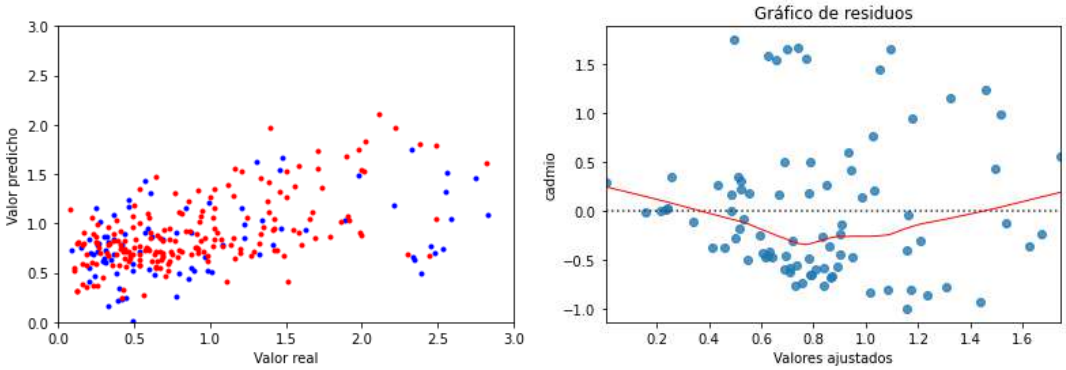
Tabla 5. Resultados de SVR

Modelo	R^2	MSE-V
Conjunto de datos 1	0.341	0.516
Conjunto de datos 2	0.362	0.174
Conjunto de datos 3	0.601	0.121

3.2.2 PLSR

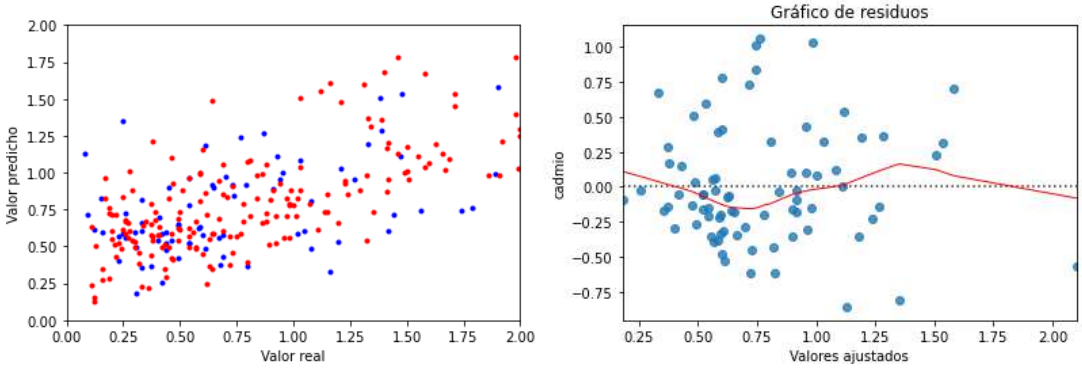
Para la aplicación de este algoritmo a cada conjunto de datos se eligieron diferentes números de componentes: 16, 18 y 16 debido a que se obtuvieron los mejores resultados con esos valores en cada caso (Tabla 6).

Figura 54. Resultados de PLSR con conjunto de datos 1



En la Figura 54, se muestran los resultados de PLSR aplicado al conjunto de datos 1. En la parte a) se grafican los valores reales frente a los valores predichos de contenido de cadmio en µg/g obtenidos aplicando PLSR. Se obtuvo un coeficiente de determinación, R^2, con un valor de 0.442. En comparación con SVR con el mismo conjunto de datos, se obtuvieron mejores resultados aplicando PLSR, lo cual puede deberse a que en este algoritmo selecciona las componentes principales de las variables de entrada lo que disminuye la multicolinealidad y la información redundante del conjunto de datos.

Figura 55. Resultados de PLSR con conjunto de datos 2



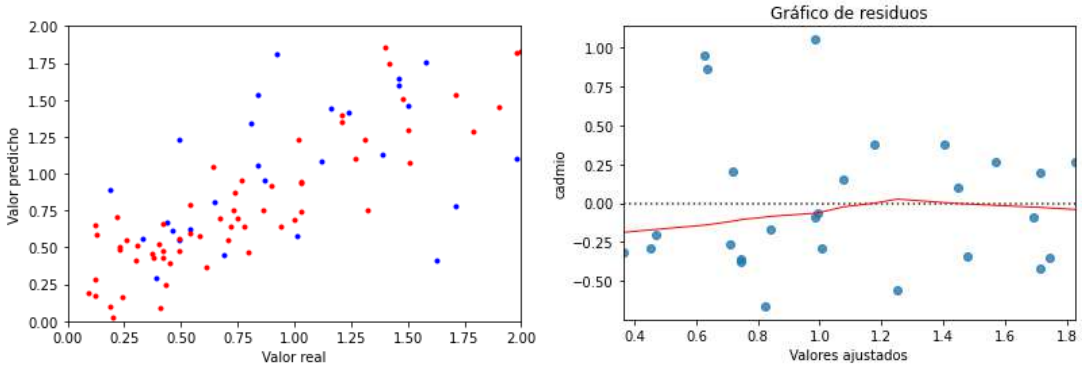
Sin embargo, este modelo no puede explicar de manera óptima la variabilidad de la variable respuesta, debido a que el valor de R^2 se encuentra por debajo de 0.50. En la parte b) de la Figura 54 se muestra el gráfico de residuos versus valores ajustados. Se aprecia una

gran dispersión en la gráfica de los valores ajustados con valores de errores de ± 1.5 y un MSE-V de 0.358. Lo cual reafirma la inefectividad del modelo para predecir de manera óptima la variable de respuesta.

En la Figura 55, se muestran los resultados de PLSR aplicado al conjunto de datos 2. En la parte a) se grafican los valores reales frente a los valores predichos de contenido de cadmio en µg/g obtenidos aplicando PLSR. Se obtuvo un coeficiente de determinación, R^2, con un valor de 0.457 un valor similar en comparación con el modelo anterior. Por lo que, este modelo no puede explicar de manera óptima la variabilidad de la variable respuesta, debido a que el valor de R^2 se encuentra por debajo de 0.50.

En la parte b) de la Figura 55 se muestra el gráfico de residuos versus valores ajustados. Se aprecia una menor dispersión en la gráfica de los valores ajustados con valores de errores de ± 1.0 y un MSE-V de 0.194. Lo cual reafirma la inefectividad del modelo para predecir de manera óptima la variable de respuesta.

Figura 56. Resultados de PLSR con conjunto de datos 3



En la Figura 56, se muestran los resultados de PLSR aplicado al conjunto de datos 3. En la parte a) se grafican los valores reales frente a los valores predichos de contenido de cadmio en µg/g obtenidos aplicando PLSR. Se obtuvo un coeficiente de determinación, R^2, con un valor de 0.777. En comparación a los resultados anteriores, con este conjunto de datos se obtuvieron mejores resultados. Esto refleja que este modelo puede explicar mejor la variabilidad de la variable respuesta, que los dos anteriores modelos de SVR.

En la parte b) de la Figura 56, se muestra el gráfico de residuos versus valores ajustados. Se aprecia una menor dispersión en los valores ajustados que los dos modelos anteriores con valores de errores en su gran mayoría de ± 0.5 µg/g y se obtuvo un MSE-V de 0.139. Estos resultados demuestran que este modelo de PLSR es capaz de explicar de manera más óptima el comportamiento de la variable respuesta. Esto se reafirma la mejor calidad de variables de entrada del conjunto de datos 3 como lo observado con el método de SVR.

Tabla 6. Resultados de PLSR

Modelo	R^2	MSE-V
Conjunto de datos 1	0.442	0.358
Conjunto de datos 2	0.457	0.194
Conjunto de datos 3	0.777	0.139

3.3 Modelos de Deep Learning

3.3.1 Modelo 1

Conformado por una estructura de red de MLP en donde se tiene como variables de entrada las 240 bandas espectrales y como variable de salida el contenido de cadmio en µg/g.

Está conformado por una capa de entrada de a neuronas, 3 capas ocultas que contienen b, c y d de neuronas respectivamente y una capa de salida con una neurona. El detalle del número de neuronas dependió de los conjuntos de datos y se describen en la Tabla 7. En las capas ocultas se usó la función de activación ReLU, la cual elimina el problema de gradiente de desvanecimiento (vanishing gradient). Para la capa de salida se usó una activación lineal debido a que se tiene un parámetro de respuesta.

Tabla 7. Parámetros en arquitectura de red del modelo 1

Modelo	# de neuronas en la capa de entrada "a"	# de neuronas en la capa oculta "b"	# de neuronas en la capa oculta "c"	# de neuronas en la capa oculta "d"
Conjunto de datos 1	55	25	15	5
Conjunto de datos 2	35	25	15	5
Conjunto de datos 3	30	20	10	5

Asimismo, para combatir el sobreajuste proveniente de la alta colinealidad de las variables de entrada se usaron técnicas de regularización como la regularización L1, dropout en la capa de entrada al 60%, un tamaño de lote de 32 y la técnica de early stopping, la cual detiene el entrenamiento si el resultado de la función de pérdida en los datos de entrenamiento no sigue disminuyendo.

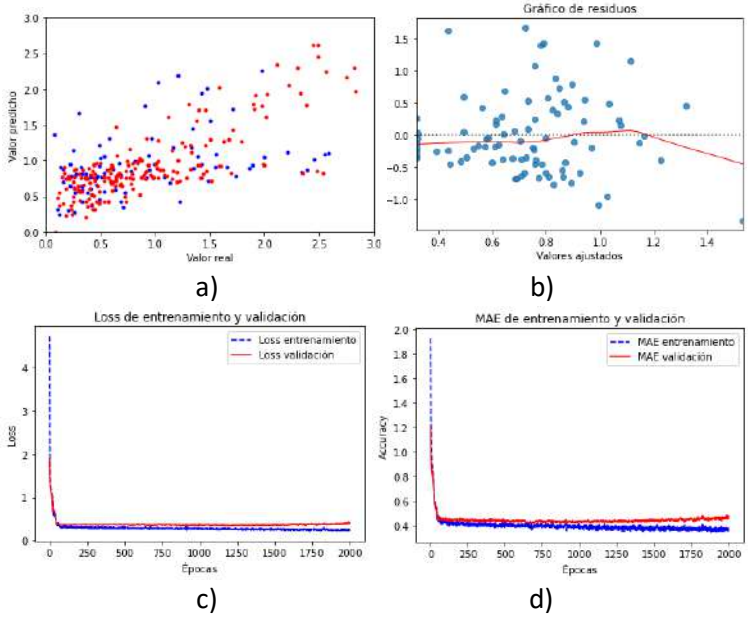
Tabla 8. Resultados del modelo 1

Modelo	R^2	MSE-T	MSE-V	MAE-T	MAE-V
Conjunto de datos 1	0.611	0.197	0.385	0.337	0.485
Conjunto de datos 2	0.616	0.117	0.169	0.266	0.327
Conjunto de datos 3	0.817	0.056	0.152	0.180	0.301

Además, se utilizó el optimizador Adam para optimizar la función de pérdida, la cual fue MSE, buscando acelerar la convergencia del algoritmo, evitar caer en los valores extremos locales y simplificar la configuración manual de la tasa de aprendizaje. Finalmente, se entrenó el modelo y luego de 1500 épocas se obtuvieron los resultados para cada uno de 3 conjuntos de datos (Tabla 8).

En la Figura 57, se muestran los gráficos de los resultados del modelo 1 aplicado al conjunto de datos 1. En la parte a) se grafican los valores reales frente a los valores predichos de contenido de cadmio en µg/g obtenidos aplicando MLP, del cual se obtuvo un coeficiente de determinación, R^2, con un valor de 0.611. Se puede observar cómo desde el primer modelo de MLP se obtienen mejores resultados que los encontrados con los métodos de Machine Learning. En la parte b) de la Figura 57, se muestra el gráfico de residuos versus valores ajustados, en donde se aprecia una gran dispersión en los valores ajustados, se tienen valores de errores en su gran mayoría de ± 1.0 µg/g. Estos resultados son mejores a los obtenidos con los algoritmos de Machine Learning sin embargo pueden mejorar con los otros conjuntos de datos.

Figura 57. Resultados de MLP con conjunto de datos 1



En cuanto a la parte c) y d) se tienen los gráficos de la función de pérdida MSE y precisión con la métrica MAE, error medio absoluto, respectivamente. Se obtuvieron valores de MSE-V, es decir un error cuadrático medio en los datos de validación de 0.385 y un MAE-V, es decir un error medio absoluto en los datos de validación de 0.485. Lo cual es un error bastante alto, por lo que este modelo no es capaz de describir la variabilidad de los datos de entrada.

En la Figura 58, se muestran los gráficos de los resultados del modelo 1 aplicado al conjunto de datos 2. En la parte a) se grafican los valores reales frente a los valores predichos de contenido de cadmio en µg/g obtenidos aplicando MLP, del cual se obtuvo un coeficiente de determinación, R^2, con un valor de 0.616. En este caso se obtuvo un valor de R^2 similar que el modelo anterior, lo cual indica la similitud de ambos algoritmos para predecir el contenido de cadmio. En la parte b) de la Figura 58, se muestra el gráfico de residuos versus valores ajustados, en donde se aprecia una gran dispersión en los valores ajustados, se tienen valores de errores en su gran mayoría de ± 0.8 µg/g menores que los del modelo anterior.

En cuanto a la parte c) y d) se tienen los gráficos de la función de pérdida MSE y precisión con la métrica MAE, error medio absoluto, respectivamente. Se obtuvieron valores de MSE-V, es decir un error cuadrático medio en los datos de validación de 0.169 y un MAE-V, es decir un error medio absoluto en los datos de validación de 0.327. Lo cual sigue siendo un error bastante alto, por lo que este modelo no es capaz de describir con tanta precisión la variabilidad de los datos de entrada.

Figura 58. Resultados de MLP con conjunto de datos 2

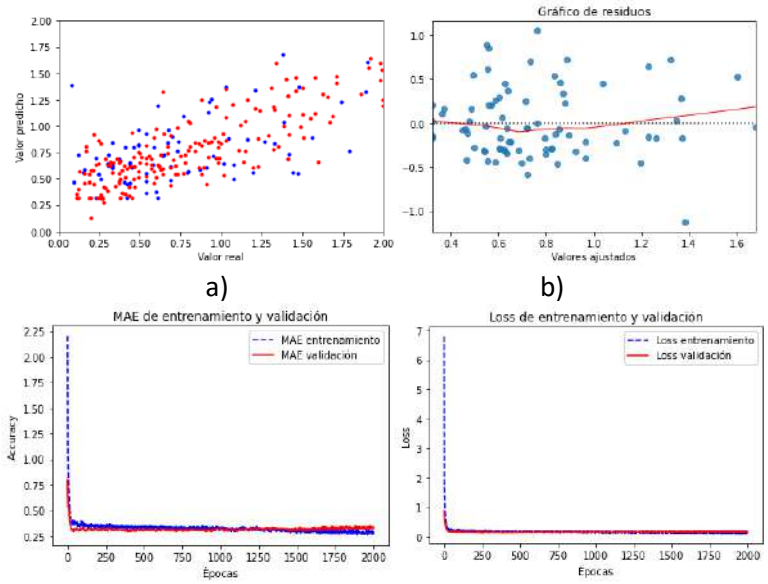
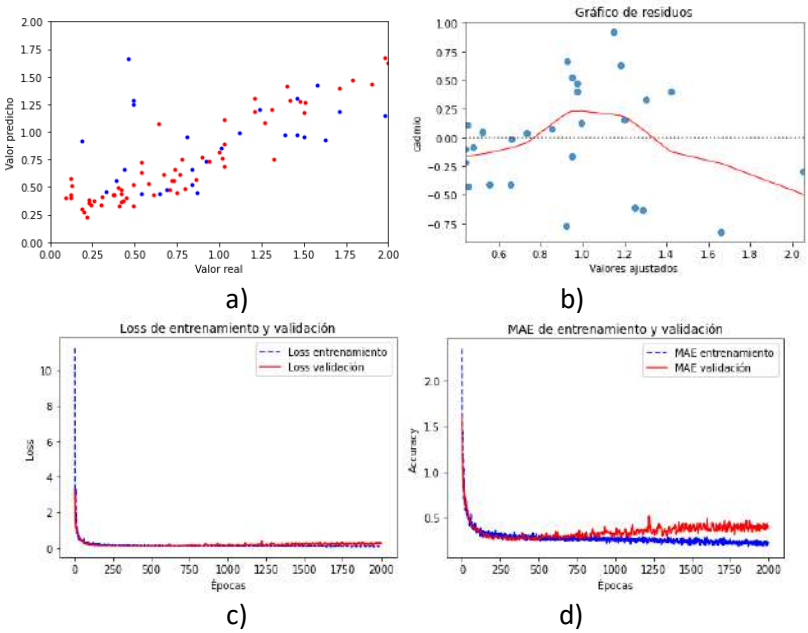


Figura 59. Resultados de MLP con conjunto de datos 3



En la Figura 59, se muestran los gráficos de los resultados del modelo 1 aplicado al conjunto de datos 1. En la parte a) se grafican los valores reales frente a los valores predichos de contenido de cadmio en µg/g obtenidos aplicando MLP, del cual se obtuvo un coeficiente de determinación, R^2, con un valor de 0.817. Desde este resultado, se observa un mejor

rendimiento de este modelo con respecto a los anteriores modelos de Machine Learning y se reafirma la mejor calidad de los datos de entrada.

En la parte b) de la Figura 59, se muestra el gráfico de residuos versus valores ajustados, en donde se aprecia una menor dispersión en los valores ajustados, se tienen valores de errores en su gran mayoría de ± 0.75 µg/g.

En cuanto a la parte c) y d) se tienen los gráficos de la función de pérdida MSE y precisión con la métrica MAE, error medio absoluto, respectivamente. Se obtuvieron valores de MSE-V, es decir un error cuadrático medio en los datos de validación de 0.152 y un MAE-V, es decir un error medio absoluto en los datos de validación de 0.301. Estos errores son menores que los modelos anteriores sin embargo continúan siendo un error bastante alto, por lo que este modelo no es capaz de describir con tanta precisión la variabilidad de los datos de entrada.

3.3.2 Modelo 2

Está conformado por una estructura de red MLP en donde se tiene como variables de entrada las bandas espectrales encontradas por el algoritmo de selección de bandas SPA. Este algoritmo toma un tiempo de aproximadamente 2 min. y 45 seg. para encontrar las bandas óptimas. En la Tabla 9, se muestra el detalle de las bandas seleccionadas por este algoritmo para cada uno de los conjuntos de datos, las cuales se encuentran cercanamente correlacionadas entre los diferentes conjuntos de datos.

Tabla 9. Longitudes de onda óptimas seleccionadas por SPA

Modelo	Variables*	Bandas óptimas
Conjunto de datos 1	19	397.52, 403.79, 407.96, 412.14, 418.4, 424.66, 443.45, 470.58, 487.28, 531.12, 560.34, 579.13, 616.70, 685.59, 754.47, 804.57, 815.00, 869.28, 892.24
Conjunto de datos 2	16	395.44, 405.87, 414.22, 424.66, 428.84, 437.19, 474.76, 516.51, 549.91, 574.95, 587.48, 639.66, 702.29, 762.82, 863.02, 892.24
Conjunto de datos 3	14	399.61, 412.14, 424.66, 435.10, 455.97, 506.07, 587.48, 608.35, 666.80, 750.30, 783.69, 863.02 ,871.36, 892.24

*La columna variables se refiere al número de bandas espectrales que se han considerado como variables de entrada al modelo.

Este modelo está conformado por una capa de entrada de a neuronas, 1 capa oculta que contiene b neuronas y una capa de salida con una neurona detalladas en la Tabla 10. En este caso, se ha disminuido el ancho de la arquitectura de red debido a que se tienen un número menor de variables de entrada con la eliminación de la información espectral redundante y la selección de las variables más relevantes para la predicción del contenido de cadmio según SPA. Por lo que se han reducido el número de neuronas en la capa de entrada y las capas ocultas comparado en el modelo 1. En cuanto a los demás hiperparámetros se utilizaron los mismos que el modelo 1 a excepción del tamaño de lote a 20. Luego se entrenaron 1500 épocas y se obtuvieron los resultados para cada uno de 3 conjuntos de datos Tabla 11.

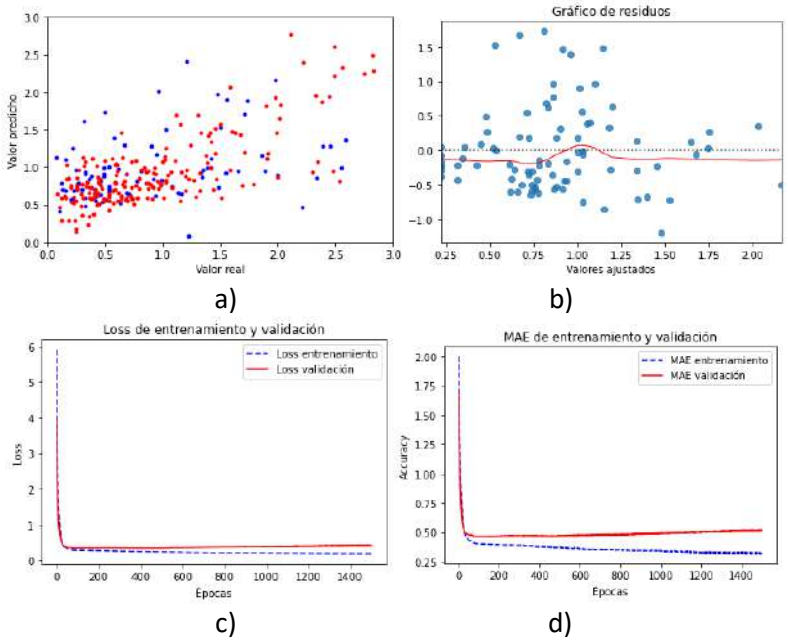
Tabla 10. Parámetros en arquitectura de red del modelo 2

Modelo	# de neuronas en la capa de entrada "a"	# de neuronas en la capa oculta "b"
Conjunto de datos 1	6	4
Conjunto de datos 2	6	4
Conjunto de datos 3	4	2

Tabla 11. Resultados del modelo 2

Modelo	R^2	MSE-T	MSE-V	MAE-T	MAE-V
Conjunto de datos 1	0.619	0.204	0.483	0.341	0.528
Conjunto de datos 2	0.643	0.126	0.224	0.264	0.360
Conjunto de datos 3	0.844	0.036	0.142	0.174	0.285

Figura 60. Resultados de SPA-MLP con conjunto de datos 1

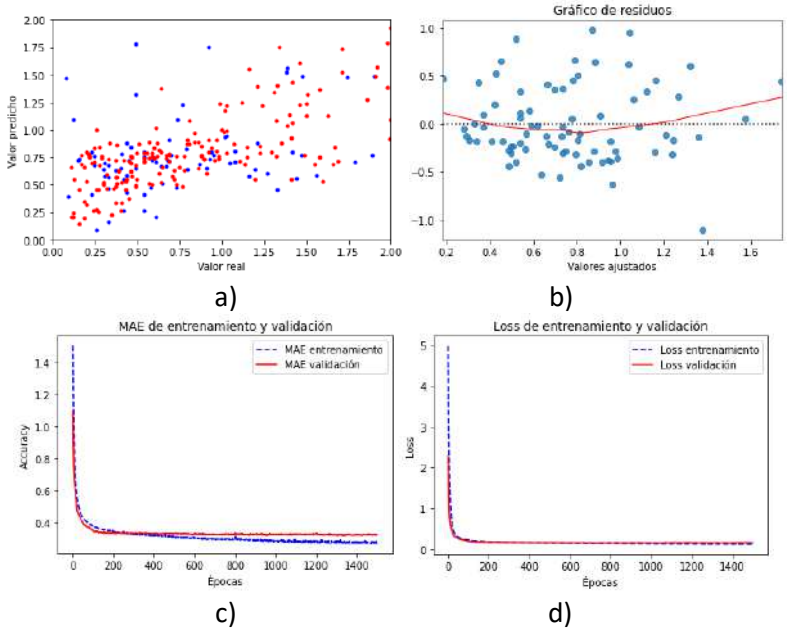


En la Figura 60, se muestran los gráficos de los resultados del modelo 2 aplicado al conjunto de datos 1. En la parte a) se grafican los valores reales frente a los valores predichos de contenido de cadmio en µg/g obtenidos aplicando MLP, del cual se obtuvo un coeficiente de determinación, R^2, con un valor de 0.619. Se puede observar que este resultado es ligeramente mejor que el obtenido por el modelo 1 para el mismo conjunto de datos. Esto se debe a que se han reducido las variables de entrada y por ende se ha disminuido la redundancia de información en las mismas.

En la parte b) de la Figura 60, se muestra el gráfico de residuos versus valores ajustados, en donde se aprecia una gran dispersión en los valores ajustados, se tienen valores de errores en su gran mayoría mayores de ± 1.0 µg/g. En cuanto a la parte c) y d) se tienen los gráficos de la función de pérdida MSE y precisión con la métrica MAE, error medio absoluto, respectivamente. Se obtuvieron valores de MSE-V, es decir un error cuadrático medio en los

datos de validación de 0.483 y un MAE-V, es decir un error medio absoluto en los datos de validación de 0.528. Además, se observa un sobreajuste de los datos debido a que la curva de la pérdida de validación se encuentra por encima de la de entrenamiento. Esto se puede dar, debido a que no usó dropout en este modelo debido a que era una red ya bastante pequeña para usarlo. Por lo que, debido a la magnitud de los errores, este modelo no es capaz de describir la variabilidad de los datos de entrada.

Figura 61. Resultados de SPA-MLP con conjunto de datos 2



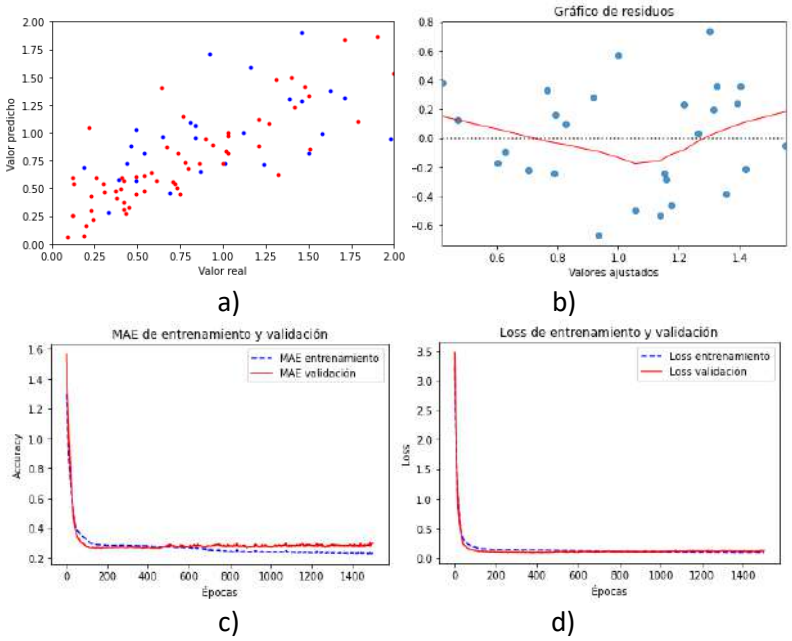
En la Figura 61, se muestran los gráficos de los resultados del modelo 2 aplicado al conjunto de datos 2. En la parte a) se grafican los valores reales frente a los valores predichos de contenido de cadmio en µg/g obtenidos aplicando MLP, del cual se obtuvo un coeficiente de determinación, R^2, con un valor de 0.643. Se puede observar que este resultado es ligeramente mejor que el obtenido por el modelo 1 para el mismo conjunto de datos.

En la parte b) de la Figura 61, se muestra el gráfico de residuos versus valores ajustados, en donde se aprecia una gran dispersión en los valores ajustados, se tienen valores de errores en su gran mayoría de ± 1.0 µg/g. En cuanto a la parte c) y d) se tienen los gráficos de la función de pérdida MSE y precisión con la métrica MAE, error medio absoluto, respectivamente. Se obtuvieron valores de MSE-V, es decir un error cuadrático medio en los datos de validación de 0.224 y un MAE-V, es decir un error medio absoluto en los datos de validación de 0.360. Además, se observa un ligero sobreajuste en comparación con el modelo anterior. Por lo que, debido a la magnitud de los errores, este modelo no es capaz de describir la variabilidad de los datos de entrada.

En la Figura 62, se muestran los gráficos de los resultados del modelo 2 aplicado al conjunto de datos 2. En la parte a) se grafican los valores reales frente a los valores predichos de contenido de cadmio en µg/g obtenidos aplicando MLP, del cual se obtuvo un coeficiente

de determinación, R^2, con un valor de 0.844. Se puede observar que este resultado es mejor que el obtenido por el modelo 1 para el mismo conjunto de datos.

Figura 62. Resultados de SPA-MLP con conjunto de datos 3



En la parte b) de la Figura 62, se muestra el gráfico de residuos versus valores ajustados, en donde se aprecia menor dispersión en los valores ajustados, se tienen valores de errores en su gran mayoría de ± 0.8 µg/g. En cuanto a la parte c) y d) se tienen los gráficos de la función de pérdida MSE y precisión con la métrica MAE, error medio absoluto, respectivamente. Se obtuvieron valores de MSE-V, es decir un error cuadrático medio en los datos de validación de 0.142 y un MAE-V, es decir un error medio absoluto en los datos de validación de 0.285. Además, se observa un ligero sobreajuste en comparación con el modelo anterior. Si bien estos resultados son mejores que el modelo 1 para el mismo conjunto de datos, debido a la magnitud de los errores este modelo no es capaz de describir de manera óptima la variabilidad de los datos de entrada.

3.3.3 Modelo 3

Está conformado por una estructura de red MLP en donde se tiene como variables de entrada las bandas espectrales encontradas por el algoritmo de selección de bandas CARS. Este algoritmo toma un tiempo de aproximadamente 1 min para encontrar las bandas óptimas. En la Tabla 12, se muestra el detalle de las bandas seleccionadas por este algoritmo para cada uno de los conjuntos de datos.

Este modelo está conformado por una capa de entrada de a neuronas, 1 capa oculta que contiene b neuronas y una capa de salida con una neurona detalladas en la Tabla 13. En este caso, se ha disminuido el ancho de la arquitectura de red, similar a los modelos SPA-MLP, debido a que se tienen un número menor de variables de entrada con la eliminación de la información espectral redundante y la selección de las variables más relevantes para la

predicción del contenido de cadmio según CARS. Por lo que se han reducido el número de neuronas en la capa de entrada similar al modelo 2. En cuanto a los demás hiperparámetros se utilizaron los mismos que el modelo 2. Se entrenaron 1500 épocas y se obtuvieron los resultados para cada uno de 3 conjuntos de datos (Tabla 14).

Tabla 12. Longitudes de onda óptimas seleccionadas por CARS

Modelo	Variables*	Bandas óptimas
Conjunto de datos 1	16	395.44, 399.61, 433.01, 443.45, 487.28, 531.12, 551.99, 574.95, 579.13, 604.18, 633.4, 698.11, 710.63, 733.6 812.92, 877.63
Conjunto de datos 2	15	395.44, 412.14, 424.66, 433.01, 474.76, 529.03, 549.91, 579.13, 614.61, 643.84 662.62, 860.93, 873.45, 888.06, 892.24
Conjunto de datos 3	16	403.79, 412.14, 433.01, 472.67, 503.98, 529.03, 543.64, 589.57, 602.09, 681.41, 700.2, 735.68, 823.35, 846.32, 877.63, 881.8

*La columna variables se refiere al número de bandas espectrales que se han considerado como variables de entrada al modelo.

Tabla 13. Parámetros en arquitectura de red del modelo 3

Modelo	# de neuronas en la capa de entrada "a"	# de neuronas en la capa oculta "b"
Conjunto de datos 1	6	3
Conjunto de datos 2	6	3
Conjunto de datos 3	4	2

Tabla 14. Resultados del modelo 3

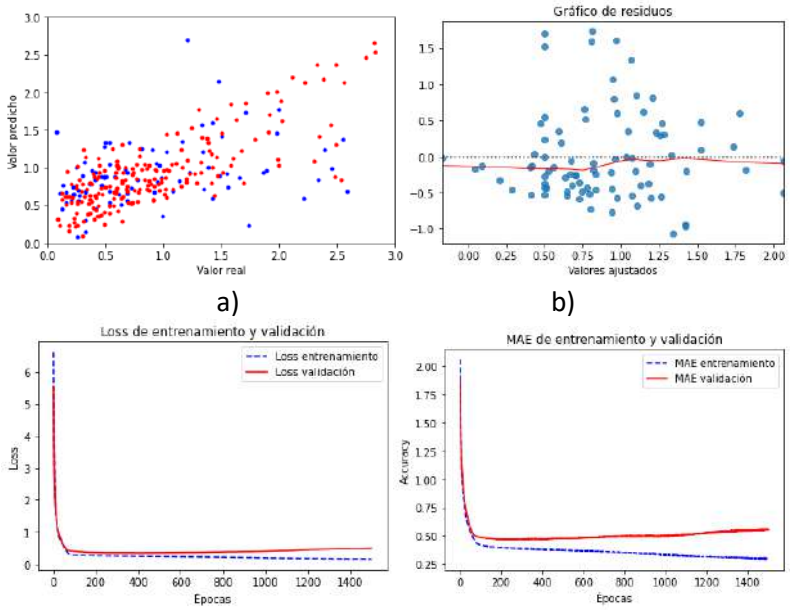
Modelo	R^2	MSE-T	MSE-V	MAE-T	MAE-V
Conjunto de datos 1	0.632	0.158	0.431	0.297	0.495
Conjunto de datos 2	0.641	0.112	0.181	0.261	0.332
Conjunto de datos 3	0.862	0.050	0.128	0.167	0.294

En la Figura 63, se muestran los gráficos de los resultados del modelo 3 aplicado al conjunto de datos 1. En la parte a) se grafican los valores reales frente a los valores predichos de contenido de cadmio en µg/g obtenidos aplicando MLP, del cual se obtuvo un coeficiente de determinación, R^2, con un valor de 0.632. Se puede observar que es el mejor resultado para este conjunto de datos. Por lo que desde el rendimiento del primer conjunto de datos se puede evidenciar la superioridad del modelo CARS-MLP.

En la parte b) de la Figura 63, se muestra el gráfico de residuos versus valores ajustados, en donde se aprecia una gran dispersión en los valores ajustados, se tienen valores de errores en su gran mayoría mayores de ± 1.0 µg/g. En cuanto a la parte c) y d) se tienen los gráficos de la función de pérdida MSE y precisión con la métrica MAE, error medio absoluto, respectivamente. Se obtuvieron valores de MSE-V, es decir un error cuadrático medio en los datos de validación de 0.431 y un MAE-V, es decir un error medio absoluto en los datos de validación de 0.495. Además, se observa un sobreajuste de los datos debido a que la curva de la pérdida de validación se encuentra por encima de la de entrenamiento. Esto se puede dar, debido a que no usó dropout, similar al modelo anterior, debido a que era una red ya bastante

pequeña para usarlo. Por lo que, debido a los resultados obtenidos, este modelo puede mejorar con los otros conjuntos de datos.

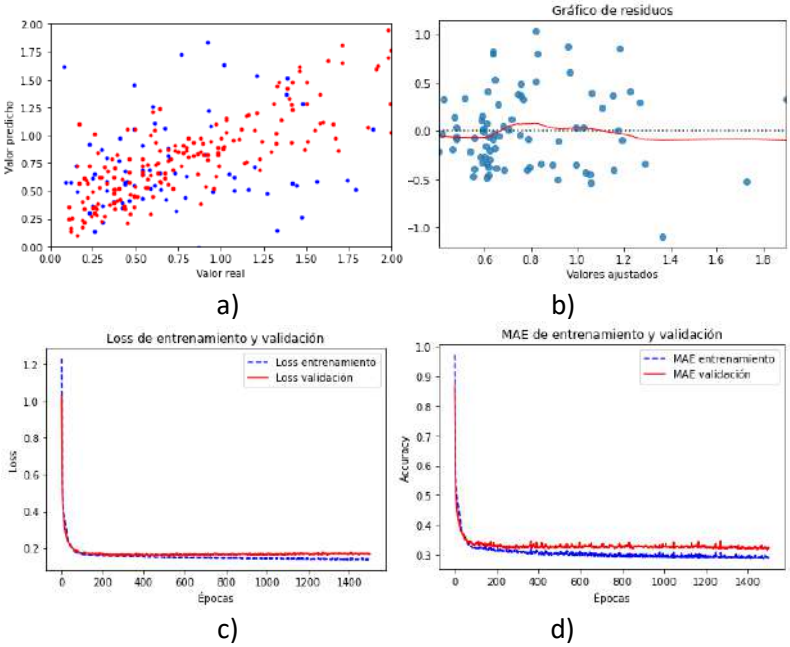
Figura 63. Resultados de CARS-MLP con conjunto de datos 1



a) b) c) d)

En la Figura 64, se muestran los gráficos de los resultados del modelo 3 aplicado al conjunto de datos 2. En la parte a) se grafican los valores reales frente a los valores predichos de contenido de cadmio en µg/g obtenidos aplicando MLP, del cual se obtuvo un coeficiente de determinación, R^2, con un valor de 0.641. Se puede observar que este resultado es ligeramente mejor que el obtenido por el modelo 2 para el mismo conjunto de datos. Por lo que, debido al rendimiento del modelo se reafirma la superioridad del CARS-MLP.

Figura 64. Resultados de CARS-MLP con conjunto de datos 2



a) b) c) d)

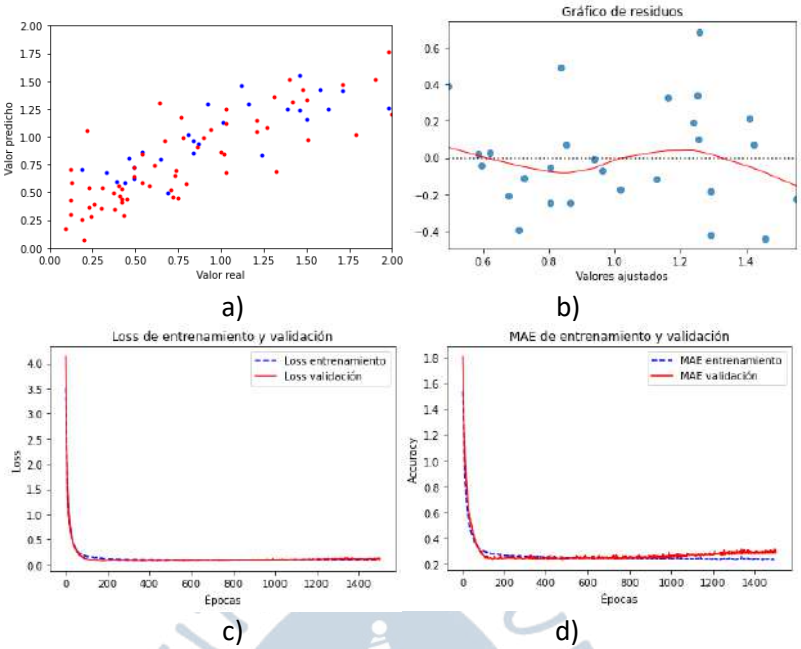
En la parte b) de la Figura 64, se muestra el gráfico de residuos versus valores ajustados, en donde se aprecia una gran dispersión en los valores ajustados, se tienen valores de errores en su gran mayoría menores de ± 1.0 µg/g. En cuanto a la parte c) y d) se tienen los gráficos de la función de pérdida MSE y precisión con la métrica MAE, error medio absoluto, respectivamente. Se obtuvieron valores de MSE-V, es decir un error cuadrático medio en los datos de validación de 0.181 y un MAE-V, es decir un error medio absoluto en los datos de validación de 0.332. Además, se observa un ligero sobreajuste de los datos debido a que la curva de la pérdida de validación se encuentra por encima de la de entrenamiento. Por lo que, debido a los resultados obtenidos, este modelo puede mejorar con el conjunto de datos 3.

En la Figura 65, se muestran los gráficos de los resultados del modelo 3 aplicado al conjunto de datos 3. En la parte a) se grafican los valores reales frente a los valores predichos de contenido de cadmio en µg/g obtenidos aplicando MLP, del cual se obtuvo un coeficiente de determinación, R^2, con un valor de 0.862. Se puede observar que este resultado es el mejor que se ha obtenido en general en comparación con los otros modelos. Por lo que, debido al rendimiento del modelo se destaca la superioridad del CARS-MLP y del conjunto de datos 3.

En la parte b) de la Figura 65, se muestra el gráfico de residuos versus valores ajustados, en donde se aprecia una gran dispersión en los valores ajustados, se tienen valores de errores en su gran mayoría menores de ± 1.0 µg/g. En cuanto a la parte c) y d) se tienen los gráficos de la función de pérdida MSE y precisión con la métrica MAE, error medio absoluto, respectivamente. Se obtuvieron valores de MSE-V, es decir un error cuadrático medio en los datos de validación de 0.128 y un MAE-V, es decir un error medio absoluto en los datos de validación de 0.294. Por lo que, debido a los resultados obtenidos, este modelo es el mejor de todos los modelos lo que se debe tanto a la superioridad de CARS al encontrar las bandas espectrales que mejor describen el comportamiento de la variable respuesta, como también por la calidad del conjunto de datos 3.

Los resultados anteriormente encontrados son compatibles con lo encontrado por (Checa et al., 2019), en donde se utilizaron metodologías de selección de longitudes de banda óptimas para predecir el contenido de cadmio en granos de cacao y se encontraron bandas espectrales similares y en el mismo rango del espectro electromagnético a las encontradas en este estudio. Además, de acuerdo con los hallazgos de (Feng et al., 2019), la mayoría de las longitudes de onda óptimas se encontraron en la región visible (380-780 nm). Esto se debe a que las longitudes de onda a 450 y 680 nm están relacionadas con la luz azul y roja las cuales son causadas por la clorofila. Por lo que, si se encuentran bandas espectrales en este rango se debe a que la bioacumulación de cadmio en la muestra de estudio influye en los parámetros de la clorofila ya que tiene efectos nocivos sobre la fotosíntesis y por ende en el crecimiento de la toda la planta.

Figura 65. Resultados de CARS-MLP con conjunto de datos 3



Conclusiones

El presente estudio evidencia la gran capacidad de las imágenes hiperespectrales para la determinación de contenido de cadmio aplicando modelos de Machine Learning y Deep Learning. Los resultados obtenidos de los modelos propuestos demuestran la aplicabilidad de esta metodología confiable, rápida y basada en métodos no destructivos y supone una oportunidad para utilizarse en la práctica para la determinación del contenido de cadmio en granos de cacao secos.

Los algoritmos de selección de bandas espectrales SPA y CARS lograron destacables desempeños haciendo uso de parte de la información espectral de la muestra, ya que mejoraron significativamente la precisión de los modelos de MLP al combinarse con estos algoritmos. Con su aplicación, se encontraron como máximo 19 bandas óptimas, lo cual representa aproximadamente un 8% de las 240 bandas espectrales iniciales. Esto demuestra la gran capacidad de estos algoritmos para encontrar la información espectral más relevante para la predicción de la variable de respuesta. El algoritmo de CARS-MLP fue el mejor modelo debido a la superioridad de este algoritmo en encontrar las bandas espectrales que contienen la información esencial para la determinación del contenido de cadmio. Se caracterizó principalmente por el uso de 16 bandas óptimas que se concentran en el rango de 400 a 880 nm, lo cual es compatible los resultados obtenidos en anteriores estudios. Por lo que este rango es crucial para la determinación de contenido de cadmio en granos de cacao.

Con los resultados obtenidos con el conjunto de datos 3, los cuales describieron mejor la variabilidad de la variable respuesta, se demostró la importancia del procesamiento de los datos en el desarrollo de modelos predictivos, ya que, al tener una mejor calidad de datos como variables de entrada, mejores serán los resultados del modelo.

En cuanto a trabajos futuros, se podría trabajar con muestras de igual cantidad y de diferentes variedades de cacao en las diversas zonas cacaoteras del Perú. Además, se podrían aplicar imágenes multiespectrales, las cuales se caracterizan por un menor costo de inversión y una información espectral más específica. De esta manera podría ser más factible la aplicación de esta metodología en la práctica y ser una herramienta de gran ayuda para los pequeños y medianos agricultores del Perú.

UNIVERSITAS STUDIORUM PIURENSIS

Referencias bibliográficas

Agile X. (2022). *Scout 2.0*. https://global.agilex.ai/products/scout-2-0

Aparicio, J., Morales, J., & Martínez, A. (2004). *Modelos Lineales Aplicados en R* (Número December 2016) [Universidad Miguel Hernández]. https://www.researchgate.net/profile/Juan-Aparicio-5/publication/311562518_Modelos_Lineales_Aplicados_en_R/links/584cfd7508ae4bc8992c44eb/Modelos-Lineales-Aplicados-en-R.pdf

Araújo, M. C. U., Saldanha, T. C. B., Galvão, R. K. H., Yoneyama, T., Chame, H. C., & Visani, V. (2001). The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. *Chemometrics and Intelligent Laboratory Systems*, *57*(2), 65–73. https://doi.org/10.1016/S0169-7439(01)00119-8

Arévalo-Gardini, E., Arévalo-Hernández, C. O., Baligar, V. C., & He, Z. L. (2017). Heavy metal accumulation in leaves and beans of cacao (Theobroma cacao L.) in major cacao growing regions in Peru. *Science of the Total Environment*, *605–606*(December), 792–800. https://doi.org/10.1016/j.scitotenv.2017.06.122

Augstburger, F., Berger, J., Censkowsky, U., Heid, P., Milz, J., & Streit, C. (2000). *Agricultura Orgánica en el Trópico y Subtrópico* (Asociación Naturland (ed.); 1ª ed.). Asociación Naturland. www.naturland.de

Awad, M., & Khanna, R. (2015). Support Vector Regression. En Springer (Ed.), *Efiicient Learning Machines*. Apress. https://link.springer.com/chapter/10.1007/978-1-4302-5990-9_4

Awais, M., Li, W., Hussain, S., Cheema, M. J. M., Li, W., Song, R., & Liu, C. (2022). Comparative Evaluation of Land Surface Temperature Images from Unmanned Aerial Vehicle and Satellite Observation for Agricultural Areas Using In Situ Data. *Agriculture (Switzerland)*, *12*(2). https://doi.org/10.3390/agriculture12020184

Ayala Martini, D. (2018). *Automatización del análisis de imágenes hiperespectrales para identificación de aptitud de patatas*. Universidad Pública de Navarra.

Banco de Desarrollo de América Latina. (2018). Observatorio del cacao fino y de aroma para América Latina. *Iniciativa Latinoamericana del Cacao*, *3*, 19. http://scioteca.caf.com/bitstream/handle/123456789/1258/OLC_CAF_boletin_3_Español-final.pdf

Bao, Y., Mi, C., Wu, N., & Liu, F. (2019). Rapid Classification of Wheat Grain Varieties Using Hyperspectral Imaging and Chemometrics. *applied sciences*. https://doi.org/10.3390/app9194119

Bawa, U., Ahmad, A., Ahmad, J. N., & Ezra, A. G. (2021). Assessment of health risks from

consumption of food crops fumigated with metal based pesticides in Gwadam, Gombe State, Nigeria. *Bayero Journal of Pure and Applied Sciences*, *14*(1), 100–110. https://doi.org/10.4314/bajopas.v14i1.14

Burkov, A. (2020). The Hundred-Page Machine Learning Book. *Journal of Information Technology Case and Application Research*, *22*(2), 136–138. https://doi.org/10.1080/15228053.2020.1766224

Cardona Velásquez, L. M. (2016). *Influencia del proceso de fermentación sobre las características de calidad del grano de cacao (Theobroma cacao)* [Universidad Nacional de Colombia]. https://doi.org/24 - 28

Castro-Bedriñana, J., Chirinos-Peinado, D., Ríos-Ríos, E., Machuca-Campuzano, M., & Gómez-Ventura, E. (2021). Dietary risk of milk contaminated with lead and cadmium in areas near mining-metallurgical industries in the Central Andes of Peru. *Ecotoxicology and Environmental Safety*, *220*(June). https://doi.org/10.1016/j.ecoenv.2021.112382

Chang, C. I. (2006). Hyperspectral Data Exploitation: Theory and Applications. En *Hyperspectral Data Exploitation: Theory and Applications*. John Wiley & Sons, Inc. https://doi.org/10.1002/9780470124628

Checa, K., Gamarra, M., Soto, J., Ipanaque, W., & Rosa, G. La. (2019, noviembre 1). Preliminary study of the relation between the content of cadmium and the hyperspectral signature of organic cocoa beans. *IEEE CHILEAN Conference on Electrical, Electronics Engineering, Information and Communication Technologies, CHILECON 2019*. https://doi.org/10.1109/CHILECON47746.2019.8987991

Compañía Nacional de Chocolates. (2018). *Protocolo para la caracterización morfológica de árboles élite de cacao (Theobroma cacao L.)*. www.chocolates.com.co/fomento-cacaotero/

Cubero-Castan, M., Schneider-Zapp, K., Bellomo, M., Shi, D., Rehak, M., & Strecha, C. (2018). Assessment of the radiometric accuracy in a targetless workflow using pix4d software. *2018 9th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, 1–4. https://assets.ctfassets.net/go54bjdzbrgi/4QnOSmyXBuoeeqEQuUs2gw/96f2b080855aadec31ab868ed722bf7f/whispers2018_pix4d_targetless.pdf

Reglamento UE N° 488/2014, Diario Oficial de la Unión Europea 1; 4 (2014). https://doi.org/10.2903/j.efsa.2011.1975

Dostert, N., Roque, J., Cano, A., La Torre, M., & Weigend, M. (2012). *Hoja botánica: Cacao - Theobroma cacao L.* (p. 20). Proyecto Perú diverso. https://doi.org/10.13140/RG.2.2.31228.44165

El Rasafi, T., Oukarroum, A., Haddioui, A., Song, H., Kwon, E. E., Bolan, N., Tack, F. M. G., Sebastian, A., Prasad, M. N. V., & Rinklebe, J. (2022). Cadmium stress in plants: A critical review of the effects, mechanisms, and tolerance strategies. *Critical Reviews in Environmental Science and Technology*, *52*(5), 675–726. https://doi.org/10.1080/10643389.2020.1835435

End, M. ., & Dand, R. (2015). *COBISCO/ECA/FCC Cocoa Beans:Chocolate and Cocoa Insdustry Quality Requirements.* http://www.cocoaquality.eu/data/Cacao en Grano Requisitos de Calidad de la Industria Apr 2016_es.pdf

ESA. (2015). *SENTINEL 2*. The European Space Agency. https://www.esa.int/Space_in_Member_States/Spain/SENTINEL_2

Feng, X., Chen, H., Chen, Y., Zhang, C., Liu, X., Weng, H., Xiao, S., Nie, P., & He, Y. (2019). Rapid detection of cadmium and its distribution in Miscanthus sacchariflorus based on visible and near-infrared hyperspectral imaging. *Science of The Total Environment*, *659*, 1021–1031. https://doi.org/10.1016/J.SCITOTENV.2018.12.458

Fundación Charles Darwin. (2022). *Theobroma cacao L.* https://www.darwinfoundation.org/es/datazone/checklist?species=793

García Carrión, L. F. (2010). *Cultivares de cacao en el Perú*. http://agroaldia.minagri.gob.pe/biblioteca/download/pdf/manuales-boletines/cacao/catalogo_cultivares_cacao.pdf

Geotop. (2022). *Sensor Multiespectral Micasense RedEdge MX*. https://www.geotop.la/producto/sensor-termico-micasense-rededge-mx/

Géron, A. (2019). Hands-on Machine Learning with Scikit-Learn, Keras, and Tensorflow. En *Physical and Engineering Sciences in Medicine* (2da edició, Vol. 43, Número 3). O'REILLY. https://doi.org/10.1007/s13246-020-00913-z

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 and Technology*. https://doi.org/10.1016/j.tifs.2007.06.001

Gupta, N., Yadav, K. K., Kumar, V., Kumar, S., Chadd, R. P., & Kumar, A. (2019). Trace elements in soil-vegetables interface: Translocation, bioaccumulation, toxicity and amelioration - A review. *Science of the Total Environment*, *651*, 2927–2942. https://doi.org/10.1016/j.scitotenv.2018.10.047

Haider, F. U., Liqun, C., Coulter, J. A., Cheema, S. A., Wu, J., Zhang, R., Wenjun, M., & Farooq, M. (2021). Cadmium toxicity in plants: Impacts and remediation strategies. *Ecotoxicology and Environmental Safety*, *211*, 111887. https://doi.org/10.1016/j.ecoenv.2020.111887

Huawei. (2020a). *AI Overview*. Huawei Learning. https://talent.huaweiuniversity.com/portal/courses/course-v1:HuaweiX+EBG2020CCHW1100087+Self-paced/about

Huawei. (2020b). *Deep Learning Overview*. Huawei Learning. https://talent.huaweiuniversity.com/portal/courses/course-v1:HuaweiX+EBG2020CCHW1100087+Self-paced/about

INACAL. (2015). *Resolución Directoral Nº 004-2015-INACAL/DN* (Resolución Directoral Nº 004-2015-INACAL/DN). https://cdn.www.gob.pe/uploads/document/file/1675717/004-2015.pdf.pdf

Isla Ramírez, E., & Andrade Adaniya, B. (2009). Propuestas para el manejo de cacao orgánico. En Fundación Conservación Internacional (Ed.), *Proyecto "Paz y Conservación Binacional en la Cordillera del Cóndor, Ecuador-Perú-Fase II (Componente Peruano)"* (1ª ed.). Ministerio del Ambiente. https://web.conservation.org/global/peru/publicaciones/Documents/Propuesta_de_manejo_de_cafe_organico.pdf

Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. *Electronic*

Markets, *31*(3), 685–695. https://doi.org/10.1007/s12525-021-00475-2

Jia, B., Wang, W., Ni, X., Lawrence, K. C., Zhuang, H., Yoon, S. C., & Gao, Z. (2020). Essential processing methods of hyperspectral images of agricultural and food products. *Chemometrics and Intelligent Laboratory Systems*, *198*(17), 103936. https://doi.org/10.1016/j.chemolab.2020.103936

Jun, S., Xin, Z., Xiaohong, W., Bing, L., Chunxia, D., & Jifeng, S. (2019). Research and analysis of cadmium residue in tomato leaves based on WT-LSSVR and Vis-NIR hyperspectral imaging. Spectrochimica Acta - Part A: Molecular and Biomolecular Spectro. *Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy*, *212*, 215–221. https://doi.org/10.1016/j.saa.2018.12.051

Kubier, A., & Pichler, T. (2019). Cadmium in groundwater − A synopsis based on a large hydrogeochemical data set. *Science of the Total Environment*, *689*, 831–842. https://doi.org/10.1016/j.scitotenv.2019.06.499

Lagunes Gálvez, S., Loiseau, G., Paredes, J. L., Barel, M., & Guiraud, J.-P. (2007). Study on the microflora and biochemistry of cocoa fermentation in the Dominican Republic. *International Journal of Food Microbiology*, 124–130. https://doi.org/10.1016/j.ijfoodmicro.2006.10.041

Lewis, C., Lennon, A. M., Eudoxie, G., & Umaharan, P. (2018). Genetic variation in bioaccumulation and partitioning of cadmium in Theobroma cacao L. *Science of The Total Environment*, *640–641*, 696–703. https://doi.org/10.1016/J.SCITOTENV.2018.05.365

Li, H., Liang, Y., Xu, Q., & Cao, D. (2009). Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration. *Analytica Chimica Acta*, *648*(1), 77–84. https://doi.org/10.1016/j.aca.2009.06.046

Liu, Z., Chen, M., Lin, M., Chen, Q., Lu, Q., Yao, J., & He, X. (2022). Cadmium Uptake and Growth Responses of Seven Urban Flowering Plants: Hyperaccumulator or Bioindicator? *Sustainability (Switzerland)*, *14*(2), 1–12. https://doi.org/10.3390/su14020619

Llatance, W. O., Gonza Saavedra, C. J., Guzmán Castillo, W., & Pariente Mondragón, E. (2018). Bioacumulación de cadmio en el cacao (Theobroma cacao) en la Comunidad Nativa de Pakun, Perú. *Revista Forestal del Perú*, *33*(1), 63. https://doi.org/10.21704/rfp.v33i1.1156

Lu, B., Dao, P. D., Liu, J., He, Y., & Shang, J. (2020). Recent advances of hyperspectral imaging technology and applications in agriculture. *Remote Sensing*, *12*(16), 1–44. https://doi.org/10.3390/RS12162659

Lutheran World Relief. (2013). *Caja de herramientas para cacao: Aprendiendo e Innovando sobre el Manejo Sostenible del Cultivo de Cacao en Sistemas Agroforestales. Cosecha, fermentación y secado del cacao*. Aprendiendo e Innovando sobre el Manejo Sostenible del Cultivo de Cacao en Sistemas Agroforestales.

Lv, W., & Wang, X. (2020). Overview of Hyperspectral Image Classification. *Journal of Sensors*, *2020*. https://doi.org/10.1155/2020/4817234

Mahajan, P., & Kaushal, J. (2018). Role of Phytoremediation in Reducing Cadmium Toxicity in Soil and Water. *Journal of Toxicology*, *2018*. https://doi.org/10.1155/2018/4864365

METRICS. (2022). *2022 ACRE 1st Field Campaign*. https://metricsproject.eu/agri-food/2022-

acre-1st-field-campaign/

Ministerio de Agricultura y Riego. (2018). *Resolucion Ministerial N°0451-2018. Lineamientos de muestreo para la determinacion de niveles de cadmio en suelos, hojas, granos y productos derivados de cacao* (pp. 1–22). MINAGRI. https://www.gob.pe/institucion/midagri/normas-legales/221785-451-2018-minagri

Ministerio de Agricultura y Riego del Perú. (2018a). *G/SPS/GEN/1602*. https://docs.wto.org/dol2fe/Pages/FE_Search/FE_S_S009-DP.aspx?language=E&CatalogueIdList=242051&CurrentCatalogueIdIndex=0&FullTextHash=1&HasEnglishRecord=True&HasFrenchRecord=True&HasSpanishRecord=True

Ministerio de Agricultura y Riego del Perú. (2018b). *G/SPS/GEN/1624*. https://docs.wto.org/dol2fe/Pages/FE_Search/FE_S_S009-DP.aspx?language=S&CatalogueIdList=246256,246257,246240,246241,246242,246243,246244,246225,246226,246219&CurrentCatalogueIdIndex=8&FullTextHash=371857150&HasEnglishRecord=True&HasFrenchRecord=True&HasS

Ministerio de Desarrollo Agrario y Riego. (2020). Observatorio de COMMODITIES - Cacao 2020. En *Boletín de publicación trimestral Octubre - Diciembre*. https://cdn.www.gob.pe/uploads/document/file/1782245/Commodities Cacao%3A oct-dic 2021.pdf

Observatorio de Commodities, 1 (2021). https://cdn.www.gob.pe/uploads/document/file/2825303/Commodities Cacao - Abril - setiembre 2021.pdf

Ministerio de Desarrollo Agrario y Riego. (2021a). *Reporte estadístico Junio 2021 - Cacao*. https://cdn.www.gob.pe/uploads/document/file/2189328/REPORTE ESTADÍSTICO 2021 CACAO JUNIO.pdf

Ministerio de Desarrollo Agrario y Riego. (2021b). *Reporte estadístico Enero 2021 - Cacao*. https://cdn.www.gob.pe/uploads/document/file/1862253/REPORTE ESTADÍSTICO CACAO 2021 ENERO.pdf

Morin, M., Lawrence, R., Repasky, K., Sterling, T., McCann, C., & Powell, S. (2017). Agreement analysis and spatial sensitivity of multispectral and hyperspectral sensors in detecting vegetation stress at management scales. *Journal of Applied Remote Sensing*, *11*(04), 1. https://doi.org/10.1117/1.JRS.11.046025

Neo. (2022). *Hyspex*. https://www.hyspex.com/

Nigam, P. S., & Singh, A. (2014). Cocoa and Coffee Fermentations. En *Encyclopedia of Food Microbiology* (pp. 485–492). Elsevier. https://doi.org/10.1016/B978-0-12-384730-0.00074-4

Ozdemir, A., & Polat, K. (2020). Deep Learning Applications for Hyperspectral Imaging: A Systematic Review. *Journal of the Institute of Electronics and Computer*, *2*(1), 39–56. https://doi.org/10.33969/jiec.2020.21004

Paiva, H. M., Soares, S. F. C., Galvão, R. K. H., & Araújo, M. C. U. (2012). A graphical user interface for variable selection employing the Successive Projections Algorithm. *Chemometrics and Intelligent Laboratory Systems*, *118*, 260–266. https://doi.org/10.1016/j.chemolab.2012.05.014

Pandey, P. C., Srivastava, P. K., Balzter, H., Bhattacharya, B., & Petropoulos, G. P. (2020). *Hyperspectral Remote Sensing: Theory and Applications*. Elsevier.

Paoletti, M. E., Haut, J. M., Plaza, J., & Plaza, A. (2019). Deep learning classifiers for hyperspectral imaging: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, *158*(November), 279–317. https://doi.org/10.1016/j.isprsjprs.2019.09.006

Parra Rosero, P. (2017). Modelación de un proceso de secado de cacao utilizando una cámara rotatoria cilíndrica y flujo de aire caliente [Universidad de Piura]. En *Universidad de Piura*. https://pirhua.udep.edu.pe/handle/11042/3488

PIX4D. (2022). *Parrot Sequoia+*. https://www.pix4d.com/es/producto/sequoia

PromPerú. (2018). *Cacao : Propiedades y beneficios del cacao peruano | Perú Info*. https://peru.info/es-pe/superfoods/detalle/super-cacao

PromPerú. (2021). *Chocolate y cacao peruanos distinguidos con categorías "Oro" y "Campeón de campeones" en importante concurso internacional*. https://www.gob.pe/institucion/promperu/noticias/492966-chocolate-y-cacao-peruanos-distinguidos-con-categorias-oro-y-campeon-de-campeones-en-importante-concurso-internacional

Ramoelo, A., Skidmore, A. K., Cho, M. A., Mathieu, R., Heitkönig, I. M. A., Dudeni-Tlhone, N., Schlerf, M., & Prins, H. H. T. (2013). Non-linear partial least square regression increases the estimation accuracy of grass nitrogen and phosphorus using in situ hyperspectral and environmental data. *ISPRS Journal of Photogrammetry and Remote Sensing*, *82*, 27–40. https://doi.org/10.1016/j.isprsjprs.2013.04.012

Resonon. (2017). *SpectrononPro Manual*. 1–135.

Rui, J., Zhang, H., Zhang, D., Han, F., & Guo, Q. (2019). Total organic carbon content prediction based on support-vector-regression machine with particle swarm optimization. *Journal of Petroleum Science and Engineering*, *180*, 699–706. https://doi.org/10.1016/j.petrol.2019.06.014

Ruiz, J. (2016). *Estudio de la visión hiperespectral en el proceso de fermentación del cacao*. Universidad de Piura.

Saddik, A., Latif, R., El Ouardi, A., Elhoseny, M., & Khelifi, A. (2022). Computer development based embedded systems in precision agriculture: tools and application. *Acta Agriculturae Scandinavica Section B: Soil and Plant Science*, *72*(1), 589–611. https://doi.org/10.1080/09064710.2021.2024874

Saeidan, A., Khojastehpour, M., Golzarian, M. R., Mooenfard, M., & Khan, H. A. (2021). Detection of Foreign Materials in Cocoa Beans by Hyperspectral Imaging Technology. *Food Control*, 108242. https://doi.org/10.1016/j.foodcont.2021.108242

Saltini, R., Akkerman, R., & Frosch, S. (2013). Optimizing chocolate production through traceability: A review of the influence of farming practices on cocoa bean quality. *Food Control*, *29*(1), 167–187. https://doi.org/10.1016/j.foodcont.2012.05.054

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, *2*(3), 1–21. https://doi.org/10.1007/s42979-021-00592-x

Satarug, S., Vesey, D. A., & Gobe, G. C. (2017). Health risk assessment of dietary cadmium

intake: Do current guidelines indicate how much is safe? *Environmental Health Perspectives*, *125*(3), 284–288. https://doi.org/10.1289/EHP108

Schwan, R. F., & Wheals, A. E. (2004). The Microbiology of Cocoa Fermentation and its Role in Chocolate Quality. *Critical Reviews in Food Science and Nutrition*, *44*(4), 7–9. https://doi.org/10.1080/10408690490464104

Sistema de gestión de información sanitaria y fitosanitaria. (2019). *Preocupaciones comerciales específicas: sanitarias y fitosanitarias*. http://spsims.wto.org/en/SpecificTradeConcerns/View/430

Specim. (2022). *Spectral imaging made easy*. https://www.specim.fi/

TECHPRESS. (2014). *Visión hiperespectral para inspección de calidad*. https://techpress.es/vision-hiperespectral-para-inspeccion-de-calidad/

USGS, & NASA. (2019). Landsat 9. *Fact Sheet*, 2. http://pubs.er.usgs.gov/publication/fs20193008

Viera, G. (2018). Aplicación de procesamiento de imágenes para clasificación de granos de cacao según su color interno. En *Universidad de Piura*. Universidad de Piura.

Wu, D., & Sun, D.-W. (2013). Advanced applications of hyperspectral imaging technology for food quality and safety analysis and assessment: A review — Part I: Fundamentals. *Innovative Food Science & Emerging Technologies*, *19*, 1–14. https://doi.org/10.1016/j.ifset.2013.04.014

Xin, Z., Jun, S., Yan, T., Quansheng, C., Xiaohong, W., & Yingying, H. (2020). A deep learning based regression method on hyperspectral data for rapid prediction of cadmium residue in lettuce leaves. *Chemometrics and Intelligent Laboratory Systems*, *200*(March), 103996. https://doi.org/10.1016/j.chemolab.2020.103996

Xu, Z. M., Li, Q. S., Yang, P., Ye, H. J., Chen, Z. S., Guo, S. H., Wang, L. L., He, B. Y., & Zeng, E. Y. (2017). Impact of osmoregulation on the differences in Cd accumulation between two contrasting edible amaranth cultivars grown on Cd-polluted saline soils. *Environmental Pollution*, *224*, 89–97. https://doi.org/10.1016/j.envpol.2016.12.067

Zea, M. (2020). *Using hyperespectral imaging to quantify cadmium stress and estimate concentration in plan leaves* (Número August). West Lafayette, Indiana.