UNIVERSIDAD DE PIURA

FACULTAD DE INGENIERÍA

Estimación de variables de importancia de un modelo basado en Machine Learning (ML) con imágenes hiperespectrales

Tesis para optar el Título de
Ingeniero Mecánico - Eléctrico

César Alejandro Cruz Ruiz
Eduardo Renato Grados Portocarrero

Asesor:
Mgtr. Ing. Juan Carlos Soto Bohórquez
Mgtr. Ing. Juan Junior Valdiviezo Espinoza

Piura, marzo de 2023

NOMBRE DEL TRABAJO

Estimación de variables de importancia d e un modelo basado en Machine Learnin g (ML) con imágenes hiperespectrales

RECUENTO DE PALABRAS

14042 Words

RECUENTO DE CARACTERES

77006 Characters

RECUENTO DE PÁGINAS

59 Pages

TAMAÑO DEL ARCHIVO

8.5MB

FECHA DE ENTREGA

Mar 8, 2023 9:49 AM GMT-5

FECHA DEL INFORME

Mar 8, 2023 9:53 AM GMT-5

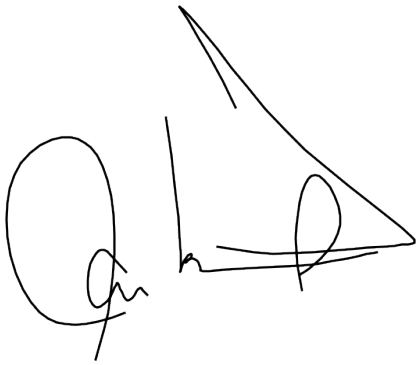
15% de similitud general

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

- 14% Base de datos de Internet
- Base de datos de Crossref
- 7% Base de datos de trabajos entregados
- 2% Base de datos de publicaciones
- Base de datos de contenido publicado de Crossr

Excluir del Reporte de Similitud

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



Mgtr Ing. Juan Carlos Soto Bohorquez

Dedicatoria

A mis queridos padres Carlos y Marisa y a mi hermana Silvana por su incesante apoyo.

César Cruz Ruiz

A mis padres Cristina y Eduardo por su incondicional apoyo.

Eduardo Grados Portocarrero

UNIVERSITAS STUDIORUM PIURENSIS

Agradecimientos

Nuestros agradecimientos especiales a la facultad de Ingeniera de la Universidad de Piura, la cual brindo sus puertas para nuestra formación profesional.

A nuestros docentes que nos acompañaron a lo largo de nuestros estudios universitarios, haciendo una mención especial al Mgtr. Ing. Juan Carlos Soto Bohórquez y Mgtr. Ing. Juan Junior Valdiviezo Espinoza por su apoyo incondicional a la realización de este trabajo investigación.

Y a todas aquellas personas que nos acompañaron a lo largo de este camino alentándonos en todo momento.

UNIVERSITAS STUDIORUM PIURENSIS

Resumen

La agroindustria en el país ha crecido de manera significativa en la última década, gracias a esto el Perú ha logrado ser considerado como uno de los más importantes productores de cacao fino y de aroma, y es el segundo productor de cacao orgánico a nivel mundial.

La Unión Europea ha establecido parámetros de calidad por la presencia de metales pesados, como lo es el cadmio, pues éstos generan efectos nocivos para la salud humana. Esta medida amenaza de forma directa la sostenibilidad de la producción de cacao en el suroeste de América, sin embargo, también tiene como consecuencia mejorar el control de calidad del producto.

La tecnología de Machine Learning se ha convertido en el eje de la cuarta revolución industrial y actualmente viene siendo aplicada en diferentes sectores. Este método consiste en el análisis de datos a partir de procesos analíticos, buscando que los sistemas aprendan, con la ayuda de datos suministrados, identificando patrones en éstos para su posterior toma de decisiones sin la necesidad de la intervención humana. Con la finalidad de mejorar el proceso de aprendizaje se analiza las muestras suministradas identificando cuales son los datos que aportan de manera significativa al modelo, éstas se definen como variables de importancia.

El presente trabajo de investigación propone estimar el porcentaje de cadmio en una muestra de cacao. Con la finalidad de lograr reemplazar el método tradicional de medición invasivo por un método que aplica los algoritmos de Machine Learning, además de la tecnología de Imágenes Hiperespectrales. El método planteado es no destructivo y de resultados inmediatos, cumpliendo el objetivo de reducir la intervención durante el proceso productivo.

UNIVERSITAS STUDIORUM PIURENSIS

Tabla de contenido

Lista de tablas

UNIVERSITAS STUDIORUM PIURENSIS

Lista de figuras

Introducción

La visión hiperespectral son tecnologías emergentes que comenzaron con aplicaciones en la astronomía y teledetección. A su vez el aprendizaje automático nace en 1943 con la publicación de un artículo en el cual se presenta el primer modelo matemático de una red neuronal. Gracias al constante avance tecnológico, esta técnica se ha aplicado en varios campos de la industria, en especial la industria alimenticia. Actualmente, existen investigaciones sobre la evaluación de la calidad de diversos productos agrícolas.

El objetivo principal de este trabajo es aplicar las metodologías de Machine Learning combinada con la Visión Hiperespectral para la agroindustria, con la finalidad de diseñar estrategias para la detección en tiempo real del contenido de cadmio de muestras de Cacao de una manera eficiente utilizando variables de importancia.

Para el desarrollo de la presente investigación se ha empleado la Cámara Hiperespectral Resonon, perteneciente al Laboratorio de Sistemas Automáticos de Control de la Facultad de Ingeniería de la Universidad de Piura, y se han tomado muestras proporcionadas por dicho laboratorio, con la finalidad de obtener las firmas hiperespectrales de diversas muestras de cacao, las cuales servirán de datos para el modelo de predicción.

Durante el desarrollo de este trabajo de Investigación, se indicarán los resultados obtenidos con algoritmos de predicción con Machine Learning, que nos permitirá determinar el porcentaje de cadmio de una muestra de cacao con un margen de error aceptable. Finalmente, se determinarán las variables de importancia de los datos de entrada, esto ayudará al modelo de predicción a tomar las medidas correctivas de manera más eficiente y así, mitigar la contaminación por metales pesados como el cadmio.

UNIVERSITAS STUDIORUM PIURENSIS

Capítulo 1

Identificación del problema a resolver

1.1 Justificación

En los últimos años, las exportaciones de cacao en grano han aumentado de manera sostenida a un porcentaje promedio anual de 4.3%, siendo principalmente empleado para productos como chocolates, confites, alimentos procesados, entre otros. Los principales importadores son la Unión Europea (principalmente Países Bajos, Alemania, Bélgica, Francia, España y Reino Unido) y los Estados Unidos. Cabe destacar que, en la campaña 2020/2021, a nivel mundial, las importaciones de cacao en grano aumentaron 11.8% respecto a la campaña anterior (MIDAGRI, 2022).

En la tabla 1, según un el informe realizado por el Ministerio de Desarrollo Agrario y Riego (MIDAGRI, 2022), el Perú ocupa el octavo puesto en exportaciones de cacao en grano a nivel mundial y el tercer puesto a nivel de Latinoamérica, habiendo exportado en la campaña 2020/2021 un total de 51 mil toneladas de cacao en grano.

Tabla 1

Principales exportadores de cacao en grano en miles de toneladas, 2013 - 2021

Nº	Exportadores	2013/14	2014/15	2015/16	2016/17	2017/18	2018/19	2019/20	2020/21[1/]
	Mundo	3188	2626	2980	3892	3472	3420	3099	3495
1	Costa de Marfil	1117	1286	1056	1562	1530	1578	1539	1646
2	Ghana	-	-	581	611	525	501	417	502
3	Ecuador	199	236	227	285	288	311	329	324
4	Camerún	193	238	264	236	178	220	175	211
5	Bélgica	135	161	187	97	114	169	161	169
6	Países bajos	197	172	139	222	110	192	148	157
7	Malasia	94	71	91	90	104	119	82	106
8	Perú	47	59	62	78	66	65	55	51
9	República Dominicana	68	80	74	66	82	67	74	69
10	Sierra Leona	0	4	10	23	15	13	20	14

Subtotal	2051	2308	2691	3270	3012	3235	3000	3249
Otros	1137	318	289	622	460	185	99	246

Nota. Observatorio de Commodities – Cacao (MIDAGRI, 2022)

En el ámbito nacional, la producción de cacao en grano crece a una tasa de 12.6% en promedio anual. En la tabla 2, se puede observar que la región San Martín es el mayor productor a nivel nacional (35.6% de participación), es importante resaltar que la mayor producción de cacao en grano se da en la sierra peruana; sin embargo, regiones como Piura han aumentado su producción de cacao en 25%, por año, en los últimos diez años.

Tabla 2

Producción de cacao en grano por regiones en toneladas, 2015 - 2022

Dptos.	2015	2016	2017	2018	2019	2020	2021	2021[1/]	2022[1/]
Perú	84814	107922	121825	134676	141775	158944	160222	31940	33177
San Martín	37319	45996	51440	56136	54184	66786	63601	14601	15466
Junín	15334	21400	21801	24755	25560	27536	29774	4057	4653
Huánuco	5292	6491	8912	10392	13403	14395	15958	3705	3974
Cusco	8048	10788	8707	8192	9915	7476	7684	3597	3807
Ucayali	4201	8622	13245	16587	17031	21705	20046	2159	1647
Amazonas	4718	4224	6352	4514	5108	5052	5335	1146	1220
Pasco	1144	1338	1835	3881	4407	4033	4707	1016	1151
Cajamarca	1320	1001	996	955	1121	1137	1263	473	432
Piura	768	658	599	1009	1438	1385	1501	568	71
Ayacucho	4973	5544	5056	5113	5998	5634	6190	42	19
Otros dptos.	1696	1858	2881	3141	3612	3803	4163	31940	33177

Nota. La producción de cacao se vio afectada en 2021 y 2022 debido a malas condiciones ambientales y a la pandemia por Covid-19.

Nota. Observatorio de Commodities – Cacao (MIDAGRI, 2022)

El Perú destaca por su cacao fino y aromático, reconocido por la Organización Internacional del Cacao (ICCO), siendo el segundo productor de cacao orgánico a nivel mundial. En la tabla 3, de acuerdo con un censo realizado por el Instituto Nacional de Estadística e Informática (INEI) en 2022, los cultivos de cacao se establecen como el segundo cultivo permanente de mayor superficie agrícola en el país.

Las exportaciones de cacao y sus derivados han ido incrementando en los últimos años, alcanzando, en el año 2021, un total de 303 millones de dólares, de los cuales el 51.16% corresponden a exportaciones de cacao en grano, siendo los principales destinos de exportaciones de cacao en grano países como Holanda, Indonesia, Bélgica, Estados Unidos y Malasia.

Una de las principales dificultades de este cultivo se encuentra la presencia de metales pesados (principalmente cadmio), tanto en la semilla como en sus productos derivados. La intoxicación por cadmio y otros metales pesados representa un gran riesgo para la salud.

Tabla 3

Superficie de principales productos agrícolas 2022

Cultivo Transitorio	Superficie Cosechada (ha)	Cultivo Permanente	Superficie Cosechada (ha)
Arroz	417 650	Café pergamino	449 528
Papa	330 604	Cacao	184 781
Maíz amarillo duro	252 614	Palta	55 056
Alfalfa	198 468	Vid	35 708
Maíz amiláceo	189 399	Naranja	32 852
Plátano	175 575	Espárrago	31 285
Cebada	126 605	Mango	30 691
Trigo	118 205		
Yuca	104 644		
Palma aceitera	94 902		
Caña de azúcar	84 852		
Frijol grano seco	70 691		
Quinua	68167		

Nota. Información extraída del Compendio Estadístico Perú 2022 (INEI, 2022)

El cadmio es un metal pesado, el cual se puede encontrar en la corteza terrestre en forma de óxidos, asociado con minerales como el zinc, plomo y cobre. Dicho metal se puede encontrar tanto en el aire, suelo, agua, así como también en diversos alimentos, en mayor o menor concentración.

Los principales medios de exposición al cadmio son por riesgos laborales, sobre todo por inhalación de polvos y vapores producto de fundición de minerales, gases de combustión, refinación de metales no ferrosos, disposición e incineración de basura, entre otros; y por presencia de cadmio en alimentos, ya que estos absorben el cadmio del suelo y lo almacenan en los tejidos vegetales y frutos.

El cadmio se acumula principalmente en los riñones, hígado y pulmones. El cuerpo humano es capaz de procesar pequeñas cantidades de cadmio, sin embargo, altas concentraciones de este metal pueden causar efectos negativos para la salud tales como fallas en los riñones, sistema respiratorio, sistema esquelético, y además el cadmio es clasificado como cancerígeno por la Agencia Internacional de Investigación de Cáncer (IARC), causante de cáncer a los pulmones.

Esto ha llevado a instituciones internacionales como la Unión Europea (UE), la Organización de las Naciones Unidas para la Alimentación y Agricultura (FAO), la Comisión Mixta de Expertos en Aditivos Alimentarios (JECFA) y la Organización Mundial de la Salud (OMS) a establecer límites máximos permisibles en cuanto a la concentración de cadmio presente en el cacao y sus productos finales. En la tabla 4, se muestra un comparativo sobre los límites máximos permisibles de las principales instituciones.

Tabla 4

Límites máximos permisibles de cadmio en productos derivados del cacao

Reglamento de la UE 1323/2021 vigente a partir del 28 de Febrero 2022		CODEX ALIMENTARIUS - Estándar Internacional de Alimentos de la FAO	
Descripción	Límite máximo permisible (mg/kg peso húmedo)	Descripción	Límite máximo permisible
Chocolate con leche con contenido <30% de cacao total seco.	0.10	Límite máximo por ingesta de productos en un periodo mensual, designado por la JECFA.	25 µg/kg peso corporal
Chocolate con contenido <50% de cacao total seco. Leche con chocolate con contenido >30% de cacao total seco.	0.30	Chocolate de contenido ≥ 50% y < 70% de cacao total seco.	0.8 mg/kg
Chocolate con contenido ≥50% de cacao total seco.	0.8	Chocolate de contenido ≥ 70% de cacao total seco.	0.9 mg/kg
Cocoa en polvo vendida al consumidor final o como ingrediente de cocoa en polvo endulzada vendido al consumidor final (chocolate para beber).	0.6		

Nota. Información extraída de Unión Europea (2021) y FAO (2019)

Actualmente se cuentan con diversos métodos de laboratorio para determinar la concentración de cadmio en el cacao, entre los cuales se encuentra la espectrometría de Absorción Atómica y Espectrometría de masa. Estos son análisis destructivos en los cuales se

requiere enviar muestras a un laboratorio, éstos deben contar con equipos especializados, ocasionando un alto costo y tiempo para los agricultores de cacao.

Frente a los parámetros planteados por la Unión Europea, con la finalidad de evitar daños a la salud de sus consumidores, se crea la necesidad de establecer métodos para el control de la calidad de este producto y así poder detectar o cuantificar la cantidad de cadmio que el cacao posee durante su proceso de extracción, por medio de pruebas en tiempo real y no destructivas.

En los últimos años, el procesamiento de imágenes ha surgido como una alternativa muy fiable para la evaluación de la calidad de diferentes productos, pues permite el análisis rápido, efectivo, de bajo costo, no destructivo y es capaz de determinar características externas e internas del producto a partir de una imagen digital.

El ML es una de las herramientas más utilizadas en la industria para el procesamiento imágenes hiperespectrales para un análisis confiable, pues permite obtener procesos automatizados para el control de la calidad con muy buena precisión.

1.2 Antecedentes

El presente trabajo se centra en poder determinar las bandas de importancia de un modelo basado en ML con la finalidad de predecir el contenido de cadmio presente en el cacao, utilizando imágenes hiperespectrales, por ello, se ha realizado una revisión de artículos y trabajos de investigación que analizan información de aplicaciones similares de carácter agroindustrial.

La norma ISO 1114:1977 *Cocoa Beans* describe el procedimiento para poder conocer la calidad del grano del cacao obtenido luego del proceso de fermentación. Esta normativa plantea un método tradicional, el cual consiste en realizar un corte de manera longitudinal sobre el grano con la finalidad de visualizar la máxima superficie del interior del cotiledón, tal como se puede apreciar en la figura 1. En dicha normativa se sugiere analizar 300 granos por cada tonelada de cacao para obtener así un análisis certero.

Figura 1

Prueba de corte realizada con una guillotina y prueba de corte realizada con una navaja

Nota. Tomado de “Manual para la evaluación de la calidad del grano de cacao. La Lima, Honduras” (Aguilar H., 2016).

Durante la prueba de corte, es posible identificar y clasificar cada grano etiquetándolo como: bien fermentado, ligeramente violeta, violeta, sobre fermentado, mohoso, pizarroso, y dañado por insectos y roedores. A partir de esta clasificación es posible conocer la calidad de un lote de una tonelada de cacao debido a que determinará el porcentaje de granos bien fermentados de la muestra analizada.

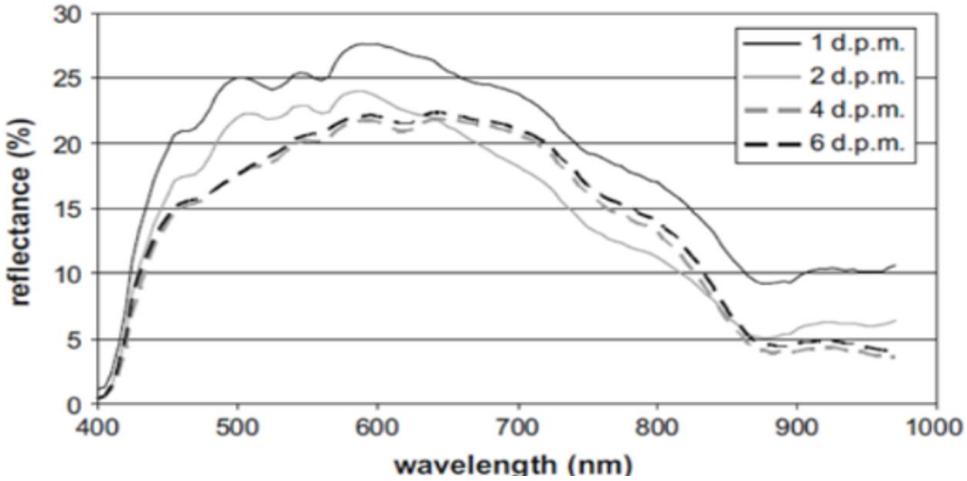
La problemática principal de la prueba de corte es que no cuenta con una precisión certera, debido a que depende principalmente del análisis visual realizado por el evaluador, el cual está sujeto a distintos factores como las condiciones sobre las cuales se realizó la prueba, y la experiencia y percepción visual por parte del evaluador. Además, esta prueba no permite conocer atributos relacionados con la inocuidad del grano, por tal motivo es imposible determinar la concentración de metales pesados presentes en los granos de cacao.

Las imágenes hiperespectrales son comúnmente usadas para analizar elementos de manera no destructiva ni invasiva. Sun (2010) llega a determinar la calidad de la carne de pescado, conociendo el grado de frescura de esta mediante imágenes hiperespectrales e inteligencia artificial, se analizó cada píxel de la imagen hiperespectral para predecir la concentración de agua y grasa, multiplicando el espectro de cada píxel con el vector de coeficiente obtenido a partir del modelo PLSR (Partial Least Squares Regression).

En la figura 2, se muestra como varia el nivel de frescura de la carne de pescado tras varios días post mortem (d.p.m) medida con la reflectancia espectral VIS/NIR (Espectroscopia del infrarrojo cercano).

Figura 2

Firma espectral promedio en días post mortem de una Caballa tras conservarse por congelación



Nota. Tomado de *"Hyperspectral Imaging for Food Quality Analysis and Control. Academic Press"* (Sun, 2010)

En Mundaca (2016), se logró determinar que existe una gran eficacia con el uso de las técnicas de procesamiento de imágenes hiperespectrales, se evaluó la calidad de granos de cacao permitiendo conocer a través de los análisis de índices espectrales, correlacionar el índice de antocianina AR12. En figura 3, se puede observar la imagen hiperespectral de un grano de cacao.

Figura 3

Imagen hiperespectral en NIR y RGB

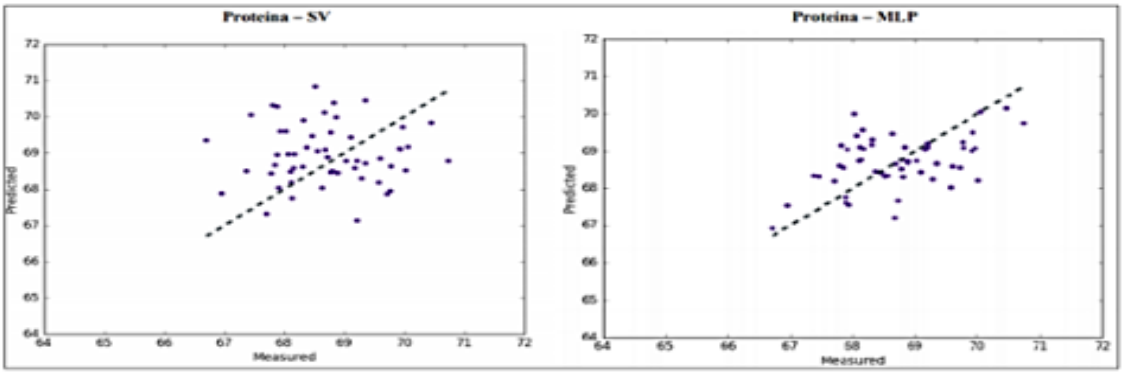
Nota. Tomado de "Análisis de la calidad del grano de cacao mediante imágenes hiperespectrales usando técnicas de visión artificial" (Mundaca Vidarte, 2016).

Cherre Pupuche (2019) aplicó un procedimiento a la salida del proceso de producción de harina de pescado, el cual logra establecer la identificación de parámetros fisicoquímicos de la sustancia en estudio, utilizando la tecnología de Imágenes hiperespectrales mediante procesos de visión artificial, procesamiento de señales y modelos de regresión basados en algoritmos de ML con aprendizaje supervisado. En este estudio, se logra comprobar una mejor calidad de predicción con el algoritmo de Multilayer Perceptrón por encima del algoritmo Support Vector Regression (SVR). En la figura 4, se puede ver un cuadro comparativo de dichos resultados.

Para lograr obtener una correcta información de las imágenes hiperespectrales es indispensable que las imágenes sean capturadas de la mejor calidad posible. Los métodos para el procesamiento y análisis de imágenes digitales RGB, pueden ser utilizadas en el procesamiento y análisis de las imágenes hiperespectrales.

Figura 4

Comparación entre los modelos SVR y Multilayer Perceptrón para estimar el parámetro de la proteína



Nota. Tomado de "Medición de parámetros de calidad de harina de pescado usando imágenes hiperespectrales e inteligencia artificial" (Cherre Pupuche, 2019).

Neyra Hau Yon (2021) comparó distintos métodos de predicción para estimar, en tiempo real, el nivel de cadmio en el cacao, tras aplicar múltiples algoritmos de ML a partir de imágenes hiperespectrales y analizó los resultados obtenidos. Para el desarrollo de este proyecto el autor empleó una cámara Hiperespectral Resonon, dentro la infraestructura del Departamento de Automática y Control de la Universidad de Piura, disponiendo de muestras de Cacao tomadas de algunas locaciones de diversas regiones del Perú.

En el trabajo de investigación se mostraron también los resultados obtenidos con diversos algoritmos de predicción basados en ML, tales como regresión lineal multivariable, ecuación normal, Support Vector Machine (SVM), Partial Least Square (PLS). Estos se han comparado a través del error obtenido entre el valor real y el valor obtenido por el algoritmo de predicción, para esto se utilizó la fórmula del Error Cuadrático Medio (MSE) y la fórmula de Raíz del Error Cuadrático Medio (RMSE).

Peltroche Saavedra (2022) utilizó un algoritmo de redes neuronales artificiales perceptrón multicapa e imágenes hiperespectrales para predecir los valores de contenido de cadmio del cacao, con un margen de error aceptable, utilizando muestras de cacao pertenecientes a la región de Huánuco, Perú.

El trabajo de investigación parte de la generación del cubo hiperespectral, a partir de, escoger distintas muestras de cacao pertenecientes al mismo lote y colocarlas sobre distintos recipientes para su posterior escaneo con la cámara hiperespectral, con la finalidad de encontrar su respectiva firma hiperespectral. La firma hiperespectral representa el valor más relevante presente en cada longitud de onda que la cámara es capaz de detectar. La cámara Resonon modelo Pika II genera 240 imágenes hiperespectrales, por tanto, la firma hiperespectral está formada por los 240 valores más relevantes de cada longitud de onda, las cuales son los datos de entrada de la red neuronal.

Luego de una gran cantidad de pruebas empíricas, el autor determinó que la arquitectura que generaba los mejores modelos fue la arquitectura de red de 240-20-20-5-1. Posteriormente, para este trabajo de investigación, la data fue dividida de forma aleatoria usando la librería SkyLearn de Python, distribuyendo estos como: Datos de entrenamientos 70%, Datos de validación 20%, Datos de testeo 10%. Los resultados en la etapa de testeo se muestran en la tabla 5.

Tabla 5

Resultados del modelo de predicción propuesto por Peltroche Saavedra (2922)

Cd en grano (ppm SGS)	Predicción del modelo	Error en la predicción	Error porcentual
3.78	3.14	0.64	16.93%
2.59	2.23	0.36	13.90%
1.69	1.48	0.21	12.43%
1.6	1.42	0.18	11.25%
1.55	1.55	0	0.00%
1.23	0.91	0.32	26.02%
1.03	0.99	0.04	3.88%
1.01	1	0.01	0.99%
0.98	0.98	0	0.00%
0.95	1.14	-0.19	20.00%
0.86	0.9	-0.04	4.65%
0.81	0.96	-0.15	18.52%
0.58	0.51	0.07	12.07%
0.54	0.47	0.07	12.96%
0.49	0.48	0.01	2.04%
0.45	0.47	-0.02	4.44%
0.43	0.46	-0.03	6.98%
0.42	0.4	0.02	4.76%
0.38	0.38	0	0.00%
0.28	0.39	-0.11	39.29%
0.24	0.29	-0.05	20.83%

Nota. Tomado de “Diseño e implementación de algoritmos inteligentes basados en aprendizaje de máquina para la detección de cadmio en granos de cacao mediante imágenes hiperespectrales” (Peltroche Saavedra, 2022).

1.3 Objetivos

1.3.1 Objetivo general

Desarrollar un modelo basado en ML que permita predecir el contenido de cadmio en el cacao, a partir de imágenes hiperespectrales.

1.3.2 Objetivos específicos

Estudiar y comparar distintos algoritmos de Machine Learning.

Definir las principales bandas de importancia, por métodos de reducción, para estimar el contenido de cadmio en el cacao.

Desarrollar una metodología que permita relacionar las variables del contenido de cadmio con la firma espectral.

1.4 Metodología

En base a la normativa revisada y teniendo en cuenta las regulaciones internacionales establecidas, se definirán los parámetros máximos permisibles para el contenido de cadmio en el cacao.

Se cuenta con un conjunto de datos de distintas muestras de granos de cacao proporcionados por el Instituto Biodiversity y por el medio del Laboratorio de Sistemas Automáticos y de Control de la Universidad de Piura. Esos datos contemplan imágenes hiperespectrales, firmas espectrales y valores de contenido de cadmio para cada muestra.

Se estudiarán distintos tipos de algoritmos de Machine Learning y se escogerá el más adecuado para el caso de estudio, cuyos códigos serán desarrollados en el lenguaje de programación Python. Posteriormente se realizará el proceso de entrenamiento y obtener así un modelo inicial que permita estimar el contenido de cadmio en el cacao, con un error aceptable.

Una vez se tenga un primer modelo establecido, se determinarán las principales bandas de importancia, con métodos de reducción, con la finalidad de requerir menos variables de entrada para el modelo, manteniendo un bajo error en la medición con respecto a la data recopilada.

Capítulo 2

Marco teórico

2.1 Machine Learning

El término Machine Learning hace referencia al procesamiento y clasificación de data e información. Machine Learning utiliza algoritmos capaces de aprender automáticamente basándose en la información provisionada, para así poder generar estimaciones o toma de decisiones. Machine Learning es una rama de la Inteligencia Artificial (IA), en la cual se busca predecir variables de interés, a través del entrenamiento de un modelo, a partir de, una data conocida.

En la actualidad, muchos sistemas utilizan algoritmos de Machine Learning en su funcionamiento, como por ejemplo sistemas de reconocimiento facial, filtros de información, detección de fraudes bancarios, protección de tarjetas de crédito, reconocimientos de voz, entre otros. Dentro de las principales ventajas de utilizar Machine Learning se destacan:

Aprendizaje basado en la "experiencia": El sistema va mejorando a medida que recibe y procesa la información, permitiéndonos realizar tareas de manera más fácil y resolver problemas complejos.

Adaptabilidad: Uno de los grandes problemas de la programación es su rigidez; en cambio, con las herramientas que nos proporciona Machine Learning, el sistema se puede adaptar a distintas entradas sin necesidad de modificar el código.

2.1.1 Tipos de aprendizaje

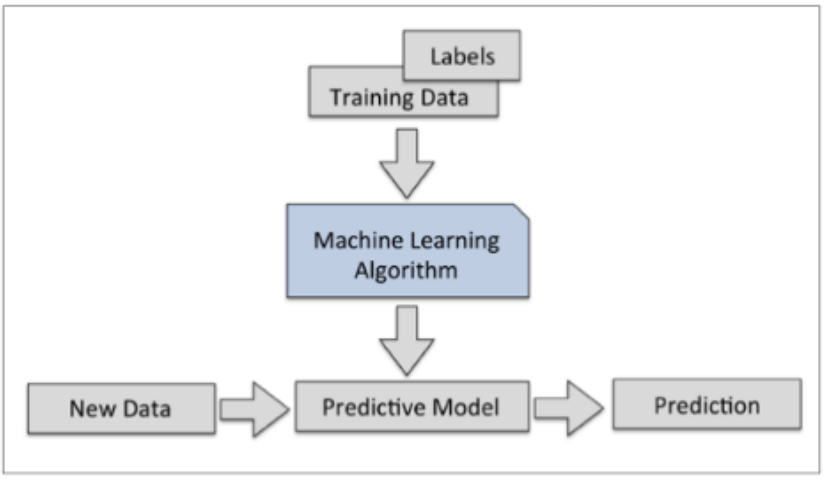
Existen diversas formas de entrenar un modelo de Machine Learning, estas se pueden clasificar en cinco grupos, dependiendo del tipo de interacción que tenga el modelo con el usuario al momento del entrenamiento. El tipo de aprendizaje utilizado dependerá de la data adquirida y de las funciones que realizará dicho modelo.

2.1.1.1 Aprendizaje supervisado. En este tipo de aprendizaje se tiene un conjunto de datos en los cuales se conoce tanto las entradas como salidas o valores deseados del sistema. Este tipo de entrenamiento consiste en proporcionarle al modelo ejemplos para su aprendizaje, a estos ejemplos se le conoce como datos de prueba o datos etiquetados (labels). Se le conoce como supervisado ya que se busca guiar al modelo proporcionándole información

de datos de entrada y salida para el diseño del modelo y de esta forma prediga nuevos datos de entrada. En la figura 5, se muestra la estructura básica del aprendizaje supervisado.

Figura 5

Diagrama de flujo del aprendizaje supervisado



Nota. Tomado de "*Python: Deeper Insights into Machine Learning*" (Hearty et al., 2016).

Este tipo de aprendizaje se puede subdividir en regresión o clasificación dependiendo si los datos de salida son de tipo real (ordinales) o discretos (categóricos). Algunos ejemplos de algoritmos de aprendizaje supervisado son: Redes Neuronales Artificiales, Árbol de Decisiones, Random Forest, Support Vector Machines, Partial Least Squares, entre otros.

2.1.1.2 Aprendizaje no supervisado. Este tipo de aprendizaje se basa en buscar patrones en los valores de entrada utilizando datos sin etiquetar, es decir, se cuenta con un conjunto de datos en los cuales no se conocen los valores de salida.

En el aprendizaje no supervisado, los datos son clasificados en subgrupos, una forma de clasificación es basándose en similitudes o relaciones entre ellos, a este tipo de aprendizaje no supervisado se le conoce como clustering o agrupación. Otro tipo de aprendizaje no supervisado es la reducción dimensional, en la cual se busca reducir o comprimir la información proporcionada, identificando conjuntos de datos relevantes para el sistema y eliminado aquellos menos importantes. El aprendizaje no supervisado nos permite analizar conjuntos de datos más extensos y complejos en comparación con el aprendizaje supervisado; sin embargo, al no conocer los datos de salida, es más difícil determinar si la clasificación o predicción del modelo es acertada y fiable. Algunos ejemplos de aprendizaje no supervisado son: k-means Clustering, Independent Component Analysis (ICA), Principal Component Analysis (PCA), entre otros.

2.1.1.3 Aprendizaje semi-supervisado. Este tipo de aprendizaje combina el aprendizaje supervisado y no supervisado. En este caso se cuenta con un conjunto de datos etiquetados y sin etiquetar. Los datos sin etiquetar sirven como información para que el

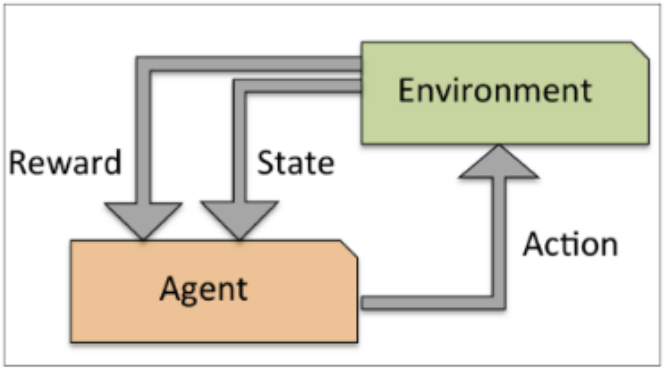
modelo pueda agrupar los datos por homogeneidad, y así poder relacionar mejor las variables de entrada con las variables de salida de los datos etiquetados. Algunos de los algoritmos más comunes de aprendizaje semi-supervisado son: Self-training, Co-training y Mincut.

2.1.1.4 Aprendizaje activo. La palabra activo hace referencia a que el usuario interacciona con el modelo durante la etapa de entrenamiento. Este tipo de aprendizaje parte de un conjunto de datos sin etiquetar. El usuario selecciona algunos datos para etiquetarlos manualmente y añadirlos al modelo como ejemplos. Esto mejora la interacción del modelo en relación a las variables de entrada y salida del sistema. Esto deberá repetirlo hasta que el modelo alcance el rendimiento deseado.

2.1.1.5 Aprendizaje por refuerzo. Este último tipo de aprendizaje nace de analizar como el medio ambiente interacciona con el ser humano y los animales (agentes). Los seres humanos y animales aprenden, en ausencia de un mentor o guía, por un sistema de prueba y error. El aprendizaje por refuerzo consiste en desarrollar un sistema que mejore el rendimiento del modelo por medio de "recompensas", es decir utilizar un algoritmo que pueda medir que tan bien el modelo está trabajando y en base a ello actuar (recompensa), tal como se muestra en la figura 6.

Figura 6

Diagrama de flujo del aprendizaje por refuerzo



Nota. Tomado de "*Python: Deeper Insights into Machine Learning*" (Hearty et al., 2016).

Una forma de comprender este tipo de aprendizaje es relacionándolo con un juego de ajedrez, en el cual se deben tomar decisiones, a partir de la posición de las piezas (agente) en el tablero (medio ambiente).

2.1.2 Algoritmos de Machine Learning

Por la naturaleza de los datos obtenidos en el presente trabajo de investigación, se cuenta con un conjunto de datos complejos, en los cuales conocemos tanto las entradas (reflectancia de cada muestra) y salida (contenido de cadmio). Se estudiarán los distintos tipos de modelos del tipo de aprendizaje supervisado o semi-supervisado.

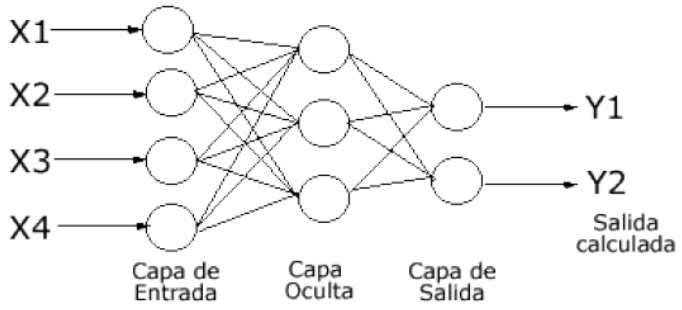
En esta sección se describirán algunos de los algoritmos o modelos más utilizados en ML para clasificar y predecir variables.

2.1.2.1 Redes neuronales artificiales. Una Red Neuronal Artificial es un modelo estadístico no lineal que relaciona las variables de entrada y las de salida para descubrir patrones. Este modelo nace al tratar imitar como trabaja el cerebro humano al momento de tomar decisiones y resolver problemas.

Se llegó a la conclusión que una neurona es una unidad básica del pensamiento lógico, la cual recibe una variable o señal de entrada y entrega un valor de salida a la siguiente neurona. Las neuronas se pueden dividir en tres grupos o capas: capas de entrada, las cuales reciben la información; capas internas ocultas, que procesan dicha información, y capas de salida, aquellas que proporcionan los resultados; tal como se muestra en la figura 7.

Figura 7

Estructura básica de una red neuronal artificial



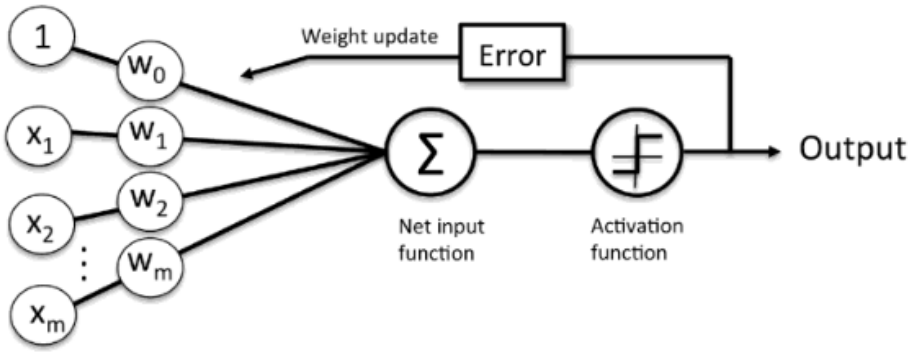
Nota. Tomado de "Determinación en tiempo real de presencia de cadmio en cultivo de cacao aplicando *Machine Learning*" (Neyra Hau Yon, 2021).

Cada una de las neuronas tiene un número de conexiones de entrada y puede tomar ciertos estados de salida. Además, cada neurona cuenta con un peso dependiendo de la relevancia de dicha neurona en la salida del sistema.

En conclusión, una red neuronal es una estructura, en la cual las neuronas reciben datos de entrada, los cuales son multiplicados por sus respectivos pesos y posteriormente son sumados, una vez realizado este proceso se aplica una función de activación con lo cual se obtiene una salida, tal como se puede ver en la figura 8.

Figura 8

Diagrama de aprendizaje de una red neuronal artificial



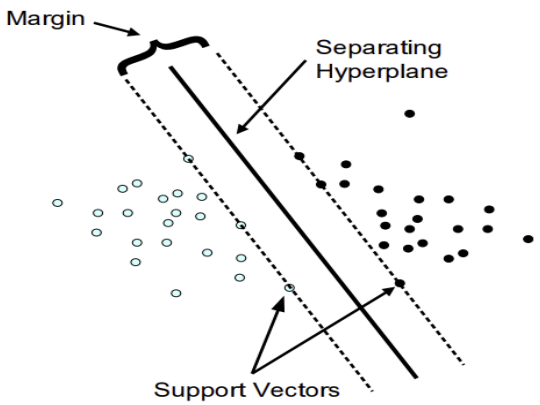
Nota. Tomado de "*Python Machine Learning*" (Raschka y Mirjalili, 2016).

El proceso de entrenamiento de una Red Neuronal Artificial consiste en ir ajustando o variando los pesos de las neuronas, en base a los datos de pruebas anteriores. A medida que la red es entrenada se van obteniendo resultados más precisos hasta que el modelo sea capaz de replicar su funcionamiento con datos los cuales se desconoce el resultado.

2.1.2.2 Support Vector Machines. Support Vector Machines es un algoritmo de tipo de aprendizaje supervisado, el cual permite realizar tareas de clasificación, regresión y detección de valores atípicos. SVM se basa en generar un hiperplano para dividir, de manera óptima, los datos en dos categorías (ver figura 9). Para generar el hiperplano se necesita maximizar la distancia entre los valores más cercanos, a esta distancia se le conoce como margen. A los datos ubicados en el margen de separación de cada clase se les conoce como vectores de soporte.

Figura 9

Separación de datos por medio de un hiperplano



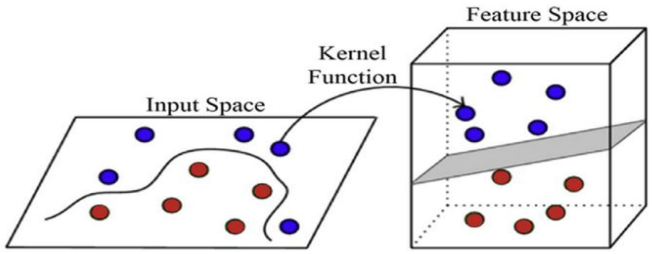
Nota. Tomado de "*Support Vector Machines*" (Meyer, 2015).

Para aquellos datos que no sean separables linealmente, se utilizan funciones kernel para proyectar los datos en un espacio (normalmente de dimensión mayor) en donde el

hiperplano sea lineal y por lo tanto poder realizar la clasificación. En la figura 10, se muestra cómo se proyectan los datos a un espacio de tres dimensiones, de manera que es posible dividirlos por medio de un hiperplano.

Figura 10

Función kernel aplicada para clasificación por Support Vector Machines



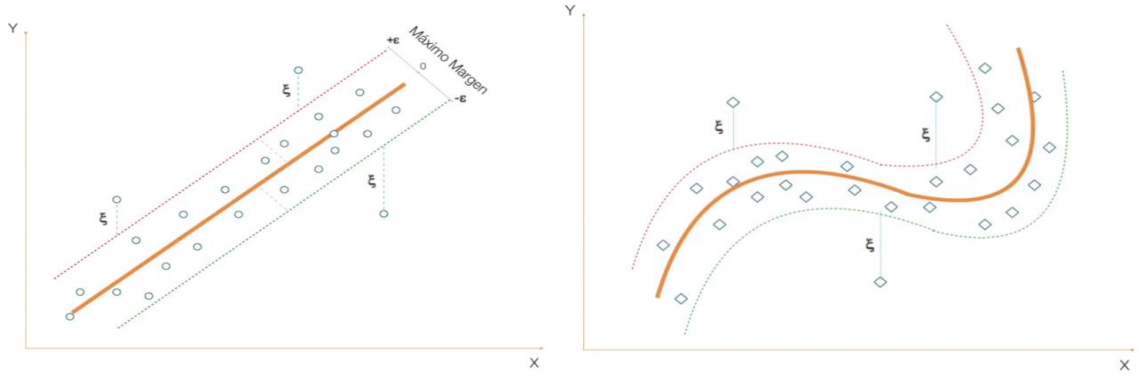
Nota. Tomado de "*Machine Learning; Methods and Applications to Brain Disorders*" (Pisner y Schnyer, 2020).

Para el caso de utilizar el algoritmo de SVM para regresión, o también conocido como Support Vector Regression, se busca seleccionar un hiperplano que mejor se ajuste a los datos, en base a su tendencia. A partir del hiperplano se crearán dos bandas paralelas, una positiva y otra negativa. El margen será la distancia entre ambas bandas. Aquellos datos que no se encuentren entre las bandas serán considerados como slack variables o variables de holgura.

La idea es generar una región o rango que abarque la mayor cantidad de datos posibles y maximizar la distancia épsilon (ξ) entre los datos o variables de holgura y la banda más cercana. En el caso la función de la tendencia sea no lineal, por ejemplo, una curva, las bandas tendrán la misma forma del hiperplano. En la figura 11, se muestran ejemplos de hiperplanos y sus respectivos márgenes para regresión por Support Vector Machines.

Figura 11

Support Vector Machines aplicado para regresión



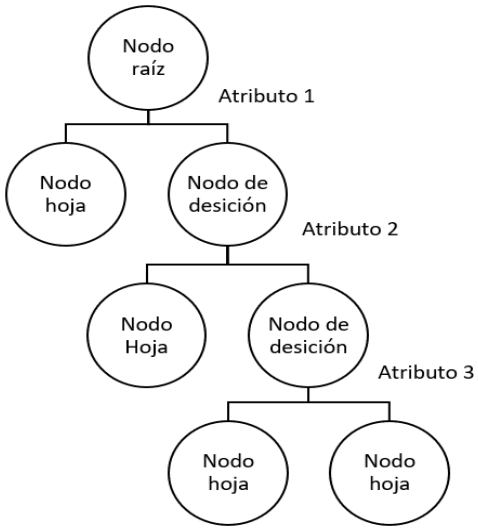
Nota. Tomado de "Aprendizaje Supervisado: Support Vector Machines" (AprendeIA con Lidgi Gonzales, 2018)

2.1.2.3 Árbol de decisiones. Es uno de los modelos de regresión y clasificación más utilizados. Este modelo se basa en realizar predicciones a partir de construcciones lógicas binarias, en otras palabras, por medio de preguntas. Como su nombre lo indica, este modelo representa la estructura de un árbol invertido, tal como se puede apreciar en la figura 12, en el cual se tienen:

- Nodos de decisión: Cada uno representa una pregunta sobre un atributo o condición en particular.
- Nodo raíz o principal: Atributo principal a partir del cual se empieza la clasificación.
- Nodo hoja o final: Representan la clase asignada o resultado de la predicción.
- Ramas: Cada una representa un camino o resultado de una pregunta.

Figura 12

Estructura de árboles de decisión



Al momento de realizar una clasificación, se busca dividir los datos en clases o categorías, que cumplen ciertas características. El objetivo es identificar que reglas o preguntas que contengan la mayor información posible, es decir que mientras menos reglas o ramas tenga el modelo, más eficiente será. Para poder optimizar el modelo se utilizan una serie de indicadores, los cuales son la Entropía, Impurezas de Gini y la Ganancia de Información.

La entropía busca medir la proporcionalidad o cantidad de impurezas en un conjunto de datos. El valor de la entropía varía entre cero y uno. Por ejemplo, si se tienen diez manzanas en una canasta la entropía será cero, ya que solo existe la clase manzana; por otro lado, si en la canasta hay cinco manzanas y cinco peras, la entropía será uno, ya que existe la misma

proporcionalidad para ambas clases; de manera que es más difícil determinar a qué clase pertenece un dato aleatorio.

Impurezas de Gini es un indicador, el cual permite medir la probabilidad con la que el modelo realizará una clasificación incorrecta. También permite determinar la impureza de los datos obtenidos de un nodo. Se va a priorizar utilizar atributos que tengan un valor de impurezas de Gini bajo. Así mismo, el índice de Gini permitirá determinar que atributo colocar en un nodo superior o inferior.

Por último, la Ganancia de Información se puede definir como la variación de la entropía entre un nodo superior y uno inferior. A medida que vamos agregando nodos al modelo, el sistema se va dividiendo cada vez más en subclases o subconjuntos, por lo que la aleatoriedad de los datos va disminuyendo, por ende, también disminuye su entropía (error); es decir, aumenta su ganancia de información.

Para entrenar este modelo es necesario contar con un conjunto de datos de entrenamiento o muestra, los cuales serán divididos en clases. Para clasificar los datos, será necesario generar una serie de reglas base que el modelo deberá seguir. La metodología se centra en crear y ubicar nodos de decisión que permitan obtener la mayor ganancia de información posible e ir calculando los indicadores antes mencionados.

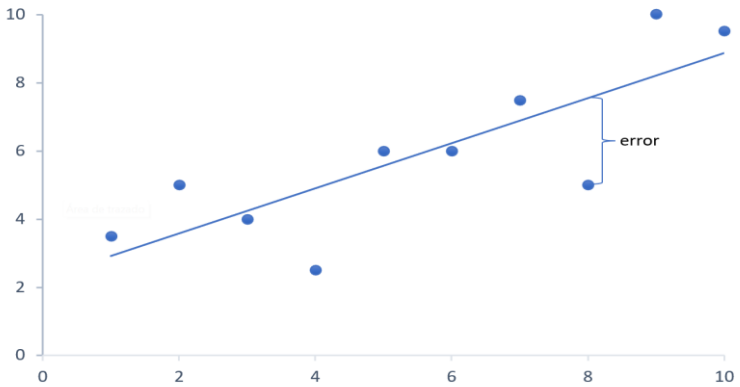
2.1.2.4 Regresión Lineal Multivariable. La regresión lineal simple es un algoritmo de ML el cual permite predecir una variable a partir de un conjunto de variables independientes. Como su nombre lo indica, se busca obtener una función lineal para predecir variables.

$$y = ax + b \tag{1}$$

Para que el modelo permita predecir valores de manera óptima, se debe minimizar el error entre los valores reales y la predicción. Para ello se busca una función que minimice la distancia entre los valores y la función lineal. En la figura 13, se muestra un conjunto de datos con su respectiva función lineal obtenida por regresión.

Figura 13

Ejemplo de regresión lineal aplicada a un conjunto de datos aleatorios



En el caso de regresión lineal multivariable, este modelo permite estimar o predecir una variable a partir de múltiples variables independientes y además estudiar cómo influye cada variable independiente en la estimación. La función de regresión lineal multivariable tendrá la siguiente estructura:

$$y = a_1x_1 + a_2x_2 + \cdots + a_nx_n + b \tag{2}$$

Ya que se tiene más de una variable, es importante determinar qué variables son significativas para el modelo; es decir, aquellas que contribuyen en mayor medida a predecir la variable dependiente. Es importante realizar este paso previo porque es posible que al agregar variables independientes aleatoriamente el error del modelo aumente.

Otro paso a considerar es la Normalización de las variables, lo cual consiste en transformar las variables de diferentes magnitudes a una escala o magnitud en común para su posterior análisis.

2.1.2.5 Partial Least Squares. Partial Least Squares o algoritmo de Mínimos Cuadrados Parciales es un método con el cual se puede construir ecuaciones de regresión para sistemas de una o múltiples variables. Este método permite identificar qué variables influyen realmente en el modelo y qué variables presentan colinealidad.

La regresión por mínimos cuadrados parciales adquiere mayor relevancia cuando tenemos un conjunto de datos cuyo número de variables independientes sea mayor al número de muestras. El principal objetivo de la regresión por PLS es reducir el número de variables independientes a un conjunto de variables significativas que no tengan correlación entre sí.

La regresión por PLS parte de un modelo lineal multivariable, en el cual se quiere predecir una variable dependiente a partir de m variables independientes.

$$Y_{nx1} = X_{mxn} \cdot \beta_{mx1} + E_{nx1} \tag{3}$$

El método de regresión por PLS parte de una reducción de variables, teniendo en cuenta la correlación presente entre dos o más variables independientes. A estas variables se les conoce como variables latentes, las cuales son combinaciones lineales de las variables observadas. Estas variables pueden ser expresadas en forma de matrices, por lo que la estructura del modelo de regresión por PLS tendrá la siguiente forma (Valdéz, 2010).

$$X_{nxm} = T_{nxa} \cdot P_{axn}^{T} + E_{nxm} \tag{4}$$

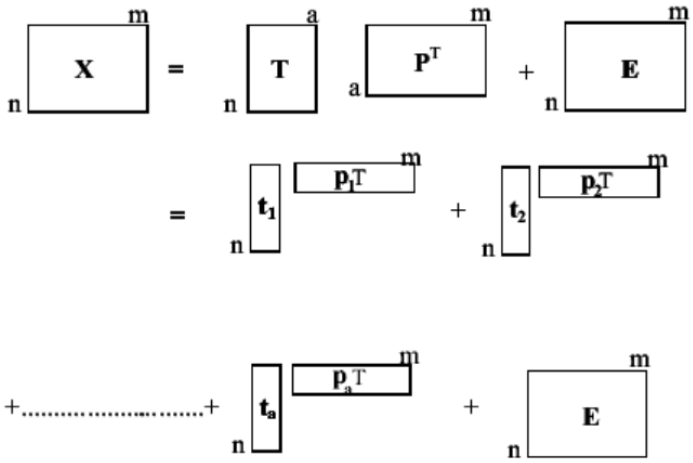
En la ecuación (4), X representa la variable dependiente, la cual queda expresada en función de matriz T, que es la matriz de variables latentes o “resultados”, la matriz P, que representa los pesos o “cargas”, y el error o desviación, representado por la matriz E.

Si se descompone la ecuación (4) para un número de variables latentes a, de manera que a < m, se tendría la siguiente ecuación (5). En la figura 14, se muestra, de manera gráfica, la estructura de la ecuación (5).

$$X_{nxm} = t_{1_{nx1}} \cdot p^{T}_{1_{1xn}} + t_{2_{nx1}} \cdot p^{T}_{2_{1xn}} + \cdots + t_{a_{nx1}} \cdot p^{T}_{a_{1xn}} + E_{nxm} \quad (5)$$

Figura 14

Representación gráfica del método de Mínimos Cuadrados Parciales



Nota. Tomado de "*Partial Least Squares (PLS) regression and its application to coal analysis*". (Carlos et al., 2003).

Si se reemplaza el valor de X en la ecuación (5) se puede hallar una relación entre el vector variable dependiente y y la matriz T.

$$y = T \cdot b + f \quad (6)$$

En la ecuación (6), el vector b, también conocido como "sensibilidades" o coeficiente de regresión, se calcula para minimizar el vector error f. Otro vector calculado durante la etapa de construcción del modelo es el vector w (vector de "pesos" de X), cuyo valor viene dado en función del error E y f.

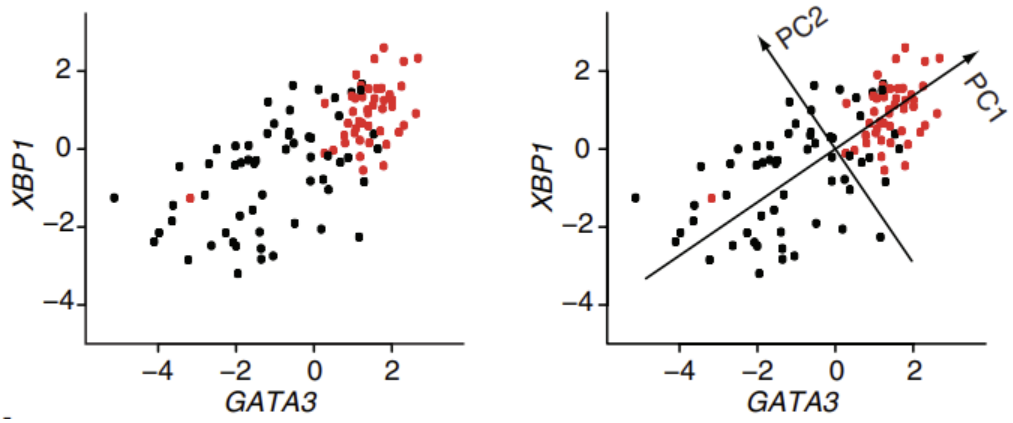
$$W^{T}_{h} = (f^{T}_{h-1} \cdot E^{1}_{h-1})/(f^{T}_{h-1} \cdot f^{1}_{h-1}) \quad (7)$$

2.1.2.6 Principal Components Analysis. Principal Components Analysis o Análisis de Componentes Principales, es un método similar al PLS, el cual nos permite analizar un conjunto de datos, cuyas variables independientes están correlacionadas entre sí.

El objetivo de PCA es el de extraer las variables de mayor relevancia y crear, a partir de ellas, nuevas variables ortogonales; es decir, que no presente colinealidad. A ese nuevo sistema de variables se le conoce como Principal Components (PC) o Componentes principales. En la figura 15, se puede ver un ejemplo de este método.

Figura 15

Ejemplo de componentes principales de un conjunto de datos



Nota. Tomado de "*What is Principal Components Analysis?*" (Ringnér, 2008).

Cada componente principal busca maximizar la varianza de la muestra, de manera que serán ordenados en función de cuanta varianza presente; es decir, el primer componente será aquel que tenga la mayor varianza posible y así sucesivamente con los siguientes componentes.

Sea X una matriz de variables independientes, la matriz de componentes principales Y se puede expresar de la siguiente manera. La matriz Q, o también conocida como matriz de "cargas", representa factores o pesos que hay que aplicar a la matriz X para obtener los componentes principales Y.

$$Y = X \cdot Q \tag{8}$$

2.2 Imágenes hiperespectrales

El contenido de esta sección permitirá un mayor entendimiento respecto a las imágenes hiperespectrales, su aplicación, y como esta técnica se ha convertido en una de las principales formas de adquisición de datos para detección de parámetros de calidad en áreas de control de alimentos, agricultura de precisión y otras ramas. Con la finalidad de comprender mejor la técnica antes mencionada, se presenta la teoría en la que se basa. Por lo tanto, se proporcionará algunos principios fundamentales.

2.2.1 Visión artificial

La visión artificial es la primera etapa en un proceso de inteligencia artificial, es un conjunto de metodologías, conceptos teóricos y técnicas que tienen como finalidad simular el proceso de visión biológica, con el cual se tiene la capacidad de extraer y analizar, de manera inmediata, información de las imágenes obtenidas. Esto permite plantear algoritmos matemáticos y como consecuencia aplicaciones para identificar el contenido de una determinada imagen, con este proceso se puede obtener información de un objeto en el espacio a partir de la adquisición de datos de una o varias imágenes digitales de dicho objeto.

Con la ayuda de este proceso se es capaz de evaluar características propias de un objeto tales como su color, dimensiones, características de su superficie; sin embargo, esta técnica presenta ciertas limitaciones, pero sólo en determinados casos.

2.2.2 Procesamiento

Durante el proceso de obtención de imágenes digitales, suele presentarse distintos factores que afectan directamente a la información de la imagen, tales como ruido entre otras interferencias. Es por este motivo que se aplican diversos procesos con la finalidad de mejorar la calidad de la imagen como la disminución del ruido, cambio el contraste, ajuste de brillo, suavizar o realzar los bordes, encuadrar el enfoque, entre otros.

Según González & Woods (2002), el proceso de mejora de calidad imágenes digitales se dividen en dos grandes categorías las cuales son el dominio espacial y dominio frecuencial. Es posible combinar estas dos categorías antes mencionadas en el procesamiento de una imagen digital.

El conjunto de métodos para la mejora de imagen en el dominio espacial se aplica al mismo plano de la imagen, estos métodos consisten en la manipulación directa de los píxeles de una imagen digital, ya que los pixeles pueden someterse a transformaciones de intensidad individualmente y/o filtrarse de manera individual a aquellos pixeles que opera en las vecindades pertenecientes a la imagen. Por otro lado, los métodos en el dominio frecuencial se aplican en la transformada de Fourier de una imagen.

A continuación, se mencionará las principales técnicas de procesamiento digital de imágenes digitales, estas son aplicables para imágenes multidimensionales y de imágenes hiperespectrales.

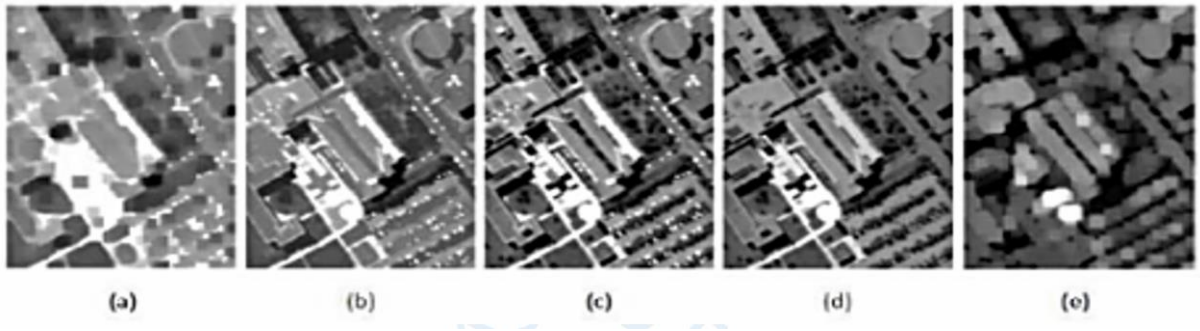
2.2.2.1 Morfología. La técnica de morfología matemática consiste en extraer elementos de la imagen que representan y describen el perímetro de la misma, tales como bordes, límites de forma, esqueletos, etc. Los procesos matemáticos binarios en esta técnica son las operaciones de deslizamiento de un elemento estructurante sobre la imagen digital. La forma y el tamaño escogido del elemento estructurante depende directamente del objeto en estudio.

En la técnica de morfología existen dos principales operaciones las cuales se definen como dilatación y erosión. La dilatación consiste en incorporar en un objeto todos los puntos de fondo que se relacionan de manera directa al mismo, por otro lado, el proceso de erosión consiste en suprimir todos los puntos del perímetro del objeto, con la finalidad de diferenciar pixeles que se encuentran en el objeto de otros pertenecientes a un objeto distinto.

En resumen, es posible indicar que, la técnica de dilatación reduce las diferencias entre dos objetos separados y la erosión suprime detalles muy pequeños, por medio de técnicas como por ejemplo transformada Acierto-o-Fracaso, extracción de límites, adelgazamiento, espesamiento. En la Figura 16, se muestra otras técnicas de morfología.

Figura 16

Ejemplos de técnicas de morfología para el procesamiento de imágenes



Nota. En la imagen se puede apreciar distintas técnicas morfológicas aplicadas a la misma imagen. (a) Cierre morfológico, (b) Cierre por reconstrucción, (c) Imagen pancromática VHR original, (d) Apertura por reconstrucción, (e) Apertura morfológica. Tomado de Ngadi & Liu (2010). Tomado de "Hyperespectral Image Processing Techniques" (Liu y Ngadi, 2010).

2.2.2.2 Histograma. Gonzáles & Woods (2013) indican que el histograma de una imagen digital corresponde la frecuencia relativa de aparición de distintos valores de la escala de grises de la misma imagen. El histograma de una imagen digital corresponde a la función discreta $\mathrm{h(rk) = nk}$ en la cual rk es el valor de la intensidad de orden k y nk es el número de píxeles de la imagen con intensidad rk, sabiendo que los valores de intensidad se encuentran en el rango $[0, \mathrm{L} - 1]$. Posteriormente, se sigue un proceso de ecualización del histograma, el cual consiste en redistribuir los niveles de gris de la imagen digital por la reasignación de los niveles de brillo de cada uno de los píxeles. Uno de los principales problemas en la ecualización de histograma es la presencia de disturbios y ruido, pues estos aumentan la distorsión y contraste de la imagen.

2.2.2.3 Filtro. Un filtro digital es un conjunto de algoritmos donde redefine el valor de un píxel a partir de su valor anterior y el de los pixeles vecinos. Con método no se modifica la geometría de la imagen resultante. El filtro consiste en una función de transferencia que, al aplicarse a una señal de entrada, tiene como resultado una señal de salida según las necesidades planteadas.

Estos, según su tipología, se clasifican en filtro paso bajo, filtro pasa alto y filtro pasa banda. Dependiendo a su dominio de trabajo, estos se pueden clasificar como filtros en el dominio del espacio y filtros en el dominio de la frecuencia.

2.2.3 Segmentación

Otro de los procesos importantes de la visión artificial es la segmentación, esta consiste en extraer una imagen de otra con la finalidad de facilitar el análisis de la misma. La segmentación es empleada en su mayoría para identificar objetos, además de encontrar límites dentro de una imagen digital. La segmentación genera un conjunto de elementos segmentados, que al agruparlos es posible formar un conjunto de curvas de nivel de la imagen.

El proceso de segmentación se basa principalmente en identificación de discontinuidades de nivel en la escala de grises, usada principalmente en detección de perímetros de zonas de interés de una imagen. La segmentación es aplicada en distintos campos como lo son pruebas médicas, localización de objetos en imágenes de satélite, sensores de huella digital, reconocimiento facial, sistemas de control de tráfico, etc.

2.2.4 Firma espectral

La firma espectral de un cuerpo se genera con la respuesta de la energía irradiada de una fuente de luz (energía de luz con amplio rango en el espectro electromagnético) con la materia en estudio. Esta energía al contacto con el cuerpo y su capacidad de reflejar, absorber o transmitir la energía se divide en energía de Reflectancia, Absorbancia y Transmitancia. Los valores de estas energías separadas estarán en función de la composición químicas de la materia. Las firmas espectrales quedan como valores de energía en función de la longitud de onda correspondiente al espectro electromagnético, éstas se establecen como huellas únicas de un objeto pues dependen directamente de su estructura química.

2.2.4.1 Espectrometría. La espectroscopia se define como el estudio de la relación entre la radiación electromagnética y la materia en función de la longitud de onda (λ), con la finalidad medir e identificar la información química y física, cualitativa y cuantitativa de la materia basándose en la Ley de Beer. La evaluación espectral busca detectar la absorción o emisión de radiación electromagnética a diferentes longitudes de onda.

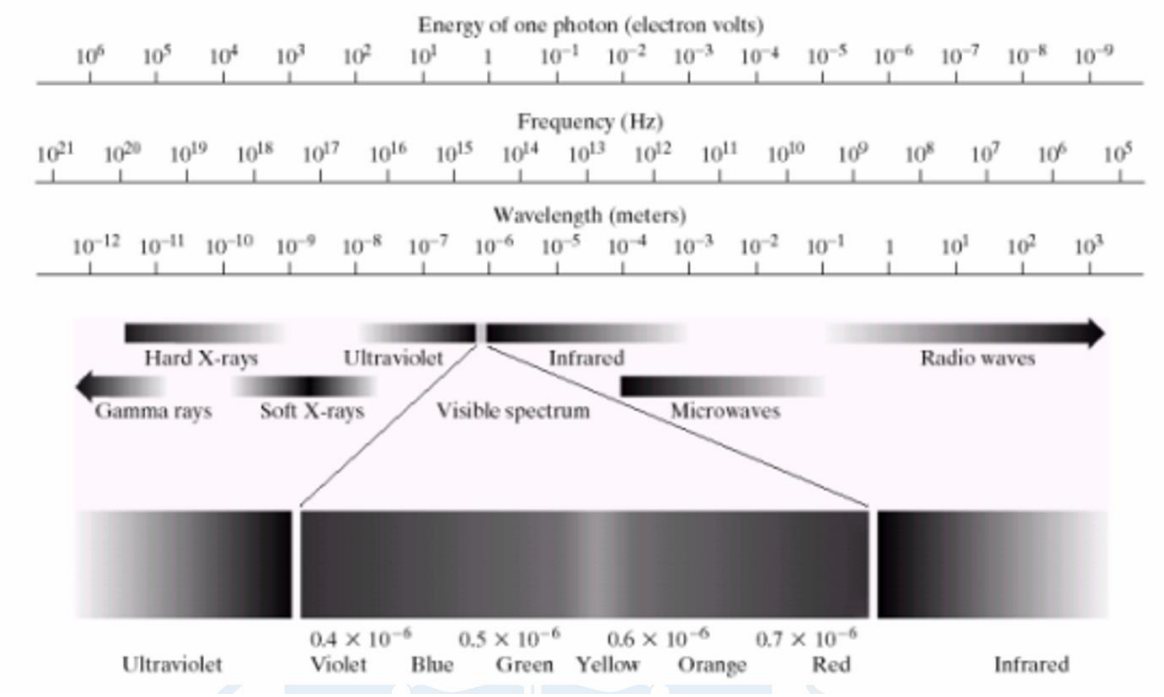
La técnica de espectrometría se originó en 1665, pues Isaac Newton descubrió que cuando un destello de luz atraviesa un prisma de cristal, ésta se descompone en un espectro continuo de colores.

2.2.4.2 Espectro Electromagnético. La energía luminosa, la cual viaja a la velocidad de la luz en forma de ondas, se puede detectar a través de su interacción con el medio ambiente. Esta energía en función de la frecuencia puede ser representada por el espectro electromagnético.

Al identificar todo el rango de espectro electromagnético, se determina que en un extremo del espectro se encuentran las ondas de radio con longitudes de onda miles de millones de veces más grandes que las de la luz visible. En cambio, en el otro extremo del espectro localizan los rayos gamma con longitudes de onda millones de veces más pequeñas que las de la luz visible. En la figura 17 se muestra el espectro electromagnético expresada en energía por fotón, esta comprende desde rayos gamma en la zona más alta y en el otro extremo las ondas de radio.

Figura 17

Espectro electromagnético



Nota. El espectro visible se muestra ampliado. Tomado de "Digital Image Processing" (Gonzáles y Woods, 2013).

Las ondas electromagnéticas sinusoidales propagadas de longitud de onda λ, se definen también como una corriente de partículas sin masa que viajan en forma de onda a la velocidad de la luz. Es posible expresar espectro electromagnético en términos de longitud de onda, frecuencia y energía.

La longitud de onda λ es expresada como λ=c/v, donde v corresponde a la frecuencia y donde c es la velocidad de la luz (2.998×108 m/s). A si mismo la energía de los diversos componentes del espectro electromagnético está dada por la expresión E = hv, donde h corresponde a la constante de Planck.

El ojo humano solo puede identificar una porción muy pequeña del espectro electromagnético, La gama de colores que percibimos se define como luz visible, este rango va desde aproximadamente 0.43 µm (violeta) hasta alrededor de 0.79 µm (rojo). El espectro de color se divide en seis grandes regiones para su mejor identificación, estas son violeta, azul, verde, amarillo, naranja y rojo.

Una forma de describir las características de una fuente de luz cromática es clasificar en tres categorías básicas: radiancia, luminancia, y el brillo. Radiancia se define como la cantidad total de energía que fluye de la fuente de luz, luminancia es la cantidad de energía que un observador percibe a partir de una fuente de luz medida en lúmenes, y el brillo es un medidor subjetivo de percepción de la luz.

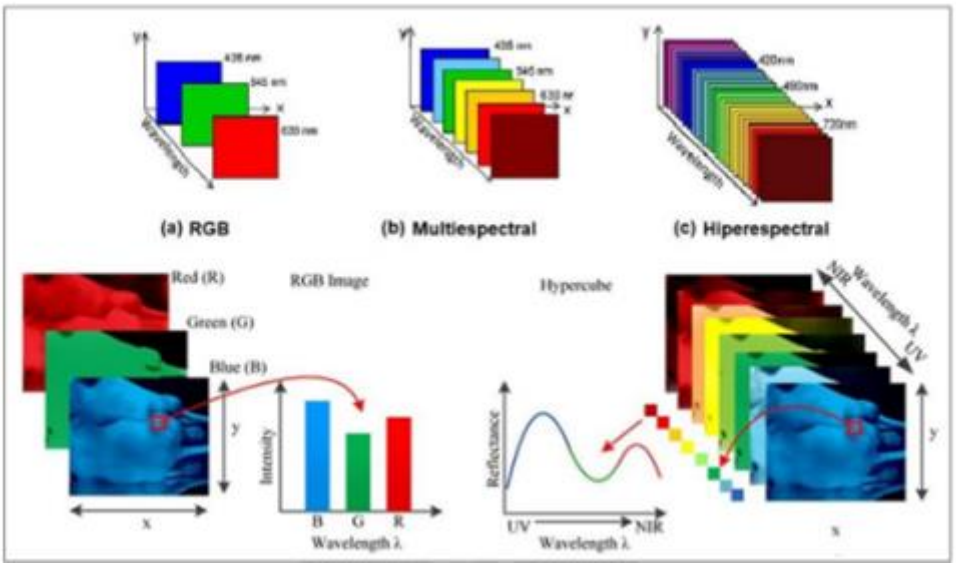
2.2.4.3 Imágenes multiespectrales. Las imágenes multiespectrales se identifican como aquellas que capturan información de un rango determinado del espectro electromagnético, que va más allá del espectro visible identificado como RGB. Estas imágenes pueden estar formadas por varios planos y cada plano corresponde a una campana dentro de un rango del espectro electromagnético. Al valor de luz obtenido producto de la incidencia de la misma sobre el cuerpo se le conoce como reflectancia.

2.2.4.4 Imágenes Hiperespectrales. Las imágenes hiperespectrales, son aquellas cuyo rango cubre un amplio rango del espectro electromagnético. Es posible lograr recopilar información de los espectros electromagnéticos consecuentes de la emisión de la luz sobre los cuerpos, también llamados reflectancia, absorbancia y transmitancia.

La espectrometría de imagen, técnica que usa las imágenes hiperespectrales, es una práctica cada vez más utilizadas en el estudio de propiedades de la materia, usando sensores que no estén en contacto con la materia en estudio, en este caso el sensor es una cámara el cual es un elemento pasivo que recibe la energía reflejada por el objeto, esta energía se conoce como Reflectancia. En la figura 18, se puede apreciar las distintas bandas espectrales que es capaz de obtener una cámara de acuerdo a su capacidad.

Figura 18

Comparación entre imágenes de distintas bandas espectrales



Nota. Tomado de "El análisis de calidad de semillas en un nuevo escenario tecnológico" (Arango et al., 2020).

2.2.4.5 Medición de parámetros. El modelo empleado para el presente informe es el Resonon Pika – II, este está constituido por un sistema integral de hardware y software que es capaz de la captura y analizar respectivamente imágenes hiperespectrales. El modelo Resonon Pika II captura imágenes por medio de un escaneo en línea, además cubre el rango

en el espectro electromagnético visible y parte del infrarrojo cercano, tomando así un rango de longitudes de onda desde 400 nm a 900 nm.

El equipo Pika II, mostrado en la figura 19, es un dispositivo compacto y de gran capacidad, debido a que puede proporcionar imágenes de alta calidad, buena proporción de señal y ruido también presenta ligeras distorsiones.

Figura 19

Cámara Resonon Pika II

Nota. Cámara hiperespectral perteneciente al Laboratorio de Sistemas Automáticos y de Control de la Universidad de Piura. Tomado de "Predicción de parámetros de calidad de la harina de pescado utilizando Imágenes Hiperespectrales y Redes Neuronales Artificiales" (Moscol et al., 2021).

El sistema incluye una fuente iluminación con cuatro lámparas halógenas de cuarzo, estas se encuentran especialmente diseñadas para soportar elevadas temperaturas como la generada por el ciclo halógeno ideal para un buen muestreo, Estas lámparas por sus características proporcionan iluminación adecuada en todas las longitudes de onda para adquirir datos hiperespectrales de alta calidad, las luces son controladas por una fuente de alimentación estabilizada que minimiza la variación debido a las fluctuaciones.

Este dispositivo presenta también una plataforma de desplazamiento llamado stage y una torre de montaje de aluminio acanalada para graduar la posición del soporte de la cámara y las lámparas.

2.3 Lenguajes de programación para Machine Learning

La Inteligencia Artificial y Machine Learning se centran en el diseño de modelos capaces de realizar tareas que el ser humano no podría o le supondría mucha dificultad. A diferencia de la programación convencional, la programación de modelos de ML busca que el mismo modelo pueda aprender y mejorar automáticamente. Esto puede suponer una difícil tarea para cualquier desarrollador, es por ello que varios lenguajes de programación han desarrollado herramientas y librerías que facilitan el trabajo de programar un modelo y su posterior entrenamiento.

Cada lenguaje de programación tiene sus ventajas y restricciones, algunos incluso han desarrollado entornos específicos para resolver determinados problemas. Entre los principales lenguajes o entornos de programación se encuentran R, Matlab y Python.

2.3.1 R

R es el lenguaje de programación para modelos estadísticos por excelencia. Es un software libre, por lo que sus usuarios pueden utilizar y modificar algoritmos ubicados en su librería para adaptarlos a sus necesidades. Además, R tiene una gran cantidad de herramientas estadísticas, lo cual permite trabajar y analizar bases de datos con mayor facilidad, así como generar gráficos.

Otra ventaja de R es que es un software que mejora continuamente, lanzando nuevas versiones, desarrollando nuevas herramientas y funcionalidades, incorporando nuevas librerías, etc.

En cuanto a Machine Learning, R posee entre sus principales herramientas al paquete "Caret". El paquete Caret (Classification and Regression Training) fue desarrollado por Max Kuhn y ofrece un conjunto de funciones que permiten preprocesar los datos, subdividirlos en clases, categorizar los datos, estimar variables de importancia, entre otros.

2.3.2 Matlab

Matlab es uno de los entornos de programación más utilizados en Ingeniería y Estadística. Matlab posee una amplia biblioteca de herramientas integradas, o también conocidas como Toolbox. Dentro de las herramientas que ofrece Matlab se encuentra MathWorks, el cual posee, dentro de sus principales aplicaciones, un conjunto de paquetes destinados para el desarrollo de modelos de Machine Learning.

MathWorks provee de algoritmos para la adquisición y análisis de datos, así como para su posterior entrenamiento. Además, MathWorks puede ser utilizado en conjunto con otras herramientas que ofrece Matlab. (Paluszek & Thomas, 2017)

Entre las principales herramientas que tiene MathWorks tenemos a:

- Statistics and Machine Learning Toolbox: Permite el análisis y la Adquisición de grandes conjuntos de datos. Permite clasificar datos, aplicar regresiones y clustering.

- Neural Network Toolbox: Permite crear, entrenar y simular redes neuronales. También tiene herramientas con las que se puede visualizar y graficar los resultados para un mejor entendimiento.

- Computer Vision System Toolbox: Permite desarrollar funciones para el procesamiento de datos de imágenes tanto 2D como 3D.

- System Identification Toolbox: En conjunto con Simulink, permite crear modelos matemáticos para Machine Learning, tanto para sistemas lineales como no lineales.

2.3.3 Python

Python es uno de los lenguajes de programación más utilizados para el desarrollo de modelos de Inteligencia Artificial, Ciencia de Catos y Machine Learning. El lenguaje de programación Python ha crecido a gran escala en los últimos años y posee una gran comunidad de programadores activos, los cuales comparten sus códigos y algoritmos gratuitamente. Esto ha hecho que Python cuente con un número elevado de librerías para Machine Learning que otros entornos de programación.

Además, Python es un software libre y gratuito, ofrece un lenguaje de programación altamente interpretativo e intuitivo, lo cual lo hace fácil de aprender y tiene un entorno sencillo y atractivo para muchos desarrolladores.

Entre las principales librerías tenemos:

- Pandas: Permite cargar bases de datos de diferentes formatos, para posteriormente analizarlos, filtrarlos, transformarlos, etc.

- NumPy: Permite guardar información en forma de arrays y ejecutar cálculos matemáticos con ellos, tales como operaciones lógicas, extracción de datos, aplicar funciones estadísticas, entre otras funciones.

- SciPy: Permite visualizar, optimizar los datos, realizar operaciones estadísticas, operaciones algebraicas con los arrays proporcionados por Numpy.

- Scikit-learn: Esta librería posee una gran cantidad de algoritmos de Machine Learning, como por ejemplo clasificaciones, regresiones, clustering, reducción dimensional, entre otros.

UNIVERSITAS STUDIORUM PIURENSIS

Capítulo 3

Evaluación y resultados

En el presente capítulo se describe con mayor detalle la metodología utilizada en la obtención del modelo de predicción que permita estimar el contenido de cadmio en el cacao con las muestras obtenidos de las imágenes hiperespectrales y contenido de cadmio. Así mismo, se explica los modelos de Machine Learning desarrollado en el software Python, los resultados obtenidos y su posterior análisis.

3.1 Descripción del caso de estudio

Para el caso de estudio desarrollado se cuenta con un conjunto de datos de distintas muestras de cacao, de las cuales se obtuvo sus respectivas imágenes hiperespectrales, obtenidas con la cámara hiperespectral Resonon Pika-II del Laboratorio de Sistemas Automáticos y de Control de la Universidad de Piura. Posteriormente, se analizaron dichas muestras de cacao para obtener la concentración de cadmio, expresada en partes por millón (ppm), de cada muestra.

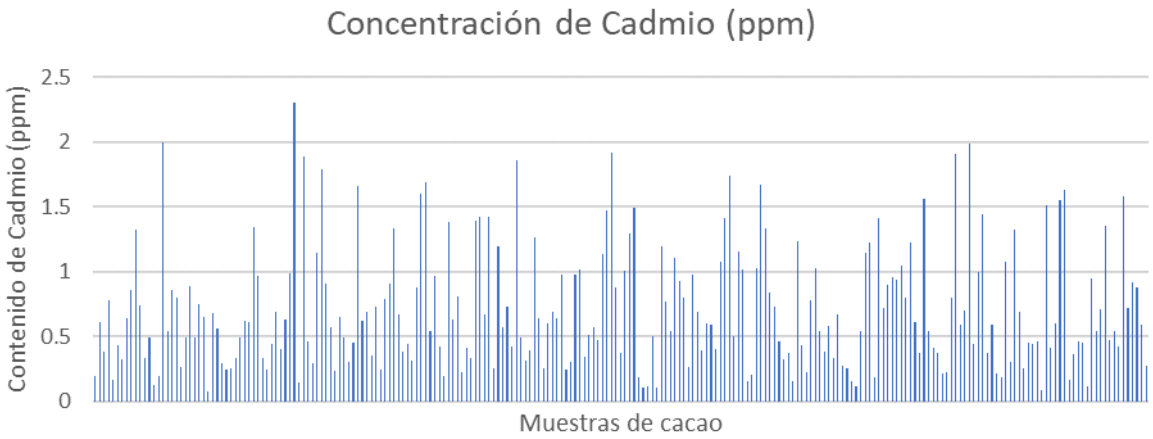
En la figura 20, se puede observar las concentraciones de cadmio de cada muestra, las cuales tienen una media de 0.71 ppm de cadmio. Así mismo, se puede evidenciar qué, según la normativa internacional analizada previamente, el 58% de las muestras superan las 0.5 ppm, lo cual puede derivar en limitaciones en las exportaciones a los países que acaten dicha normativa.

A partir de estos datos iniciales, es posible analizar la data obtenida de la toma de imágenes hiperespectrales para estimar dichos valores de cadmio de las muestras por medio de un modelo basado en Machine Learning desarrollado en el lenguaje de programación Python.

Se ha determinado el uso del software Python como entorno de programación debido a la gran cantidad de librerías de algoritmos de Machine Learning, lectura y procesamiento de datos, y gráficos que posee.

Figura 20

Concentración de Cadmio en las muestras de granos de cacao, en ppm



Para el desarrollo del modelo de predicción se ha utilizado varios modelos de ML, teniendo en cuenta la cantidad de muestras con las que se dispone y a la gran cantidad de variables independientes que proporcionan las imágenes hiperespectrales.

Adicionalmente, se estudiará el algoritmo VIP para determinar aquellas variables de importancia, o también llamadas variables significativas, para la predicción del modelo y así poder estudiar los resultados provenientes de utilizar un conjunto de variables reducidas dimensionalmente.

3.2 Imágenes hiperespectrales de granos de cacao

La primera parte para el desarrollo del presente trabajo consiste en la generación del cubo hiperespectral del elemento en estudio. Para la obtención del cubo espectral se procedió a tomar distintas muestras de cacao, las cuales se colocan sobre distintos recipientes para su escaneo con una cámara hiperespectral.

Para el desarrollo de este trabajo de investigación se utilizó la cámara hiperespectral Resonon modelo Pika II, la cual cuenta con un rango espectral de 400nm hasta los 900nm y una resolución espectral de 2.1 nm.

En nuestro caso, un cubo espectral está formado por 240 imágenes de la misma escena correspondiente a su respectiva longitud de onda.

El cubo espectral puede ser visualizado mediante el software Spectronon, Python u otro. La imagen hiperespectral se guarda en dos archivos, con extensión .hdr y .bill que corresponden a la metadata y al cubo espectral respectivamente.

El archivo denominado con la extensión .hdr brinda información del software y configuración de la cámara durante la toma de la muestra, a esta también se le denomina metadata. La información guardada en los archivos .hdr será la misma en todos los casos

debido a que todas las muestras fueron tomas bajo las mismas condiciones. Por otro lado, el archivo con la extensión .bill representa la información del cubo hiperespectral. Este puede ser abierto con el software Spectronon, Matlab, Python u otro.

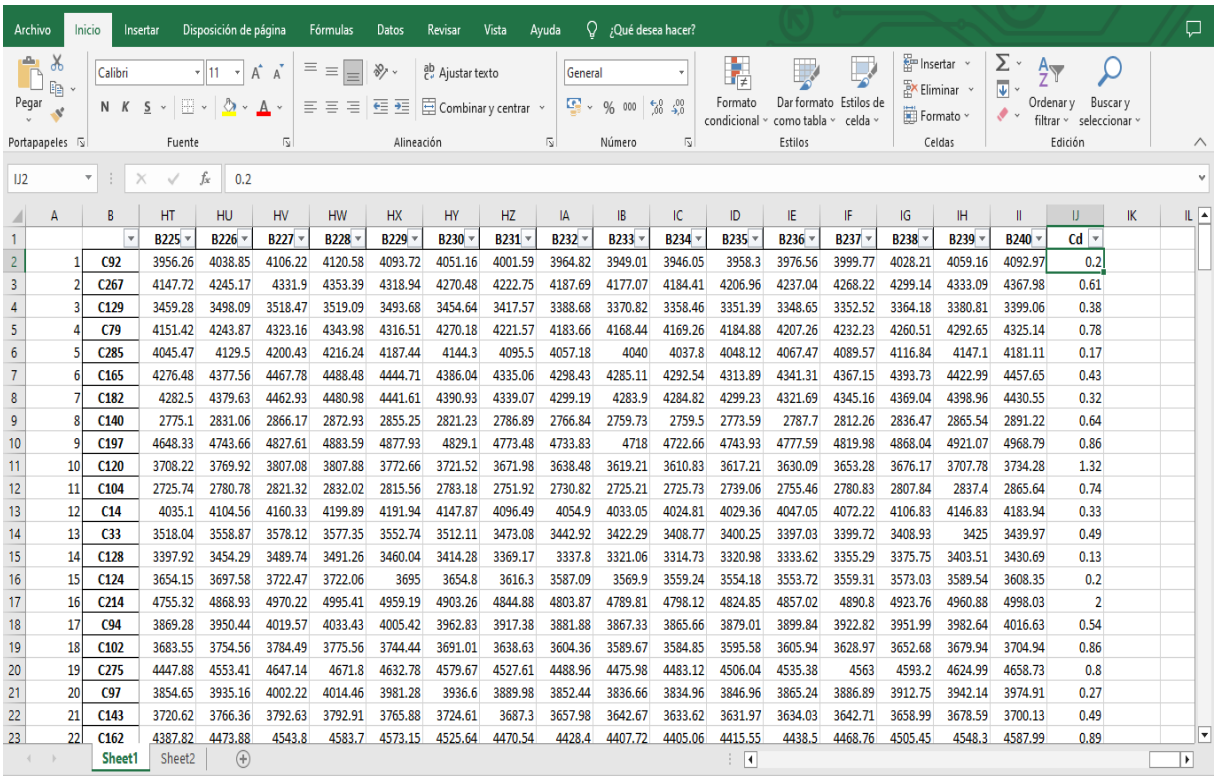
3.2.1 Firma espectral

La firma espectral representa los valores de reflectancia respecto a su longitud de onda. Para la generación de la firma espectral de nuestra muestra, se determinó una región de interés en el grano conocida como Region of Interest (ROI), a este ROI se determina su firma espectral promedio, el cual consta de 240 bandas o longitudes de onda.

Las firmas hiperespectrales son usadas en la generación del modelo de ML. Los datos de las firmas espectrales son almacenados en una hoja de cálculo con extensión .xlsx, pudiendo también ser archivo .csv tal como se muestra en la figura 21. Haremos uso de la librería.

Figura 21

Datos obtenidos de las imágenes hiperespectrales en archivo excel

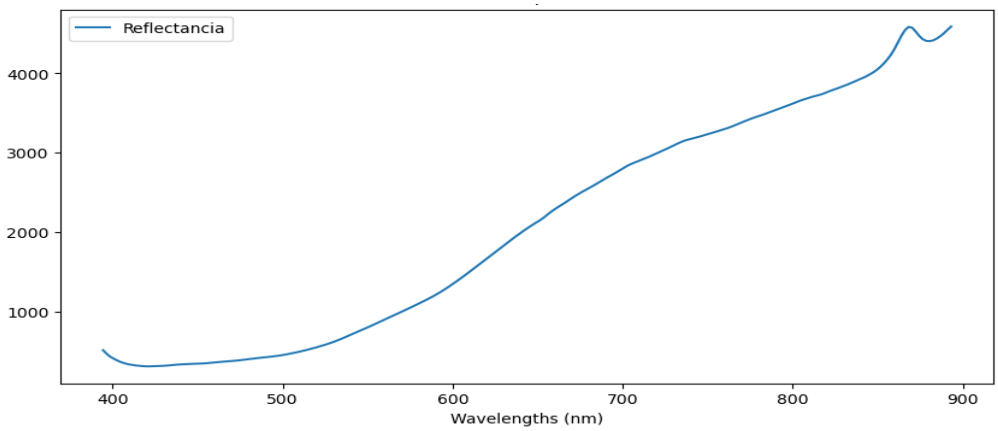


	A	B	HT	HU	HV	HW	HX	HY	HZ	IA	IB	IC	ID	IE	IF	IG	IH	II	IJ	IK	IL
1			B225	B226	B227	B228	B229	B230	B231	B232	B233	B234	B235	B236	B237	B238	B239	B240	Cd		
2	1	C92	3956.26	4038.85	4106.22	4120.58	4093.72	4051.16	4001.59	3964.82	3949.01	3946.05	3958.3	3976.56	3999.77	4028.21	4059.16	4092.97	0.2		
3	2	C267	4147.72	4245.17	4331.9	4353.39	4318.94	4270.48	4222.75	4187.69	4177.07	4184.41	4206.96	4237.04	4268.22	4299.14	4333.09	4367.98	0.61		
4	3	C129	3459.28	3498.09	3518.47	3519.09	3493.68	3454.64	3417.57	3388.68	3370.82	3358.46	3351.39	3348.65	3352.52	3364.18	3380.81	3399.06	0.38		
5	4	C79	4151.42	4243.87	4323.16	4343.98	4316.51	4270.18	4221.57	4183.66	4168.44	4169.26	4184.88	4207.26	4232.23	4260.51	4292.65	4325.14	0.78		
6	5	C285	4045.47	4129.5	4200.43	4216.24	4187.44	4144.3	4095.5	4057.18	4040	4037.8	4048.12	4067.47	4089.57	4116.84	4147.1	4181.11	0.17		
7	6	C165	4276.48	4377.56	4467.78	4488.48	4444.71	4386.04	4335.06	4298.43	4285.11	4292.54	4313.89	4341.31	4367.15	4393.73	4422.99	4457.65	0.43		
8	7	C182	4282.5	4379.63	4462.93	4480.98	4441.61	4390.93	4339.07	4299.19	4283.9	4284.82	4299.23	4321.69	4345.16	4369.04	4398.96	4430.55	0.32		
9	8	C140	2775.1	2831.06	2866.17	2872.93	2855.25	2821.23	2786.89	2766.84	2759.73	2759.5	2773.59	2787.7	2812.26	2836.47	2865.54	2891.22	0.64		
10	9	C197	4648.33	4743.66	4827.61	4883.59	4877.93	4829.1	4773.48	4733.83	4718	4722.66	4743.93	4777.59	4819.98	4868.04	4921.07	4968.79	0.86		
11	10	C120	3708.22	3769.92	3807.08	3807.88	3772.66	3721.52	3671.98	3638.48	3619.21	3610.83	3617.21	3630.09	3653.28	3676.17	3707.78	3734.28	1.32		
12	11	C104	2725.74	2780.78	2821.32	2832.02	2815.56	2783.18	2751.92	2730.82	2725.21	2725.73	2739.06	2755.46	2780.83	2807.84	2837.4	2865.64	0.74		
13	12	C14	4035.1	4104.56	4160.33	4199.89	4191.94	4147.87	4096.49	4054.9	4033.05	4024.81	4029.36	4047.05	4072.22	4106.83	4146.83	4183.94	0.33		
14	13	C33	3518.04	3558.87	3578.12	3577.35	3552.74	3512.11	3473.08	3442.92	3422.29	3408.77	3400.25	3397.03	3399.72	3408.93	3425	3439.97	0.49		
15	14	C128	3397.92	3454.29	3489.74	3491.26	3460.04	3414.28	3369.17	3337.8	3321.06	3314.73	3320.98	3333.62	3355.29	3375.75	3403.51	3430.69	0.13		
16	15	C124	3654.15	3697.58	3722.47	3722.06	3695	3654.8	3616.3	3587.09	3569.9	3559.24	3554.18	3553.72	3559.31	3573.03	3589.54	3608.35	0.2		
17	16	C214	4755.32	4868.93	4970.22	4995.41	4959.19	4903.26	4844.88	4803.87	4789.81	4798.12	4824.85	4857.02	4890.8	4923.76	4960.88	4998.03	2		
18	17	C94	3869.28	3950.44	4019.57	4033.43	4005.42	3962.83	3917.38	3881.88	3867.33	3865.66	3879.01	3899.84	3922.82	3951.99	3982.64	4016.63	0.54		
19	18	C102	3683.55	3754.56	3784.49	3775.56	3744.44	3691.01	3638.63	3604.36	3589.67	3584.85	3595.58	3605.94	3628.97	3652.68	3679.94	3704.94	0.86		
20	19	C275	4447.88	4553.41	4647.14	4671.8	4632.78	4579.67	4527.61	4488.96	4475.98	4483.12	4506.04	4535.38	4563	4593.2	4624.99	4658.73	0.8		
21	20	C97	3854.65	3935.16	4002.22	4014.46	3981.28	3936.6	3889.98	3852.44	3836.66	3834.96	3846.96	3865.24	3886.89	3912.75	3942.14	3974.91	0.27		
22	21	C143	3720.62	3766.36	3792.63	3792.91	3765.88	3724.61	3687.3	3657.98	3642.67	3633.62	3631.97	3634.03	3642.71	3658.99	3678.59	3700.13	0.49		
23	22	C162	4387.82	4473.88	4543.8	4583.7	4573.15	4525.64	4470.54	4428.4	4407.72	4405.06	4415.55	4438.5	4468.76	4505.45	4548.3	4587.99	0.89		

Las muestras utilizadas en el presente trabajo de investigación fueron proporcionadas por el Laboratorio de Sistemas Automáticos de Control de la Facultad de Ingeniería de la Universidad de Piura, la cual está compuesta por 233 muestras de cacao y de la cuales se obtuvieron las firmas espectrales respectivas. En la figura 22, se logra apreciar la firma espectral de una muestra de cacao.

Figura 22

Firma espectral de una muestra de cacao



3.3 Desarrollo de método de predicción

Analizando el caso de estudio, se ha determinado que corresponde a un problema o caso de regresión, ya que el objetivo principal es el de hallar un valor numérico continuo. Es por ello que, para desarrollar dicho modelo, se han utilizado algoritmos de ML para regresión. Para obtener dicha predicción se ha tenido que seguir algunos pasos previos, con la finalidad de determinar el modelo que pueda estimar los valores con el menor error posible.

3.3.1 Importación de librerías

Python cuenta con un gran número de módulos, librerías y paquetes los cuales nos facilitan el desarrollo de códigos por medio de funciones que podemos importar. Estos paquetes o librerías son un conjunto de códigos desarrollados por la comunidad de software libre o empresas de investigación, siendo compartidos para su uso gratuito en algunos casos. El primer paso para desarrollar cualquier modelo de ML es importar las librerías que serán utilizadas en el código. Para el actual caso de estudio se han utilizado las siguientes librerías:

- Pandas: Permite la importación de la data de las Imágenes Hiperespectrales, del formato Excel al entorno de programación Python, así como la selección de columnas, filtro de datos, eliminación de datos no relevantes, selección de variables de entrada y salida, entre otros.

- Numpy: Permite el uso de arrays y matrices para el almacenamiento de información, y su posterior análisis u operación.

- Matplotlib: Es una librería de visualización de Python, la cual permite la creación de gráficos, visualizar datos, crear tablas, entre otros.

- Scikit-learn: Posee una gran variedad de paquetes y funciones, entre los cuales se han utilizado Mean Square Error, R2_score, PLS Regression, SVR, Cross Val Predict y Train – Test Split.

3.3.2 Definir las variables independientes

Las firmas espectrales obtenidas de las imágenes hiperespectrales han sido guardadas en un archivo Excel. Para importar los datos a Python se ha utilizado la librería Pandas.

Los datos importados son almacenados en un dataframe, el cual contiene información adicional que no utilizaremos en este estudio, así tenemos: Cadmio, fecha de medición, lote de la muestra, ubicación GPS, nombre de distrito y provincia, entre otros.

Los datos de entrada y salida corresponden a la firma espectral y valor de cadmio correspondientemente. La variable de entrada cuenta con 240 parámetros independientes, que representan el valor de reflectancia según su longitud de onda específico. La variable de salida corresponde a una columna que contiene el valor de cadmio admisible en ppm.

De esta manera se define la matriz de variables de entrada y salida para el desarrollo del modelo.

3.3.3 Separación de datos para entrenamiento y prueba

Para propósitos de entrenamiento y validación del modelo, se separó la matriz de variables en dos partes. Una primera parte fue utilizada para el entrenamiento del modelo, con la cual se obtuvo posteriormente la configuración del modelo que presenta el mejor desempeño, es decir el menor error. Y la segunda parte servirá de datos de prueba o validación, con los cuales se verificará si el modelo responde de manera óptima ante la entrada de nuevos datos que nunca antes ha visto el modelo.

Para dividir los datos se utilizó la función de Train – Test Split de la librería Scikit-Learn de Python, con la cual las muestras fueron divididas aleatoriamente en datos de entramiento y datos de prueba, siendo el 75% y 25% respectivamente.

3.3.4 Partial Least Squares

3.3.4.1 Cálculo y selección del número de componentes. Como se ha descrito previamente en el presente documento, el modelo de Partial Least Squares se basa en reducir el número de variables del conjunto de datos, combinándolas y generando nuevas variables latentes o componentes, los cuales son una combinación lineal de las variables originales.

Para ello es necesario optimizar el modelo con la finalidad de obtener el número de componentes que mejor se ajuste, teniendo en cuenta los siguientes parámetros:

- R2 o coeficiente de determinación: Es un indicador estadístico, cuyo valor va entre 0 y 1. Cuantifica el ajuste del modelo a la variable dependiente, es decir mide la precisión con la cual el modelo estima los valores de salida del modelo. Mientras mayor sea su valor mejor ajuste o precisión tendrá el modelo.

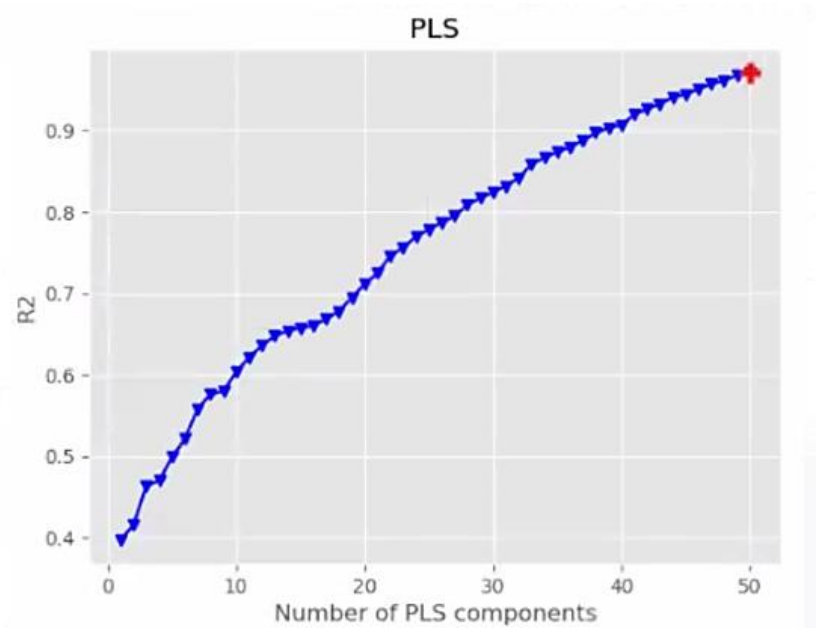
- Q2 o relevancia predictiva: Mide la capacidad del modelo de predecir valores. Si su valor es menor a cero, significa que el modelo tiene una baja precisión y que las variables independientes no pueden predecir la variable dependiente. Por el contrario, si su valor es mayor a cero, indica que el modelo tiene es relevante para predecir la variable dependiente. A medida que aumenta el valor de Q2, se puede decir que el modelo tiene una mejor predicción.

- RMSE o Root Mean Square Error o Raíz del Error Cuadrático Medio: Es un indicador del rendimiento del modelo de predicción, el cual mide la desviación de los valores predichos con respecto a los valores reales, en otras palabras, mide el error de la predicción. Valores bajos de RMSE indican una mayor precisión por parte del modelo.

Aplicando una función de iteración, en la cual se varía el número de componentes del modelo de PLS, se obtuvieron los parámetros antes indicados. A continuación, se mostrarán las gráficas obtenidas, en las cuales se puede observar cómo varían los indicadores de regresión antes mencionados en función del número de componentes. En la figura 23, se muestra como varía el coeficiente de determinación del modelo.

Figura 23

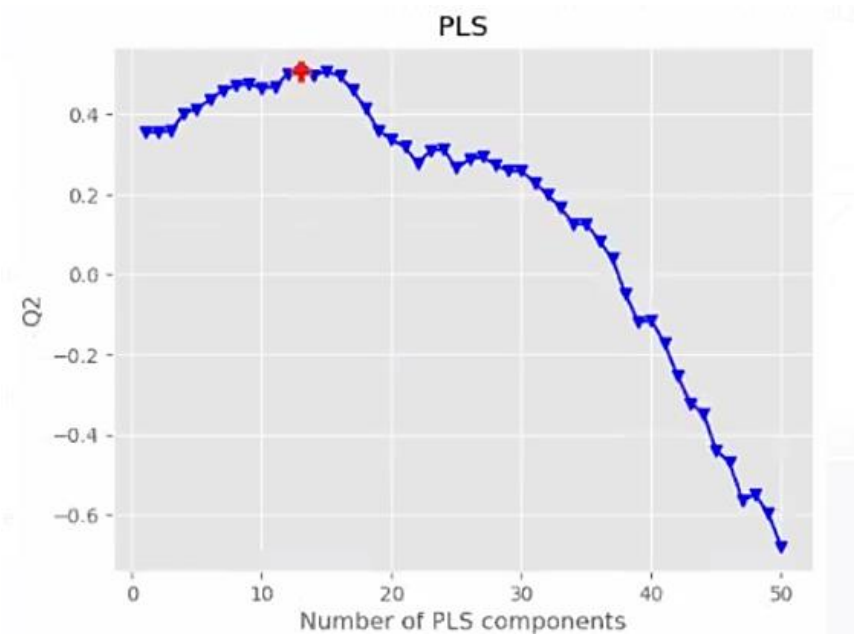
Valor de R2 del modelo en función del número de componentes



En el caso de la relevancia predictiva, en la figura 24, se puede apreciar como el indicador del Q2 alcanza su valor máximo considerando 16 componentes.

Figura 24

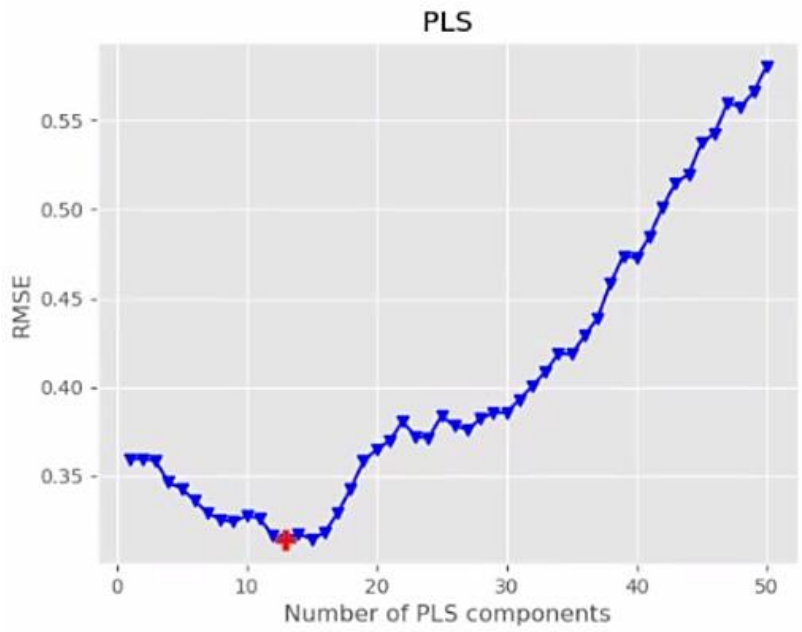
Valor de Q2 del modelo en función del número de componentes



Por último, se obtuvo la Raíz del Error Cuadrático Medio, la cual se muestra en la figura 25, donde vemos como el modelo presenta un error mínimo para quince componentes; sin embargo, lo ideal hubiera sido que dicho valor se acerque más a cero. De igual forma se puede observar cómo entre quince y veinte componentes presenta un valor bajo de RMSE.

Figura 25

Valor de RMSE del modelo en función del número de componentes



Con estos valores y analizando las gráficas se pudo determinar que el número de componentes ideal para el modelo es 19 componentes.

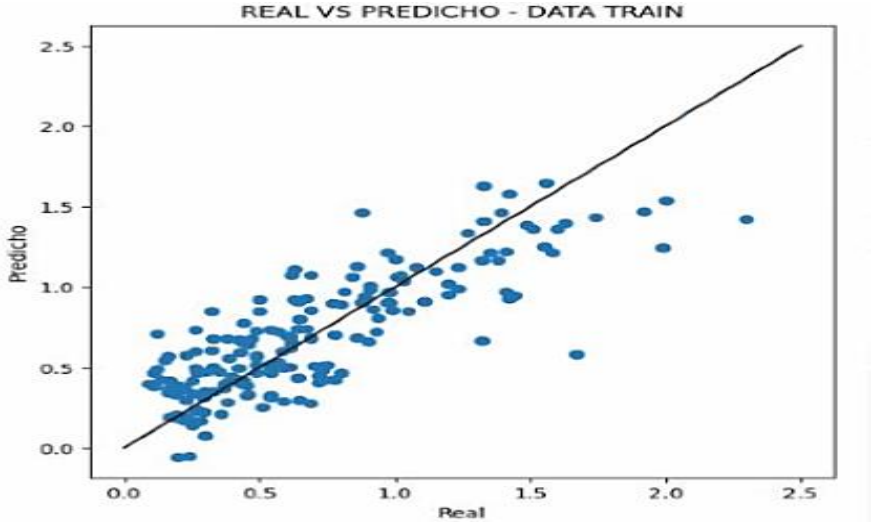
3.3.4.2 Validación del modelo. Para realizar la validación del modelo se ha tenido en cuenta el valor del R2, considerando 19 componentes, cuyos resultados fueron:

$$\mathrm{R2}_{\mathrm{entrenamiento}} = 67.78\%$$
$$\mathrm{R2}_{\mathrm{prueba}} = 67.92\%$$

En la figura 26 se, muestran los resultados del modelo con la data de entrenamiento, se puede observar que los valores predichos son muy cercanos a los valores reales, con excepción de algunas muestras, las cuales si presentan mayor desviación.

Figura 26

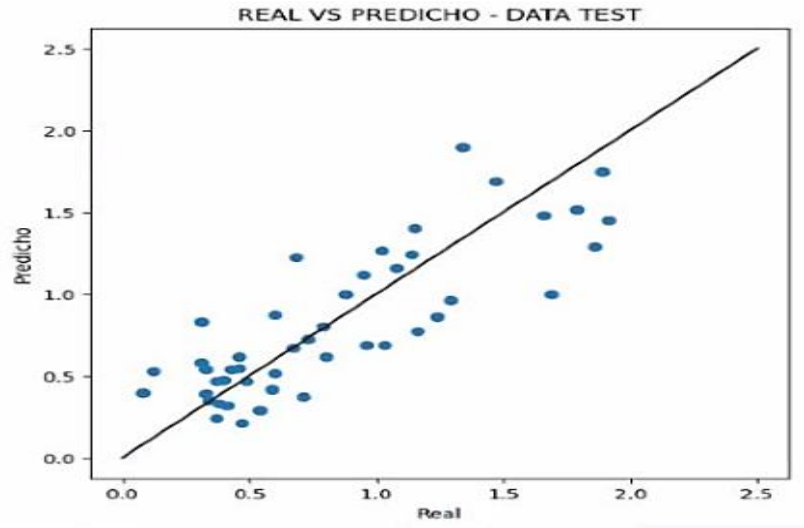
Valor predicho y valor real del conjunto de datos de entrenamiento



Así mismo, en la figura 27, se obtuvo la gráfica de los resultados del modelo de predicción, utilizando la data de prueba. Se puede observar cómo los valores son cercanos a los datos reales, con una pequeña desviación u error.

Figura 27

Valor predicho y valor real del conjunto de datos de prueba



3.3.4.3 Principales bandas de importancia. En el caso actual contamos con un conjunto de datos de entrada, el cual cada muestra es un vector de 240 parámetros independientes. Por lo que es importante identificar cuales variables realmente aportan de manera significativa al modelo. Esto con la finalidad de reducir el ruido, mejorar el tiempo de entrenamiento y reducir la cantidad las bandas necesarias para la predicción sin afectar las métricas del modelo.

Por nombrar algunos algoritmos o métodos de reducción de variables tenemos: Filtro de baja varianza, Filtro de alta correlación, Análisis de Componentes Principales, Variable Importance in Projection (VIP), entre otros.

3.3.4.4 Cálculo de variables de importancia VIP. En actual modelo se ha utilizado el método de Variable Importance in Projection o Variables de Importancia para la Proyección, el cual consiste en medir la influencia de cada variable independiente en el modelo de PLS. Los valores de VIP se calculan teniendo en cuenta los pesos de las variables latentes del modelo de PLS, cuantificando el aporte en el resultado de la predicción. Para seleccionar las variables significativas se tendrá en cuenta aquellas variables cuyo valor VIP sea igual o superior a uno, lo cual indica que tienen alta relevancia para el modelo.

En la figura 28, se muestran los valores VIP de las variables en función de la longitud de onda del modelo. Reemplazando en número de la banda a la cual corresponde dicha longitud de onda se obtienen las regiones o conjuntos de bandas que son significativas para el modelo de predicción.

Figura 28

Valor VIP de las variables de entrada

En la tabla 6, se muestran las regiones a considerar para el nuevo modelo de predicción.

Tabla 6

Zonas o regiones de bandas de importancia del modelo

Descripción	Límite inferior	Límite Superior
Zona 1	(B1) 394.35 nm	(B4) 400.61 nm
Zona 2	(B68) 534.01 nm	(B113) 665.71 nm
Zona 3	(B154) 713.72 nm	(B191) 790.95 nm
Zona 4	(B201) 811.83 nm	(B206) 822.26 nm
Zona 5	(B231) 874.45 nm	(B240) 893.23 nm

3.3.4.5 Validación de modelo reducido. Para validar el nuevo modelo con las variables independientes reducidas, se ha seguido el mismo procedimiento antes descrito. De igual forma es necesario realizar el análisis para determinar el número de componentes ideal del modelo.

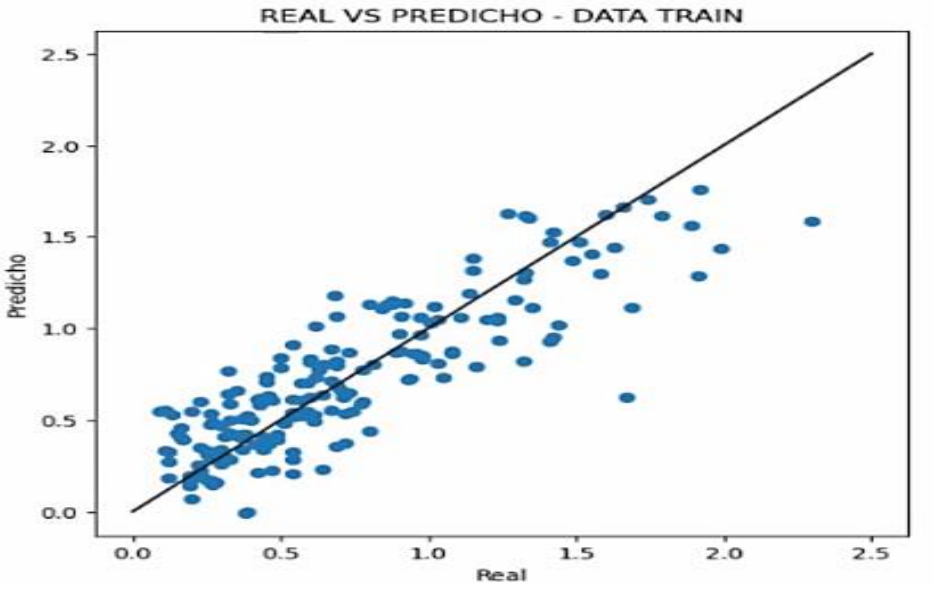
Graficando los indicadores R2, Q2 y RMSE del nuevo modelo se obtuvieron que el número de componente para el cual el modelo presenta el mejor desempeño es de 23 componentes. Y con el cual se obtuvieron los siguientes valores

$$R2^*_{\text{entrenamiento}} = 73.87\%$$
$$R2^*_{\text{prueba}} = 72.39\%$$

En la figura 29 se muestran los resultados de la estimación de los datos de entrenamiento. Se puede observar que los valores predichos se encuentran muy cercanos a los valores reales del conjunto de entrenamiento.

Figura 29

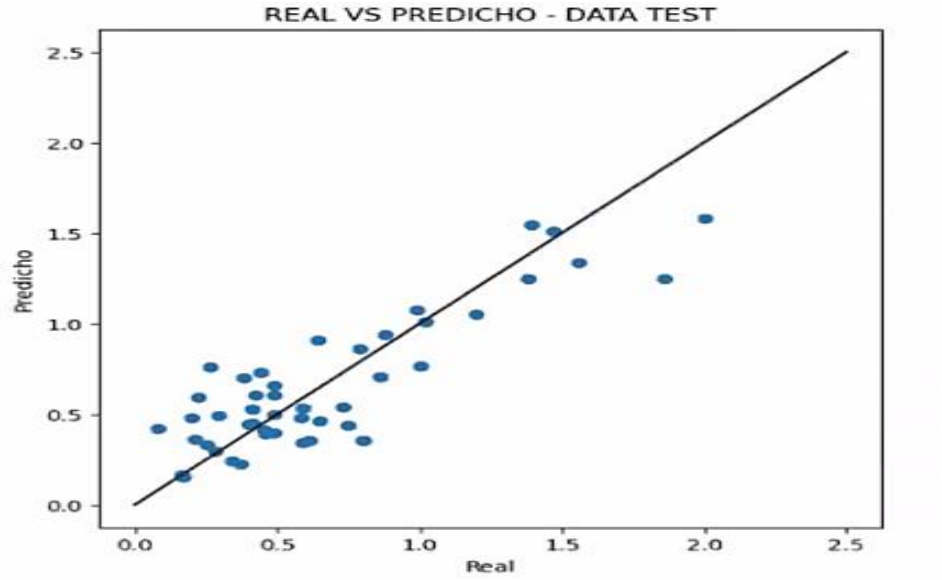
Valor predicho vs valor real del conjunto de datos de entrenamiento del modelo de estimación con variables de importancia



En la figura 30 se muestran los resultados obtenidos de la predicción de los valores de prueba. Se puede observar que presentan una desviación similar al modelo anterior.

Figura 30

Valor predicho vs valor real del conjunto de datos de prueba del modelo de estimación con variables de importancia



3.3.5 Support Vector Regression

Para el desarrollo del modelo de regresión basado en Support Vector Machines, se utilizó la función SVR de la librería Scikit – Learn, en la cual se han definido los siguientes parámetros para el entrenamiento del modelo de predicción.

- Kernel: Define el tipo de algoritmo kernel a ser utilizado para el reconocimiento de patrones en la data. Este parámetro depende de la tendencia que siguen los datos, por ejemplo, una tendencia lineal. En caso actual, al no conocer el comportamiento de los datos, se ha utilizado el valor “rbf”, el cual es para datos no lineales.

- C: Es una constante que determina el ancho del margen máximo utilizado para el entrenamiento del modelo; es decir, representa la desviación máxima permisible a la cual se encontrarán los vectores de soporte.

- Epsilon: Controla el error cometido por la función de regresión al aproximar aquellos datos ubicados fuera del margen a los vectores de soporte, en otras palabras, reduce la influencia de datos lejanos a los vectores de soporte.

Para la obtención del modelo de predicción por SVR se ha variado de forma empírica los valores de C y Epsilon. El mejor desempeño se presentó para los valores de C y Epsilon igual a 0.8 y 0.1 respectivamente, obteniendo los siguientes valores.

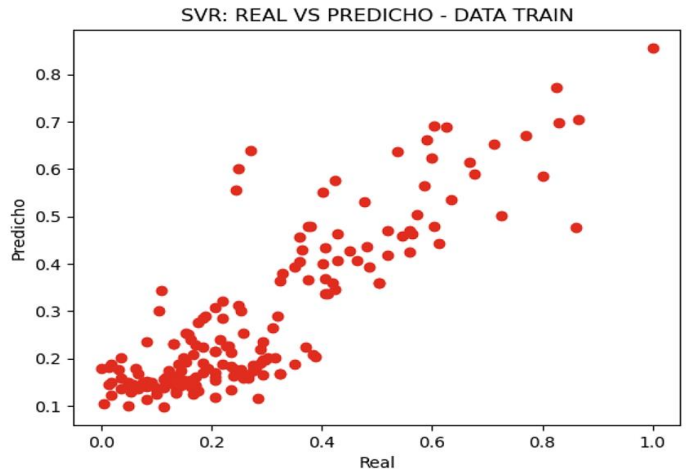
$$\mathrm{R2}_{\mathrm{entrenamiento,SVR}} = 73.34\%$$

$$R2_{\text{prueba,SVR}} = 58.61\%$$

En la figura 31 se muestran los resultados del modelo de predicción del modelo de SVR utilizando el conjunto de datos de entrenamiento, en cual se puede observar que la mayoría de los datos presentan una ligera desviación.

Figura 31

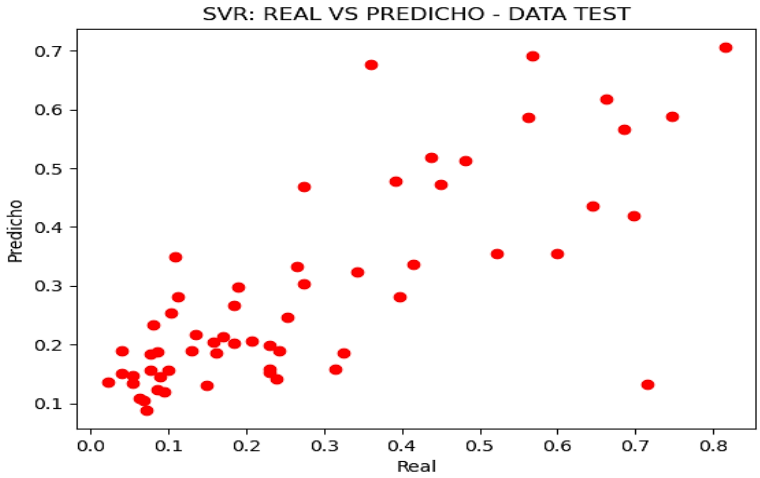
Valor predicho vs valor real del conjunto de datos de entrenamiento del modelo de estimación por SVR



En la figura 32 se muestran los resultados obtenidos de la predicción del modelo de SVR de los valores de prueba. Se puede observar que los valores presentan una mayor desviación.

Figura 32

Valor predicho vs valor real del conjunto de datos de prueba del modelo de estimación por SVR



3.4 Discusión de resultados

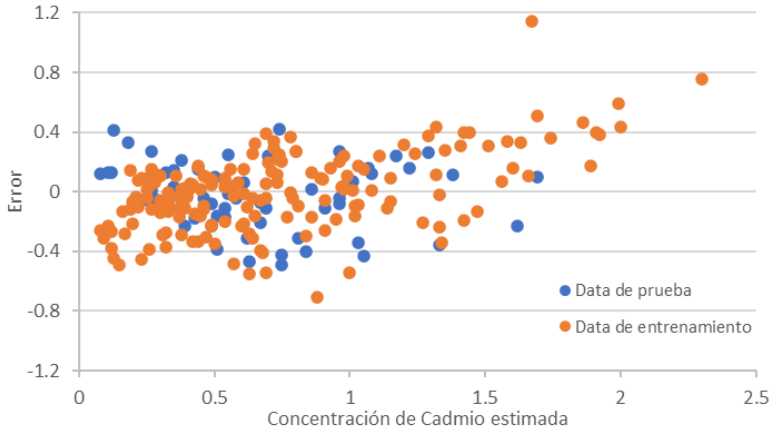
En esta sección se presentan los resultados obtenidos de aplicar los algoritmos de PLS y SVR para obtener modelos de predicción. En el caso del modelo de PLS se tienen los resultados tanto del modelo inicial, el cual cuenta con todas las bandas como parámetros de

entrada, como del modelo reducido dimensionalmente, el cual utiliza solo las bandas de importancia antes definidas.

En la figura 33 se muestra una gráfica del error cometido por el modelo de predicción para cada uno los valores de cadmio predichos.

Figura 33

Gráfica del error en la estimación realizada por el modelo de predicción inicial

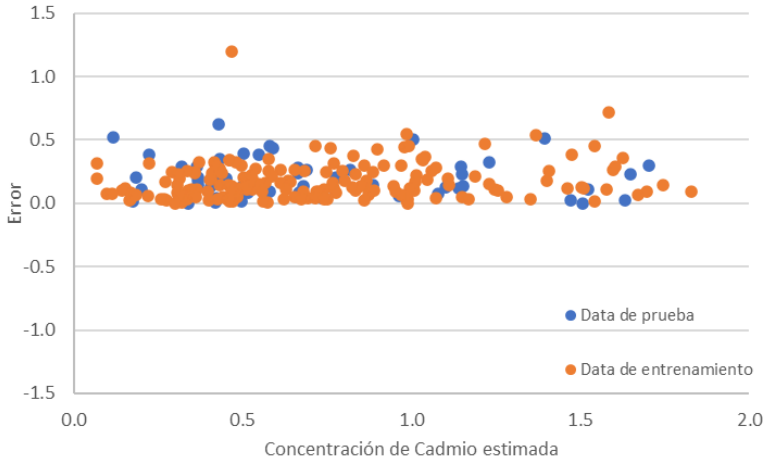


Se puede observar como la mayoría de los datos se encuentran entre el rango de error de ± 0.4. Además, se puede evidenciar que el modelo presenta mayor error para valores de cadmio superiores a una ppm.

De igual forma, en la figura 34 se muestra una gráfica del error cometido por el modelo de predicción de PLS utilizando solo las bandas de importancia como variables de entrada.

Figura 34

Gráfica del error en la estimación realizada por el modelo de predicción reducido dimensionalmente



Se puede evidenciar que este modelo presenta, en su mayoría de datos, una desviación de ± 0.5, con excepción de algunos valores. Así mismo se puede observar menor dispersión en los datos mayores a una ppm de cadmio.

En la tabla 7 se muestra una recopilación de los principales indicadores de interés sobre el desempeño de los modelos desarrollados en el presente documento.

Tabla 7

Comparación de resultados de los modelos de ML obtenidos con los datos etiquetados

Modelo	R2 Entrenamiento	R2 Prueba	Desviación mínima (ppm)	Desviación máxima (ppm)	Desviación promedio (ppm)
PLS inicial	67.78%	67.92%	0.00210	1.14	0.19
PLS reducido dimensionalmente	73.87%	72.39%	0.00160	1.20	0.18
SVR	73.34%	58.61%	0.00002	0.58	0.09

Conclusiones

En el presente trabajo de investigación se demuestra que es posible obtener la estimación del contenido de Cadmio en el grano de cacao haciendo uso de un análisis no destructivo utilizando algoritmos de ML, en tiempo real.

Se encontró un modelo basado en Partial Least Square (PLS) que predice los valores del contenido de cadmio en muestras de cacao, con un error aceptable. Las entradas del modelo son las firmas espectrales obtenidas de las imágenes hiperespectrales.

A partir del modelo inicial, tomando la totalidad de las variables de entrada, es decir considerando las 240 bandas de la firma espectral, se obtuvo un índice de R2 de 67.78% y un R2 de prueba de 67.92%, lo cual indica que el modelo de predicción es válido.

Por medio del análisis del análisis de la Variable Importance in Projection o Variables de Importancia para la Proyección VIP, fue posible diseñar un modelo de predicción con un valor de R2 de 73.87% y un R2 de prueba de 72.39%, utilizando menor cantidad de datos de entrada, con esto se demuestra que es posible construir un modelo usando estas bandas de seleccionadas, mostrando este último un mejor desempeño.

A partir de los valores obtenidos con el método VIP se llegó a determinar las variables relevantes para el modelo, las cuales fueron seleccionadas cuyos valores sean mayor o igual a uno.

Al comparar los valores predichos obtenidos por medio del algoritmo PLS con un algoritmo de SVR se observa que ambos algoritmos presentan un buen desempeño para el conjunto de datos de prueba; sin embargo, al analizar los resultados, se puede evidenciar que el modelo de SVR tiene un menor error en la predicción de la concentración de cadmio de las muestras.

UNIVERSITAS STUDIORUM PIURENSIS

Referencias bibliográficas

Abdi, H., & Williams, L. (2010). Principal Component Analysis. *WIREs Computational Statistics, vol 2*, 433-459. doi:https://doi.org/10.1002/wics.101

Aguilar, H. (2016). *Manual para la evaluación de la calidad del grano de cacao.* La Lima, Honduras: Editorial FHIA.

AprendeIA con Lidgi Gonzales. (8 de Abril de 2018). *Aprendizaje Supervisado: Support Vector Machine [Archivo de Video].* Obtenido de https://www.youtube.com/watch?v=_lXhf_uoZGQ

AprendeIA con Lidgi Gonzales. (7 de Abril de 2018). *Aprendizaje Supervisado: Support Vector Regression [Archivo de Video]*. Obtenido de https://www.youtube.com/watch?v=wvLIcYk2kgQ

Arana, C. (2021). *Modelos de Aprendizaje Automático Mediante Árboles de Decisión. Universidad del Centro de Estudios Macroeconómicos de Argentina (UCEMA), Buenos Aires.* Obtenido de https://www.econstor.eu/bitstream/10419/238403/1/778.pdf

Arango, M., Gallo, C., Gomes, J., & Martinez, M. (2020). El análisis de calidad de semillas en un nuevo escenario tecnológico. . *EEA Oliveros, INTA*.

Barrientos, R., Cruz, N., Acosta, H., Rabatte, I., Gogeascoechea, M., Pavón, P., & Blázquez, S. (2009). *Árboles de decisión como herramienta en el diagnóstico médico.* Obtenido de http://www.soporte.uv.mx/rm/num_anteriores/revmedica_vol9_num2/articulos/arboles.pdf

Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., . . . Jones, Z. (2016). Machine Learning in R. *Journal of Machine Learning Research 17*, 1-5.

Camacho, C., López, A., & Arias, M. (2006). *Regresión Lineal Simple.* Obtenido de https://personal.us.es/vararey/regresion-simple.pdf

Cámara, F., & Borysov, S. (2019). *Machine Learning Fundamentals.* Obtenido de https://doi.org/10.1016/B978-0-12-812970-8.00002-6

Carlos, A., Marcos, E., De La Cruz, C., & Rincón, C. (2003). *Partial Least Squares (PLS) regression and its application to coal analysis.* Obtenido de http://ve.scielo.org/scielo.php?script=sci_arttext&pid=S0254-07702003000300006#e1

Cherre Pupuche, C. (2019). *Medicion de parametros de calidad de harina de pescado usando imagenes hiperespectrales e inteligencia artificial.* Universidad de Piura.

Food and Agriculture Organization of the Unites Nations (FAO). (2019). *CODEX ALIMENTARIUS International Food Standards. Estándares Generales para Contaminantes y Toxinas en Alimentos y Comestibles CXS 193-1995.* Obtenido de https://www.fao.org/fao-who-codexalimentarius/sh-proxy/en/?lnk=1&url=https%253A%252F%252Fworkspace.fao.org%252Fsites%252Fcodex%252FStandards%252FCXS%2B193-1995%252FCXS_193e.pdf

Gonzáles, R., & Woods, R. (2013). Digital Image Processing.

Hearty, J., Raschka, S., & Julian, D. (2016). *Python: Deeper Insights into Machine Learning.* Packt Publishing.

Instituto Nacional de Estadística e Informática (INEI). (2022). *Compendio Estadístico Perú - Agrario.* Obtenido de https://www.inei.gob.pe/media/MenuRecursivo/publicaciones_digitales/Est/Lib1872/cap13/cap13.pdf

Liu, L., & Ngadi, M. (2010). Hyperespectral Image Processing Techniques. *Chapter 4*.

Meyer, D. (2015). *Support Vector Machines: The Interface to libsvm in package e1071.* Obtenido de https://mran.revolutionanalytics.com/snapshot/2016-03-14/web/packages/e1071/vignettes/svmdoc.pdf

Ministerio de Desarrollo Agrario y Riego (MIDAGRI). (2022). *Observatorio de Commodities - Cacao.* Obtenido de https://cdn.www.gob.pe/uploads/document/file/3561419/Commodities%20Cacao%3A%20ene-mar%202022.pdf

Moscol, I., Peltroche, G., & Ruesta, V. (2021). *Predicción de parámetros de calidad de la harina de pescado utilizando Imágenes Hiperespectrales y Redes Neuronales Artificiales.*

Mundaca Vidarte, G. (2016). *Analisis de la calidad del grano de cacao mediante imagenes hiperespectrales usando tecnicas de vision artificial.* Universidad de Piura.

Negrete, A. (2021). *Redes Neuronales.* Obtenido de http://ofeliayorquesta.com/articulos/Redes_Neuronales_001.pdf

Neyra Hau Yon, J. (2021). *Determinación en tiempo real de presencia de cadmio en cultivo de cacao aplicando Machine Learning.* Universidad de Piura.

Ngadi, M., & Liu, L. (2010). *Hyperespectral Image Processing Techniques.* Quebec: McGill University.

Okwuashi, O., & Ndehedehe, C. (2017). Tide modelling using support vector machine regression Vol. 62. *Journal of Spatial Science*, 29-46.

Paluszek, M., & Thomas, S. (2017). *MATLAB Machine Learning.* Apress. Obtenido de https://books.google.com.pe/books?hl=es&lr=&id=3kXODQAAQBAJ&oi=fnd&pg=PR6&dq=matlab+machine+learning&ots=ZMNwTG58rP&sig=X3_N-ggA-IcF6RyKu70ONQ-hGY4&redir_esc=y#v=onepage&q&f=false

Peltroche Saavedra, G. (2022). *Diseño e implementación de algoritmos inteligentes basados en aprendizaje de máquina para la detección de cadmio en granos de cacao mediante imágenes hiperespectrales.* Universidad de Piura.

Pisner, D., & Schnyer, D. (2020). *Machine Learning; Methods and Applications to Brain Disorders. Chapter 6: Support Vector Machine.* Obtenido de https://doi.org/10.1016/B978-0-12-815739-8.00006-7

Raschka, S., & Mirjalili, V. (2016). *Python Machine Learning.* Packt Publising.

Ringnér, M. (2008). What is Principal Component Analysis? *Nature Biotechnology 26*, 303-304. doi:https://doi.org/10.1038/nbt0308-303

Saha, D., & Manickavasagan, A. (2021). *Machine learning techniques for analysis of hyperspectral images to determine quality of food products: A review.* Obtenido de https://doi.org/10.1016/j.crfs.2021.01.002

Shai Shalev, S., & Shai Ben, D. (2014). *Understanding Machine Learning: From Therory to Algorithms.* Obtenido de https://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/understanding-machine-learning-theory-algorithms.pdf

Sun, D. W. (2010). *Hyperspectral Imaging for Food Quality Analysis and Control.* Academic Press.

Suthaharan, S. (2016). *Machine Learning Models and Algorithms for Big Data Classification. Chapter 10: Decision Tree Learning.* doi:https://doi.org/10.1007/978-1-4899-7641-3_10

Unión Europea. (2021). *Reglamento (UE) 2021/1323 de la Comisión de 10 de agosto de 2021 que modifica el Reglamento (CE) No. 1881/2006 por lo que respecta al contenido máximo de cadmio en determinados productos alimentcios.* Obtenido de https://eur-lex.europa.eu/legal-content/EN/TXT/

Ustav, G., Sildomar, M., & Eli, S. (2019). *Machine learning based hyperspectral image analysis: a survey.* Obtenido de https://arxiv.org/pdf/1802.08701.pdf

Valdéz, D. (2010). *Regresión por Mínimos Cuadrados Parciales. Revista Varianza.* Obtenido de http://www.revistasbolivianas.ciencia.bo/scielo.php?script=sci_arttext&pid=S9876-67892010000100005&lng=en&nrm=iso&tlng=es