UNIVERSIDAD DE PIURA

FACULTAD DE INGENIERÍA

Determinación en tiempo real de presencia de cadmio en cultivo de cacao aplicando Machine Learning

Tesis para optar el Grado de
Máster en Ingeniería Mecánico -Eléctrica con mención en Sistemas Energéticos y Mantenimiento

Jorge Luis Neyra Hau Yon

Asesor(es):
Ph D. Ing. William Ipanaqué Alama; Mgtr. Ing. Juan Carlos Soto Bohórquez

Piura, abril de 2021

UNIVERSITAS STUDIORUM PIURENSIS

A mis adoradas Josselyn, Romina y Doménica, por su paciencia, comprensión y motivación diaria en cada meta propuesta. A mis padres y hermano, por confiar siempre en mí y aconsejarme siempre el terminar lo que se empieza.

UNIVERSITAS STUDIORUM PIURENSIS

Resumen

El objetivo de este proyecto de tesis está enfocado en Desarrollar métodos de Predicción basados en *Machine Learning*, para estimar el Contenido de Cadmio en granos de Cacao, a partir de sus firmas hiperespectrales.

La presenta tesis se desarrolla en cuatro capítulos. El primero permitirá conocer la importancia del proceso productivo del Cacao, cómo afecta la presencia del Cadmio en este producto, y las consecuencias en la salud del consumidor final en caso de no regularlo. En el segundo capítulo se estudia la metodología principal de esta tesis, qué es *Machine Learning*, y los algoritmos basados en esta metodología para la predicción del contenido de Cadmio en los granos de Cacao.

Para el tercer capítulo se desarrolla el marco teórico sobre el cual se basa la obtención de firmas hiperespectrales del Cacao, y cómo pueden éstas, ayudar en la obtención de características propias de cada producto o elemento. En el cuarto y último capítulo, se mostrarán los resultados de los modelos de predicción con base en *Machine Learning* para este proyecto.

UNIVERSITAS STUDIORUM PIURENSIS

Prefacio

En el Perú, la agroindustria ha crecido progresivamente a través de los años, generando un incremento en el nivel económico del país. Dentro de las diversas actividades agroindustriales, el procesamiento de Cacao ha tenido una situación particular e importante, ya que, debido a la excelente calidad del Cacao peruano, éste ha sido reconocido como el mejor del mundo en diversas ediciones de competencias internacionales. Este reconocimiento conlleva a la responsabilidad de mantener la calidad del Cacao dentro de las expectativas del consumidor final.

El mercado europeo es el principal consumidor de nuestro Cacao prime, y ya en el año 2014, publicó un reglamento para regular el contenido de algunos metales pesados en los productos alimentarios de consumo humano, entre ellos, el Cadmio presente en el Cacao. Este reglamento está vigente desde el año 2019 y esto genera preocupación en los productores de Cacao, que, debido a situaciones de suelo, ambiente, agua y otros factores que influyen en el contenido de Cadmio en el cacao, muchas veces exceden estos parámetros permisibles.

En la búsqueda de mejorar los procesos y reducir la contaminación del producto final, los productores han buscado intervenir en las diversas etapas de la producción el Cacao, controlando el nivel de Cadmio en estas etapas. Sin embargo, controlar el nivel de Cadmio no es tan ágil para anticipar su presencia. Es por ello, que, se busca un método para determinar en tiempo real, el contenido de cadmio de una muestra de Cacao.

Para lograrlo, se ha previsto utilizar las firmas espectrales del Cacao como producto en evaluación, y como futuros datos de entrada para determinación del contenido de Cadmio. Además, se plantean metodologías de predicción basadas en *Machine Learning*, para que nuestro sistema se enriquezca con la experiencia de predicción, y podamos definir una propuesta adecuada para el éxito de este proyecto.

Esta tesis ha sido financiada por el FONDECYT y BM, en el proyecto: “Fortalecimiento de capacidad de investigación en medición, en tiempo real de físico químicas de productos de agroindustria y elaboración de alimentos usando imágenes hiperespectrales”. Código:

005-2018-FONDECYT-BM. Quisiera además expresar mi agradecimiento a mis asesores Ph.D. Ing. William Ipanaqué Alama y Mgtr. Juan Soto Bohórquez, por su apoyo, orientación, y guía que contribuyeron para sacar adelante esta tesis. A mi familia por su comprensión, apoyo incondicional y el ánimo que constantemente me brindan.

Tabla de contenido

UNIVERSITAS STUDIORUM PIURENSIS

Lista de tablas

UNIVERSITAS STUDIORUM PIURENSIS

Lista de figuras

Introducción

El *Machine Learning* o Aprendizaje Automático es una tecnología de la Inteligencia Artificial cuyo objetivo es que las máquinas aprendan a partir de situaciones ya acontecidas o experiencia previa. Muchas aplicaciones en la actualidad como YouTube, Facebook, Google, Netflix, entre otras muchas, utilizan *Machine Learning* para 'aprender' de los usuarios y mostrar la publicidad o sugerencias que a cada 'cliente' podría interesarle.

Las aplicaciones del *Machine Learning* están ampliándose hacia diversos campos como: medicina, marketing, seguridad nacional, prevención de fraudes, agronomía, industrias alimentarias, entre otras. Es por ello que, la aplicación de esta metodología en procesos industriales que generan un fuerte impacto económico a nivel nacional, es una prioridad.

Por otro lado, la tecnología de la Visión Hiperespectral, que inició con aplicaciones de astronomía y teledetección, y que ahora es aplicada a muchos otros campos como: medicina, bioquímica, industria alimentaria se ha desarrollado de tal manera, que permite determinar diversas características fundamentales de cada producto o elemento.

El objetivo de este proyecto es incorporar la estrategia de *Machine Learning* y la tecnología de la Visión Hiperespectral para el diseño de estrategias para la detección en tiempo real del Contenido de Cadmio de muestras de Cacao, con el objetivo de poder intervenir en el proceso productivo del Cacao, y así, reducir el riesgo de contaminación con Cadmio hen el producto final.

Para el desarrollo de este proyecto se ha empleado la Cámara Hiperespectral Resonon, dentro la infraestructura del Departamento de Automática y Control de la Universidad de Piura, y se ha contado con las muestras de Cacao tomadas de algunas locaciones de la región, con la finalidad de obtener las firmas hiperespectrales de la muestra, como para hacer la modelación de la predicción.

Durante el desarrollo de esta tesis, se mostrarán también los resultados obtenidos con diversos algoritmos de predicción con *Machine Learning*, que nos permitirá comparar el mejor modelo de predicción para nuestro caso particular y específico. Finalmente, con un monitoreo

en tiempo real, esto ayudará a tomar las medidas correctivas y así, mitigar la contaminación por metales pesados como el cadmio.

Capítulo 1

El problema de la presencia de cadmio en el cultivo del cacao

1.1 Prefacio

En diversas regiones de la Amazonía, el cultivo del cacao se ha convertido en una actividad económica principal, debido a la demanda en mercados extranjeros.

Esto ha permitido que dicha actividad agrícola sea llevada a gran escala, y, por ende, industrializada, por lo que sus regímenes de cuidado y protección del suelo han mejorado en la calidad de su administración.

Esto no ha sido impedimento para que las plantaciones de cacao se vean afectadas por contaminación de diversos agentes, entre ellos, metales pesados como el cadmio y el plomo.

Las opiniones de diversos expertos son contradichas respecto a esta contaminación, pero sin entrar en detalles del debate, ha quedado demostrado por la Agencia para Sustancias Tóxicas y Registro de Enfermedades (ATSDR), que los índices elevados de concentración de cadmio incrementarían el riesgo de sufrir enfermedades relacionadas a los riñones, sistemas ósea y respiratorio.

Es por ello, que actualmente, los controles de estos niveles de contaminación son más rigurosos, que a la larga genera que el producto sea trabajado con mayores estándares de calidad.

1.2 El cacao

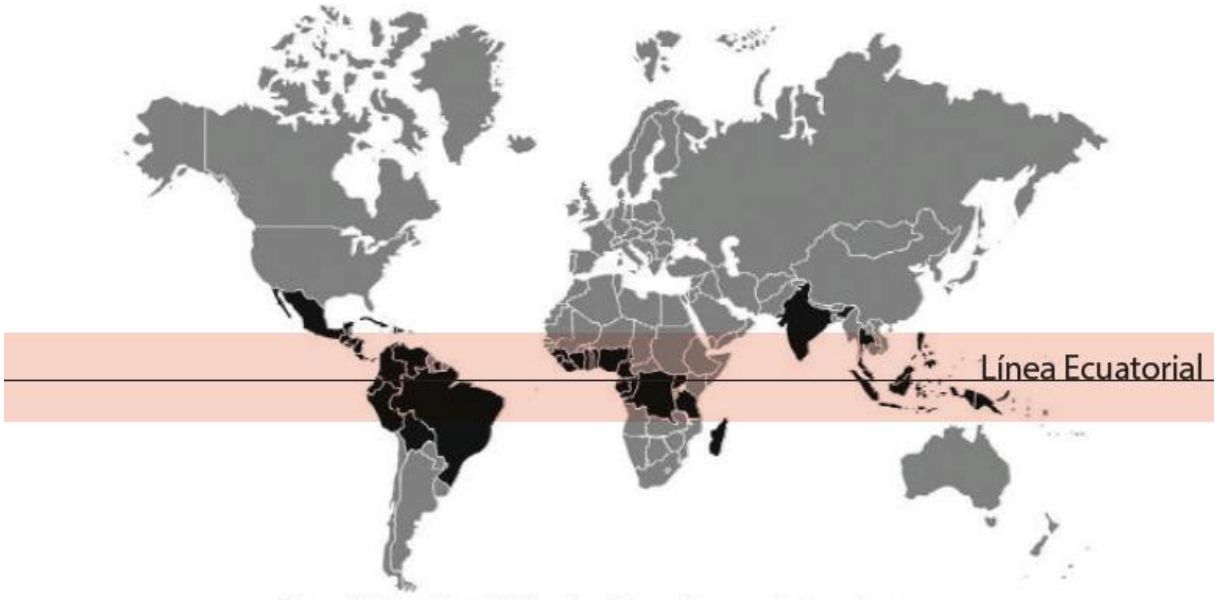
Theobroma Cacao L. es el nombre científico que recibe el árbol de cacao, que es una especie de origen tropical, conocido principalmente porque de su fruto, se extraen las semillas que son el ingrediente base para la elaboración del chocolate. Theobroma es una palabra griega que significa: “alimento de los dioses”.

El cacao es una planta de origen en la cuenca del Amazonas, y su uso tiene evidencias de hace 5,500 años en lo que actualmente se ubica la Amazonía ecuatoriana.

El árbol de cacao, o cacaotero, se encuentra en estado natural en selvas húmedas, y requiere para un buen crecimiento, de un clima cálido y húmedo (idealmente entre 20°C y 30°C de temperatura media, y con una mínima de 16°C), y de una altura que no supere los 1250m. En la figura 1 puede observarse entre líneas rojas, las franjas en las que el cacao podría crecer (18°N y 15°S).

Suele crecer hasta una altura de entre 5m a 10m. El fruto es de color rojo o amarillo purpúreo con un peso promedio aproximado de 450gr cuando éste ha madurado y con unas dimensiones aproximadas que van desde los 15 hasta los 30 centímetros de largo, y desde los 7 hasta los 10 centímetros de ancho. El árbol tarda 5 años en madurar y su vida es en promedio de 30 años.

Figura 1. Franjas de latitudes posibles para cultivo de cacao.



Fuente. Romero, 2016.

1.2.1 Breve reseña del cacao y el chocolate

La historia del cacao, pese a tener su origen en la Amazonía, fue extendida a América Central y desarrollada en esta zona hace más de 2,500 años.

La palabra cacao, posiblemente proviene de la lengua indígena Maya: “CAC” que significa “Rojo”, como el color de la mazorca y “CAU” que significa fuerza, fuego.

Los antepasados mayas creían que este árbol era un regalo del dios Quetzalcoatl para los hombres, por lo que años más tarde, fue bautizado con el nombre científico Theobroma Cacao, que significa en griego: “Alimento de los dioses”.

El cacao en la sociedad azteca fue utilizado como un alimento de gran importancia, y también como moneda de cambio. El cacao en su forma líquida era mezclado con especias, y obtenían una bebida energética, espesa y espumosa a la que llamaban “tchocolatl”.

Figura 2. Representación gráfica. Origen del chocolate en cultura Azteca.

Fuente: Nestlé, 2017.

Se cuenta en la historia, que cuando Hernán Cortés desembarcó en México en 1591, los aztecas liderados por el emperador Moctezuma II, creyeron que era la reencarnación del dios Quetzacoatl, y fue agasajado con “tchocolatl”. Después de la colonización, las semillas de cacao fueron aún utilizadas como moneda de cambio.

Los europeos añadieron azúcar como endulzante, y canela como aromatizante. El chocolate para esta época fue difundido en su forma compacta.

La primera producción mecánica de chocolate data de 1777, en Barcelona, marcando la “revolución industrial” para este producto. Para el siglo XIX el chocolate se hace más popular gracias al crecimiento de la Industria Chocolatera. En 1819, el suizo Francois- Louis Cailler crea la primera fábrica de chocolate, en la que este producto era una mezcla de granos de cacao molidos con azúcar añadido. Producto que solo podían disfrutar la clase pudiente de la época.

En 1830, el suizo Charles-Amédée Kohler abre también su fábrica, en la que inventa el chocolate con avellanas. Para 1875, otro chocolatero suizo Daniel Peter, tras varios años de experimentos, inventa el chocolate con leche, luego de mezclar el chocolate con harina lacteada, inventada por Henri Nestlé en 1867.

En la actualidad, el cacao y sus productos derivados son producidos y consumidos a nivel mundial, siendo un producto de alto valor nutritivo.

1.2.2 Variedad del cacao

Existen tres variedades principales de cacao: Criollo, Forastero y Trinitario.

- **Cacao criollo:** Es un cacao reconocido como de gran calidad, utilizado para la producción de los chocolates más finos. Con un árbol de bajo rendimiento y muy frágil (sensible a enfermedades), tiene un grano aromático de cáscara suave y fina y con bajo contenido de taninos (sustancia que debe ser eliminadas para evitar el sabor astringente en el chocolate).

Este tipo de cacao representa aproximadamente el 10% de la producción mundial y es cultivado principalmente en la zona norte de Sudamérica (Perú, Venezuela, Colombia, Ecuador), Centroamérica (Nicaragua, Guatemala, El Salvadora, Jamaica, México, República Dominicana), y algunas regiones asiáticas.

Figura 3. Cacao Criollo.

Fuente: Nestlé, 2017.

- **Cacao forastero o campesino.** Es un cacao con una cáscara más gruesa y resistente, menos aromático que el criollo y con mayor presencia de taninos. Sus imperfecciones suelen neutralizarse con un tostado más intenso, de ahí procede su sabor y aroma.

Es un cacao originario en la alta Amazonía o cuenca amazónica, y es el más cultivado en el oeste de África y Brasil y a su vez, el de mayor producción mundial (aproximadamente 70%).

Mientras que este cacao forastero otorga cuerpo y amplitud al chocolate, la acidez y equilibrio, es brindada por el cacao criollo.

Figura 4. Cacao forastero.

Fuente: Nestlé, 2017.

- **Cacao trinitario.** Es un tipo de cacao híbrido, obtenido por el cruce entre los tipos Criollo y Forastero, con un nivel de calidad similar a este último. Es originario de Trinidad, y tiene las características de calidad del cacao criollo, y la robustez del forastero.

Este tipo de cacao representa aproximadamente el 20% de la producción mundial.

Figura 5. Cacao trinitario.

Fuente: Nestlé, 2017.

1.2.3 Partes del árbol de cacao

Las partes principales del árbol de cacao se muestran son principalmente:

- **Raíz.** La raíz principal del cacao crece aproximadamente 2 metros hacia abajo, y las raíces secundarias hacia los lados en los primeros 30 centímetros del suelo. La raíz tiene la función de fijar la planta al suelo, y de absorber y conducir el agua y nutrientes necesarios.

- **Tronco.** El tronco del árbol de cacao es recto, y dependiendo de la poda, adquiere su forma. La corteza es oscura, con ramas café y vellosas. Sobre el tronco crece el fruto del cacao.

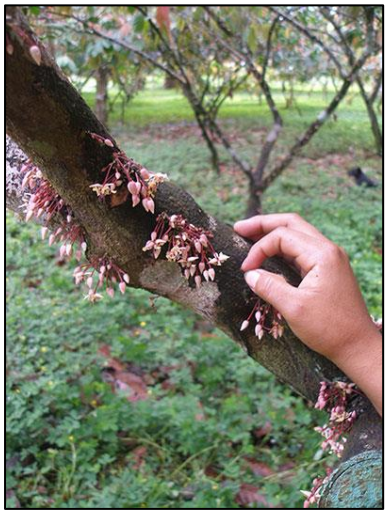
- **Hojas.** Las hojas del cacao suelen ser angostamente ovadas, alargadas, normalmente con colores que van desde el café claro, morado, rojizo, hasta verde claro.

Figura 6. Hojas de cacao.

Fuente. Lutheran Wolrld Relief, 2017.

Flor. La flor del cacao es pequeña, estrellada con cinco pétalos, y brotan sobre el tronco y ramas. Son sostenidas por un pedúnculo. Dependiendo de la variedad del cacao, las flores tienen distinto color, pueden ser blancas, rosadas, púrpura, o similares.

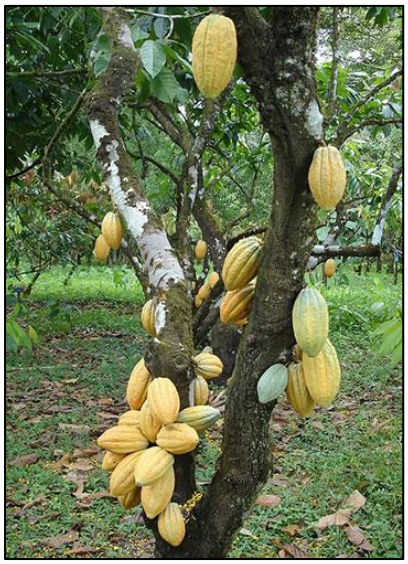
Figura 7. Flor del cacao.



Fuente. Lutheran World Relief, 2017.

- **Fruto.** El fruto o mazorca suele tener formas y tamaños distintos según sea su variedad. De largo miden entre 15 a 30 centímetros, y de ancho entre 7 y 10 centímetros. El fruto puede ser de cáscara lisa o arrugada, alargada o redonda, y de diversos colores como: rojo, amarillo, verdes, moradas, café.

Figura 8. Fruto del cacao.



Fuente: Lutheran World Relief, 2017.

- **Semilla o grano de cacao.** La mazorca de cacao contiene de 20 a 40 semillas de 2 a 3 centímetros de largo. Están cubiertas por una capa muy fina conocido como mucílago, que, según su variedad, puede tener sabor dulce o ácido.

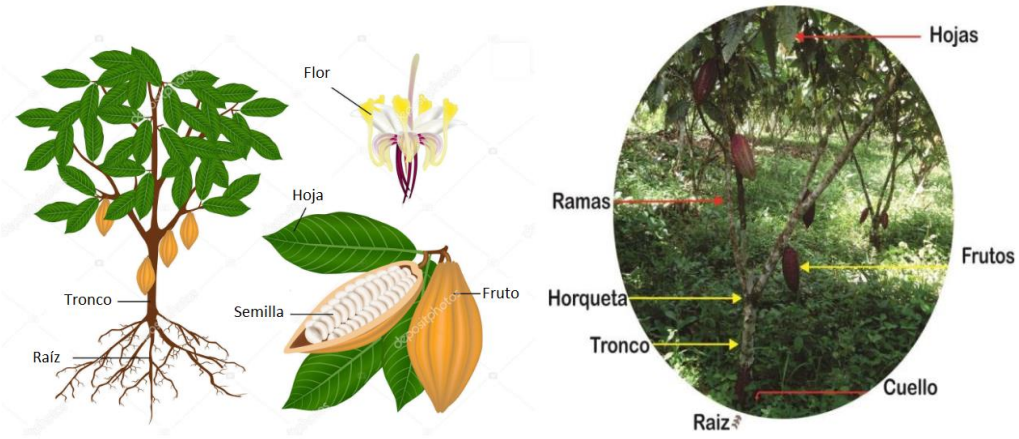
El interior de la semilla está formado por dos cotiledones de forma ovalada y aplanada.

Figura 9. Grano de cacao.

Fuente. De la Cruz M, Vargas A, 2010.

En la figura 10, puede verse un esquema de la planta de cacao y una imagen en la que se pueden visualizar las partes del árbol.

Figura 10. Partes del árbol de cacao. Izquierda en esquema y derecha en fotografía.



Fuente. Izquierda: Havryliuk N, 2018. Derecha: Cabana A, 2017.

1.2.4 Productos del fruto de cacao

Del fruto del cacao, se extraen principalmente productos como:

- **Polvo de cacao:** Producto seco y de color oscuro, que tiene el sabor característico del cacao. Es la parte del cacao resultante de reducir la manteca mediante el uso de prensas hidráulicas y disolventes, hasta lograr la textura pulverulenta.

El polvo de cacao es amargo y libre de impurezas, olores y sabores extraños. Es usado para aromatizar galletas, bebidas y postres, además de ser usado para la elaboración del chocolate. Este producto es un alimento muy calórico, altamente proteico y con bajo contenido de hidratos de carbono y grasa.

Figura 11. Cacao en Polvo.

Fuente. Pazmina C, 2019.

Del fruto del cacao, se extrae principalmente el polvo de cacao y grasa o manteca de cacao; ambos fundamentales para la producción del chocolate, que producto final para el que se destina más del 60% de la producción mundial de cacao.

- **Manteca de cacao:** Es la grasa natural comestible de la semilla de cacao, extraída mediante presión. Esta manteca conserva el aroma suave y el sabor del cacao. Por su textura, es utilizada en la industria alimentaria (chocolate), farmacéutica (medicamentos) y cosmética (productos de belleza como limpiadores, jabones, mascarillas).

Con la manteca de cacao pura, leche y azúcar, se elabora el conocido: chocolate blanco.

Figura 12. Manteca de cacao.

Fuente: Ramos E, 2019.

- **Pulpa de cacao:** Es la capa que envuelve a la semilla de cacao, antes considerada como un desecho, pero que ahora se usa como ingrediente en la gastronomía.

Figura 13. Pulpa de cacao.

Fuente. Tinoco P, 2018.

- **Cáscara del cacao:** Aprovechada para hacer infusiones, preparar mermeladas, o con fines de alimentación animal.

Figura 14. Cáscara de cacao.

Fuente. Espín M. 2020.

1.2.5 Procesamiento del cacao

El fruto del cacao es la materia prima para la elaboración de diversos productos como ya se vio en los puntos anteriores. Sin embargo, la materia prima no es tomada directamente para su aplicación, sino que debe seguir una secuencia de procesos posteriores a la cosecha.

- **Cosecha:** Es la etapa de extracción de las mazorcas de sus respectivos árboles. Los recolectores suelen guiarse por el color de la vaina y por el sonido que genera el interior de la mazorca cuando es golpeada para saber si el fruto ha madurado. La época de cosecha varía según la región en la que se cultivando, y en algunas regiones, la recolección se da durante todo el año.

- **Apertura:** En esta etapa, se abren las mazorcas cosechadas para extraer los granos de cacao, es importante cortar las mazorcas sin estropear las semillas. Por tener contacto directo con el producto, los estándares fitosanitarios deben ser lo suficientemente buenos para evitar contaminación por contacto directo, de hongos o insectos.

Los granos de cacao extraídos son apilados sobre hojas de plátano, depósitos de madera y/o plásticos, luego se envuelven completamente estos grupos, para iniciar el siguiente proceso: la fermentación.

- **Fermentación:** En esta etapa inicia el proceso biológico. Las bacterias y levaduras presentes en el ambiente se multiplican en la pulpa que rodea los granos debido a su concentración de azúcares. Ésta se descompone y forma un líquido ácido y alcohol. Esta reacción bioquímica y de oxidación permite, por un lado, que la semilla se hinche, y por otro lado, la disminución del grado de acidez y astringencia en el sabor del cacao, que definirán también el aroma del cacao.

La calidad de los granos depende de este proceso de fermentación. Si es excesivo, el cacao puede arruinarse; si es insuficiente, puede adquirir un sabor de patatas crudas. Según el grado de acidez y sabor requerido, la fermentación suele durar entre 3 y 7 días.

- **Secado:** En esta etapa, los granos son extendidos y rastrillados constantemente para desecarse. Este proceso puede realizarse de manera natural (con exposición solar), o de manera artificial (mediante el uso de secadores). Producto de este proceso, la masa de los granos disminuye aproximadamente en una cuarta parte de su peso original.

- **Selección:** La selección de los granos de cacao suele realizarse visualmente mediante una prueba de corte. Ésta revela la presencia de granos defectuosos y el grado de fermentación de los granos del lote. La normativa ISO (*International Organization for Standardization*) (1114:1977- *Cut Test*), sugiere que, por cada tonelada de cacao, debe analizarse en lo mínimo 300 gramos de cacao y 30% de los sacos de cacao para determinar el grado de fermentación.
- **Torrefacción, descascarillado y desgrasado:** Esta siguiente etapa afina más los aromas y reduce su amargura y astringencia. Los granos son separados de su cascarilla y por molturación, se obtiene una pasta de cacao con un 55% de manteca promedio. Luego, con las prensas, se extrae gran parte de la manteca de cacao, quedando como subproducto, la "torta de cacao". Esta última conserva una porción residual de manteca, que oscila entre el 10 y 20% según la aplicación destinada para cada lote. Nuevamente por molturación, a partir de la torta de cacao, se consigue el cacao magro o cacao en polvo desgrasado.

1.2.6 Países productores de cacao

El cacao es cultivado principal y mayoritariamente en África Occidental, América Central, Sudamérica y Asia.

En estas regiones se produce poco más del 90% de la producción mundial de cacao. En África, la producción principal está en los países de Camerún, Costa de Marfil, Ghana y Nigeria. En América está principalmente en Brasil, Ecuador, Perú; y en Asia se centra, principalmente, en Indonesia.

La producción de Latinoamérica es considerada como 'cacao prime' y corresponde a un porcentual ubicado entre el 70 y 100% de la exportación de: Bolivia, Colombia, Ecuador, Perú, México.

En la tabla N°1 se puede visualizar los niveles de producción y en la tabla N°2 se muestra la demanda de cacao en grano para molienda.

Tabla 1. *Producción de Cacao en miles de toneladas.*

		2014/2015	2015/2016	2016/2017	2017/2018*	2018/2019**
	Total Mundo	**4 252**	**3 994**	**4 731**	**4 652**	**4 835**
1	Costa de Marfil	1 796	1 581	2 020	1 964	2 200
2	Ghana	740	778	970	905	870
3	Indonesia	325	320	270	240	220
4	Brasil	230	141	174	204	195
5	Nigeria	195	200	245	250	245
6	Ecuador	261	232	290	287	298
7	Camerún	232	211	246	250	260
8	**Peru**	**92**	**105**	**115**	**134**	**120**
9	República Dominicana	82	80	57	85	75
10	Colombia	51	53	55	55	60
	Subtotal	**4 004**	**3 700**	**4 441**	**4 373**	**4 543**
	Otros	248	294	290	280	292

Fuente: Huamán O, 2019.

Tabla 2. *Demanda de Cacao en miles de toneladas.*

		2014/2015	2015/2016	2016/2017	2017/2018*	2018/2019**
	Mundo	**4 152**	**3 127**	**4 397**	**3 596**	**4 750**
1	Países Bajos	501	534	565	590	605
2	Costa de Marfil	558	492	577	559	585
3	Indonesia	335	382	455	483	490
4	Alemania	415	430	410	448	450
5	Estados Unidos	400	398	390	385	400
6	Ghama	234	202	250	310	300
7	Malasia	195	194	2 016	236	260
8	Brasil	224	225	227	231	230
9	Francia	130	138	143	152	158
10	España	99	112	115	100	102
	SUBTOTAL	**3 091**	**3 106**	**5 149**	**3 494**	**3 580**
	Otros	1 061	1 021	-752	1 102	1 170

Fuente: Huamán O, 2019.

1.2.7 Cacao peruano

Perú es considerado uno de los principales productores de cacao fino y aromático, y es el segundo productor de cacao orgánico a nivel mundial. La producción de cacao en grano ha tenido un crecimiento sostenido con tasa de 15.6% promedio anual en los últimos 10 años.

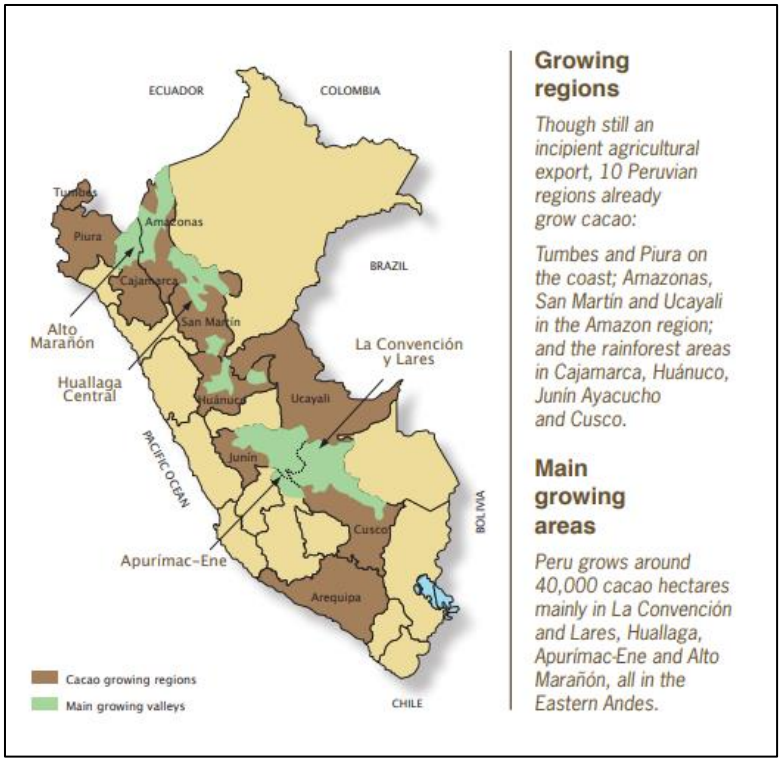
Las principales regiones productoras de cacao son: Piura, Pasco, Ucayali y Cajamarca. Además de San Martín, Ayacucho, Cusco y Tumbes.

En Perú crecen variedades de cacao trinitario, amazónico y criollo, siendo este último el de mayor calidad debido a su aroma, sabor y contenido de grasa.

Los mercados de exportación de cacao en grano en 2018 fueron principalmente: Indonesia, Holanda, Italia, Bélgica y España.

El cacao peruano ha sido considerado como un tipo de cacao 'prime' por su elevada calidad, y ha sido merecedor de diversos premios de reconocimiento mundial. En el 2014, obtuvo el Cacao International Award en el Salón Du Chocolat, en París. En 2017, también fue reconocido como el mejor del mundo en el Salón del Chocolate, en Londres, en el International Chocolate Awards. Continuamente ha recibido premios por su calidad, por lo que en Perú inclusive ha sido merecedor de un reconocimiento nacional al ser declarado Patrimonio Nacional.

Figura 15. Principales regiones de producción de Cacao.



Fuente: Romero C, 2016.

Los principales países importadores de este producto son: Países Bajos y Bélgica, con un consumo del 19.3% y 18.4% respectivamente de nuestra exportación en grano.

Si bien el consumo per cápita es de 0.5 kilogramos por año, en países europeos como Suiza y Alemania, el consumo llega a 9 kilogramos por año y 7.9 kilogramos por año respectivamente.

Entre los beneficios del cacao se encuentra el contenido de fitoesteroles, que bloquean la absorción del colesterol alimentario, la reducción de la posibilidad de hipertensión y control de arterioesclerosis por los efectos vasodilatadores. Además de contener polifenoles, que son antioxidantes que previenen procesos ateroscleróticos.

Podemos visualizar en la tabla a continuación, los valores nutricionales del cacao.

Tabla 3. *Valor nutricional del cacao y algunos chocolates.*

Alimento	Energía (kcal)	Agua (g)	Hidratos de carbono (g)	Azúcar (g)	Proteínas (g)	Grasas (g)	Ácidos grasos saturados (g)	Fibra (g)	Fósforo (mg)
Semilla de cacao (tras el tostado)	(> 550 kcal)	2	7 - 7.5	1 - 1.5	11.8	52 - 56	(aprox. 60% del total)	9.3	--
Cacao en polvo desgrasado	255	--	16	3	23	11	6.5	23	600
Cacao en polvo azucarado	390	1.5	81.02	(la mayoría)	5.88	4	2.5	7 - 8	315
Chocolate	518	7	56.4	(la mayoría)	7.8	30.6	18.2	0	287
Chocolate negro	532	5	63	(la mayoría)	2	30	16.8	--	147
Chocolate negro con almendras	535	3.8	41.5	(la mayoría)	8.2	37.4	16.3	9.1	219
Chocolate blanco	547	1.5	58.8	(la mayoría)	8	30.9	18.2	0.8	230
Chocolate con leche	538	2.9	54.1	(la mayoría)	9.19	31.5	18.96	0.8	261
Chocolate con leche y almendras	549	2.7	49.8	(la mayoría)	8.6	35	16.6	3.9	246
Chocolate con leche con polialcoholes	478	--	51.5	8.7	8.4	33.1	21.1	3.1	--

Fuente: Ropero A, 2016.

1.3 Cadmio.

El cadmio (*cadmia* en latín y *kadmeia* en griego) es un metal pesado de origen natural que fue descubierto en 1817 en Alemania por Friedrich Stromeyer, como una impureza del carbonato de cinc. El cinc tiene aplicaciones metalúrgicas, aplicaciones en galvanoplastia, pinturas y pigmentos, baterías de níquel y cadmio, conductores eléctricos, componentes para automóviles, y sistemas de calefacción.

El cadmio no se encuentra en el ambiente como un metal puro, sino en forma de óxidos complejos, sulfuros y carbonatos de cinc, plomo y cobre. Al igual que otros metales pesados como el plomo, mercurio y arsénico, el cadmio constituye un riesgo considerable para la salud por el contacto ambiental y labora.

El cadmio en el ambiente se presenta en niveles bajos, sin embargo, la actividad humana ha incrementado esos niveles drásticamente. Desde la fuente emisora, el cadmio puede recoger grandes distancias a través del aire hasta el receptor. Se encuentran en moluscos, crustáceos, vegetales, tubérculos, y también en plantaciones como el cacao.

El cadmio provoca efectos tóxicos en los riñones, sistema respiratorio y sistema óseo. Además, es considerado como un agente cancerígeno para el hombre. El contacto con el cadmio se produce principalmente por: inhalación de humo de tabaco, consumo de alimentos contaminados, inhalación y contacto con emisiones industriales, exposición a actividades naturales como actividad volcánica, combustión de combustibles fósiles, incineración de basura, entre otros.

El cadmio contenido en el suelo y en el agua puede ser captada por ciertos cultivos y organismos acuáticos y almacenarse para ser parte de la cadena alimenticia. Esta es la principal fuente de contaminación de cadmio para no fumadores. Si bien la mayor cantidad

de cadmio es encontrada en hígado y riñón de los mamíferos alimentados con dieta alta en cadmio, la alimentación por cultivos también podría tornarse importante.

Es por ello, que la Agencia Estadounidense para el Registro de Sustancias Tóxicas y Enfermedades (ATSDR) ubica al cadmio como un metal con efectos de entre los más peligrosos.

Tabla 4. *Fuentes de Cadmio.*

Antropogénicas	*Naturales*
Lodos residuales y estiércol	Actividad volcánica
Fertilizantes fosfatados y nitrogenados	Rocas
Industria de plateado y galvanizado	
Minería del cinc, cobre, plomo y otros metales	
Industria de fundición de metales	
Incineración	
Industria de alimentos fosfatados para animales	

Fuente. Pérez P, 2012.

En los terrenos agrícolas, el cadmio llega por deposición aérea (41%), fertilizantes fosfatados (54%), y por aplicación de abono de estiércol (5%).

1.3.1 Propiedades físico - químicas del cadmio

El cadmio es un elemento químico relativamente raro, con número atómico 48 y de símbolo Cd, tiene estrecha relación con el Zinc, elemento con el que se encuentra asociado en la naturaleza.

El cadmio es un metal dúctil de color blanco con matiz azulado, más blando y maleable que el zinc y más duro que el estaño. No se encuentra en estado libre en la naturaleza, y el sulfuro de cadmio (único mineral de cadmio), no es fuente comercial del metal.

El cadmio es producido normalmente como subproducto de la fundición y refinamiento de los minerales de zinc. Anteriormente el zinc tenía un uso importante como cubierta anticorrosiva sobre hierro o acero. Otra aplicación es en baterías níquel-cadmio y otra es como reactivo químico. Los compuestos de cadmio se emplean como estabilizadores de plásticos y en la producción de cadmio fosforado.

El cadmio es liberado al ambiente aproximadamente sobre las veinticinco mil toneladas al año, este cadmio es liberado en los ríos a través de la descomposición de rocas, y parcialmente liberado al aire a través de fuego forestal y volcanes. La manufactura también genera un aporte de cadmio al ambiente.

Las aguas residuales que contienen cadmio, procedentes de la industria terminan normalmente en los suelos. Este cadmio puede ingresar también en el aire a través de la quema de residuos urbanos o combustibles fósiles.

Otra fuente importante de emisión de cadmio es la producción de fertilizantes fosfatados, que son aplicados en las granjas y transportados por aguas superficiales y/o subterráneas. El cadmio es absorbido por la materia orgánica del suelo, lo que se convierte en una situación extremadamente peligrosa, ya que, a través de la alimentación de las plantas y posterior consumo de los animales, el cadmio es luego acumulado en los cuerpos de los animales.

1.3.2 Perjuicios en la salud

El cadmio se acumula de manera primaria en el riñón, que puede desencadenar en una disfunción renal, que resultará en mayor excreción de proteínas en la orina.

Por otro lado, puede generar alteraciones en el metabolismo del calcio y también la formación de cálculos renales. Con exposición permanente al cadmio, se genera ablandamiento de huesos y osteoporosis.

Existe evidencia de que la exposición en largo plazo al cadmio contribuye al desarrollo de cáncer de pulmón, así como la posibilidad de desarrollar cáncer de riñón y próstata. Inclusive, el cadmio ya fue clasificado como uno de los componentes cancerígenos del grupo 1 para el hombre.

La Agencia Estadounidense para el Registro de Sustancias Tóxicas y Enfermedades, que cataloga riesgos de desechos tóxicos según su prevalencia y nivel de intoxicación que originan. Los más peligrosos son: plomo, mercurio, arsénico y cadmio. Este último, por tanto, es de alto interés toxicológico.

1.3.3 Niveles de cadmio permitidos

Algunas organizaciones internacionales han realizado propuestas para estimar una ingesta tolerable. Una propuesta establece un límite de 7ug/semana por kg de peso. Consumos de hasta 100gr por vía digestiva, generan síntomas gastrointestinales, y superiores a los 350gr, se considera potencialmente mortal.

Inhalaciones superiores a 500ug/m3 producen neumonitis química, y encima de los 5,000ug/m3, se hace mortal. Para 2009, la Autoridad Europea de Seguridad Alimentaria, estableció una ingesta semanal tolerable de 2.5ug/kg de peso corporal. Un año después, la Organización Mundial de la Salud (OMS), el Comité de Expertos en Aditivos Alimentarios (JECFA) y la Organización de Agricultura y Comida de Estados Unidos (FAO), determinaron un consumo de 25ug/kg de peso corporal de manera mensual.

Es importante resaltar, que las personas con bajas reservas de hierro son más vulnerables a los efectos adversos del Cadmio.

1.3.4 Toxicidad

Por lo expuesto en puntos anteriores, el cadmio produce efectos tóxicos en los organismos vivos, inclusive si las concentraciones son pequeñas.

En las plantas, el cadmio altera la absorción, transporte y utilización de los elementos esenciales: calcio (Ca), magnesio (Mg), fósforo (P) y potasio (K) y del agua, lo que genera desequilibrios nutricionales e hídricos en la planta. Las plantas expuestas a suelos contaminados por cadmio, presenten modificaciones en su apertura, fotosíntesis, y transpiración.

En los humanos, cuando se produce la ingestión, el cadmio ingresa al torrente sanguíneo por absorción, y llegará hasta el hígado, en donde se une a una proteína de bajo peso molecular. Finalmente son transportados a los riñones para ser reabsorbidos y almacenados en el riñón.

Entre las consecuencias más resaltantes por las exposiciones prolongadas, se pueden citar:

- Daño en los riñones de las personas que han estado expuestas a exceso de cadmio en su dieta o por el aire.
- Daños en pulmones, en personas que trabajan en fábricas con concentración de cadmio elevado.
- Cáncer pulmonar en animales expuestos a períodos largos de inhalación por cadmio.
- Elevada presión arterial, aún no comprobada para personas, pero sí comprobada en animales expuestos al cadmio.

Es importante acotar, además, que el cadmio está calificado como agente cancerígeno de tipo I por la Agencia Internacional para la Investigación del Cáncer (IARC), por lo que ya existe suficiente evidencia del riesgo de ser carcinógeno en personas.

1.4 Cadmio en plantaciones de cacao

El árbol del cacao absorbe el cadmio existente en los suelos, y los concentra en las semillas. La absorción sucede por las raíces hasta la distribución en toda la planta, y no se elimina inclusive después del lavado o lixiviación del suelo.

Estas plantas sufrirán de una reducción en su actividad fotosintética, de absorción de minerales, nutrientes y de agua, y podrían terminar en la muerte de la planta.

El Ministerio del Ambiente, en el anexo 1 del DS 002-2013, establece como parámetro de calidad para suelos de aplicación agrícola, que la máxima concentración permitida de cadmio es de 1.4 mg/kg. Por lo que plantaciones en suelos con niveles de cadmio superiores, generarán problemas en la planta y en los frutos que se produzcan.

El cacao peruano no es ajeno a esta contaminación de cadmio, debido a que nuestro territorio presenta gran probabilidad de contaminación por la industrialización, contaminación del aire y agua, y por actividad volcánica.

En el año 2006, se publicó un reglamento en el que se establece el contenido máximo de determinados contaminantes en productos alimenticios, este es: Reglamento CE N°1881/2006. Para el año 2014, la Unión Europea puso en vigor el reglamento UE N°488/2014, que modificaba el Reglamento anterior, y en el que se establecen los niveles máximos de Cadmio permitidos en diversos productos alimenticios (el cacao dentro de ellos). Es por ello, que, los nuevos límites quedaron regulados según el cuadro siguiente:

Tabla 5. *Reglamentación N°488/2014 para contenido de Cadmio en Cacao.*

Producto	Nivel máximo permisible (mg/kg)
Chocolate con leche con un contenido de materia seca total de cacao < 30%	0.10
Chocolate con un contenido de materia seca total de cacao < 50%; chocolate con leche con un contenido de materia seca total de cacao ≥ 30%	0.30
Chocolate con un contenido de materia seca total de cacao ≥ 50%	0.80
Cacao en polvo vendido al consumidor final o como ingrediente en cacao en polvo edulcorado vendido al consumidor final (chocolate para beber)	0.60

Fuente: Meter A, 2019.

Después de los 04 años transcurridos para ajustarse a esta medida, esta reglamentación entró en vigor desde el 01 de enero de 2019, con el objetivo de proteger a los consumidores más vulnerables y al público mayoritario, que suelen ser los niños.

Definitivamente, la aplicación de este reglamento ha complicado las exportaciones hacia la Unión Europea, pero permitirá a su vez, elevar los estándares de calidad en la manipulación y cultivo de este producto.

Para el Perú, el cacao es el octavo producto de mayor exportación y el segundo en línea agroindustrial, siendo Europa, el principal mercado de destino. Es por ello que, el cierre de este mercado de exportación reducirá la proyección de facturación de los próximos años, mientras que la contaminación con cadmio no sea controlada.

De la tabla N°05, de niveles máximos permisibles para producto final, se observa que no es aplicable de manera directa para materia prima. Sin embargo, es posible relacionar el contenido de cadmio en materia prima según el producto o concentración de cacao que será aplicado. La manteca de cacao contiene niveles mínimos de cadmio, y la concentración de cadmio en la masa de cacao es similar a la del licor de cacao (primer producto derivado de los granos de cacao después de la fermentación, secado y tostado). Así, conociendo el porcentaje

de masa de cacao en producto final de chocolate, se puede utilizar la ecuación a continuación para estimar el nivel máximo de cadmio en el cacao.

$$m_{CM} = \frac{m_{UE}}{X_{\%P}} \quad (01)$$

Donde:

m_{CM}: Nivel máximo de cadmio en la masa de cacao (mg/kg)

m_{UE}: Nivel máximo permitido de la UE en el producto terminado (mg/kg)

$X_{\%P}$: Porcentaje de masa de cacao en producto terminado

1.4.1 Mitigación del cadmio en el cacao.

El cadmio, por estar presente en el ambiente (de forma natural o por actividad antropogénica), es fácilmente absorbido por las plantaciones, y, por tanto, ser luego consumidas por animales y seres humanos, con lo que podría desencadenarse efectos nocivos en caso se superen los niveles máximos tolerables.

Es poco probable que exista una solución única para reducir la acumulación de cadmio en los granos de cacao debido a la diversidad de condiciones ambientales y de suelo de la región, así como la diversidad de fuentes de contaminación de cadmio y a los costos que su implementación implique.

Según recomendaciones de la comisión del Codex Alimentario sobre normas alimentarias, en las que participó la FAO (Organización de las Naciones Unidas para la Alimentación y la Agricultura) y la OMS (Organización Mundial de la Salud), de febrero de 2019, se puede optar por algunas prácticas recomendadas.

Antes de la siembra:

- Las nuevas plantaciones deberán ser realizadas en suelos de bajo contenido de cadmio (no deben contener más de 1.4 mg/kg de Cd).
- Reemplazar el monocultivo de cacao sin sombra por un diseño de plantaciones mixtas con diversas variedades de cacao.
- Las plantaciones deben estar alejadas de carreteras, de tal manera de evitar el contacto entre cacaotales y gases producto de combustión de vehículos.
- De igual manera, deberán permanecer alejadas de zonas mineras e industriales.
- Evitar los suelos inundables que podrían ser fuente de cadmio.
- El uso de cultivos de cobertura de leguminosas, ayudan a proteger al suelo de la erosión, y así, mejorar su calidad de materia orgánica, y reducir la toxicidad de los metales pesados.

Durante la etapa de producción hasta la cosecha:

- Monitorear la distribución de cadmio en el suelo y las posibles fuentes.
- Realizar mediciones de nivel de cadmio con laboratorios acreditados, analizando el suelo y la planta.
- Monitorear la salinidad del suelo y del agua de riego. A mayor salinidad, mayor absorción de cadmio.

Etapa de poscosecha:

- Escurrir el mucílago por 12 horas reduce significativamente el contenido de cadmio en los granos de cacao sin afectar su calidad.
- Durante el almacenamiento, debe impedirse la contaminación de los granos por derrames de combustibles, gases de escape o humo.

Algunas estrategias para inmovilizar el cadmio en el suelo:

- Usar sulfato de zinc para reducir el contenido de cadmio en los granos de cacao mediante fertilización balanceada.
- La deficiencia de zinc y manganeso incrementa la probabilidad de ingreso de cadmio a la Planta, por lo tanto, debe nivelarse para evitarlo.
- El encalado del suelo (aumentar el pH del suelo con aporte de calcio y magnesio), es una actividad muy importante sobre todo para pH medido inferior a 5.5. No excederse de ello, de lo contrario el efecto también se hace negativo.
- Usar abono orgánico, con la finalidad de aminorar la absorción del cadmio.
- De preferencia, usar fertilizantes nitrogenados y potásicos, ya que suelen tener bajo contenido de cadmio. Los fertilizantes fosfatados suelen contener cadmio como impureza.

Existen condiciones de suelo y de agua, que favorecen la bioacumulación de cadmio en granos de cacao, y que deben mitigarse para evitar su contaminación. En el cuadro a continuación se señalan algunas de estas condiciones:

Tabla 6. *Condiciones de suelo y agua en bioacumulación de Cadmio – Medidas de mitigación.*

Condiciones de suelo y agua que favorecen la bioacumulación de cadmio en los granos de cacao	**Medidas de mitigación propuestas**
La FDA de los EE.UU. indica que es muy poco probable que las aguas de riego con solución salina sean un problema en la producción de cacao y Cd. Es apropiado proporcionar una advertencia sobre los altos niveles de cloruro en el agua de riego, los fertilizantes y otras enmiendas del suelo. Y es específicamente el cloruro del suelo, no la salinidad, lo que provoca una mayor acumulación de Cd en todas las especies de plantas. Además, los consejos sobre el pH del suelo deberían ser más específicos. Las encuestas sobre las propiedades del suelo en las plantaciones de cacao en varias naciones informaron un pH del suelo tan bajo como 4,5 que promovió fuertemente la acumulación de Cd (ver la lista de publicaciones a continuación). Los suelos en la zona de enraizamiento deben ser encalados para alcanzar un pH de 6,5 si es necesario reducir los niveles de Cd de cacao. Y si el suelo tiene niveles naturalmente altos de Cd, los suelos deben ser encalados para minimizar la acumulación de Cd, u otros cultivos deben sembrarse.	
Suelos de baja fertilidad natural	Fertilizar el suelo con buen contenido de nutrientes
Bajo contenido de materia orgánica en el suelo	Incrementar la materia orgánica (> 4% MOS).
Baja concentración de Zn y Mn	Incorporación de Zn y Mn
Suelos arenosos	Evitar sembrar en suelos arenosos de preferencia utilizar suelos francos a arcillosos
Aguas salinas (2 mS/cm) con alto contenido de cloruros. En la unidad mencionada S significa Siemens.	Tratar el agua para bajar su salinidad y disminuir los cloruros
Suelos fuertemente ácidos	Encalar los suelos hasta niveles moderadamente ácidos a neutros

Fuente: Comisión CODEX Alimentos, 2019.

1.4.2 Concentración de metales en suelos por departamento y provincia

Debido a que el cacao ha tenido un crecimiento elevado en su producción en los últimos años, y también en sus niveles de exportación, la medición de los metales pesados en territorio nacional es de vital importancia. En la tabla N°07, se puede visualizar la distribución zonificada de los distintos tipos de metales pesados.

Tabla 7. *Valores de metales pesados totales en plantaciones de Cacao por Departamento y provincias. (0 – 20cm de profundidad).*

Departamento / Provincia	Fe*	Cu	Zn	Mn	Cd	Ni	Pb
	---------- (µg g $^{-1}$) ----------						
Tumbes							
Tumbes	3.31 ± 0.03a**	43.16 ± 0.98a	111.72 ± 0.92a	419.84 ± 9.99c	0.50 ± 0.03a	27.67 ± 1.08a	15.84 ± 0.20a
Zarumilla	1.55 ± 0.03b	16.48 ± 0.98b	45.83 ± 0.92b	379.86 ± 9.99c	0.05 ± 0.03c	13.54 ± 1.08b	9.67 ± 0.20b
Piura							
Huamcabamba	3.04 ± 0.02a	49.42 ± 0.66a	72.42 ± 0.61a	725.76 ± 6.66b	0.14 ± 0.02b	42.12 ± 0.72a	7.95 ± 0.14b
Morropón	2.11 ± 0.02a	27.46 ± 0.66b	64.80 ± 0.61a	429.73 ± 6.66c	0.53 ± 0.02a	10.05 ± 0.72b	9.18 ± 0.14b
Piura	3.32 ± 0.01a	49.14 ± 0.28a	85.93 ± 0.26a	671.85 ± 2.86b	0.48 ± 0.01a	25.30 ± 0.30a	12.67 ± 0.06a
Cajamarca							
Jaen	2.9 ± 0.02a	53.58 ± 0.66a	70.56 ± 0.61a	795.8 ± 6.66b	0.01 ± 0.02c	6.25 ± 0.72b	11.09 ± 0.14b
San Ignacio	2.21 ± 0.03a	29.59 ± 0.98b	61.00 ± 0.92a	553.19 ± 9.99c	0.00 ± 0.00c	8.29 ± 1.08b	8.47 ± 0.20b
Amazonas							
Bagua	2.50 ± 0.00a	25.5 ± 0.13b	75.34 ± 0.12a	524.87 ± 1.32c	0.11 ± 0.00b	16.56 ± 0.14a	10.05 ± 0.03b
Condorcanqui	3.06 ± 0.01a	43.56 ± 0.32a	81.54 ± 0.30a	739.84 ± 3.31b	0.01 ± 0.01c	22.75 ± 0.36a	15.92 ± 0.07a
San Martín							
Bellavista	1.66 ± 0.07b	15.44 ± 1.96b	52.85 ± 1.85b	486.64 ± 19.89c	0.20 ± 0.06b	13.69 ± 2.13b	7.18 ± 0.41b
El Dorado	0.21 ± 0.07b	2.43 ± 1.96b	3.76 ± 1.85b	82.63 ± 19.89c	0.00 ± 0.00c	1.64 ± 2.13b	5.52 ± 0.41b
Huallaga	1.56 ± 0.07b	14.44 ± 1.96b	43.56 ± 1.85b	482.68 ± 19.89c	0.13 ± 0.06b	10.89 ± 2.13b	5.76 ± 0.41b
Mariscal							
Cáceres	1.58 ± 0.01b	14.44 ± 0.4b	49.00 ± 0.37b	517.56 ± 4.00c	0.21 ± 0.01b	11.63 ± 0.44b	9.36 ± 0.08b
Tocache	1.26 ± 0.01b	6.86 ± 0.32b	35.16 ± 0.30b	477.42 ± 3.31c	0.00 ± 0.00c	3.50 ± 0.36b	6.71 ± 0.07b
Huánuco							
Huamalíes	4.11 ± 0.07a	49 ± 1.96a	87.24 ± 1.85a	919.3 ± 19.89b	0.00 ± 0.00c	43.03 ± 2.13a	15.37 ± 0.41a
Leoncio Prado	2.34 ± 0.02a	21.34 ± 0.66b	55.06 ± 0.61b	876.16 ± 6.66b	0.00 ± 0.00c	17.22 ± 0.72a	6.50 ± 0.14b
Junin							
Satipo	2.32 ± 0.02a	21.53 ± 0.49b	73.10 ± 0.46a	615.04 ± 4.97c	0.10 ± 0.01b	19.10 ± 0.53a	12.89 ± 0.10a
Cuzco							
La Convención	4.28 ± 0.01a	34.46 ± 0.4a	96.83 ± 0.37a	1275.20 ± 4.00a	0.00 ± 0.00c	24.70 ± 0.44a	21.81 ± 0.08a
Pv	***<0.0001***	***0.0035***	***<0.0001***	***0.0001***	***0.0001***	***0.005***	***0.0005***

*, Los valores de Fe están expresados en 10^4 miles de µg g $^{-1}$.
**, Promedios unidos con igual letra en columna no difieren significativamente, según Scott & Knott Alfa=0.05.

Fuente: Arévalo-Gardini E, 2016.

En la tabla, podemos identificar para el cadmio, que los valores nulos se encontraron en Cajamarca, San Martín, Huánuco y Cusco. El mayor valor fue encontrado en Piura (Morropón), con 0.53+-0.02 $ug\ g^{-1}$. Sectorizando, podríamos afirmar que la zona norte es la que presenta un índice de mayor concentración de cadmio, seguido por la zona central y finalmente la zona sur.

Es por ello, que el control y medidas preventivas/correctivas en los cultivos de la zona norte, y sobre todo en Piura, son de vital importancia para no generar contaminación con cadmio.

Capítulo 2

Machine Learning

2.1 Prefacio

La evolución tecnológica ha permitido que los sistemas digitales y electrónicos estén ahora al alcance de un universo de usuarios, que antes no podríamos haber imaginado. Hoy en día es posible disponer en casa de asistentes personales como Siri o Alexa, es posible reconocer matrículas de vehículos en circulación, es posible traducir expresiones completas utilizando *Google Translate* (antes era posible solo por palabras), y muchas otras alternativas que ahora se manejan inclusive desde celulares con Inteligencia Artificial.

En la búsqueda de seguir evolucionando de manera autónoma, nace una rama de la Inteligencia Artificial conocida como *Machine Learning*, que desarrolla principalmente la capacidad de un ordenador de aprender de manera autónoma la reacción a diversos estímulos. Esta tecnología puede aplicarse en diversos entornos con soluciones innovadoras para aplicaciones profesionales en investigación, cultura y otros. Con *Machine Learning* se puede conocer las preferencias de los clientes, clasificar de manera autónoma diversos productos, hacer una cadena de suministro logístico e inventarios autónomos, predecir o detectar enfermedades, entre otros.

2.2 Inteligencia Artificial

Inteligencia Artificial (IA) está compuesta por las palabras “Inteligencia” y “Artificial”. Podemos definir Inteligencia como la capacidad para adquirir y aplicar conocimiento a través de experiencia, y Artificial, como cualquier elemento no natural, o fabricado por el hombre.

La Inteligencia Artificial es usada en muchos campos, tales como optimización lógico-matemática, neurociencia, bioingeniería, automatización de procesos, entre otros, y es necesaria para situaciones en las que se requiera encontrar soluciones a problemas de gran complejidad.

Los sistemas de Inteligencia Artificial (IA) deben ser capaces de percibir señales externas (a través de sensores), procesarlas de manera correcta (modelarlas según el sistema en el que deba influenciar), y generar un efecto o reacción a través de sus actuadores.

Pero no es lo único, puesto que un sistema IA debe ir adquiriendo y procesando esta experiencia para futuras predicciones y nuevas reacciones (decisiones) del sistema inteligente.

La investigación en Inteligencia Artificial ha permitido desarrollar otras herramientas para resolver problemas de mayor dificultad, tales como:

- Búsqueda y Optimización
- Lógica
- Métodos probabilísticos y estadísticos
- Redes Neuronales
- Teoría de Control
- Idiomas
- Psicología Informática

Podemos encontrar ejemplos de Inteligencia Artificial de diferente complejidad en:

- Traductor de Google
- Asistentes personales como Siri o Alexa.
- Máquinas seleccionadoras
- Vehículos autónomos
- Drones
- Buscadores por imágenes
- Predicción de decisiones
- Juegos de razonamiento

2.3 Machine Learning

El Aprendizaje, similar al concepto dado antes de Inteligencia, es definido como la obtención de conocimiento, entendimiento de algo, o ganar habilidad para algo.

Machine Learning es el fundamento principal de la Inteligencia Artificial, definido también como el estudio de algoritmos informáticos y modelos estadísticos que mejoran un sistema automáticamente a través de la experiencia e inferencia.

Los algoritmos del *Machine Learning* construyen un modelo matemático basado en Datos de Entrenamiento para elaborar predicciones o decisiones sin ser programados de

manera explícita para optimizar una tarea que, con algoritmos convencionales, no generaría un procesamiento efectivo.

Tom M. Mitchell generó una definición más formal de los algoritmos estudiados en Machine Learning: Un programa de computadora aprende a partir de la experiencia E, sobre una tarea T, y con medida de rendimiento R, si su rendimiento medido por R, sobre un tema T, mejora con las experiencias E.

Por la metodología de entrenamiento también está relacionado con estadística pero con la diferencia del enfoque de cada ciencia, Estadística busca inferencias de una muestra, mientras que *Machine Learning* busca patrones predictivos que sean generalizables.

El objetivo de un sistema *Machine Learning* está enfocado en ser capaz de reaccionar de manera precisa ante estímulos nuevos, después de haber adquirido conocimiento producto de experiencia previa (datos de aprendizaje). La experiencia se obtiene a partir de una distribución probabilística desconocida, y el sistema debe construir el modelo genérico para generar predicciones precisas.

Por ejemplo, si un Sistema de Juego de Ajedrez, incrementa su habilidad y dificultad después de diversos ejemplos de juego, podría decirse que el Sistema ha aprendido. O, si un Sistema de Reconocimiento de voz, incrementa su eficiencia después de escuchar varios ejemplos de voces de varias personas, podríamos decir, que el sistema ha logrado su objetivo.

2.3.1 Principales algoritmos de Machine Learning

La diversidad de tipos de Algoritmos de *Machine Learning* tienen su base en el enfoque, el tipo de datos de entrada-salida, y el tipo de tarea que quiere resolverse.

2.3.1.1 Aprendizaje Supervisado (Supervised learning). En este tipo de aprendizaje, el algoritmo construye un modelo matemático desde un paquete de datos que contiene entradas y salidas esperadas.

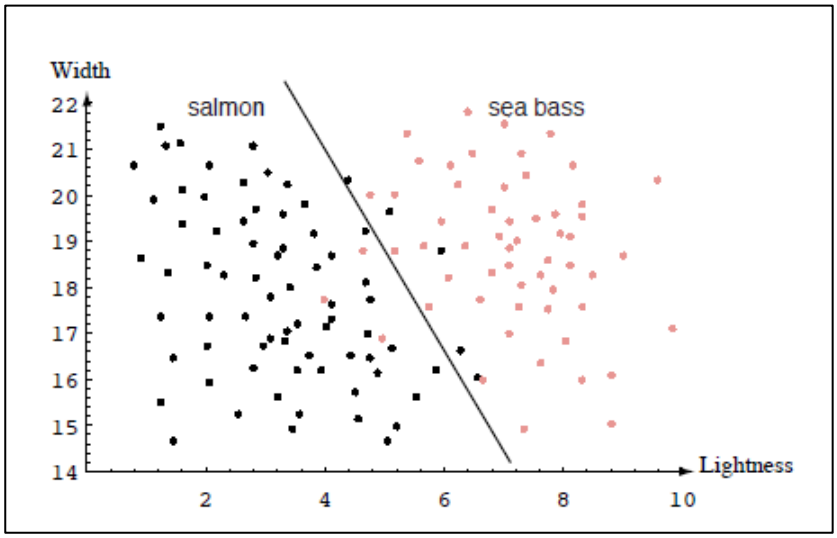
Estos datos son conocidos como datos de entrenamiento, y consiste en un conjunto de ejemplos de entrenamiento con diversas entradas y su respectiva salida esperada, también conocida como señal de supervisión.

Si consideramos que un ejemplo de entrenamiento es representado por un *array* (o vector), los datos de entrenamiento estarían representados por una matriz. Después de haber generado reiterativas pruebas de optimización para alcanzar la Función Objetivo, los algoritmos de Aprendizaje Supervisado aprenden una función para predecir salidas asociadas a nuevas entradas (inclusive entradas que no formen parte de las entradas de los datos de Entrenamiento).

Un ejemplo de este tipo de aprendizaje puede describirse en la situación en la que se busca que un equipo pueda clasificar un producto (pescado) entre dos clases, como salmón y lubina. Brindando características de luminosidad y dimensión, podríamos ir clasificándolo

hasta minimizar el error. Podrían seguir incluyéndose características para que la clasificación sea más exacta, de igual manera, la clasificación podría no ser lineal, dependiendo del grado de complejidad y precisión con la que se requiera alcanzar el objetivo.

Figura 16. Clasificación de salmón y lubina para el caso.



Fuente. Duda O.R, 2015.

Según el formato de la salida, los algoritmos de Aprendizaje Supervisado se subdividen en algoritmos de Clasificación y de Regresión.

Los algoritmos de Clasificación son usados cuando la salida recae en un conjunto finito de resultados posibles (opción limitada). Dependiendo de las opciones de salida, se conocen como de clasificación binaria (dos opciones) o multicategoría (varias opciones).

Por ejemplo: Predicción de si un producto será comprado o no (binario), y clasificación de tipos de pescado para productos de conserva (multicategoría).

Los algoritmos de regresión son usados cuando la salida puede obtener un valor diverso, dentro de un rango. Por ejemplo, la presión y velocidad de un objeto, reflectividad de un objeto, que es indicado como valor porcentual.

2.3.1.2 Aprendizaje No Supervisado (*Unsupervised learning*). Este tipo de aprendizaje ayuda a encontrar patrones previamente desconocidos en el conjunto de datos, sin etiquetas preexistentes. El algoritmo crea un modelo matemático solamente con datos de entrada, para encontrar una estructura o relación en los conjuntos de datos.

En este caso, no responden a la retroalimentación, sino que se identifican elementos comunes en los datos y reaccionan dependiendo de la presencia o ausencia, en cada nuevo conjunto de datos.

Por ejemplo: a partir de un conjunto de restos arqueológicos, encontrar la forma origen del modelo original.

Los principales métodos usados en Aprendizaje No Supervisado son: *Principal Components Analysis y Cluster Analysis*.

El *Principal Components Analysis* (PCA) es un procedimiento estadístico que utiliza una transformación ortogonal para convertir un conjunto de observaciones de variables posiblemente correlacionadas (entidades que adquieren varios valores numéricos) en un conjunto de valores de variables linealmente no correlacionadas llamadas componentes principales. Esta transformación se define de tal manera que el primer componente principal tiene la mayor varianza posible (es decir, representa la mayor variabilidad posible en los datos), y cada componente subsiguiente a su vez tiene la mayor varianza posible bajo la restricción que es ortogonal a los componentes anteriores. Los vectores resultantes (cada uno de los cuales es una combinación lineal de las variables y contiene 'n' observaciones) son un conjunto de bases ortogonales no correlacionadas.

Cluster Analysis se utiliza en el aprendizaje no supervisado para agrupar o segmentar conjuntos de datos con atributos compartidos para extrapolar las relaciones algorítmicas. *Cluster Analysis* es una rama del aprendizaje automático que agrupa los datos que no han sido etiquetados, clasificados o categorizados. Identifica elementos comunes en los datos y reacciona en función de la presencia o ausencia de dichos elementos comunes en cada nueva pieza de datos. Este enfoque ayuda a detectar puntos de datos.

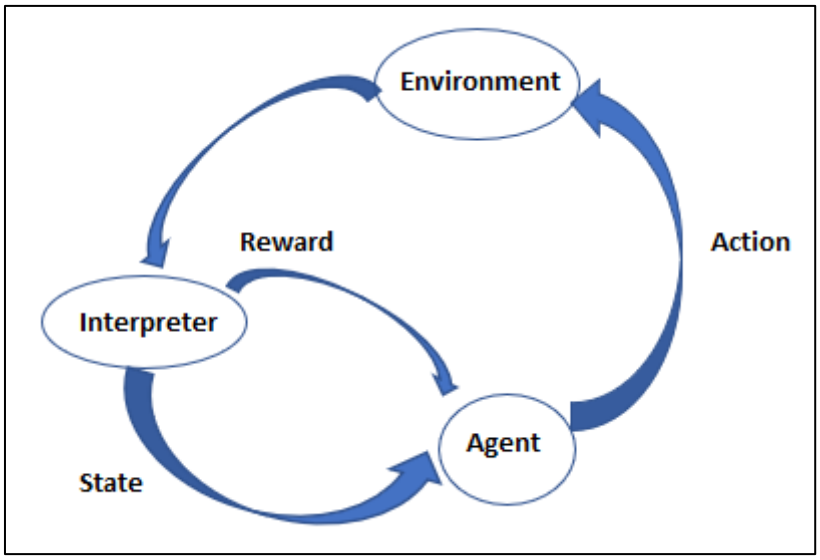
2.3.1.3 Reinforcement Learning. Partiremos del concepto de Reforzamiento de Aprendizaje en Psicología, en el que Reforzamiento se refiere al incremento de la probabilidad de una ocurrencia o acción particular, por ejemplo: destaques, premios, elogios, u otros.

En el plano de *Machine Learning*, el Reforzamiento del Aprendizaje está relacionado con la forma en la que los agentes informáticos toman acciones dentro de un entorno para maximizar el aprendizaje.

Los agentes del *Reinforcement Learning*, interactúan con su entorno en pasos de tiempo discretos, para cada tiempo t, el agente recibe la observación O_t, que incluye una recompensa H_t, luego se escoge una acción A_t, de un conjunto de acciones disponibles, que consecuentemente envía al entorno. Al pasar a otro escenario, un nuevo hallazgo es determinado. El objetivo de este aprendizaje es recolectar tantas recompensas como sean posibles.

En la imagen a continuación, se esquematiza un escenario de *Reinforcement Learning*, el agente ejecuta acciones en un entorno que es interpretado en una recompensa y representación del estado y retroalimentado hacia el agente.

Figura 17. Esquema de Reinforcement Learning.



Fuente. Elaboración propia.

El modelo *Reinforcement Learning* no solo trabaja con acciones inmediatas, sino que evalúa el rendimiento a largo plazo, es decir, maximiza los resultados futuros, aunque la inversión (o resultado inicial) sea negativo o menor. Para optimizar el rendimiento se emplean muestras y aproximación de funciones para entornos grandes.

2.3.2 Modelos

Se presentará a continuación, los principales modelos con los que trabaja *Machine Learning*:

2.3.2.1 Regresión lineal multivariable. La regresión lineal multivariable es una estrategia de Predicción de *Machine Learning*, basado en la minimización del error entre una función de predicción lineal y la salida real. Para llegar a la generalización multivariable, se presentará primero las ecuaciones para regresión lineal de una variable.

Para regresión lineal de una variable, podemos definir la hipótesis de predicción según:

$$h_\theta(x) = \theta_0 + \theta_1 . x_1 = \theta^T . x \qquad (02)$$

Notando la forma de ecuación lineal, en la que la pendiente estaría dada por θ_1, y θ_0 sería el término independiente. Con esta información, podemos definir la función de Costo como:

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)} \right)^2 \qquad (03)$$

El objetivo de la regresión será minimizar la función de costo hasta encontrar la ecuación lineal que mejor aproxime nuestro modelo en el espacio Rn. Para ello, aplicaremos optimización por derivación de la función de costo respecto de theta.

Con la ecuación de gradiente descendente, el parámetro θ_j, se actualizará hasta lograr el menor valor de $J(\theta)$.

Para el caso de regresión Multivariable, un paso previo al cálculo de los predictores, es la Normalización de datos de entrada. Esto debido a, por ser características o variables distintas, no siempre la regresión permitirá hacer los cálculos a una escala adecuada. Para mejorar la convergencia, se suelen normalizar para evitar diferencias por orden de magnitud.

Realizada la Normalización, obtendremos como datos de entrada, la matriz de características o de variables independientes y como salida, el vector de respuestas al sistema. La función de Costo será determinada por la ecuación:

$$J(\theta) = \frac{1}{2m}(X.\theta - \vec{y})^T.(X.\theta - \vec{y}) \quad (04)$$

Donde:

$$X = \begin{bmatrix} -\left(x^{(1)}\right)^T - \\ -\left(x^{(2)}\right)^T - \\ \vdots \\ -\left(x^{(m)}\right)^T - \end{bmatrix}, \text{ e } \quad \vec{y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(m)} \end{bmatrix}$$

Una opción de convergencia inmediata (sin iteraciones), es aplicar la fórmula de Ecuación Normal:

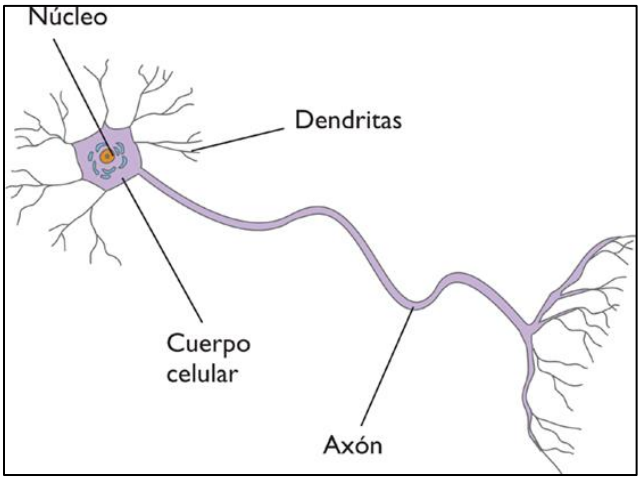
$$\theta = (X^T.X)^{-1}X^T\vec{y} \quad (05)$$

Con esta fórmula, no es requerida tampoco la Normalización, pues es una ecuación de aplicación directa. Para Regresión Multivariable, el objetivo es encontrar la Matriz de valores θ, que permitirá para diferentes entradas de X, predecir la salida.

2.3.2.2 Redes Neuronales. Las redes neuronales artificiales, dentro de *Machine Learning*, serán fundamentales para utilizarlas como una herramienta para resolver situaciones o problemas de Regresión o Clasificación Supervisada.

Primero debemos recordar, que las redes neuronales artificiales se basan en el funcionamiento de las redes neuronales del cuerpo humano. Y en estas pueden encontrarse tres elementos fundamentales: órganos receptores que reciben información del exterior, sistema nervioso que analiza y transporta la información y los órganos efectores, que convierten dichas señales en acciones.

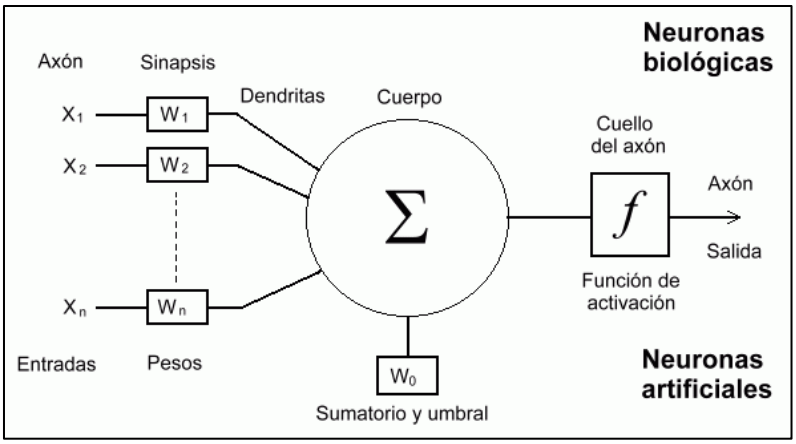
Figura 18. Representación de la neurona.



Fuente: Cruces E, 2016.

La unidad fundamental del sistema nervioso es la neurona, y ésta se une a otras formando redes. Su composición consta de un núcleo, de una ramificación de salida (axón), y de ramificaciones de entrada (dendritas). Funciona de la siguiente manera: las señales de entrada llegan a la neurona a través de sinapsis; mediante dendritas, la información llega al núcleo de la neurona, y se procesan salidas que serán propagadas por el axón. El sistema neuronal es un conjunto de neuronas conectadas entre sí, reciben información de unas neuronas, elaboran y transmiten información a otras neuronas.

Figura 19. Similitud entre neuronas biológicas y artificiales.



Fuente: Navarrete G, 2019.

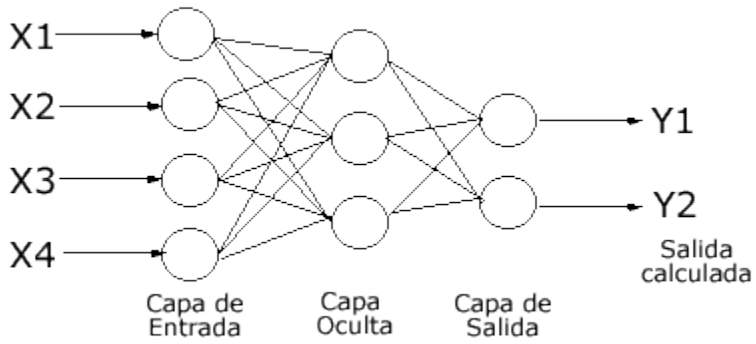
En el caso de las neuronas artificiales, éstas tratan de emular las características de las biológicas, y con un funcionamiento muy similar: cada neurona recibe un valor de entrada, y con una función de activación, la señal pasa a ser la salida de la neurona.

Dentro de la arquitectura de las neuronas, ellas se conectan entre sí siguiendo una arquitectura: las conexiones tienen un determinado peso, por lo que cada entrada a la

neurona es ponderada, es decir: la entrada de la neurona es la suma de las salidas de las neuronas conectadas a ella, multiplicadas por su respectivo peso (ver figura 19).

En la arquitectura de las redes neuronales artificiales, las neuronas son agrupadas en capas de entrada (*input layer*), capas de salida (*output layer*), y capas internas ocultas (*hidden layers)*.

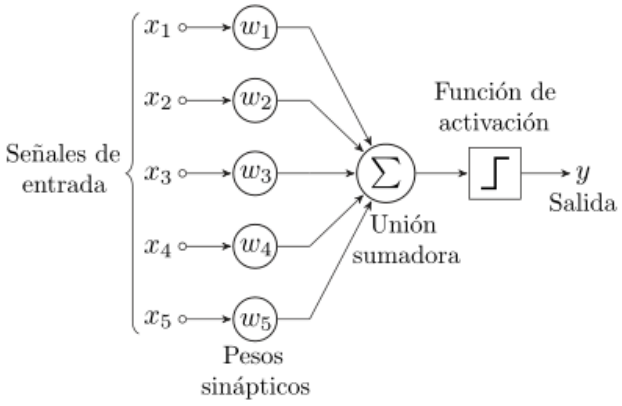
Figura 20. Arquitectura de red neuronal artificial.



Fuente: Mañas A, 2019.

El perceptrón es el modelo matemático más simple de una neurona, que por sí solo, no tiene aplicación, sin embargo, se hace de vital importancia y funcionalidad cuando se asocia con otras neuronas, al formar una red.

Figura 21 Esquema de funcionamiento de una red neuronal.



Fuente: Ramírez F, 2018.

De la figura 21, podemos visualizar las entradas X_i, que son ponderadas por los pesos sinápticos W_i, para luego ser adicionados para aplicarles una función de activación, con lo que se obtendrá la salida.

El perceptrón multicapa, es la generalización del perceptrón simple que tiene la particularidad de resolver los problemas de resolución del espacio R^n.

Las funciones de activación, normalmente usadas, son:

Función identidad: f(x) = x;

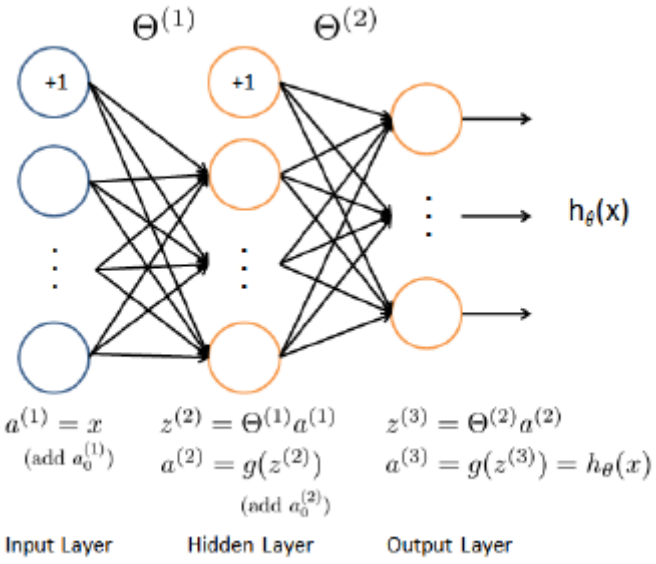
Función sigmoide: $f(x) = 1/(1+e^{-x})$

Función tangente hiperbólica: $f(x) = (1-e^{-x})/(1+e^{-x})$

El objetivo del modelo matemático de Redes Neuronales es obtener los pesos que serán aplicados a cada una de las entradas y que serán la base del modelo matemático. Esto es: $y_i = \sum_j w_{ij}.y_j$

Representación del modelo matemático:

Figura 22. Representación de una red neuronal.



Fuente: Tkachenko P, 2019.

Para la representación de esta red neuronal, se muestra la capa de entrada, una capa oculta y la capa de salida. Para encontrar los valores de peso de cada capa, se aplicará la metodología de la retropropagación (*Backpropagation*).

Para nuestro esquema, daremos nomenclatura de matriz Theta, a la matriz de pesos que conecta nuestra capa de entrada con la capa oculta y como matriz Theta 2, a la matriz obtenida por los pesos de conexión entre las neuronas de la capa de salida y la capa oculta. Esto es: θ_{ji}: peso ponderado entre la neurona de entrada i y la neurona de capa oculta j, y a θ_{kj}; peso ponderado entre la neurona oculta j y la neurona de salida k.

El aprendizaje de la red se da en medida en que la función de error (función de Costo) sea minimizada.

Feedforward y función de costo.

Vamos a suponer una red neuronal con *K* elementos en la capa de salida y 'm' elementos en la capa de entrada.

Definiremos a la matriz de data de entrada como *X*, y al vector de salidas como *y*.

$$X = \begin{bmatrix} -\left(x^{(1)}\right)^T - \\ -\left(x^{(2)}\right)^T - \\ \vdots \\ -\left(x^{(m)}\right)^T - \end{bmatrix} \quad , \qquad \vec{y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(m)} \end{bmatrix}$$

La función de activación a considerar será la función sigmoide. Por lo que quedaría según:

$$h_\theta(x) = g(\theta^T x) \tag{06}$$

La regla de propagación h_i, a partir de entradas y pesos sinápticos, es definida por:

$$h_i\left(x_1, \dots, x_n, w_{i_1}, \dots, w_{i_n}\right) = \sum_{i=1}^{n} w_{ij} x_j \tag{07}$$

A esta función se le añade un término conocido como Bias, que será el elemento externo añadido a la red en cada capa. Por lo que quedaría como:

$$h_i\left(x_1, \dots, x_n, w_{i_1}, \dots, w_{i_n}\right) = \sum_{i=1}^{n} w_{ij} x_j - \theta_i \tag{08}$$

Considerando que los índices i y j empiezan en 0, y denotando a $w_{i0} = \theta_i$ y $x_0 = -1$, expresaríamos la regla de propagación según:

$$h_i\left(x_1, \dots, x_n, w_{i_1}, \dots, w_{i_n}\right) = \sum_{j=0}^{n} w_{ij} x_j = \sum_{i=1}^{n} w_{ij} x_j - \theta_i \tag{09}$$

Donde la función de activación sería:

$$y_i = f_i(h_i) = f_i(\sum_{j=0}^{n} w_{ij} x_j) \tag{10}$$

Y la función de Costo sin regularización para Clasificación vendría dada por:

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \sum_{k=1}^{K} \left[-y_k^{(i)} . log\left(\left(h_\theta\left(x^{(i)}\right) \right)_k \right) - \left(1 - y_k^{(i)}\right) \log\left(1 - \left(h_\theta\left(x^{(i)}\right) \right)_k \right) \right] \tag{11}$$

Minimizando la función de Costo, se generará el aprendizaje de la red neuronal.

$$sigmoid(z) = g(z) = \frac{1}{1 + e^{-z}} \tag{12}$$

Sigmoid gradient: La función gradiente del sigmoide, que sería la derivada de la función anterior, quedaría representada por:

$$g'(z) = \frac{d}{dz}g(z) = g(z)\big(1 - g(z)\big) \tag{13}$$

Y la función de Costo con Regularización, quedaría como:

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{K}\left[-y_k^{(i)}.\,log\left(\left(h_\theta\left(x^{(i)}\right)\right)_k\right) - \left(1 - y_k^{(i)}\right)\log\left(1 - \left(h_\theta\left(x^{(i)}\right)\right)_k\right)\right] + \frac{\lambda}{2m}\left[\sum_{j=1}^{m}\sum_{k=1}^{K}\left(\Theta_{j,k}^{(1)}\right)^2 + \sum_{j=1}^{m}\sum_{k=1}^{K}\left(\Theta_{j,k}^{(2)}\right)^2\right] \tag{14}$$

La optimización de las funciones de costo, nos permitirán encontrar el mínimo de la solución, y así los parámetros o pesos deseados.

2.3.2.3 Support Vector Machines O máquinas de Vectores de Soporte, son un conjunto de algoritmos de Aprendizaje Supervisado codesarrollados por Vladimir Vapnik en los laboratorios AT&T.

La aplicación de los **SVMs** separa los puntos de muestra del espacio y subdivide las clases en más espacios mediante uno o más hiperplanos de separación. Al colocar las nuevas muestras sobre los vectores de soporte, estos nuevos elementos son reagrupados en las nuevas clases.

Un algoritmo basado en SVM construye un modelo capaz de predecir y separar de forma óptima a puntos de una clase u otra. Es acá donde radica la característica principal de los SVM, se busca un hiperplano en el que se encuentre la mayor distancia (margen) con los puntos más cercanos al mismo. Así, los puntos del vector etiquetados con una categoría se encontrarán de uno de los lados del hiperplano, y de otra categoría, del otro lado del hiperplano. Este algoritmo puede ser utilizado para Clasificación y para Regresión.

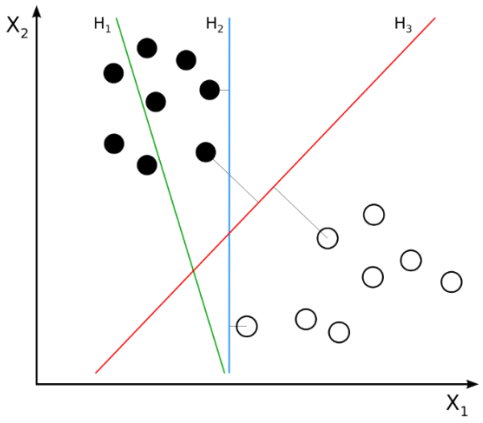
La organización del SVM está relacionada con la idea básica de redes neuronales de una capa o multicapa. Para patrones separables linealmente, se utiliza un hiperplano óptimo, y para patrones no separables linealmente, se puede transformar la data para un nuevo espacio usando las funciones de Kernel.

Suponiendo que los datos de entrada son los mostrados en el plano de la figura 23, se pueden generar diversos Hiperplanos, y no todos cumplirán la función de separar totalmente en clases distintas a los datos. Para elegir la mejor solución, se debe determinar aquel hiperplano que mantenga la mayor distancia o margen máximo entre los elementos de ambas categorías.

De lo que podemos observar, el Hiperplano 1 no separa correctamente las clases, el Hiperplano 2 las separa completamente, pero mantiene algunos márgenes bajos para algunos

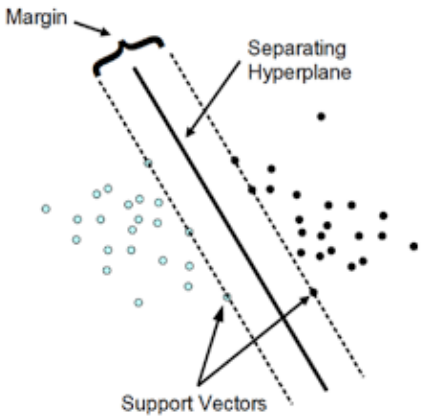
puntos clasificados. Para este caso, el Hiperplano 3 (H3), es el que mejor clasifica los datos y con un margen superior a los anteriores Hiperplanos. Sobre este hiperplano se trazarán dos líneas paralelas hacia ambos lados, a una distancia definida y máxima, de tal manera, que la unión de los puntos que las conforman, definirá a estos Vectores de Soporte.

Figura 23. Hiperplanos posibles para una distribución de datos.



Fuente. Chakraborty K, 2019.

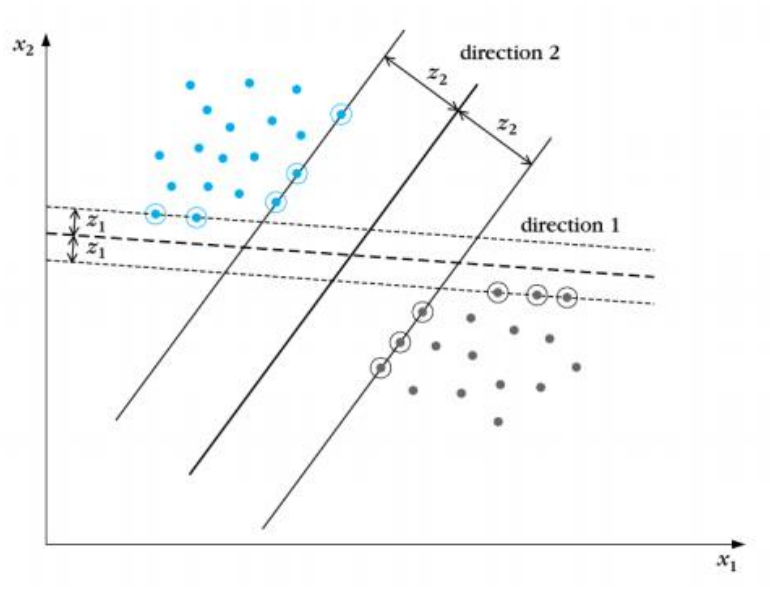
Figura 24. Vectores de soporte, paralelos y equidistantes del hiperplano con un margen.



Fuente: Roobini M.S, 2018.

Podemos visualizar en la figura 25, la presentación de dos hiperplanos que separan perfectamente las clases con márgenes distintos. El Vector de Soporte del modelo, serán los correspondientes a los que permitan mayor margen entre los datos, para este caso, la dirección 2.

Figura 25. Hiperplanos con sus márgenes de datos.

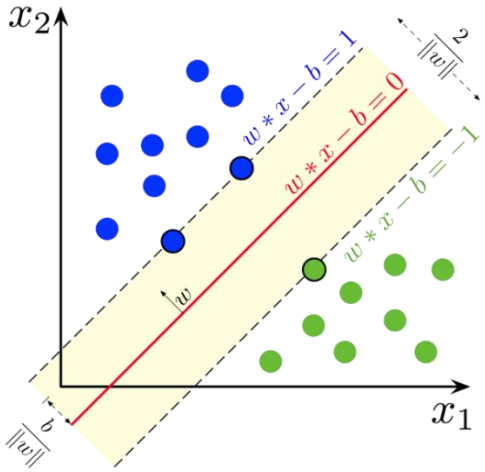


Fuente. Bonafini B, 2018.

Modelación matemática:

A partir del esquema de la figura 26, se desarrollará el modelo matemático.

Figura 26. Esquema de hiperplanos para SVM.



Fuente: Van Acoleyen, K, 2019.

Vamos a definir los hiperplanos H mostrados en la figura como:

$$w.x_i + b \geq +1, para\ y_i = +1 \quad (15)$$

$$w.x_i + b < -1, para\ y_i = -1 \quad (16)$$

Los vectores de soporte serían los definidos por las ecuaciones:

$$H1: w.x_i + b = +1 \quad (17)$$

$$H2: w.x_i + b = -1 \quad (18)$$

Recordemos que la distancia de un punto (x_0, y_0)a una recta: $A.x + B.y + c = 0$, es definida como: $\frac{|A.x_0+B.y_0+c|}{sqrt(A^2+B^2)}$. De esto se desprende que la distancia entre H_0 y H_1 es: $\frac{|w.x+b|}{||w||} = \frac{1}{||w||}$.

Entonces, 1/||w|| sería definida como la menor distancia hasta el punto más cercano de la clasificación, donde:

- w: vector de pesos.
- x: vector de entrada.
- b: bias

La modelación en SVM consiste en resolver un problema de aproximación y minimización, y suele utilizarse para ello: *Lagrangian function.*

La función a minimizar, será nuestro costo, definido como:

$$J(w, b) = \frac{1}{2}||w||^2 \quad (19)$$

Bajo la premisa: $|y_i - \langle w, x_i \rangle - b| \leq \varepsilon$

Funciones de Kernel.

Una forma simple de clasificar es mediante línea recta, plano recto, o hiperplano N-dimensional. Sin embargo, no todos los casos o situaciones son presentadas de esta manera, sino que también se pueden presentar:

- Más de dos variables predictoras.
- Curvas no lineales de separación.
- Conjuntos de datos que no pueden ser separados completamente.
- Clasificaciones en más de dos categorías.

Para estos casos se desarrollaron las funciones de Kernel, que permiten proyectar nuestra información hacia un espacio de características de mayor dimensión, y así incrementar la capacidad computacional de las máquinas de aprendizaje lineal.

Las funciones típicas de Kernel son las siguientes:

- Lineal:
- Polinómico:
- Gaussiano:

Funciones de Kernel típicos:

Kernel líneal: $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_j^T \boldsymbol{x}_j$

Kernel polinómico: $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i^T \boldsymbol{x}_j + c)^p$

Kernel Gaussiano: $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\sigma^2}\right)$

Podríamos mencionar algunas características a manera de comparativo entre Redes Neuronales artificiales y *Support Vector Machine* como sigue:

Tabla 8. *Principales características de Neural Networks y Support Vector Machines.*

Neural Networks	**SVM**
• Capas ocultas transforman a espacios de cualquier dimensión	• Kernels transforman a espacios de dimensión superior
• Entrenamiento costoso	• Entrenamiento eficiente
• Clasificación eficiente	• Clasificación eficiente
• El diseñador define número de capas ocultas y nodos	• Se diseña función Kernel y parámetro de Coste C
	• Menos necesidad de entrenamiento

Fuente: Elaboración propia.

2.3.2.4 K-means Clustering. El algoritmo *K-means* es un método para generar agrupaciones de datos similares dentro de un conjunto de muestras. Dado un conjunto de entrenamiento $\{x^{(1)}, \dots x^{(m)}\}$ donde $x^{(i)} \in R^n$ y queremos formar pequeños grupos, la intuición para el K-means inicia por suponer los centroides iniciales, y de manera iterativa, se van suponiendo nuevos centroides hasta encontrar los más cercanos, luego se repite el proceso.

El Algoritmo *K-means*, siempre convergerá hacia un conjunto de centroides principales. Sin embargo, no siempre la solución convergente es la ideal, sino que dependerá también de la elección de los centroides iniciales. Una forma de elegir entre las diferentes soluciones a partir de inicializaciones diferentes es, escoger la que tenga el valor menor en la función de costo.

Determinación de los centroides más cercanos.

En la etapa de asignación de clúster para el algoritmo *K-means*, el algoritmo asigna cada ejemplo a su centroide más cercano, dada la posición actual de los centroides. Se determina según:

$c^{(i)} := j$ que minimiza la función $\left\|x^{(i)} - \mu_j\right\|^2$, donde

$c^{(i)}$ es el índice del centroide más cercano a $x^{(i)}$, y μ_j es la posición del centroide j'ésimo

Cálculo de los centroides principales.

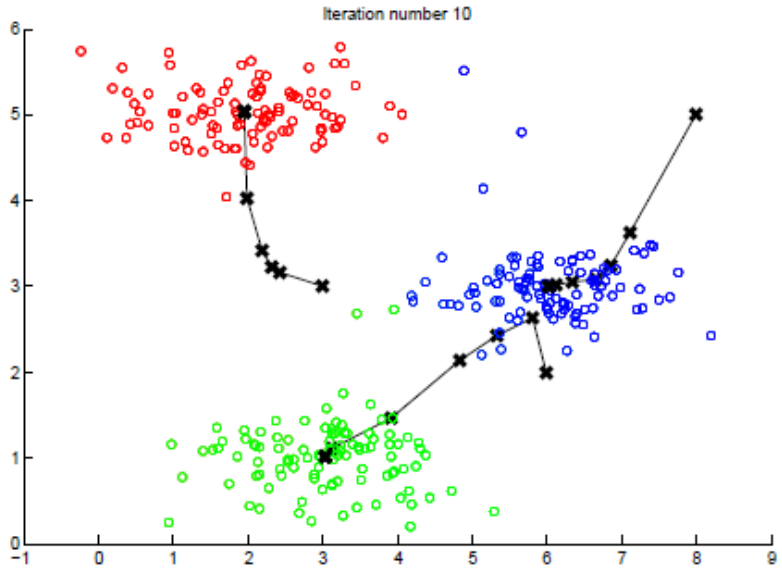
Considerando la distancia de cada punto hacia el centroide, la segunda fase del algoritmo recalcula para cada centroide, la media de los puntos que le fueron asignados. Esto es:

$$\mu_k = \frac{1}{|C_k|}\sum_{i \in C_k} x^{(i)} \tag{20}$$

Donde C_k es el conjunto de ejemplos asignados al centroide k. Por ejemplo, Si dos puntos $x^{(3)}$ y $x^{(5)}$ son asignados al centroide k=2, debe actualizarse según: $\mu_2 = \frac{1}{2}\left(x^{(3)} + x^{(5)}\right)$.

Un ejemplo de *K-means* Clustering se muestra en la figura 27, en la que se han asignado 03 centroides para la clasificación de las muestras.

Figura 27**.** Ejemplo de Centroides para K-means.



Fuente: Tkachenko P, 2019.

2.3.2.5 Principal Components Analysis. El método de *Principal Components Analysis* (PCA) es una forma de identificar patrones en datos, destacando sus similitudes y diferencias. PCA es una herramienta muy poderosa para análisis de datos cuando es complicado encontrar patrones en datos de altas dimensiones, o donde la representación gráfica no está disponible. Una ventaja de esta metodología es que, una vez encontrados los patrones de datos, reduce la cantidad de dimensiones sin sacrificar mucha información.

Esta técnica estadística tiene aplicaciones principalmente en procesamiento de imágenes (reconocimiento de rostros, compresión de imágenes) y en la búsqueda de patrones en data de gran dimensión.

PCA consiste en dos pasos computacionales, primero debe calcularse la matriz Covarianza de la data, luego debe calcularse los Autovectores, que corresponderán a los Componentes Principales de variación en la data.

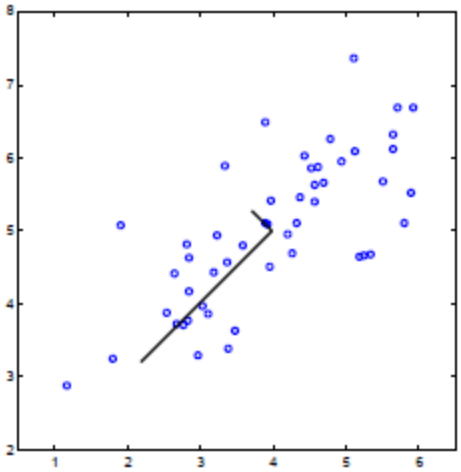
Antes de usar PCA, es importante primero normalizar los datos sustrayendo el valor medio (*mean value*), para cada valor del conjunto de datos, y llevando a la misma escala de valores los conjuntos de datos. La matriz covarianza de la matriz de datos, viene dada por:

$$\sum = \frac{1}{m} X^T . X \tag{21}$$

Donde X es la matriz de conjunto de datos en filas y m es la cantidad de muestras.

Según el software desde el cual se trabaje, se podrá obtener la matriz de Componentes Principales. Usando Matlab, el comando SVD permite obtener la Matriz U (de componentes principales), y la matriz S (matriz diagonal). Esto es: [U, S, V] = svd (Sigma).

Figura 28. Ejemplo gráfico de autovectores calculados del conjunto de datos.



Fuente. Tkachenko P, 2019.

Reducción de dhimensiones con PCA.

Con la matriz de componentes principales calculada, ésta puede utilizarse para reducir las dimensiones del conjunto de datos, proyectando cada muestra hacia un espacio dimensional menor. La reducción de dimensiones permite ejecutar operaciones más eficientemente.

2.3.2.6 Partial Least Squares Regression. La regresión por mínimos cuadrados es un método estadístico que tiene relación con la Regresión de Componentes Principales (PCR: *Principal Components Regression*), pero en lugar de buscar hiperplanos entre la respuesta y

variables independientes, busca un modelo de regresión lineal proyectando las variables predichas y las observables en un nuevo espacio. Este método es conocido como un modelo de factor bilineal. PLS es usado normalmente para encontrar relaciones fundamentales entre dos matrices *X* e *Y* para poder modelar la estructura de covarianza en esos dos espacios.

Modelación matemática:

Denominaremos nuestra entrada o matriz de variables independientes a **X**, como en los casos anteriores, y a la variable o variables dependientes como **Y** (**y** en caso sea un vector), por lo tanto: Y=f(X).

Si modelamos la variable dependiente como una función lineal de la independiente X, podríamos representar según:

$$\boldsymbol{y} = \boldsymbol{X}.\beta + \mathbf{f} \tag{22}$$

Donde:

- y: variable dependiente
- X: variable independiente.
- β: vector solución.
- f: error o residual.

Puede suceder que la cantidad de variables independientes sea muy superior a la cantidad de muestras. Esto puede ser trabajado de diversas maneras: selección de variables, reducción de la dimensión de la matriz de muestras, transformación de variables y otros, con el fin de que la matriz de entradas X sea de menor dimensión y de esta manera, facilitar los cálculos e incrementar la robustez del sistema.

Para trabajar con PLS, vamos a representar la matriz X según:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \tag{23}$$

En la que definimos:

T: Matriz de resultados o scores.

P: Matriz de cargas o resultados.

De esta manera, la matriz X podría ser descompuesta en un número de variables latentes caracterizadas por un vector **t** y un vector **p**.

Si se incluyen todas las variables latentes, el error se hace “cero” (E=0). Esto es equivalente a encontrar la matriz de coeficientes del modelo de regresión lineal múltiple incluyendo todas las variables. En la aplicación del modelo, suele representarse la matriz X (con un cierto margen de error), por una matriz T con menor cantidad de variables (o columnas).

El número de variables latentes necesarias para desarrollar la matriz X (dentro de un margen de error permitido) determina de la complejidad del modelo.

Dentro del diseño del modelo, son calculados otros vectores w^T (pesos) y b (sensibilidad). La relación entre la variable dependiente **y** y **T**, es:

$$\boldsymbol{y} = \mathbf{Tb} + \mathbf{f} \tag{24}$$

El objetivo de esta ecuación es encontrar b, de tal manera de que f sea minimizado. Podemos mencionar dentro de las ventajas de este modelo, que: los vectores t (de la matriz T), son linealmente independientes y pueden utilizarse para hacer aproximaciones; el modelo permite hacer una reducción de variables independientes cuando éstas sean muy grandes. Las primeras variables latentes guardarán la información más relevante, y las demás, serán modeladas por las últimas independientes.

Una desventaja del modelo es que éste es correlativo (No causal), por ello los modelos obtenidos no ofrecen necesariamente información fundamental respecto al sistema estudiado.

2.3.2.7 Anomaly Detection. Un algoritmo dentro de *Machine Learning* nos permite detectar anomalías en sistemas informáticos. Dada una muestra total de datos colectados, el sistema aprende o adquiere conocimiento del comportamiento normal de sus datos en operación, por lo que cuando los datos difieren significativamente de los datos normales, el sistema lo detecta como una anomalía. Este tipo de anomalías pueden ser: fraude bancario, problemas médicos, ruido en señales, u otros.

Las aplicaciones principales incluyen los dominios referidos a: detección de intrusiones, detección de fraude, detección de fallas en un sistema de monitoreo de salud, sensor de detección de redes, entre otros.

Supongamos una distribución Gaussiana para un conjunto de entrenamiento $\{x^{(1)}, \dots, x^{(m)}\}$, donde $x^{(i)} \in R^n$, para definir los parámetros μ_i y σ_i^2 para cada elemento, la distribución estaría dada por:

$$p(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \tag{25}$$

Donde μ es la media, y σ^2 controla la varianza, y vienen definidos por:

$$\mu_i = \frac{1}{m}\sum_{j=1}^{m} x_i^{(j)}$$

$$\sigma_i^2 = \frac{1}{m}\sum_{j=1}^{m} \left(x_i^{(j)} - \mu_i\right)^2$$

(27)

Una vez diseñada la Distribución Gaussiana, se define el límite ϵ, para definir la alta o baja probabilidad de que un elemento pertenezca a un conjunto de datos o de que sea una anomalía. El índice F1, nos permitirá evaluar el nivel de precisión de nuestro modelo, según:

$$F1 = \frac{2.prec.rec}{prec+rec} \tag{28}$$

Definido como:

$$prec = \frac{tp}{tp + fp} \tag{29}$$

$$rec = \frac{tp}{tp + fn} \tag{30}$$

Donde:

tp: número de verdaderos positivos. Indica el número de elementos que son anomalías, correctamente clasificados como anomalías.

fp: número de falsos positivos. Indica el número de elementos que no son anomalías, pero que el algoritmo clasificó como anomalía.

fn: número de falsos negativos. Indica el número de elementos que son anomalías, pero que han sido erróneamente clasificados como No anomalías.

2.3.2.8 Recommender Systems. *Machine Learnin*g es también utilizado para diseñar sistemas de Recomendación. Los algoritmos predicen las preferencias de los usuarios. Esta es una de las aplicaciones más exitosas y extendidas, pues este servicio permite interactuar en diversos entornos, tales como: compras por internet, video *on demand*, música por *streaming*, periódicos, revistas, recursos para eventos, viajes, educación, entre otros.

Recommender Systems, son típicamente clasificados en dos categorías: *Collaborative Filtering Systems*, que analizan solo las interacciones históricas, y *Content-based Filtering*, que está basado en atributos del perfil. Las técnicas híbridas combinan ambos diseños.

Vamos a definir las matrices Y y R para ejemplificar el modelo. La matriz Y almacenará la calificación $y^{(i,j)}$, mientras que la matriz R será un indicador binario que referirá si un ítem j fue evaluado por un usuario i ($R(i,j) = 1$ ó $R(i,j) = 0$ respectivamente.

Definiremos también las matrices X y Theta, donde X contendrá al vector de características $x^{(i)}$ referidas a la cantidad de ítems y el vector Theta contendrá a los parámetros $\theta^{(j)}$ referidos a la cantidad de usuarios. La función de Costo estará definida por:

$$J\left(x^{(i)}, \dots, x^{(n_m)}, \theta^{(1)}, \dots, , \theta^{(n_u)}\right) = \frac{1}{2} \sum_{(i,j):r(i,j)=1} \left((\theta^{(j)})^T x^{(i)} - y^{(i,j)}\right)^2 \tag{31}$$

Y los gradientes de la función de Costo sin regularización, estarán definidos por:

$$\frac{\partial J}{\partial x_k^{(i)}} = \sum_{j:r(i,j)=1} \left((\theta^{(j)})^T x^{(i)} - y^{(i,j)} \right) \theta_k^{(j)} \tag{32}$$

$$\frac{\partial J}{\partial \theta_k^{(j)}} = \sum_{i:r(i,j)=1} \left((\theta^{(j)})^T x^{(i)} - y^{(i,j)} \right) x_k^{(i)} \tag{33}$$

Al ejecutar el modelo respectivo en el código del programa de Recomendación, se obtendrá como resultado una valuación que entregará como resultado, la Recomendación según los datos de ingreso asignados.

2.3.3 Campos de aplicación de Machine Learning

Partiendo de los modelos que han sido mostrados en las líneas precedentes, podemos identificar diversos campos de aplicación del *Machine Learning*.

En Medicina: Aplicado a predicción de enfermedades y monitoreo de condiciones para identificar patrones en enfermedades antes de diagnosticarlas.

Ciberseguridad bancaria: Aplicada a evitar fraudes o ciberataques en las transacciones bancarias. Busca identificar transacciones legítimas respecto de ilegítimas.

Preferencias de selección: Aplicado por ejemplo a programas de Adquisición/Renta de películas y/o música. Por ejemplo: Netflix, Spotify, etc.

Autos inteligentes: Aplicado inclusive a los vehículos de conducción autónoma y a los autos con sistemas de detección de fallas y a los de conducción asistida.

Clasificación: Aplicado a diversos procesos industriales en los que pueda definirse criterios de clasificación de productos.

Como podemos ver, *Machine Learning* puede aplicarse en una diversidad de necesidades bastante amplia, identificando correctamente el problema a mejorar y la metodología a utilizar para hacerlo óptimo.

Vamos a añadir algunos comentarios respecto a las diferencias entre Inteligencia Artificial y *Machine Learning*.

Tabla 9. *Principales diferencias entre Inteligencia Artificial y Machine Learning.*

Inteligencia Artificial	Machine Learning
Habilidad de adquirir y aplicar conocimiento	Adquisición de conocimiento o habilidad
Máquina capaz de imitar el razonamiento humano	Subconjunto de IA donde las personas entrenan máquinas para reconocer patrones basados en experiencia y hacer predicciones.
Objetivo: incrementar la chance de éxito y no precisión	Objetivo: incrementar precisión, no necesariamente éxito.
Trabaja como programa de computación que hace trabajo inteligente.	Concepto simple en el que la máquina aprende de datos.
Busca simular inteligencia natural para resolver problemas complejos	Busca aprender de datos referidos a cierto tema para maximizar el performance de la maquina en dicho tema
IA lidera para inteligencia o sabiduría	ML lidera para conocimiento
IA es generador de decisiones	ML permite al sistema aprender nuevas cosas de los datos

Fuente. Elaboración propia.

UNIVERSITAS STUDIORUM PIURENSIS

Capítulo 3

Visión Hiperespectral

3.1 Prefacio

De la misma manera en la que el hombre puede utilizar la vista para clasificar un producto, resolver un problema, o tomar una decisión, es posible procesar las capturas de imágenes para resolución de problemas del día a día, esto gracias a los avances tecnológicos y dispositivos que son desarrollados para esta adquisición de datos.

Las cámaras que podemos encontrar en cualquier supermercado, catalogadas como domésticas, y que utilizamos para obtener imágenes estáticas, trabajan en unas bandas espectrales en las que se encuentra el umbral de visión humana.

Existen otras, utilizadas a nivel científico, que permiten registrar fenómenos que no son observables directamente, como: los eventos desarrollados en tiempo muy corto, o los desarrollados en tiempo muy largo, los microscópicos, los que impliquen espectros no visibles para el ojo humano, entre otros.

Lo que desarrollaremos a lo largo del capítulo nos permitirá un mayor entendimiento respecto a las imágenes hiperespectrales y su aplicación, y como se va convirtiendo en una de las principales formas de adquisición de datos sobre todo para detección de parámetros de calidad, plagas y otros en agricultura y otras ramas.

3.2 Imágenes Hiperespectrales

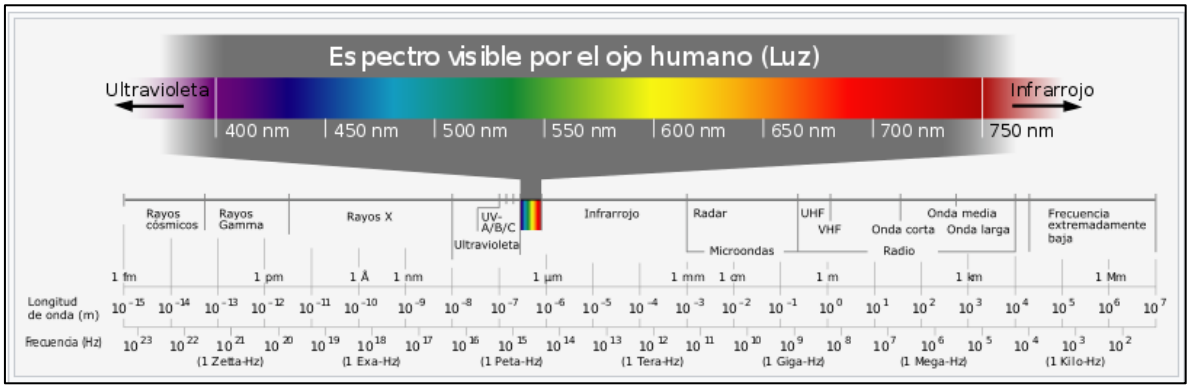
Las imágenes obtenidas por fotografía tradicional, almacena la información en un mapa bidimensional, lo que permite la proyección de un objeto sobre un plano. Las propiedades del elemento estudiado, como reflectancia o fluorescencia son grabadas por la imagen.

En las fotografías en blanco y negro, se sensa la intensidad total de cada píxel, integrando la luz recibida en un amplio rango espectral.

En el caso de las imágenes RGB (Red Green Blue), se obtiene información para las dimensiones espaciales (x,y), con tres bandas espectrales para rojo, verde, azul (así, se obtienen 03 imágenes monocromáticas de la misma escena).Las imágenes RGB permiten añadir a las mismas, una sensación real del color en el espectro visible por el ojo humano.

En la figura 29, podemos visualizar el espectro visible por el ojo humano, dentro de un rango de longitud de onda.

Figura 29. Espectro visible por el ojo humano.



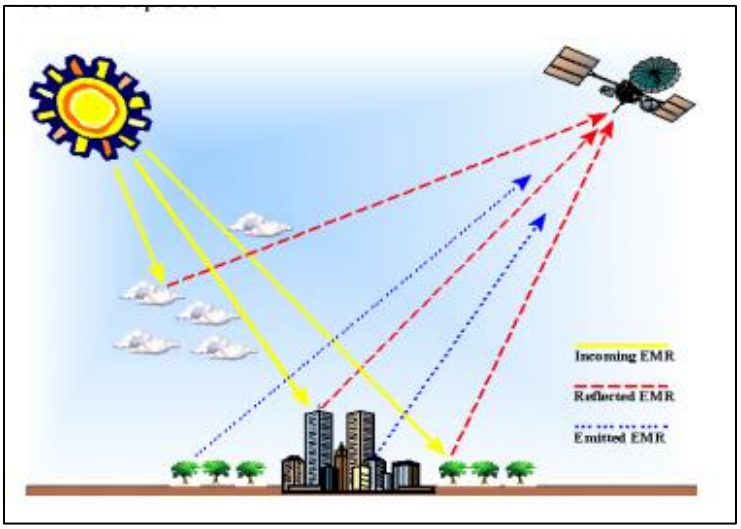
Fuente: Sánchez A, 2015.

Existen diversas aplicaciones científicas en las que inclusive con el color, no se hace suficiente para obtener su información relevante. Para ello, se busca conocer el espectro completo del objeto.

Diversos estudios e investigaciones orientadas principalmente a teledetección y espectroscopía espacial, fueron base para la obtención de propiedades físicas a partir de conocer la firma espectral del elemento de estudio.

Conocer el espectro completo de un cuerpo, permite conocer propiedades invisibles al ojo humano o aplicando fotografía convencional. Puede conocerse propiedades químicas y cinemáticas de estrellas distantes en el caso de astronomía, o conocer estado de salud de las plantas en campo de la agronomía, o conocer el estado físico o de salud de un órgano humano en el campo de la medicina.

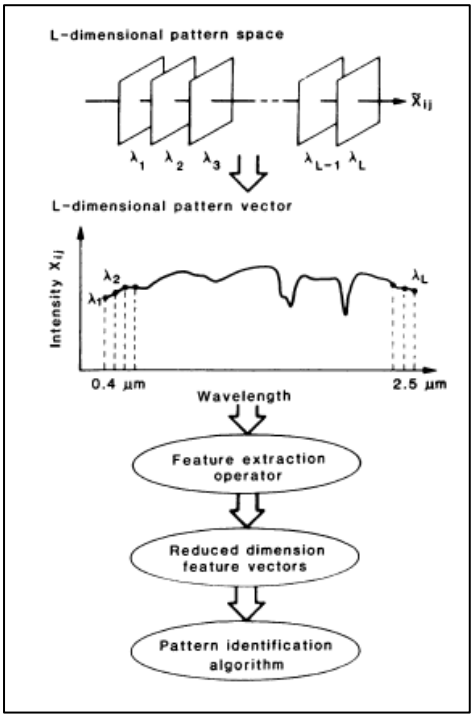
Figura 30. Adquisición de imágenes en teledetección.



Fuente: Roman-Gonzales A, 2013.

La imagen hiperespectral es formada por la recopilación y procesamiento de imágenes a lo largo de diversas bandas de espectro electromagnético. En la figura 31 podemos visualizar cómo se forman las imágenes hiperespectrales, según la intensidad de reflectancia en cada longitud de onda o banda espectral. A partir de ese conjunto de información o características, podemos definir nuestro algoritmo de identificación de parámetros.

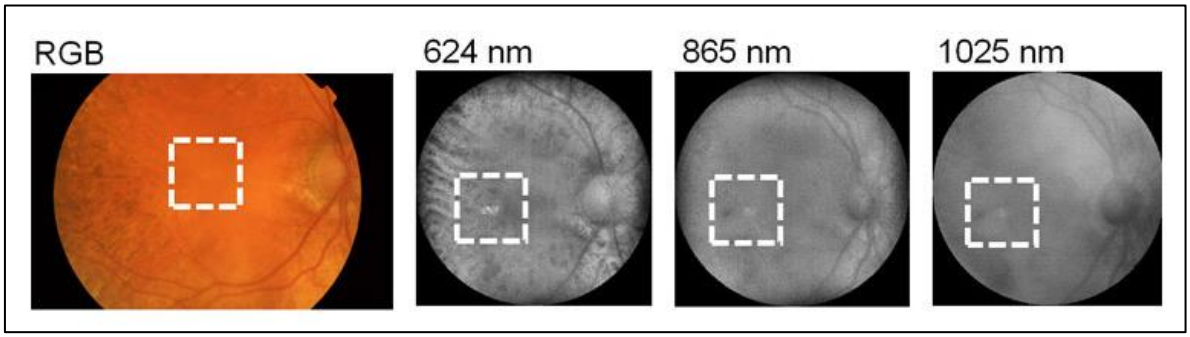
Figura 31. Recopilación de planos espectrales según longitudes de onda.



Fuente. Goetz A, Vane G, 1985.

Se puede observar en la imagen 32, un ejemplo de cómo un mismo cuerpo, es representado por diversas imágenes según la longitud de onda en la que se adquiera el espectro electromagnético.

Figura 32. Imagen espectral de una retina en diversas longitudes de onda (624, 885, 1025 nm).



Fuente. Alterini T., Díaz-Doutón F, 2019.

Mediante esta técnica, se establece una correlación entre la distribución espacial de un cuerpo y su espectro, aplicando espectroscopia para cada píxel. Los sensores de imagen que pueden tomar mediciones espectrales colectan la información como un conjunto de fotografías. Cada una de ellas representa a un rango estrecho de longitudes de onda del espectro electromagnético (banda espectral). La información obtenida (x, y, lambda), representan a las dos dimensiones espaciales de la imagen (x, y), y lambda representa la firma espectral para cada longitud de onda.

3.3 Visión Artificial

La visión artificial es una rama de la Inteligencia Artificial que simula el sentido de la 'vista', analizando e interpretando la imagen adquirida de un elemento u objeto captado.

Replicar este proceso humano, mediante el cual se puede adquirir información de: forma, color, defectos externos, textura superficial, permite realizar diversos procesos de manera autónoma como: clasificación de productos, detección de objetos, análisis de imágenes satelitales, huellas dactilares (*Touch iD*), reconocimiento de rostros (*Face iD*), entre otros. Sin embargo, puede presentar limitaciones cuando se presentan casos de colores similares, clasificaciones complejas, encontrar propiedades internas, y otros.

3.3.1 Procesamiento de imágenes

Durante la formación de imágenes digitales, suele presentarse ruido u otras interferencias que degradan la calidad de la imagen. Es por ello que se aplican técnicas de mejora como:

- Disminución del ruido

- Modificar el contraste de imagen.
- Ajustar el brillo
- Suavizar o realzar los bordes,
- Modificar enfoque
- Otros

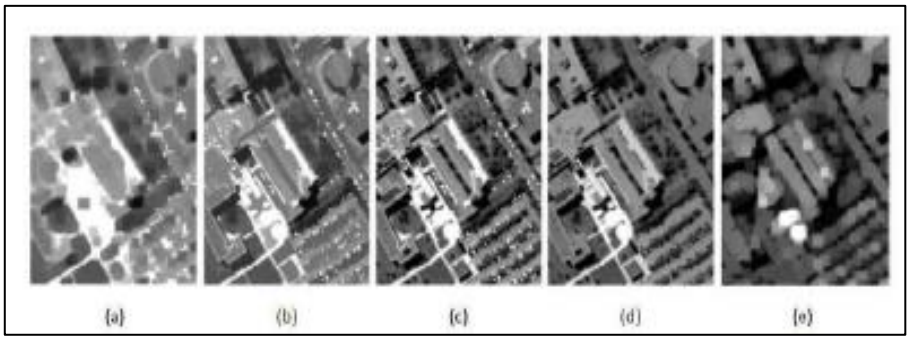
La mejora de imágenes puede realizarse en el **dominio espacial**, cuando se aplican al propio plano de la imagen, manipulando directamente los píxeles de la imagen, o en el **dominio de frecuencia**, cuando se aplica en la transformada de Fourier de una imagen, en lugar de su propia imagen.

Mencionaremos algunas técnicas de procesamiento digital de imágenes aplicables para imágenes multidimensionales y de imágenes hiperespectrales.

a. Morfología matemática. Esta es una técnica que permite extraer elementos de imagen que representan y describen la forma de la región, como: bordes, límites, etc. Dependiendo del objeto a analizar, se define la forma y tamaño. Son dos operaciones morfológicas principales: dilatación y erosión. La primera consiste en incorporar todos los puntos de fondo conectados al objeto, además de reducir las diferencias entre dos objetos separados.

Mientras que la erosión consiste en eliminar todos los puntos de contorno del objeto y también elimina detalles pequeños. Existen otras operaciones morfológicas como las mostradas en la figura 33.

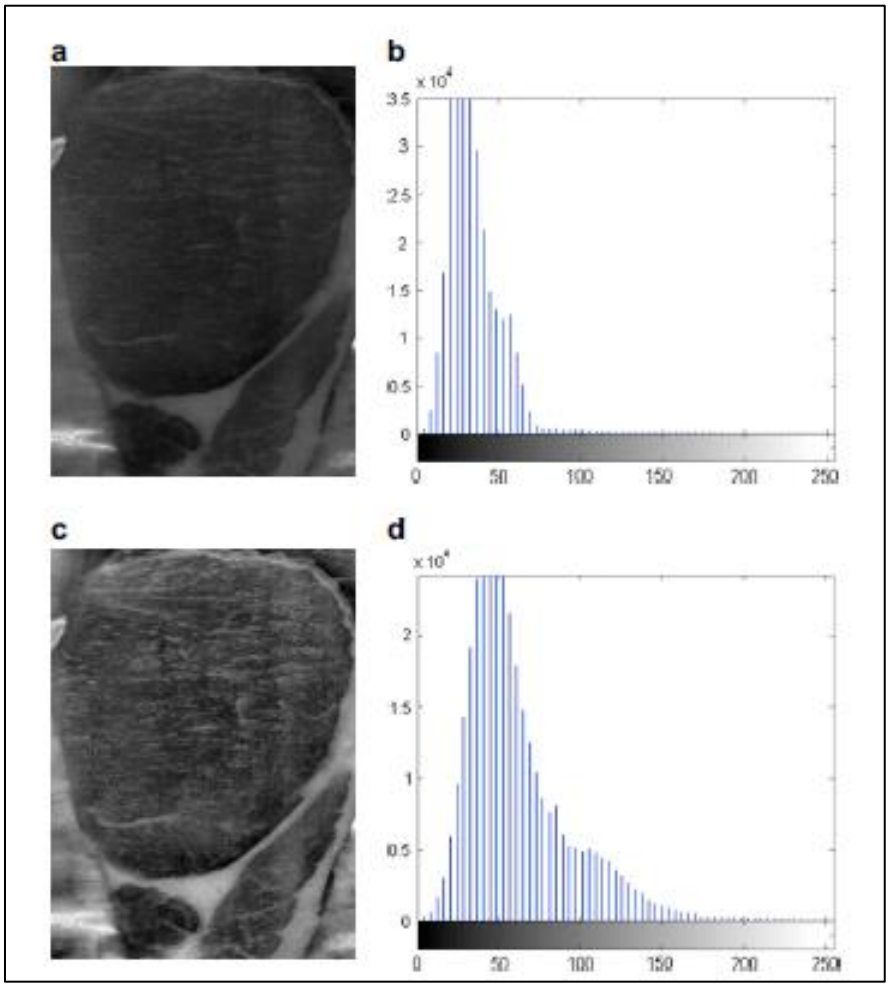
Figura 33 a) Cierre morfológico; b) Cierre por reconstrucción; c) Imagen pancromática VHR original; d) Apertura por reconstrucción e) Apertura morfológica.



Fuente: Ngadi & Liu, 2010.

b. Ecualización de histograma. El histograma de la imagen en escala de grises es la frecuencia relativa de aparición de cada nivel de gris en la imagen. El proceso de ecualización del histograma consiste en redistribuir los niveles de gris de la imagen por la reasignación de los valores de brillo de los píxeles. El principal problema en esta ecualización es el ruido, pues genera distorsión en la imagen e incremento de contraste.

Figura 34. Calidad de imagen mejorada usando ecualización de histograma: (a)Imagen espectral de una muestra de cerdo; (b)Histograma de la imagen en (a); (c) Imagen obtenida después de ecualización de histograma de (a); (d)Histograma de imagen en (c).



Fuente: Ngadi M, Liu L,2010.

Filtrado. El filtrado engloba a un conjunto de operaciones en las que el valor de un píxel depende de su valor anterior y el de los píxeles contiguos, sin alterar la geometría de la imagen resultante. Los filtros que pueden aplicarse son: **pasabajos**, que atenúan las frecuencias altas y mantiene sin variación las bajas, **pasaaltos**, que atenúa las frecuencias bajas manteniendo invariables las frecuencias altas, y **pasabanda**, que atenúa las frecuencias muy altas y muy bajas, manteniendo una banda de rango medio invariable.

Según el dominio de trabajo, también son clasificados como: filtros en dominio del espacio y filtros en dominio de la frecuencia. Los objetivos buscados con esta aplicación son:

- Suavizar la imagen
- Eliminar ruido.
- Realizar bordes.

- Detectar bordes.
- Optimizar la imagen y enfatizar información.

3.3.2 Segmentación

En el campo de la visión artificial, Segmentación es el proceso de dividir una imagen de una imagen, en otra más fácil de analizar. La segmentación es usada para localizar objetos y para encontrar límites dentro de una imagen.

La segmentación genera un conjunto de elementos segmentados, que, en conjunto, formarán nuevamente toda la imagen, o un conjunto de curvas de nivel de la imagen.

Los algoritmos de segmentación se basan en los principios de: *discontinuidades del nivel de gris*, usadas en detección de líneas, bordes y puntos aislados; *y, similitud de nivel de gris*, usadas normalmente en umbralización, crecimiento de regiones, etc.

La segmentación puede aplicarse en:

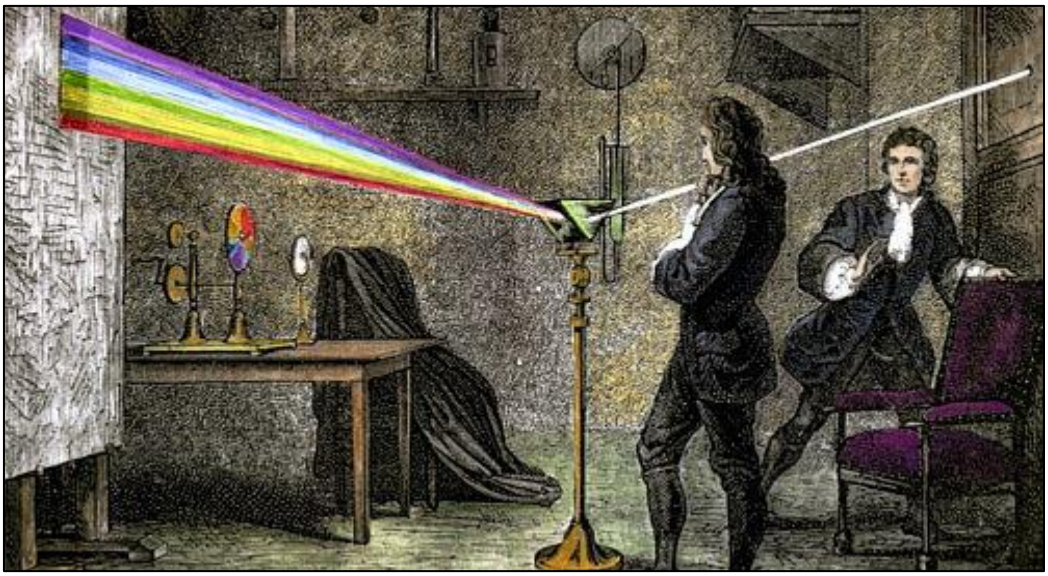
- Pruebas médicas
- Localización de objetos en imágenes de satélite.
- Sensor de huella digital.
- Reconocimiento facial.
- Sistemas de control de tráfico, etc.

3.4 Espectroscopia

La espectroscopia es el estudio de la interacción entre la radiación electromagnética y la materia en función de la longitud de onda (λ), con la finalidad de proporcionar información química y física, cualitativa y cuantitativa de la materia (basado en la Ley de Beer). La evaluación espectral busca detectar la absorción o emisión de radiación electromagnética a diferentes longitudes de onda.

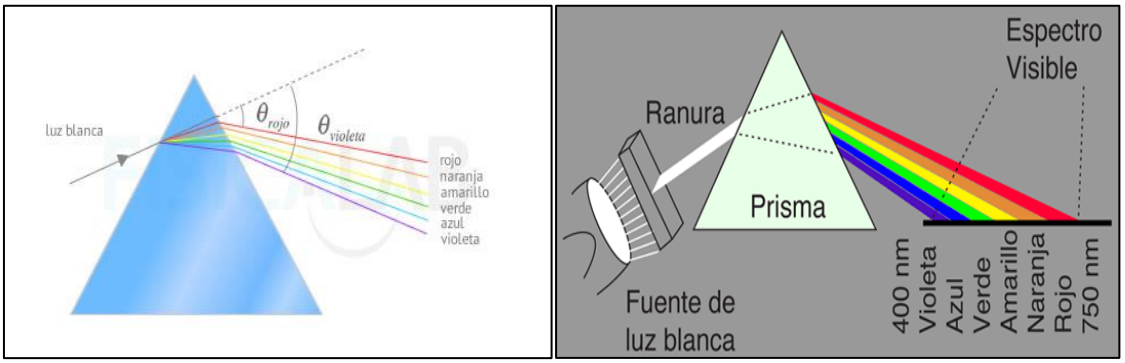
Isaac Newton, en 1665 descubrió que cuando un haz de luz atraviesa un prisma de cristal, ésta se descompone en un espectro continuo de colores. Esto dio origen a la técnica de la espectrometría.

Figura 35. Representación de la descomposición de luz de Newton.



Fuente. Villatoro F, 2016

Figura 36. Descomposición de la luz a través de un prisma.

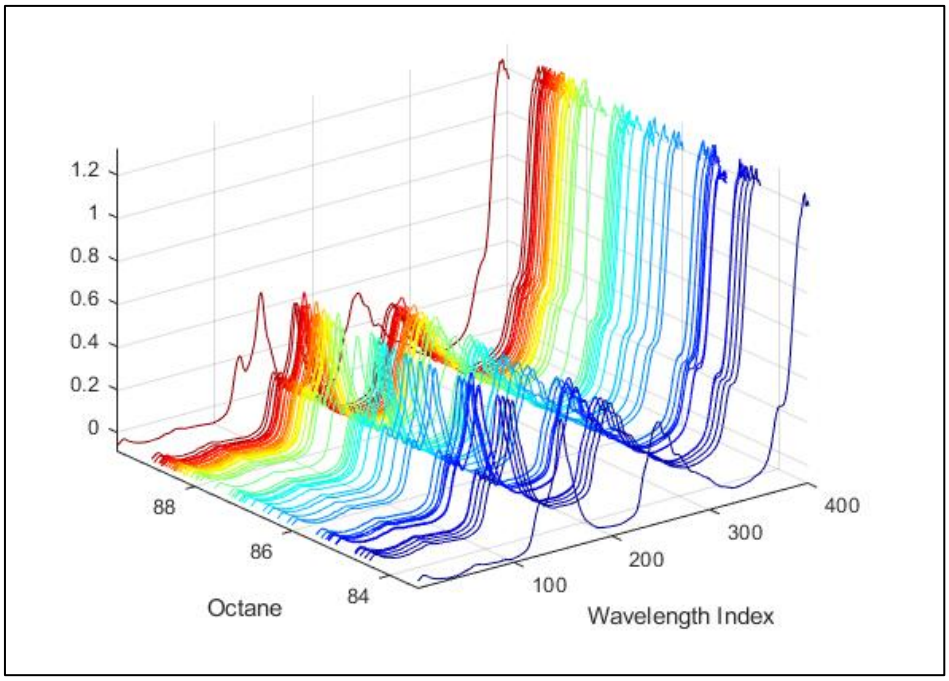


Fuente: Martínez W, 2020.

El análisis espectral estudia la luz en función de la longitud de onda absorbida o reflejada por un material. Cuando los fotones del haz de luz tocan la superficie de la materia, algunos son reflejados desde la superficie, otros atraviesan la materia y otros son absorbidos.

A nivel molecular, la dispersión o absorción de la luz depende de los enlaces atómicos o de la estructura molecular. Cada cuerpo tiene propiedades ópticas especiales, con patrón de espectros únicos, que indican su composición química y sus propiedades físicas y estructurales.

Figura 37. Intensidad espectral de 60 muestras de gasolina con 401 longitudes de onda.



Fuente: Kalivas J, 1997.

Ley de Beer-Lambert

La ley de Beer-Lambert en óptica, conocida también como ley de Beer, es una relación empírica que relaciona la absorción de la luz con propiedades del material atravesado.
Wilhel Beer y Johann Lambert propusieron que la absorbancia de una muestra a determinada longitud de onda depende de la cantidad de especie absorbente con la que se encuentra la luz al pasar por la muestra. La ley de Beer-Lambert relaciona la intensidad de luz entrante en un medio con la intensidad saliente después de que en dicho medio se produzca absorción. La relación puede expresarse según:

$$\text{Líquidos: } \frac{I_1}{I_0} = 10^{-A} \quad \text{Gases: } \frac{I_1}{I_0} = 10^{-\alpha'} \tag{34}$$

Donde: I_1, I_0; son intensidades saliente y entrante respectivamente.

$$A = \alpha . l;$$

l: longitud atravesada por la luz en el cuerpo o medio

α: coeficiente de absorción

$\alpha' = \frac{4\pi k_\lambda}{\lambda^2}$: coeficiente de absorción, donde λ es longitud de onda de luz absorbida y k_λ: es el coeficiente de extinción

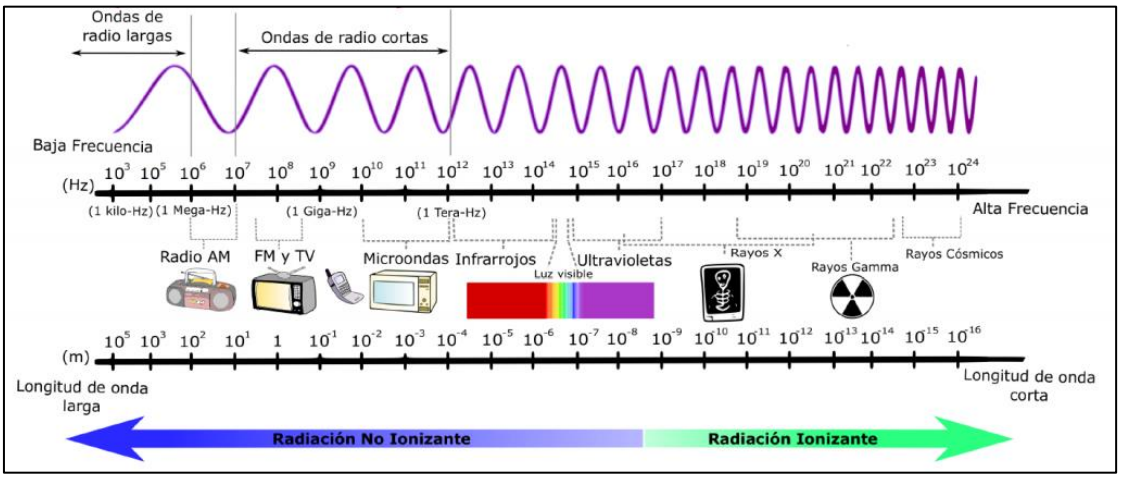
Podemos entender con esta ley, que existe una relación exponencial entre la transmisión de la luz a través de una sustancia y la concentración de la sustancia, así como entre la transmisión y longitud del cuerpo que la atraviesa.

3.4.1 Espectro electromagnético

Espectro electromagnético se denomina a la distribución energética del conjunto de ondas electromagnéticas que un objeto absorbe o emite. Esta radiación permite identificar la sustancia u objeto como si se tratase de su huella dactilar.

Este espectro permite obtener sus características físicas como intensidad de radiación, longitud de onda y frecuencia. Se extiende desde la radiación de menor longitud de onda, hasta las ondas electromagnéticas de mayor longitud de onda (baja frecuencia), o sea: rayos gamma, rayos X, radiación ultravioleta, luz visible, radiación infrarroja, hasta las ondas de radio – que son las de mayor longitud de onda-.

Figura 38. Espectro electromagnético.



Fuente. Gonzáles G, 2018.

La energía electromagnética para una longitud de onda λ está relacionada directamente con una frecuencia f y una energía de fotón E. Por lo que el espectro puede representarse por las ecuaciones:

$$\lambda = \frac{c}{f} \tag{35}$$

$$E = h.f \text{ o su equivalente: } E = \frac{h.c}{f} \tag{36}$$

Donde:

C= 299 792 458 m/s (velocidad de la luz), y

h=6.62607x10^{-34} J.s (constante de Planck)

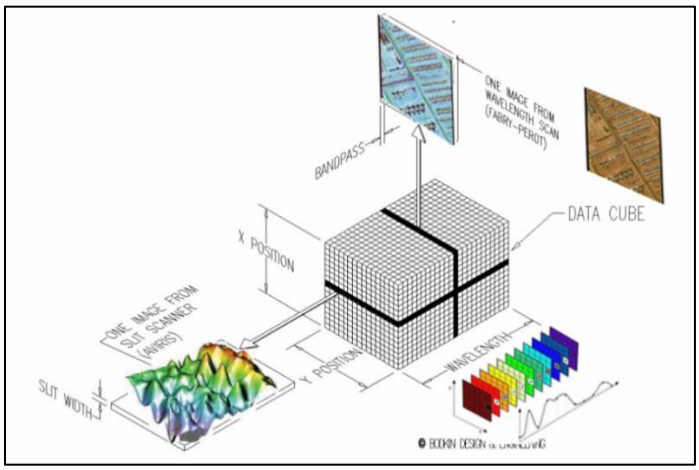
De esto se puede deducir que las ondas electromagnéticas de alta frecuencia tienen longitud de onda corta y mucha energía, mientras que las ondas de baja frecuencia tienen poca energía y grandes longitudes de onda.

3.4.2 Representación de imagen hiperespectral

Las imágenes hiperespectrales son representadas por un cubo de datos tridimensional denominado: cubo hiperespectral tridimensional, hipercubo, cubo espectral o volumen espectral. Al plano bidimensional de representación de las imágenes digitales, se le añade una nueva dimensión para el espectro, de esta manera, el hipercubo quedará plenamente representado con la representación de imágenes bidimensionales (x, y) para cada longitud de onda unidimensional (lambda).

Para cada longitud de onda, la intensidad espectral en diferentes lugares dentro de la imagen representará la absorción de luz.

Figura 39. Cubo espectral.



Fuente. Bodkin, Sheinis, 2009.

3.4.2.1 Resolución. La resolución refiere al nivel de detalle con el que se pueden capturar las imágenes, su frecuencia temporal y su 'finura espectral'. Se pueden considerar así, cuatro tipos de resolución: espectral, espacial, radiométrica y temporal.

- **Resolución espacial**: Este tipo de resolución refiere a la finura de los detalles visibles en una imagen. Mientras menor es el área representada por cada píxel, mayores son los detalles que pueden ser captados y mayor es la resolución espacial (designa al objeto más pequeño que puede distinguirse en la imagen).

Esta resolución depende mucho de los sensores utilizados, el poder del sistema óptico, influencia atmosférica, presencia de humo, nivel de iluminación, etc. Otros factores importantes a son el buen contraste y bordes nítidos del objeto, así se favorece la detección. El incremento de resolución lleva consigo un encarecimiento del equipo de detección y mayor peso para el procesamiento de información.

Figura 40. Resolución de 1m – imagen satelital.

Fuente. QuickBird, 2005.

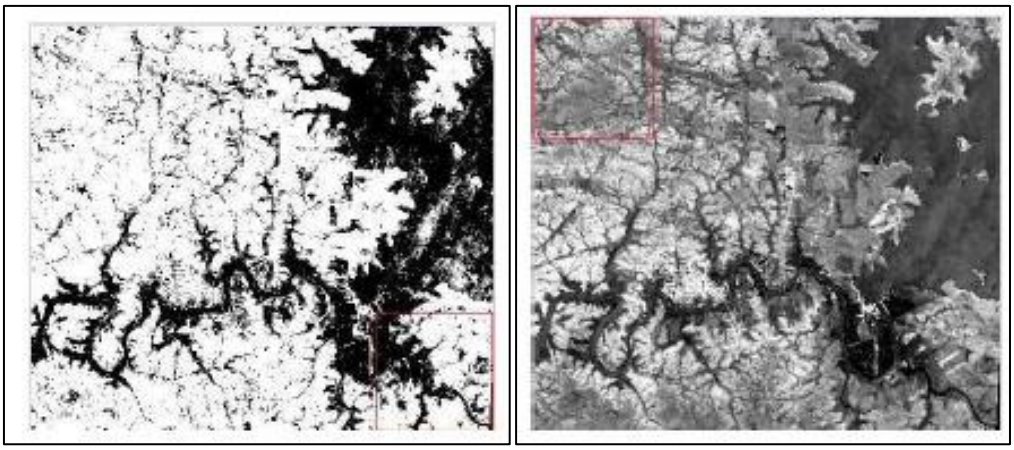
- **Resolución espectral:** Esta resolución se refiere al número y ancho de bandas espectrales registradas por el sensor de adquisición. Mientras más estrechas sean estas bandas, mayor será su resolución espectral. Por la banda de los sensores de percepción, suele distinguirse los sistemas multiespectrales y los hiperespectrales. Por ejemplo: el LANDSAT y SPOT son sistemas multiespectrales, caracterizados por un número no muy elevado de bandas espectrales. El sensor de ASTER, de la NASA, posee 14 bandas en región visible, infrarroja y térmica del espectro.

Los sistemas hiperespectrales registran imágenes en cientos de bandas espectrales estrechas. Mientras más estrecha es la banda, menor es la energía transmitida al detector.

- **Resolución radiométrica:** Esta resolución hace referencia al número de niveles digitales utilizados para expresar los datos recogidos por el sensor. Mientras mayor es número de niveles, la información se podrá expresar con mayor detalle.

En la imagen 41, se puede visualizar la misma toma considerando pocos niveles digitales, y a su costado, más niveles digitales.

Figura 41. Ejemplificación de resolución radiométrica.



Fuente: Teledet, 2007.

- **Resolución temporal:** Esta resolución es una medida de la frecuencia con la que un satélite es capaz de obtener imágenes de una determinada área. Cuando los procesos o eventos cambian en períodos cortos, como: incendios, inundaciones, cosechas, y otros, es necesario tener una alta resolución temporal.

3.4.2.2 Firma espectral. La firma espectral o huella digital espectral, es el patrón de reflexión, absorbancia, transmitancia, y/o emisión de energía electromagnética en diferentes longitudes de onda. Esta firma, debido a la diferente composición química y estructura física de los materiales, es distinta para cada material y/u objeto.

3.4.2.3 Imágenes multiespectrales y ultraespectrales. Las técnicas de adquisición de imágenes espectrales, pueden clasificarse como: multiespectrales, hiperespectrales y ultraespectrales,

Los sistemas multiespectrales trabajan con imágenes en bandas discretas y estrechas. Suelen ser de costo bajo en comparación con las hiperespectrales, y suelen diseñarse de manera más específica. Por el tamaño de procesamiento de información, es beneficioso en aplicación de tiempo real (es más rápido), podemos ver ejemplo de aplicación en la figura 42.

Los sistemas hiperespectrales, tratan con imágenes en bandas estrechas en un rango de longitud de onda continuo, resultando en el “espectro” de todos los píxeles de la escena. El sistema hiperespectral es usado para determinar un conjunto de información más apropiado para longitudes de onda en sistema multiespacial. Podemos ver un ejemplo de ello en las cámaras hiperespectrales de Resonon, en la figura 43.

Los sistemas ultraespectrales son usados con sensores de imágenes tipo interferómetro (aplica al fenómeno de interferencia de las ondas), con una resolución espectral muy fina. Estos sensores suelen tener una resolución espacial de pocos píxeles, debido a la alta velocidad de datos.

Figura 42. Ejemplo de cámaras multiespectrales usadas en agricultura.

Fuente: Hobbytuxtla, 2020.

Figura 43. Cámaras hiperespectrales Resonon.

Fuente. SpectrononPro, 2020.

3.5 Procesamiento de Imágenes Hiperespectrales

La calidad, eficiencia y correcto procesamiento de análisis con sistemas Hiperespectrales, dependerá en primera instancia, de la selección adecuada de instrumentos, diseño del sistema y su calibración.

3.5.1 Adquisición de Imágenes Hiperespectrales

Para la adquisición de imágenes y formación de los cubos hiperespectrales es puede utilizarse métodos de detección y escaneo, según la configuración de los elementos conformantes del sistema de adquisición.

3.5.1.1 Métodos de detección o captura de Imágenes hiperespectrales. Según las características a obtener de la muestra en análisis, existen tres modelos de detección para la

formación de imágenes hiperespectrales, según se dé la disposición de la fuente de luz y del detector óptico (cámara, espectrógrafo, lente).

a. **Reflectancia.** Este método es uno de los más usados, y con un costo relativamente bajo, utilizado comúnmente en el análisis de muestras sólidas y semisólidas. En este método, el detector captura la luz reflejada por la muestra iluminada, con una disposición con la que pueda evitarse la reflexión especular o regular (normalmente formando 45° con la fuente de luz, ver figura OO).

Con este método, puede determinarse la composición química de frutos, y características externas de los mismos como: color, forma, tamaño, textura superficial, y posibles defectos en la superficie.

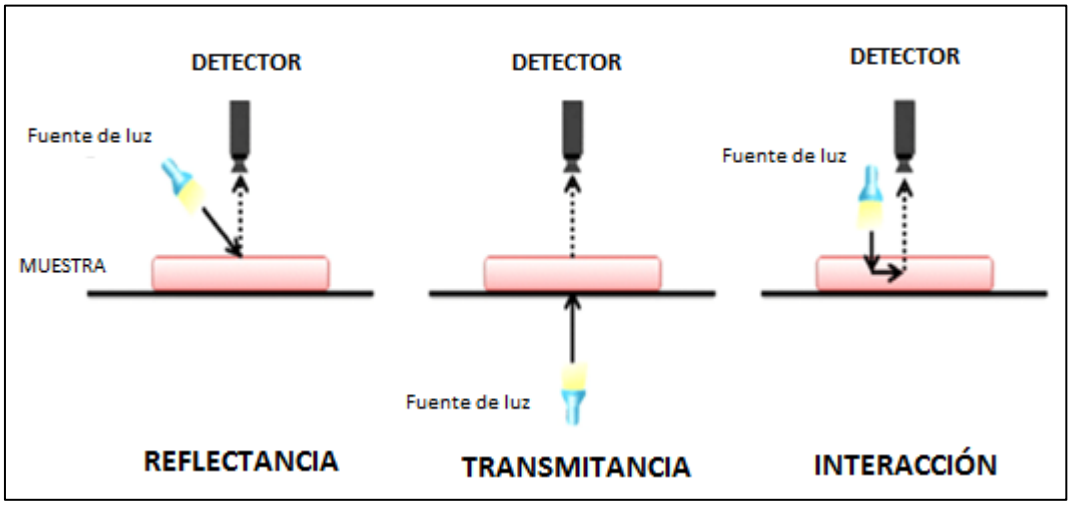
b. **Transmitancia.** Bajo esta metodología, el detector es colocado del lado opuesto a la fuente de luz, y capta la luz transmitida a través de la muestra. Su objetivo es determinar la cantidad de energía que atraviesa un cuerpo por unidad de tiempo.

Este método es empleado para analizar sólidos de baja densidad, líquidos y semilíquidos, sin embargo, no es usual su empleo para analizar alimentos debido a que la intensidad de la luz podría causar daños térmicos en la muestra y esto a su vez, cambiaría sus propiedades espectrales.

c. **Interacción.** En esta metodología se dispone el detector, paralelo a la fuente de luz, y del mismo lado de la muestra, así, la reflexión especular o regular no alcanzará al detector.

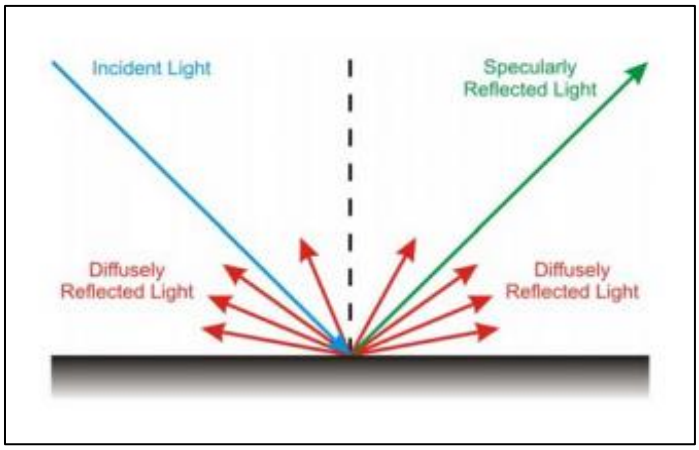
Con esta distribución, la información que se obtiene contiene información más profunda que la que podría obtenerse mediante reflectancia. En la figura 44, se muestra la distribución para los métodos antes mencionados:

Figura 44. Métodos de captura de imágenes hiperespectrales.



Fuente. Wu and Sun, 2013.

Figura 45. Reflexión especular y difusa.



Fuente. Analytik, 2019.

3.5.1.2 Métodos de escaneo. De manera convencional, se manejan las siguientes metodologías para construir una imagen hiperespectral:

a. **Escaneo espectral.** En este método de escaneo, cada lectura de un sensor bidimensional representa un mapa espacial (x, y) monocromático, lo que lo hace conocido como escaneo por área. El objeto es escaneado espectralmente mediante el intercambio de un filtro tras otro y la plataforma estática. Las capturas pueden realizarse utilizando:

- **Cámara de filtros intercambiables.** Mediante la que es posible tomar imágenes monocromáticas de la misma escena, usando varios filtros. Así, se procesa la misma escena bajo distintas longitudes de onda.

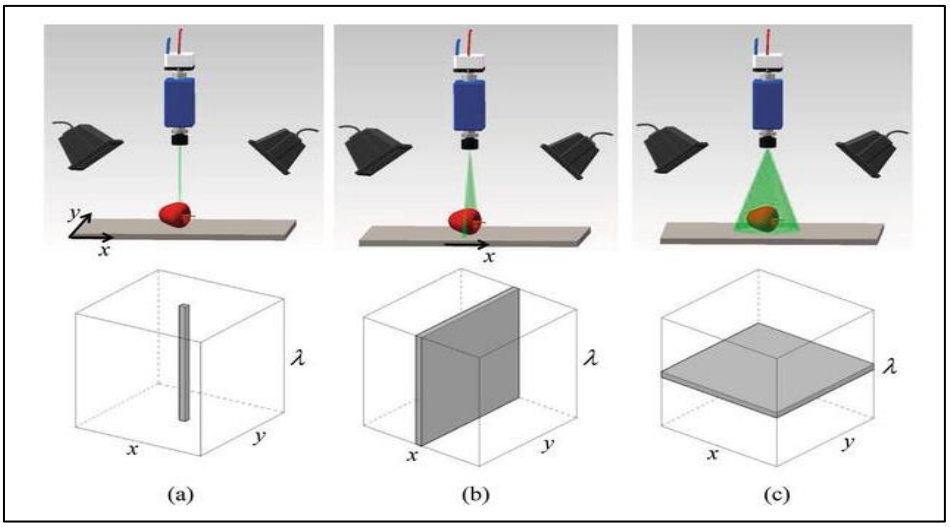
- **Cámara de filtros alternables.** Esta se caracteriza por utilizar filtros ajustables mecánica o eléctricamente.

b. **Escaneo espacial.** En esta técnica, cada lectura de un sensor bidimensional se corresponde con un corte completo de espectro (x, lambda). Los dispositivos de adquisición de imágenes hiperespectrales, obtienen un corte del espectro proyectando una franja de la escena y dispersándola mediante un prisma de difracción. Podemos destacar, en esta técnica, según el tipo de recorrido:

- **Escaneo de barrido o lineal.** En este tipo de escaneo, se adquieren los espectros de líneas registrados por un detector de array. La luz proveniente de la escena cruza la abertura lineal y luego dispersada a través de una matriz bidimensional de detectores, así, los puntos a lo largo de la línea son muestreados simultáneamente.

- **Escaneo de puntos.** Este sistema puede considerarse como un caso particular del escaneo lineal, en la que la abertura a utilizar es por punto, en lugar de una franja. El sensor es unidimensional, en lugar de bidimensional.

Figura 46. (a) Escaneo de puntos; (b) Escaneo lineal, (c) Escaneo por área.



Fuente. Xiaona Li, 2017.

3.5.2 ***Componentes de una cámara hiperespectral.*** Los componentes básicos para el diseño de un sistema de visión hiperespectral son: fuentes de luz, dispositivos de dispersión de longitud de onda, detectores de área, software de procesamiento.

- **Fuente de Luz:** La selección de la fuente de luz y su posicionamiento es de gran importancia para el tipo de sistema hiperespectral a emplear. Las fuentes pueden componerse de lámparas halógenas, diodos LED y lásers.

- **Dispositivo de dispersión:** Son instrumentos ópticos y electroópticos utilizados para generar la dispersión de la longitud de onda de la luz, en diferentes bandas. Estos dispositivos pueden ser: espectrógrafos de imágenes, ruedas de filtros, filtros sintonizables, espectrómetros de imágenes por Transformada de Fourier, cámaras de un disparo, etc.

- **Detectores de área:** El detector de área cuantifica la intensidad de luz recogida, mediante la conversión de la energía de esta radiación, en señales eléctricas. Los principales son: detectores de estado sólido, y semiconductor complementario de óxido metálico. Es posible superponer detectores para mejorar la sensibilidad en regiones de longitud de onda.

- **Software:** Para controlar las etapas de adquisición, procesamiento y análisis de imágenes hiperespectrales, se hace necesario contar con un software que facilite este proceso. Uno de los principales softwares para procesamiento es ENVI, que está orientado principalmente para misiones de teledetección hiperespectral. Es, un software completo para procesamiento y análisis de imágenes, con herramientas para gestionar datos hiperespectrales, con detección e identificación espectral.

3.6 Ventajas y desventajas en el uso de Imágenes Hiperespectrales

Podemos mencionar las siguientes ventajas en el uso de esta tecnología de Imágenes hiperespectrales:

- Es un método “no invasivo” y “no destructivo”.
- No afecta ni genera contaminación ambiental (evaluación libre de químicos).
- Se requiere de una leve preparación de muestra.
- Comparado con otros métodos de inspección, este es más económico.
- Se puede obtener información cualitativa y cuantitativa.
- Nos permite identificar componentes bioquímicos presentes en la muestra a través de su firma espectral.
- A partir de una firma hiperespectral, puede seleccionarse las bandas más eficientes para el caso evaluado, y en adelante, trabajar con firmas espectrales, que suelen ser más económicas y menos pesadas en información.

Dentro de las limitaciones que ofrece este sistema Hiperespectral, podemos mencionar:

- El paquete de datos obtenido por la firma hiperespectral suele disponer de mucha información, y normalmente contiene redundancia. La información suele ser tan grande que el tiempo de proceso normalmente es elevado.
- La inversión en el sistema hiperespectral es alta en la etapa de adquisición e implementación.
- La visión hiperesepectral es un método indirecto, requiere de calibración y procesos de transferencia de datos.
- El software comercial especializado ENVI, fue desarrollado principalmente para teledetección.

3.7 Aplicaciones

La detección hiperespectral, aplicada para detección, clasificación y cuantificación, puede aplicarse a diversas necesidades del mercado. Algunas de ellas que podemos citar son:

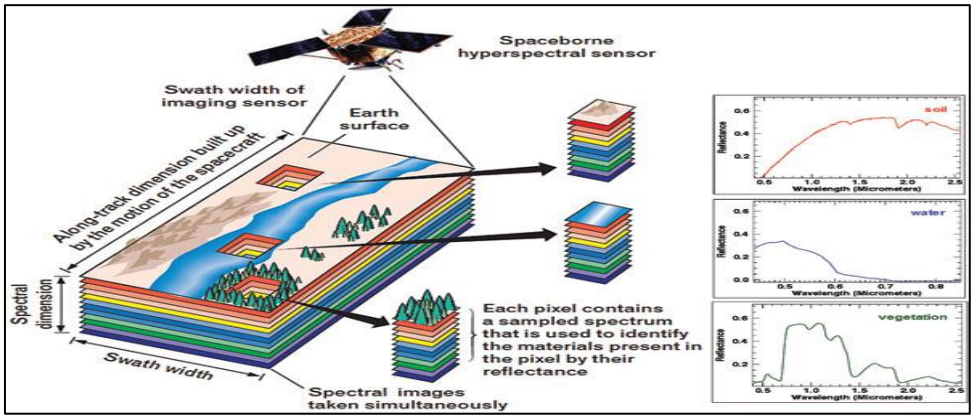
a. Teledetección. Fue la aplicación principal para el desarrollo del sistema hiperespectral.
b. Ciencia Forense. Para evaluación de escenas de crimen de manera no invasiva.
c. Medicina. Con el desarrollo de métodos no invasivos para identificación de enfermedades.
d. Bioquímica. Optimiza y acelera el diseño de fármacos.
e. Industrial Alimentaria. En los últimos años, ha sido una industria que ha desarrollado esta metodología de análisis y detección. Entre las aplicaciones se puede destacar: determinación de contenido de metales, detección de enfermedades, evaluación de acidez, entre otros.

f. Industria minera y petrolera. En la exploración de posibles zonas de extracción de minerales o petróleo.

g. Ambiental. La detección hiperespectral puede utilizarse en el monitoreo de emisiones de gases de efecto invernadero, medición de calidad del agua y sus fuentes, además de la identificación de recursos naturales e impacto ambiental.

h. Astronomía. Esta industria está desarrollándose con el objetivo de enviar satélites con dispositivos hiperespectrales para misiones de investigación y seguridad

Figura 47. Aplicación de satélites con visión hiperespectral.



Fuente. Telematica, 2019.

Figura 48. a) Ejemplificación de aplicación agrícola b) Proyectos de identificación en astronomía c) Detección de enfermedades agrícolas con imágenes multiespectrales.

Fuente: a) Isamil N, 2017 b) Telematica, 2019 c) Innovatione, 2019.

3.8 Uso de imágenes Hiperespectrales en producción de cacao

En el capítulo 1 de la presente tesis, fue presentado el problema de la contaminación de plantaciones de cacao con cadmio, que, pese a la búsqueda de su mitigación, requiere ser controlado para que la exportación de este producto al mercado europeo siga siendo viable.

En los siguientes puntos, se desarrollará la metodología propuesta para la detección de cadmio en los frutos de cacao, con el empleo de una cámara hiperespectral del Laboratorio de Automática y Control de la Universidad de Piura

3.8.1 Cámara Hiperespectral de Universidad de Piura

En las instalaciones del Laboratorio de Automática y Control de la Universidad de Piura, se cuenta con un Sistema de Adquisición de Imágenes Hiperespectrales de la marca Resonon, modelo Pika II, y con una visión hiperespectral que trabaja en los rangos de espectros en el rango de 400 a 900 nm.

Resonon es una empresa estadounidense fundada en 2002, que está ubicada en Bozeman, Montana y se dedica al suministro completo de los sistemas de imágenes hiperespectrales, software y hardware para este proceso. El sistema de adquisición de Imágenes Hiperespectrales consta de:

* Cámara de imágenes hiperespectrales.Pika II (400 a 900 nm)

* Sistema de iluminación con 4 lámparas halógenas (de cuarzo).

* Fuente de alimentación (power supply).

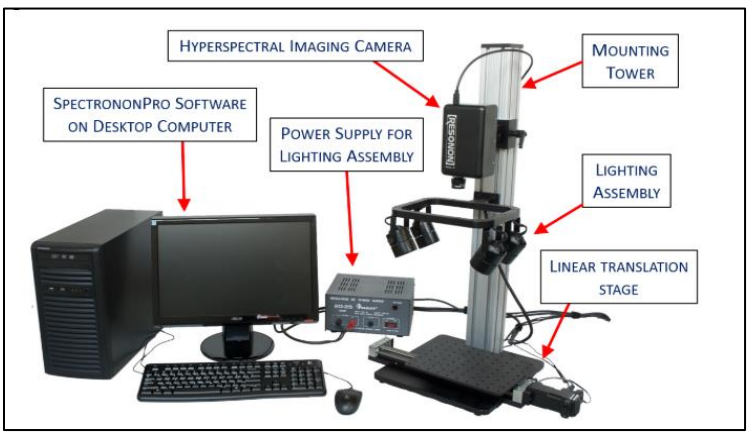
* Plataforma de desplazamiento de muestra.

* Torre de montaje de aluminio.

* Computador con software SpectrononPro, para adquisición y procesamiento de datos.

Estos elementos pueden visualizarse en la figura 50.

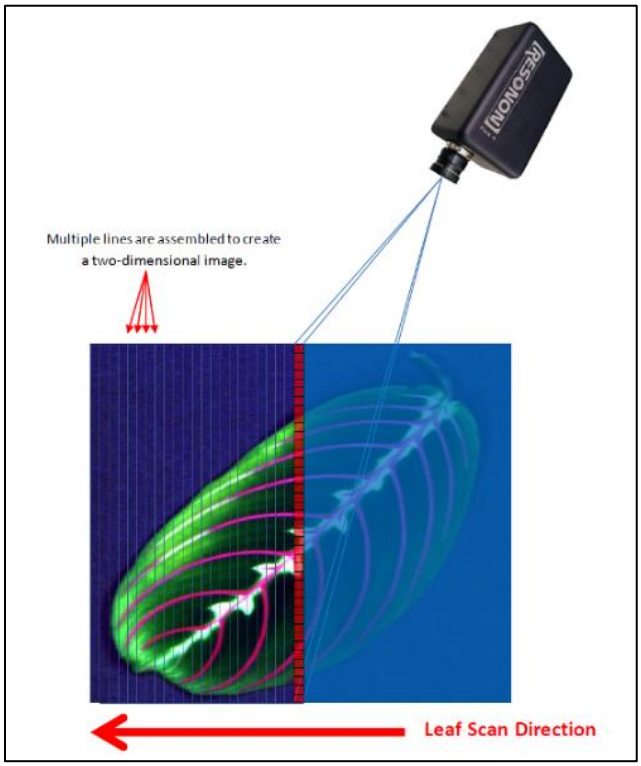
Figura 49. Esquemático del sistema de imágenes hiperespectrales.



Fuente: SpectrononPro, 2020.

La forma de escaneo de esta cámara hiperespectral es lineal, en la figura 50 podemos visualizar un esquema de este proceso.

Figura 50. Escaneo en línea de cámara Resonon.



Fuente: SpectrononPro, 2020.

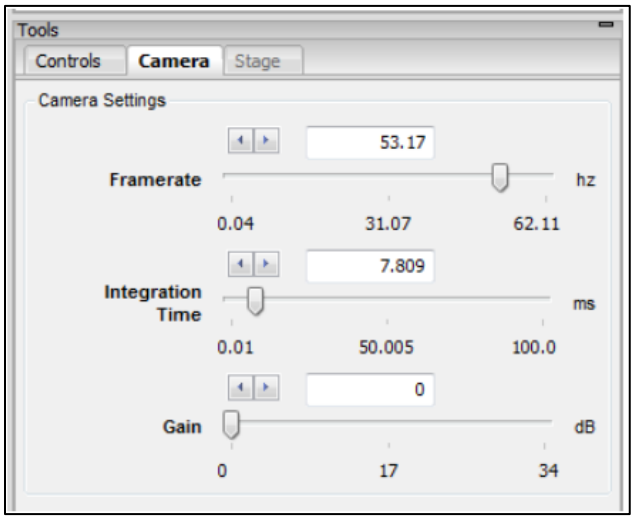
3.8.1.1 Calibración del equipo e inicio del sistema.

Resonon, a través de su manual de usuario, brinda la información básica para una correcta calibración del Sistema.

Lo primero a ejecutarse es conectar la cámara y sistema de escaneo al computador, que debe tener instalado el software Spectronon Pro, para poder acceder a la interfaz del programa.

a. **Controles de cámara.** Que pueden ser modificados y controlados en el panel de herramientas.

Figura 51. Controles de cámara.

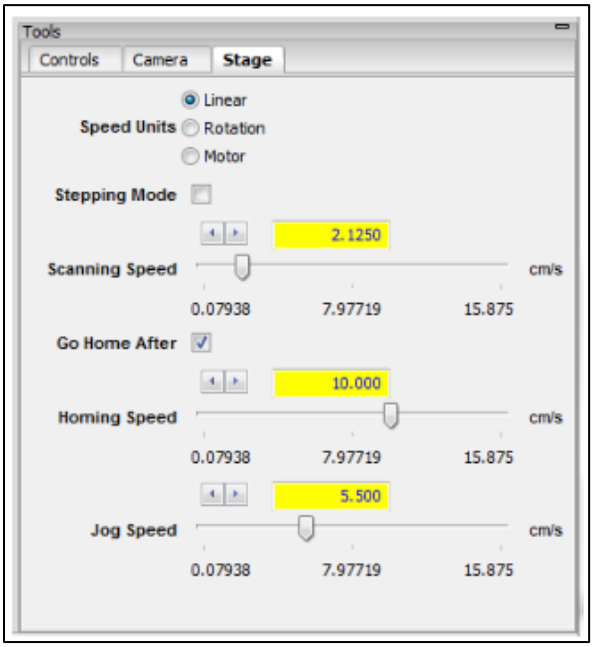


Fuente: SpectrononPro, 2020.

- Frame rate. Es el número de imágenes adquiridas por segundo.
- Integration time. O tiempo de exposición, es la duración de adquisición de datos para cada línea individual de imagen.
- Gain. Es un factor de incremento de señal, pero que genera ruido sobre la señal.

b. Controles de escenario. El control del escenario puede realizarse manualmente haciendo clic en los botones de la barra de herramientas, o realizarlo de manera más profunda desde el panel de control.

Figura 52. Controles de escenario.



Fuente: SpectrononPro, 2020.

- **Speed Units.** Permite seleccionar el tipo de escenario, siendo: ***Linear***, para traslación estándar en la mayoría de sistemas, ***Rotation***, usada para escaneo de escenario en rotación (aplicaciones exteriores normalmente), ***Motor***, usado para mostrar la velocidad en pulsos de motor por segundo.

- **Stepping Mode.** Este modo de pasos controla la forma en la que el escenario se mueve en relación al generador de imágenes. Pueden ejecutarse continuamente durante la exploración, o de manera incremental, así, se adquiere una imagen cuando el escenario está paralizado, pero en cada toma, el escenario se ha desplazado incrementalmente.

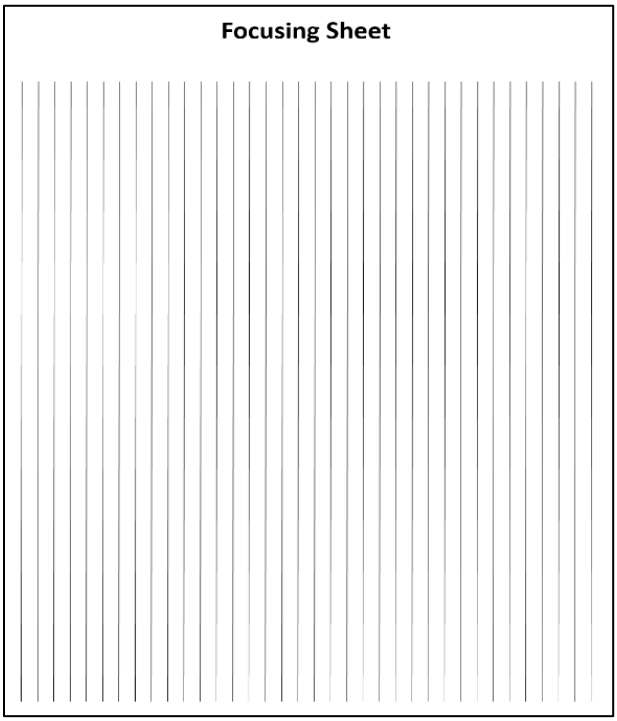
- **Scanning Speed.** Es la velocidad lineal del escenario durante el escaneo.

- **Homing Speed.** Es la velocidad en la que el escenario retorna a su punto inicial.

- **Jog Speed.** Es la velocidad con la que se puede desplazar manualmente el escenario.

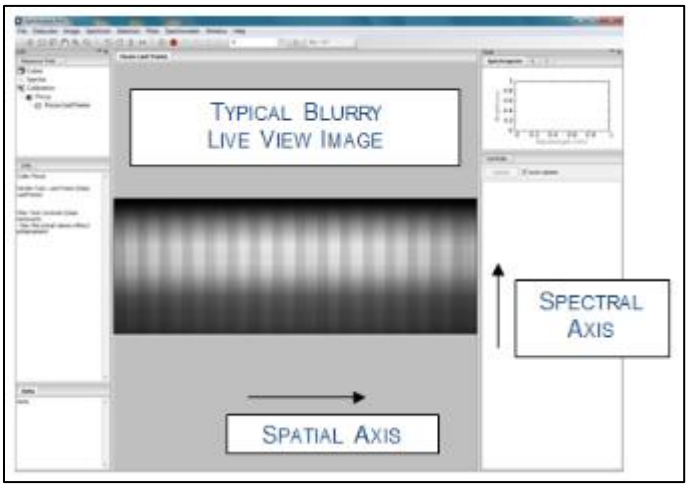
c. Enfoque de lente de cámara. Utilizando la herramienta de Enfoque en Spectronon, se podrá visualizar la imagen en vivo a través de la cámara. Para la calibración, es necesario utilizar una plantilla con líneas oscuras, como la mostrada en la figura 54. Antes de calibrarse, se visualizará una imagen como la de la figura 55, donde un eje de la imagen representa al eje espacial del objeto, y el otro representa al eje espectral (longitud de onda).

Figura 53. Plantilla para calibración de lente.



Fuente. SpectrononPro, 2020.

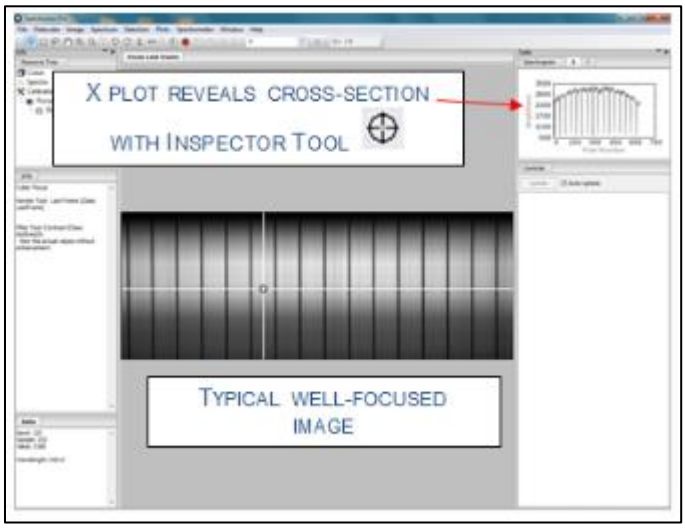
Figura 54. Eje espacial y espectral.



Fuente: SpectrononPro, 2020.

Para ajustar el enfoque, deberá desbloquear el ajuste. Luego, rotar el lente objetivo hasta que las líneas oscuras sean visibles, buscando maximizar la nitidez de estas líneas, como se puede ver en la siguiente figura.

Figura 55. Ajuste del enfoque del lente.



Fuente: SpectrononPro, 2020.

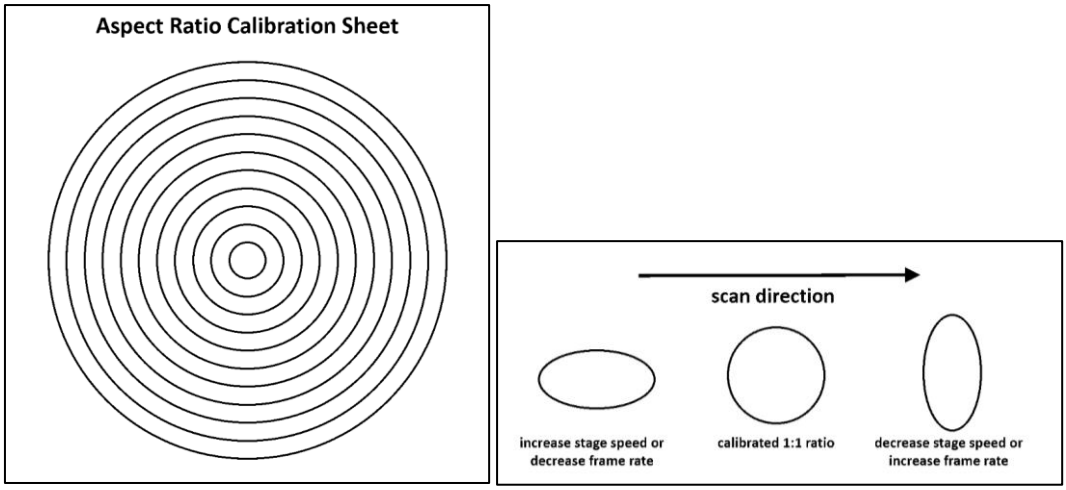
d. Calibración de imágenes. Se mostrará cómo definir el sistema para escanear la Reflectancia, escalada a un objeto de referencia.

- Remover la corriente oscura (*Remove dark current*). Existe una corriente residual en el dispositivo fotoeléctrico cuando no existe iluminación incidente. Esta corriente (*Dark current*), es eliminada con una herramienta del Software SpectrononPro. El programa indicará tapar el lente (*lens cap*) hasta que el indicador se torne rojo, luego puede retirarse la tapa negra para habilitar el lente.

- Definir la reflectancia de referencia. La medición de la reflectancia requiere corrección para tener en consideración los efectos de la iluminación. Siendo así, necesita también de material de referencia que sea uniforme, como el teflón blanco en láminas o Spectralon. Una vez se complete esta verificación y se tome en consideración la referencia, el indicador marcará listo y el sistema estará calibrado para reflectancia.

- Ajuste en la relación de aspecto. Para ajustar la relación de aspecto del escaneo, suele ser útil primero, crear una imagen que permita identificar distorsión fácilmente. Debe usarse la hoja de calibración de la relación de aspecto de píxeles del manual SpectrononPro y grabarse un escaneo con líneas suficientes para ver el círculo completo. Si la imagen está distorsionada en la dirección de escaneo, debe cambiarse la velocidad de escaneo, si la imagen se alarga en la dirección de escaneo, se requiere incrementar la velocidad de escaneo, y si la imagen se compacta en la dirección de escaneo, se debería disminuir la velocidad de escaneo.

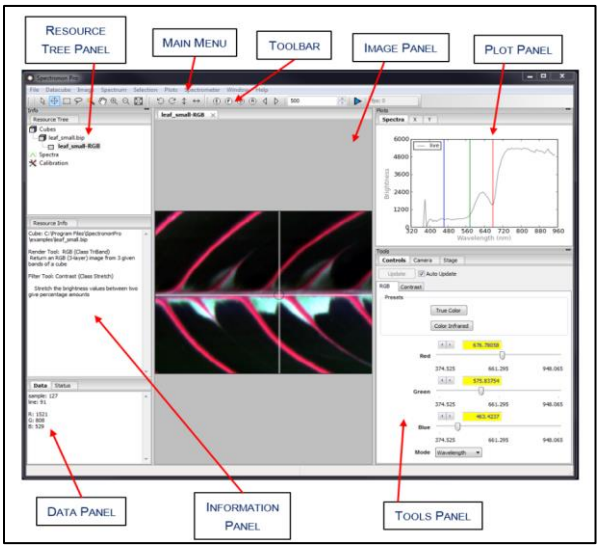
Figura 56. Plantilla de calibración de aspecto.



Fuente: SpectrononPro, 2020.

3.8.1.2 Software de adquisición de datos. El Software disponible para usar con la Cámara Hiperespectral de la Universidad de Piura, es el SpectrononPro, de RESONON. Este software es utilizado para el control de la cámara hiperespectral, para la etapa de calibración de la cámara, y también para usar las herramientas para procesamiento y análisis de imágenes hiperespectrales adquiridas por la cámara.

Figura 57. Software Spectronon Pro.



Fuente: SpectrononPro, 2020.

3.9 Aplicación de Imágenes Hiperespectrales en vegetación

Las imágenes hiperespectrales se han desarrollado rápidamente como herramienta para la evaluación no destructiva de productos agroindustruales y vegetación, debido a su buena correlación con parámetros como densidad, biomasa, estrés, sanidad vegetal, concentración de pigmentos, entre otros.

Si consideramos los índices de vegetación, gran número están basados en el contraste existentes entre la banda del Rojo y la del Infrarojo Cercano.

El índice más utilizado en la estimación del contenido de clorofila es el NDVI, y es el indicador más aceptable para conocer la madurez del fruto. Este índice queda definido según:

$$NDVI = \frac{NIR - Red}{NIR + Red} \tag{37}$$

Donde:
NIR: Reflactancia del infrarrojo cercano
RED: Reflactancia del infrarojo

Los índices Hiperespectrales usados en vegetación (HVI's: *Hyperspectral Vegetation Indices*), son una herramienta utilizada para análisis de datos hiperespectrales de cultivos agrícolas. Estos índices están formulados por combinaciones aritméticas de reflectancia espectral en determinadas longitudes de onda.

En la siguiente tabla, se muestran los índices hiperespectrales para vegetación y su ecuación en el software empleado: Spectronon Pro.

Tabla 10. *Cálculo de índices hiperespectrales para vegetación.*

Índice Hiperespectral de Vegetación	Ecuación SpectrononPro
Índice de Reflectancia de Antocianina 1	$ARI1 = \frac{1}{\rho_{550}} - \frac{1}{\rho_{700}}$
Índice de Reflectancia de Antocianina 2	$ARI2 = \rho_{800}\left(\frac{1}{\rho_{550}} - \frac{1}{\rho_{700}}\right)$
Índice de Vegetación Resistente a la Atmósfera	$ARVI = \frac{NIR - (Red - \gamma(Blue - Red))}{NIR + (Red - \gamma(Blue - Red))}$
Índice de Reflectancia de Carotenoide 1	$CRI2 = \frac{1}{\rho_{510}} - \frac{1}{\rho_{550}}$
Índice de Reflectancia de Carotenoide 2	$CRI2 = \frac{1}{\rho_{510}} - \frac{1}{\rho_{700}}$
Índice de Vegetación Mejorado	$EVI = \frac{NIR - Red}{NIR + 6.Red - 7.5.Blue + 1}$
Índice Modificado de Reflectancia de Absorción de Clorofila	$MCARI = \rho_{700} - \rho_{670} - 0.2.(\rho_{700} - \rho_{550}).(\frac{\rho_{700}}{\rho_{670}})$
Índice de Vegetación Normalizado Modificado de Borde Rojo	$MRENDVI = \frac{\rho_{750} - \rho_{705}}{\rho_{750} + \rho_{705} - 2.\rho_{445}}$
Índice Modificado de Relación Simple de Borde Rojo	$MRESR = \frac{\rho_{750} - \rho_{445}}{\rho_{705} - \rho_{445}}$
Índice de Vegetación de Diferencia Normalizada	$NDVI = \frac{NIR - Red}{NIR + Red}$
Índice de Reflectancia Fotoquímica	$PRI = \frac{\rho_{531} - \rho_{570}}{\rho_{531} + \rho_{570}}$
Índice de Reflectancia de la Senescencia de la Planta	$PRI = \frac{\rho_{680} - \rho_{500}}{\rho_{750}}$
Índice de Vegetación de Diferencia Normalizada de Borde Rojo	$RENDVI = \frac{\rho_{750} - \rho_{705}}{\rho_{750} + \rho_{705}}$
Índice de Relación Simple	$SR = \frac{NIR}{Red}$
Índice de Pigmentación Insensible a la Estructura	$SIPI = \frac{\rho_{800} - \rho_{445}}{\rho_{800} + \rho_{680}}$
Índice Transformado de Reflectancia de Absorción de Clorofila	$TCARI = 3\left[\rho_{700} - \rho_{670} - 0.2.(\rho_{700} - \rho_{550})..(\frac{\rho_{700}}{\rho_{670}})\right]$
Índice Vogelmann Borde Rojo 1	$VREI = \frac{\rho_{740}}{\rho_{720}}$
Índice Vogelmann Borde Rojo 2	$VREI2 = \frac{\rho_{734} - \rho_{747}}{\rho_{715} - \rho_{726}}$
Índice Vogelmann Borde Rojo 3	$VREI3 = \frac{\rho_{734} - \rho_{747}}{\rho_{715} + \rho_{720}}$
Índice de Banda de Agua	$WBI = \frac{\rho_{970}}{\rho_{900}}$

Fuente: Elaboración propia.

Visualizamos en los índices, que hacen referencia a Antocianina y Carotenoide, por lo que se detallará brevemente a qué refieren estos componentes.

- **Antocianina:** Las antocianinas son compuestos fenólicos sintetizados por las plantas, que se acumulan en los órganos vegetativos a manera de respuesta a la radiación ultravioleta. Las antocianinas son las responsables por las coloraciones roja, azul y púrpura de algunas flores, frutos, hojas y raíces de plantas (Aceituno, 2010; Viera, 2018). La estructura de las antocianinas, acidez del medio, y presencia de metales afecta su coloración.

El consumo de antocianina presenta beneficios por su capacidad antioxidante y quimiopreventiva, pues disminuye el riesgo de padecimiento de enfermedades crónicas, inhibe el crecimiento de células cancerígenas y previene enfermedades cardíacas, es por ello, que tienen gran aplicación en productos de la industria farmacéutica y alimentaria.

- **Carotenoides** Los carotenoides son pigmentos orgánicos que se encuentran en plantas y algunas clases de hongos y bacterias. Su coloración va desde el amarillo hasta el rojo oscuro según la estructura. En organismos fotosintéticos, los carotenoides participan en el proceso de transferencia de energía o evitando la auto oxidación del centro de reacción, y en organismos no fotosintéticos, funcionan como mecanismo de protección de la oxidación.

3.10 Principales Índices Hiperespectrales en vegetación

Para la evaluación no destructiva del contenido de clorofila de hojas y frutos, se han diseñado modelos conceptuales utilizados exitosamente en análisis no destructivo de contenido de carotenoides y antocianina. Los principales son:

A. ARI. Índice específico para determinar el contenido de Antocianina de una muestra. Su cálculo es basado en los índices de reflectancia en las longitudes de onda 550 y 700nm.

B. CRI. Índice que permite estimar el contenido total de Carotenoides de una muestra a partir de tomar en cuenta los índices de reflectancia en las longitudes de onda 510, 550 y 700nm. El índice CRI no es aplicable en tejidos vegetales que contienen antocianina (Merzlyak, 2003b).

C. PRI. Índice que correlaciona el estado de epoxidación de los pigmentos del ciclo de xantofila, eficacia de la fotosíntesis y estrés del Nitrógeno de los doseles. Su cálculo toma en consideración los índices de reflectancia a 531 y 570nm.

Capítulo 4

Resultados experimentales

4.1 Prefacio

En el presente capítulo se expondrán los resultados experimentales obtenidos a partir de la toma de muestras de campo, toma de imágenes hiperespectrales, y la programación de códigos que permiten usar *Machine Learning* para la predicción del contenido de Cadmio a partir de los resultados espectrales de su muestreo.

El objetivo de este control será brindar herramientas que faciliten la detección del Contenido de Cadmio en las plantaciones de Cacao en tiempo real, y se pueda tomar acción sobre la marcha, para evitar que el producto de Cacao quede contaminado con Cadmio al final del proceso productivo.

La programación en *Machine Learning* se ha realizado en el Software Matlab, que presenta una variedad de funciones que permiten realizar códigos de programa para diversas metodologías de predicción.

Las muestras han sido tomadas en diversas ubicaciones de la región, para poder tener una correlación respecto a las pruebas.

4.2 Imágenes Hiperespectrales de muestras de cacao

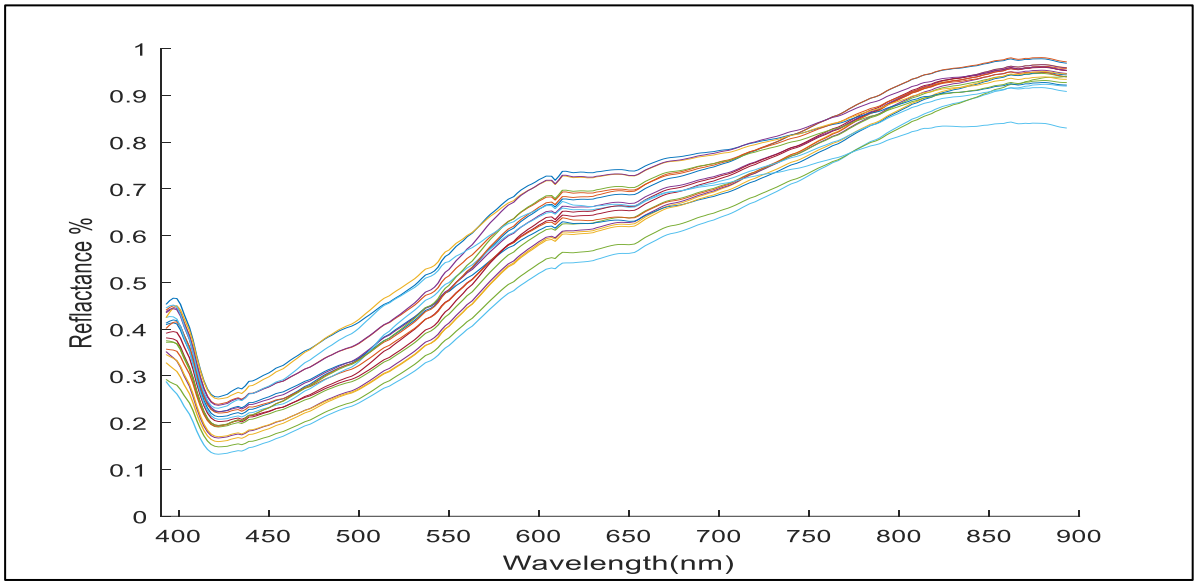
Las muestras empleadas para la obtención de las imágenes hiperespectrales fueron tomadas en distintas regiones, como:

- Platanal Bajo
- Las Lomas
- Buenos Aires
- Tambogrande

Se puede visualizar en los resultados de la firma hiperespectral de la figura 58. En ella se distingue que la forma de la firma espectral muestra un comportamiento muy similar, y cada una corresponde a un nivel de Cadmio variable, y a una región de obtención distinta.

Cada firma hiperespectral, tiene 240 lecturas de reflectancia en el rango de 400 y 900 nm.

Figura 58. Imágenes hiperespectrales de muestras.



Fuente. Elaboración propia.

A partir de esta firma hiperespectral, se aplicarán diversos métodos de predicción de contenido de cadmio para identificar la que mejor representa nuestro problema.

4.3 Regressión Lineal Multivariable

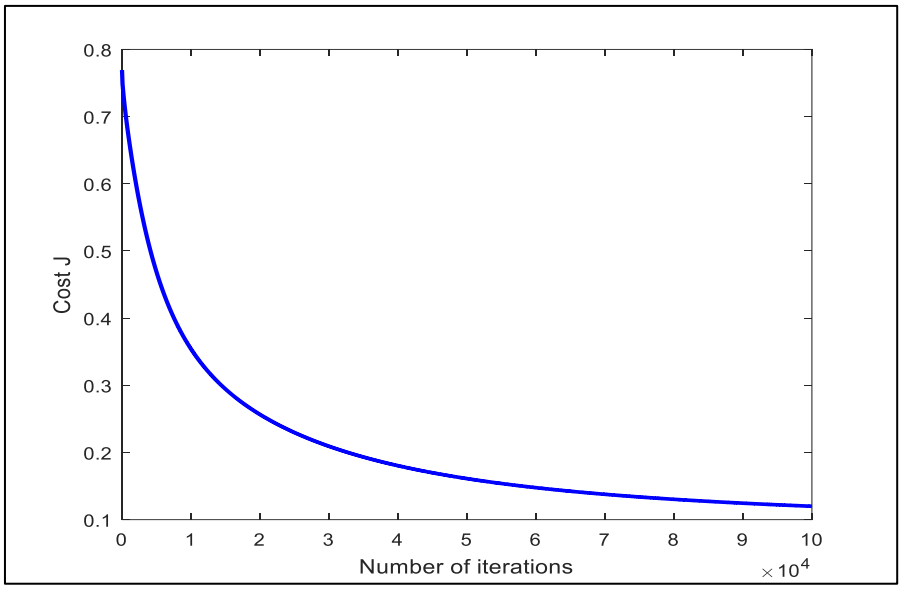
El primer método de predicción será la Regresión Lineal Multivariable, en la que usaremos *Machine Learning* para diseñar un modelo lineal de múltiples variables, que permitirá predecir el contenido de Cadmio a partir de una firma hiperespectral como dato de ingreso.

En la primera parte realizaremos la modelación con todas las características de la firma, y más adelante, se comparará los resultados como un muestreo filtrado, con menor cantidad de características y menor peso de data a procesar (menor cantidad de datos).

La matriz de características de datos de entrada será la matriz “X”, y los valores de Cadmio serán la salida Real “y” a predecir.

Así, podremos identificar la gráfica de función de Costo y cómo va disminuyendo el error a mayor cantidad de iteraciones. Podemos visualizar que debido a la cantidad de características consideradas como entrada (240 elementos), se requiere altas iteraciones para la convergencia.

Figura 59. Función de costo para la matriz de entrada completa (alpha=0.01).

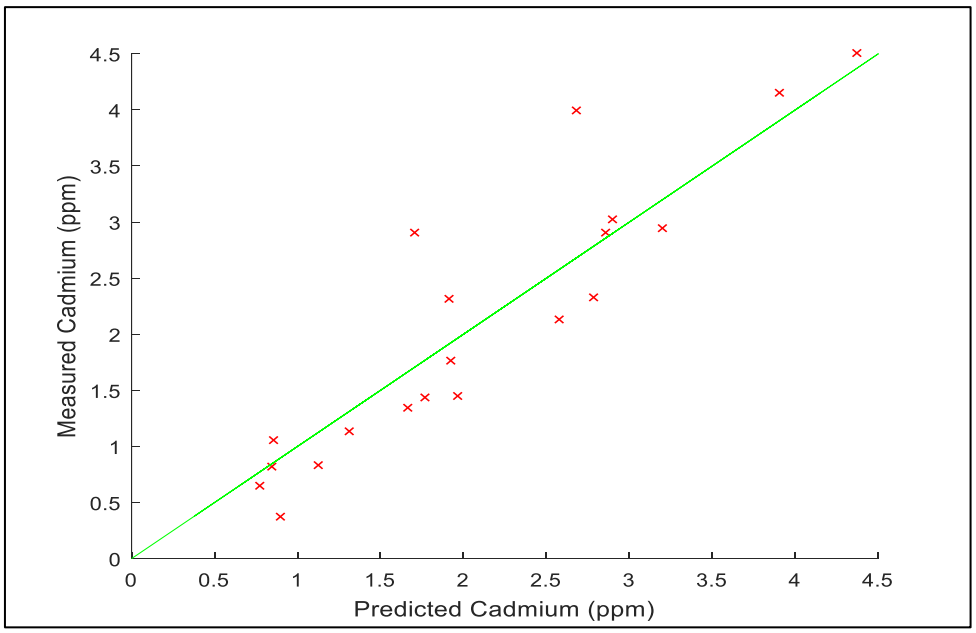


Fuente: Elaboración propia.

Con el modelo de predicción, podemos evaluar el funcionamiento o aprendizaje de esta modelación por regresión, probando las salidas que se obtendrían con las entradas hiperespectrales.

En la figura 60 podemos visualizar gráficamente la diferencia existente entre la salida predicha y la real. Mientras los puntos identificados se encuentren más cerca de la función identidad, mejor será la predicción realizada.

Figura 60. Diferencia entre salidas real y predicha.



Fuente: Elaboración propia.

De igual manera, es posible calcular el error utilizando la fórmula del Error cuadrático medio (MSE), y también el RMSE (raíz del error cuadrático medio), definidas por:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_i)^2 \tag{38}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{39}$$

Donde:

y_i: Valor obtenido del modelo de predicción.

$\bar{y}_i$: Valor real en el punto de predicción.

Aplicando ambas fórmulas aplicadas al cálculo del error cuadrático, obtenemos:

MSE: 0.2399

RMSE: 0.4898

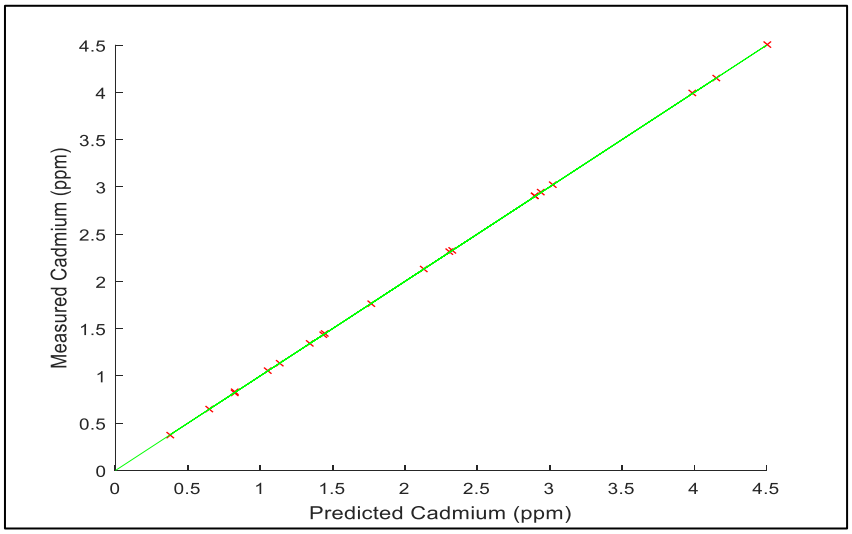
La otra manera de definir el modelo de regresión lineal es usando la Ecuación Normal, definida por:

$$\theta = (X^T X)^{-1} X^T \vec{y} \tag{40}$$

Donde θ será el vector de resultados de la predicción.

Utilizando la Ecuación Normal, se obtiene el resultado mostrado en la figura 61, debido a que el cálculo fue formulado de manera directa. Sin embargo, es necesario evaluar también, qué tan eficiente es el modelo para entradas diferentes.

Figura 61. Diferencia entre predicción de Ecuación Normal y salida real.



Fuente: Elaboración propia.

De igual manera, aplicando las ecuaciones para definir el error en los resultados, obtenemos lo siguiente:

MSE = 4.5366e-18

RMSE = 2.1299e-09

Como podemos visualizar en la figura 62, las salidas predichas son prácticamente las mismas que las reales. Sin embargo, esto no garantiza que el modelo sea preciso con situaciones diferentes a las del entrenamiento. Al generar un modelo complejo, al buscar predicciones de valores diferentes a los del entrenamiento, el error podría ser muy grande.

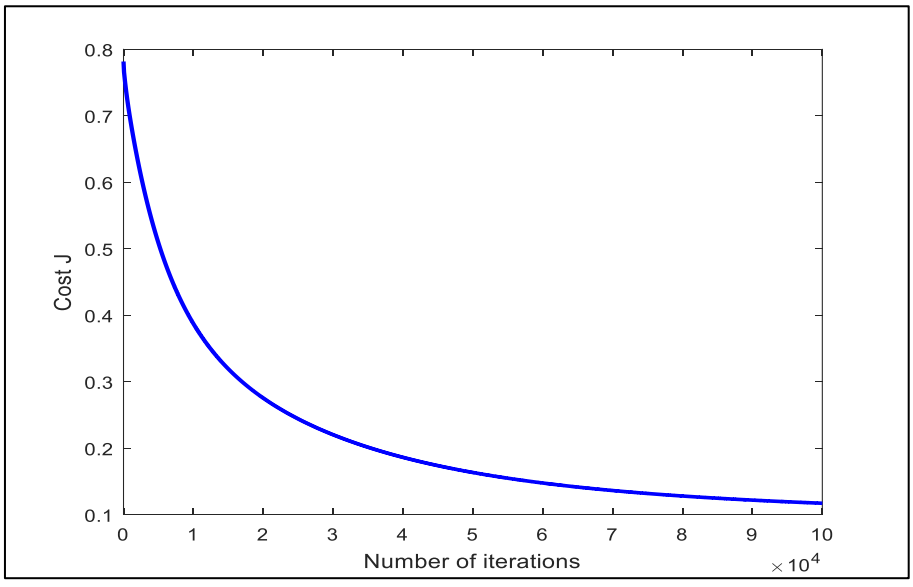
4.3.1 Validación cruzada

En esta etapa, podremos probar los algoritmos de predicción, y compararlos contra las salidas reales. Para esta Validación cruzada, usaremos el 70% de los datos para realizar la predicción y el 30% de la data para hacer la validación.

Usando el Conjunto de entrenamiento, se obtienen los siguientes resultados:

Gráfico de la función de costo. Podemos notar de la figura 62, que la cantidad de muestras ha disminuido, y la cantidad de características se mantiene, por lo que la cantidad de iteraciones para encontrar un error bajo se mantendrá alta, con tendencia a incrementarse.

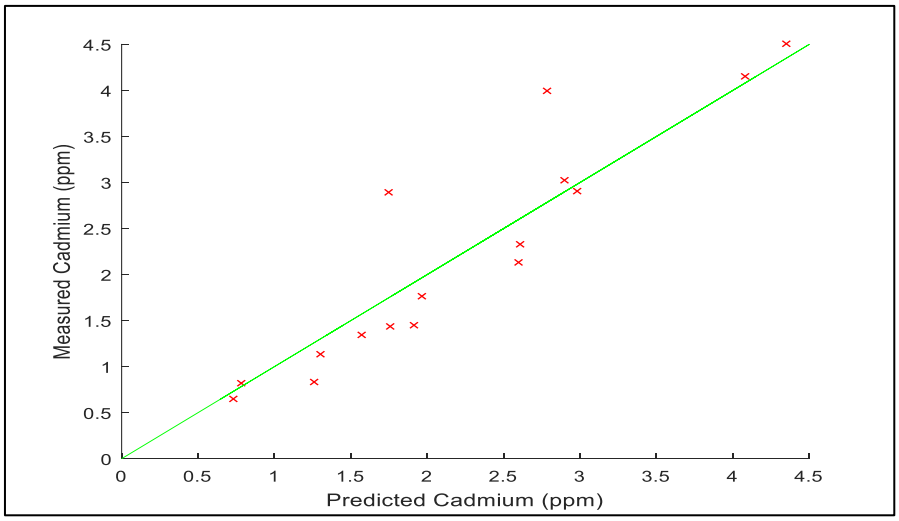
Figura 62. Función de Costo con menor cantidad de entradas.



Fuente: Elaboración propia.

De igual manera, graficaremos los resultados en la figura 63, donde podrá visualizarse la diferencia en la predicción para el Conjunto de datos de entrenamiento.

Figura 63. Predicción utilizando conjunto de entrenamiento.



Fuente: Elaboración propia.

Visualmente se observa que la predicción es bastante buena para los índices que buscamos conocer, de igual manera existen algunos puntos que quedan con mayor distanciamiento respecto del valor real.

Si calculamos el error numéricamente, obtenemos los siguientes resultados para este modelo de menor cantidad de muestras:

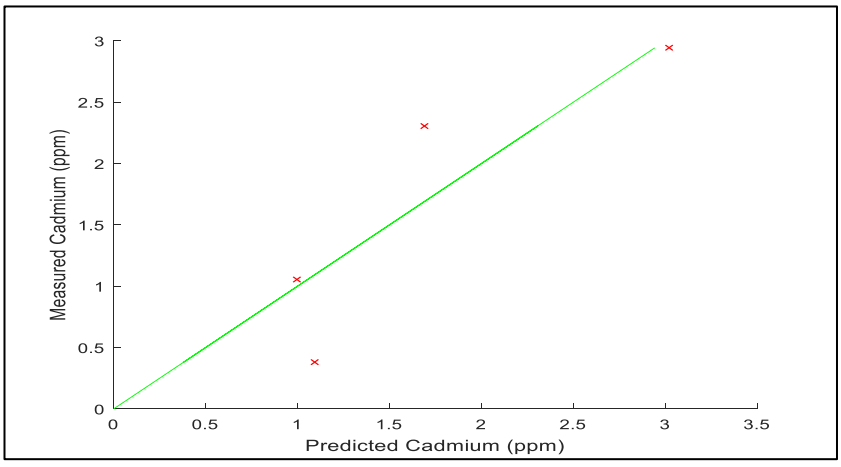
MSE = 0.2346

RMSE = 0.4844

Con estas premisas, es posible emplear el modelo obtenido para encontrar una predicción de salidas para variables de ingreso diferentes (no entrenadas). Esto permitirá hacer una evaluación pronta de qué tan bueno es el modelo obtenido.

Para los datos de validación, utilizaremos el 30% de la cantidad de datos de entrada, con lo que se obtiene el resultado gráfico de la figura 64.

Figura 64. Predicciones para conjunto de datos de validación.



Fuente: Elaboración propia.

Visualmente podemos ver que el conjunto de validación ha sido correctamente predicho, y los valores obtenidos para los errores son:

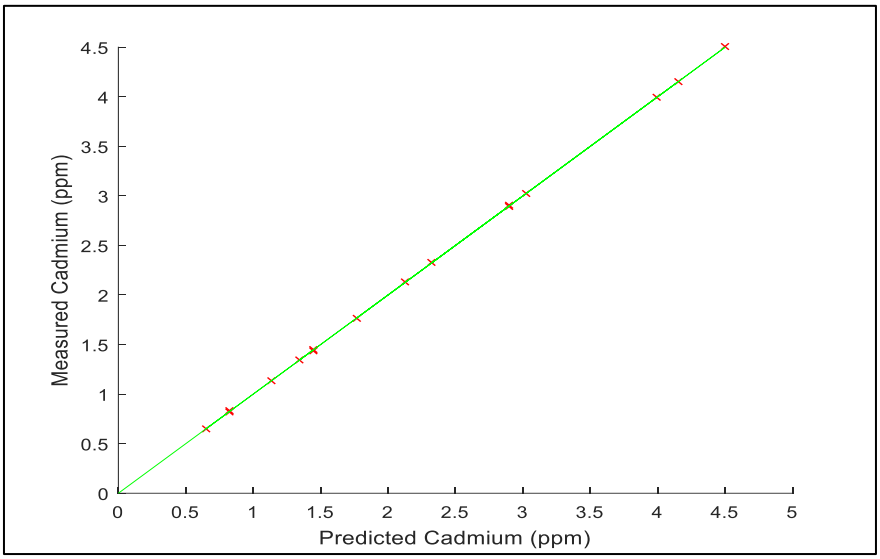
MSE = 0.2234

RMSE=0.4727

Realizaremos la misma verificación para la modelación con Ecuación Normal, y comprobaremos si la predicción es buena o no.

Utilizando la Ecuación Normal, y los datos definidos para el Conjunto de Entrenamiento, obtenemos como resultado, valores prácticamente idénticos a los reales, como se observa en la figura 65.

Figura 65. Predicción del conjunto de entrenamiento con Ecuación Normal.



Fuente: Elaboración propia.

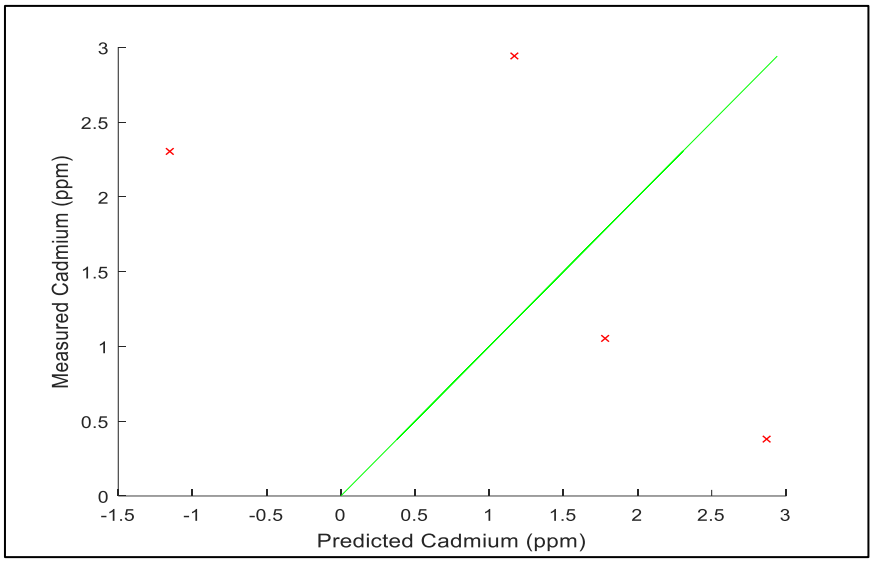
Los valores de error para este caso son prácticamente insignificantes, similar al resultado obtenido con la totalidad de los datos de entrada.

MSE= 1.50e-18

RMSE = 1.23e-09

Sin embargo, esto no garantiza que el modelo de predicción sea realmente idóneo. Para ello es necesario probar con datos diferentes a los de entrenamiento. En la figura 66 podemos visualizar los resultados obtenidos con el Conjunto de datos de Verificación.

Figura 66. Predicción de datos de Verificación con Ecuación Normal.



Fuente: Elaboración propia.

Como se puede notar, los valores de error obtenidos son visualmente altos, con diferencias muy grandes, que conlleva a un error de predicción. Las fórmulas de error nos muestran los mismos resultados, como sigue:

MSE = 5.47

RMSE = 2.34

Como podemos ver, el uso de la Ecuación Normal para este caso no aproxima correctamente nuestro modelo, o por lo menos no para la cantidad de datos de entrenamiento y características.

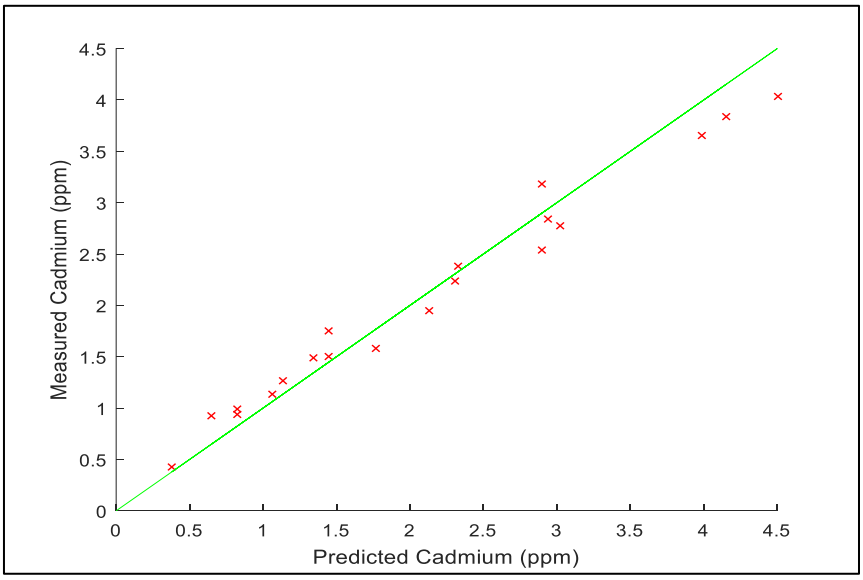
Podemos reconocer que, utilizando la Ecuación Normal para esta aproximación, no encontraremos la mejor solución, debido a su Alta Varianza.

4.4 Support Vector Machines

Con la aplicación de Machine Learning y este método de predicción de niveles de Cadmio, realizaremos las pruebas de predicción para la data de alimentación completa.

El algoritmo de predicción con SVM, permite el uso de Kernell, y para el caso específico, usaremos el Kernell Gaussiano, para facilitar la modelación. Podemos visualizar en la figura 67, el resultado de predicción con la data recolectada.

Fgigura 67. Predicción de resultados usando Gaussian Kernell.



Fuente: Elaboración propia.

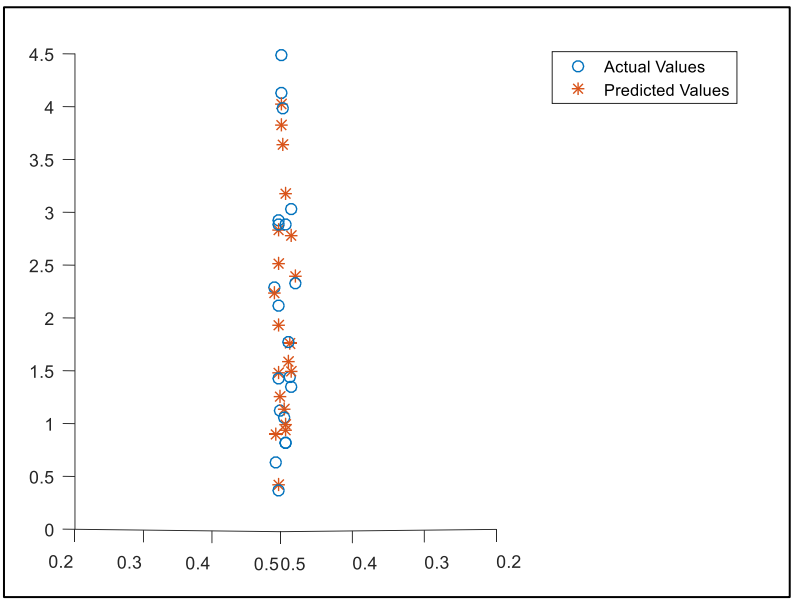
Así como se hizo en las predicciones anteriores, los valores de error para este caso, son:

MSE = 0.0531

RMSE = 0.23

Visual y numéricamente podemos notar que la aproximación es bastante buena y cercana a los valores reales. En la figura 68, podremos visualizar de manera similar, gráficamente, los resultados obtenidos comparados con los resultados reales.

Figura 68. Predicción usando Gaussian Kernell.

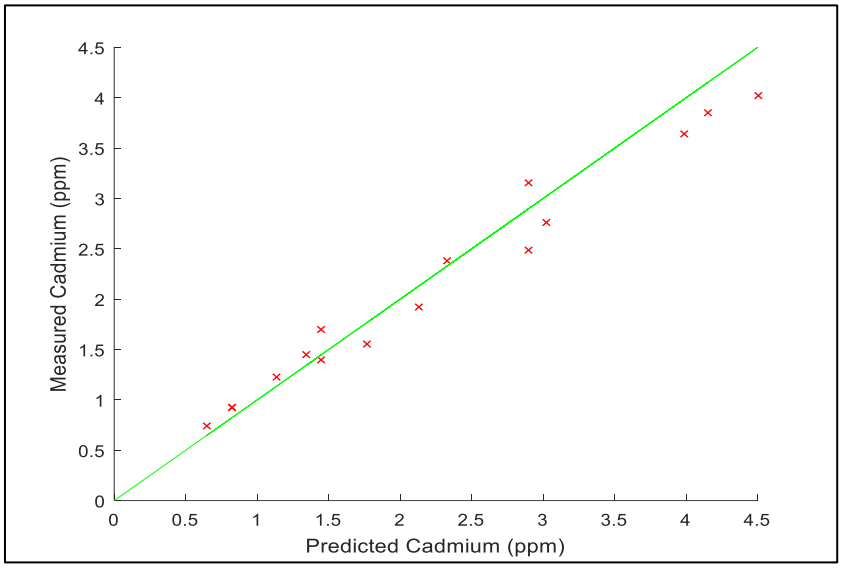


Fuente: Elaboración propia.

4.4.1 Validación cruzada para Support Vector Machines

Tal y como se probó con validación cruzada para Regresión Multivariable, evaluaremos los resultados de esta validación para SVM. Podemos verificar gráficamente (figura 69) que el resultado de la predicción para los datos de entrenamiento es visiblemente buena.

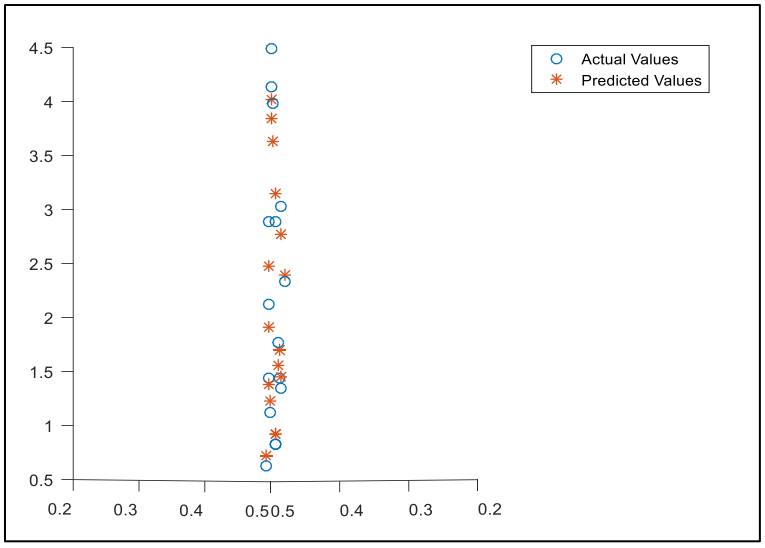
Figura 69. Predicción con datos de entrenamiento y SVM.



Fuente. Elaboración propia.

Se puede apreciar un resultado similar con el comparativo de valores predichos y valores reales de la figura 70.

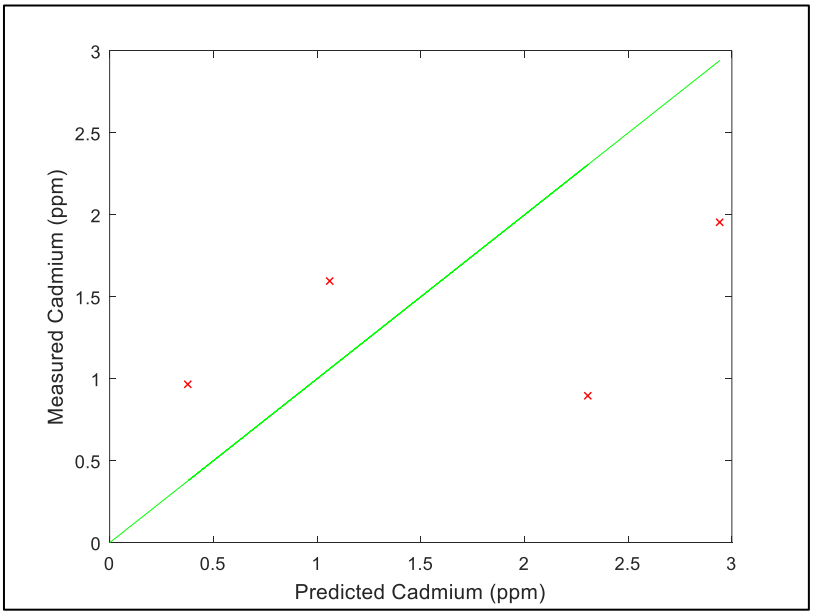
Figura 70. Predicción de valores vs. Valores reales (data entrenamiento).



Fuente. Elaboración propia.

El resultado de la validación cruzada se visualiza en la figura 71, en la que se aprecia un resultado adecuado, pero no tan bueno como el obtenido en Regresión Lineal. Esto es probablemente por la cantidad de datos de entrenamiento comparado con la cantidad de datos de validación.

Figura 71. Resultado gráfico de predicción de resultados vs. Valores reales.



Fuente. Elaboración propia.

Los errores obtenidos en la Validación Cruzada de SVM son:

MSE=0.89

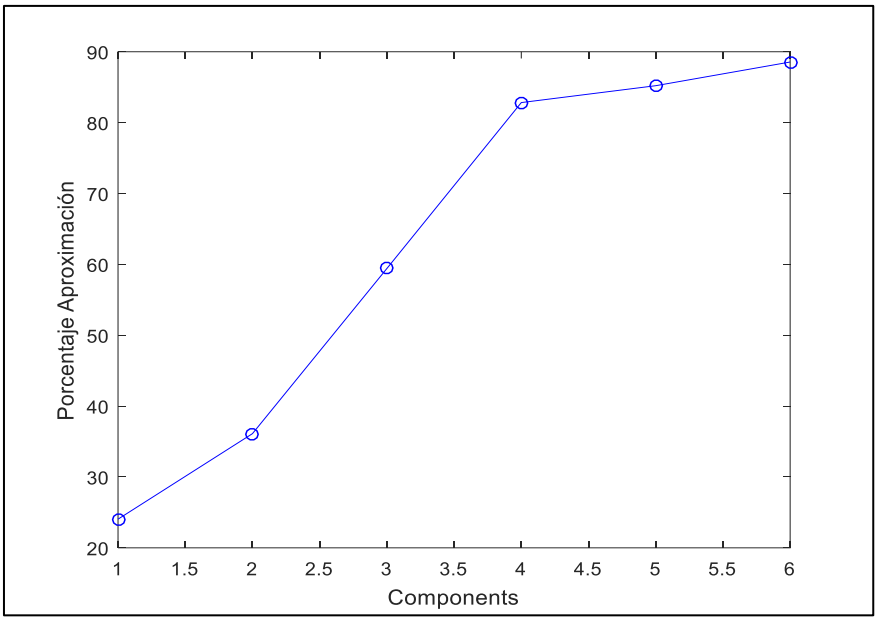
RMSE=0.94

4.5 Regresión con Partial Least Squares (PLS)

La regresión con *Partial Least Squares* mantiene una relación con *Principal Components Analysis*, pero busca un modelo de regresión lineal proyectando variables predichas y variables observables hacia un nuevo espacio, en lugar de encontrar hiperplanos.

Mostraremos a continuación, los resultados con la modelación *Machine Learning* para PLS. Debido a que el modelo nos permite evaluar el grado del polinomio a partir de los componentes, podemos evaluar un grado adecuado, que sea rápido para el procesamiento, y a la vez nos dé un porcentaje de aproximación alto. Para nuestro caso, hemos previsto dar un grado 6, y conseguimos con esto una aproximación de casi 90% como se ve en la figura 72.

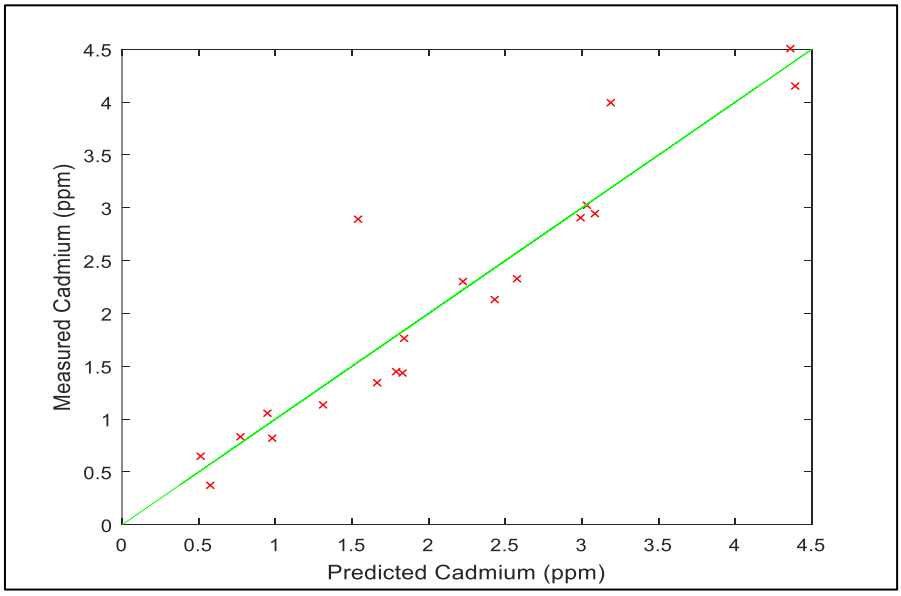
Figura 72. Nivel de predicción según los componentes.



Fuente. Elaboración propia.

Con este grado de predicción, procederemos a verificar y comparar con los resultados reales. En la figura 73 podemos ver que los valores de la predicción son bastante cercanos a los valores reales.

Figura 73. Predicción usando un grado polinómico 6.



Fuente. Elaboración propia.

Los valores de error obtenido son:

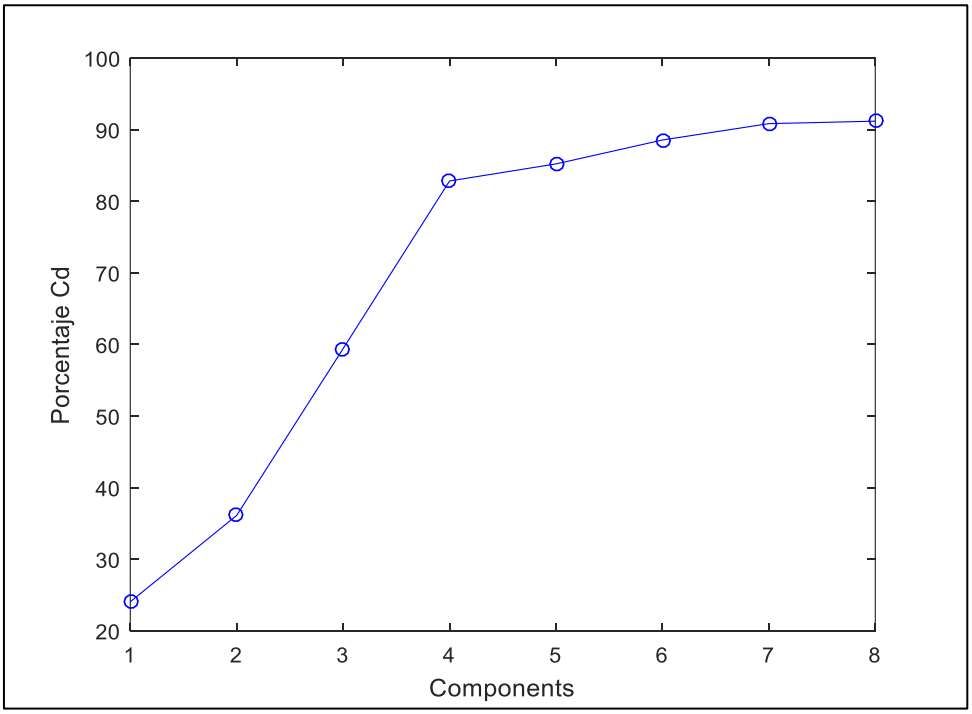
MSE = 0.1623

RMSE = 0.4028

Para comparar cómo varía el resultado según el grado polinómico escogido, mudaremos ahora para un grado 8.

En la figura 74 se muestra la variación en el nivel de predicción con un incremento de grado a 8. Se consigue aproximadamente 92% con este nuevo grado. Por lo tanto, dependiendo del nivel de precisión requerido, se debe evaluar el grado del polinomio a considerar.

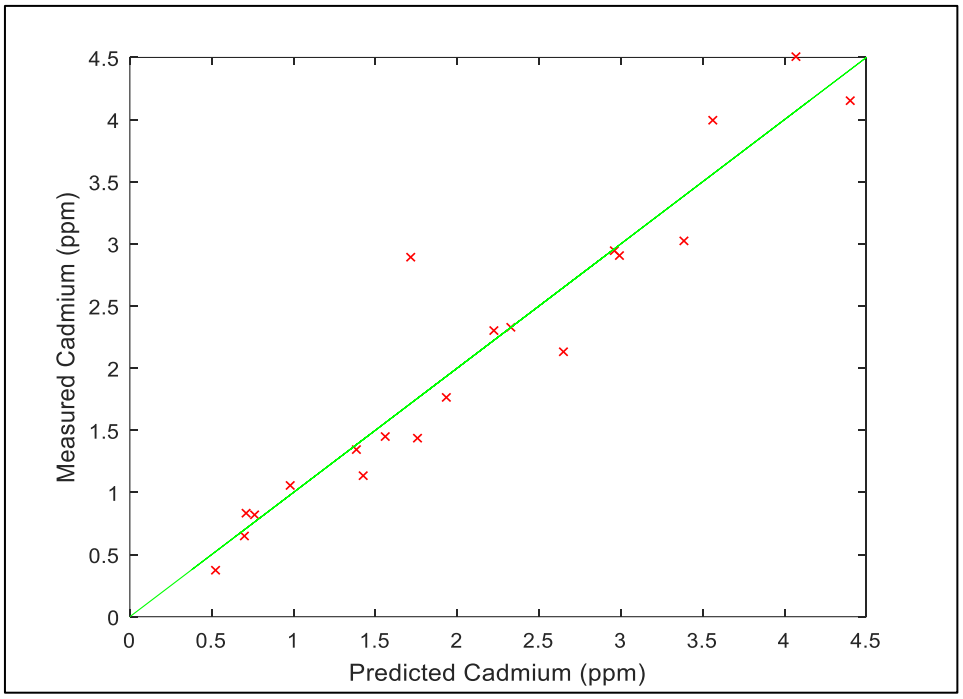
Figura 74. Nivel de predicción con grado 8.



Fuente. Elaboración propia.

Los resultados para este nuevo grado polinómico se muestran en la figura 75, donde podemos visualizar una predicción bastante semejante a los valores reales.

Figura 75. Predicción usando polinomio de grado 8.



Fuente. Elaboración propia.

Los valores de error para este algoritmo de predicción y este grado de polinomio son de:

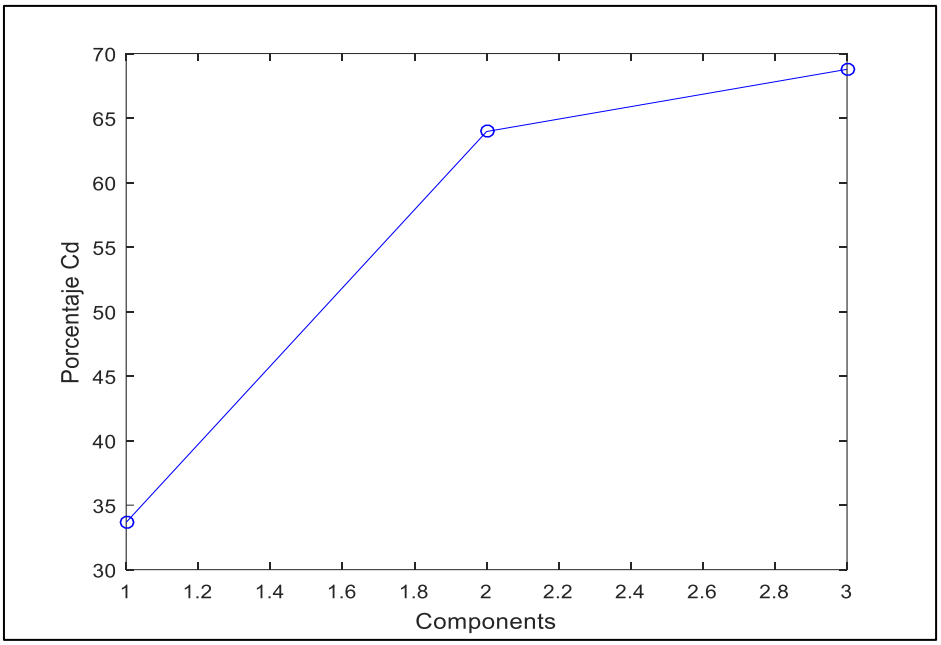
MSE = 0.125

RMSE = 0.354

4.5.1 Validación cruzada para Partial Least Squares

Para la modelación con *Partial Least Squares* se hizo también la modelación, y compararemos una validación para número de componentes 3 y otra con 4 componentes. Para 3 componentes, se visualiza una aproximación del 70% (como se visualiza en la figura 76).

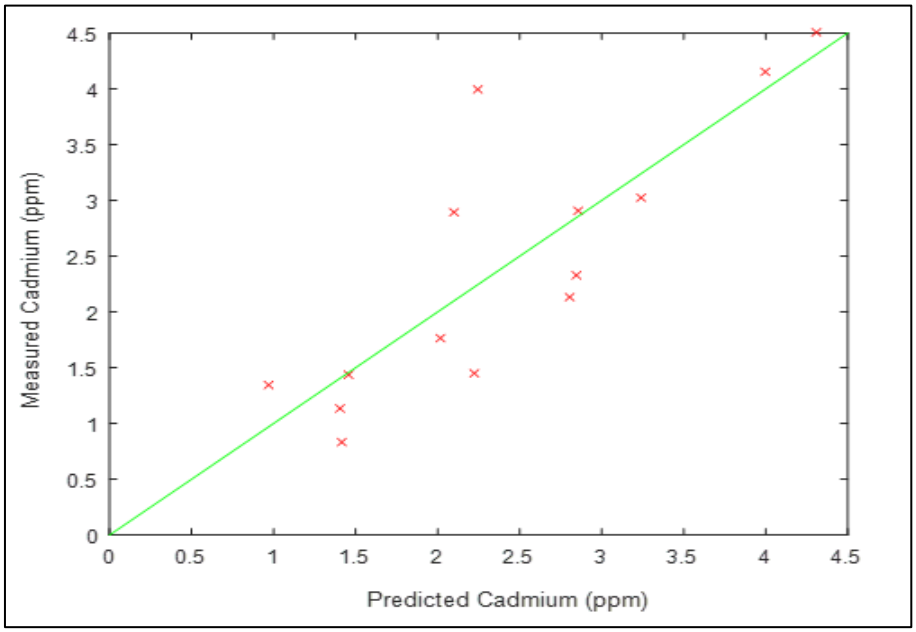
Figura 76. Aproximación con PLS de N°Comp=3.



Fuente. Elaboración propia.

Así mismo, se puede visualizar en la figura 77, el resultado de la predicción con la data de entrenamiento para número de componentes 3.

Figura 77. Resultado de la predicción con PLS (Training set) N°comp=3.



Fuente. Elaboración propia.

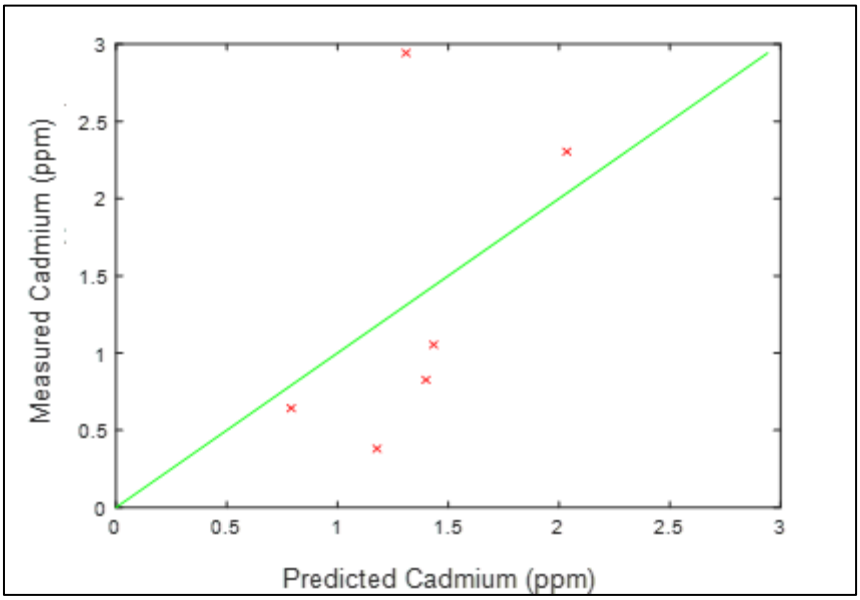
El cálculo de error para este caso es de:

MSE: 0.2899

RMSE: 0.5384

Adicionalmente, se calculó el error obtenido en la predicción con la data de validación cruzada, obteniendo como error: MSE: 0.6158 / RMSE: 0.7847. En la figura 78 se visualiza la aproximación de la predicción.

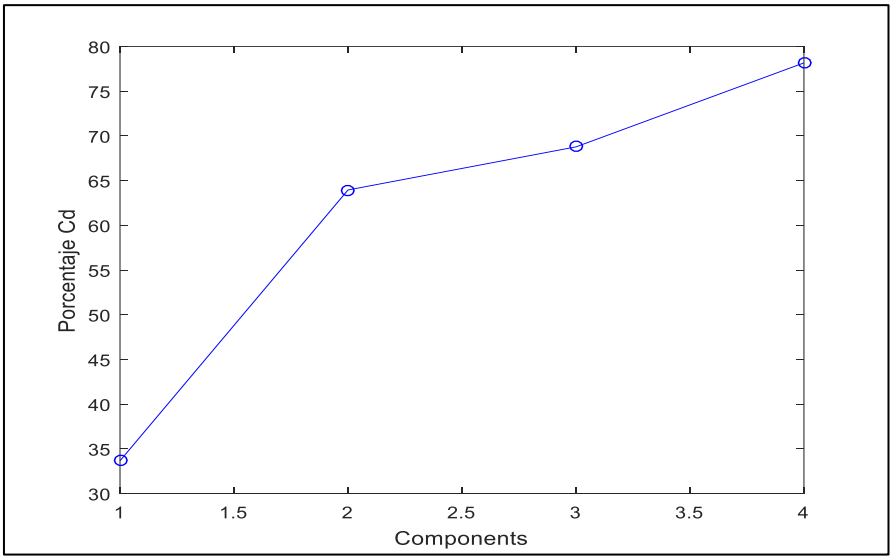
Figura 78. Resultado de validación cruzada con N°comp = 3.



Fuente. Elaboración propia.

El mismo análisis se realizó para la modelación de validación cruzada con 4 componentes. En la figura 79 se ve el incremento en la efectividad del sistema con un componente más.

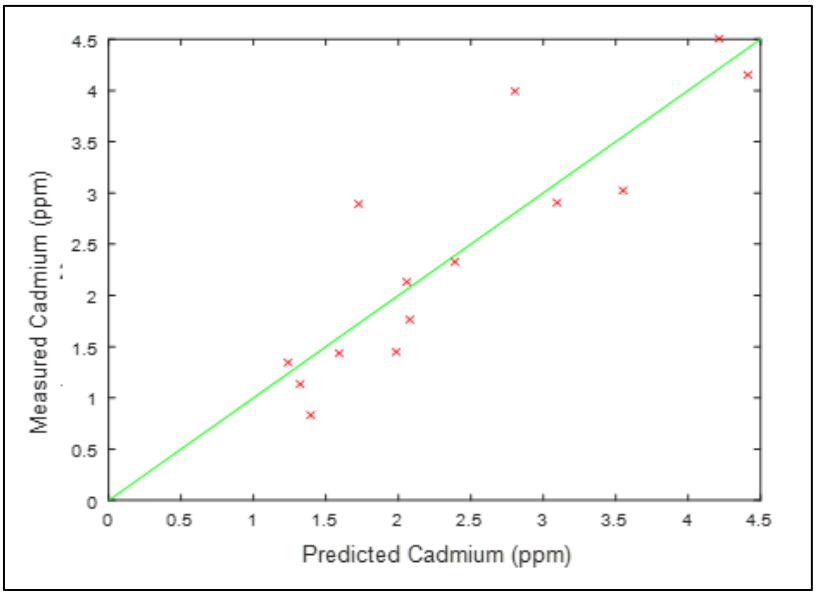
Figura 79. Aproximación con PLS de N°Comp=4.



Fuente. Elaboración propia.

El resultado gráfico es visible para la data de entrenamiento en la figura 80

Figura 80. Resultado de la predicción con PLS (Training set) N°comp=4.



Fuente. Elaboración propia.

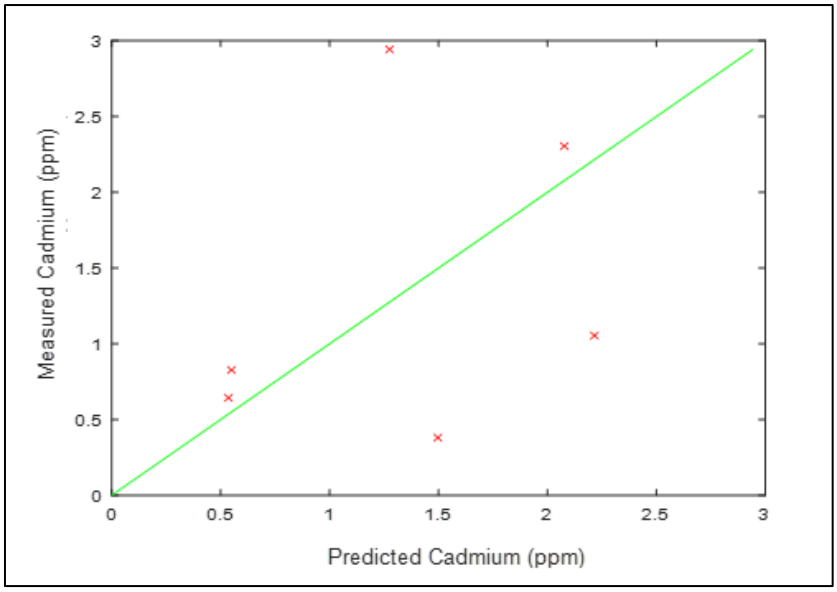
Como se pudo prever, el cálculo de error para este caso es menor que en el anterior, con valores de:

MSE: 0.2027

RMSE: 0.4502

Sin embargo, el resultado para la predicción de Validación Cruzada presenta un error ligeramente mayor, con errores de: MSE: 0.87 y RMSE: 0.9327. Gráficamente, se puede apreciar la predicción en la figura 81.

Figura 81. Resultado de validación cruzada con N°comp = 4.



Fuente. Elaboración propia.

4.6 Análisis del método de predicción

Un concepto importante en *Machine Learning* es la compensación de *Bias-Variance*, que nos ayudará a entender el funcionamiento de la modelación que hemos aplicado y cómo podría mejorarse.

Error de Bias. Este tipo de error se da cuando los modelos de predicción son muy simples para la complejidad del conjunto de datos. Este caso se da por ejemplo cuando se intenta ajustar a una regresión lineal un conjunto de datos que no presenta un patrón lineal, por ejemplo, curvas de mayor grado. Los algoritmos con alto *Bias* son rápidos para aprender y entender, pero no son flexibles para las predicciones futuras, que genera un menor rendimiento en la predicción. Dentro de *Machine Learning,* los algoritmos con Bajo *Bias*, suelen ser: *Support Vector Machines* y Árboles de Decisión, mientras que los de Alto *Bias* suelen ser: Regresión lineal, regresión logística.

Error de Varianza. Este error refiere a la calidad de la estimación de la función objetivo utilizando diferentes datos de entrenamiento. Una baja Varianza requiere de pequeños cambios en la estimación de la función objetivo.

El objetivo en el diseño de algoritmos con *Machine Learning* es generar uno que presente Bajo Bias y Baja Varianza.

En resumen, los modelos que generan alto *Bias,* no son lo suficientemente complejos para la data y tienden a converger rápidamente, mientras que los modelos con Alta Varianza, sobre ajustan los datos de entrenamiento.

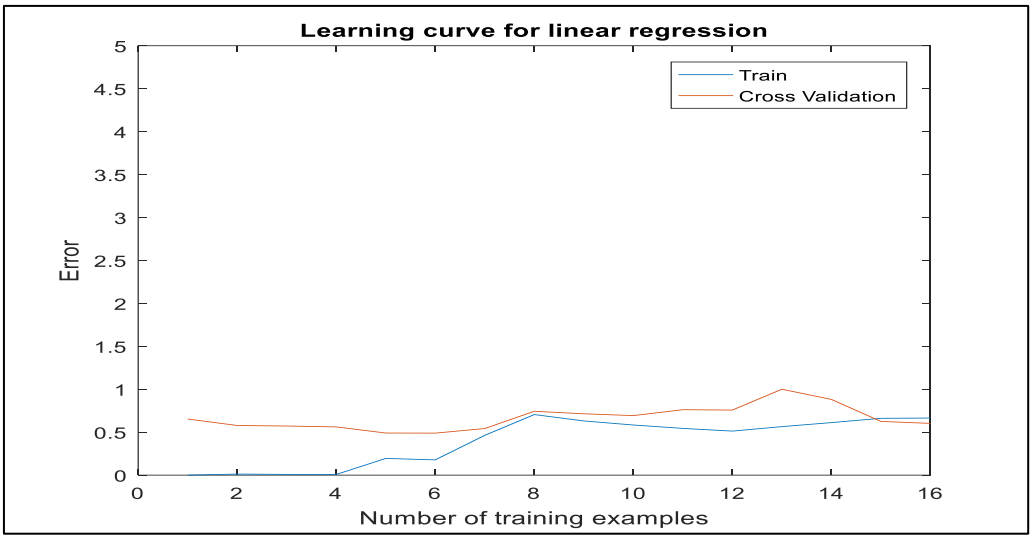
Es posible implementar un código para generar curvas de aprendizaje que serán muy útiles para depurar los algoritmos de aprendizaje. Para ello se requiere definir conjuntos de datos de aprendizaje y de validación cruzada. El error para el entrenamiento es definido por la fórmula siguiente:

$$J_{train}(\theta) = \frac{1}{2m}\left[\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2\right]$$

Aplicaremos este criterio para evaluar la calidad del algoritmo de regresión analizado.

Para Regresión Lineal se obtuvieron las curvas de aprendizaje mostradas en la figura 82.

Figura 82. Evolución del error según conjuntos de datos de entrenamiento.



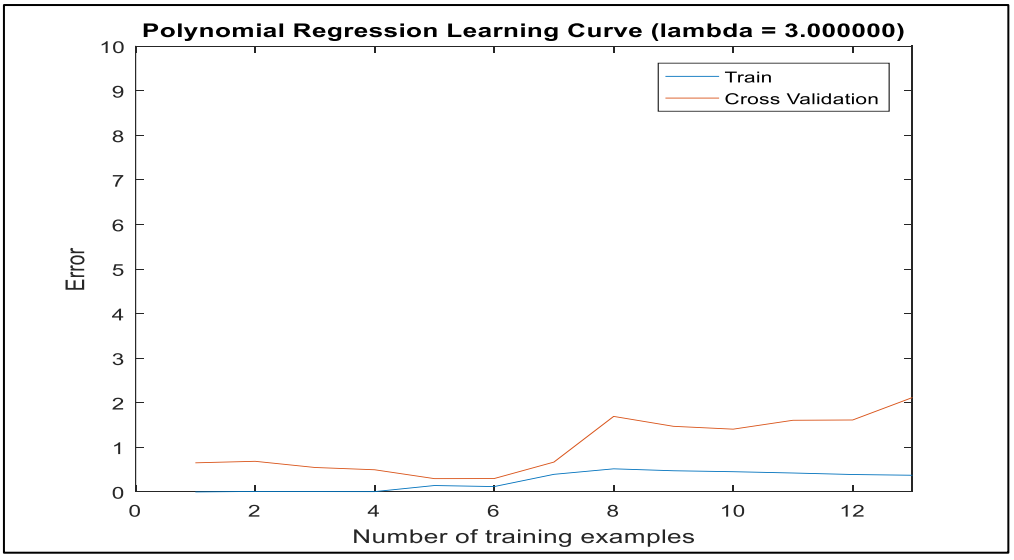
Fuente. Elaboración propia.

Podemos ver en la figura 82, que por más ejemplos o data de entrenamiento, el error no reducirá más de manera significativa. (El modelo lineal es simple, pero el error generado tampoco es muy grande, presenta ligero Bias, pero atiende para nuestra modelación).

Se evaluará adicionalmente cómo se comporta la Regresión Polinomial comparado con la Regresión lineal para la predicción de nuestros datos.

Para este caso, modelará con un polinomio de regresión de grado 8 y luego de grado 6. Podemos ver los resultados de la aplicación de un polinomio de grado 8 en la figura 83.

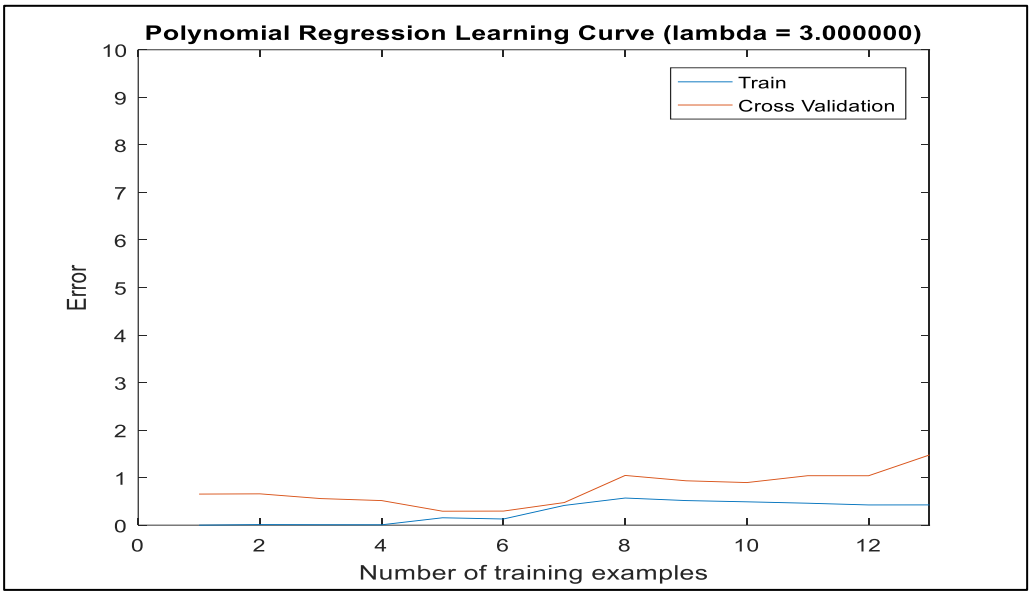
Figura 83. Curva de Aprendizaje por cantidad de datos entrenamiento Reg.polinomial (P=8).



Fuente. Elaboración propia.

El resultado de la modelación con grado 6 podemos visualizarlo en la figura 84.

Figura 84. Curva de aprendizaje por cantidad de datos entrenamiento. Reg. polinomial (P=6).

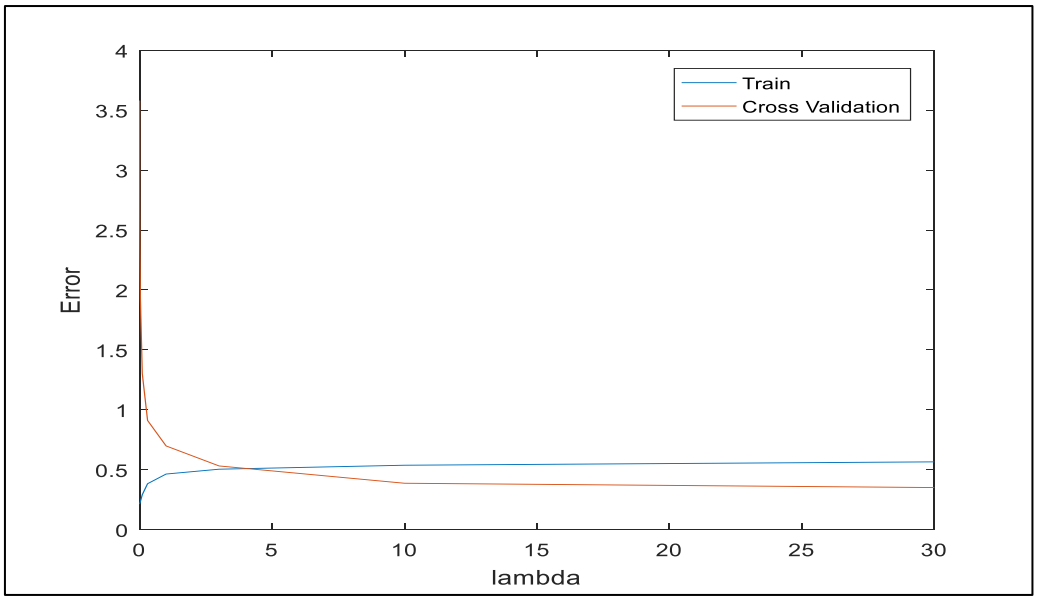


Fuente. Elaboración propia.

Como podemos apreciar, en ambos casos conseguimos una buena modelación, con un error que podría considerarse bajo para la cantidad de datos de entrenamiento de prueba.

Para poder definir la velocidad de convergencia más adecuada para nuestra modelación, debemos evaluar el valor correcto de Lambda, y esto lo podemos visualizar en la figura 85, en la que podemos destacar que con un valor de lamba de 10, conseguimos el menor error para la Validación Cruzada.

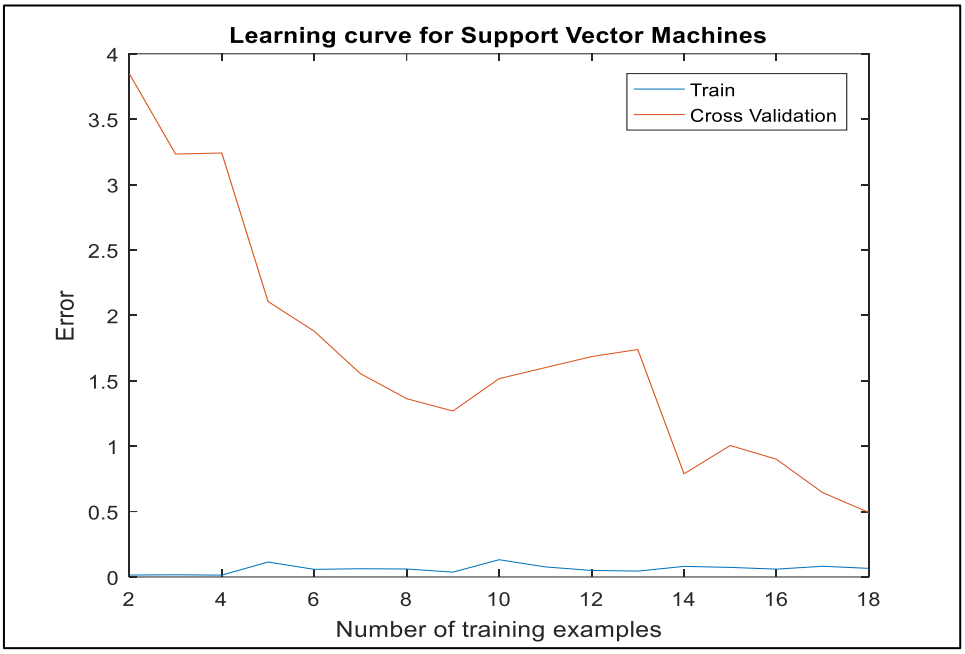
Figura 85. Error según valor de lambda.



Fuente. Elaboración propia.

Aplicando el concepto de *Bias-Variance*, ahora para la modelación con *Support Vector Machines*, obtenemos el resultado mostrado en la figura 86, en la que podemos ver que a mayor cantidad de datos de entrenamiento, el modelo tiende a converger. Esto nos indica que la modelación no presenta alto *Bias*.

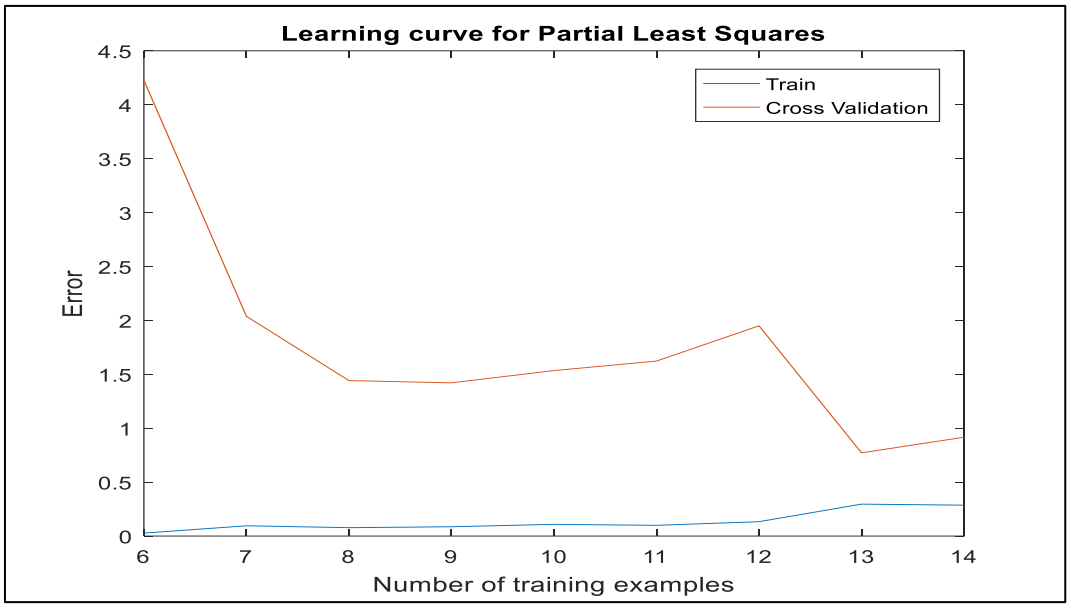
Figura 86. Curva de aprendizaje según cantidad de datos entrenamiento (SVM).



Fuente. Elaboración propia.

La aplicación del mismo concepto para la modelación con *Partial Least Squares*, nos genera el resultado de la figura 87, en la que se visualiza que el error tiende a converger a medida que la data de entrenamiento incrementa.

Figura 87. Curva de aprendizaje según cantidad de datos entrenamiento (PLS).



Fuente. Elaboración propia.

Como hemos podido notar (con la convergencia en el incremento de datos de entrenamiento), los modelos escogidos para realizar las predicciones mejorarán su calidad de predicción con mayor entrenamiento, por lo que presentan bajo *Bias* y son adecuados para esta situación.

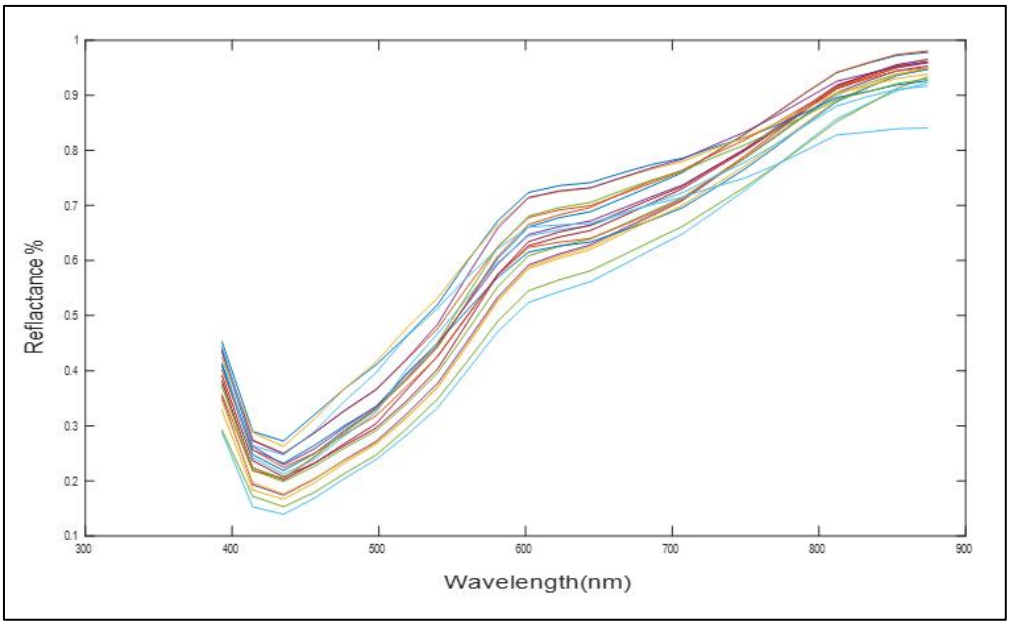
4.7 Reducción de características o datos de entrada (features)

Muchas veces la velocidad de desarrollo de un algoritmo o cantidad de datos de entrada puede entorpecer el aprendizaje del modelo en lugar de favorecerlo. En el modelo que hemos trabajado en los apartados anteriores, la matriz de características presentaba 240 lecturas de reflectancia para sus respectivas longitudes de onda.

Lo que probaremos en los siguientes puntos, es cómo varía la aproximación de las salidas con una reducción importante de características o lecturas de reflectancia, usando solo el 10% de las tomas de longitud de onda.

En la figura 88 se podrá visualizar la gráfica de características reducidas para las firmas hiperespectrales.

Figura 88. Datos filtrados (24 características).



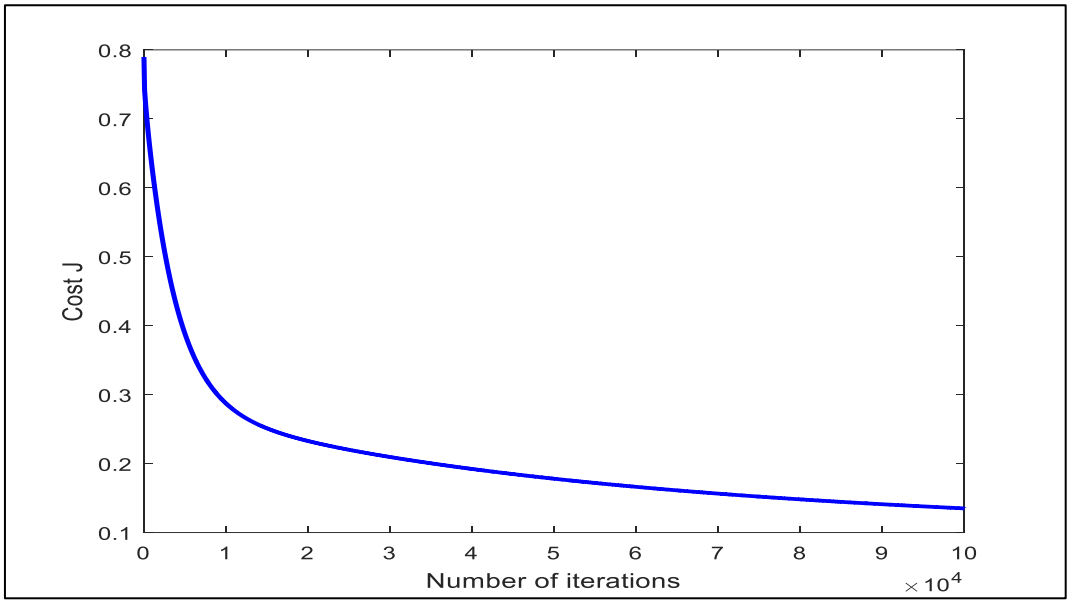
Fuente. Elaboración propia.

La forma de la firma hiperespectral mantiene su esquema (similar al de 240 características), sin embargo, es necesario probar cómo reaccionan las predicciones de salida.

4.7.1 Regresión Multivariable

Los resultados con menor cantidad de características para regresión multivariable, permite la convergencia de la función de Costo, como se muestra en la figura 89.

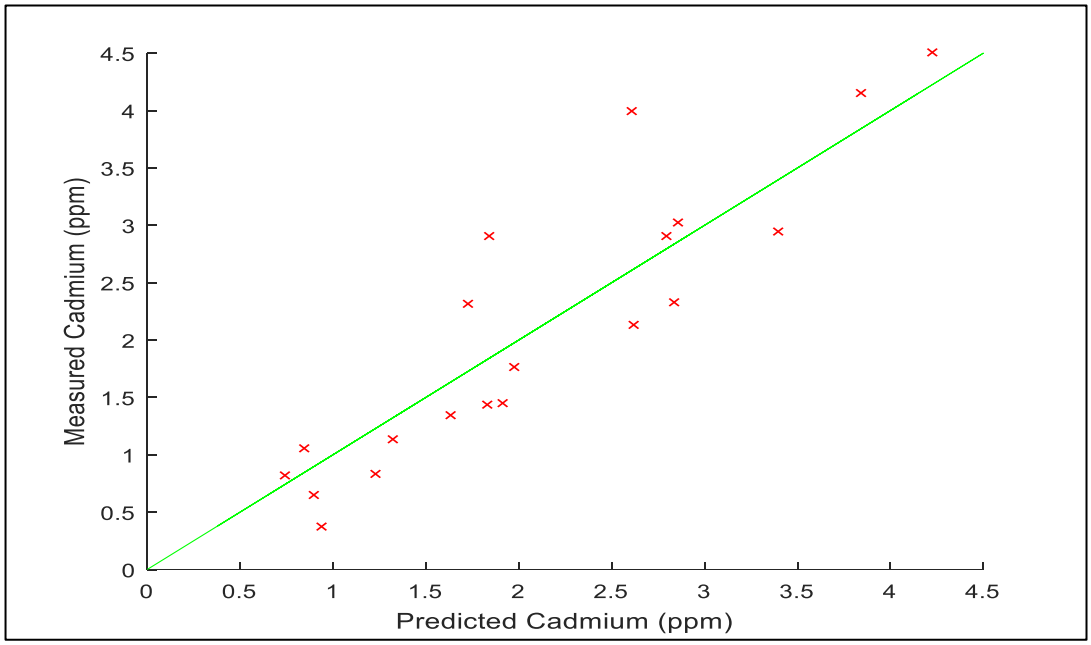
Figura 89. Iteraciones para convergencia de función de costo (alpha=0.1).



Fuente. Elaboración propia.

Los resultados con los datos de entrenamiento y verificación nos generan los resultados mostrados en la figura 90, en la que se puede visualizar una predicción de valores bastante buena, muy cercana inclusive a los obtenidos con todas las características de la matriz de entrada.

Figura 90. Salida real y salida prevista (alpha=0.1).



Fuente. Elaboración propia.

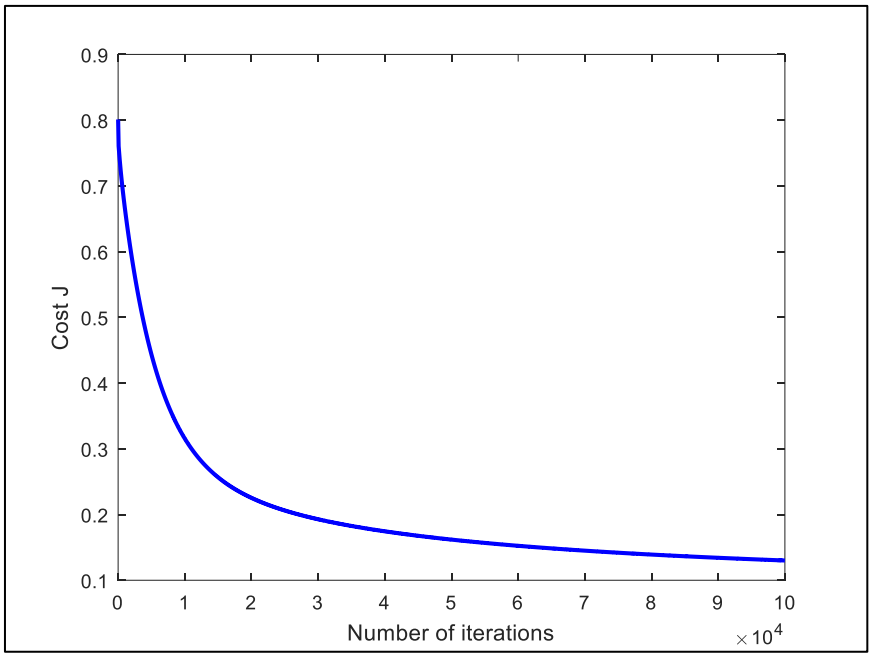
Los valores de error incrementan solamente en 10% respecto del error medido con todas las características. Esto es:

MSE= 0.27

RMSE= 0.5197

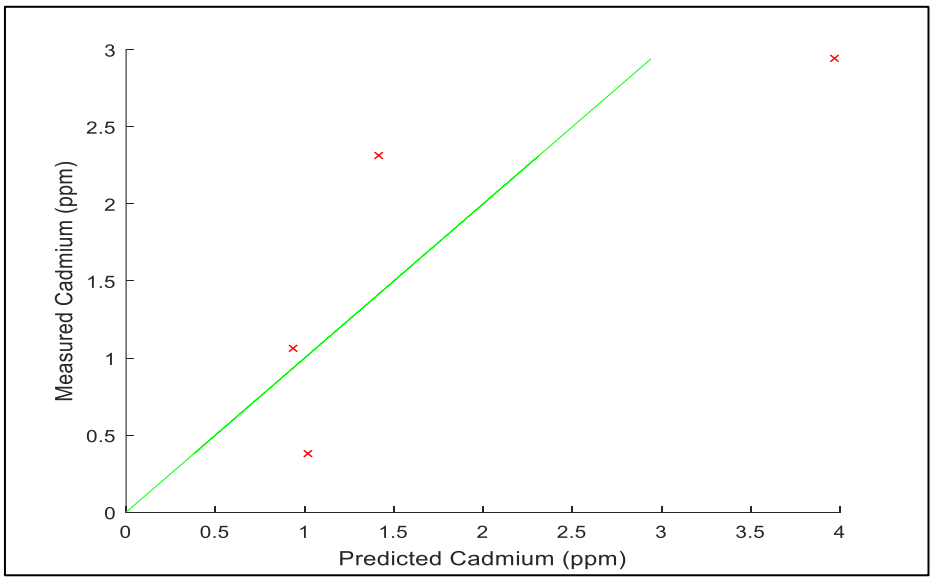
4.7.1.1 Validación cruzada de regresión multivariable. Una vez asignados los grupos de data, para entrenamiento y para validación, ejecutamos el algoritmo para visualizar la función de costo (figura 91). Respecto a los valores de predicción, podemos notar tanto gráficamente como con el cálculo de error, que los valores de salida son menos precisos que con la consideración de todas las características.

Figura 91. Iteraciones para convergencia de la función de Costo.



Fuente. Elaboración propia.

Figura 92. Valores de predicción y valores reales.



Fuente. Elaboración propia.

La medida del error para esta predicción es:

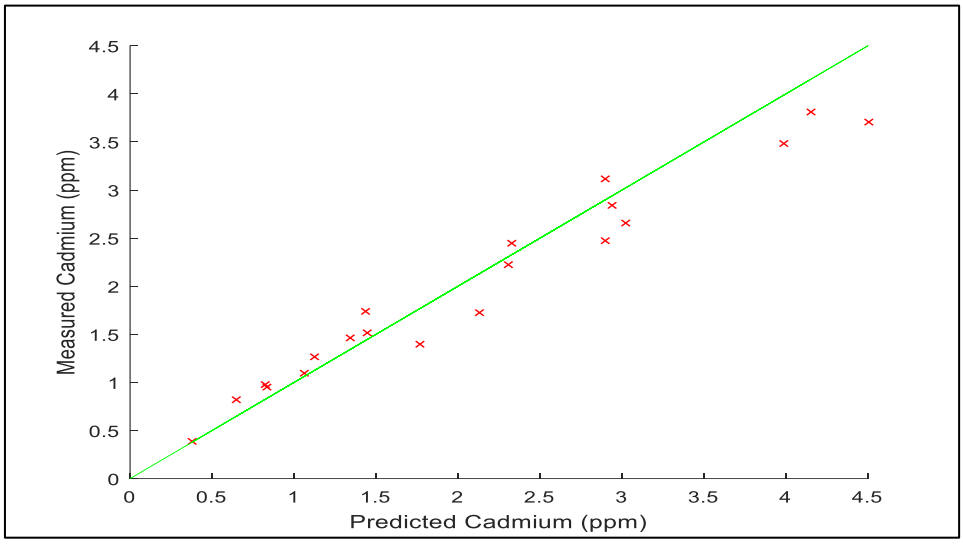
MSE = 0.5705

RMSE = 0.755

4.7.2 Support Vector Machines

Para SVM, encontramos los resultados mostrados en la figura 93, en la que se muestra un buen nivel de predicción de valores de salida.

Figura 93. Predicted output vs actual output.



Fuente. Elaboración propia.

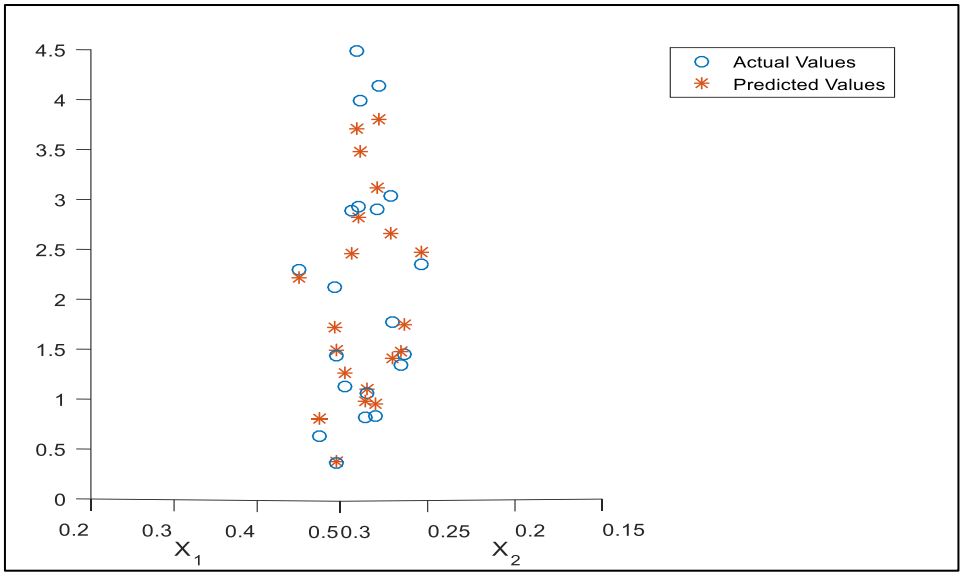
El error para esta metodología de predicción es:

MSE=0.0949

RMSE = 0.308.

Gráficamente, la precisión de valores predichos y reales se muestra en la figura 94.

Figura 94. Valores reales y valores predichos por modelación SVM.



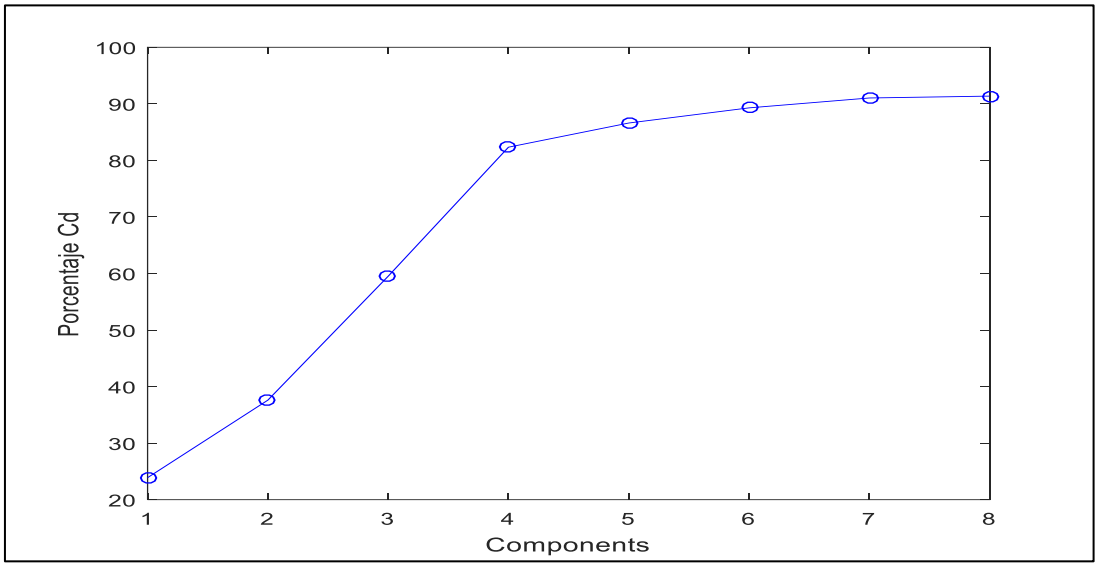
Fuente. Elaboración propia.

4.7.3 Regresión con Partial 2Least Squares:

Aplicando regresión con PLS con data de entrada con características reducidas, obtenemos resultados similares a los anteriores:

Para una proyección de 6 a 8 componentes, el porcentaje de precisión ya es bueno para nuestro caso.

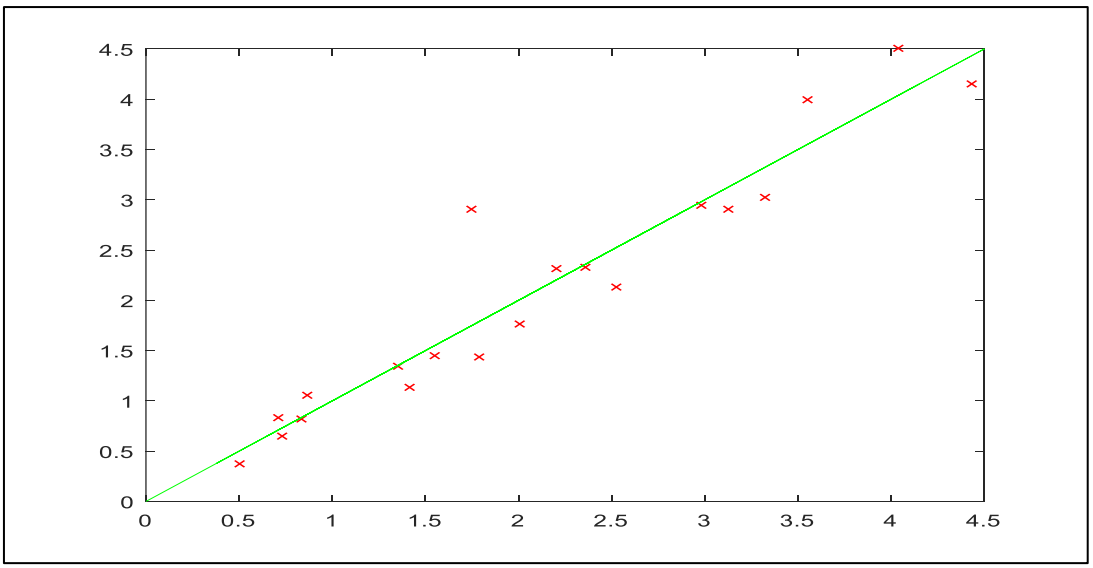
Figura 95. Precisión de resultados según cantidad de componentes.



Fuente. Elaboración propia.

Se identifica visualmente (figura 96), que los resultados de la predicción son bastante aproximados a los reales.

Figura 96. Predicción usando Regresión Polinomial.



Fuente. Elaboración propia.

El error medido para esta metodología de predicción es:

MSE= 0.1223

RMSE=0.3498

4.8 Comparativo de resultados

En las tablas comparativa a continuación N°11 y N°12, se mostrará, en resumen, los valores de error obtenidos en cada una de las metodologías de predicción de salidas, tanto para los datos de entrenamiento como para la Validación o prueba con datos diferentes a los de entrenamiento.

Tabla 11. *Resultados comparativos base al error de predicción.*

PREDICCIÓN DE DATOS DE ENTRENAMIENTO						Reducción de características: 24 features		
ERROR	REGRESIÓN LINEAL MULTIVARIABLE	ECUACIÓN NORMAL	SVM	PLS k=6	PLS k=8	REGRESIÓN LINEAL MULTIVARIABLE	SVM	PLS
MSE	0.2399	4.53E-18	0.053	0.1623	0.125	0.27	0.0949	0.1223
RMSE	0.4898	2.13E-09	0.23	0.4028	0.354	0.52	0.308	0.35

Fuente. Elaboración propia.

Tabla 12. *Resultados comparativos base al error de predicción.*

PREDICCIÓN DE DATOS POR VALIDACIÓN CRUZADA						24 features
ERROR	REGRESIÓN LINEAL MULTIVARIABLE	ECUACIÓN NORMAL	SVM	PLS k=3	PLS k=4	Regresión lineal multivariable
MSE	0.2234	5.47	0.89	0.2899	0.2027	0.5705
RMSE	0.4727	2.34	0.94	0.5384	0.4502	0.755

Fuente. Elaboración propia.

Conclusiones

El resultado de la modelación para la predicción del nivel de cadmio en el cacao, utilizando algoritmos de *Machine Learning* resultó exitosa, pues se consiguió un alto nivel de predicción pese a la baja cantidad de muestras de entrada.

La Ecuación Normal, sin embargo, por más que los resultados de entrenamiento generen muy bajo error en su predicción, cuando se aplica el modelo para data de validación, la distorsión es muy alta, debido a que el modelo presenta Alta Varianza.

Para nuestra aplicación, la modelación con Redes Neuronales presenta un problema, requiere de muchos datos de entrenamiento para conseguir resultados correctos, y debido a nuestra elevada cantidad de características (reflectancia por longitud de onda), la necesidad de data es mucho mayor para conseguir resultados positivos (por lo menos mil datos de entrenamiento).

Podemos destacar que, dentro de las pruebas realizadas, se consiguió el menor error en los resultados de validación para *Partial Least Squares*, seguido por la validación de Regresión Lineal. Sin embargo, evaluando la predicción con los datos de entrenamiento, el mejor resultado se obtiene con *Support Vector Machines*.

A excepción de la modelación con Ecuación Normal, El análisis de Bias y Varianza otorga buenos resultados para los modelos aplicados para nuestro caso. En todos obtuvimos convergencia a mayor cantidad de datos de entrenamiento.

La modelación polinomial genera una mejor predicción que la modelación con regresión lineal.

Como se esperaba por la forma de las imágenes hiperespectrales, usando una cantidad de características reducidas, obtendremos modelos simples, con resultados también bastante buenos. Esto es, para el caso, un incremento de 12% en el error usando 90% menos de características, puede ser un costo importante para mejorar la velocidad de desarrollo del modelo. Esto debido a que a pesar de que las entradas parecen curvas paralelas, ellas no son proporcionales.

Adicionalmente, se ha demostrado para la situación que, con una reducción en el número de características, la modelación se hace más sencilla y requiere menos iteración para la convergencia. Sin embargo, con mayor cantidad de data, podemos incrementar el rango de acción del modelo.

Referencias bibliográficas

[1] Alciaturi C, Escobar M, De la Cruz C, Rincón C, 2003 “La Regresión de mínimos cuadrados parciales (PLS) y su aplicación al análisis del carbón mineral”.

[2] Alterini T, Díaz-Doutón F, 2019. “Fast visible and extended near-infrared multispectral fundus camera”

[3] Arévalo-Gardini E, Meyier E., Zúñiga L, Arévalo-Hernández C, Baligar V. y Zhenli He. 2016 “Metales pesados en suelos de plantaciones de cacao (Theobroma cacao L.) en tres regiones del Perú”.

[4] Bodkin, Sheinis, 2009. “Snapshot hyperspectral imaging: The hyperpixel array camera”.

[5] Bonafini B. 2018 “Um estudo sobre reconhecimento de padrões aplicado a detecção de câncer do tipo melanoma maligno”

[6] CAOBISCO/ECA/FCC 2015, “Cocoa Beans: Chocolate and Cocoa Industry Quality Requirements”.

[7] Chakraborty K, 2019 “A complete guide to support vector machines”

[8] Checa K, Gamarra M, Torres C, Soto J. 2019 “Relación entre el contenido de Cadmio y la firma hiperespectral de granos de Cacao peruanos”.

[9] Comisión CODEX Alimentos, 2019. Programa conjunto FAO/OMS sobre normas alimentarias.

[10] Cruces, E. 2016. “Los neonicotinoides y su uso seguro en la agricultura”.

[11] De la Cruz M, Vargas A, 2010. “CACAO: Operaciones poscosecha”.

[12] Duda R.O, 2015 “Pattern Classification”. Second Edition

[13] Espín M, 2020. “La cascara de los granos de cacao”.

[14] Goetz A, Vane G, Solomon J, Rock B, 1985 “Imaging Spectrometry for Earth Remote Sensing”.

[15] Gonzáles G, 2018. El espectro detrás de la información

[16] Gonzáles R, Woods R, 2013, Digital Image Processing.

[17] Huamán O. 2019. Observatorio de Commodities: Cacao”

[18] Kalivas, John H. 1997. “Two data sets of near infrared spectra”.

[19] Loor, Motamayor, Schnell et. al 2008. Geographic and Genetic Population Differentiation of the Amazonian Chocolate Tree.

[20] Lutheran World Relief, 2017. “Manejo sostenible de cultivo de Cacao”
[21] MacDonald J., Ustin S, Schaepman M, 2008 “The Contributions of Dr. Alexander Goetz to imaging spectrometry”.
[22] Mañas A, 2019. “Pronóstico del flujo de tráfico en la ciudad de Madrid”.
[23] Martínez W, 2020 “Desarrollo de un prototipo para la medición de partículas PM2.5 en la ciudad de Bogotá”
[24] Melville P, Sindhwani V, 2010 “Recommender Systems”
[25] Meter A., Atkinson RJ, Laiberte B., 2019 “Cadmio en el cacao de América Latina y el Caribe – Análisis de la investigación y soluciones potenciales para la mitigación”.
[26] Michael O. Ngadi, Li Liu, 2010 “Hyperespectral Image Processing Techniques” Chapter 4.
[27] Navarrete G, 2019. “Análisis y desarrollo de una aplicación para predecir toneladas de caña de azúcar anual mediante redes neuronales artificiales.
[28] Nestlé, 2017 “El origen del Chocolate”, “Cómo se cultiva el Cacao”.
[29] Nils N, 1998. “Introduction to Machine Learning”.
[30] Pazmina C, 2019. “Polvo de Cacao: Por qué y cómo usarlo”. Minka-Ecuador.
[31] Pérez P, 2012 “Los efectos del cadmio en la salud”.
[32] Ramírez F, 2018 “Historia de la IA”.
[33] Ramos E, 2019. “Exportaciones de manteca de cacao alcanzaron los US$ 4millones durante primer bimester”.
[34] Rodríquez-Serrano M., Martínez-de la Casa N., Romero-Puertas, M.C., L.A. del Río, L.M. Sandalio, 2008, “Toxicidad del Cadmio en Plantas”.
[35] Roman-Gonzales A, 2013. Plataforma de Energy & Geoscience Institute (EGI). “Análisis de imágenes hiperespectrales”.
[36] Romero, C. 2016 “Estudio del Cacao en Perú y en el Mundo”.
[37] Roobini M.S., 2018 “Comparative study of data mining methodologies for prediction of parkinson disease by statical methods.
[38] Ropero A, 2016 “Cacao y chocolate”
[39] Samuel A, 1959. "Some Studies in Machine Learning Using the Game of Checkers". IBM Journal of Research and Development.
[40] Sánchez A, 2015. “Desarrollo de técnicas de visión hiperespectral y tridimensional para el sector agroalimentario”
[41] Shalev-Shwartz, Shai, Ben-David, Shai, 2014. "Decision Trees". Understanding Machine Learning. Cambridge University Press.
[42] SpectrononPro, 2020, Datasheet
[43] Thyssen, Keil, Wolff, 2018 “Bioimaging of the elemental distribution in cocoa beans by means of LA-ICP-TQMS.
[44] Ticono P, 2018. “La pulpa de cacao: un descubrimiento en beneficio de la salud”.
[45] Tkachenko P, 2019. “Curso Machine Learning – Neural Networks”.
[46] Tkachenko P, 2019. “K-means clustering”.

[47] UE, 2014, “Reglamento N°488/2014 que modifica al reglamento (CE) 1881/2006 por lo que respecta al contenido máximo de cadmio en los productos alimenticios”

[48] Van Acoleyen, K. 2019. “Machine Learning with tensor networks”

[49] Viera-Maza, 2018 “Aplicación de Procesamiento de Imágenes para Clasificación de Granos de Cacao según su color intenso”.

[50] Villatoro F, 2016. “Newton y la dualidad onda-corpúsculo para la luz”.

[51] Wu and Sun (2013). Modes of hyperspectral data acquisition.

[52] Xiaona Li, 2017. “Hyperspectral Imaging and their Applications in Nondestructive Quality Assesment of Fruits and Vegetables”